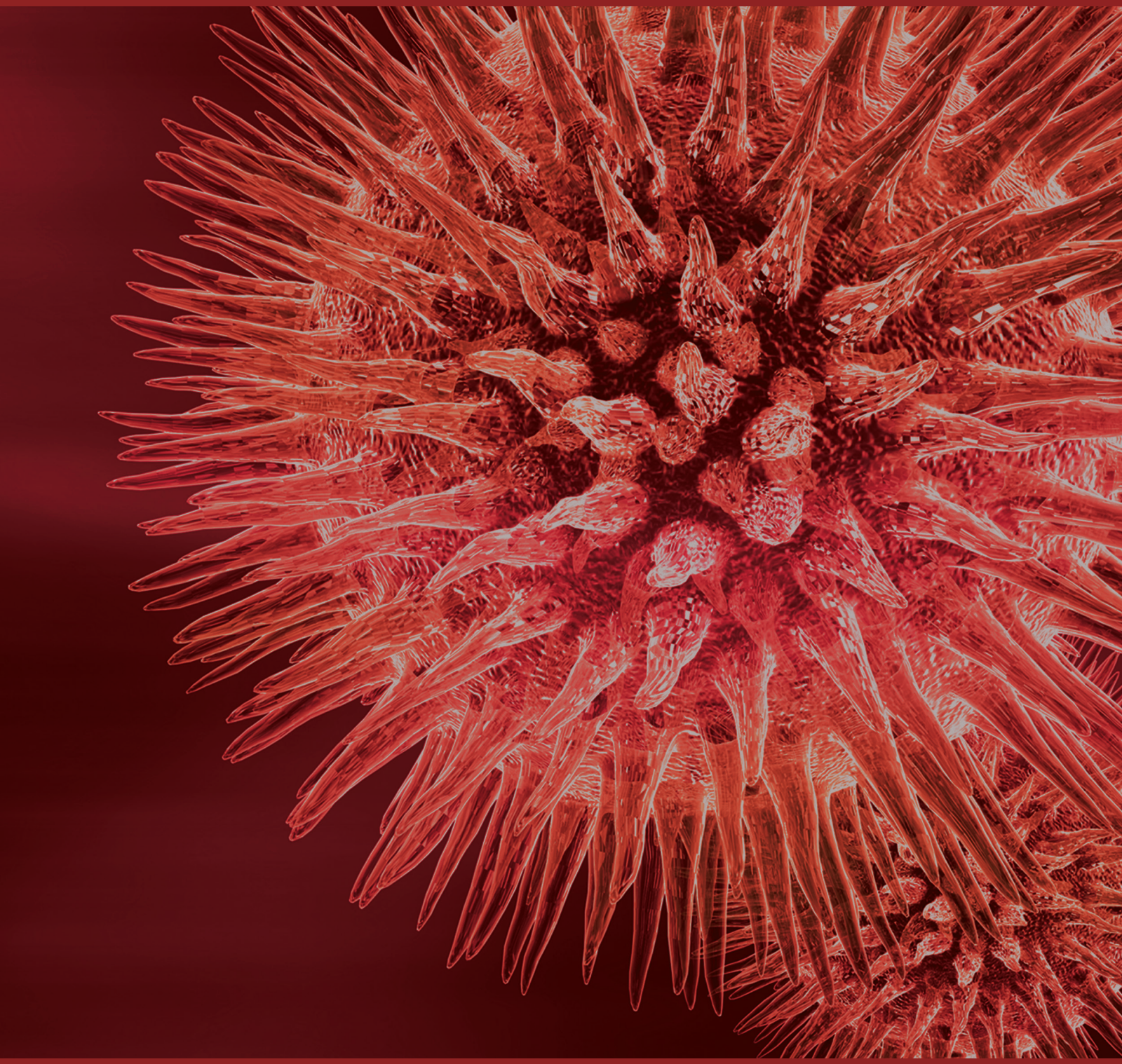
BioMed Research International

Molecular Phylogenetics 2014

Guest Editors: Vassily Lyubetsky, William H. Piel, and Peter F. Stadler



Molecular Phylogenetics 2014

BioMed Research International

Molecular Phylogenetics 2014

Guest Editors: Vassily Lyubetsky, William H. Piel, and Peter F. Stadler

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 919251, 2 pages
http://dx.doi.org/10.1155/2015/919251

Editorial
Molecular Phylogenetics 2014

Vassily Lyubetsky,[1] William H. Piel,[2] and Peter F. Stadler[3]

[1]*Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, 127994 Moscow, Russia*
[2]*Yale-NUS College & National University of Singapore, Singapore 138614*
[3]*Institut für Informatik, University of Leipzig, 04109 Leipzig, Germany*

Correspondence should be addressed to Vassily Lyubetsky; lyubetsk@iitp.ru

Received 14 April 2015; Accepted 14 April 2015

Knowledge of phylogeny is of fundamental importance for understanding evolution. It has become an indispensable tool in modern genomics as a framework for interpreting genomes and metagenomes, for understanding evolution of genes, proteins, and noncoding RNAs as well as different types of regulations including secondary RNA and protein structures, and for reconstructing ancestral genomes [1]. The era of next-generation sequencing (NGS) brought an influx of data but posed theoretical challenges, for example, in reducing systematic errors and increasing robustness under much less sampling error [2].

An important avenue of modern phylogenetics is the study of coevolution. Here a major focus is the developing of effective algorithms to infer common history of genes, proteins, molecular machines (ribosomes, RNA polymerases, and transporter systems), signal transductions, metabolic pathways, chromosomal structures and gene synteny, and so forth. An important direction is to define and reconstruct the evolutionary scenario on the basis of polytomous trees, also in application to study fast evolving bacteria and viruses with high medical impact, and in shaping the primary genomes. Dating of the phylogeny immediately depends on progress in detecting zones of active horizontal gene transfers or other genomic events, defining time slices to describe tree-like phylogenies, and so forth.

Mathematics of phylogenetics will grow to foster alternative ways to describe evolution: generalization of trees into nets and developing models based on stochastic and partial differential equations. An important focus is to develop low polynomial complexity algorithms for exact models that are usually solved heuristically. Contemporary methods of clustering and discrete optimization have already proved very effective. Thus, graphs with 10^{20} vertices, 10^9 parts, and the same average vertex degree can be processed on a multiprocessor machine in reasonable time. A high priority for testing modern algorithms is to obtain primary data with known predefined solutions. The call is for realistic computational models that allow simulating large-scale phylogenies and can serve for future studies, as well as real data with known phylogenies or at least few data sets where everybody would agree what the correct phylogeny is.

The contents of the special issue of 2015 cover research of various groups that use the toolkit of phylogenetics to tackle a spectrum of evolutionary questions. Apart from classic molecular systematic applications to infer taxon phylogenies, the trend is obvious to approach molecular and biodiversity assessment at different levels in various communities, at intraspecific level and in environmental samples, including systematic studies of bacterial and viral pathogenic agents. Molecular markers such as mobile elements are being developed and exploited in studies of population polymorphisms, and RNA secondary structures are used to detect signatures of selection.

A historical profile of molecular phylogenetics with some extrapolations into the future, as well as a brief outline of hot spots in this field, can be found in this special issue [3]. We truly hope that these contributions will be of use to scientists in various areas in possibly helping them to find answers and pose new questions in their own research.

Vassily Lyubetsky
William H. Piel
Peter F. Stadler

References

[1] E. V. Koonin, *The Logic of Chance: The Nature and Origin of Biological Evolution*, FT Press, 2011.

[2] J. W. Wägele and T. Bartolomaeus, Eds., *Deep Metazoan Phylogeny: The Backbone of the Tree of Life: New Insights from Analyses of Molecules, Morphology, and Theory of Data Analysis*, Walter de Gruyter, Berlin , Germany, 2014.

[3] V. Lyubetsky, W. H. Piel, and D. Quandt, "Current advances in molecular phylogenetics," *BioMed Research International*, vol. 2014, Article ID 596746, 2 pages, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 452958, 6 pages
http://dx.doi.org/10.1155/2015/452958

Research Article

Comparative Analysis of Apicoplast-Targeted Protein Extension Lengths in Apicomplexan Parasites

Alexandr V. Seliverstov, Oleg A. Zverkov, Svetlana N. Istomina, Sergey A. Pirogov, and Philip S. Kitsis

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoy Karetny Pereulok 19, Moscow 127994, Russia

Correspondence should be addressed to Alexandr V. Seliverstov; slvstv@iitp.ru

Received 10 September 2014; Accepted 25 December 2014

Academic Editor: William H. Piel

In general, the mechanism of protein translocation through the apicoplast membrane requires a specific extension of a functionally important region of the apicoplast-targeted proteins. The corresponding signal peptides were detected in many apicomplexans but not in the majority of apicoplast-targeted proteins in *Toxoplasma gondii*. In *T. gondii* signal peptides are either much diverged or their extension region is processed, which in either case makes the situation different from other studied apicomplexans. We propose a statistic method to compare extensions of the functionally important regions of apicoplast-targeted proteins. More specifically, we provide a comparison of extension lengths of orthologous apicoplast-targeted proteins in apicomplexan parasites. We focus on results obtained for the model species *T. gondii*, *Neospora caninum*, and *Plasmodium falciparum*. With our method, cross species comparisons demonstrate that, in average, apicoplast-targeted protein extensions in *T. gondii* are 1.5-fold longer than in *N. caninum* and 2-fold longer than in *P. falciparum*. Extensions in *P. falciparum* less than 87 residues in size are longer than the corresponding extensions in *N. caninum* and, reversely, are shorter if they exceed 88 residues.

1. Introduction

In general, the mechanism of protein translocation through the apicoplast membrane requires a specific extension of a functionally important region of the apicoplast-targeted proteins. In *T. gondii* signal peptides are either much diverged or their extension region is processed, which in either case makes the situation different from other studied apicomplexans. We propose a statistic method to compare extensions of the functionally important regions of apicoplast-targeted proteins. More specifically, we provide a comparison of extension lengths of orthologous apicoplast-targeted proteins in apicomplexan parasites. We ground on the notion that the majority of cyanobacterial proteins lack such extensions (including signal peptides) and consist of only functional sequences.

Sporozoans comprise a monophyletic lineage of apicomplexan parasites. Among them, *Toxoplasma gondii* is an important medical and veterinary pathogen commonly causing morbidity in HIV patients [1, 2]. The study [3, 4] describes the propagation mechanism of *T. gondii* in various hosts worldwide, including several aquatic mammal species, where it may provoke abortion and lethal systemic disease. The observation that the *apicoplast* of *T. gondii* significantly varies in shape and protein expression patterns at different life stages of the parasite suggests its important role in virulence; the apicoplast of *T. gondii* is also involved in the pathogen stage conversion and the parasite proliferation [5]. Due to bacterial origin of the apicoplast proteins, they present a natural target for selective treatment in the eukaryotic host. The sporozoans contain a semiautonomous organelle, the apicoplast, acquired by secondary endosymbiosis with ancient red algae; plastid organelles of the red algae originate from cyanobacteria [6–8].

Elucidating the molecular mechanism that underlies the role of apicoplast in the parasite invasion, conversion, and proliferation is important for development of novel therapeutics to control infection and reactivation of the parasite.

Further analysis of unique features of apicoplast-targeted proteins (particularly, regions involved in translocation processes) in *T. gondii* can add to the effective design of drug-based or genetic strategies to control the pathogen development and proliferation. At the reported stage of the study, we analyze only extensions in length of orthologs among apicomplexan parasites.

Note that the coccidian *Cryptosporidium parvum* lacks the apicoplast [9], and the apicoplast in piroplasmids *Babesia bovis* and *Theileria parva* largely differs from that in common coccidians and the haemosporidian *Plasmodium* spp. [10, 11].

The majority of apicoplast proteins are encoded in the nucleus and only few in its own genome. Most of these proteins can be identified due to their cyanobacterial origin. Transport of nuclear-encoded proteins to the apicoplast in *T. gondii* is significantly less documented experimentally compared to *Plasmodium falciparum*. Among the documented cases is the nuclear-encoded lipoic acid synthetase LipA [12]; other examples are described in [13–15]. A mechanism of protein import into secondary plastids is also described in [16], where many orthologous proteins involved in this process were shown to be presented in the sporozoans *P. falciparum* and *T. gondii*, cryptophyte alga *Guillardia theta*, and diatom *Phaeodactylum tricornutum*. Plastids also possess the bacterial system to translocate folded proteins [17, 18].

A variety of protein localization prediction methods are used to identify apicoplast-targeted proteins. Some of them utilize the notion that translocation across the four membranes surrounding the apicoplast is mediated by an N-terminal bipartite targeting sequence, a special N-terminal signal, and a transit peptide [13]. The algorithm ApicoAP described in [19] predicts apicoplast-targeted proteins containing the signal peptide, because it trains on a learning sample of signal peptide-containing proteins. Other apicoplast-targeted proteins are predicted neither with this algorithm nor with ApicoAMP [20]. The comprehensive ToxoDB database constructed using the SignalP algorithm contains proteins with the information on presence/absence of the signal peptide. According to this database, some nuclear-encoded proteins in *T. gondii* that are experimentally shown to reach the apicoplast should do not contain signal peptides, albeit bearing housekeeping functions in the apicoplast. The methods in application to *Plasmodium* spp. are described, for example, in [19, 21–23]. The PlasmoAP algorithm [24] is designed specifically for *Plasmodium* spp. and is of little applicability to coccidians. Hence, these two widely used databases may be considered of limited use to identify apicoplast-targeted proteins not containing the standard signal peptide in coccidians. We therefore applied a crude technique to compare apicomplexan proteins with their orthologs in a cyanobacterium. Namely, orthology between nuclear-encoded sporozoan proteins and cyanobacterial proteins is used as a basis to suggest the apicoplast-targeted nature of the proteins. As our study relies on statistic estimates, its predictions are hopefully not affected by the chosen parameters of global protein alignment.

In this work, the lengths of sporozoan proteins are compared with each other and with the length of their orthologs in the cyanobacterium *Synechocystis* sp. PCC 6803. We consider lengths of the sporozoan proteins that extend outside the conserved alignment region, which usually covers the entire cyanobacterial sequence. We focus on results obtained for the model species *T. gondii*, *Neospora caninum* (the two coccidian sporozoans with completed genome projects, as per the end of 2013), and the malaria agent *P. falciparum* from the Haemosporidia.

Based on the total comparison, we conclude that *T. gondii* in most cases contains longer proteins compared to both *N. caninum* and *P. falciparum*. We also surmised that at least some of them undergo processing in the cytoplasm to facilitate transporting into the apicoplast. The extended portions of proteins may also be involved in gene expression regulation at the level of protein-protein interaction.

As an argument, the regulation of plastid-encoded genes *ycf24* and *rps*4 affects the general functionality of the apicoplast in *T. gondii* [25]. The expression regulation of *ycf24* (the SufB factor mediating the Fe-S cluster assembly in many nuclear proteins) was suggested to take place in the apicomplexans *Eimeria tenella*, *T. gondii* RH, and *Plasmodium* spp., as well as in *Gracilaria tenuistipitata*, *Porphyra purpurea*, and *Porphyra yezoensis* [25]. The same type of regulation was suggested for *rps*4 (ribosomal protein S4) in *T. gondii* RH [25].

2. Materials and Methods

Protein data for *T. gondii* and *N. caninum* was extracted from the ToxoDB database (version 8.2), data for *Plasmodium* spp. from the PlasmoDB database (version 9.3), and data for *Synechocystis* sp. PCC 6803 from GenBank, NCBI [26]. ToxoDB and PlasmoDB are specialized, regularly updated, and nonoverlapping databases. Conserved domains were detected according to the Pfam database [27]. The location of regions enriched with a certain amino acid was established using the PROSITE database [28].

We compared the proteomes of three apicomplexan parasites (*T. gondii* ME49, *N. caninum* Liverpool, and *P. falciparum* 3D7) and the cyanobacterium *Synechocystis* sp. PCC 6803. For each pair of proteomes, pairs of orthologous proteins were computed on the basis of an alignment quality score using the Needleman-Wunsch method and BLOSUM62 matrix [29, 30].

Our method to study the lengths of the apicomplexan protein extensions is as follows. For each cyanobacterial protein with length s and its two orthologs (from a fixed pair of apicomplexans) with lengths a and a', the point with coordinates $(a - s, a' - s)$ is computed. In some cases, one or both of the coordinates are negative, which indicates a sporadic case of a shorter length of the sporozoan protein versus the cyanobacteria.

In Figures 1–3, the cases of *N. caninum*-*T. gondii* (*N*-*T*), *P. falciparum*-*T. gondii* (*P*-*T*), and *P. falciparum*-*N. caninum* (*P*-*N*) are analyzed using three sets of points. Each coordinate is the difference in lengths between the sporozoan and cyanobacterial orthologs: *T. gondii* versus *Synechocystis* (*T*-*S*), *N. caninum* versus *Synechocystis* (*N*-*S*), and *P. falciparum* versus *Synechocystis* (*P*-*S*). The sets of points are then statistically analyzed. The following statistic was used to test

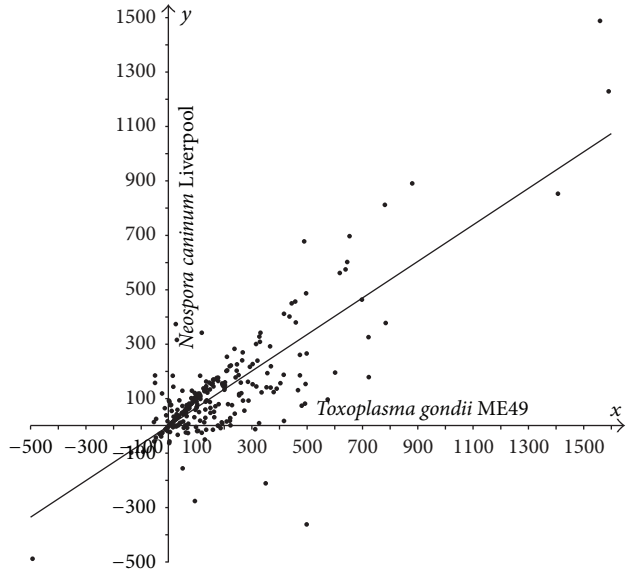


Figure 1: Plot of protein length extensions in *T. gondii* ME49 versus *N. caninum* relative to their orthologs in *Synechocystis* sp. PCC 6803. Differences in lengths between sporozoan and cyanobacterial orthologous proteins are plotted as follows: *T*-*S* on the *x*-axis, *N*-*S* on the *y*-axis.

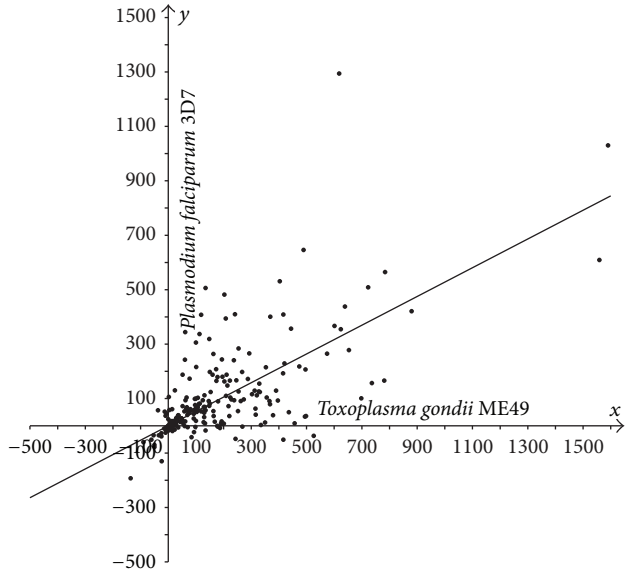


Figure 2: Plot of protein length extensions in *T. gondii* ME49 versus *P. falciparum* 3D7 relative to their orthologs in *Synechocystis* sp. PCC 6803. Differences in lengths between sporozoan and cyanobacterial orthologous proteins are plotted as follows: *T*-*S* on the *x*-axis, *P*-*S* on the *y*-axis.

the hypotheses that "a constant is better compatible with the set of points $\{(x_i, y_i) \mid 1 \le i \le n\}$ than the nontrivial affine function $y = kx + b$, $k \neq 0$" and "the linear function $y = kx$ is better compatible with this set of points than the affine function $y = kx + b$, $b \neq 0$": $F = \sqrt{((\sum_i (y_i - \widehat{y}_0(x_i))^2 / \sum_i (y_i - \widehat{y}(x_i))^2) - 1) \cdot (n-2)}$, where $\widehat{y}_0$ in the numerator is a constant (mean $\overline{y}$ over all y) or linear regression $y = kx$ and $\widehat{y}$ in the denominator is affine regression $y = kx + b$. This statistic can be explained more clearly: it determines whether there is a correlation between the difference $y_i - \widehat{y}(x_i)$ and x_i. This statistic is standard and substantiated in [31, 32]. The value of F was compared against a threshold defined as the Student random variable at significance level α, $t(n-2, \alpha)$. Under the number of degrees of freedom $n - 2 > 30$, the Student and standard Gaussian distributions approximate each other, and the threshold $t(266, 0.05)$ equals 1.96. An analogous statistic was used to test the hypothesis "affine function versus general polynomial of second degree." The confidence interval radius and the radius of the intercept (further referred to as *radius*) for the affine regression slope as well as the slope coefficient radius for linear regression were calculated in a standard fashion [32]. The Student test statistic S was used as well [32]. Deming regression and screening singular points were tested as well.

Regions of proteins with a predominance of one amino acid were determined by using the PROSITE program. The distribution of amino acid pairs separated by a fixed distance k in a given set of amino acid sequences was established using the simple computer program available from http://lab6.iitp.ru/utils/aapf/. Namely, frequencies of all amino acid pairs occurring in the given sequences at the distance of k residues (specified in the interval from 0 to 255) are computed and averaged over all sequences. The output is a frequency matrix of amino acid pairs. This matrix can be used to characterize nonstandard types of the putative signal peptide. This way it also appeared impossible to determine specificity of the N-terminus of apicoplast-targeted proteins in *T. gondii*; refer to Figure 4.

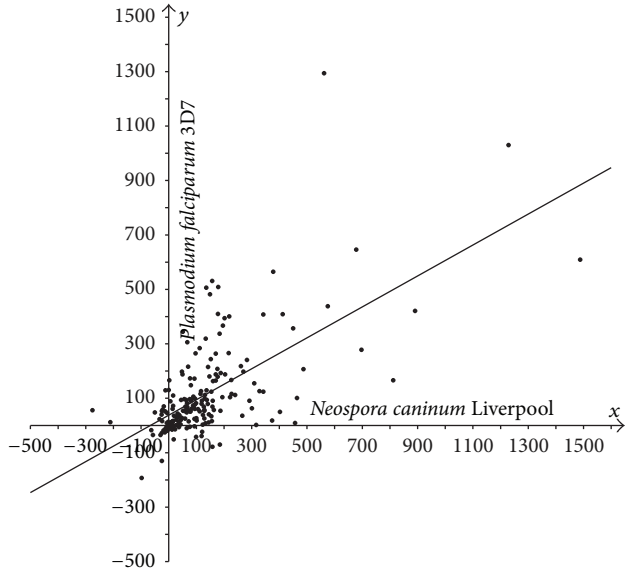


Figure 3: Plot of protein length extensions in *N. caninum* Liverpool versus *P. falciparum* 3D7 relative to their orthologs in *Synechocystis* sp. PCC 6803. Differences in lengths between sporozoan and cyanobacterial orthologous proteins are plotted as follows: *N*-*S* on the *x*-axis, *P*-*S* on the *y*-axis.

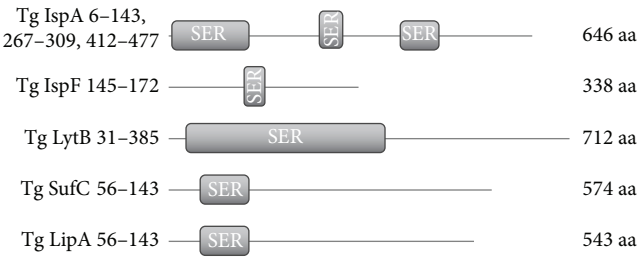


Figure 4: Some apicoplast-targeted proteins in *T. gondii* containing serine-rich regions. The region coordinates indicated on the left and the protein lengths on the right. The serine-rich regions are arranged irregularly, which is confirmed on a larger number of proteins for different pairs, triples, and so forth of amino acids.

3. Results and Discussion

Orthologs of *Synechocystis* sp. PCC 6803 were identified for 515 of 8319 (~6%) nuclear proteins in *T. gondii*, 560 of 7122 (~8%) nuclear proteins in *N. caninum*, and 390 of 5538 (~7%) nuclear proteins in *P. falciparum*. Only 877 of 3179 (~28%) proteins in *Synechocystis* sp. were found to be orthologous against at least one of the three apicomplexan species (see Supplementary Material available online at http://dx.doi.org/10.1155/2015/452958).

The identified orthologs are putative apicoplast-targeted proteins. Among them are proteins with either experimentally shown or anticipated apicoplast affinity, such as the bacterial type RNA polymerase sigma subunit (RpoD), DNA ligase, aminoacyl-tRNA synthetases, cell-cycle-associated protein kinase PRP4, enzymes IspA, IspB, IspE, IspF, IspG (GpcE), and IspH (LytB) of the mevalonate-independent pathway of isoprenoid biosynthesis, sulphur mobilization protein SufC from a Fe-S cluster assembly pathway, and LipA and LipB enzymes of lipoic acid synthesis (refer to the Introduction [5, 12, 14, 15]). In pairwise alignments, the sporozoan and cyanobacterial proteins usually align well at their C-termini, and the cyanobacterial sequence is fully covered by the alignment. In many cases, the N-termini of sporozoan proteins extend outside the alignment (data not shown).

In most cases, sporozoan proteins are longer compared to their bacterial orthologs, Figures 1–3. We demonstrate statistically that the majority of proteins in *T. gondii* are considerably longer compared to their orthologs in *N. caninum* Liverpool and *P. falciparum* 3D7, which was evidenced previously only for selected proteins [12, 33].

The hypotheses "a constant is better than the nontrivial affine function" and "affine function versus general polynomial of second degree" were rejected for every three sets of points shown in Figures 1–3. The hypothesis "the linear function is better than the affine function" was compatible with the first two sets ($F = 0.15$ and $F = 1.57$; refer to the designation in Materials and Methods section) and was rejected for the third set ($F = 3.40$). Thus, the third set was tested against the hypothesis "the mean over all x-coordinates coincides with the mean over all y-coordinates"; this hypothesis was accepted with the Student test statistic S at the same significance level α (with $S = 1.547$) [32].

Hence, the following regressions were justified. For set 1, $y = 0.6711x$ with radius 0.0468; for set 2, $y = 0.528x$ with radius 0.0590; for set 3, $y = 0.5685x + 37.756$ (linear regression rejected with $F = 3.40$) with radii 0.0926 and 21.7521, respectively.

The Deming regression gives approximately the same estimates; screening singular points does not significantly affect the results (data not shown).

So, the following conclusions can be drawn for the apicoplast protein orthologs that have orthologs in the cyanobacterium.

(1) Protein extensions in *T. gondii* are on average 1.5-fold longer compared to the corresponding extensions in *N. caninum*, with almost 1.0 confidence (Figure 1).

(2) Protein extensions in *T. gondii* are on average 2-fold longer compared to the corresponding extensions in *P. falciparum*, with high confidence (Figure 2).

(3) Set 3 (Figure 3) is compatible with the hypothesis that the average of protein extension lengths in *N. caninum* equals that in *P. falciparum*. Extension lengths in *P. falciparum* being less than 87 residues are longer than the corresponding extensions in *N. caninum* and, reversely, are shorter if they exceed 88 residues. In units, the dependency between extension lengths in *P. falciparum* versus *N. caninum* is an affine function $y = 0,5685x + 37,756$, where y runs over extension lengths in *P. falciparum* and x in *N. caninum*. The affinity, but not the linearity, of the regression testifies on behalf of the difference of *T. gondii* from her immediate species *P. falciparum* and *N. caninum* once again.

Among other specific features of apicoplast-targeted proteins is the abundance of serine-rich regions revealed in analyses with PROSITE (Figure 4). Each of the 3551 proteins in *T. gondii* ME49 possesses at least one 27 amino acid-long region with at least 9 serine residues, and 39 proteins possess at least one region with 27 or more continuous serine residues. Contrary to our expectations, larger-scale searching for serine-rich motifs in *T. gondii* showed their presence in various protein families, thus suggesting a selectively neutral nature of their origin. In other words, serine-rich regions are not specific to N-termini of apicoplast-targeted proteins. The same is also observed for other amino acids. This approach does not allow detecting a novel type of the N-terminal signal.

Earlier preliminary results are reported in [33].

4. Conclusions

For apicomplexan parasites, we suggest a statistically based method to compare the extension lengths of orthologous proteins that have orthologs in the cyanobacterium.

With this method, we demonstrate that the majority of cyanobacterium orthologs in *Toxoplasma gondii* are significantly longer compared to those in both *Neospora caninum* and *Plasmodium falciparum*. These proteins commonly lack

signal sequences typical for *Plasmodium* spp. [34]. The corresponding extensions might be essential for regulation of the apicoplast proteins and their translocation into the apicoplast. This notion conforms well with the observation that the apicoplast membrane in *T. gondii* is known to be less permissible, at least against drugs, compared to that in *P. falciparum* (personal communication with Gamaleya Research Institute of Epidemiology and Microbiology). Differences in protein extension lengths between *T. gondii* and other apicomplexan species may suggest different membrane transport mechanisms in these sporozoan groups. Mechanism of regulation and translocation in *T. gondii* may be based on protein processing in the cytoplasm to mature their extended N-termini.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors are deeply respectful to the editor for valuable comments that improved the paper. They also thank L. Rusin for valuable discussions and help with preparing the paper. Research was funded by the Russian Scientific Fund (Project 14-50-00150).

References

[1] T. N. Ermak, A. B. Peregudova, V. I. Shakhgil'dian, and D. B. Goncharov, "Cerebral toxoplasmosis in the pattern of secondary CNS involvements in HIV-infected patients in the Russian federation: clinical and diagnostic features," *Meditsinskaia Parazitologiia i Parazitarnye Bolezni*, no. 1, pp. 3–7, 2013.

[2] H. N. Luma, B. C. N. Tchaleu, Y. N. Mapoure et al., "Toxoplasma encephalitis in HIV/AIDS patients admitted to the Douala general hospital between 2004 and 2009: a cross sectional study," *BMC Research Notes*, vol. 6, article 146, 2013.

[3] N. Andenmatten, S. Egarter, A. J. Jackson, N. Jullien, J. P. Herman, and M. Meissner, "Conditional genome engineering in *Toxoplasma gondii* uncovers alternative invasion mechanisms," *Nature Methods*, vol. 10, no. 2, pp. 125–127, 2013.

[4] G. Di Guardo and S. Mazzariol, "*Toxoplasma gondii*: clues from stranded dolphins," *Veterinary Pathology*, vol. 50, no. 5, p. 737, 2013.

[5] R. J. M. Wilson, K. Rangachari, J. W. Saldanha et al., "Parasite plastids: maintenance and functions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1429, pp. 155–164, 2003.

[6] S. Köhler, C. F. Delwiche, P. W. Denny et al., "A plastid of probable green algal origin in Apicomplexan parasites," *Science*, vol. 275, no. 5305, pp. 1485–1489, 1997.

[7] J. Janouškovec, A. Horák, M. Oborník, J. Lukeš, and P. J. Keeling, "A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 24, pp. 10949–10954, 2010.

[8] S. Sato, "The apicomplexan plastid and its evolution," *Cellular and Molecular Life Sciences*, vol. 68, no. 8, pp. 1285–1296, 2011.

[9] G. Zhu, M. J. Marchewka, and J. S. Keithly, "Cryptosporidium parvum appears to lack a plastid genome," *Microbiology*, vol. 146, no. 2, pp. 315–321, 2000.

[10] K. A. Brayton, A. O. T. Lau, D. R. Herndon et al., "Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa," *PLoS Pathogens*, vol. 3, no. 10, pp. 1401–1413, 2007.

[11] M. J. Gardner, R. Bishop, T. Shah et al., "Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes," *Science*, vol. 309, no. 5731, pp. 134–137, 2005.

[12] N. Thomsen-Zieger, J. Schachtner, and F. Seeber, "Apicomplexan parasites contain a single lipoic acid synthase located in the plastid," *FEBS Letters*, vol. 547, no. 1–3, pp. 80–86, 2003.

[13] C. Y. He, B. Striepen, C. H. Pletcher, J. M. Murray, and D. S. Roos, "Targeting and processing of nuclear-encoded apicoplast proteins in plastid segregation mutants of *Toxoplasma gondii*," *Journal of Biological Chemistry*, vol. 276, no. 30, pp. 28436–28442, 2001.

[14] M. J. Crawford, N. Thomsen-Zieger, M. Ray, J. Schachtner, D. S. Roos, and F. Seeber, "Toxoplasma gondii scavenges host-derived lipoic acid despite its de novo synthesis in the apicoplast," *EMBO Journal*, vol. 25, no. 13, pp. 3214–3222, 2006.

[15] J. Mazumdar, E. H. Wilson, K. Masek, C. A. Hunter, and B. Striepen, "Apicoplast fatty acid synthesis is essential for organelle biogenesis and parasite survival in *Toxoplasma gondii*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 35, pp. 13192–13197, 2006.

[16] S. Agrawal and B. Striepen, "More membranes, more proteins: complex protein import mechanisms into secondary plastids," *Protist*, vol. 161, no. 5, pp. 672–687, 2010.

[17] T. Brüser and C. Sanders, "An alternative model of the twin arginine translocation system," *Microbiological Research*, vol. 158, no. 1, pp. 7–17, 2003.

[18] D. Mehner, H. Osadnik, H. Lünsdorf, and T. Brüser, "The Tat system for membrane translocation of folded proteins recruits the membrane-stabilizing Psp machinery in *Escherichia coli*," *Journal of Biological Chemistry*, vol. 287, no. 33, pp. 27834–27842, 2012.

[19] G. Cilingir, S. L. Broschat, and A. O. T. Lau, "ApicoAP: the first computational model for identifying apicoplast-targeted proteins in multiple species of apicomplexa," *PLoS ONE*, vol. 7, no. 5, Article ID e36598, 2012.

[20] G. Cilingir, A. O. T. Lau, and S. L. Broschat, "ApicoAMP: the first computational model for identifying apicoplast-targeted transmembrane proteins in Apicomplexa," *Journal of Microbiological Methods*, vol. 95, no. 3, pp. 313–319, 2013.

[21] K. E. Jackson, J. S. Pham, M. Kwek et al., "Dual targeting of aminoacyl-tRNA synthetases to the apicoplast and cytosol in *Plasmodium falciparum*," *International Journal for Parasitology*, vol. 42, no. 2, pp. 177–186, 2012.

[22] B. Kumar, S. Chaubey, P. Shah et al., "Interaction between sulphur mobilisation proteins SufB and SufC: evidence for an iron-sulphur cluster biogenesis pathway in the apicoplast of *Plasmodium falciparum*," *International Journal for Parasitology*, vol. 41, no. 9, pp. 991–999, 2011.

[23] E. V. S. R. Ram, A. Kumar, S. Biswas, S. Chaubey, M. I. Siddiqi, and S. Habib, "Nuclear *gyrB* encodes a functional subunit of the *Plasmodium falciparum* gyrase that is involved in apicoplast DNA replication," *Molecular and Biochemical Parasitology*, vol. 154, no. 1, pp. 30–39, 2007.

[24] B. J. Foth, S. A. Ralph, C. J. Tonkin et al., "Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*," *Science*, vol. 299, no. 5607, pp. 705–708, 2003.

[25] T. A. Sadovskaya and A. V. Seliverstov, "Analysis of the 5′-Leader regions of several plastid genes in protozoa of the phylum apicomplexa and red algae," *Molecular Biology*, vol. 43, no. 4, pp. 552–556, 2009.

[26] T. Kaneko and S. Tabata, "Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803," *Plant and Cell Physiology*, vol. 38, no. 11, pp. 1171–1176, 1997.

[27] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D290–D301, 2012.

[28] C. J. A. Sigrist, E. De Castro, L. Cerutti et al., "New and continuing developments at PROSITE," *Nucleic Acids Research*, vol. 41, no. 1, pp. D344–D347, 2013.

[29] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[30] O. A. Zverkov, A. V. Seliverstov, and V. A. Lyubetsky, "Plastid-encoded protein families specific for narrow taxonomic groups of algae and protozoa," *Molecular Biology*, vol. 46, no. 5, pp. 717–726, 2012.

[31] J.-R. Barra, *Notions Fondamentales de Statistique Mathématique: Maîtrise de Mathématiques et Applications Fondamentales*, Dunod, Paris, France, 1971.

[32] G. A. F. Seber, *Linear Regression Analysis*, John Wiley & Sons, New York, NY, USA, 1977.

[33] N. V. Kobets, D. B. Goncharov, A. V. Seliverstov, O. A. Zverkov, and V. A. Lyubetsky, "Comparative analysis of apicoplast - targeted proteins in *Toxoplasma gondii* and other Apicomplexa species," in *Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB '13)*, Moscow, Russia, July 2013.

[34] L. Lim and G. I. McFadden, "The evolution, metabolism and functions of the apicoplast," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1541, pp. 749–763, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 786861, 11 pages
http://dx.doi.org/10.1155/2015/786861

Research Article

Lengths of Orthologous Prokaryotic Proteins Are Affected by Evolutionary Factors

Tatiana Tatarinova,[1] Bilal Salih,[2,3] Jennifer Dien Bard,[1] Irit Cohen,[2,4] and Alexander Bolshoy[2]

[1] *Children's Hospital Los Angeles, Keck School of Medicine, University of Southern California, Los Angeles, CA 90027, USA*
[2] *Department of Evolutionary and Environmental Biology and Institute of Evolution, University of Haifa, 3498838 Haifa, Israel*
[3] *Department of Computer Science, University of Haifa, 3498838 Haifa, Israel*
[4] *The Tauber Bioinformatics Research Center, University of Haifa, 3498838 Haifa, Israel*

Correspondence should be addressed to Tatiana Tatarinova; tatiana.tatarinova@usc.edu and Alexander Bolshoy; bolshoy@research.haifa.ac.il

Received 8 September 2014; Accepted 2 November 2014

Academic Editor: Vassily Lyubetsky

Proteins of the same functional family (for example, kinases) may have significantly different lengths. It is an open question whether such variation in length is random or it appears as a response to some unknown evolutionary driving factors. The main purpose of this paper is to demonstrate existence of factors affecting prokaryotic gene lengths. We believe that the ranking of genomes according to lengths of their genes, followed by the calculation of coefficients of association between genome rank and genome property, is a reasonable approach in revealing such evolutionary driving factors. As we demonstrated earlier, our chosen approach, Bubble-sort, combines stability, accuracy, and computational efficiency as compared to other ranking methods. Application of Bubble Sort to the set of 1390 prokaryotic genomes confirmed that genes of Archaeal species are generally shorter than Bacterial ones. We observed that gene lengths are affected by various factors: within each domain, different phyla have preferences for short or long genes; thermophiles tend to have shorter genes than the soil-dwellers; halophiles tend to have longer genes. We also found that species with overrepresentation of cytosines and guanines in the third position of the codon (GC_3 content) tend to have longer genes than species with low GC_3 content.

1. Introduction

To better understand the interaction between the environment and bacteria, whether in a human host or any other ecosystem, one must know the laws governing prokaryotic evolution and adaptation to environment. For example, it is essential to study how a change in pH or external temperature affects a bacterial genome and especially its coding sequences. Unfortunately, the laws of prokaryotic coding sequence evolution remain unclear. Orthologous proteins may drastically differ in both codon usage and length across species. When a gene length changes, a protein may acquire a new function or lose an existing one, hence, changing the entire ecosystem. Many studies have analyzed the relationship between codon usage and the environment [1–3], but a few efforts were made to predict the effect of a changing environment on gene length. The main results were related to comparative analysis between protein lengths in eukaryotes and prokaryotes. Detailed comparison of protein length distributions in eukaryotes and prokaryotes can be found in [4, 5]. Wang et al. [6] proposed that "molecular crowding" effect and evolution of linker sequences can explain differences between length of orthologous sequences in super-kingdoms. Our study is focused on protein lengths in prokaryotes, exclusively.

How does gene length change occur in prokaryotes? The main driving force in shaping gene length is a point mutation [7]. Point mutations may cause a stop codon shift, when the existing stop codon is destroyed and gene length is increased, a start codon drift, or appearance of a premature stop codon. To understand trends of fixation of mutations

changing protein lengths we performed a comparative study of lengths of paralogs. We explore the use of seriation of genomes based on paralogs' lengths.

In recent papers [8, 9], we formulated the genome ranking problem, listed several approaches to solve it, described a novel method for genome ranking according to gene lengths, and demonstrated preliminary results from the ranking of prokaryotic genomes. These results indicated that hyperthermophilic species have shorter genes than mesophilic organisms. We hypothesize that gene lengths are not randomly distributed; instead they are affected by a number of environmental, genomic and taxonomic factors. In this paper we present a framework for analysis of gene lengths and evaluate effects of environmental factors.

In order to analyze evolutionary pressures acting on genes it is necessary to group them into well-defined functional categories. There are several existing approaches. First of all, there is the most popular database of Clusters of Orthologous Groups (COG) of proteins, which is a comprehensive collection of prokaryotic gene families. This database was created to classify the complete complement of proteins encoded by complete genomes based on evolutionary development. The data in COGs are updated continuously following the sequencing of new prokaryotic genomic sequences. As described by Tatusov et al. [10], the COGs database is a growing and useful resource to identify genes and groups of orthologs in different species that are related by evolution. Sixteen years ago, the database was started with only seven Bacterial genomes; in 2010 the database consisted of proteins from 52 Archaeal and 601 Bacterial genomes (a total of 653 complete genomes) that were assigned to 5,663 COGs; currently it contains approximately 2 K genomes.

The COG database is not the only possible data compilation to classify prokaryotic proteins. Since its publication over a decade ago, additional classifications have appeared. In 2007, *Archaea* were grouped into the acCOG database [11]. Another alternative, the eggNOG database [12, 13], grouped gene families at the universal level, covering all three domains of life.

Recently, Bolshoy et al. introduced a "gene-length based" model [14, 15], representing genomes as vectors of genes. The set of genomes is represented as a matrix, in which each row stands for a genome and each column stands for a gene family. Therefore, each element of this matrix stands for the length of a member of a gene family i in a genome j. In our study, the objects are annotated prokaryotic genomes; the descriptors are the lengths of the genome proteins indexed according to the COG database.

A ranking is a relationship between a set of objects such that, for any two objects, the first is either ranked "higher than," "lower than," or "equal to" the second. Gene ranking is a useful approach to answer biological questions, however it is sometimes difficult to implement. Here we bring examples of usage this measure in biologic sciences. A prioritization or ranking is used in bioinformatics to aid in the discovery of disease-related genes. Computational methods are employed for ranking the genes according to their likelihood of being associated with the disease. A variety of methods have been conceived by the researchers for the prioritization of the disease candidate genes. A review of various aspects of computational disease gene prioritization and related problems is presented in Gill et al. [16].

In our case, the goal is to order the genomes that are represented as rows of a gene length matrix. There are different possible approaches to define the optimal rank of rows in the matrix. We have previously determined [9] that Bubble Sort method (B-Sort, see Section 5) is more accurate than Average Sort and Simple Additive Ranking and it is as accurate and significantly faster than the Simulated Annealing Procedure.

The complexity of the ranking problem using matrices with missing values was discussed in detail [17]. The same ranking problem appears in several areas of operations research, such as in the context of group decision making [18] and country-credit risk rating [19]. Missing data as well as variable relative importance of different gene families make the problem increasingly complex. To the best of our knowledge, genome ranking problem has been addressed for the first time in [8, 9].

Establishing ordered lists of genomes using lengths of coding sequences of orthologous genes, we aim to find an association between a genome rank and a genome property of interest, such as its role in virulence and adaptation. There are many different types of such properties: a prokaryote can be either Archaea or Bacteria; an organism may be hyperthermophile, thermophile, psychrophile, or mesophile; a genome has a certain GC-content, and so on. In summary, the goal is to find out whether gene lengths of a genome are associated with various genome properties and to measure the magnitude of this association. These findings will allow us to determine important factors such as virulence, biofilm formation, and antimicrobial resistance that may be associated with the pathogenesis of a specific species and the ability to cause serious infections in patients.

2. Results

We used a dataset of 1390 genomes (the "big" dataset) and a randomly selected subset of 100 genomes (the "small" dataset). For each dataset we used complete and filtered versions. The filtering procedure (see Section 5) removes those COGs that are present in only a small number of genomes and are likely to skew the ordering results. We set the frequency threshold to be 35%, meaning that the filtering procedure removes COGs present in less than 35% of analyzed genomes. After filtering we obtained the filtered dataset containing 1474 COGs.

We assessed the consistency of ranks of genomes of the small dataset in two orderings: of the entire collection of 1390 genomes and of the subset of 100 genomes (Figure 1). We determined corresponding ranks of 100 genomes in the B-sorted dataset of 1390 genomes and discovered that the two orderings of 100 genomes were highly consistent (with correlation coefficient of 0.95). This confirms that the ranking procedure is stable. However, random selection of a small subset may cause wrong ranks of a few isolated genomes. Indeed, there are some genomes that show differences in 100 and 1390 genomes rank, for example, bacteria *Sodalis glossinidius*, which is ranked 42 in 100 genomes and 162

Table 1: B-sort results (one run) for 1390 genomes, archaea.

Phylum	Average rank	StDev	Median rank	Rank range	Number of genomes
Crenarchaeota	189	179	77	7–492	35
Euryarchaeota	312	297	233	5–1263	74
Korarchaeota	169	NA	169	169–169	1
Nanoarchaeota	5	NA	5	5–5	1
Thaumarchaeota	347	239	347	178–516	2
Unclassified archaea	771	NA	771	771–771	1

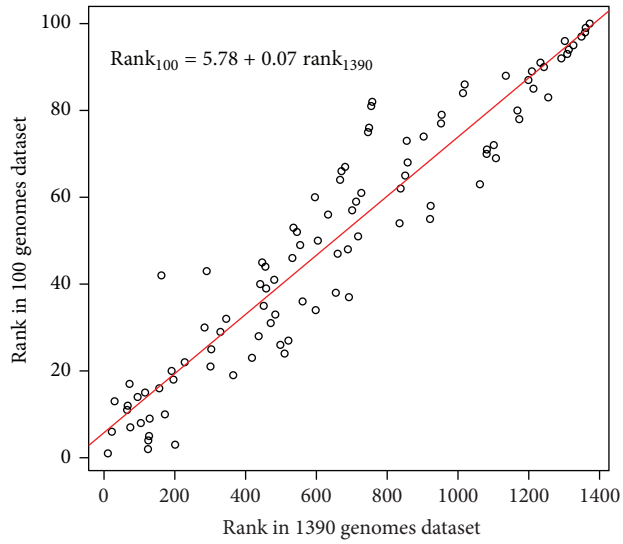


Figure 1: Consistency of Bubble Sort ranks in 1390 and 100 genomes datasets. Pearson's correlation coefficient between two ranks is 0.95; Kendall tau correlation coefficient is 0.82.

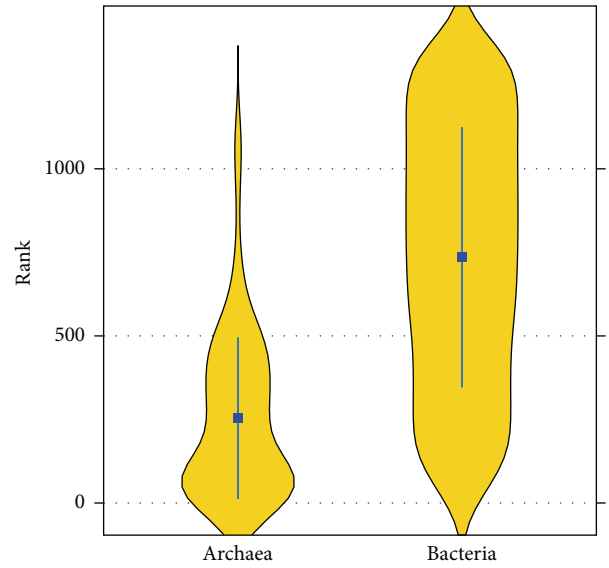


Figure 2: Violin plots of Bubble sort ranks of Archaea and Bacteria. Average rank of 1276 *Bacterial* genomes is 735 and average rank of 114 *Archaeal* genomes is 254.

in 1390 genomes dataset. Therefore, the ranks' consistency found for the huge majority of ranks is an additional support to the chosen method of ranking.

Let us start with an overview of the orderings; let us compare ranks of Bacteria and Archaea. (Larger value of a genome rank means longer genes in this genome.) Bacterial genomes have a broader distribution of ranks than Archaeal genomes (Figure 2). Overall, Bacterial ranks are larger than Archaeal ranks in the 1390 genome, as well as in 100 genome datasets. This observation can be illustrated using the violin plot of ranks' distributions, as shown in Figure 2. Average rank of 1276 *Bacterial* genomes was 735 and average rank of 114 Archaeal genomes was 254. This visual observation is also supported by a simple statistical procedure. Using the Wilcoxon rank test and $\alpha = 0.01$, we calculated the test statistic T_a, equal to the sum of the ranks for the ordered data that belong to Archaea. T_a was 28,913. For large samples T_a is approximately normal with expected value and standard deviation calculated as

$$E(T_a) = \frac{n_a(n_b + n_a + 1)}{2} = 79,287,$$
$$\sigma(T_a) = \sqrt{\frac{n_a n_b(n_a + n_b + 1)}{12}} = 4106.3. \quad (1)$$

Therefore,

$$Z = \frac{T_a - E(T_a)}{\sigma(T_a)} = -12.27,$$
$$P(Z < -12.27) \approx 10^{-34} < 0.01. \quad (2)$$

Hence, we conclude that *Bacterial* genomes rank significantly higher than *Archaeal* genomes. Tables 1 and 2 show the summary statistics for the ordering of Archaeal and Bacterial genomes. These tables show mean, median, range, and standard deviation of Archaeal and Bacterial ranks of 1390 genomes stratified by phylum. In the Bacterial domain, Firmicutes and Thermotogae have shorter genes and Actinobacteria have longer ones. In the Archaeal domain, Euryarchaeota have longer genes than Crenarchaeota.. These results are consistent with our earlier findings from analysis of 100 prokaryotic genomes [8].

Next, we considered the nucleotide composition of coding regions. In prokaryotes, the nucleotide composition of coding regions varies significantly between species. GC_3 (frequency of cytosine and guanine in the third position of the codon) is one of the variable features. Across the Bacterial domain, GC_3 ranges from 10% to 90% [20]. Tatarinova et al.

Table 2: B-sort results for 1390 genomes, bacteria.

Phylum	Average rank	STD	Median rank	Rank range	Number of genomes
Actinobacteria	1223	166	1260	343–1390	137
Aquificae	182	79	168	82–306	8
Bacteroidetes/Chlorobi	992	188	1071	502–1359	71
Candidatus Cloacamonas	1054	NA	1054	1054–1054	1
Chlamydiae/Verrucomicrobia	1076	81	1079	835–1223	25
Chloroflexi	774	520	1109	70–1274	15
Chrysiogenetes	545	NA	545	545–545	1
Cyanobacteria	938	209	975	619–1276	40
Deferribacteres	205	NA	205	205–205	1
Deinococcus-Thermus	607	282	566	263–1126	12
Dictyoglomi	207	49	207	172–242	2
Elusimicrobia	412	143	412	311–513	2
Fibrobacteres/Acidobacteria	1171	172	1240	839–1293	6
Firmicutes	307	188	286	21–1387	271
Fusobacteria	462	100	461	361–564	4
Gemmatimonadetes	1214	NA	1214	1214–1214	1
Nitrospirae	563	418	563	267–858	2
Planctomycetes	1364	29	1368	1319–1389	5
Proteobacteria	759	325	775	1–1379	588
Spirochaetes	1050	155	1066	700–1317	31
Synergistetes	466	40	466	438–494	2
Tenericutes	657	223	631	92–1092	36
Thermobaculum	1049	NA	1049	1049–1049	1
Thermodesulfobacteria	458	32	458	435–480	2
Thermotogae	253	165	203	45–566	12

previously demonstrated [21, 22] that, within one eukaryotic species, GC_3 content can be used to distinguish two classes (housekeeping and stress-specific) genes. Currently, we sought to evaluate mutation pressure acting on the entire prokaryotic genome by examining how the average GC_3 content, calculated across all genes in a genome, is related to the position of the genome in a global ordering. We calculated the GC_3 content of coding regions across all analyzed genomes and discovered that the genome rank and cytosine/guanine content of the third codon position of genes are positively correlated (Spearman rank correlation coefficient $\rho(GC_3, \text{rank}) = 0.62$ for *Bacteria* and $\rho(GC_3, \text{rank}) = 0.59$ for *Archaea*). For example, the average GC_3 content in Actinobacteria (0.70) is twice the amount seen in Firmicutes (0.35).

3. Discussion

The ability of some species to grow at high temperatures has been a long-term fascination of microbiologists. Proteins of hyperthermophilic species are more resilient to heat and are shorter than proteins of mesophilic species. Understanding this effect is important for biotechnology [23].

Up to now, less than a dozen studies were devoted to protein length distribution. Among those, there were only four relevant publications: [4, 5, 8, 24]. In 2000, using an early version of the COG database, Zhang compared 22 species in three domains of life [4] and found that the average gene length is smallest for *Archaea* and greatest for eukaryotes. Similarly, Skovgaard et al. [24] analysed 34 prokaryotic genomes and discovered that, for the vast majority of functional families, Bacterial proteins were longer than *Archaeal* ones. In 2005, Brocchieri and Karlin [5] confirmed these findings using a larger collection of genomes (16 *Archaeal* and 67 *Bacterial* species). They found that bacteria were enriched in functional families with longer genes. In addition, they described a negative correlation between protein length and optimal growth temperature of *Archaea* and *Bacteri*a. By grouping proteins into broad functional classes (information storage and processes; cellular processes; metabolism; poorly characterized; not characterized) and comparing their median lengths, Brocchieri and Karlin concluded that "information storage and processes" proteins are shorter than "cellular processes" and "metabolism" proteins. They also found that Archaea have more of the shorter and poorly characterized proteins.

The above mentioned studies, performed on relatively small sets of genomes, share the same deficiency of using average (mean or median) lengths of genes in a genome to reach their conclusions. As we illustrated in [8, 9] this approach can substantially distort results. In [8, 9] we proposed a systematic framework to analyse the relationship of prokaryotic gene lengths and environmental conditions that is not based on analysis of average lengths of proteins. This

framework, further investigated in the current paper, allows more flexibility and produces more meaningful results than the previous approaches.

Hyperthermophilic species of Archaea and Bacteria, living in extreme environments (such as volcanic hot springs) occupy the top portions of the ranking lists of the small and big datasets. At a first glance it appears that we could hypothesize that extremophiles have shorter genes than species living under normal conditions. However, the situation appears to be more complex. For illustration we consider Sulfolobales, Thermoproteales, and Halobacteriales. Sulfolobales grow in volcanic hot springs at pH 2-3 and a temperature of 75–80 degrees Celsius. In the ordered list of 1390 genomes, Sulfolobales occupy positions from 12 to 94, which means that as a rule Sulfolobales have very short genes. *Thermoproteales* (extremely thermoacidophilic anaerobic Archaea isolated from Icelandic solfataras) also have very short genes, their genomes are found in positions from 7 to 77 but also in positions from 412 to 460 in the ordered list, which are positions of genomes with moderately short genes. Halobacteriales (found in water saturated or nearly saturated with salt) are placed in positions from 541 to 1263 which are not considered genomes with short genes. From these observations follows that stress of living in *an arbitrary* extreme environment is not the factor, while, probably, hyperthermophilicity and halophilicity are the factors affecting orderings in opposite directions.

We also showed that, as a group, Bacterial genomes are ranked significantly higher than the Archaeal ones according to the length of their genes (Figure 2). This observation may be explained by the fact that the vast majority of completely sequenced Archaeal genomes are hyperthermophiles, which tend to have shorter genes as compared to psychrophiles and mesophiles. Our previous speculations obtained on relatively small datasets [8] and our current results on 1390 genome dataset are consistent with the hypothesis that high temperature environment is a factor causing reduction of gene length. In the 100-genome dataset hyperthermophiles occupy positions in the top portion of the list: top 20 in the 100 genomes list. They are also ranked in the top of the 1390 genome dataset.

We also observed that 34% of the shortest (first 100 positions in the ordered list) of 1390 genomes are occupied by hyperthermophilic species, while none are found in the longest (last 100 in the ordered list). Furthermore, 90% of thermophiles are placed in the top third of the list. Moderately thermophilic species are not restricted to the top positions. For example, *Thermobifida fusca* (a moderately thermophilic soil bacterium growing at 55°C and a major degrader of plant cell walls in heated organic materials such as compost heaps, rotting hay, manure piles or mushroom growth medium) occupies position 1260 in the ordered list. *Anaerolinea thermophila*, with similar growth temperature, has a close position of 1173.

There are several remarkable features that appeared as a result of the 1390 genome ordering. Campylobacterales (belonging to the phylum Proteobacteria) have an average position of 203, with the smallest position of 10 (*Helicobacter bizzozeronii ciii-1*) and the largest position of 392 (*Helicobacter hepaticus atcc 51449*). Most species in this family are human and animal pathogens. Namely, *Campylobacter jejuni* is a microaerophilic bacterium frequently associated with gastroenteritis in humans. Complications such as meningitis [25], septicemia [26], and Guillain-Barré syndrome have also been reported [27]. In addition, *Helicobacter bizzozeronii* (position 10) has been implicated in gastric infections, similar to *Helicobacter pylori*, referred to as *non-Helicobacter pylori Helicobacter* (NHPH) infections in humans [28]. It appears that all known Campylobacterales have short genomes. It is tempting to speculate that there are evolutionary pressures to keep genes in short pathogenic genomes as short as possible.

However, not all pathogens have short genes. Not even all pathogens with short genomes have short genes. Common obligate intracellular prokaryotic pathogens from the phylum of Chlamydiae are very small (measuring 0.3–0.6 μm in diameter) and grow by infecting eukaryotic host cells. This phylum is comprised of several major intracellular pathogens of humans and animals, causing a variety of diseases. These bacteria can cause keratoconjunctivitis, pneumonitis, and sexually transmitted infections. In spite of its small physical dimensions, Chlamydiae have exceptionally long genes: the ranks of 21 members of this class are located from positions 835 to 1127 in the ranking list. We speculate that there are certain evolutionary factors (yet to be discovered) that keep Chlamydiae genes so long.

As we see, Campylobacterales (of the phylum Proteobacteria), have short genes. At the opposite end of the length spectrum we find the phylum Actinobacteria, tending to have longer genes. Only 8 out of 137 species of Actinobacteria have positions below 1000 in the ordered set. One of the species, a pathogenic bacterium *Renibacterium salmoninarum* [29], was placed among species with characteristically short genes in the position 343. The genome of *R. salmoninarum* has extended regions of synteny to the *Arthrobacter* sp. strain FB24 and *Arthrobacter aurescens* TC1 genomes, but it is approximately 1.9 Mb smaller than two sequenced *Arthrobacter* genomes and has a lower GC content [29]. In the Bubble Sort list, *Arthrobacters* occupy positions 1230, 1301, 1342, 1343, and 1354. Our results show that significant genome reduction, which has occurred since divergence from the last common ancestor, affected not only gene content but also lengths of remaining genes. It is possible that factors affecting gene lengths of Actinobacteria are different from the factors acting on Chlamydiae, while resulting in keeping proteins longer in both cases.

Relationships between gene length and codon bias have been previously studied by [30–33]. Oliver and Marín [30] and Xia et al. [32] observed a positive correlation between length and GC composition of coding sequences in prokaryotes, attributing the effect to reduced frequency of stop codons in GC-rich species. Later Xia et al. [33] mentioned that the correlation is weak for a number of species, with 4 species showing a negative correlation. Thus Xia et al. formulated a more general hypothesis incorporating selection against cytosine (C) usage. In [33] they described two additional factors giving rise to this selection: transcription efficiency and "insurance" against cytosine deamination.

Third positions in codon are largely degenerate; 70% of changes at third codon positions are synonymous [34]. Therefore, it makes sense to analyze adaptation effects using GC composition in the third position of the codon, GC_3. We showed that adaptation to higher temperatures affects the genome in two ways: first, GC_3 content of genes tends to increase with growth temperature [35]; at the same time, hyperthermophilic species tend to have shorter genes as it can be seen from the ranks of these species both in the 100-genome dataset and in the larger dataset. Several factors may compete for placement of the Bacterial species in the ordering rank. Adaptation to high temperatures and pathogenicity may tend to place an organism into lower ranks. High GC_3 composition and adaptation to high salinity environments places an organism into higher ranks. However, future research is needed to determine important factors, both environmental and genomic, that may affect the rank of the genome. This information will allow us to further understand and possibly predict the invasive or virulent nature of a particular species compared to a nonpathogenic organism that is part of the normal commensal flora of an individual. Further exploration of these factors may also answer questions on the emerging mechanisms of resistance that may be associated with specific organisms and on prediction of resistance using novel methods other than conventional susceptibility tests.

We will continue updating our collection of prokaryotic genome orderings. When a new genome is sequenced, it is not necessary to repeat the entire ranking procedure from an unordered dataset. In order to incorporate a newly sequenced genome in our analysis, it is necessary to (1) predict genes and (2) assign COG categories. Then the new, completely annotated, genome can be added to the presorted data matrix, using average gene length as a rough indicator of the new genome position. Then the ranking procedure should be applied to the updated matrix. Since all but one of the genomes is already in the correct place, the ranking procedure will have to make only a small number of steps to determine the rank of a new genome.

4. Conclusions

We applied Bubble Sort to the set of 1390 prokaryotic genomes and revealed several interesting trends. We demonstrated that hyperthermophiles may be always characterized as having short proteins. Also, the resulting ordering showed that Archaea have shorter genes than Bacteria, and we speculate that this can be attributed to the prevalence of hyperthermophiles among the sequenced Archaea. Within each domain, different phyla have preferences for short or long genes. Another interesting observation is the significant correlation between gene length and GC composition of coding regions. Therefore, we suggest that gene lengths are not randomly distributed across species but are shaped by environmental and genomic factors.

The genome ranking procedure is stable. Inclusion of additional genomes does not distort the relative ranking of genomes. The correlation coefficient between the ranks of the 100 genomes in the 100-genome dataset and in the larger (1390) dataset is 0.95. Hyperthermophilic species are ranked on top in both 100 and 1390-genome lists; soil dwelling species are consistently at the bottom of the list.

Our results show that environmental factors constitute a strong force that groups evolutionary distant species together in protein-lengths' ranking. On the other hand, evolutionary history and phylogenetic closeness group certain organisms together as well. Relative influence of these factors varies between organisms. For example, we demonstrated that hyperthermophilic species have shorter genes than mesophilic organisms, which implies that environmental factors may affect gene length. However, not every environmental stress has the gene shortening effect. For example, high salinity represents an extreme environment that relatively few organisms have been able to adapt to and occupy. Halophiles are a type of extremophile organisms that live in high salt concentrations. Seemingly, high salinity opposite to high temperature does not cause protein-length decrease; the extreme halophiles (or halobacteria), tend to have pretty long genes.

5. Materials and Methods

All four ranking algorithms discussed in this paper were applied to input matrices based on the database of Clusters of Orthologous Groups of proteins (COG) [10, 36–38]. As of October 2012, there were 5664 COGs, 1276 Bacterial and 114 Archaeal genomes sequences in the NCBI database. The sequences were processed according to the procedures described below.

5.1. COGs Database. Information about every completely sequenced and annotated prokaryotic genome is stored as tables of protein features, called PTT files, prepared by the National Center for Biotechnology Information (NCBI). The complete collection of current PTT files can be found at ftp://ftp.ncbi.nih.gov/genomes/.

From every prokaryotic NCBI PTT file, we extracted information about each gene length, COG and added the genome index (tax id). We created a combined gene-length matrix, where rows correspond to genomes, identified by taxonomy id, and columns correspond to COGs. Each element (i, j) of this matrix is a length of gene belonging to COG i in genome j. All currently available genomes were described in these two files. To check the ranking methods described below we used small subsets (100 genomes) of this dataset.

5.2. Preprocessing Procedures. To get an input file for further ranking the following preprocessing procedures developed by Bolshoy et al. [9, 15, 39] were applied.

(1) *Selection of Genome Subsets*. A subset may be defined applying different criteria: it may be either a representative sample, a taxaspecific subset, or randomly chosen genomes.

(2) *Application of a Filtering Parameter (An Entry Threshold) on a Selected Subset*. Only COGs containing more than a threshold number of genomes are considered for further processing. For example, if the filtering value is equal to 20% and an amount of genomes in

Table 3: List of Archaeal (A) and Bacterial (B) genomes in the Bubble Sort ordering rank, 100 genomes dataset. Hyperthermophiles, Streptococci, and Enterococci are marked in the Note column.

Rank	Domain	Note	Organism
1	A	Hyperthermophile	*Archaeoglobus fulgidus dsm 4304*
2	A	Hyperthermophile	*Thermoplasma volcanium gss1*
3	B	Hyperthermophile	*Thermotoga sp. rq2*
4	A	Hyperthermophile	*Thermoplasma acidophilum dsm 1728*
5	B	Hyperthermophile	*Thermotoga neapolitana dsm 4359*
6	A	Hyperthermophile	*Thermococcus onnurineus na1*
7	B		*Campylobacter concisus 13826*
8	B		*Campylobacter curvus 525.92*
9	B	Hyperthermophile	*Aquifex aeolicus vf5*
10	B	Hyperthermophile	*Dictyoglomus thermophilum h-6-12*
11	B		*Bacillus cereus atcc 14579*
12	B		*Bacillus cytotoxicus nvh 391-98*
13	B		*Melissococcus plutonius atcc 35311*
14	A	Hyperthermophile	*Thermococcus sibiricus mm 739*
15	B		*Listeria monocytogenes clip81459*
16	B		*Bacillus amyloliquefaciens dsm 7*
17	B		*Rickettsia canadensis str. Mckiel*
18	A	Hyperthermophile	*Pyrococcus abyssi ge5*
19	B		*Helicobacter felis atcc 49179*
20	A	Hyperthermophile	*Pyrococcus horikoshii ot3*
21	B	Streptococcus	*Streptococcus pneumoniae p1031*
22	B	Streptococcus	*Streptococcus agalactiae a909*
23	B		*Caldicellulosiruptor bescii dsm 6725*
24	B		*Mycoplasma fermentans m64*
25	B	Streptococcus	*Streptococcus agalactiae 2603v/r*
26	A		*Methanosalsum zhilinae dsm 4017*
27	B		*Francisella sp. tx077308*
28	B	Streptococcus	*Streptococcus equi subsp. zooepidemicus*
29	B		*Bacillus pumilus safr-032*
30	B		*Pediococcus pentosaceus atcc 25745*
31	B		*Geobacter lovleyi sz*
32	B	Enterococcus	*Enterococcus faecalis v583*
33	B		*Natranaerobius thermophilus jw/nm-wn-lf*
34	B		*Mycoplasma pulmonis uab ctip*
35	B		*Brevibacillus brevis nbrc 100599*
36	B		*Mycoplasma genitalium g37*
37	B		*Mycoplasma leachii pg50*
38	B		*Ureaplasma parvum serovar 3*
39	B		*Bacillus thuringiensis str. al hakam*
40	B		*Neisseria meningitidis 053442*
41	B		*Legionella pneumophila str. paris*
42	B		*Sodalis glossinidius str. "morsitans"*
43	B		*Candidatus riesia pediculicola usda*
44	B		*Lactobacillus gasseri atcc 33323*
45	B		*Coxiella burnetii rsa 331*
46	B		*Laribacter hongkongensis hlhk9*
47	B		*Ruminococcus albus 7*
48	B		*Mycoplasma pneumoniae m129*
49	A		*Halalkalicoccus jeotgali b3*

Table 3: Continued.

Rank	Domain	Note	Organism
50	B		*Geobacter uraniireducens rf4*
51	B		*Brachyspira pilosicoli 95/1000*
52	B		*Pseudogulbenkiania sp. nh8b*
53	B		*Dechloromonas aromatica rcb*
54	B		*Maribacter sp. htcc2170*
55	B		*Zobellia galactanivorans*
56	B		*Escherichia coli bw2952*
57	B		*Erwinia amylovora atcc 49946*
58	B		*Gramella forsetii kt0803*
59	B		*Klebsiella variicola at-22*
60	B		*Salmonella enterica subsp. arizonae serovar*
61	B		*Yersinia enterocolitica subsp. enterocolitica 8081*
62	B		*Methylomonas methanica mc09*
63	B		*Borrelia turicatae 91e135*
64	B		*Cronobacter turicensis z3032*
65	B		*Yersinia pseudotuberculosis pb1/+*
66	B		*Xanthomonas oryzae pv. oryzae maff 311018*
67	B		*Tropheryma whipplei tw08/27*
68	B		*Spirochaeta smaragdinae dsm 11293*
69	B		*Sphingobacterium sp. 21*
70	B		*Dyadobacter fermentans dsm 18053*
71	B		*Eubacterium eligens atcc 27750*
72	B		*Chlamydophila pneumoniae ar39*
73	B		*Pelodictyon phaeoclathratiforme bu-1*
74	B		*Desulfovibrio vulgaris str. hildenborough*
75	B		*Prosthecochloris aestuarii dsm 271*
76	B		*Dinoroseobacter shibae dfl 12*
77	B		*Acidiphilium cryptum jf-5*
78	B		*Anaerolinea thermophila uni-1*
79	B		*Thauera sp. mz1t*
80	B		*Magnetococcus sp. mc-1*
81	B		*Sinorhizobium meliloti 1021*
82	B		*Bordetella petrii dsm 12804*
83	B		*Chloroflexus aggregans dsm 9485*
84	B		*Corynebacterium glutamicum r*
85	B		*Cyanothece sp. pcc 7822*
86	B		*Starkeya novella dsm 506*
87	B		*Arcanobacterium haemolyticum dsm 20595*
88	B		*Rhodopseudomonas palustris dx-1*
89	B		*Rhodospirillum centenum sw*
90	B		*Xanthobacter autotrophicus py2*
91	B		*Mycobacterium leprae br4923*
92	B		*Gluconacetobacter diazotrophicus pal 5*
93	B		*Streptomyces griseus subsp. griseus nbrc 13350*
94	B		*Streptomyces scabiei 87.22*
95	B		*Intrasporangium calvum dsm 43043*
96	B		*Burkholderia rhizoxinica hki 454*
97	B		*Haliangium ochraceum dsm 14365*
98	B		*Salinibacter ruber m8*
99	B		*Rothia dentocariosa atcc 17931*
100	B		*Bifidobacterium animalis subsp. lactis ad011*

a subset is equal to 500, then only COGs containing at least 100 genomes are considered (passed the entry threshold).

(3) *Sampling*. If there are multiple instances of a COG related to the same genome, a median length value for all paralogs from the same genome and from the same COG is used for further processing.

5.3. Sets of Genomes. As of May 2012, there were approximately 1500 NC-numbers, corresponding to 1390 annotated prokaryotic genomes at NCBI. Multiple NC numbers occur for prokaryotes with more than one chromosome, such as *Burkholderia cepacia* (Tax id 269483). This large set was used for the final Bubble sort analysis. In that set, 114 genomes are Archaeal and 1276 are Bacterial. To compare performance of the methods, we used a small subset of this dataset, same as we used previously [8]. Then, we had randomly selected 100 prokaryotic genomes out of a possible 1390, contained at the NCBI COG database. This small set contains 9 Archaeal and 91 Bacterial genomes. The list of selected genomes is shown in Table 3. After the selection of genomes, we discarded those COGs that were present in less than 35% of those selected genomes. Upon filtering, our input contained 1455 COGs. Note, that the input file is a sparse matrix.

5.4. Bubble Sort Ranking (B-Sort). As a LOPI strategy [40] we apply here the regular "bubble sort" procedure [41] interchanging the rows of a given matrix. (In a simulation study on graphs [42], the LOPI strategies found a global maximum of the goal function defined on edges in the majority of the cases.) The criterion by which the procedure decides whether rows would be interchanged is as follows. Comparing two genomes we take into account only those COGs that both genomes have members in them. Comparing pairs of lengths of genes from relevant COGs we count which genome in a pair has longer genes more frequently. In other words, if a genome associated with a row i has longer genes than has a genome associated with a row $i+1$, then these rows would be interchanged. We note that due to application to a sparse matrix this procedure would not necessarily lead to the optimal ordering.

5.5. Solving of the Optimization Problem. The three methods above are pretty intuitive. They do not have a goal to find an optimal ranking but the results have a good chance to be close to the optimal ranking. In our review [8] we described several procedures to find a nearly optimal ranking using approach from the field of combinatorial optimization. Maximization of an average Kendall tau rank correlation coefficient is one of them. As we presented it, the goal is to assign each genome i to a scale x such that x_i most accurately recovers the across-genome gene lengths. "Most accurately" here means achieving the maximum of the function x^τ:

$$x^\tau = \max_x \left[\sum_{k=1}^{K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} C_{ij}\left(\vec{x}, \overrightarrow{r^k}\right) \right], \quad (3)$$

where given a rating vector $\vec{x}$ and an "individual" vector $\overrightarrow{r^k}$ of the gene lengths of COG k, $C_{ij}(\vec{x}, \overrightarrow{r^k})$ is equal to 1, if $(r^k_{x_i} < r^k_{x_j})$, equal to 1/2, if $(r^k_{x_i} = r^k_{x_j})$, and 0-otherwise.

5.6. Kemeny-Optimal Ranking. Kemeny-Optimal Ranking is an *optimal rank aggregation* approach. In [43, 44] the authors proposed a precise criterion for determining the "best" aggregate ranking. Given n objects and k permutations of the objects, $\{\pi_1, \pi_2, \ldots, \pi_k\}$, a *Kemeny optimal* ranking of the objects is the ranking π that minimizes a "sum of distances" $P = \sum_{i=1}^{k} d(\vec{x}, \overrightarrow{r^k})$, where $d(\vec{x}, \overrightarrow{r^k})$, denotes a distance between a rating vector $\vec{x}$ and an "individual" vector $\overrightarrow{r^k}$ based on *Kendall's $\boldsymbol{\tau}$ rank-correlation*. From the properties of Kendall's $\boldsymbol{\tau}$ rank-correlation it follows that a Kemeny optimal ranking minimizes the number of pairwise *disagreements* with the given k rankings x^τ and maximizes sortedness.

It is known that finding a Kemeny optimal ranking is NP-hard [45] and remains NP-hard even when there are only four input lists to aggregate [46]. This motivates the problem of finding a ranking that *approximately* minimizes the number of disagreements with the given input rankings.

Conflict of Interests

The authors declare that there is not conflict of interests regarding the publication of this paper.

Authors' Contribution

Tatiana Tatarinova curated the datasets, implemented the algorithm, carried out the comparison of algorithms, and wrote the paper. Bilal Salih and Irit Cohen implemented the algorithms and participated in writing of the Materials and Methods section. Jennifer Dien Bard provided clinical insight and participated in paper preparation. Alexander Bolshoy led the project, designed the framework, and wrote the paper. All authors read and approved the final paper.

Acknowledgments

Tatiana Tatarinova was supported by NIH: GM068968 and NIH-NICHD: HD070996. The authors would like to thank Professor Roger Jelliffe, USC, for proofreading the paper.

References

[1] H. Willenbrock, C. Friis, A. S. Juncker, and D. W. Ussery, "An environmental signature for 323 microbial genomes based on codon adaptation indices," *Genome Biology*, vol. 7, no. 12, article R114, 2006.

[2] M. Botzman and H. Margalit, "Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles," *Genome Biology*, vol. 12, no. 10, article R109, 2011.

[3] M. Roller, V. Lucić, I. Nagy, T. Perica, and K. Vlahoviček, "Environmental shaping of codon usage and functional adaptation

across microbial communities," *Nucleic Acids Research*, vol. 41, no. 19, pp. 8842–8852, 2013.

[4] J. Zhang, "Protein-length distributions for the three domains of life," *Trends in Genetics*, vol. 16, no. 3, pp. 107–109, 2000.

[5] L. Brocchieri and S. Karlin, "Protein length in eukaryotic and prokaryotic proteomes," *Nucleic Acids Research*, vol. 33, no. 10, pp. 3390–3400, 2005.

[6] M. Wang, C. G. Kurland, and G. Caetano-Anollés, "Reductive evolution of proteomes and protein structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 29, pp. 11954–11958, 2011.

[7] A. A. Vakhrusheva, M. D. Kazanov, A. A. Mironov, and G. A. Bazykin, "Evolution of prokaryotic genes by shift of stop codons," *Journal of Molecular Evolution*, vol. 72, no. 2, pp. 138–146, 2011.

[8] A. Bolshoy and T. Tatarinova, "Methods of combinatorial optimization to reveal factors affecting gene length," *Bioinformatics and Biology Insights*, vol. 6, pp. 317–327, 2012.

[9] A. Bolshoy, B. Salih, I. Cohen, and T. Tatarinova, "Ranking of prokaryotic genomes based on maximization of sortedness of gene lengths," *Journal of Data Mining in Genomics & Proteomics*, vol. 5, article 151, 2014.

[10] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, no. 5338, pp. 631–637, 1997.

[11] Y. I. Wolf, K. S. Makarova, N. Yutin, and E. V. Koonin, "Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer," *Biology Direct*, vol. 7, article 46, 2012.

[12] L. J. Jensen, P. Julien, M. Kuhn et al., "eggNOG: automated construction and annotation of orthologous groups of genes," *Nucleic Acids Research*, vol. 36, no. 1, pp. D250–D254, 2008.

[13] S. Powell, D. Szklarczyk, K. Trachana et al., "eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges," *Nucleic Acids Research*, vol. 40, no. 1, pp. D284–D289, 2012.

[14] A. Bolshoy and Z. Volkovich, "Whole-genome prokaryotic clustering based on gene lengths," *Discrete Applied Mathematics*, vol. 157, no. 10, pp. 2370–2377, 2009.

[15] A. Bolshoy, Z. Volkovich, V. Kirzhner, and Z. Barzily, *Genome Clustering: From Linguistic Models to Classification of Genetic Texts*, Springer, Berlin, Germany, 2010.

[16] N. Gill, S. Singh, and T. C. Aseri, "Computational disease gene prioritization: an appraisal," *Journal of Computational Biology*, vol. 21, no. 6, pp. 456–465, 2014.

[17] D. S. Hochbaum, E. Moreno-Centeno, P. Yelland, and R. A. Catena, "Rating customers according to their promptness to adopt new products," *Operations Research*, vol. 59, no. 5, pp. 1171–1183, 2011.

[18] D. S. Hochbaum and A. Levin, "Methodologies and algorithms for group-rankings decision," *Management Science*, vol. 52, no. 9, pp. 1394–1408, 2006.

[19] D. S. Hochbaum and E. Moreno-Centeno, "Country credit-risk rating aggregation via the separation-deviation model," *Optimization Methods and Software*, vol. 23, no. 5, pp. 741–762, 2008.

[20] A. Muto and S. Osawa, "The guanine and cytosine content of genomic DNA and bacterial evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 1, pp. 166–169, 1987.

[21] T. V. Tatarinova, N. N. Alexandrov, J. B. Bouck, and K. A. Feldmann, "GC3 biology in corn, rice, sorghum and other grasses," *BMC Genomics*, vol. 11, no. 1, article 308, 2010.

[22] T. Tatarinova, E. Elhaik, and M. Pellegrini, "Cross-species analysis of genic GC_3 content and DNA methylation patterns," *Genome Biology and Evolution*, vol. 5, no. 8, pp. 1443–1456, 2013.

[23] I. Lasa and J. Berenguer, "Thermophilic enzymes and their biotechnological potential," *Microbiologia*, vol. 9, no. 2, pp. 77–89, 1993.

[24] M. Skovgaard, L. J. Jensen, S. Brunak, D. Ussery, and A. Krogh, "On the total number of genes and their length distribution in complete microbial genomes," *Trends in Genetics*, vol. 17, no. 8, pp. 425–428, 2001.

[25] K. Tsoni, E. Papadopoulou, E. Michailidou, and I. Kavaliotis, "Campylobacter jejuni meningitis in a neonate: a rare case report," *Journal of Neonatal-Perinatal Medicine*, vol. 6, no. 2, pp. 183–185, 2013.

[26] H. Nadorlik, M. Marcon, K. Koranyi, O. Ramilo, and A. Mejias, "A 2-month-old with bacteremia and gastroenteritis," *Pediatric Infectious Disease Journal*, vol. 31, no. 2, pp. 210–216, 2012.

[27] M. Zhang, L. He, Q. Li et al., "Genomic characterization of the Guillain-Barre syndrome-associated *Campylobacter jejuni* ICDCCJ07001 isolate," *PLoS ONE*, vol. 5, no. 11, Article ID e15060, 2010.

[28] B. Flahou, F. Haesebrouck, A. Smet, H. Yonezawa, T. Osaki, and S. Kamiya, "Gastric and enterohepatic non-Helicobacter pylori Helicobacters," *Helicobacter*, vol. 18, supplement 1, pp. 66–72, 2013.

[29] G. D. Wiens, D. D. Rockey, Z. Wu et al., "Genome sequence of the fish pathogen *Renibacterium salmoninarum* suggests reductive evolution away from an environmental arthrobacter ancestor," *Journal of Bacteriology*, vol. 190, no. 21, pp. 6970–6982, 2008.

[30] J. L. Oliver and A. Marín, "A relationship between GC content and coding-sequence length," *Journal of Molecular Evolution*, vol. 43, no. 3, pp. 216–223, 1996.

[31] E. N. Moriyama and J. R. Powell, "Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*," *Nucleic Acids Research*, vol. 26, no. 13, pp. 3188–3193, 1998.

[32] X. Xia, Z. Xie, and W.-H. Li, "Effects of GC content and mutational pressure on the lengths of exons and coding sequences," *Journal of Molecular Evolution*, vol. 56, no. 3, pp. 362–370, 2003.

[33] X. Xia, H. Wang, Z. Xie, M. Carullo, H. Huang, and D. Hickey, "Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes," *Molecular Biology and Evolution*, vol. 23, no. 7, pp. 1450–1454, 2006.

[34] E. Elhaik and T. Tatarinova, "GC3 biology in eukaryotes and prokaryotes," in *DNA Methylation—From Genomics to Technology*, T. Tatarinova and O. Kerton, Eds., 2012, http://www.intechopen.com/books/dna-methylation-from-genomics-to-technology/gc3-biology-in-eukaryotes-and-prokaryotes.

[35] S. Basak, T. Banerjee, S. K. Gupta, and T. C. Ghosh, "Investigation on the causes of codon and amino acid usages variation between thermophilic Aquifex aeolicus and mesophilic Bacillus subtilis," *Journal of Biomolecular Structure and Dynamics*, vol. 22, no. 2, pp. 205–214, 2004.

[36] R. L. Tatusov, N. D. Fedorova, J. D. Jackson et al., "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, article 41, 2003.

[37] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, 2000.

[38] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev et al., "The COG database: New developments in phylogenetic classification of proteins from complete genomes," *Nucleic Acids Research*, vol. 29, no. 1, pp. 22–28, 2001.

[39] K. Korenblat, Z. Volkovich, and A. Bolshoy, "Robustness of the whole-genome prokaryotic clustering based on gene lengths," *Computational Biology and Chemistry*, vol. 40, pp. 20–29, 2012.

[40] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling, Theory and Applications*, Springer, New York, NY, USA, 2005.

[41] D. E. Knuth, *Art of Computer Programming*, Addison-Wesley, New York, NY, USA, 1973.

[42] P. Groenen, *The majorization approach to multidimensional scaling: some problems and extensions [Ph.D. thesis]*, University of Leiden, 1993.

[43] J. G. Kemeny, "Mathematics without numbers," *Daedalus*, vol. 88, pp. 571–591, 1959.

[44] J. G. Kemeny and J. L. Snell, *Mathematical Models in the Social Sciences*, The MIT Press, Cambridge, UK, 1972.

[45] I. Bartholdi, C. A. Tovey, and M. A. Trick, "Voting schemes for which it can be difficult to tell who won the election," *Social Choice and Welfare*, vol. 6, no. 2, pp. 157–165, 1989.

[46] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*, pp. 613–622, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 510598, 5 pages
http://dx.doi.org/10.1155/2015/510598

Research Article

A Database of Plastid Protein Families from Red Algae and Apicomplexa and Expression Regulation of the *moeB* Gene

Oleg A. Zverkov, Alexandr V. Seliverstov, and Vassily A. Lyubetsky

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoy Karetny Pereulok 19, Moscow 127994, Russia

Correspondence should be addressed to Oleg A. Zverkov; zverkov@iitp.ru

Received 18 June 2014; Revised 29 August 2014; Accepted 13 September 2014

Academic Editor: William H. Piel

We report the database of plastid protein families from red algae, secondary and tertiary rhodophyte-derived plastids, and Apicomplexa constructed with the novel method to infer orthology. The families contain proteins with maximal sequence similarity and minimal paralogous content. The database contains 6509 protein entries, 513 families and 278 nonsingletons (from which 230 are paralog-free, and among the remaining 48, 46 contain at maximum two proteins per species, and 2 contain at maximum three proteins per species). The method is compared with other approaches. Expression regulation of the *moeB* gene is studied using this database and the model of RNA polymerase competition. An analogous database obtained for green algae and their symbiotic descendants, and applications based on it are published earlier.

1. Introduction

The concept of orthology and construction of orthology databases are important areas of bioinformatic research. However, the orthology relationship is not yet decisively formalized and some of its important features may depend on taxonomic context of the data and properties of particular organelles. Mathematically, identification of orthologs corresponds to building clusters in a graph with its vertices assigned gene or protein sequences. The majority of clustering methods utilize various strategies to weight the graph edges with subsequent construction of "highly connected components," that is, clusters resulting from a certain clustering procedure.

The edge weight reflects similarity of amino acid sequences generated in various pairwise alignment procedures, intron content and positioning, protein domain architecture, gene synteny, and so forth. Usually the weights are computed with global alignment using the Needleman-Wunsch algorithm, or local alignment using BLAST. Various clustering approaches were proposed, from specifically organized partitioning of the spanning tree of the initial graph (the originally proposed algorithm ClusterZSL, refer to [1]) to time estimation of random walk on a graph (the OrthoMCL algorithm). In the latter algorithm based on Markov clustering, walk within a cluster is long, and jumps between clusters are rare [2]. Due to heuristic nature of these processes, comparison of the algorithms cannot be formalized, especially in the absence of standard benchmarking data. The description of OrthoMCL implicitly states that its convergence is difficult to discuss even in hypothesis.

The algorithm ClusterZSL essentially differs from commonly employed methods, including OrthoMCL, by not using the mutual-best-hit criterion. For a pair of genomes, a gene may produce none or many best hits; the latter is especially the case when considering suboptimal hits that may in fact represent true orthologs. In contrast with other methods, ClusterZSL also minimizes the amount of paralogs in each cluster that in general seems a reasonable property. ClusterZSL can consider gene positioning in DNA and orthologous context of the gene neighborhood. A version of this algorithm that uses gene synteny was applied to various chordate animals and will be described in a separate publication.

The algorithm ClusterZSL and its computer program implementation possess the computational complexity of maximum n^2 accurate to a coefficient. The OrthoMCL uses matrix multiplication, the operation with the minimal complexity

Table 1: Orthologs of *moeB* in plastids of rhodophyte algae as inferred with ClusterZSL and their genomic neighborhoods.

Class	Species	Locus	Protein MoeB	Genomic context
Bangiophyceae	*Porphyra purpurea*	NC_000925	NP_053945.1	(*trnW*)-*ORF75*-*moeB*
Bangiophyceae	*Porphyridium purpureum*	NC_023133	YP_008965710.1	(*ORF144*)-*ycf38*-*moeB*
Bangiophyceae	*Pyropia haitanensis*	NC_021189	YP_007947865.1	(*trnW*)-*ORF75*-*moeB*
Bangiophyceae	*Pyropia perforata*	NC_024050	YP_009027619.1	(*trnW*)-*ORF75*-*moeB*
Bangiophyceae	*Pyropia yezoensis*	NC_007932	YP_537017.1	(*trnW*)-*moeB*
Bangiophyceae	*Cyanidioschyzon merolae*	NC_004799	NP_849016.1	(*trnW*)-*moeB*
Bangiophyceae	*Cyanidium caldarium*	NC_001840	NP_045115.1	(*trnW*)-*moeB*
Florideophyceae	*Calliarthron tuberculosum*	NC_021075	YP_007878185.1	(*trnW*)-*moeB*
Florideophyceae	*Chondrus crispus*	NC_020795	YP_007627343.1	(*trnW*)-*moeB*
Florideophyceae	*Gracilaria salicornia*	NC_023785	YP_009019560.1	(*trnW*)-*moeB*
Florideophyceae	*Gracilaria tenuistipitata*	NC_006137	YP_063552.1	(*trnW*)-*moeB*
Florideophyceae	*Grateloupia taiwanensis*	NC_021618	YP_008144807.1	(*trnW*)-*moeB*

The *moeB* orthologs are also denoted by *moeB*, irrespective of corresponding original annotations. Genes on the opposite strand to *moeB* are given in brackets.

n^{ω}, where the exponent ω is a parameter. For the Gauss algorithm $\omega = 3$ and for the Strassen algorithm $\omega = \log_2 7 \approx 2.81$ [3]. An asymptotically faster algorithm is known, which, however, takes advantage only with matrices of very high order and is practically of little use [4]; also refer to [5, 6]. Further concerns with the OrthoMCL algorithm are the estimation of the number of iterations (including matrix multiplications) and proof of convergence. The convergence requirement is obviously met with ClusterZSL. The running time of OrthoMCL appears to be much longer than that of ClusterZSL, at least with our testing data. Due to high scalability, performance of ClusterZSL does not depend on the amount of CPUs, which is a valuable practical property; the authors are unaware of attempts to assess the scalability of OrthoMCL.

Compare ClusterZSL with the algorithm used in the Ensembl database. Both start from the spanning tree. On later stages, the Ensembl algorithm relies in many respects on multiple alignments of leaf proteins, the task exponential in computational complexity if the alignment is optimized [7]. For alignment construction, the algorithm integrates the *M*-Coffee algorithm [8] or Mafft for larger data [9]. Both mentioned alignment procedures are heuristic and do not guarantee global minimization of the used functional. The ClusterZSL algorithm does not utilize multiple alignment.

Worth mentioning is another clustering method to establish orthology that was previously used by the authors. When the size of the clusters is known, for example, in studies of multicomponent systems where the length of the orthologous series is known for one component, the most dense cluster of the known size is constructed using the algorithm described in [10, 11]. We do not compare with phylogenetic methods here; for instance, refer to [12]. Note that the phylogenetic position of a species or protein belonging to any species is not always known.

The problem of the transcription factor regulon definition is of great interest. In red algae, the only plastid-encoded transcription factors are Ycf27, Ycf28, Ycf29, and RbcR (Ycf30). Of little information on them, the RbcR binding sites are known to vary even among close species [13], which hampers their detection. We will consider this problem on the example of the factor Ycf28, which, as it turned out, regulates the expression of the gene *moeB*.

In this study, the gene *moeB*, which is itself an important object of research, is tackled in a case study of gene expression regulation using ClusterZSL. This gene encodes an E1-like family enzyme involved in molybdopterin and thiamine biosynthesis. This family includes proteins that catalyze the adenylation by ATP of the carboxyl group of the C-terminal glycine in sulfur carrier proteins, for example, MoaD or ThiS. Bacterial proteins with domains characteristic for this family are described in [14]. The *moeB* gene is present in plastids of all sequenced Rhodophyta; refer to Table 1. Its ortholog in *Porphyra purpurea* and *Pyropia* spp. is *ORF382*, in *Cyanidium caldarium chlN*. In *P. perforata* the neighboring genes *moeB* and *ORF382* encode the N- and C-termini of the MoeB protein.

As evident from Table 1, the neighbor of *moeB* on the opposite strand is *trnW* that encodes the tryptophanyl-tRNA. In *Porphyra purpurea*, *Pyropia haitanensis*, and *Pyropia perforata* the genes *trnW* and *moeB* are separated by the short coding frame *ORF75*. The only exception is *Porphyridium purpureum*, where the neighborhood of *moeB* lacks a reliably highly transcribed gene on the opposite strand; refer to [15–23].

In this study we describe a database ClusterZSL of orthologous plastid proteins in red algae, secondary and tertiary rhodophyte-derived plastids, and Apicomplexa (the RedLine at May 2014 from the GenBank; also refer to http://lab6.iitp.ru/ppc/redline50/), constructed with the same algorithm ClusterZSL.

We use it in a case study of transcription regulation of the *moeB* gene. An analogous database obtained for green algae and their symbiotic descendants (the green line) and its applications are published in [1, 24–26].

Some recent papers ([27] et al.) glance upon plastid proteins the database CpBase, http://chloroplast.ocean.washington.edu/. It represents 35 plastomes from RedLine in comparison with 50 plastomes represented in the database ClusterZSL. The authors are not aware of the description of

Table 2: Orthologs of Ycf28 in plastids of Rhodophyta as inferred with ClusterZSL.

Class	Species	Locus	Protein Ycf28	Bit score	*E*-value
Bangiophyceae	*Porphyra purpurea*	NC_000925	NP_053952.1	50.9	$9.5e-14$
Bangiophyceae	*Porphyridium purpureum*	NC_023133	YP_008965713.1	48.6	$5.0e-13$
Bangiophyceae	*Pyropia haitanensis*	NC_021189	YP_007947872.1	52.9	$2.2e-14$
Bangiophyceae	*Pyropia perforata*	NC_024050	YP_009027626.1	53.3	$1.7e-14$
Bangiophyceae	*Pyropia yezoensis*	NC_007932	YP_537023.1	55.2	$4.4e-15$
Bangiophyceae	*Cyanidioschyzon merolae*	NC_004799	NP_849012.1	29.5	$4.5e-07$
Bangiophyceae	*Cyanidium caldarium*	NC_001840	NP_045121.1	55.8	$2.8e-15$
Florideophyceae	*Calliarthron tuberculosum*	NC_021075	YP_007878179.1	43.9	$1.5e-11$
Florideophyceae	*Chondrus crispus*	NC_020795	YP_007627337.1	31.7	$9.1e-08$
Florideophyceae	*Gracilaria salicornia*	NC_023785	YP_009019566.1	29.0	$6.5e-07$
Florideophyceae	*Gracilaria tenuistipitata*	NC_006137	YP_063558.1	32.6	$4.8e-08$
Florideophyceae	*Grateloupia taiwanensis*	NC_021618	YP_008144797.1	33.6	$2.4e-08$

The last two columns contain estimates for the Pfam Crp-like helix-turn-helix domain (PF13545).

the method, which the CpBase has been constructed with, as well as the details related to it.

2. Materials and Methods

All plastid proteins are available in GenBank [28]. Orthology was established with the ClusterZSL algorithm described in [1] and applied previously in [24–26]. The algorithm parameters were set to $H = 0.6$, $L = 0$. Gene annotations were verified with the Pfam [29] and Prosite [30] databases.

Promoters were predicted using an algorithm described in [24, 31, 32]. For different σ-subunits of bacterial type RNA polymerases it utilizes data on mutation profiles of the *psbA* promoter in *Sinapis alba* [33] and other experimentally studied promoters [34].

In searches for motifs in the $5'$-leader regions of *moeB* we used the original algorithm published in [35, 36] and the WEB service MEME [37], although the motifs were not detected.

The notion of the phylogenetic distribution (profile) is defined in [26]: for a given gene/protein g, it is a function on a given set S of species that equals (for all s from S) $+1$ if g is present in s, and -1 otherwise.

In Section 3 we essentially exploit the originally proposed model of RNA polymerase competition [38, 39]. The model describes the following situation. In DNA locus transcription many RNA polymerases involved simultaneously bind with the promoters of their type and elongate along their chains, possibly towards each other. This leads to the interaction of RNA polymerases, both between each other and with various protein and structural factors on DNA and RNA. As a result, the transcription levels of the genes significantly change, right up to inability to initiate the transcription of the divergent located gene (below in this role *moeB*), when an actively transcribed gene (resp., *trnW*) plays against it, provided the intergenic region is not organized in a special way.

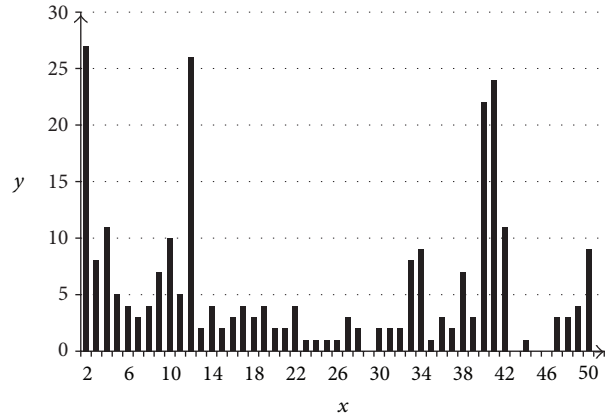


Figure 1: Distribution of cluster (y-axis, the ordinate) versus species (x-axis, the abscissa) numbers.

3. Results

We report the database ClusterZSL (http://lab6.iitp.ru/ppc/) of plastid protein families from red algae, secondary and tertiary rhodophyte-derived plastids, and Apicomplexa (the RedLine). The families contain proteins with maximal sequence similarity and minimal paralogous content and are built using the ClusterZSL algorithm. The database contains 6005 protein entries, 513 families, and 278 nonsingletons (from which 230 are paralog-free, and among the remaining 48, 46 contain at maximum two proteins per species and 2 at maximum three proteins per species). The comparison of the obtained protein families with the biological annotations indicates their good conformity.

Tables 1 and 2 describe two clusters of the database. Figure 1 presents a diagram of species content in inferred clusters.

Standard bacterial type promoters were not detected in the $5'$-leader regions of *moeB*. However, the {A, T}-rich

regions found upstream *moeB* may represent functioning −10 promoter boxes. Based on modeling RNA polymerases competition we suggest that the promoters of *moeB* are located in between *moeB* and *trnW* (refer to the Conclusions) and differ distinctly from the common template.

The presented database allows comparing a cluster of a gene (e.g., *moeB*) with all other clusters. Phylogenetic distributions of *moeB* and Ycf28 coincide; for example, there is a unique transcription factor, which is encoded in a plastid if and only if *moeB* is encoded in it; it is Ycf28. That indicates that the best hit against *moeB* is Ycf28, a transcription factor.

The lack of detected −35 box for *moeB* naturally suggests that Ycf28 is an activator. Based on the same modeling, we surmise that the Ycf28 binding sites are located in between genes *moeB* and *trnW*. The only exception might be *Porphyridium purpureum*. The Ycf28 binding motif itself was not identified, probably due to the variability of binding sites.

Note that the 5′-UTRs of *moeB* are usually short and allow for very limited secondary RNA folding [40]. No conserved structures potentially regulating translation initiation were found that also suggests presence of transcription regulation.

4. Conclusions

The Ycf28 proteins are present in plastids of all Rhodophyta; refer to Table 2. In *Cyanidioschyzon merolae* and *Porphyridium purpureum* this protein is notably shorter.

In the presented database, phylogenetic distributions of *moeB* and transcription factor Ycf28 coincide. This observation leads to the suggestion that Ycf28 is a transcription regulation factor for *moeB*. The factor Ycf28 is a close homolog of the cyanobacterial transcription factor NtcA involved in regulation of nitrogen metabolism [41, 42]. Among cyanobacterial genes under the NtcA regulation only two have homologs in plastids. These are the genes of the factor itself and the regulatory protein GlnB from the family PII [43]. However, GlnB is rarely found in plastids, and the corresponding 5′-UTRs lack the conserved motif typically binding NtcA in cyanobacteria [41, 42]. This may suggest that the plastid-encoded Ycf28 and cyanobacterial NtcA are involved in different regulations.

In most species, presence of the actively transcribed tRNA gene *trnW* on the opposite strand precludes *moeB* transcription from a promoter located upstream that of *trnW* due to inevitable strong RNA polymerase competition. An important role of such competition in expression of closely located antidirected genes is substantiated in modelling and various experiments on gene expression. Such evidence includes data on bacterial type RNA polymerases σ-subunit knockout in plastids of *Arabidopsis thaliana* and data for mitochondrial RNA polymerases of the phage type [38, 39]. Therefore, the *moeB* promoter is likely to be located in between genes *moeB* and *trnW* and requires transcription initiation due to absence of an evident −35 box. Considering polymerase competition at these genes, the transcription factor binding site is likely to occur in the same region between the genes. Indeed, a binding site within an intensively transcribed region is unlikely effective due to interference of the factor with RNA polymerases.

Notably, short conserved motifs adjoining {A, T}-rich regions at their 3′-end are commonly found upstream *moeB*. This may be related to a low GC-content in plastids of most species. However, the predicted location of the binding site makes the putative mechanism of expression regulation specific to *moeB*.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

The authors are deeply grateful to the editor of the paper for his valuable comments. Also the authors would like to thank L. Rusin for valuable discussions and help with preparing the paper. Research was partly funded by the Russian Foundation for Basic Research (Grant 13-04-40196-H).

References

[1] V. A. Lyubetsky, A. V. Seliverstov, and O. A. Zverkov, "Elaboration of the homologous plastid-encoded protein families that separate paralogs in magnoliophytes," *Mathematical Biology and Bioinformatics*, vol. 8, no. 1, pp. 225–233, 2013 (Russian).

[2] S. van Dongen and C. Abreu-Goodger, "Using MCL to extract clusters from networks," *Methods in Molecular Biology*, vol. 804, pp. 281–295, 2012.

[3] V. Strassen, "Gaussian elimination is not optimal," *Numerische Mathematik*, vol. 13, pp. 354–356, 1969.

[4] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Journal of Symbolic Computation*, vol. 9, no. 3, pp. 251–280, 1990.

[5] F. Le Gall, "Powers of tensors and fast matrix multiplication," in *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation (ISSAC '14)*, pp. 296–303, July 2014.

[6] A. V. Smirnov, "The bilinear complexity and practical algorithms for matrix multiplication," *Computational Mathematics and Mathematical Physics*, vol. 53, no. 12, pp. 1781–1795, 2013.

[7] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney, "EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates," *Genome Research*, vol. 19, no. 2, pp. 327–335, 2009.

[8] I. M. Wallace, O. O'Sullivan, D. G. Higgins, and C. Notredame, "M-Coffee: Combining multiple sequence alignment methods with T-Coffee," *Nucleic Acids Research*, vol. 34, no. 6, pp. 1692–1699, 2006.

[9] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[10] A. E. Galashov and A. V. Kel'manov, "A 2-approximate algorithm to solve one problem of the family of disjoint vector subsets," *Automation and Remote Control*, vol. 75, no. 4, pp. 595–606, 2014.

[11] A. V. Kel'manov and S. M. Romanchenko, "FPTAS for solving a problem of search for a vector subset," *Diskretnyi Analiz i Issledovanie Operatsii*, vol. 21, no. 3, pp. 41–52, 2014 (Russian).

[12] C. M. Zmasek and S. R. Eddy, "RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs," *BMC Bioinformatics*, vol. 3, article 14, 2002.

[13] A. Minoda, A. P. M. Weber, K. Tanaka, and S.-Y. Miyagishima, "Nucleus-independent control of the rubisco operon by the plastid-encoded transcription factor Ycf30 in the red alga *Cyanidioschyzon merolae*," *Plant Physiology*, vol. 154, no. 3, pp. 1532–1540, 2010.

[14] M. S. Cortese, A. B. Caplan, and R. L. Crawford, "Structural, functional, and evolutionary analysis of *moeZ*, a gene encoding an enzyme required for the synthesis of the *Pseudomonas* metabolite, pyridine-2,6-bis(thiocarboxylic acid)," *BMC Evolutionary Biology*, vol. 2, no. 1, article 8, 2002.

[15] J. Collén, B. Porcel, W. Carré et al., "Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 13, pp. 5247–5252, 2013.

[16] M. S. Depriest, D. Bhattacharya, and J. M. López-Bautista, "The plastid genome of the red macroalga *Grateloupia taiwanensis* (Halymeniaceae)," *PLoS ONE*, vol. 8, no. 7, Article ID e68246, 2013.

[17] G. Glöckner, A. Rosenthal, and K. Valentin, "The structure and gene repertoire of an ancient red algal plastid genome," *Journal of Molecular Evolution*, vol. 51, no. 4, pp. 382–390, 2000.

[18] J. C. Hagopian, M. Reis, J. P. Kitajima, D. Bhattacharya, and M. C. De Oliveira, "Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids," *Journal of Molecular Evolution*, vol. 59, no. 4, pp. 464–477, 2004.

[19] J. Janouškovec, S.-L. Liu, P. T. Martone et al., "Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers," *PLoS ONE*, vol. 8, no. 3, Article ID e59001, 2013.

[20] N. Ohta, M. Matsuzaki, O. Misumi et al., "Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*," *DNA Research*, vol. 10, no. 2, pp. 67–77, 2003.

[21] M. E. Reith and J. Munholland, "Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome," *Plant Molecular Biology Reporter*, vol. 13, no. 4, pp. 333–335, 1995.

[22] L. Wang, Y. Mao, F. Kong et al., "Complete sequence and analysis of plastid genomes of two economically important red algae: *Pyropia haitanensis* and *Pyropia yezoensis*," *PLoS ONE*, vol. 8, no. 5, Article ID e65902, 2013.

[23] M. A. Campbell, G. Presting, M. S. Bennett, and A. R. Sherwood, "Highly conserved organellar genomes in the Gracilariales as inferred using new data from the Hawaiian invasive alga *Gracilaria salicornia* (Rhodophyta)," *Phycologia*, vol. 53, no. 2, pp. 109–116, 2014.

[24] V. A. Lyubetsky, A. V. Seliverstov, and O. A. Zverkov, "Transcription regulation of plastid genes involved in sulfate transport in viridiplantae," *BioMed Research International*, vol. 2013, Article ID 413450, 6 pages, 2013.

[25] O. A. Zverkov, L. Y. Rusin, A. V. Seliverstov, and V. A. Lyubetsky, "Study of direct repeats in micro evolution of plant mitochondria and plastids based on protein clustering," *Moscow University Biological Sciences Bulletin*, vol. 68, no. 2, pp. 58–62, 2013.

[26] O. A. Zverkov, A. V. Seliverstov, and V. A. Lyubetsky, "Plastid-encoded protein families specific for narrow taxonomic groups of algae and protozoa," *Molecular Biology*, vol. 46, no. 5, pp. 717–726, 2012.

[27] S. R. Starkenburg, K. J. Kwon, R. K. Jha et al., "A pangenomic analysis of the Nannochloropsis organellar genomes reveals novel genetic variations in key metabolic genes," *BMC Genomics*, vol. 15, no. 1, article 212, 2014.

[28] D. A. Benson, M. Cavanaugh, K. Clark et al., "GenBank," *Nucleic Acids Research*, vol. 41, no. 1, pp. D36–D42, 2013.

[29] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D290–D301, 2012.

[30] C. J. A. Sigrist, E. de Castro, L. Cerutti et al., "New and continuing developments at PROSITE," *Nucleic Acids Research*, vol. 41, no. 1, pp. D344–D347, 2013.

[31] V. A. Lyubetsky, L. I. Rubanov, and A. V. Seliverstov, "Lack of conservation of bacterial type promoters in plastids of Streptophyta," *Biology Direct*, vol. 5, article 34, 2010.

[32] A. V. Seliverstov, E. A. Lysenko, and V. A. Lyubetsky, "Rapid evolution of promoters for the plastome gene *ndhF* in flowering plants," *Russian Journal of Plant Physiology*, vol. 56, no. 6, pp. 838–845, 2009.

[33] A. Homann and G. Link, "DNA-binding and transcription characteristics of three cloned sigma factors from mustard (*Sinapis alba* L.) suggest overlapping and distinct roles in plastid gene expression," *European Journal of Biochemistry*, vol. 270, no. 6, pp. 1288–1300, 2003.

[34] E. A. Lysenko, "Plant sigma factors and their role in plastid transcription," *Plant Cell Reports*, vol. 26, no. 7, pp. 845–859, 2007.

[35] V. A. Lyubetsky and A. V. Seliverstov, "Some algorithms related to finite groups," *Information Processes*, vol. 3, no. 1, pp. 39–46, 2003 (Russian).

[36] V. A. Lyubetsky and A. V. Seliverstov, "Note on cliques and alignments," *Information Processes*, vol. 4, no. 3, pp. 241–246, 2004.

[37] T. L. Bailey, M. Boden, F. A. Buske et al., "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. 2, pp. W202–W208, 2009.

[38] V. A. Lyubetsky, O. A. Zverkov, L. I. Rubanov, and A. V. Seliverstov, "Modeling RNA polymerase competition: the effect of σ-subunit knockout and heat shock on gene transcription level," *Biology Direct*, vol. 6, article 3, 2011.

[39] V. A. Lyubetsky, O. A. Zverkov, S. A. Pirogov, L. I. Rubanov, and A. V. Seliverstov, "Modeling RNA polymerase interaction in mitochondria of chordates," *Biology Direct*, vol. 7, article 26, 2012.

[40] A. A. Vladimirov, "Non-crossing matchings," *Problems of Information Transmission*, vol. 49, no. 1, pp. 54–57, 2013.

[41] M. I. Muro-Pastor and F. J. Florencio, "Regulation of ammonium assimilation in cyanobacteria," *Plant Physiology and Biochemistry*, vol. 41, no. 6-7, pp. 595–603, 2003.

[42] K. V. Lopatovskaya, A. V. Seliverstov, and V. A. Lyubetsky, "NtcA and NtcB regulons in cyanobacteria and rhodophyta chloroplasts," *Molecular Biology*, vol. 45, no. 3, pp. 522–526, 2011.

[43] K. Forchhammer, "PII signal transducers: novel functional and structural insights," *Trends in Microbiology*, vol. 16, no. 2, pp. 65–72, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 962637, 7 pages
http://dx.doi.org/10.1155/2015/962637

Research Article

miR-1322 Binding Sites in Paralogous and Orthologous Genes

Raigul Niyazova, Olga Berillo, Shara Atambayeva, Anna Pyrkova, Aigul Alybayeva, and Anatoly Ivashchenko

National Nanotechnology Laboratory, Al-Farabi Kazakh National University, 71 Al-Farabi, Almaty 050038, Kazakhstan

Correspondence should be addressed to Anatoly Ivashchenko; a_ivashchenko@mail.ru

Received 10 September 2014; Accepted 19 November 2014

Academic Editor: Vassily Lyubetsky

We searched for 2,563 microRNA (miRNA) binding sites in 17,494 mRNA sequences of human genes. miR-1322 has more than 2,000 binding sites in 1,058 genes with $\Delta G/\Delta G_m$ ratio of 85% and more. miR-1322 has 1,889 binding sites in CDSs, 215 binding sites in 5′ UTRs, and 160 binding sites in 3′ UTRs. From two to 28 binding sites have arranged localization with the start position through three nucleotides of each following binding site. The nucleotide sequences of these sites in CDSs encode oligopeptides with the same and/or different amino acid sequences. We found that 33% of the target genes encoded transcription factors. miR-1322 has arranged binding sites in the CDSs of orthologous *MAMLD1*, *MAML2*, and *MAML3* genes. These sites encode a polyglutamine oligopeptide ranging from six to 47 amino acids in length. The properties of miR-1322 binding sites in orthologous and paralogous target genes are discussed.

1. Introduction

Interest in microRNAs (miRNAs) is constantly growing, and new data supplement existing knowledge about the role of these molecules in key biological processes. The main objective of these studies is to identify miRNA binding sites and evaluate their binding affinities. The characteristics of binding sites shed light on the biological role of miRNAs and have practical applications. It is possible to predict interactions between miRNAs and mRNAs and their properties by using computational methods [1]. It has been established that miRNAs bind to mRNAs predominantly in 3′-untranslated regions (3′ UTRs) [2]. They can also bind to 5′-untranslated regions (5′ UTRs) and coding domain sequences (CDSs) [3, 4]. Moreover, some miRNAs have binding sites in 5′ UTRs, CDSs, and 3′ UTRs [5]. For example, miR-3960 binding sites are mainly in CDSs, and many are positioned adjacent to each other (through one, two, three, or more nucleotides) [6]. Such mRNA fragments can consist of 2–17 binding sites. Discussed in this paper is miR-1322 which also contains multiple sites in CDSs. Clusters of miRNAs binding sites located in the CDS of genes are unexpected because proteins have specific amino acid sequences that are evolutionarily conserved. The presence of multiple binding sites in close proximity significantly increases the probability of interactions between miRNAs and mRNAs, even if mutations occur. Many miRNAs regulate the expression of genes involved in tumorigenesis [7–11]. For example, changes in miRNA concentrations are observed during the development of lung cancer [7, 8], breast cancer [9], gastrointestinal cancer [10], and other cancers [11]. The serum level of miR-1322 is a potential diagnostic biomarker for squamous cell carcinoma of the esophagus [12]. We studied the arrangement and evolution of miR-1322 binding sites in genes involved in disease.

2. Materials and Methods

The nucleotide sequences of precursor mRNAs (pre-mRNAs) of human genes (*Homo sapience* (Hsa)) and mammal genes (*Bos mutus* (Bmu), *Bos taurus* (Bta), *Cricetulus griseus* (Cgr), *Cavia porcellus* (Cpo), *Equus caballus* (Eca), *Felis catus* (Fca), *Gorilla gorilla* (Ggo), *Heterocephalus glaber* (Hgl), *Macaca mulatta* (Mul), *Macaca fascicularis* (Mfa), *Nomascus leucogenys* (Nle), *Pongo abelii* (Pab), *Papio anubis* (Pan), *Pan paniscus* (Ppa), *Pan troglodytes* (Ptr), *Rattus norvegicus* (Rno), and *Tupaia chinensis* (Tch)) were downloaded from NCBI GenBank (http://www.ncbi.nlm.nih.gov) in FASTA format. Nucleotide sequences of human mature miR-1322 were

```
     3′      GUCGUAGUCGUCGUAGUAG  5′  mir-1322
5′           ||||| |||||||| || |          3′
GAAUUUCUACCAGCAGCAGCAGCAGCAGCAACAACAACA  AAK1 2101
.G.GAAGG.G.............A........G..G..   ABCF1 284
CC.GCAAGGG.....................G.GG..G..  AFF3 1471
...GCAG.UA...................CC.C.UCG..   AKAP2 1221
.CCGCAGCCG.............C.....CC..G.CG.C  ALX4 425
A.UUCAACUU.....A........A...UCUUGC.....  ANKRD17 2811
AUGUGGAG.G.................UCGGG.GG.AC   ANKS3 1281
.UGACAGUGA.....A......UCAUCAUC.ACCUCGUG  ANO2 1801
.CUGC.GCUG.....................G..G..G..  AR 1286
AUUGAAGA.G.............AUCAGC.A.G....G   ARGFX 500
CU.CCC.CCG.....................G....G...  ARID1A 4351
.G.GCCACUU.....................G..G.....  ARID3B 213
A..CCAG.U.....................G.....G..   ARID1B 339
..GCGCGC.G.....................G..G..G..  ASCL1 721
.C.GCAGC.G.....................U..CGG.A.  AUN1 1725
.C.GCAGC.G...................CG.CG.CU.C   AUXN7 652
C.GC.G...G.....................UG.UC.C   B4GALU2 538
ACGAGG.CU.....................G..G..G..   BCL6B 763
.C.GCGGCGG.....................G..G..G..  BHLHE22 1227
UCC.CA.C.G.....................G.....G..  BMP2K 1543
.C.GCAGG.GG...................G.UG.CU.G   BMP6 514
UG.G.GAG.G...................UCAUCAG.GU   BRDU 2624
UC.ACAGC.A.....................U.UGAG.AG  BUBD7 2874
.C.GCAGCGG.....................GA.GA.     C9ORF43 1295
ACCACCUC.A.....................G..G.....   CELF3 1871
.G.GCAGC.G.................U...U.G.CCGG   FAM104A 504
C.GC.AUG.....................G..G..G..     RAI1 1300
AGGGGAGC.G.....................CU.CC.C    SOCS7 661
A.GACCAA.G.....................G..G..C.U   SRP14 394
AUU.C.AG.G.....................CU..GUC     SUSD4 226
.C...A.AUG.....................G..G..       UFEB 404
AC.GCAGC.G...................GGAGGGGCGC     UMEM245 1048
..G.GCACUG.....................G..GAGG..   UNRC6B 4171
..UGCAA.UG.....................             UOX3 1508
.UGG.GGAGG.....................G.GCUUUCU.   USC1 3341
C.UGAA.C.....................GA..GCGGG      USP7 208
U..CCAAA.G.....................G..G..G..   VEZF1 1151
.C.ACC.CC.....................G....C...     ZFHX3 10261
CU.CCG.AGU.....................GGGC.UGGC    ZFP36L2 1473
.C.GCAGC.G.....................G.C..C       ZNF384 1794
```

Figure 1: The arrangement of miR-1322 binding sites in CDSs of human target genes.

downloaded from the miRBase database (http://mirbase.org) [13].

Target genes for miR-1322 were determined using the MirTarget program [6], which was developed in our laboratory. This program defines the following features of binding sites: (a) the start position of an miRNA binding site with respect to the mRNA sequence; (b) the localization of miRNA binding sites in 5′ UTRs, CDSs, and 3′ UTRs of genes; (c) the free energy of hybridization (ΔG, kJ/mole); and (d) the schemes of nucleotide interactions between miRNAs and mRNAs. The ratio $\Delta G/\Delta G_m$ (%) was estimated for each binding site, where ΔG_m is equal to the value of free energy of an miRNA binding with its perfect complementary nucleotide sequence. One family of miRNAs have nucleotide sequences with the level of homology of 85% or more. Therefore we used the $\Delta G/\Delta G_m$ ratios of 85% or more. We also noted the positions of the binding sites on the mRNA, beginning from the first nucleotide of the 5′ UTR. The MirTarget program predicts interactions between the nucleotides of miRNAs and those of target gene mRNAs. It found bonds between adenine (A) and uracil (U), guanine (G) and cytosine (C), and G and U, as well as between A and C via a hydrogen bond [14]. The TmiROSite program was used to identify mRNA fragments that have miRNA binding sites and to define the corresponding amino acid sequences [15].

3. Results and Discussion

3.1. Features of miR-1322-3p Binding Sites. miR-1322 has a length of 19 nucleotides (nt) and a GC-content of 53%. The maximum free energy of miR-1322 binding with mRNAs is −101.9 kJ/mole. We found that miR-1322 has 2,264 binding sites on 1,058 target mRNAs with a $\Delta G/\Delta G_m$ ratio of 85% or more. Of those, 160 miR-1322 binding sites are located in the 3′ UTRs of 130 genes, 215 binding sites are located in the 5′ UTRs of 109 genes, and 1,889 binding sites are located in the CDSs of 819 genes. The average number of binding sites in the CDS of a single gene is 2.3, which is almost two times higher than the average number of binding sites in 3′ UTRs.

The maximum number of sites observed in 3′ UTR is eight in *CACN1A* and five in *PDYN* and *S100A16*. The maximum number of sites in 5′ UTR was 13 in *MAB21L1*, and the *AMOT*, *BACH2*, *CAPNG*, *PIM1*, *RBM39*, and *STC1* genes have five sites. Characteristics of the clusters of five or more binding sites located in CDSs are shown in Table 1. The start points of several miR-1322 binding sites are located through three nucleotides of each other. Several such sites in mRNA form a cluster and increase the probability of binding and the ability to inhibit protein synthesis. Oligonucleotides of binding sites located in CDSs can encode polyglutamine, polyalanine, or polyserine depending on the open reading frame (Table 1). These data indicate the importance of conserved nucleotide sequences of miR-1322 binding sites and not only the amino acid sequence corresponding to oligopeptides of the encoded protein.

The arranged nucleotide sequences of the CDSs contain binding sites for miR-1322 (Figure 1). The conservation of binding sites relative to the adjacent regions of CDSs is shown in Figure 2. It is important to establish the presence of miR-1322 binding sites for paralogous and orthologous mRNA sequences. Additionally, the properties of binding sites were studied for mRNA sequences of both human and other animal species.

The $\Delta G/\Delta G_m$ ratio for all miR-1322 binding sites of the *ANO2* gene is 95.8%. The nucleotide fragment alignments of the CDSs containing miR-1322 binding sites for 38 genes are shown in Figure 1. Characteristics of the binding sites with start points located through three nucleotides in 5′ UTRs and 3′ UTRs are shown in Table 2. The number of binding sites in 5′ UTRs ranged from five to 13. Consequently, these untranslated regions have an increased probability of binding with miR-1322. The $\Delta G/\Delta G_m$ ratio ranged from 85.4% to 91.7% (Table 2). Therefore, expression of these genes can be controlled extensively by miR-1322.

Transcription factors represent 33% of all target genes in this study (Figure 1 and Tables 1 and 2). Inhibition of the synthesis of proteins can cause diseases, including cancer. Unfortunately, experimental data on miR-1322 binding sites are insufficient; however, some previous studies confirm the high efficacy of the predictions of the MirTarget program

Figure 2: Nucleotide variation in the miR-1322 binding sites in the CDSs of human target genes.

```
1  Ggo SLTPTSNLLSQQQQQQQQQQQQQQQQQQQQQQQQQQQQANAIFKPMSSNSSKTLSMIMQQGMASSSPGATEPFTF
2  Hsa SLTPTSNLLSQQQQQQQQQQQ-----------------ANVIFKPISSNSSKTLSMIMQQGMASSSPGATEPFTF
3  Ptr SLTPTSNLLSQQQQQQQQQQQ-----------------ANAIFKPMSSNSSKTLSMIMQQGMASSSPGATEPFTF
4  Pab SLTPTSNLLSQQQQQQQQQQQQQQQQQQQQQQ------ANAIFKPMSSNSSKTLSMIMQQGMASSSPGATEPFTF
5  Mul SLTPTNNLLSQQQQQQQQQQQQQQQ-------------ANAIFKPMNSNSSKTLSMIMQQGMASSSPGATEPFTF
6  Pan SLTPTSNLLSQQQQQQQQQQQQQQQ-------------ANAIFKPMNSNSSKTLSMIMQQGMASSSPGATEPFTF
7  Ppa SLTPTSNLLSQQQQQQQQQQQ-----------------ANAIFKPMSSNSSKTLSMIMQQGMASSSPGATEPFTF
8  Nle SLTPTSNLLSQQQQQQQQQQQQQQQQQQQQQQ------ANAIFKPMSSNSSKTLSMIMQQGMASSSPGATEPFTF
9  Cpo SLTPTSNLLSQQQQQQQQQQQQQQQQQQQQQQQQ----SNSIFKPMTSNSSKTLSMLMHQGLASSSPEAPEPFTF
10 Eca SLTPASNPLSQQQQQQQQQQQQQQ--------------ANAVFKPMVTNSPKTLSMIMHQGLASPSPGAPEPFSF
11 Hgl SLTPTSNLLGQQQQQQQQQ-------------------ANAIFKPMTSNSSKTLSMLMHQGLASSSPEASEPFTF

1  GNTKPLSHFVSEPGPQKMPSMPTTSRQPSLLHYLQQPTPTQASSATASSTATATLQLQQQQQQQQQQPDHSSFLLQQMM
2  GNTKPLSHFVSEPGPQKMPSMPTTSRQPSLLHYLQQPTPTQASSATASSTATATLQLQQQQQQQQQQPDHSSFLLQQMM
3  GNTKPLSHFVSEPGPQKMSSMPTTSRQPSLLHYLQQPTPTQASSATASSTATATLQLQQQQQQQQQQPDHSSFLLQQMM
4  GNTKPLSHFVSEPGPQKMPSMPATSRQPSLLHYLQQPTPTQASSATASSTATATLQLQQQQQQQQ--PDHSSFLLQQMM
5  GNTKPLSHFVSEPGPQKMPSMPATSRQPSLLHYLQQPTPTQASSATASSTATATLQQQQQQQQQQ--PDHSSFLLQQMM
6  GNTKPLSHFVSEPGPQKMPSMPATSRQPSLLHYLQQPTPTQASSATASSTATATLQLQQQQQQQQQQPDHSSFLLQQMM
7  GNTKPLSHFVSEPGPQKMSSMPTTSRQPSLLHYLQQPTPTQASSATASSTATATLQLQQQQQQQQQQPDHSSFLLQQMM
8  GNTKPLSHFVSEPGPQKMPSIPATSRQPSLLHYLQQPTPTQASSATASSTATATLQQQQQQQQQQQ-PDHSSFLLQQMM
9  ANTKPLSHFASEPAPQKMPSMPAASRQASLLHYLQQPISAQASSATASSTATATLQLQPQPQQQQPQPEHS-FLLQQMM
10 GNTKPLSHFIAEPGPQKLPSMPATSRQPSLLHYLQQPTPAQASSATASSTATTSLQLPPQQ------PDHSAFLLQQMM
11 GNTKPLSHFISEPAPPKMPSMPATSRQASLLHYLQQPLPAQASSATASSTATATLQLQPQPQQQ---PEHS-FLLQQMM
```

Figure 3: Conserved amino acid sequences containing polyglutamine in orthologous *MAMLD1*.

```
1  Hsa QNKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQSSISAQQQQQQQ
2  Ppa QNKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQQQQ------SSISAQQQQQQQ
3  Mfa QNKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ-----SSISAQQQQQQQ
4  Mul QNKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ-----SSISAQQQQQQQ
5  Pab QNKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQQ--------SSISAQQQQQQQ
6  Ptr QNKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQ---------SSISAQQQQQQQ
7  Ggo QNKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQQ--------SSISAQQQQQQQ
8  Tch QNKPSLLHYTQQQQQQPQ--------------------------SSISAQQQQQQH
9  Eca QNKPSLLHYTQQQQQQQ---------------------------GSISAQQQQQQQ
10 Bmu QSKPSLLHYTQQQPQQ----------------------------------------
11 Rno QSKPSLLHYTQQQQPQLQPQSQQQQQQQQQQQQQQQQQQ-----GSLAAQQQQQAQ
12 Mmu QSKPSLLHYTQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ15QGSLAAQQQQQQP
13 Hgl QSKPSLLHYTQQQQHQQQQQQQQQQP------------------------------

1  ----SSISAQQQQQQQQQQQQQQQQQQQQQQQQQQQ--------PSSQPAQSLPSQ Hsa
2  ----SSISAQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ-----PSSQPAQSLPSQ Ppa
3  ----SSISAQQQQQQQQQQQQQQQQQQQQQQQQQQQ--------PSSQSAQSLPSQ Mfa
4  ----SSISAQQQQQQQQQQQQQQQQQQQQQQQQQQQ--------PSSQSAQSLPSQ Mul
5  QQ--SSISAQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ---PSSQPAQSLPSQ Pab
6  ----SSISAQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ-----PSSQPAQSLPSQ Ptr
7  ----SSISAQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ----PSSQPAQSLPSQ Ggo
8  QQQQSSISAQQQQQQQQQQQQQQQQQQQQQQRQQQQQQQQQQQ-PPSQSSQPLSSQ Tch
9  QQQQ-----QQQQQQQQQQQQQQQQQQQQQQQQPQQ--------PTSQPTQPLSSQ Eca
10 ----SSITVQPQQQQQQPQQQQQPQQQQQPQPQQQQQQQPQAQQPAAQPTQPLSNQ Bmu
11 ----SSLAAQQQQQQQQQQQQQQQQQQQQQQQQ-----------PS-QPTQALSSQ Rno
12 QPQ-SSLVAQQQQQQQQQQQQQQGSLTAQQQQQQQQQQQ-----PS-QPTHALSSQ Mmu
13 ------QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQPPSQPTQPLPTQ Hgl
```

Figure 4: Conserved amino acid sequences containing polyglutamine in orthologous *MAML2*. Note: the number “15” indicates the number of glutamine residues in a site position of Mmu protein.

Table 1: Characteristics of miR-1322 binding sites located through three nucleotides in the CDSs of some mRNAs. The number of binding sites in the mRNA fragment is indicated within parentheses.

Gene	Position of binding sites, nt	$\Delta G/\Delta G_m$, %	Oligopeptide
AFF3	1471–1486 (6)	85.4 ÷ 87.5	SSSSSSSGSSS
AR	1286–1334 (17)	87.5	QQQQQQQQQQQQQQQQQQQQQ
ARID3B	213–225 (5)	87.5	QQQQQQQQQQ
ASCL1	724–742 (8)	87.5	QQQQQQQQQQQA
ATN1	1692–1731 (13)	87.5 ÷ 91.7	QQQQQQQQQQQQQQQQQH
ATXN1	1559–1592 (12)	85.4 ÷ 91.7	QQQQQQQQQQQHQHQQ
	1604–1631 (10)	87.5 ÷ 89.6	QQQQQQQQQQQQQH
ATXN2	657–684 (10)	87.5	QQQQQQQQQQQQQQ
	699–714 (6)	87.5 ÷ 89.6	QQQQQQQQPP
ATXN7	637–658 (8)	85.4 ÷ 89.6	QQQQPPPPQPQ
BCL6B	763–775 (5)	85.4 ÷ 87.5	SSSSSSSSS
BHLHE22	1224–1236 (5)	85.4 ÷ 87.5	GSSSSSSSS
BMP2K	1543–1555 (5)	87.5	QQQQQQQQQ
	1600–1615 (6)	85.4 ÷ 91.7	QQQQQQQHHH
C9orf43	1283–1298 (6)	85.4 ÷ 87.5	QQQRQQQQQQ
CELF3	1871–1883 (5)	87.5	QQQQQQQQQ
E2F4	980–1007 (10)	85.4 ÷ 87.5	SSSSSSSSSSSSNS
EP400	8333–8363 (11)	87.5	QQQQQQQQQQQQQQQ
FAM155A	732–753 (8)	85.4 ÷ 87.5	QQQQRQQQQQQ
FAM157A	408–432 (9)	87.5	QQQQQQQQQQQQQ
FAM157B	414–435 (8)	85.4 ÷ 87.5	RQQQQQQQQQQ
HTT	196–247 (19)	85.4 ÷ 89.6	QQQQQQQQQQQQQQQQQQQQQQQQ
IRF2BPL	1249–1267 (7)	87.5	QQQQQQQQQQQ
IRS1	2088–2100 (5)	85.4 ÷ 87.5	SSSSSSSNAV
KCNN3	512–539 (10)	87.5 ÷ 91.7	QQQQQQQQQQQQQP
KIAA2018	4794–4815 (8)	87.5	QQQQQQQQQQQQ
MAGI1	1759–1771 (5)	87.5	QQQQQQQQQ
MAML2	3064–3091 (10)	87.5	QQQQQQQQQQQQQQ
MAML3	2219–2264 (16)	87.5	QQQQQQQQQQQQQQQQQQHSN
	2678–2690 (5)	87.5 ÷ 91.7	QQQQQPPPPQ
MED15	710–722 (5)	87.5 ÷ 91.7	QQQQQQQHL
	830–848 (7)	87.5	QQQQQQQQQQQ
MEF2A	1836–1860 (9)	85.4 ÷ 89.6	GFQQQQQQQQQQP
MLLT3	729–741 (5)	85.4 ÷ 87.5	SSSSSSSSS
	762–774 (5)	85.4 ÷ 87.5	SSSSSSSSS
MN1	2524–2539 (6)	87.5	QQQQQQQQQQ
MPRIP	622–643 (8)	87.5 ÷ 91.7	SSSSSSSSSSIP
NAP1L3	549–561 (5)	85.4 ÷ 87.5	GSGSSSSSG
NCOA3	4023–4038 (6)	87.5	QQQQQQQQQQ
NCOA6	1126–1138 (5)	87.5	QQQQQQQQQ
NCOR2	1812–1830 (7)	87.5 ÷ 91.7	QQQQQQQQQQQ
POLG	408–429 (8)	85.4 ÷ 87.5	QQQQQQQQQQQQ
POU6F2	701–719 (7)	85.4 ÷ 89.6	QQQQQQQQQPP
PRPF40A	785–797 (5)	85.4	AAAAAAAAAA
RAI1	1300–1324 (9)	87.5	QQQQQQQQQQQQQ
SALL1	590–602 (5)	85.4 ÷ 87.5	SSSSSSSSG
SCAF4	3303–3315 (5)	87.5 ÷ 91.7	QQQQQQPPP

Table 1: Continued.

Gene	Position of binding sites, nt	$\Delta G/\Delta G_m$, %	Oligopeptide
SMARCA2	765–795 (11)	87.5	QQQQQQQQQQQQQQQ
SRP14	394–406 (5)	87.5 ÷ 91.7	AAAAAAAAAP
TBP	468–480 (5)	87.5	QQQQQQQQQ
	501–546 (16)	87.5	QQQQQQQQQQQQQQQQQQQQ
THAP11	611–629 (7)	87.5 ÷ 91.7	QQQQQQQQQQQ
TNS1	2348–2363 (6)	87.5 ÷ 91.7	QQQQQQQQPR
VEZF1	1151–1175 (9)	87.5	TSNQKQQQQQQQQ
ZNF384	1770–1806 (13)	87.5 ÷ 91.7	QQQQQQQQQQQQQQQPP

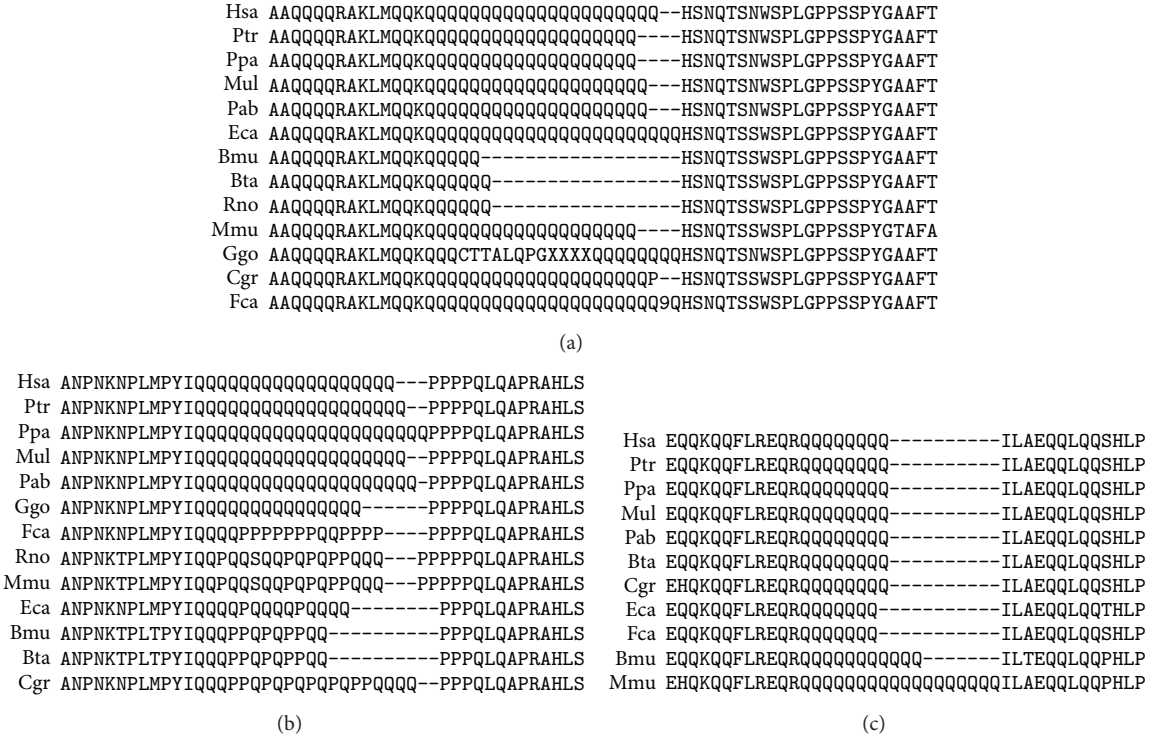


Figure 5: Conserved amino acid sequences containing polyglutamine in orthologous *MAML3*. Note: the number "9" indicates the number of glutamine residues in a site position of Fca protein (a).

developed in our laboratory. For example, downregulation of *ECRG2* and *TCA3* is associated with squamous cell carcinoma of the esophagus (ESCC) via miR-1322 [12]. ECRG2 can act as a tumor suppressor, regulating protease cascades during carcinogenesis and the migration and invasion of esophageal cancer cells [16].

3.2. Binding Sites in Paralogous and Orthologous mRNAs of the MAML Gene Family. The relationship between paralogous and orthologous mRNAs of the *MAML* gene family was considered an example of adaptation of gene expression to the action of miR-1322. *MAMLD1* encodes a mastermind-like domain-containing protein, which can act as a transcriptional coactivator [17]. Both *MAML2* and *MAML3* stabilize the DNA-binding complex RBP-J/CBF-1 and the Notch intracellular domains that are signaling intermediates [18]. Higher *MAML2* expression is observed in several B cell-derived lymphoma types, including classical Hodgkin's lymphoma cells, more than in normal B cells [19].

Various paralogous genes are targets for miR-1322. Two regions contain multiple miR-1322 binding sites in *MAMLD1* (Figure 3). The first region consists of eight sites and the second region consists of four sites. They were in domains (oligopeptides) consisting of 11 and 10 glutamine residues in the corresponding proteins, respectively.

The number of amino acids in orthologous proteins depends on the species (Figure 3). For example, for the first region, there are 28 glutamine residues in Ggo and nine residues in Hgl. Ten glutamine residues of Hsa, Ggo, and Ptr mRNAs to six of Eca mRNA were identified in the second region. In this case, the binding site of horse mRNA encoded proline in the associated protein.

miR-1322 binding sites in orthologous *MAML* mRNAs are highly conserved. Orthologous *MAML* proteins have

Table 2: The arrangement of miR-1322 binding sites in 5′ UTRs and 3′ UTRs human target genes.

Gene	Position of binding sites, nt	$\Delta G/\Delta G_m$, %
BACH2	25–43 (7)	85.4 ÷ 87.5
CACNA1A[*]	7170–7191 (8)	87.5
CAPN6	118–136 (7)	85.4 ÷ 87.5
CNKSR2	178–199 (8)	87.5 ÷ 89.6
GLS	53–86 (12)	85.4 ÷ 89.6
MAB21L1	342–378 (13)	87.5
PDYN[*]	1413–1425 (5)	85.4 ÷ 87.5
PIM1	103–118 (7)	85.4 ÷ 89.6
RBM39	323–335 (5)	87.5 ÷ 91.7

Note: the symbol "[*]" indicates binding site localization in the 3′ UTR.

conserved amino acid sequences containing polyglutamine (Figures 3–5). Orthologous miRNAs are not identified in most animals except *Pan troglodytes* (chimpanzee) and *Pongo pygmaeus* (orangutan); however, some other miRNAs are identical or very similar to the corresponding human miRNAs. Therefore, human miRNAs were used for the subsequent identification of conserved binding sites. Oligonucleotides containing CAG repeats represent the miR-1322 binding site of the mRNA that encoded a long polyglutamine sequence in the corresponding protein. Oligonucleotides encoding polyglutamine are located in the conserved protein domain.

The CDS of the human *MAML2* gene also has two regions with miR-1322 binding sites and encodes oligopeptides containing 47 and 27 glutamine residues (Figure 4). The number of glutamine residues in the oligopeptides is varied depending on the species. For example, there are six glutamine residues in the first oligopeptide region of the cow protein and 24 residues in the second region of the rat protein.

The CDS of the human *MAML3* gene has three regions that contain miR-1322 binding sites, and it encodes oligopeptides containing 21, 18, and eight glutamine residues. Some amino acids were lacking in the domains of *MAML3*, depending on the species (Figures 5(a)–5(c)).

The presence of multiple miR-1322 binding sites in *MAMLD1*, *MAML2*, and *MAML3* demonstrates their interactions. The expression of these genes has become increasingly important because the studied organisms were separated by tens of millions of years. The presence of multiple regions containing miR-1322 binding sites in *MAMLD1*, *MAML2*, and *MAML3* genes shows a strong dependence of their expression via miR-1322.

The glutamine-containing regions play an important role in the development of different diseases, according to previous literature. It is possible that changes in the dependence of the interactions between miR-1322 and *MAMLD1*, *MAML2*, and *MAML3* are interconnected.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] A. Issabekova, O. Berillo, M. Regnier, and A. Ivashchenko, "Interactions of intergenic microRNAs with mRNAs of genes involved in carcinogenesis," *Bioinformation*, vol. 8, no. 11, pp. 513–518, 2012.

[2] E. C. Lai, "Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation," *Nature Genetics*, vol. 30, no. 4, pp. 363–364, 2002.

[3] A. T. Ivashchenko, A. S. Issabekova, and O. A. Berillo, "MiR-1279, miR-548j, miR-548m, and miR-548d-5p binding sites in CDSs of paralogous and orthologous *PTPN12*, *MSH6*, and *ZEB1* genes," *BioMed Research International*, vol. 2013, Article ID 902467, 10 pages, 2013.

[4] L. da Sacco and A. Masotti, "Recent insights and novel bioinformatics tools to understand the role of MicroRNAs binding to 5' untranslated region," *International Journal of Molecular Sciences*, vol. 14, no. 1, pp. 480–495, 2013.

[5] O. Berillo, M. Régnier, and A. Ivashchenko, "Binding of intronic miRNAs to the mRNAs of host genes encoding intronic miRNAs and proteins that participate in tumourigenesis," *Computers in Biology and Medicine*, vol. 43, no. 10, pp. 1374–1381, 2013.

[6] A. Ivashchenko, O. Berillo, A. Pyrkova, R. Niyazova, and S. Atambayeva, "MiR-3960 binding sites with mRNA of human genes," *Bioinformation*, vol. 10, no. 7, pp. 423–427, 2014.

[7] H. Wang, Y. Zhu, M. Zhao et al., "miRNA-29c suppresses lung cancer cell adhesion to extracellular matrix and metastasis by targeting integrin β1 and matrix metalloproteinase2 (MMP2)," *PLoS ONE*, vol. 8, no. 8, Article ID e70192, 2013.

[8] A. Subramani, S. Alsidawi, S. Jagannathan et al., "The brain microenvironment negatively regulates miRNA-768-3p to promote K-ras expression and lung cancer metastasis," *Scientific Reports*, vol. 3, article 2392, 2013.

[9] S. Zhong, W. Li, Z. Chen, J. Xu, and J. Zhao, "miR-222 and miR-29a contribute to the drug-resistance of breast cancer cells," *Gene*, vol. 531, no. 1, pp. 8–14, 2013.

[10] Y. Shi, C.-Z. Wang, Y.-Y. Hou et al., "Screening of differentially expressed microRNAs in borderline and malignant gastrointestinal stromal tumors," *Zhonghua Bing Li Xue Za Zhi*, vol. 42, no. 1, pp. 20–25, 2013.

[11] M. Luo, D. Shen, X. Zhou, X. Chen, and W. Wang, "MicroRNA-497 is a potential prognostic marker in human cervical cancer and functions as a tumor suppressor by targeting the insulin-like growth factor 1 receptor," *Surgery*, vol. 153, no. 6, pp. 836–847, 2013.

[12] T. Zhang, D. Zhao, Q. Wang et al., "MicroRNA-1322 regulates ECRG2 allele specifically and acts as a potential biomarker in patients with esophageal squamous cell carcinoma," *Molecular Carcinogenesis*, vol. 52, no. 8, pp. 581–590, 2013.

[13] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Research*, vol. 34, pp. D140–D144, 2006.

[14] E. T. Kool, "Hydrogen bonding, base stacking, and steric effects in DNA replication," *Annual Review of Biophysics and Biomolecular Structure*, vol. 30, pp. 1–22, 2001.

[15] O. Berillo, M. Regnier, and A. Ivashchenko, "TmiRUSite and TmiROSite scripts: searching for mRNA fragments with miRNA binding sites with encoded amino acid residues," *Bioinformation*, vol. 10, no. 7, pp. 472–473, 2014.

[16] Y. Cui, M. Bi, T. Su, H. Liu, and S.-H. Lu, "Molecular cloning and characterization of a novel esophageal cancer related gene," *International Journal of Oncology*, vol. 37, no. 6, pp. 1521–1528, 2010.

[17] M. Fukami, Y. Wada, M. Okada et al., "Mastermind-like domain-containing 1 (*MAMLD1* or *CXorf6*) transactivates the Hes3 promoter, augments testosterone production, and contains the SF1 target sequence," *The Journal of Biological Chemistry*, vol. 283, no. 9, pp. 5525–5532, 2008.

[18] S.-E. Lin, T. Oyama, T. Nagase, K. Harigaya, and M. Kitagawa, "Identification of new human mastermind proteins defines a family that consists of positive regulators for notch signaling," *The Journal of Biological Chemistry*, vol. 277, no. 52, pp. 50612–50620, 2002.

[19] K. Köchert, K. Ullrich, S. Kreher et al., "High-level expression of Mastermind-like 2 contributes to aberrant activation of the NOTCH signaling pathway in human lymphomas," *Oncogene*, vol. 30, no. 15, pp. 1831–1840, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 208346, 12 pages
http://dx.doi.org/10.1155/2015/208346

Research Article

Phyloproteomic Analysis of 11780 Six-Residue-Long Motifs Occurrences

O. V. Galzitskaya and M. Yu. Lobanov

Institute of Protein Research, Russian Academy of Sciences, 4 Institutskaya Street, Pushchino, Moscow Region 142290, Russia

Correspondence should be addressed to O. V. Galzitskaya; ogalzit@vega.protres.ru

Received 11 July 2014; Accepted 3 November 2014

Academic Editor: Peter F. Stadler

How is it possible to find good traits for phylogenetic reconstructions? Here, we present a new phyloproteomic criterion that is an occurrence of simple motifs which can be imprints of evolution history. We studied the occurrences of 11780 six-residue-long motifs consisting of two randomly located amino acids in 97 eukaryotic and 25 bacterial proteomes. For all eukaryotic proteomes, with the exception of the Amoebozoa, Stramenopiles, and Diplomonadida kingdoms, the number of proteins containing the motifs from the first group (one of the two amino acids occurs once at the terminal position) made about 20%; in the case of motifs from the second (one of two amino acids occurs one time within the pattern) and third (the two amino acids occur randomly) groups, 30% and 50%, respectively. For bacterial proteomes, this relationship was 10%, 27%, and 63%, respectively. The matrices of correlation coefficients between numbers of proteins where a motif from the set of 11780 motifs appears at least once in 9 kingdoms and 5 phyla of bacteria were calculated. Among the correlation coefficients for eukaryotic proteomes, the correlation between the animal and fungi kingdoms (0.62) is higher than between fungi and plants (0.54). Our study provides support that animals and fungi are sibling kingdoms. Comparison of the frequencies of six-residue-long motifs in different proteomes allows obtaining phylogenetic relationships based on similarities between these frequencies: the Diplomonadida kingdoms are more close to Bacteria than to Eukaryota; Stramenopiles and Amoebozoa are more close to each other than to other kingdoms of Eukaryota.

1. Introduction

By the middle of the XXth century, it had become clear that all living organisms of cellular texture are divided into two groups or kingdoms, prokaryotes and eukaryotes, according to structural peculiarities of their cells. It was long believed that the terms "prokaryotes" and "bacteria" are synonyms for the same independent evolutionary branch of living organisms. However, about 30 years ago, molecular comparisons of base sequences of ribosomal RNAs provided grounds to divide prokaryotes into at least two independent branches, Eubacteria and Archaebacteria, which differ in their origin [1]. Later, these data were generalized and the term DOMAIN was suggested, which is the branch that has the highest rank in the hierarchic taxonomy [2]. These DOMAINS are Bacteria, Archaea, and Eukaryota.

Protein phylogeny was developed simultaneously with RNA phylogeny [3, 4]. Protein phylogeny is similar to the developed RNA phylogeny because it is based on the division of living organisms into three DOMAINS. RNA and protein phylogenies are based on the alignments of sequences from different organisms, and most phylogenetic methods are based on comparison of protein or nucleic acid sequences in their aligned parts. The conventional tree-building methods for phylogenetic reconstructions are neighbor joining (NJ) [5], maximum parsimony (MP) [6], and maximum likelihood (ML) [7]. Moreover, there is an additional approach as alignment-free phylogeny methods based on k-mer appearance in genomic DNA [8–12].

The understanding of how different major groups of organisms are related to each other and the tracing of their evolution from the common ancestor remains controversial and unsolved. In recent years, the wealth of new information based on a large number of gene and protein sequences has become available. At present, a phylogenetic analysis can be carried out based on either nucleic acid or protein sequences.

Nonetheless, the phylogenetic relationship among the kingdoms Animalia, Plantae, and Fungi remains uncertain despite extensive attempts to clarify it. The first hypothesis states that Animalia is more closely related to Plantae [13–15]. The second one supports Plantae and Fungi grouping [16]; the third one, Animalia and Fungi [17–23]. To elucidate evolutionary relationships among different proteomes we will consider the occurrence of some simple motifs which can be imprints of evolution history.

What candidates can be stated as simple motifs? We have done several investigations in this direction. First, by combining the motif discovery and disorder protein segment identification in the Protein Data Bank (PDB: http://www.rcsb.org/), we have compiled the largest database of disordered patterns (171) from the clustered PDB where identity between chains inside a cluster is larger than or equal to 75% using simple rules of selection [21–24]. Second, among these patterns, the patterns with low complexity are more abundant and the length of these motifs is six residues. Third, the patterns with frequent occurrence in proteomes have low complexity (PPPPP, GGGGG, EEEED, HHHH, KKKKK, SSTSS, and QQQQQP), and the type of patterns varies across different proteomes [21]. It is supposed that if an amino acid motif possesses no definite spatial structure in most protein structures, it is likely to be disordered in a protein with an unknown spatial structure [21]. Therefore, the patterns with the length of six residues and low complexity, which are, for example, homorepeats of 20 amino acids, are the major candidates for this role. The length of six residues is important: (1) the experiments performed demonstrated that a minimum repeat size of 6 histidine residues was required for efficient protein translocation to nuclear speckles [25]; (2) six-residue patches affect the folding/aggregation features of proteins, and they are important "words" for the understanding of protein dynamics [26]; (3) nucleation sites are constrained by patches of approximately six residues [27, 28].

It has been found that homorepeats of some amino acids (runs of a single amino acid) occur more frequently than others and the type of homorepeats varies across different proteomes [21]. For example, EEEEEE appears to be the most frequent for all considered proteomes for Chordata, QQQQQQ for Arthropoda, and SSSSSS for Nematoda. A comparative analysis of the number of proteins containing 6-residue-long homorepeats and the 109 disordered selected patterns in 123 proteomes has demonstrated that the correlation coefficients between numbers of proteins are higher inside the considered kingdom than between them [21]. In these proteins a six-residue-long homorepeat occurs at least once for each of the 20 types of amino acid residues and 109 disordered patterns from the library appearing in 9 kingdoms of Eukaryota and 5 phyla of Bacteria.

Here, we present a new phyloproteomic criterion which is based on the peculiarities of amino acid sequences which is an occurrence of some simple motifs which can be imprints of evolution history. In this work, we focus our attention on studying the frequency of six simple amino acid motifs consisting of two randomly located amino acids (11780 motifs) in 122 eukaryotic and bacterial proteomes.

2. Materials and Methods

2.1. Construction of the Library of Six-Residue-Long Motifs . We constructed the library of all possible motifs composed of two amino acids, with the assumption that each amino acid could be at any position and at any ratio and that such a motif was six amino acids long [29]. There were $11780 = (2^6-2)\cdot C_{20}^2$ such motifs in total (excluding two homorepeats for every amino acid pair). The obtained motifs could be divided into three groups. The first group contains the motifs where one of the two amino acids occurs only once and occupies the first or sixth (i.e., outside) position. The second group includes motifs where the second amino acid also occurs once but is inside the motif. The third group contains all the other motifs where each of the two amino acids occurs at least twice and in any order.

2.2. Database of Proteomes. We considered 3279 proteomes from the EBI site (ftp://ftp.ebi.ac.uk/pub/databases/SPproteomes/last_release/uniprot/proteomes/). A preliminary analysis showed that the number of proteins with at least one occurrence of homorepeats, 6 residues long, is less than 500 for proteomes with an overall number of residues below 2,500,000. Even so, only 22 proteomes out of 3156 have more than 100 proteins with at least one occurrence of 6-residue homorepeats. These data provided grounds for our research involving only proteomes with an overall number of residues exceeding 2,500,000.

We obtained 122 proteomes taking into account the length of proteomes representing 9 kingdoms of eukaryotes and 5 phyla of Bacteria (see Table 1 in [21]). Unfortunately, only three kingdoms of eukaryotes (Metazoa, Viridiplantae, and Fungi) are given at http://www.ncbi.nlm.nih.gov/Taxonomy. In other cases, the rank of kingdom is missing. In such situations, we chose the highest taxonomic category following from the subkingdom of eukaryotes instead of the kingdom. We chose 97 out of 120 eukaryotic proteomes and a small number of bacterial proteomes. The smallest eukaryotic proteome belongs to *Hemiselmis andersenii*, class Cryptophyta. It is evident that 498 proteins with an overall number of 167,452 amino acid residues are not sufficient for reliable statistics. Historically, the superkingdom of Bacteria is divided into phyla but not kingdoms. We preferred to consider such phyla separately.

Among 97 eukaryotic proteomes, 17 belong to the kingdom of Metazoa or animals: *Homo sapiens* (51778 protein sequences), *Bos taurus* (18405), *Mus musculus* (42120), *Rattus norvegicus* (28166), *Gallus gallus* (12954), *Danio rerio* (21576), and *Tetraodon nigroviridis* (27836) belong to Chordata phylum; *Drosophila melanogaster* (15101), *Drosophila pseudoobscura* (16000), *Aedes aegypti* (16042), *Anopheles darlingi* (11437), and *Anopheles gambiae* (12455) to arthropods; *Caenorhabditis briggsae* (18531), *Caenorhabditis elegans* (23817), Loa loa (16271), and *Trichinella spiralis* (16040) to nematodes; *Nematostella vectensis* (24435) belongs to Cnidaria phylum.

2.3. Calculation of Correlation Coefficient. The vectors of 11780 values for each type of motif are compared between

Table 1: 11780 motifs that frequently occur in 123 proteomes.

11780		The first group		The second group		The third group	
EEEEED	6744	EEEEED	6744	EDEEEE	4248	APAPAP	3543
QQQQQP	6300	QQQQQP	6300	STSSSS	4166	DDEEEE	3464
DEEEEE	6165	DEEEEE	6165	NNNNSN	4030	SGSGSG	3423
TSSSSS	6135	TSSSSS	6135	EEEEDE	3995	PAPAPA	3392
SGGGGG	6117	SGGGGG	6117	NSNNNN	3992	EEEEDD	3292
AAAAAG	5863	AAAAAG	5863	EEDEEE	3959	GSGSGS	3240
PSSSSS	5813	PSSSSS	5813	SSSSTS	3953	DEDEDE	3127
NNNNNS	5811	NNNNNS	5811	GGGGSG	3934	EDEDED	3045
QQQQQH	5798	QQQQQH	5798	AAVAAA	3768	RSRSRS	2983
SSSSST	5780	SSSSST	5780	AAAVAA	3758	DDDDEE	2953
DDDDDE	5611	DDDDDE	5611	GSGGGG	3690	EEEDDD	2845
SNNNNN	5585	SNNNNN	5585	SSTSSS	3660	DDDEEE	2822
ASSSSS	5581	ASSSSS	5581	SSSTSS	3652	RGRGRG	2817
SAAAAA	5405	SAAAAA	5405	EEEDEE	3627	EEDDDD	2754
APPPPP	5325	APPPPP	5325	AAAAVA	3616	AAAAGG	2743
AAAAAS	5322	AAAAAS	5322	GGGSGG	3556	EDEDEE	2651
AAAAAV	5277	AAAAAV	5277	SPSSSS	3459	DDEDED	2570
GGGGGS	5118	GGGGGS	5118	NNSNNN	3429	RGGRGG	2537
GGGGGA	4862	GGGGGA	4862	NNNSNN	3418	DEDEDD	2489
PQQQQQ	4819	PQQQQQ	4819	AVAAAA	3391	SSSSTT	2448

different proteomes. The correlation coefficient (r) was calculated using the equation

$$r = \frac{(1/n)\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}, \quad (1)$$

where S_x and S_y are the standard deviations for variables x and y.

For 20 homorepeats, the standard error in determining the correlation coefficient is less than $1/\sqrt{20-2} \cong 0.24$. The standard error of correlation coefficient is $\mathrm{se}_r = \sqrt{(1-r^2)/(n-2)}$ where n is the number of points; for 109 disordered patterns it is less than $1/\sqrt{109-2} \cong 0.1$, and for 11780 patterns it is less than 0.01. Therefore, in Tables 3–7 the correlation coefficients range as follows: less than 0.5, from 0.5 to 0.75, and larger than 0.75.

3. Results and Discussion

3.1. Occurrences of Motifs in 122 Proteomes. We constructed the library of all possible motifs consisting of the two amino acids, with the assumption that each amino acid could be at any position and at any ratio and that such a motif was six amino acid residues long. There were 11780 such motifs in total. The obtained motifs were divided into three groups (see Section 2). The numbers of motifs in the first, second, and third groups were 760 (6%), 1520 (13%), and 9500 (81%), respectively. We estimated the occurrences of these motifs in 122 proteomes.

The most often occurrences of simple motifs for 122 proteomes from the three groups are presented in Table 1. Among the motifs from the first group, the leaders from the human proteome were EEEEED (422 times), DEEEEE (370), LPPPPP (327), APPPPP (264), PLLLLL (251), and PPPPPL (216). It should be noted that such motifs as LPPPPP, PLLLLL, and PPPPPL are not leaders among the occurrences of 122 proteomes (see Table 1). Among the motifs in which one amino acid occurred once and only inside the motif, the leaders from the human proteome were EEEEDE (288), EDEEEE (279), EEDEEE (248), EEEDEE (250), PLPPPP (239), and PPPPLP (207). Among the leaders in which the two amino acids occurred were SGSGSG (135), EEEEDD (157), GPPGPP (162), and RSRSRS (153). The following rare motifs that appeared only in two proteins should be noted for the human proteome: FFFFFN, FFFFFP, CHHHHH, MVVVVV, IHHHHH, WKKKKK, NNNNNS, and IIIIIF from the first group; IIMIII, RRFRRR, YLYYYY, NNCNNN, HHTHHH, and DDQDDD from the second group; and CCCRRR, MMMGGG, TTTDDD, FFSFFS, FFPFFP, VVRVVR, QQKQQK, and DDHDDH from the third group. At the same time, the NNNNNS motif is among the leader motifs for 122 proteomes and it occurs 146 times in the *Drosophila melanogaster* proteome and 473 times in the *Plasmodium falciparum* proteome (Alveolata kingdom). An analogous situation is observed for SNNNNN. It does not occur in the human proteome and appears in 489 proteins for the *Plasmodium falciparum* proteome. PQQQQQ occurs 52 times in the human proteome and 413 times in the *Dictyostelium discoideum* proteome.

In frequently occurring motifs from the *Drosophila melanogaster* proteome, the leading amino acids were glutamine, alanine, and glycine. Among the motifs from the first group, the leaders were QQQQQH (470), HQQQQQ (410), LQQQQQ (359), QQQQQL (359), QQQQQP (276),

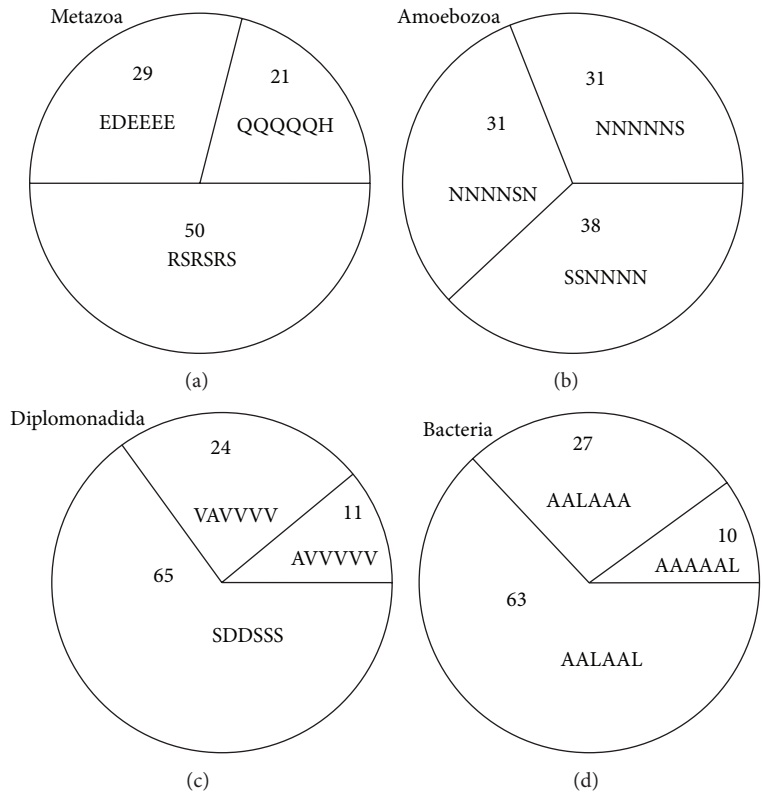


Figure 1: Statistics of occurrence of motifs, six residues long. Statistics of occurrence of motifs, six residues long, consisting of two amino acids in the three groups for 3 kingdoms of Eukaryota and for 5 phyla of Bacteria in percentage terms: (a) the Metazoa kingdom (17 proteomes), (b) the Amoebozoa kingdom (2 proteomes), (c) the Diplomonadida kingdom (3 proteomes), and (d) 26 bacterial proteomes. For each kingdom, the motif with the frequent occurrence in the group is presented.

PQQQQQ (260), QQQQQA (221), AQQQQQ (219), and SAAAAA (224). Among the motifs of the second group, the leaders were QQQQHQ (319), QQQHQQ (297), QQHQQQ (290), QHQQQQ (284), QQQQLQ (243), QLQQQQ (229), and QQLQQQ (218). Among the motifs of the third group, the leaders were GSGSGS (174), SGSGSG (157), HHQQQQ (163), QQQQHH (166), SSGGGG (110), and GGSGSG (105).

Out of 11780 motifs, 865 were not found in 122 proteomes.

We estimated the occurrence of the motifs from the three groups in 9 kingdoms of Eukaryota and 5 phyla of Bacteria (see Table 2). Interestingly, for all eukaryotic proteomes with the exception of the Amoebozoa and Diplomonadida kingdoms, the number of proteins containing at least one motif from the first group was about 20%; in the case of motifs from the second and third groups, 30% and 50%, respectively (see Table 2). For bacterial proteomes this relationship is 10%, 27%, and 63%, respectively. One can see that proteomes from the Diplomonadida kingdom are more close to bacterial proteomes than to eukaryotic ones (see Figure 1). It should be noted that diplomonads are a group of flagellates, most of which are parasitic. At the same time, the proteomes from the Amoebozoa kingdom have different statistics: 31%, 31%, and 38%, respectively. For the Metazoa, Amoebozoa, Diplomonadida, and Bacteria kingdoms, the motifs with the frequent occurrence in the groups are presented in Figure 1.

Among animal proteomes, one can see some deviation from the average values for *Nematostella vectensis* (class Anthozoa, phylum Cnidaria): 14%, 27%, and 59%, correspondingly. This is more close to the statistics for the bacterial proteomes. Another deviation from the average values is observed for phylum Arthropoda, especially for class Insecta (29%, 30%, and 41% for *Anopheles darlingi* and 26%, 29%, and 45% for *Anopheles gambiae*).

Table 2: Occurrence of 11780 motifs from the three groups in 9 kingdoms of Eukaryota and for 5 phyla of Bacteria in percentage terms.

Kingdom	<x>	Error	<x>	Error	<x>	Error
	First group		Second group		Third group	
Metazoa (17)	21	3	29	1	50	4
Viridiplantae (5)	21	4	28	2	51	5
Stramenopiles (1)	28	—	32	—	41	—
Choanoflagellida (1)	18	—	27	—	55	—
Euglenozoa (4)	22	3	29	2	49	4
Alveolata (6)	23	4	29	1	48	5
Amoebozoa (2)	31	1	31	0	38	2
Diplomonadida (3)	11	1	24	1	65	2
Fungi (58)	18	3	28	1	53	4
Bacteria (25)	10	1	27	2	63	3
All 11780 motifs	6	0	13	0	81	0

Table 3: Averaged correlation coefficients (in percentage terms) between numbers of proteins where a simple motif, six residues long, from the whole set of **11780** motifs appears at least once in 9 kingdoms of Eukaryota and 5 phyla of Bacteria.

Metazoa (17)	Viridiplantae (5)	Stramenopiles (1)	Choanoflagellida (1)	Euglenozoa (4)	Alveolata (6)	Amoebozoa (2)	Diplomonadida (3)	Fungi (58)	Acidobacteria (1)	Actinobacteria (14)	Proteobacteria (8)	Bacteroidetes (2)	Chloroflexi (1)	
65*	49	46	29	47	22	30	35	62*	27	25	27	26	24	Metazoa (17)
49	61*	60*	28	47	14	13	27	54*	41	43	42	21	22	Viridiplantae (5)
46	60*	—	20	52*	8	8	16	41	49	48	44	23	17	Stramenopiles (1)
29	28	20	—	30	7	13	23	32	20	21	22	12	15	Choanoflagellida (1)
47	47	52*	30	47	10	17	27	45	35	39	35	24	22	Euglenozoa (4)
22	14	8	7	10	69*	37	8	25	−1	−1	0	13	3	Alveolata (6)
30	13	8	13	17	37	90**	13	35	−1	−1	−1	5	3	Amoebozoa (2)
35	27	16	23	27	8	13	68*	38	21	22	22	25	27	Diplomonadida (3)
62*	54*	41	32	45	25	35	38	71*	28	26	28	21	21	Fungi (58)
27	41	49	20	35	−1	−1	21	28	—	70*	67*	32	39	Acidobacteria (1)
25	43	48	21	39	−1	−1	22	26	70*	87**	74*	29	38	Actinobacteria (14)
27	42	44	22	35	0	−1	22	28	67*	74*	72*	29	39	Proteobacteria (8)
26	21	23	12	24	13	5	25	21	32	29	29	39	39	Bacteroidetes (2)
24	22	17	15	22	3	3	27	21	39	38	39	39	—	Chloroflexi (1)

Table 4: Averaged correlation coefficients (in percentage terms) between numbers of proteins where a simple motif, six residues long, from the first group (760 motifs) appears at least once in 9 kingdoms of Eukaryota and 5 phyla of Bacteria.

Metazoa (17)	Viridiplantae (5)	Stramenopiles (1)	Choanoflagellida (1)	Euglenozoa (4)	Alveolata (6)	Amoebozoa (2)	Diplomonadida (3)	Fungi (58)	Acidobacteria (1)	Actinobacteria (14)	Proteobacteria (8)	Bacteroidetes (2)	Chloroflexi (1)	
64*	49	45	61*	50*	13	26	37	65*	34	29	35	29	32	Metazoa (17)
49	62*	61*	60*	48	6	7	28	54*	48	47	52*	22	22	Viridiplantae (5)
45	61*	—	45	53*	0	1	14	40	68*	64*	63*	28	20	Stramenopiles (1)
61*	60*	45	—	60*	7	28	35	67*	42	46	47	26	29	Choanoflagellida (1)
50*	48	53*	60*	51*	3	16	30	49	47	49	48	31	32	Euglenozoa (4)
13	6	0	7	3	74*	29	4	16	−6	−6	−5	8	−1	Alveolata (6)
26	7	1	28	16	29	92**	12	33	−5	−5	−5	1	−1	Amoebozoa (2)
37	28	14	35	30	4	12	67*	40	17	16	19	29	31	Diplomonadida (3)
65*	54*	40	67*	49	16	33	40	74*	29	25	32	20	22	Fungi (58)
34	48	68*	42	47	−6	−5	17	29	—	75*	73*	40	38	Acidobacteria (1)
29	47	64*	46	49	−6	−5	16	25	75*	90**	76**	30	28	Actinobacteria (14)
35	52*	63*	47	48	−5	−5	19	32	73*	76**	74*	33	33	Proteobacteria (8)
29	22	28	26	31	8	1	29	20	40	30	33	52*	57*	Bacteroidetes (2)
32	22	20	29	32	−1	−1	31	22	38	28	33	57*	—	Chloroflexi (1)

Table 5: Averaged correlation coefficients (in percentage terms) between numbers of proteins where at least once a simple motif, six residues long, from the second group (1520 motifs) appears in 9 kingdoms of Eukaryota and 5 phyla of Bacteria.

Metazoa (17)	Viridiplantae (5)	Stramenopiles (1)	Choanoflagellida (1)	Euglenozoa (4)	Alveolata (6)	Amoebozoa (2)	Diplomonadida (3)	Fungi (58)	Acidobacteria (1)	Actinobacteria (14)	Proteobacteria (8)	Bacteroidetes (2)	Chloroflexi (1)	
64*	45	42	54*	47	20	23	39	63*	26	24	27	26	27	Metazoa (17)
45	64*	62*	58*	50*	10	5	28	52*	51*	52*	52*	22	24	Viridiplantae (5)
42	62*	—	46	55*	6	2	16	40	62*	57*	56*	22	23	Stramenopiles (1)
54*	58*	46	—	55*	11	23	32	61*	44	51*	52*	28	29	Choanoflagellida (1)
47	50*	55*	55*	50	10	13	32	48	41	46	45	27	26	Euglenozoa (4)
20	10	6	11	10	66*	40	6	23	−3	−4	−3	13	−1	Alveolata (6)
23	5	2	23	13	40	90**	10	30	−4	−4	−4	0	−3	Amoebozoa (2)
39	28	16	32	32	6	10	72*	42	17	19	20	25	31	Diplomonadida (3)
63*	52*	40	61*	48	23	30	42	71*	28	26	29	19	19	Fungi (58)
26	51*	62*	44	41	−3	−4	17	28	—	70*	69*	33	40	Acidobacteria (1)
24	52*	57*	51*	46	−4	−4	19	26	70*	92**	80**	30	35	Actinobacteria (14)
27	52*	56*	52*	45	−3	−4	20	29	69*	80**	77**	32	38	Proteobacteria (8)
26	22	22	28	27	13	0	25	19	33	30	32	47	56*	Bacteroidetes (2)
27	24	23	29	26	−1	−3	31	19	40	35	38	56*	—	Chloroflexi (1)

Table 6: Averaged correlation coefficients (in percentage terms) between numbers of proteins where a simple motif, six residues long, from the third group (9500 motifs) appears at least once in 9 kingdoms of Eukaryota and 5 phyla of Bacteria.

Metazoa (17)	Viridiplantae (5)	Stramenopiles (1)	Choanoflagellida (1)	Euglenozoa (4)	Alveolata (6)	Amoebozoa (2)	Diplomonadida (3)	Fungi (58)	Acidobacteria (1)	Actinobacteria (14)	Proteobacteria (8)	Bacteroidetes (2)	Chloroflexi (1)	
59*	44	44	23	37	25	31	31	56*	25	26	25	21	22	Metazoa (17)
44	58*	54*	22	41	17	15	26	53*	36	40	38	17	22	Viridiplantae (5)
44	54*	—	15	47	11	11	16	41	43	45	39	18	16	Stramenopiles (1)
23	22	15	—	24	6	10	21	25	15	16	16	8	12	Choanoflagellida (1)
37	41	47	24	40	10	14	23	38	30	34	29	17	19	Euglenozoa (4)
25	17	11	6	10	61*	38	8	28	0	−1	0	13	5	Alveolata (6)
31	15	11	10	14	38	88**	13	36	0	1	0	8	6	Amoebozoa (2)
31	26	16	21	23	8	13	67*	33	22	24	24	21	24	Diplomonadida (3)
56*	53*	41	25	38	28	36	33	68*	26	27	27	18	21	Fungi (58)
25	36	43	15	30	0	0	22	26	—	69*	65*	28	38	Acidobacteria (1)
26	40	45	16	34	−1	1	24	27	69*	84**	73*	28	40	Actinobacteria (14)
25	38	39	16	29	0	0	24	27	65*	73*	71*	26	40	Proteobacteria (8)
21	17	18	8	17	13	8	21	18	28	28	26	30	30	Bacteroidetes (2)
22	22	16	12	19	5	6	24	21	38	40	40	30	—	Chloroflexi (1)

Table 7: Averaged correlation coefficients (in percentage terms) between numbers of proteins where a simple motif, six residues long, appears at least once in 17 animal proteomes (kingdom Metazoa).

Phylum	Proteome	*H. sapiens*	*B. taurus*	*M. musculus*	*R. norvegicus*	*G. gallus*	*D. rerio*	*T. nigroviridis*	*D. melanogaster*	*D. pseudoobscura*	*A. aegypti*	*A. darlingi*	*A. gambiae*	*C. briggsae*	*C. elegans*	*L. loa*	*T. spiralis*	*N. vectensis*
Chordata	*H. sapiens*		**95****	**96****	**95****	**89****	**80****	**73***	**53***	**53***	**60***	**48**	**63***	**70***	**65***	**42**	**54***	**68***
	B. taurus	**95****		95**	95**	89**	80**	70*	**49**	49	57*	44	60*	68*	**62***	38	50*	68*
	M. musculus	**96****	95**		97**	90**	82**	75**	**56***	56*	64*	52*	66*	71*	**67***	44	57*	70*
	R. norvegicus	**95****	95**	97**		90**	83**	72*	**51***	50*	61*	46	61*	71*	**66***	43	53*	70*
	G. gallus	**89****	89**	90**	90**		86**	71*	**49**	48	61*	46	59*	74*	**68***	42	55*	73*
	D. rerio	**80****	80**	82**	83**	86**		72*	**50***	48	68*	49	60*	80**	**73***	48	60*	77**
	T. nigroviridis	**73***	70*	75**	72*	71*	72*		**46***	45	54*	43	56*	61*	**58***	50*	52*	59*
Arthropoda	*D. melanogaster*	**53***	**49**	**56***	**51***	**49**	**50***	**46**		**96****	**87****	**92****	**87****	**57***	**67***	**48**	**71***	**44**
	D. pseudoobscura	**53***	49	56*	50*	48	48	45	**96****		84**	91**	87**	53*	**63***	47	71*	42
	A. aegypti	**60***	57*	64*	61*	61*	68*	54*	**87****	84**		86**	87**	73*	**79****	54*	76**	57*
	A. darlingi	**48**	44	52*	46	46	49	43	**92****	91**	86**		91**	53*	**62***	51*	75**	40
	A. gambiae	**63***	60*	66*	61*	59*	60*	56*	**87****	87**	87**	91**		62*	**69***	48	72*	50*
Nematoda	*C. briggsae*	**70***	68*	71*	71*	74*	80**	61*	**57***	53*	73*	53*	62*		**90****	52*	64*	71*
	C. elegans	**65***	**62***	**67***	**66***	**68***	**73***	**58***	**67***	**63***	**79****	**62****	**69****	**90****		**52***	**67***	**66***
	L. loa	**42**	38	44	43	42	48	50*	**48**	47	54*	51*	48	52*	**52***		68*	46
	T. spiralis	**54***	50*	57*	53*	55*	60*	52*	**71***	71*	76**	75**	72*	64*	**67***	68*		53*
Cnidaria	*N. vectensis*	**68***	68*	70*	70*	73*	77**	59*	**44**	42	57*	40	50*	71*	**66***	46	53*	

It should be also noted that the proteins bearing motifs from the third group occurred more frequently than the proteins with motifs from the two other groups only because the third group contained a significantly larger number of motifs (12.5 times as many as in the first group). It might be noted that motifs from the first groups are the simplest, being homorepeats with an adjacent amino acid. Motifs from the second group are homorepeats with an inclusion of the other amino acid. Meanwhile, members of the third group can hardly be derived from homorepeats. The most frequent motifs are the ones most closely resembling homorepeats, that is, the motifs from the first group, whereas the motifs from the second group occur somewhat more rarely, and the motifs not resembling homorepeats are the rarest of all. Each proteome contains its characteristic leading motifs, and it is apparent that the amino acids foremost among six amino acid repeats occur most often.

3.2. Construction of Matrices of Correlation Coefficients for Proteins Containing Simple Motifs in the Studied Proteomes. For each proteome, we calculated a set of 11780 values reflecting the number of proteins containing at least one simple motif, 6 residues long. Then considering all possible pairs of proteomes, the correlation coefficients between the 11780 values have been calculated which allowed us to construct a matrix of correlation coefficients (see Table 3). As a rule, the correlation coefficients are higher inside the studied kingdom than between them. A similar conclusion follows from considering the occurrence of motifs from the three groups (see Tables 4, 5, and 6). "$**$" in Tables 3–7 is used to show the correlation higher than 75%, and "$*$" is used to show the correlation from 50% to 75%. Usually, the correlation coefficients are higher inside the considered kingdom than between them. The highest correlation is observed for the Amoebozoa kingdom in all cases (see Tables 3–6).

Most of the theories suggest that colonial naked choanoflagellate-like protists gave rise to first animals, while chitinous thecate choanoflagellate-like protists gave rise to first fungi [30, 31]. In the case of occurrence of the motifs from the first and second groups, we obtained a high correlation between the Choanoflagellida and Fungi kingdoms (0.67 and 0.61) compared to between the Choanoflagellida and animals kingdoms (0.61 and 0.54) (see Tables 4–6).

We averaged the correlation coefficients over all proteomes from the studied kingdoms. The averaged correlation coefficient is low inside such a kingdom as Metazoa (see Table 3). We decided to analyze in more detail the proteomes from the Metazoa kingdom. If the correlation coefficients for animal proteomes only (see Table 7) are to be considered, four clusters can be selected with high correlation between the numbers of proteins where a simple motif, 6 residues long, appears at least once. The first cluster corresponds to the phylum Chordata (7 proteomes), the second to Arthropoda (5 proteomes), the third to Nematoda (4 proteomes), and the fourth to Cnidaria (only 1 proteome). Again one can see that the correlation coefficients are higher inside the considered phylum than between them.

In Table 7 one can see that the correlation coefficient between zebrafish, *Danio rerio*, and pufferfish, *Tetraodon nigroviridis*, is 0.72, while on the other hand that between *D. rerio* and starlet sea anemone, *Nematostella vectensis*, is 0.77 and those between *D. rerio* and two nematodes, *Caenorhabditis elegans* and *C. briggsae*, are 0.73 and 0.80, respectively. The correlation coefficients between *T. nigroviridis* and other vertebrates are 0.70–0.75, while those between *D. rerio* and other vertebrates, except for *T. nigroviridis*, are 0.80–0.86. These values suggest that the pattern of six-residue-long motifs in *T. nigroviridis* has changed very rapidly after the separation of the lineages of pufferfish (belongs to a family of primarily marine and estuarine fish) and zebrafish (a tropical freshwater fish). This fact is not surprising in light of the last data, that horses were evolutionarily closest to Brandt's bats (*Myotis brandtii*); their divergence occurred about 81.7 million years ago, which is close to the time of the adaptive radiation of the class Mammalia [32].

In the case of the occurrence of simple motifs (all 11780 and 9500 for the third group), there is no high correlation (larger than 0.5) between eukaryotic and bacterial proteomes. Among the correlation coefficients for eukaryotic proteomes, there is a high correlation between the animal and Fungi kingdoms (0.62) compared to between the fungi and plants (0.54). This is valid also in the case of consideration of the correlation coefficients for the occurrence of the motifs from the three groups separately (see Tables 4–6). Moreover, this result agrees with the results obtained by us after analysis of loops in elongation factors EF1A using the novel informative characteristic called the "loops" method [20]. The method is based on the ability of amino acid sequences to form flexible loops in protein structure. Each kingdom displayed variations in the number of loops and their location within the three EF1A domains. It has been found that animals and fungi are sibling kingdoms [20].

4. Conclusions

One can see that some simple motifs have been maintained throughout evolution and that in the studied 122 eukaryotic and bacterial proteomes the most frequent motifs are specific for each proteome. The ratio between occurrences of the simple motifs from the three groups is practically the same for the eukaryotic proteomes. The other relationship between occurrences of the motifs is observed for the bacterial proteomes. The question about specificity of these motifs is more important for biological functioning. Our study provides support that animals and fungi are sibling kingdoms.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This study was supported by the Russian Science Foundation (Grant no. 14-14-00536) to O. V. Galzitskaya and by the Russian Academy of Sciences, Molecular and Cell Biology Program (Grant 01201353567) to M. Yu. Lobanov.

References

[1] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.

[2] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains *Archaea*, *Bacteria*, and *Eucarya*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[3] N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata, "Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 23, pp. 9355–9359, 1989.

[4] R. S. Gupta and B. Singh, "Cloning of the *HSP70* gene from *Halobacterium marismortui*: relatedness of archaebacterial *HSP70* to its eubacterial homologs and a model for the evolution of the *HSP70* gene," *Journal of Bacteriology*, vol. 174, no. 14, pp. 4594–4605, 1992.

[5] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, pp. 406–425, 1987.

[6] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specified tree topology," *Systematic Zoology*, vol. 20, no. 4, pp. 406–416, 1971.

[7] J. W. Pratt, "F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation," *The Annals of Statistics*, vol. 4, no. 3, pp. 501–514, 1976.

[8] R. A. Lippert, H. Huang, and M. S. Waterman, "Distributional regimes for the number of *K*-word matches between two random sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 13980–13989, 2002.

[9] G. Reinert, D. Chew, F. Sun, and M. S. Waterman, "Alignment-free sequence comparison I: statistics and power," *Journal of Computational Biology*, vol. 16, no. 12, pp. 1615–1634, 2009.

[10] L. Wan, G. Reinert, F. Sun, and M. S. Waterman, "Alignment-free sequence comparison (II): theoretical power of comparison statistics," *Journal of Computational Biology*, vol. 17, no. 11, pp. 1467–1490, 2010.

[11] X. Liu, L. Wan, J. Li, G. Reinert, M. S. Waterman, and F. Sun, "New powerful statistics for alignment-free sequence comparison under a pattern transfer model," *Journal of Theoretical Biology*, vol. 284, pp. 106–116, 2011.

[12] J. M. Comeron, R. Ratnappan, and S. Bailin, "The many landscapes of recombination in *Drosophila melanogaster*," *PLoS Genetics*, vol. 8, no. 10, Article ID e1002905, 2012.

[13] M. Gouy and W. H. Li, "Molecular phylogeny of the kingdoms animalia, plantae, and fungi," *Molecular Biology and Evolution*, vol. 6, no. 2, pp. 109–122, 1989.

[14] G. K. Philip, C. J. Creevey, and J. O. McInerney, "The Opisthokonta and the ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than ecdysozoa," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1175–1184, 2005.

[15] S. Yang, R. F. Doolittle, and P. E. Bourne, "Phylogeny determined by protein domain content," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 2, pp. 373–378, 2005.

[16] A. Löytynoja and M. C. Milinkovitch, "Molecular phylogenetic analyses of the mitochondrial ADP-ATP carriers: the plantae/fungi/metazoa trichotomy revisited," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 18, pp. 10202–10207, 2001.

[17] P. O. Wainright, G. Hinkle, M. L. Sogin, and S. K. Stickel, "Monophyletic origins of the metazoa: an evolutionary link with fungi," *Science*, vol. 260, no. 5106, pp. 340–342, 1993.

[18] S. L. Baldauf and J. D. Palmer, "Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 24, pp. 11558–11562, 1993.

[19] N. Nikoh, N. Hayase, N. Iwabe, K.-I. Kuma, and T. Miyata, "Phylogenetic relationship of the Kingdoms Animalia, Plantae, and Fungi, inferred from 23 different protein species," *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 762–768, 1994.

[20] V. G. Oxana, I. D. Eugeniya, and N. S. Igor, "Phylogenetic analysis of the loops in elongation factors EF1A: stronger support for the grouping of animal and fungi," *Journal of Computer Science and System Biology*, vol. 1, pp. 073–080, 2008.

[21] M. Y. Lobanov and O. V. Galzitskaya, "Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes," *Molecular BioSystems*, vol. 8, no. 1, pp. 327–337, 2012.

[22] M. Y. Lobanov and O. V. Galzitskaya, "Disordered patterns in clustered protein data bank and in eukaryotic and bacterial proteomes," *PLoS ONE*, vol. 6, no. 11, Article ID e27142, 2011.

[23] M. Y. Lobanov, E. I. Furletova, N. S. Bogatyreva, M. A. Roytberg, and O. V. Galzitskaya, "Library of disordered patterns in 3D protein structures," *PLoS Computational Biology*, vol. 6, no. 10, Article ID 1000958, 2010.

[24] M. Y. Lobanov, I. V. Sokolovskiy, and O. V. Galzitskaya, "IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model," *Journal of Biomolecular Structure and Dynamics*, vol. 31, no. 10, pp. 1034–1043, 2013.

[25] E. Salichs, A. Ledda, L. Mularoni, M. M. Albà, and S. de la Luna, "Genome-Wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment," *PLoS Genetics*, vol. 5, no. 3, Article ID e1000397, 2009.

[26] J. P. Zbilut, G. H. Chua, A. Krishnan, C. Bossa, M. Colafranceschi, and A. Giuliani, "Entropic criteria for protein folding derived from recurrences: six residues patch as the basic protein word," *FEBS Letters*, vol. 580, no. 20, pp. 4861–4864, 2006.

[27] O. V. Galzitskaya, "Search for folding initiation sites from amino acid sequence," *Journal of Bioinformatics and Computational Biology*, vol. 6, no. 4, pp. 681–691, 2008.

[28] G. V. Nikiforovich and C. Frieden, "The search for local native-like nucleation centers in the unfolded state of β-sheet proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 16, pp. 10388–10393, 2002.

[29] M. Y. Lobanov, N. S. Bogatyreva, and O. V. Galzitskaya, "Occurrence of six-amino-acid motifs in three eukaryotic proteomes," *Molecular Biology*, vol. 46, no. 1, pp. 168–173, 2012.

[30] T. Cavalier-Smith, "The origin of fungi and pseudogungi," in *Evolutionary Biology of Fungi*, A. D. M. Rayner, C. M. Braisier, and D. Moore, Eds., pp. 339–353, Cambridge University Press, Cambridge, UK, 1987.

[31] K. Buck, "Phylum zoomastigina, class choanomastigotes (choanoflagellates)," in *Handbook of Protista*, L. Margulis, J. O.

Corliss, M. Melkonian, and D. J. Chapman, Eds., pp. 194–199, Jones and Bartlett Publishers, 1990.

[32] I. Seim, X. Fang, Z. Xiong et al., "Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*," *Nature Communications*, vol. 4, article 2212, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 864804, 13 pages
http://dx.doi.org/10.1155/2015/864804

Research Article

Molecular Biogeography of Tribe Thermopsideae (Leguminosae): A Madrean-Tethyan Disjunction Pattern with an African Origin of Core Genistoides

Ming-Li Zhang,[1,2] Jian-Feng Huang,[1,3] Stewart C. Sanderson,[4] Ping Yan,[5] Yu-Hu Wu,[6] and Bo-Rong Pan[1]

[1]*Key Laboratory of Biogeography and Bioresource in Arid Land, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, Xinjiang 830011, China*
[2]*Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China*
[3]*Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming, Yunnan 650223, China*
[4]*Shrub Sciences Laboratory, Intermountain Research Station, Forest Service, U.S. Department of Agriculture, UT 84601, USA*
[5]*School of Life Science, Shihezi University, Shihezi, Xinjiang 832003, China*
[6]*Northwest Plateau Institute of Biology, Chinese Academy of Sciences, Xining, Qinghai 810001, China*

Correspondence should be addressed to Ming-Li Zhang; zhangml@ibcas.ac.cn and Bo-Rong Pan; brpan@ms.xjb.ac.cn

Received 9 September 2014; Revised 20 December 2014; Accepted 8 January 2015

Academic Editor: Peter F. Stadler

Thermopsideae has 45 species and exhibits a series of interesting biogeographical distribution patterns, such as Madrean-Tethyan disjunction and East Asia-North America disjunction, with a center of endemism in the Qinghai-Xizang Plateau (QTP) and Central Asia. Phylogenetic analysis in this paper employed maximum likelihood using ITS, *rps*16, *psb*A-*trn*H, and *trn*L-F sequence data; biogeographical approaches included BEAST molecular dating and Bayesian dispersal and vicariance analysis (S-DIVA). The results indicate that the core genistoides most likely originated in Africa during the Eocene to Oligocene, ca. 55-30 Ma, and dispersed eastward to Central Asia at ca. 33.47 Ma. The origin of Thermopsideae is inferred as Central Asian and dated to ca. 28.81 Ma. *Ammopiptanthus* is revealed to be a relic. Birth of the ancestor of Thermopsideae coincided with shrinkage of the Paratethys Sea at ca. 30 Ma in the Oligocene. The Himalayan motion of QTP uplift of ca. 20 Ma most likely drove the diversification between Central Asia and North America. Divergences in East Asia, Central Asia, the Mediterranean, and so forth, within Eurasia, except for *Ammopiptanthus*, are shown to be dispersals from the QTP. The onset of adaptive radiation at the center of the tribe, with diversification of most species in *Thermopsis* and *Piptanthus* at ca. 4-0.85 Ma in Tibet and adjacent regions, seems to have resulted from intense northern QTP uplift during the latter Miocene to Pleistocene.

1. Introduction

In the Leguminosae, the so-called core genistoides includes tribes Crotalarieae, Genisteae, Podalyrieae, Thermopsideae, Euchresteae, and Sophoreae sensu strictu [1–8]. Tribe Thermopsideae includes seven genera and about (43)-45-(46) species and occurs in the temperate regions of Eurasia and North America [6, 9]. Of them, *Pickeringia*, with one species endemic to western North America, has been transformed into *Cladrastis–Styphnolobium* [5, 10]. *Thermopsis* and *Baptisia* are two perennial herbaceous genera, respectively, in distributions of an East Asian-North American disjunction and North American endemism. *Anagyris, Piptanthus*, and *Ammopiptanthus* are shrubby and in Eurasia. *Anagyris* includes two species and occurs around the Mediterranean Basin [11, 12]. *Piptanthus* and *Ammopiptanthus* mainly occur in China, the former in Sino-Himalayan [13] and the latter in Central Asian regions [14]. New monotypic genus *Vuralia* recently is segregated from *Thermopsis* in Turkey [15].

Molecular evidence above the rank of genus has provided a foundation for Thermopsideae systematics and biogeography [4, 5, 16–19]. However, due to a lack of sufficient

species sampling, a dense addition of species at generic level is necessary. Wang et al. [14] carried out a comprehensive systematic study of Thermopsideae on the basis of dense species addition and ITS sequences. Biogeographically, the fossil record indicates that the three legume subfamilies appeared in the early Eocene, and extensive diversification and origin of most of the woody legume lineages occurred in the middle Eocene [20]. Schrire et al. [21] divided the distribution patterns of ca. 730 legume genera into four biomes, that is, succulent, grass, rainforest, and temperate, with temperate groups possessing the largest numbers. From macrofossils of leaves and pods, the origin of legumes appears unlikely to have been much before 60 Ma, and, from that time, a rapid diversification among major clades took place [22]. In contrast with a proposed West Gondwana origin of the family [23, 24] or a "moist equatorial megathermal" origin, recent studies favor an origin in the seasonally dry to arid tropical Tethyan seaway corridor [21]. Lavin et al. [22] established a comprehensive schematic chronogram of legumes based on sequence data and fossil constraint, employing a total of 324 species. However, only three species of Thermopsideae were sampled. Lavin et al. [22] estimated ca. 26.5 Ma for the time of origin of Thermopsideae, and Ortega-Olivencia and Catalán [12] dated the appearance of *Anagyris* to late Miocene (8.2 ± 4.5 Ma). Xie and Yang [25] estimated *Ammopiptanthus* to have originated in early Miocene ca. 20-21 Ma.

Exceptionally, in the Wang et al. [14] study, although there was strong support for the tribal clade, the systematic position of *Ammopiptanthus* was suspected as not being a member of Thermopsideae because of the nesting of three *Sophora* species with it, resulting in *Ammopiptanthus* being placed in a basally branching position with respect to the rest of the tribe. Many studies have speculated that *Piptanthus, Ammopiptanthus, Thermopsis*, and so forth in the tribe originated in the Tertiary [14, 26–29], but the exact time and place of origin have remained poorly understood.

In summary, Thermopsideae contains many attractive biogeographical topics, Central Asia, East Asia, and QTP endemism and Madrean-Tethyan disjunction, East Asia-North America disjunction, Tertiary origin, and so forth. Therefore, this paper attempts to reconstruct the phylogeny of the tribe using four genes and, afterward, focuses on the tribe Thermopsideae biogeography by employing biogeographical molecular dating and S-DIVA approaches to explore the spatiotemporal origin and evolution of Thermopsideae and its evolutionary dynamics; to confirm the Madrean-Tethyan disjunction using Thermopsideae; and to discuss the East Asia-North America disjunction, Central Asian endemism, QTP endemism, and so forth.

2. Materials and Methods

2.1. Taxon Sampling. We sampled 32 individuals of 20 species, mainly from China, belonging to three genera, *Thermopsis*, *Piptanthus*, and *Ammopiptanthus* of Thermopsideae; see Table 1. Outgroups were selected from *Sophora* (*S. davidii, S. flavescens,* and *S. microphylla*), *Podalyria* (Podalyrieae), and *Cytisus* (Genisteae); see Supplementary Material S 1 available online at http://dx.doi.org/10.1155/2015/864804. More outgroup species were used in ITS phylogeny; also see S 2.

2.2. DNA Sequencing. Total genomic DNA was extracted using the CTAB method [30]. The polymerase chain reaction (PCR) was used for amplification of double stranded DNA. A 25 μL reaction system contained 0.25 μL of Ex Taq, 2.5 μL of 10× Ex Taq buffer (Mg^{2+} concentration of 25 mM), 2.0 μL of dNTP mix (2.5 mM concentration for each dNTP), 1 μL of the forward and reverse primers at 5 μmol/μL, and 0.5 μL of template DNA. The following primers were used: for ITS, ITS1-F (5′-AGA AGT CGT AAC AAG GTT TCC GTA GC-3′) and ITS4-R (5′-TCC TCC GCT TAT TGA TAT GC-3′) [31], for trnL-F, trnLF (5′-CGA AAT CGG TAG ACG CTA CG-3′) and trnFR (5′-ATT TGA ACT GGT GAC ACG AG-3′) [32], for *psb*A-*trn*H, *psb*AF (5′-GTT ATG CAT GAA CGT AAT GCT C-3′) [33] and *trn*HR (5′-CGC GCA TGG ATT CAC AAT CC-3′) [34], and, for the intron of rps16, *rps*16F (5′-GTG GTA GAA AGC AAC GTG CGA CTT-3′), and for *rps*16R (5′-TCG GGA TCG AAC ATC AAT TGC AAC-3′) [35].

PCR amplifications were carried out using the following procedures: there was predenaturation at 94°C for 3 min., followed by 30 cycles of (1) denaturation at 94°C for 30 s, (2) annealing at 48°C–54°C for 30 s, and (3) extension at 72°C for 1 min.; at the end of these cycles, there was a final extension at 72°C for 10 min. PCR products were purified using the PEG precipitation procedure [36]. Sequencing reactions were performed by a company specializing in the procedure (Beijing Sanbo Biological Engineering Technology and Service Corporation, China). Sequences were aligned using CLUSTAL X software [37] and then adjusted by hand in BioEdit ver. 5.0.9 [38].

2.3. Phylogenetic Analyses. Two datasets consisting of ITS and the 4-gene sequences combined (ITS+3cpDNA) were prepared for phylogenetic analysis. The 4-gene dataset was examined using the incongruence length difference (ILD) tests [39], implemented in PAUP version 4.0b10 [40], with 100 homogeneity replicates, 10 random addition sequences, tree-bisection-reconnection (TBR) branch swapping on best only, and MULTREES on, and this was performed to test whether the four datasets could be combined. The data partitions of four genes were not significantly incongruent on the basis of the ILD tests ($P = 0.01$).

Phylogenetic analysis by Maximum Likelihood (ML) of the 4-gene combined sequences was conducted using PAUP* 4.0b10 [40].

For ML analysis, the best fitting DNA substitution model was found employing Modeltest 3.6 [41], of which the Akaike information criterion (AIC) was selected on the basis of the log likelihood scores of 56 models [41]. For the dataset, the TrN+G model was selected as the most appropriate in Modeltest 3.5, with the nucleotide frequencies A = 0.3283, C = 0.1676, G = 0.1985, T = 0.3055, the shape parameter = 0.6264, and an assumed proportion of invariable (PIV) sites = 0. Clade support for the phylogenetic tree was estimated,

Table 1: Sources of plant materials.

Species	Voucher	Source	ITS	*rps*16	*psb*A-*trn*H	*trn*L-F
Ammopiptanthus S.H. Cheng						
A. mongolicus (Maxim.) S.H. Cheng	W.J. Zhu 64004 (HNWP)	Wuda, Inner Mongolia, China		KP636600	KP636576	KP636624
A. mongolicus (Maxim.) S.H. Cheng	M.L. Zhang s.n. (XJBI)	Turpan Eremophytes Botanic Garden, China	KP636562			KP636625
A. nanus (Popov) S.H. Cheng	P. Yan, M. Ma 4280 (SHI)	Wuqia, Xinjiang, China		KP636601	KP636577	KP636626
A. nanus (Popov) S.H. Cheng	Y.H. Wu 870001 (HNWP)	Wuqia, Xinjiang, China				KP636627
A. nanus (Popov) S.H. Cheng	M.L. Zhang s.n. (XJBI)	Turpan Eremophytes Botanic Garden, China	KP636563			KP636628
Piptanthus Sweet						
P. concolor Harrow ex Craib	Tibet Medicine Exp. Team 213 (HNWP)	Jilong, Tibet, China	KP636564	KP636602	KP636578	KP636629
P. laburnifolius (D. Don) Stapf	Qinghai Exp. Team 750501 (HNWP)	Longzi-Zhunba, Tibet, China	KP636565	KP636603	KP636579	KP636630
P. leiocarpus Stapf	Tibet Medicine Exp. Team 1576 (HNWP)	Nielamu, Tibet, China	KP636566	KP636604	KP636580	
P. nepalensis Sweet	ITS: Wang HC, 0121 (KUN); Ciduo Cidan, et al. 2436 (PE)	ITS: Yunnan, China; Yadong, Tibet, China	AF215922		KP636581	KP636631
Thermopsis R. Br.						
T. alpina Ledeb.	ITS: Saren 2000; Y.H. Wu 29102 (HNWP)	ITS: Tibet, China; Maduo, Qinghai, China	AF123447	KP636605	KP636582	KP636632
T. alpina Ledeb.	Y.H. Wu 28862 (HNWP)	Yushu, Qinghai, China		KP636606		KP636633
T. alpina Ledeb.	P. Yan, J.Y. Guo 6790 (SHI)	Manasi, Xinjiang, China		KP636607	KP636583	KP636634
T. alpina Ledeb.	Pamier Exp. Team 5233 (SHI)	Tashikuergan, Xinjiang, China	KP636567	KP636608	KP636584	KP636635
T. alpina Ledeb.	J. Tao, et al. 1067 (SHI)	Manasi, Xinjiang, China			KP636585	KP636636
T. barbata Benth.	Tibet Medicine Exp. Team 4329 (HNWP)	Jiacha, Tibet, China	KP636568	KP636609		KP636637
T. inflata Cambess.	ITS: Liu JQ s.n.; Z.Y. Wu, S.K. Chen, Q. Du 75–166 (HNWP)	ITS: Qinghai, China; Bogu lake-Malashan, Tibet, China	AF123451	KP636610	KP636586	KP636638
T. inflata Cambess.	XJBI Tibet Team s.n. (XJBI)	Zada, Tibet, China		KP636611		KP636639
T. inflata Cambess.	XJBI Tibet Team s.n. (XJBI)	Gaize, Tibet, China		KP636612	KP636587	KP636640
T. kaxgarica Ch.Y. Yang	C. Yan s.n. (XJBI)	Turpan, Xinjiang, China		KP636613	KP636588	KP636641
T. lanceolata R. Br.	ITS: Saren 010 (PE); Y.H. Wu 36480 (HNWP)	ITS: Qinghai, China; Dulan, Qinghai, China	AF123448	KP636614	KP636589	KP636642
T. lanceolata R. Br.	XJBI Exp. Team s.n. (XJBI)	Sawuershan, Xinjiang, China		KP636615	KP636590	KP636643
T. licentiana E. Peter	R.F. Huang 2677 (HNWP)	Tianzhu, Gansu, China	KP636569	KP636616	KP636591	KP636644
T. licentiana E. Peter	Y.H. Wu 1014 (HNWP)	Yecheng, Xinjiang, China		KP636617		
T. licentiana E. Peter	Guoluo Exp. Team 649 (HNWP)	Jiuzhi, Qinghai, China			KP636592	KP636645
T. licentiana E. Peter	H.B.G. 198 (HNWP)	Maxin, Qinghai, China			KP636593	KP636646
T. mongolica Czefr.	P. Yan 3368 (SHI)	Hefeng, Xinjiang, China	KP636570	KP636618	KP636594	KP636647
T. przewalskii Czefr.	B.Z. Guo 10267 (HNWP)	Tongren, Qinghai, China	KP636571	KP636619	KP636595	KP636648
T. schischkinii Czefr.	S.M. Duan s.n. (XJBI)	Wuma, Tibet, China	KP636572	KP636620	KP636596	KP636649
T. smithiana E. Peter	XJBI Exp. Team s.n. (XJBI)	Geji, Tibet, China	KP636573	KP636621	KP636597	KP636650
T. turkestanica (Gand.) Ch.Y. Yang	Tibet-Xinjiang Exp. Team 1044 (HNWP)	Zhaosu, Xinjiang, China	KP636574	KP636622	KP636598	KP636651
T. yushuensis S.Q. Wei	S.W. Liu 609 (HNWP)	Yushu, Qinghai, China	KP636575	KP636623	KP636599	KP636652

HNWP (Herbarium, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining, Qinghai); PE (China National Herbarium, Institute of Botany, Chinese Academy of Sciences, Beijing); SHI (Herbarium, Shihezi University, Shihezi, Xinjiang); and XJBI (Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, Xinjiang).

Table 2: References for fossils of seven genera used to constrain ages for dating.

Taxa	Time (Ma)	Location	Reference
Cercis	60–11 Late Cretaceous-Miocene	China	Tao, 1992 [58]; Tao et al., 2000 [42]
Cercis	Eocene	N America	Lavin et al., 2005 [22]
Acacia	47–42 Eocene	Liaoning, China; Tanzania	Tao et al., 2000 [42]; Lavin et al., 2005 [22]
Acacia	15 Miocene	Dominican Rep.	Lavin et al., 2005 [22]
Bauhinia	ca. 65 later Cretaceous	Helongjiang, China,	Tao et al., 2000 [42]
Cladrastis	40–20 middle Eocene	Tennessee, N America	Herendeen et al., 1992 [20]
Cladrastis	Miocene	Inner Mongolia	Tao et al., 2000 [42]
Sophora	35–9 Oligocene-Miocene	China, Siberia, N America	Tao et al., 2000 [42]; IB & NIGP, 1978 [43]
Pueraria	17–5 Miocene	Yunnan, Shandong, China	Tao, 1992 [58]; Tao et al., 2000 [42]
Dalbergia	19.5–5 Miocene	China, N America, Europe	IB & NIGP, 1978 [43]; Tao, 1992 [58]; Tao et al., 2000 [42]

IB and NIGP: Institute of Botany and Nanjing Institute of Geology and Palaeontology, Academia Sinica.

employing bootstrap values in PAUP and posterior probability values in MrBayes software.

In order to obtain comprehensive molecular dating, ITS sequence data covered broad outgroups including the four core genistoides tribes and seven fossil genera, which came from our data and from GenBank; see S 2. The final dataset comprised 107 species and 722 bps.

2.4. Estimating Divergence Times

2.4.1. Fossil Constraints. Legumes have rich fossil records [20, 42], but there are fewer fossils assignable to the core genistoides and Thermopsideae. Most fossils of legume genera date to the Miocene, with several approaching the Eocene and Paleocene, with *Bauhinia* and *Cercis* extending even to the Late Cretaceous [42, 43]. Seven fossil genera are used as outgroups; see Table 2. The occurrence of *Sophora* in the Oligocene-Miocene is credible, since its fossils are known from the Eocene of eastern Siberia and North America [42, 43]. In China, *Sophora* fossils have been found in Oligocene strata from Heilongjiang province, the Miocene from Shandong and Yunnan provinces, and Pliocene from Shanxi province [42, 43]. *Acacia* in the Eocene is also well represented in museum collections. *Dalbergia* fossils, including leaves and fruit, have been recorded at the Eocene-Miocene boundary in North America, the Oligocene-Miocene boundary in Europe, and in the Miocene in Yunnan Province, China. Fossil leaves of *Pueraria* appeared in the Miocene in Shandong and Yunnan provinces, China [42, 43], and *Cladrastis* has been dated to middle Eocene [22].

In terms of the ancient fossil record and the phylogenetic tree, the root taxon was considered as *Cercis*. In the southern China Guangdong province, *Cercis* fossils haves been found from the Late Cretaceous to Eocene, in the Oligocene of Yunnan province, and the Miocene of Shandong and Qinghai provinces. Therefore, this genus is regarded as the root taxon and the age of its ancestor is constrained as 60 Ma. This root constraint is in agreement with Lavin et al. [22], by whom the ancestor of *Polygala* and *Cercis* was constrained at 60 Ma. Detailed information are described in Table 2.

The outgroup fossil dates were used as the constraint minimum ages; that is, the maximum fossil dates were selected as the generic minimum age of the most recent common ancestor (MRCA).

2.4.2. Dating Implementation. Currently, phylogenetic dating approaches include r8s, PAML, and BEAST. Of them, BEAST has an advantage for practical applications because of its non-dependence on a phylogenetic tree, and convenient implementation software (BEAST v1.46, http://beast.bio.ed.ac.uk). Moreover, a relaxed molecular clock and Bayesian MCMC search optima are available within it [44, 45].

BEAST was implemented [46] using a Yule process speciation prior to an uncorrelated lognormal model of rate variation and a normal distribution. Tracer v1.4 was used to measure the effective sample size of each parameter and mean and 95% credibility intervals. Two separate MCMC analyses were run for 20,000,000 generations and sampled every 1000 generations. After discarding as burn-in the first 10% of trees searched, the mean and 95% credibility intervals of MRCA nodes were calculated by TreeAnnotator v1.4.8. and visualized by FigTree v.1.2.4 [46].

2.5. Biogeographic S-DIVA. DIVA is used to infer mainly ancestral distributions and biogeographical events [47]; it is an event-based method that optimizes ancestral distributions by assuming a vicariance explanation, while incorporating the potential contributions of dispersal and extinction, despite minimizing these under a parsimony criterion [47, 48]. Nylander et al. [49] proposed a modified approach to DIVA naming it Bayes-DIVA because it integrates biogeographical reconstructions of DIVA over the posterior distribution of a Bayesian MCMC sample of tree topologies. Bayes-DIVA is also referred to as S-DIVA [50].

The BEAST dating tree (Figure 2) was treated as a fully resolved phylogram for use as a basis for S-DIVA, and 791 post burnin trees derived from the BEAST analysis were used for ancestral area reconstruction in the program S-DIVA beta version 1.9. S-DIVA was performed with constraints of maximum areas 2 at each node, to infer possible ancestral areas and potential vicariance and dispersal events.

Geographic areas were chosen to cover the distributions of the four core genistoides tribes, especially tribe Thermopsideae. Seven geographic endemic areas were defined in this

study: East Asia, Central Asia, the Mediterranean, Africa, Russia (including Central East, Caucasus, and northeastern Russia), North America, and Tibet. Because of its species richness and endemism, the QTP, Tibet is regarded as an area separated from the East Asian floristic region [13].

3. Results

3.1. Phylogenetic Analyses

3.1.1. 4-Gene Combined Analysis. The 4-gene combined dataset included 38 samples and 3099 bps; 496 variable characters were parsimony-uninformative and 421 were parsimony-informative. ML analysis resulted in three optimum trees, topologically almost equivalent; one of them is shown in Figure 1. Bootstrap support from PAUP and Bayesian posterior probability are labeled on the nodes in Figure 1.

Thermopsideae and *Ammopiptanthus*, respectively, formed a monophyletic group with high support, near 100% bootstrap (BT) and posterior probability values (PP). *Piptanthus* did not form a monophyletic group, since *P. nepalense* was placed outside of the genus (Figure 1). Even though the samples of *Thermopsis* came only from China (Figure 1), the results show that section *Thermopsis* sensu Sa et al. [28] can apparently be divided into two clades (Figure 1).

3.1.2. ITS Analysis. The ITS BEAST implementation yielded a phylogenetic tree and dating chronogram; see Figure 2. This topology of tree is in rough agreement with that of the previous ITS tree [14] and our 4-gene tree (Figure 1). It indicates that *Ammopiptanthus* and Thermopsideae are monophyletic groups, respectively. Importantly, this chronogram has a temporal evolutionary significance for the Thermopsideae, Podalyrieae, Crotalarieae, Genisteae, and so forth.

In contrast with previous phylogenies, especially Wang et al. [14], this ITS tree places *Ammopiptanthus* within Thermopsideae rather than outside of the tribe, and this is the same as in our 4-gene tree (Figure 1). The topological structure of the four tribes of core genistoides is also somewhat different from previous studies [19]; Thermopsideae is related to the cluster of Genisteae and Podalyrieae, while Crotalarieae is more isolated.

3.1.3. Estimating Divergence Times. Using seven fossil genera as constraints and outgroups, for 107 species and ITS dataset, the estimated root age of the four tribes of core genistoides was ca. 54.43 Ma and that of Thermopsideae was ca. 28.81 Ma, as presented in Figure 2. The estimated crown ages of the four tribes range from later Eocene 39.45 Ma (Crotalarieae) to Miocene 11.89 Ma (Podalyrieae).

Within Thermopsideae, five genera are well monophyletic, respectively, with credible crown and stem ages excluding *Thermopsis*. *Ammopiptanthus* has stem age ca. 28.81 Ma, namely, crown age of Thermopsideae. In order to discuss the origin and evolution of taxa, a geological scale was appended to the BEAST diagram in Figure 2.

3.1.4. Biogeographic S-DIVA. Reconstruction of ancestral areas with S-DIVA (Figure 2) suggested that the ancestral distribution area of core genistoides is Africa (A) and that of Thermopsideae is possibly Central Asia (C) and that *Ammopiptanthus* is directly derived from Thermopsideae. Extant taxa of North America and the QTP are shown to be dispersals from Central Asia; several events of dispersal and vicariance are illustrated in Figure 2. The most distinct dispersal for the core genistoides is from Africa to Central Asia. From the QTP, a dispersal was westward via the Himalayas to the Mediterranean for the genus *Anagyris*, and dispersal and adaptive radiation to East Asia and Central Asia. The eastward line is via the Bering Strait to North America (Figures 2 and 3).

4. Discussion

4.1. Systematics of Thermopsideae. In the previous ITS phylogenetic tree [14], the five genera of Thermopsideae formed a well resolved monophyly, except for the fact that *Ammopiptanthus* fell outside of Thermopsideae due to the nesting of a few species of *Sophora*. In addition, diversification into East Asian and North American groups is observed in *Thermopsis*. From our 4-gene combined (Figure 1) and ITS trees (Figure 2), the monophyly of Thermopsideae and *Ammopiptanthus* is confirmed once more, and *Ammopiptanthus* is entirely included within Thermopsideae with high confidence support. Consequently, our enhanced species sampling and 4-gene tree (Figure 1) yield a distinct result compared with Wang et al. [14], mainly, that Thermopsideae is monophyletic, since the three *Sophora* species (*S. davidii, S. flavescens*, and *S. microphylla*) are out of the tribe; two phylogenetic clades are recognized, where the previous tree only had one (see Figure 1 of Wang et al. [14]; our ITS tree also has one clade see Figure 2). This probably will be useful for the revision of classification [28, 51, 52], especially for section *Thermopsis* sensu Sa et al. [28], of which most species occur in Asia. In addition, the previous taxonomic opinion of *Ammopiptanthus* being morphologically related to *Piptanthus* [26, 53] should be considered as reflecting a convergence, since our results (Figures 1 and 2) illustrate that they are separated in the tree. *Vuralia* has only one species *V. turcica* (Kit Tan et al.) and has a narrowed distribution (marshy side of Aksehir in Turkey) Uysal et al. [15], even though it had not been joined into the present dataset of Thermopsideae; however, together with *Thermopsis chinensis* and *Th. fabacea*, all are shown to be included into North America clade node 9 (Figure 2) [15].

4.2. Age and Distribution Pattern of Thermopsideae. The age of Thermopsideae has been estimated several times by the molecular dating approach. On the basis of fossil data, the genistoides crown node was constrained at 56.42 ± 0.2 Ma. Lavin et al. [22] dated Thermopsideae to ca. 26.5 Ma, but only three species were sampled, that is, *Piptanthus nepalense, Baptisia australis*, and *Thermopsis rhombifolia*. This node of the genistoides is placed at ca. 54.43 Ma (Figure 2), near this fossil constraint, confirming the validity of our dating.

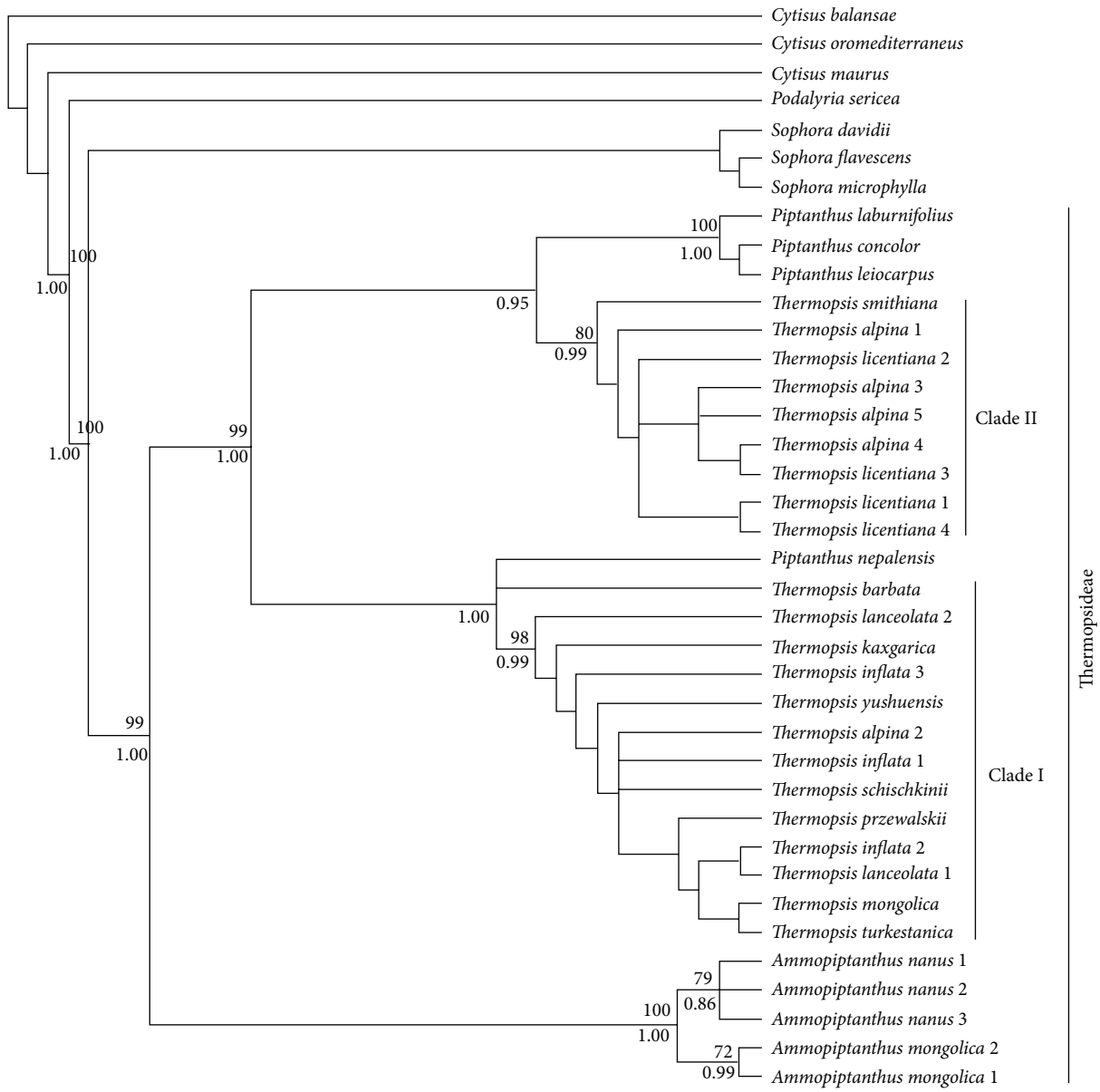


Figure 1: Phylogenetic tree resulted from maximum likelihood analysis of the combined dataset of 4 genes (ITS, *trn*L-F, *psb*A-*trn*H, and *rps*16). Bootstrap support values > 50% above branches and posterior probability support > 0.5 below branches are indicated.

To estimate the age of *Anagyris*, Ortega-Olivencia and Catalán [12] employed numerous samples of the two species *A. foetida* and *A. latifolia* and added three other species in the tribe. Their results indicated that an estimated age of Thermopsideae was 27.2 ± 4.1 Ma and of *Anagyris* was 8.2 ± 4.5 Ma. The present paper dates Thermopsideae to ca. 28.81 Ma, which approaches the dates from previous studies by Lavin et al. [22] and Ortega-Olivencia and Catalán [12]. Therefore, the middle Oligocene ca. 28.81 Ma should be treated as the diversification age of the tribe.

Along with the significant global climate cooling and increased aridity from Eocene to Oligocene, seven distinctive biomes have been recognized for the Oligocene (38~24 Ma) [54]. At ca. 30 Ma, one of seven is the warm/cool temperate biome, with a wide band of broadleaved evergreen and deciduous woodland throughout central Eurasia and North America. This biome in its northernmost part just covers the distribution range of Thermopsideae. These woodlands and forests replaced the dominantly evergreen paratropical rainforest of the middle Paleocene and much of the Eocene [54]. Therefore, we can determine that the original accompanying vegetation of Thermopsideae was woodlands and forest, with broadleaved evergreen and deciduous plants. From the Oligocene ca. 30 Ma to middle-to-late Miocene, shrinkage of the Paratethys played an important role in causing transformation of the Central Asian climate from an oceanic to a continental condition [55]. As Hrbek and Meyer [56] have reviewed, the closing of the sea near

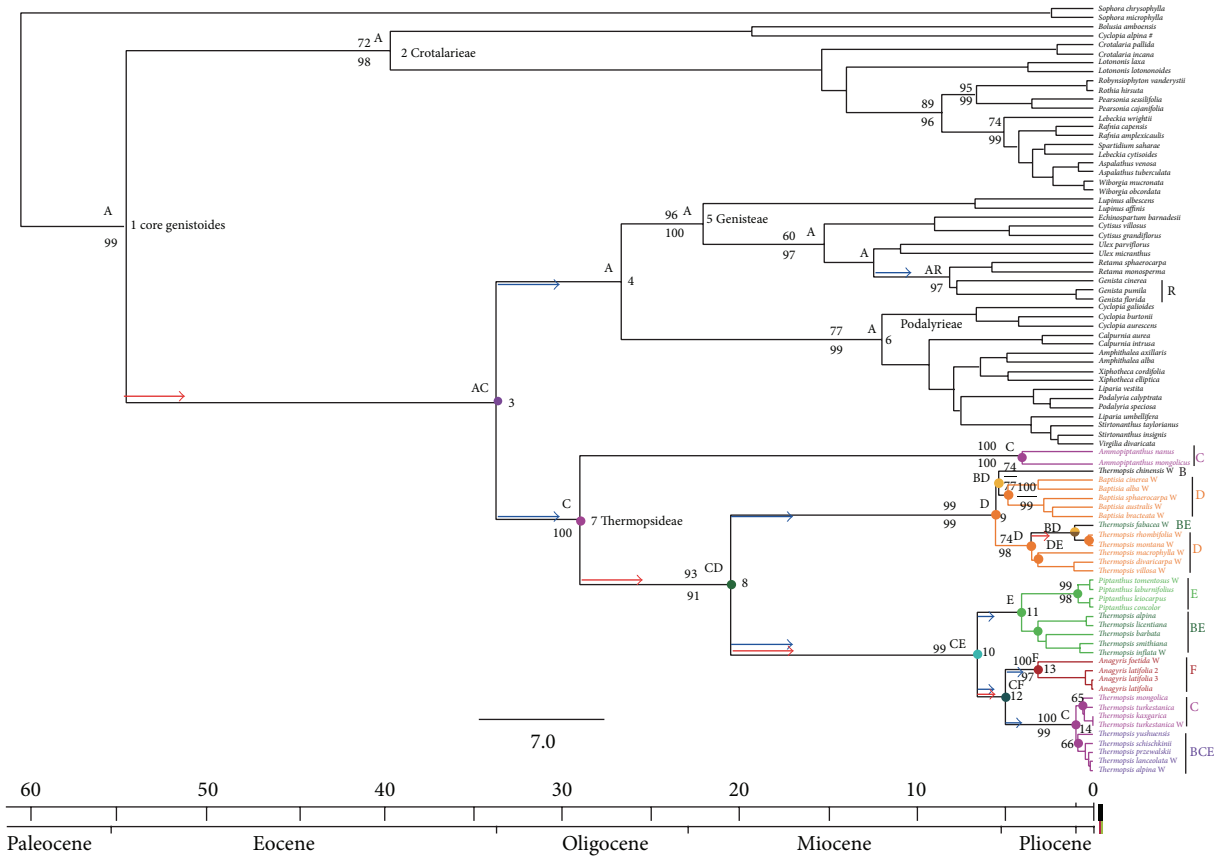


FIGURE 2: Chronogram of relaxed Bayesian BEAST on the basis of the ITS dataset. Estimated times (Ma) with 95% HPD credibility intervals at concerned nodes were 1 : 54.43 (53.04–58.85), 2 : 39.45 (16.98–54.54), 3 : 33.47, 4 : 26.51 (13.59–42.87), 5 : 21.91 (7.09–23.65), 6 : 11.89 (4.78–15.64), 7 : 28.81 (10.95–41.02), 8 : 20.32 (6.9–22.98), 9 : 5.48 (2.99–10.81), 10 : 6.5 (3.38–11.11), 11 : 4, 12 : 4.9 (1.6–6.89), 13 : 3.08 (0.26–3.2), and 14 : 0.97 (0.46–2.8). Bootstrap support values > 50% above branches and posterior probability values > 0.5 below branches are indicated. "W" behind species name means species with the sequence data come from GenBank produced by Wang et al. [14]. S-DIVA optimal reconstruction of hypothesized ancestral areas at nodes and 10 dispersals with vertical line on branches are illustrated.

the Oligocene/Miocene boundary had a major impact on the distribution of organism diversity. Therefore, origin and diversification of the Thermopsideae at ca. 30 Ma could therefore have been driven by the closing of the Paratethys, which resulted in a series of changes of environmental and ecological factors and profoundly affected the evolution of the tribe.

The Oligocene environment in Kazakhstan, with broad-leaved forest and swamps, was indicated to be a wet climate [57]. However, the Paleogene floristics of northwestern and central China evidenced by fossil data was dry and subtropical [58]. During the Oligocene, the climate of middle China is speculated to have been an arid/semiarid belt [59]. Therefore, climate in Oligocene Central Asia should have changed from western wet (relic locations of Paratethys shrinkage) in Kazakhstan to eastern dry in northwestern China. These wet environments and climates of western parts of Central Asia most likely fit the emergence of the ancestor of *Ammopiptanthus* and Thermopsideae, with a broad-leaved forest and a wet to arid climate.

Therefore, from these perspectives of time and place of origin, paleovegetation, and paleoclimate, we can confirm a balanced Oligocene Central Asian origin of Thermopsideae.

4.3. Central Asian Origin and Diversifications among Central Asia, the QTP, and North America. In terms of our inferences of dispersal and vicariance in Thermopsideae (Figures 2 and 3), we can speculate that, after origination in Central Asia, most of its broadleaved evergreen and deciduous ancestors probably soon became extinct.

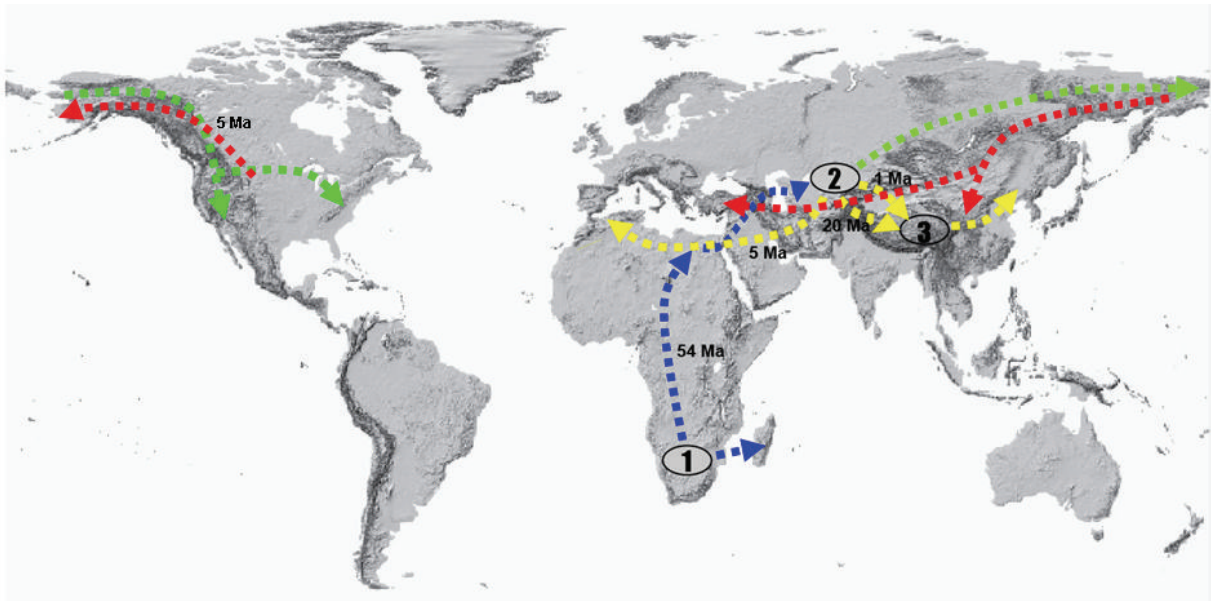


Figure 3: Scheme of dispersal routes from the biogeographical S-DIVA analysis, Figure 2. The blue dashed line indicates the origin center of core genistoides from Africa (elliptic 1), arriving in Central Asia (elliptic 2), which was the place of origin of the Thermopsideae; then there was a dispersal to North America. Routes are shown by green dashed lines. From Central Asia (elliptic 2), a dispersal westward via the Caucasus to the Mediterranean with genus *Anagyris* and dispersal and adaptive radiation to QTP (elliptic 3), from QTP to North China, are shown by yellow dashed lines. Dispersal from North America at about late Miocene eastward to East Asia and Mediterranean is shown by red lines.

Only a few survived, a case being *Ammopiptanthus*, which evolved from the ancestor of Thermopsideae (Figure 2). From the Central Asian ancestor, Thermopsideae had a dispersal to North America at ca. 28.81 Ma and to the QTP at ca. 20.32 Ma (Figures 2 and 3).

As mentioned above, shrinkage of the Paratethys starting from the Oligocene [55] was a dynamic influence for Thermopsideae and most likely also drove the dispersal to North America (dispersal from node 7 Figures 2 and 3). The diversification age between Central Asia and QTP of ca. 20.32 Ma implies a response to QTP uplift and extension as a geological event. QTP uplift is presumed to have initiated very early but had a major phase near the Oligocene-Miocene boundary when loess deposition began and strengthened thereafter. The first phase (Gangdese motion ca. 40 Ma, 45~38 Ma) is characterized by the Indian plate subducting under the Eurasian plate, resulting in the rise of the Gangdise Mountains. High altitude conifers became abundant in the QTP starting at 38 Ma [60]. The second phase (Himalayan motion ca. 21 Ma, 25~17 Ma) is characterized by westward withdrawal of the Paratethys Sea and aridification of interior Asia. The rise or expansion of the QTP became sufficient for the initiation of dust deposition due to the Asian winter monsoon [59, 61, 62]. The third phase, uplift of northern and eastern parts of the QTP at many intervals during the late Neogene to Pleistocene [63–65] is correlated with appearance of ocean upwelling connected to development of the Asian summer monsoons. A particularly intense geologic uplift during this period was recorded in parts of the northernmost QTP at ca. 3.6 Ma [63, 66–68], which was accompanied by the intensification of monsoons to present levels.

These events strikingly influenced the ecology and environment of the QTP and adjacent regions, especially northern China. These ecological and environmental settings, consequently, can hold temporally the dispersal from Central Asia (see Figure 2, node 8) into the QTP.

The QTP and adjacent regions possess many endemic species in Thermopsideae (nodes 11,14 in Figure 2); these areas are shown by the analysis to be related mechanistically to Central Asia. A vicariance between the QTP and Central Asia is estimated at ca. 6.5 Ma (node 10, Figure 2), whereas those species of the QTP, North China, and Central Asia (with BCE in Figure 2) are shown to come from the Central Asian ancestor node 14 (Figure 2) at ca. 0.97 Ma. Therefore, Central Asia shows a remarkable relation to the origin of Thermopsideae in temporal-spatial dimensions, in other words, evolution of Thermopsideae in Central Asia, was coupling with the multiple stages of uplift of the QTP since Cenozoic, which significantly affected the paleoenvironment, paleogeography, and paleoclimate in QTP and Central Asia. To explain evolutionary process of Thermopylae using uplift is reasonable. For instance, divergence of *Piptanthus* and other taxa in the QTP and North China (node 11 in Figure 2) is dated to ca. 4 Ma, during the third phase (intensive northern QTP uplift ca. 3.6 Ma). *Piptanthus* is estimated to be young, with a diversification age of ca. 0.85 Ma, which falls into the period of QTP maximum iceosphere (cryosphere) during the third phase of uplift [66, 68]. Many people have discussed the East Asia-North America disjunction in regard to *Thermopsis* [14, 28, 51, 52, 69]. Yuan et al. [69–71] and Peng and Yuan [51] considered *Thermopsis* section *Archithemopsis,* C. J. Chen, R. Sa, and P. C. Li, to be occurring in Sino-Japanese

regions, as the primitive group in the genus. Sa et al. [28] concluded that *Thermopsis* originated from the Sino-Japanese flora and then dispersed to North America via the Bering Strait. However, our analysis shows the ancestor of North American species to have come from Central Asia (node 8 in Figure 2) rather than East Asia and that the East Asia clade also on the whole evolved in Central Asia, which is different from previous hypotheses. In addition, a new genus *Vuralia* is erected from Thermopsis and related to North American node in the ITS phylogenetic tree [15], whereas *Thermopsis chinensis* and *Th. fabacea* (Figure 2) are also included in North American node; therefore, these three species in Eurasia can be regarded as the dispersals from North America and recent event in Pliocene-Pleistocene from North America crown root 5 Ma, as shown in Figures 2 and 3.

Meanwhile, our estimated diversification age of the Eurasia (Central Asia)-North America disjunction within *Thermopsis* is ca. 20.32 Ma (node 8 in Figure 2). This early Miocene time is similar to that of most other genera showing East Asia-North America disjunctions, for example, *Cercis* ca. 15.41 Ma, *Torreya* ca. 16.7 Ma [72], *Cornus* 13.1 Ma [73], *Calycanthus* 16 Ma [74], *Epimedium-Vancouveria* 9.7 Ma [75], and *Hamamelis* 7.7-7.1 Ma [76], but is different from *Kelloggia* with 5.42 Ma [77] and *Phryma* 5.23-3.68 [78]. As mentioned above, another diversification between Eurasia-North America disjunction which originates from North America at ca. 5 Ma Pliocene-Pleistocene (node 9, Figure 2) resembles *Kelloggia* and *Phryma.*

From Central Asia, a western dispersal route via the Caucasus arriving at the Mediterranean, is illustrated with *Anagyris* in Figures 2 and 3. Our dating of *Anagyris* is 3.08 Ma, which is different from the suggested age of 8.2 ± 4.5 Ma [12]. The *Anagyris* estimate of Ortega-Olivencia and Catalán [12] probably lacks denser sampling from Thermopsideae, since only 5-6 species were selected in total. In general, sufficient samples are necessary in dating. *Vuvalia,* another Mediterranean genus that belongs to the North America clade (node 9 in Figure 2), as well as *Thermopsis chinensis* and *Th. fabacea,* probably dispersed from North America in Pliocene-Pleistocene, since our estimated crown age of North America clade node 9 is about 5 Ma; see Figures 2 and 3.

4.4. Origin and Geographic Diversification of Ammopiptanthus. In view of the unique evergreen broadleaf habit of *Ammopiptanthus* in the desert region of northwestern China and Kyrgyzstan, many people have speculated that *Ammopiptanthus* is a relict of the evergreen broadleaf forest of this region from the Tertiary period [14, 26–29, 79, 80]. The two species form an obvious disjunction pattern, *A. mongolicus* distributed in western Inner Mongolia and the south Gobi desert and *A. nanus* in the western Tianshan Mts. restricted to the borders between China and Kyrgyzstan [29]. Both species of *Ammopiptanthus* are diploid [81]. Genetic diversity from ISSR analysis [29] indicated that differentiation of the two species was significant. In view of its high genetic diversity, a vicariance possibly resulted from the fragmentation of the ancestor range.

From the present analysis (Figure 2), *Ammopiptanthus* is shown to be directly derived from the common ancestor of Thermopsideae, and its divergence time is estimated at ca. 28.81 Ma of middle Oligocene. As mentioned above, during this period, the climate was cooling and increasing in aridity and the vegetation was broadleaved evergreen and deciduous woodland. Therefore, *Ammopiptanthus* spatiotemporally should be speculated to be a relict survivor of the evergreen broadleaf forest at the Tertiary Oligocene.

Much evidence indicates that CO_2 decline promoted the origin of C_4 photosynthesis in grasses in the middle Oligocene ca. 30 Ma [82–84]. Similar to C_4 grass plants, emergence of *Ammopiptanthus* just falls into this period, and since it favors cold and arid climates in an arid region, it can be regarded as a plant case of response to CO_2 decline.

The vicariance and fragmentation of the two *Ammopiptanthus* species dated to ca. 3.88 lead us to link these events to the time label 3.6 Ma of QTP intense uplift and consequently to Asian interior land aridification [66, 68, 85]. This is the same as the speciation of some species in *Caragana* [86] and *Phyllolobium* [87]. The vicariance and fragmentation of the two *Ammopiptanthus* species also likely corresponded to the low pCO_2 and cold and arid climate that resulted from glacial intensification at late Pliocene times (~3.3 to 2.4 Ma) (another time label is at middle Miocene ~14 to 10 Ma) [88].

4.5. Attribute of Madrean-Tethyan Disjunction of Thermopsideae. The Madrean-Tethyan disjunction was proposed by Axelrod [89], as reviewed recently by Wen and Ickert-Bond [90], which hypothesized a nearly continuous belt of Madrean-Tethyan dry broadleaf evergreen sclerophyllous vegetation stretching from western North America to Central Asia in the early Tertiary (from Eocene to late Oligocene) at low latitudes. The former Madrean-Tethyan belt, in fact, falls into the succulent biome locating on the two sides of the Tethys in the Tertiary flora and vegetation, one of four biomes of the legume distribution pattern [21]. The representatives of taxa forming a distinctive disjunction in Thermopsideae are *Thermopsis* section *Thermia* and *Baptisia* in North America and *Thermopsis* sections excluding section *Thermia* and *Piptanthus* in Eurasia, mainly in the QTP and its adjacent regions and *Ammopiptanthus* in Central Asia and *Anagyris* in the Mediterranean. Clearly, Thermopsideae presents a Madrean-Tethyan disjunction. From the temporal dimension, our dating of Thermopsideae to Oligocene ca. 28.81 Ma, is just consistent with Axelrod [89] time range of early Tertiary Oligocene. This is different from the ages of origin of Madrean-Tethyan disjunctions (see review of Wen and Ickert-Bond [90]), for instance, *Platanus orientalis-P. racemosa* s.l. (Platanaceae) ca. 20.5–21.9 Ma [91]; *Juniperus* was at 43.66 Ma [92]. Since Thermopsideae is illustrated to be derived from Central Asia, the evolutionary pattern of this tribe would be migration from Eurasia to North America via the Bering Strait. Moreover, types of Thermopsideae, except for members of *Thermopsis* and *Baptisia,* are perennial herbs producing rhizomes. The rest of these taxa, especially *Ammopiptanthus*, are shrubby and likely to be relicts of the Tertiary dry broadleaf evergreen sclerophyllous vegetation [89]. This, in fact, provides a relict status of dry broadleaf

evergreen sclerophyllous vegetation of Madrean-Tethyan disjunction. Therefore, from perspectives of distribution, dated age, and vegetation of Thermopsideae, it fits as a good case of Madrean-Tethyan disjunction.

4.6. African Origin and Dispersal of Core Genistoides. The so-called core genistoides are defined on the basis of molecular phylogeny [1, 3–5, 16–19]. This clade has four tribes taxonomically [19], namely, Crotalarieae, Podalyrieae, Genisteae, and Thermopsideae. Except for Thermopsideae, the tribes occur mainly in Africa, and only a few species expand to the Mediterranean, southern Europe, the Middle East, the Caucasus, and Russia [6–8].

Schrire et al. [21] stated that the derived genistoides, including Crotalarieae, Podalyrieae, and Genisteae, have their basal branching elements in warm temperate southern Africa, an ancestral crown in the southern warm temperate biome [22]. From there, they would have migrated northwards through montane tropical Africa to the Mediterranean and Macaronesian regions and sequentially to the New World, or have secondarily invaded the tropics.

Our molecular dating and S-DIVA results (Figures 2 and 3) indicate that the core genistoides originated from Africa, probably warm temperate southern Africa as mentioned above [21], from Eocene to Oligocene ca. 54.43–33.47 Ma (Nodes 1,3). The four tribes not only dispersed to the Mediterranean, West Asia, the Caucasus, northwestern Russia, Central Asia, East Asia, and North America, but also continuously diversified in Africa in situ until middle Miocene ca. 12 Ma, which is fundamentally due to the diversification of the three tribes Crotalarieae, Podalyrieae, and Genisteae in that continent (Figure 2). The exact place of origin of these three tribes will probably become less ambiguous due to discovery of fossil records and increased taxon sampling and sequencing and so forth. However, an African origin is affirmed; furthermore, the biome warm temperate southern Africa, sensu Schrire et al. [21], is accepted here. This also illuminates the origin of Thermopsideae.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Thanks are to Dr. Yong-Ming Yuan and Dr. Heng-Chang Wang for their valuable comment and suggestion to the manuscript, to Dr. Rodrigo Duno de Stefano and anonymous referee for their critical, constructive, and helpful comment and suggestions, and to Professor Peter F. Stadler for his careful corrections to manuscript. This study was financially supported by National Natural Science Foundation of China (no. 41271070), China National Key Basic Research Program (2014CB54201) and Xinjiang Institute of Ecology and Geography, CAS.

References

[1] M. D. Crisp, S. Gilmore, and B. Van Wyk, "Molecular phylogeny of the genistoid tribes of papilionoid leguminosae," in *Advances in Legume Systematics, Part 9*, P. S. Herendeen and A. Bruneau, Eds., pp. 249–276, Royal Botanic Gardens, Kew, Richmond, UK, 2000.

[2] R. T. Pennington, M. Lavin, H. Ireland, B. Klitgaard, J. Preston, and J.-M. Hu, "Phylogenetic relationships of basal papilionoid legumes based upon sequences of the chloroplast *trnL* intron," *Systematic Botany*, vol. 26, no. 3, pp. 537–556, 2001.

[3] A. Ainouche, R. J. Bayer, P. Cubas, and M. T. Misset, "Phylogenetic relationships within tribe Genisteae (Papilionoideae) with special reference to genus *Ulex*," in *Advances in Legume Systematics. Part 10: Higher Level Systematics*, B. B. Klitgaard and A. Bruneau, Eds., pp. 239–252, Royal Botanic Gardens, Kew, UK, 2003.

[4] M. F. Wojciechowski, "Reconstructing the phylogeny of legumes (Leguminosae): an early 21st century perspective," in *Advances in Legume Systematics, Part 10: Higher Level Systematics*, B. B. Klitgaard and A. Bruneau, Eds., pp. 5–35, Royal Botanic Garden, Kew, Richmond, UK, 2003.

[5] M. F. Wojciechowski, M. Lavin, and M. J. Sanderson, "A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family," *The American Journal of Botany*, vol. 91, no. 11, pp. 1846–1862, 2004.

[6] J. M. Lock, "Thermopsideae," in *Legumes of the World*, G. Lewis, B. Schrire, B. Mackinder, and M. Lock, Eds., pp. 263–265, Royal Botanic Gardens, Surrey, UK, 2005.

[7] R. M. Polhill and B. E. van Wyk, "Genisteae," in *Legumes of the World*, G. Lewis, B. Schrire, B. Mackinder, and M. Lock, Eds., pp. 283–297, Royal Botanic Gardens, Kew, UK, 2005.

[8] B. E. Van Wyk, "Podalyrieae, Crotalarieae," in *Legumes of the World*, G. Lewis, B. Schrire, B. Mackinder, and M. Lock, Eds., pp. 267–282, Royal Botanic Gardens, Surrey, UK, 2005.

[9] B. L. Turner, "Thermopsideae," in *Advances in Legume Systematics*, R. M. Polhill and P. H. Raven, Eds., vol. 1, pp. 403–407, Royal Botanic Gardens, Kew, UK, 1981.

[10] M. F. Wojciechowski, "The origin and phylogenetic relationships of the Californian chaparral "paleoendemic" Pickeringia (Leguminosae)," *Systematic Botany*, vol. 38, no. 1, pp. 132–142, 2013.

[11] K. Browicz, "Geographic distribution of some shrubsfrom the family Leguminosae in southwestern Asia," in *Arboretum Kornickie, Rocznik XXXII*, pp. 5–30, 1978.

[12] A. Ortega-Olivencia and P. Catalán, "Systematics and evolutionary history of the circum-Mediterranean genus *Anagyris* L. (Fabaceae) based on morphological and molecular data," *Taxon*, vol. 58, no. 4, pp. 1290–1306, 2009.

[13] C. Y. Wu and S. G. Wu, *A Proposal for a New Floristic Kingdom (Realm): The E. Asiatic Kingdom, Its Delineation and Character*, CHEP & Springer, Beijing, China, 1999.

[14] H. C. Wang, H. Sun, J. A. Compton, and J. B. Yang, "A phylogeny of Thermopsideae (Leguminosae: Papilionoideae) inferred from nuclear ribosomal internal transcribed spacer (ITS) sequences," *Botanical Journal of the Linnean Society*, vol. 151, no. 3, pp. 365–373, 2006.

[15] T. Uysal, K. Ertuğrul, and M. Bozkurt, "A new genus segregated from *Thermopsis* (Fabaceae: Papilionoideae): *Vuralia*," *Plant Systematics and Evolution*, vol. 300, no. 7, pp. 1627–1637, 2014.

[16] T. Kajita, H. Ohashi, Y. Tateishi, C. D. Bailey, and J. J. Doyle, "rbcL and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies," *Systematic Botany*, vol. 26, no. 3, pp. 515–536, 2001.

[17] A. Bruneau, M. Mercure, G. P. Lewis, and P. S. Herendeen, "Phylogenetic patterns and diversification in the caesalpinioid legumes," *Botany*, vol. 86, no. 7, pp. 697–718, 2008.

[18] M. A. Bello, A. Bruneau, F. Forest, and J. A. Hawkins, "Elusive relationships within order fabales: phylogenetic analyses using matK and rbcl sequence data," *Systematic Botany*, vol. 34, no. 1, pp. 102–114, 2009.

[19] G. Lewis, B. Schrire, B. Mackinder, and M. Lock, *Legumes of the World*, Royal Botanic Gardens, Kew, Richmond, UK, 2005.

[20] P. S. Herendeen, W. L. Crepet, and D. L. Dilche, "The fossil history of the Leguminosae: phylogenetic and biogeographic implications," in *Advances in Legume Systematics, Part 4*, P. S. Herendeen and D. L. Dilcher, Eds., pp. 303–316, Royal BotanicGardens, Kew, Richmond, UK, 1992.

[21] B. D. Schrire, G. P. Lewis, and M. Lavin, "Biogeography of the Leguminosae," in *Legumes of the World*, G. P. Lewis, B. D. Schrire, B. MacKinder, and M. Lock, Eds., pp. 21–54, Kew Publishing, 2005.

[22] M. Lavin, P. S. Herendeen, and M. F. Wojciechowski, "Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the Tertiary," *Systematic Biology*, vol. 54, no. 4, pp. 575–594, 2005.

[23] P. H. Raven and D. I. Axelrod, "Angiosperm biogeography and past continental movements," *Annals of the Missouri Botanical Garden*, vol. 61, no. 3, pp. 539–673, 1974.

[24] R. M. Polhill and P. H. Raven, "Papilionoideae," in *Advances in Legume Systematics, Part 1*, R. M. Polhill and P. H. Raven, Eds., pp. 191–208, Royal Botanic Gardens, Kew, UK, 1981.

[25] L. Xie and Y. Yang, "Miocene origin of the characteristic broad-leaved evergreen shrub *Ammopiptanthus* (leguminosae) in the desert flora of eastern central Asia," *International Journal of Plant Sciences*, vol. 173, no. 8, pp. 944–955, 2012.

[26] P. C. Li and Z. C. Ni, "The formation and evolution of Fabaceae in Xizang," *Acta Phytotaxonomica Sinica*, vol. 20, pp. 142–154, 1982.

[27] C. Y. Wu and H. S. Wang, *Plant Geography*, Science Press, Beijing, China, 1983, (Chinese).

[28] R. Sa, J. C. Chen, and P. C. Li, "The phytogeographical studies of *Thermopsis* (Leguminosae)," *Acta Phytotaxonomica Sinica*, vol. 38, pp. 148–166, 2000.

[29] X.-J. Ge, Y. Yu, Y.-M. Yuan, H.-W. Huang, and C. Yan, "Genetic diversity and geographic differentiation in endangered *Ammopiptanthus* (Leguminosae) populations in desert regions of northwest China as revealed by ISSR analysis," *Annals of Botany*, vol. 95, no. 5, pp. 843–851, 2005.

[30] J. Doyle and J. L. Doyle, "Genomic plant DNA preparation from fresh tissue—CTAB method," *Phytochemical Bulletin*, vol. 19, pp. 11–15, 1987.

[31] Y. Kang, M.-L. Zhang, and Z.-D. Chen, "A preliminary phylogenetic study of the subgenus Pogonophace (Astragalus) in China based on ITS sequence Data," *Acta Botanica Sinica*, vol. 45, no. 2, pp. 140–145, 2003.

[32] P. Taberlet, L. Gielly, G. Pautou, and J. Bouvet, "Universal primers for amplification of three non-coding regions of chloroplast DNA," *Plant Molecular Biology*, vol. 17, no. 5, pp. 1105–1109, 1991.

[33] T. Sang, D. J. Crawford, and T. F. Stuessy, "Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae)," *The American Journal of Botany*, vol. 84, no. 8, pp. 1120–1136, 1997.

[34] J. A. Tate and B. B. Simpson, "Paraphyly of Tarasa (Malvaceae) and diverse origins of the polyploid species," *Systematic Botany*, vol. 28, no. 4, pp. 723–737, 2003.

[35] B. Oxelman, M. Lidén, and D. Berglund, "Chloroplast rps16 intron phylogeny of the tribe *Sileneae (Caryophyllaceae)*," *Plant Systematics and Evolution*, vol. 206, no. 1–4, pp. 393–410, 1997.

[36] L. A. Johnson and D. E. Soltis, "Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (*Polemoniaceae*) using *mat*K sequences," *Annals of the Missouri Botanical Garden*, vol. 82, no. 2, pp. 149–175, 1995.

[37] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Research*, vol. 25, pp. 4876–4882, 1997.

[38] T. A. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," *Nucleic Acids Symposium Series*, vol. 41, pp. 95–98, 1999.

[39] J. S. Farris, M. Källersjö, A. G. Kluge, and C. Bult, "Testing significance of incongruence," *Cladistics*, vol. 10, pp. 315–319, 1994.

[40] D. L. Swofford, *'PAUP*,' Phylogenetic Analysis Using Parsimony, Version 4.0b10*, Sinauer, Sunderland, Mass, USA, 2002.

[41] D. Posada and K. A. Crandall, "Modeltest: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[42] J. R. Tao, Z. K. Zhou, and Y. S. Liu, *The Evolution of the Late Cretaceous-Cenozonic Floras in China*, Science Press, Beijing, China, 2000 (Chinese).

[43] Institute of Botany, Nanjing Institute of Geology and Palaeontology (IB & NIGP), and Academia Sinica, *Chinese Plant Fossils, Volume 3: Neogene Floras*, Academic Press, Beijing, China, 1978.

[44] S. S. Renner, "Relaxed molecular clocks for dating historical plant dispersal events," *Trends in Plant Science*, vol. 10, no. 11, pp. 550–558, 2005.

[45] J. J. Welch and L. Bromham, "Molecular dating when rates vary," *Trends in Ecology and Evolution*, vol. 20, no. 6, pp. 320–327, 2005.

[46] A. J. Drummond and A. Rambaut, "BEAST: bayesian evolutionary analysis by sampling trees," *BMC Evolutionary Biology*, vol. 7, no. 1, article 214, 2007.

[47] F. Ronquist, "Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography," *Systematic Biology*, vol. 46, no. 1, pp. 195–203, 1997.

[48] F. Ronquist, *DIVA Version 1.1*, (ftp.uu.se orftp.systbot.uu.se), Uppsala Univeristy, Uppsala, Sweden, 1996.

[49] J. A. A. Nylander, U. Olsson, P. Alström, and I. Sanmartín, "Accounting for phylogenetic uncertainty in biogeography: a Bayesian approach to dispersal-vicariance analysis of the thrushes (Aves: *Turdus*)," *Systematic Biology*, vol. 57, no. 2, pp. 257–268, 2008.

[50] Y. Yu, A. J. Harris, and X. He, "S-DIVA (Statistical Dispersal-Vicariance Analysis): a tool for inferring biogeographic histories," *Molecular Phylogenetics and Evolution*, vol. 56, no. 2, pp. 848–850, 2010.

[51] Z. X. Peng and Y. M. Yuan, "Systematic revision on *Thermopsideae* (Leguminosae) of China," *Acta Botanica Boreali-Occidentalia Sinica*, vol. 12, pp. 158–166, 1992.

[52] C. J. Chen, M. G. Mendenhall, and B. L. Turner, "Taxonomy of *Thermopsis* (Fabaceae) in North America," *Annals of the Missouri Botanical Garden*, vol. 81, no. 4, pp. 714–742, 1994.

[53] S. H. Cheng, "*Ammopiptanthus* Cheng f., a new genus of Leguminosae from Central Asia," *Botanicheskii Zhurnal*, vol. 44, pp. 1381–1386, 1959.

[54] K. J. Willis and J. C. McElwain, *The Evolution of Plants*, Oxford University Press, New York, NY, USA, 2002.

[55] G. Ramstein, F. Fluteau, J. Besse, and S. Joussaume, "Effect of orogeny, plate motion and land-sea distribution on Eurasian climate change over the past 30 million years," *Nature*, vol. 386, no. 6627, pp. 788–795, 1997.

[56] T. Hrbek and A. Meyer, "Closing of the Tethys Sea and the phylogeny of Eurasian killifishes (Cyprinodontiformes: Cyprinodontidae)," *Journal of Evolutionary Biology*, vol. 16, no. 1, pp. 17–36, 2003.

[57] V. A. Zubakov and I. I. Borzenkova, *Global Palaeoclimate of the Late Cenozoic*, vol. 12 of *Developments in Palaeontology and Stratigraphy*, Elsevier, Amsterdam, The Netherlands, 1990.

[58] J. R. Tao, "The Tertiary vegetation and flora and floristic regions in China," *Acta Phytotaxomica Sinica*, vol. 31, pp. 25–43, 1992.

[59] Z. T. Guo, B. Sun, Z. S. Zhang et al., "A major reorganization of Asian climate by the early Miocene," *Climate of the Past*, vol. 4, no. 3, pp. 153–174, 2008.

[60] G. Dupont-Nivet, C. Hoom, and M. Konert, "Tibetan uplift prior to the Eocene-Oligocene climate transition: evidence from pollen analysis of the Xining Basin," *Geology*, vol. 36, no. 12, pp. 987–990, 2008.

[61] Z. T. Guo, W. F. Ruddiman, Q. Z. Hao et al., "Onset of Asian desertification by 22 Myr ago inferred from loess deposits in China," *Nature*, vol. 416, no. 6877, pp. 159–163, 2002.

[62] J. L. Pei, Z. M. Sun, X. S. Wang et al., "Evidence for Tibetan Plateau uplift in Qaidam basin before Eocene-Oligocene boundary and its climatic implications," *Journal of Earth Science*, vol. 20, no. 2, pp. 430–437, 2009.

[63] Z. S. An, J. E. Kutzbach, W. L. Prell, and S. C. Porter, "Evolution of Asian monsoons and phased uplift of the Himalaya-Tibetan plateau since Late Miocene times," *Nature*, vol. 411, no. 6833, pp. 62–66, 2001.

[64] X. M. Wang, B. Y. Wang, Z. X. Qiu et al., "Danghe area (western Gansu, China) biostratigraphy and implications for depositional history and tectonics of northern Tibetan Plateau," *Earth and Planetary Science Letters*, vol. 208, no. 3-4, pp. 253–269, 2003.

[65] G. J. Li, T. Pettke, and J. Chen, "Increasing Nd isotopic ratio of Asian dust indicates progressive uplift of the north Tibetan Plateau since the middle Miocene," *Geology*, vol. 39, no. 3, pp. 199–202, 2011.

[66] J. J. Li and X. M. Fang, "Research on the uplift of the Qinghai-Xizang Plateau and environmental changes," *Chinese Science Bulletin*, vol. 43, pp. 1569–1574, 1998.

[67] X. Fang, W. Zhang, Q. Meng et al., "High-resolution magnetostratigraphy of the Neogene Huaitoutala section in the eastern Qaidam Basin on the NE Tibetan Plateau, Qinghai Province, China and its implication on tectonic uplift of the NE Tibetan Plateau," *Earth and Planetary Science Letters*, vol. 258, no. 1-2, pp. 293–306, 2007.

[68] Y. F. Shi, M. C. Tang, and Y. Z. Ma, "The relation of second rising in Qinghai-Xizang Plateau and Asia Monsoon," *Sciences in China D*, vol. 28, pp. 263–271, 1998.

[69] Y. M. Yuan and C. J. Chen, "Anatomical evidence for phylogeny of the tribe Thermopsideae (Leguminosae)," *Journal of Lanzhou University (Natural Sciences)*, vol. 29, pp. 97–104, 1993.

[70] Y. M. Yuan, Z. X. Peng, and C. J. Chen, "The systematical and ecological significance of anatomical characters of leaves in the tribe Thermopsideae (Leguminosae)," *Acta Botanica Sinica*, vol. 33, pp. 840–847, 1991.

[71] Y. M. Yuan and Z. X. Peng, "Pollen morphology and its systematic significance of the tribe Thermopsideae (Leguminosae) from China," *Acta Botanica Boreali-Occidentalia Sinica*, vol. 27, pp. 84–95, 1991.

[72] M. J. Donoghue, C. D. Bell, and J. Li, "Phylogenetic patterns in Northern Hemisphere plant geography," *International Journal of Plant Sciences*, vol. 162, pp. S41–S52, 2001.

[73] Q. Y. Xiang, D. E. Soltis, and P. S. Soltis, "Phylogenetic relationships of Cornaceae and close relatives inferred from *matK* and *rbcL* sequences," *American Journal of Botany*, vol. 85, no. 2, pp. 285–297, 1998.

[74] S. L. Zhou, S. S. Renner, and J. Wen, "Molecular phylogeny and intra- and intercontinental biogeography of Calycanthaceae," *Molecular Phylogenetics and Evolution*, vol. 39, no. 1, pp. 1–15, 2006.

[75] M.-L. Zhang, C. H. Uhink, and J. W. Kadereit, "Phylogeny and biogeography of *Epimedium/Vancouveria* (Berberidaceae): Western North American—East Asian disjunctions, the origin of European mountain plant taxa, and East Asian species diversity," *Systematic Botany*, vol. 32, no. 1, pp. 81–92, 2007.

[76] L. Xie, T.-S. Yi, R. Li, D. Z. Li, and J. Wen, "Evolution and biogeographic diversification of the witch-hazel genus (*Hamamelis* L., Hamamelidaceae) in the Northern Hemisphere," *Molecular Phylogenetics and Evolution*, vol. 56, no. 2, pp. 675–689, 2010.

[77] Z.-L. Nie, J. Wen, H. Sun, and B. Bartholomew, "Monophyly of *Kelloggia* Torrey ex Benth. (Rubiaceae) and evolution of its intercontinental disjunction between western North America and eastern Asia," *American Journal of Botany*, vol. 92, no. 4, pp. 642–652, 2005.

[78] Z.-L. Nie, H. Sun, P. M. Beardsley, R. G. Olmstead, and J. Wen, "Evolution of biogeographic disjunction between eastern Asia and eastern North America in Phryma (Phrymaceae)," *The American Journal of Botany*, vol. 93, no. 9, pp. 1343–1356, 2006.

[79] J. Q. Liu, M. X. Qiu, K. Yang, and Q. H. Shi, "Studies on the plant community of *Ammopiptanthus mongolicus*," *Journal of Desert Research*, vol. 15, pp. 109–115, 1995.

[80] B. Y. Geng, J. R. Tao, and G. P. Xie, "Early Tertiary fossil plants and paleoclimate of Lanzhou Basin," *Acta Phytaxonomica Sinica*, vol. 39, no. 2, pp. 105–115, 2001.

[81] B. R. Pan and S. P. Huang, "Cytological study of the genus *Ammopipthanthus*," *Acta Botanica Sinica*, vol. 35, pp. 314–317, 1993.

[82] R. F. Sage, "The evolution of C_4 photosynthesis," *New Phytologist*, vol. 161, no. 2, pp. 341–370, 2004.

[83] P.-A. Christin, G. Besnard, E. Samaritani et al., "Oligocene CO_2 decline promoted C_4 photosynthesis in grasses," *Current Biology*, vol. 18, no. 1, pp. 37–43, 2008.

[84] A. Vicentini, J. C. Barber, S. S. Aliscioni, L. M. Giussani, and E. A. Kellogg, "The age of the grasses and clusters of origins of C_4 photosynthesis," *Global Change Biology*, vol. 14, no. 12, pp. 2963–2977, 2008.

[85] Y. F. Shi, J. J. Li, B. Y. Li et al., "Uplift of the Qinghai-Xizang (Tibetan) Plateau and east asia environmental change during late Cenozoic," *Acta Geographica Sinica*, vol. 54, no. 1, pp. 10–21, 1999.

[86] M.-L. Zhang and P. W. Fritsch, "Evolutionary response of *Caragana* (Fabaceae) to Qinghai-Tibetan Plateau uplift and Asian interior aridification," *Plant Systematics and Evolution*, vol. 288, no. 3-4, pp. 191–199, 2010.

[87] M.-L. Zhang, Y. Kang, Y. Zhong, and S. C. Sanderson, "Intense uplift of the Qinghai-Tibetan Plateau triggered rapid diversification of *Phyllolobium* (Leguminosae) in the Late Cenozoic," *Plant Ecology and Diversity*, vol. 5, no. 4, pp. 491–499, 2012.

[88] A. K. Tripati, C. D. Roberts, and R. A. Eagle, "Coupling of CO_2 and Ice sheet stability over major climate transitions of the last 20 million years," *Science*, vol. 326, no. 5958, pp. 1394–1397, 2009.

[89] D. I. Axelrod, "Evolution and biogeography of Madrean-Tethyan sclerophyll vegetation," *Annals of the Missouri Botanical Garden*, vol. 62, no. 2, pp. 280–334, 1975.

[90] J. Wen and S. M. Ickert-Bond, "Evolution of the Madrean-Tethyan disjunctions and the North and South American amphitropical disjunctions in plants," *Journal of Systematics and Evolution*, vol. 47, no. 5, pp. 331–348, 2009.

[91] Y. Feng, S.-H. Oh, and P. S. Manos, "Phylogeny and historical biogeography of the genus *Platanus* as inferred from nuclear and chloroplast DNA," *Systematic Botany*, vol. 30, no. 4, pp. 786–799, 2005.

[92] K. S. Mao, G. Hao, J. Q. Liu, R. P. Adams, and R. I. Milne, "Diversification and biogeography of *Juniperus* (Cupressaceae): variable diversification rates and multiple intercontinental dispersals," *New Phytologist*, vol. 188, no. 1, pp. 254–272, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 675260, 25 pages
http://dx.doi.org/10.1155/2015/675260

Research Article

Phylogeography of *Pteronotropis signipinnis*, *P. euryzonus*, and the *P. hypselopterus* Complex (Teleostei: Cypriniformes), with Comments on Diversity and History of the Gulf and Atlantic Coastal Streams

Richard L. Mayden[1] and Jason Allen[2]

[1]*Department of Biology, Saint Louis University, 3507 Laclede Avenue, St. Louis, MO 63103, USA*
[2]*Department of Biology, Saint Louis Community College-Meramec, 11333 Big Bend Road, St. Louis, MO 63122, USA*

Correspondence should be addressed to Richard L. Mayden; cypriniformes@gmail.com

Received 11 July 2014; Revised 4 November 2014; Accepted 17 December 2014

Academic Editor: William H. Piel

The cyprinid genus *Pteronotropis* is endemic to southeastern Gulf of Mexico and Atlantic Ocean of North America. Never before has the genus been demonstrated to be monophyletic. We investigate both the phylogenetic relationships and the phylogeography of some species in the genus using mitochondrial ND2 sequences. In no analysis is the genus resolved as monophyletic if *Notropis harperi* is not included in the genus. Biogeographic and phylogeographic evaluations are conducted with *Pteronotropis*, including *P. signipinnis*, *P. euryzonus*, and the *P. hypselopterus* complex. Patterns of relationships and population genetic analyses support divergences within multiple clades both at the species level and within species that are tied to abiotic changes in the region. Replicated patterns across clades are observed, as well as patterns previously found in other taxa. Pteronotropis hypselopterus is likely not a natural grouping as populations from some drainages form clades more closely related to other species of the genus. The general patterns of relationships indicate likely cryptic species not currently recognized. Finally, the patterns of species relationships and clades and population structuring within species serve as another example of replicated divergences in the biodiversity east and west of the Mobile Bay.

1. Introduction

Avise [1] defines phylogeography as "...a field of study concerned with the principles and processes governing the geographic distributions of genealogical lineages, especially those within and among closely related species." Phylogeography is a subdiscipline of historical biogeography, which seeks to find historical explanations for the present distribution of organisms [2]. Within this framework, two competing hypotheses have existed for over a century to explain how species and their populations came to occupy a geographic area or aquatic system, dispersal and vicariance. Dispersalists favor the hypothesis that the present distributions of organisms are explained by movement of populations; closely related taxa separated by some type of barrier significant to them diverged once some populations were successful in overcoming this barrier and were isolated long enough to diverge from their sister group. Vicariant biogeographers seek to explain the distribution of related taxa by hypothesizing that part of the geographic range of an ancestral species became fragmented by some barrier, isolating some populations that later diverged. The hypothesis of dispersal in the past is one that is largely impossible to test. The theory behind vicariance biogeography stipulates that vicariant patterns should be used as a first-order explanation for the distribution of organisms and only if this hypothesis is rejected should dispersal be invoked. Thus, vicariance biogeography does not stipulate that dispersal does not occur. Further, this model of divergence maintains that if a variety of taxa show concordant patterns around the same barrier then the vicariant event

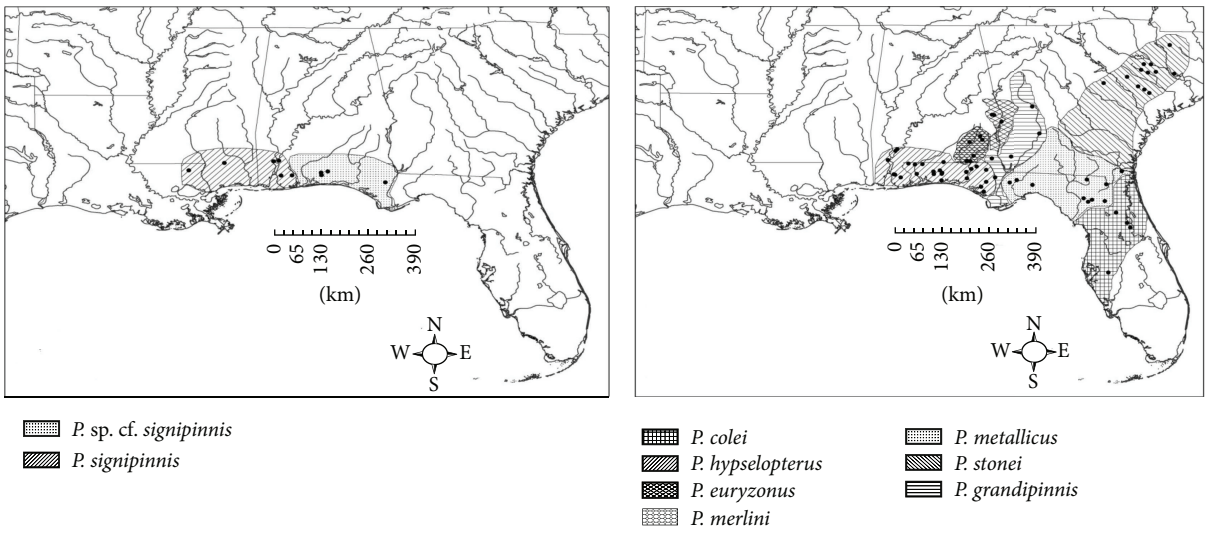


Figure 1: Distributions and sampling localities for *Pteronotropis signipinnis, P. euryzonus*, and the *P. hypselopterus* complex.

underlying the diversification event across lineages is the most parsimonious explanation. The confidence that we have in these concordant patterns is really a function of how often empirical evidence corroborates their occurrence. Fishes from rivers and streams of the southeastern United States along the Gulf and Atlantic slopes have been the attention of varied phylogeographic studies, including those by Wiley and Mayden [3], Swift et al. [4], Nagle and Simons [5], and Sandel [6] as well as studies referenced therein.

The cyprinid genus *Pteronotropis* is largely endemic to aquatic habitats of the southeastern United States and contains species with small and large distributions across this large geographic area (Figure 1). The genus was reviewed by Suttkus and Mettee [7] and Mayden and Allen [8]. Not until the latter was the genus corroborated as a monophyletic with two major lineages [8]. The distributions of species of *Pteronotropis*, combined with the many rivers systems that they occupy and independently enter the Gulf of Mexico or Atlantic Ocean, make species of the genus of particular interest for both systematic and biogeographic evaluations. The geological processes and timing of both geological and hydrological events across this region are relatively well known, the lower reaches of rivers have been inundated by fluctuating sea levels, and multiple species inhabit coastal areas [4]. The species diversity of *Pteronotropis*, distributions of species, and the history of the area provide for an attractive model to investigate their historical biogeography and compare to other species inhabiting this same area. Herein, we examine the evolutionary and biogeographic history of *Pteronotropis* using both sequences of one of the most appropriate genes for the scale of diversity being examined, ND2. We use this genetic information and appropriate statistical tests to explore hypotheses of the evolutionary history of targeted species.

Statistical parsimony, although technically a phenetic method (based upon overall similarity and not shared derived characters as in phylogenetics), can offer valuable insight into relationships of closely related individuals (i.e., populations within a single species) where the resolution of true parsimony may fail. Functionally, the method converts the haplotype tree of each species into a series of networks showing differences to the level of single mutational events [9]. This technique is widely accepted as a means of exploring phylogeographic relationships and has been used in many studies of Nearctic and Palearctic fishes [10, 11]. Mismatch distribution analysis (MDA) essentially takes the distribution of pairwise differences (mismatch distributions) calculated earlier and plots them against the number of individuals in the analysis [12]. The advantage of this method is its ability to distinguish between rapid range expansion and expansion through recurrent gene flow. If rapid range expansion has occurred in the history of a species, the expectation is that the plot will show a unimodal distribution of pairwise differences; a multimodal distribution would indicate population stability [13]. This technique was used in many demographic studies, including those of human [14] and fish populations [15]. Mitochondrial genes are very useful as phylogeographic markers because of their generally fast rate of anagenesis, uniparental mode of inheritance, and lack of recombination, all allowing researchers to detect events occurring at the population level and track events over geographical ranges and through evolutionary time [1]. In particular, ND2 is an effective marker in distinguishing relationships at many levels [16]. Two nuclear genes, the third exon of RAG 1 and the first intron of S7, used in Mayden and Allen [8], were tested for use in this analysis but lacked sufficient variation to be informative for the questions being examined herein; likewise, Chen and Mayden [17] and Mayden and Chen [18] examined variation in an additional six nuclear genes, but these genes also provided insufficient variation to test hypotheses of inter- and intraspecific relationships at this level. Finally, the southeast as a whole has served as a paradigm for phylogeographic research with many of its principles and techniques being developed from studies on

taxa endemic to this region. Consequently, many congruent patterns have emerged for numerous species ranging from freshwater fishes, amphibians, reptiles, mammals, macroinvertebrates, plants, and maritime species [1, 3, 19, 20]. Species and species complexes of *Pteronotropis* are also endemic to streams of the southeastern United States. Furthermore, the genus *Pteronotropis* has never been well corroborated as monophyletic. A previous analysis of the group by Suttkus and Mettee [7] developed specific hypotheses as to the diversification and biogeography of the group. Thus, this species complex serves as another excellent candidate group to examine the historical patterns of biodiversity along the Atlantic-Gulf slopes and to test multiple hypotheses as to dispersal, vicariance, speciation, and divergence of populations.

2. Review of the Geography and Diversity of the Region

Swift et al. [4] define the southeastern United States as areas occurring south, southeast, and southwest of the peripheral Tennessee River system, west to Lake Pontchartrain and east to the Savannah River. This includes aquatic systems from the entire states of Alabama, Florida, Georgia, Mississippi, and part of South Carolina. Geologically the southeast can be thought of as three separate areas: (1) a mountainous area consisting of Paleozoic and Mesozoic rocks, (2) a lower relief hilly terrain made up of pre-Cretaceous formations termed the Piedmont, and (3) a sedimentary terrain of very low relief termed the Coastal Plain [21]. The region separating the Piedmont from the Coastal Plain is the "fall line" and often forms the distributional boundary between many upland and lowland species. The southeast has been a model of geographic stability since the middle to late Cretaceous period with no major land movements or continental shifting occurring since this time [22]. This is also thought to be true of the river systems, which traverse this region, with most of the drainage basins being in place since the late Cretaceous to early Eocene. The southeast includes about 32 major to minor river systems. These generally have a pattern of flow to the southeast or southwest, depending on where they discharge, with the larger drainage systems having their headwaters above the fall line. The larger river systems include the Pearl, Pascagoula, Tombigbee (including Black Warrior River), Alabama (including Cahaba, Coosa, and Tallapoosa rivers), Chattahoochee, Altamaha, and Savannah rivers. The remaining rivers are much smaller in their drainage basins, have their headwaters originating below the fall line, and include the Escambia, Choctawhatchee, Ochlocknee, Suwannee, St. Marys, and Satilla Rivers [4]. Rivers originating at or above the fall line tend to have high gradient and a substrate mainly consisting of bedrock, boulders, and gravel. Streams below the fall line are typically lower gradient and have slower flows and substrates largely consisting of clay or sand, with some gravel or bedrock in some areas. These streams typically flow through forested swamps and wetlands, and it is because of this that they get their name "Blackwater streams" and are typically described as being "tannin stained," from the leaching of tannins from decaying vegetation. Even though the land formations of this region have been stable over millions of years, sea level fluctuations have been an impact on the region and proposed to have had a major impact on the biodiversity of this region [4].

It is estimated from fossil record that towards the end of the Oligocene sea levels fell dramatically and exposed a large continental shelf due to the development of polar ice caps and the filling of the Mediterranean Sea [4]. These events were followed by an elevation of sea level to about 100 meters above present levels in the mid to early Miocene. A slight drop in water level occurred again at the end of the Miocene followed by another rise in the early Pliocene, with slight fluctuations during the Pleistocene glaciations. During the high sea stands the smaller rivers below the fall line were likely completely obliterated, leaving obligate freshwater taxa to inhabit the refugia of isolated upstream sections in the larger rivers. As waters receded it is hypothesized that suitable habitat became available in lower reaches of the larger rivers coinciding with the appearance of smaller streams and rivers. Although no drainage connections exist between most of these rivers today, during periods of low sea stands there were coalescent events between many of these streams, enabling taxa to disperse throughout these systems [5, 6]. This history likely explains why many of the smaller rivers below the fall line, such as the Escambia, Blackwater, and Yellow, have no modern-day connections, yet contain most of the same taxa with few, if any, endemics [4].

One of the classical concordant patterns of diversification in this region has been the genetic distinction between Atlantic and Gulf slope taxa, with the Apalachicola drainage system being the range limit or contact zone for taxa on either side. One of the first studies to show this was that of Bermingham and Avise [19], who in their comparison of four freshwater fish species (*Amia calva*, *Lepomis punctatus*, *L. microlophus*, and *L. gulosus*) observed significantly greater genetic differentiation between Atlantic and Gulf slope populations of each species than that observed among haplotypes within each region. These patterns of diversification for freshwater fishes have been confirmed by Swift et al. [4], in their study of the zoogeography of the southeast, using a simple present-absence matrix for almost all the freshwater species and illustrated phenetically that the "oldest" split was between the Gulf Slope streams, up to and including the Apalachicola, and southeastern Atlantic Slope streams, including all of those of Florida.

Another often cited pattern in the region is that associated with the Central Gulf slope speciation hypothesis proposed by Wiley and Mayden [3]. In their work on vicariant patterns in the North American freshwater fish fauna they identified several sister species and populations within a single species that have their distributional limits defined by the Mobile Basin. Some of these taxa include the *Fundulus nottii* species group, *Ammocrypta beani* and *A. bifasciata*, *Etheostoma chlorosomum* and *E. davisoni*, and populations of *Notropis longirostris*. In a study on the *Notropis dorsalis* species group, using the mitochondrial marker cytochrome *b*, Raley and Wood [23] showed that populations of *N. longirostris* on either side of the Mobile Basin were resolved as two separate clades with as much as 8% sequence divergence. Swift et al.

[4] provide a possible mechanism for these occurrences by hypothesizing that the Alabama-Tombigbee river system had more of a westward or southwestward flow pattern in the early Miocene that was diverted directly southward during the middle to late Miocene, thus dividing populations of species on either side of this river system.

The *Pteronotropis hypselopterus* species complex occurs in streams extending across the Atlantic-Gulf slopes (Figure 1). *Pteronotropis hypselopterus* occurs from western tributaries of the Mobile Bay eastward to, but not including the Apalachicola River drainage. *Pteronotropis merlini* is endemic to the Choctawhatchee River system, including the Pea River, at and above the confluence of the east and west forks. Any *Pteronotropis* below this confluence is considered to be *P. hypselopterus* [7]. *Pteronotropis grandipinnis* is endemic to the Apalachicola River system in the lower reaches of the Chattahoochee and Flint rivers. *Pteronotropis euryzonus*, though not part of the *P. hypselopterus* complex, is closely related to this group [8] and is endemic to the middle Chattahoochee River. The two remaining species are endemic to Atlantic slope rivers and include *P. metallicus,* ranging from the Ochlocknee River to the St. Johns and Hillsborough rivers of peninsular Florida, and *P. stonei* with its northern distributional limit in the lower reaches of the Pee Dee River, South Carolina, to as far south as the Satilla River in southern Georgia.

Suttkus and Mettee [7] offered the most current biogeographical account for members of this genus. They contended that *P. euryzonus* evolved below the fall line in the middle Chattahoochee River system, where still endemic, and then spread to the adjacent Choctawhatchee River system through a temporary stream capture. These ancestral populations were then dispersed throughout the Choctawhatchee and Pea river systems, eventually giving rise to *P. merlini* above the confluence of these two rivers and *P. hypselopterus* below, possibly through a vicariant event such as habitat specialization. *Pteronotropis merlini* is thought to be more of an upland species and *P. hypselopterus* a more swampy lowland species. Thus, under the hypothesis of Suttkus and Mettee [7] *P. hypselopterus* populations then migrated as far west as the Mobile Bay area and east to the Apalachicola River Drainage, and via a stream capture with the Ochlocknee River, it dispersed further northeast.

As further described by Suttkus and Mettee [7] populations of *Pteronotropis hypselopterus* in the Apalachicola River Drainage eventually became isolated from other populations expanding to the north and east and as ancestral populations eventually gave rise to *P. grandipinnis*. Eventually, the *P. hypselopterus* stock spreading east gained access to the Suwannee and St. Mary's rivers and through interconnecting drainages spread as far north as the Pee Dee River System in South Carolina and as far south as the Myakka River in peninsular Florida. Through changes in drainage patterns the once continuous population of *P. hypselopterus* ranging from South Carolina to Florida became fragmented with the South Carolina and Georgia populations evolving into *P. stonei* and the Florida populations east of the Apalachicola diverging into *P. metallicus* [24].

3. Methods

For the phylogeographic analysis, multiple populations and multiple individuals throughout the ranges of species of the *P. hypselopterus* complex, as well as *P. signipinnis, P. euryzonus, P. hubbsi, P. welaka,* and *P. harperi,* were sequenced for the mitochondrial gene ND2. *Pteronotropis harperi* is included in this analysis as almost all previous analyses of *Pteronotropis* [8, 25–28], using both mitochondrial (Cytochrome *b,* 12S and 16S ribosomal RNA) and nuclear (RAG 1 and S7) sequences, have corroborated the hypothesis that this species is imbedded within *Pteronotropis*. Included in our analyses were *P. hypselopterus* ($n = 65$; 25 localities), *P. merlini* ($n = 10$; 4 localities), *P. grandipinnis* ($n = 9$; 4 localities), *P. metallicus* ($n = 31$; 10 localities), *P. stonei* ($n = 21$; 11 localities), *P. signipinnis* ($n = 23$; 11 localities), *P. euryzonus* ($n = 8$ individuals; 3 localities), *P. welaka* ($n = 5$; 4 localities), *P. hubbsi* ($n = 4$; 2 localities), and *P. harperi* ($n = 12$; 5 localities) for a total number of 188 individuals from 79 localities. Taxa purported earlier to be the close relatives of *Pteronotropis* were used as outgroups and included species of *Notropis*, *Cyprinella*, and *Lythrurus* (based on previous classification of *Pteronotropis* and these three genera previously in *Notropis* [29, 30]). *Pteronotropis welaka* and *P. hubbsi* also served as outgroups based on phylogenetic reconstructions by Mayden and Allen [8, 25–28]. A complete listing of sample records is provided in Table 1.

Complete genomic extractions were performed using QIAgen QIAamp tissue kits (QIAGEN, Valencia CA). The entire mitochondrial ND2 coding region was amplified using PCR with the following conditions: denaturation 94°C for 40 seconds, annealing 56°C for 60 seconds, and extension 72°C for 90 seconds. This was performed for 35 cycles with each 50 L PCR reaction consisting of 4 L of DNTPs, 5 L of 10X *Taq* buffer, 2.5 L of both forward and reverse primers, 30.7 L of dH_2O, 5 L of $MgCl_2$, and 0.3 L of *Taq.* PCR product purification was performed using either a QIAgen gel extraction kit (QIAgen, Valencia CA) or an Agencourt AMPure purification kit (Agencourt Biosciences, Beverly MA). Sequencing was performed using a big dye labeled dideoxy sequencing kit (Big Dye) and visualized on an ABI 377 automated sequencer (Auburn University Molecular Genetics Instrumentation Facility, Auburn, AL) or an ABI 3700 (Macrogen Sequencing Facility, Seoul, South Korea). Sequences were edited and aligned by eye using BioEdit versus 0.9 (Hall [31]).

Parsimony analyses were initially run in PAUPrat [32] using 5–25% character permutations. The best tree found from these analyses was used in all subsequent parsimony analyses. Maximum parsimony (MP; MPA = MP analysis) was performed in PAUP* [33] with 1000 random addition sequence replicates and tree bisection-reconnection branch swapping (TBR). All characters were equally weighted and unordered. Likelihood analyses were performed using the general algorithm for rapid likelihood inference (GARALI) [34, 35] and the GTR+ G+I model of evolutionary change; the tree with the best likelihood score was retained and is presented as the optimum ML topology. Bootstrap analysis (BA) was completed using 100 bootstrap pseudoreplicates

Table 1: Species, localities, and GenBank numbers for phylogenetic study of species of *Pteronotropis*.

(a)

Species drainage		Locality		Catalogue number	Extraction number	GenBank number
Pteronotropis welaka						
Cahaba R.	Lighseys pond	Bibb	AL	UAIC 10391	10	KP101134
Cahaba R.	Lightseys pond	Bibb	AL	UAIC 10391	11	KP101135
Apalachicola R.	Spring Cr.	Miller	GA	STL 1114.03	81	KP101136
Apalachicola R.	Spring Cr.	Miller	GA	STL 1114.03	82	KP101137
Choctawhatchee R.	Hathaway Mill	Holmes	FL	UAIC 14327.02	210	KP101138
Choctawhatchee R.	Hathaway Mill	Holmes	FL	UAIC 14327.02	211	KP101139
Choctawhatchee R.	Hathaway Mill	Holmes	FL	UAIC 14327.02	212	KP101140
Pteronotropis harperi						
Escambia R.	Hunter Cr.	Conecuh	AL	STL 862.01	65	KP101141
Escambia R.	Hunter Cr.	Conecuh	AL	STL 862.01	66	KM363660
Escambia R.	Patsaliga Cr.	Conecuh	AL	STL 367.01	67	KM363661
Escambia R.	Patsaliga Cr.	Conecuh	AL	STL 367.01	68	KM363662
Escambia R.	Patsaliga Cr.	Conecuh	AL	STL 367.01	69	KM363663
Chattahoochee R.	Kirkland Cr.	Early	GA	STL 689.03	70	KP101142
Chattahoochee R.	Kirkland Cr.	Early	GA	STL 689.03	71	KM363664
Apalachicola R.	Coolewahee Cr.	Baker	GA	STL 691.01	73	KM363665
Apalachicola R.	Coolewahee Cr.	Baker	GA	STL 691.01	74	KM363666
Apalachicola R.	Coolewahee Cr.	Baker	GA	STL 691.01	75	KM363667
Apalachicola R.	Spring Cr.	Miller	GA	STL 1114.01	83	KM363671
Apalachicola R.	Spring Cr.	Miller	GA	STL 1114.01	85	KM363672
Pteronotropis hubbsi						
Ouachita R.	Backwater pond	Ouachita	LA	UAIC 11928.01	06	KM363617
Ouachita R.	Backwater pond	Ouachita	LA	UAIC 11928.01	07	KM363617
Little R.	Little R.	McCurtin	OK	UAIC 12053	41	KM363643
Little R.	Little R.	McCurtin	OK	UAIC 12053	42	KM363644
Pteronotropis grandipinnis						
Apalachicola R.	Irwin Mill Cr.	Houston	AL	No voucher	12	KM363620
Apalachicola R,	Irwin Mill Cr.	Houston	AL	No voucher	13	KM363621
Apalachicola R,	Spring Cr.	Miller	GA	STL 1114.02	77	KM363668
Apalachicola R,	Spring Cr.	Miller	GA	STL 1114.02	78	KM363669
Apalachicola R.	Spring Cr.	Miller	GA	STL 1114.02	79	KM363670
Apalachicola R.	Beaver Cr.	Taylor	GA	STL 1129.01	116	KM363689
Apalachicola R.	Beaver Cr.	Taylor	GA	STL 1129.01	117	KM363690
Apalachicola R.	Cherokee Cr.	Lee	GA	GMNHTC 6252	141	KM363692
Apalachicola R.	Cherokee Cr.	Lee	GA	GMNHTC 6252	144	KP101143
Pteronotropis hypselopterus						
Mobile R.	Cedar Cr.	Mobile	AL	UAIC 12730.02	01	KM363612
Mobile R.	Cedar Cr.	Mobile	AL	UAIC 12730.02	02	KM363613
Mobile R.	Cedar Cr.	Mobile	AL	UAIC 12730.02	03	KM363614
Alabama R.	Little Reedy Cr.	Clarke	AL	UAIC 10850	18	KM363626
Yellow R.	Crooked Cr.	Covington	AL	UAIC 11026	19	KM363627
Choctawhatchee R.	Ponce DeLeon	Holmes	FL	UAIC 12649	20	KM363628
Yellow R.	Pond Cr.	Okaloosa	FL	UAIC 12594	25	KM363632
Escambia R.	Tenmile Cr.	Santa Rosa	FL	UAIC 12593	33	KM363636
Escambia R.	Tenmile Cr.	Santa Rosa	FL	UAIC 12593	34	KP101144
Tombigbee R.	Mill Cr.	Clarke	AL	UAIC 11050	38	KM363640
Tombigbee R.	Mill Cr.	Clarke	AL	UAIC 11050	39	KM363641
Tombigbee R.	Mill Cr.	Clarke	AL	UAIC 11050	40	KM363642
Escambia R.	Pritchett Mill Br.,	Escambia	FL	STL 684.03	55	KM363652
Choctawhatchee Bay	Garnier Cr.	Okaloosa	FL	STL 620.01	56	KM363653
Choctawhatchee Bay	Garnier Cr.	Okaloosa	FL	STL 620.01	57	KM363654
Choctawhatchee Bay	Garnier Cr.	Okaloosa	FL	STL 620.01	58	KM363655
Yellow R.	Turkey Hen Cr.	Okaloosa	FL	STL 685.02	59	KM363656

(a) Continued.

Species drainage		Locality		Catalogue number	Extraction number	GenBank number
Yellow R.	Turkey Hen Cr.	Okaloosa	FL	STL 685.02	60	KM363657
Mobile Bay	Olive Cr.	Baldwin	AL	STL 363.02	62	KM363658
Blackwater R.	Blackwater R.	Okaloosa	FL	No voucher	161	KM363697
Blackwater R.	Blackwater R.	Okaloosa	FL	No voucher	162	KM363698
Blackwater R.	Blackwater R.	Okaloosa	FL	No voucher	163	KM363699
Blackwater R.	Ates Cr.	Santa Rosa	FL	No voucher	164	KM363700
Blackwater R.	Ates Cr.	Santa Rosa	FL	No voucher	165	KM363701
Blackwater R.	Ates Cr.	Santa Rosa	FL	No voucher	166	KM363702
Fish R.	Unnamed trib.	Baldwin	AL	UAIC 14317.01	167	KM363703
Fish R.	Unnamed trib.	Baldwin	AL	UAIC 14317.01	169	KM363704
Perdido R.	Blackwater R.	Baldwin	AL	UAIC 14318	173	KM363707
Perdido R.	Blackwater R.	Baldwin	AL	UAIC 14318	174	KM363708
Perdido R.	Blackwater R.	Baldwin	AL	UAIC 14318	175	KM363709
Escambia R.	Pine Barren Cr	Escambia	FL	UAIC 14320	179	KM363710
Blackwater R.	Cobb Cr.	Santa-Rosa	FL	UAIC 14321	182	KM363711
Blackwater R.	Cobb Cr.	Santa-Rosa	FL	UAIC 14321	183	KM363712
Blackwater R.	Cobb Cr.	Santa-Rosa	FL	UAIC 14321	184	KM363713
Blackwater R.	Ates Cr.	Santa-Rosa	FL	UAIC 14322.01	185	KM363714
Yellow R.	Julian Mill Cr.	Santa-Rosa	FL	UAIC 14323.01	191	KM363715
Yellow R.	Julian Mill Cr.	Santa-Rosa	FL	UAIC 14323.01	192	KM363716
Yellow R.	Julian Mill Cr.	Santa-Rosa	FL	UAIC 14323.01	193	KM363717
Yellow R.	Juniper Cr.	Okaloosa	FL	UAIC 14324	195	KM363718
Yellow R.	Juniper Cr.	Okaloosa	FL	UAIC 14324	196	KM363719
Yellow R.	Juniper Cr.	Okaloosa	FL	UAIC 14324	197	KM363720
Yellow R.	Mill Cr.	Okaloosa	FL	UAIC 14325.01	198	KM363721
Yellow R.	Mill Cr.	Okaloosa	FL	UAIC 14325.01	199	KM363722
Yellow R.	Mill Cr.	Okaloosa	FL	UAIC 14325.01	200	KM363723
Choctawhatchee R.	Blue Cr.	Holmes	FL	UAIC 14326	204	KM363724
Choctawhatchee R.	Blue Cr.	Holmes	FL	UAIC 14326	205	KM363725
Choctawhatchee R.	Blue Cr.	Holmes	FL	UAIC 14326	206	KM363726
Choctawhatchee R.	Hathaway Mill	Holmes	FL	UAIC 14327.01	207	KM363727
Choctawhatchee R.	Hathaway Mill	Holmes	FL	UAIC 14327.01	208	KM363728
Choctawhatchee R.	Hathaway Mill	Holmes	FL	UAIC 14327.01	209	KM363729
Choctawhatchee R.	Wrights Cr.	Holmes	FL	UAIC 14328.01	213	KM363730
Choctawhatchee R.	Wrights Cr.	Holmes	FL	UAIC 14328.01	214	KM363731
Choctawhatchee R.	Seven Runs Cr.	Holmes	FL	UAIC 14329.01	217	KM363732
Choctawhatchee R.	Seven Runs Cr.	Holmes	FL	UAIC 14329.01	218	KM363733
St. Andrews Bay	Cooks Bayou	Bay	FL	UAIC 14330	219	KM363734
St. Andrews Bay	Cooks Bayou	Bay	FL	UAIC 14330	220	KM363735
St. Andrews Bay	Cooks Bayou	Bay	FL	UAIC 14330	221	KM363736
St. Andrews Bay	Unnamed trib.	Bay	FL	UAIC 14331	222	KM363737
St. Andrews Bay	Unnamed trib.	Bay	FL	UAIC 14331	223	KM363738
St. Andrews Bay	Unnamed trib.	Bay	FL	UAIC 14331	224	KM363739
Choctawhatchee Bay	Bear Cr.	Bay	FL	UAIC 14332	225	KM363740
Choctawhatchee Bay	Bear Cr.	Bay	FL	UAIC 14332	226	KM363741
Choctawhatchee R.	Spring Cr.	Geneva	AL	UAIC 14343	255	KM363756
Choctawhatchee R.	Spring Cr.	Geneva	AL	UAIC 14343	256	KM363757
Choctawhatchee R.	Spring Cr.	Geneva	AL	UAIC 14343	257	KM363758
Pteronotropis sp. cf. *hypselopterus*						
St. Johns R.	Little Orange Cr.,	Putnam	FL	UAIC 12290	28	KP101145
St. Johns R.	Juniper Cr.,	Marion	FL	GMNH5380	149	KP101146
St. Johns R.	Juniper Cr.	Marion	FL	GMNH5380	150	KP101147
St. Johns R.	Juniper Cr.	Marion	FL	GMNH5380	151	KP101148
Alafia R.	Hurrah Cr.	Hillsborough	FL	UAIC 14339	243	KP101149
Alafia R.	Hurrah Cr.	Hillsborough	FL	UAIC 14339	244	KP101150
Alafia R.	Hurrah Cr.	Hillsborough	FL	UAIC 14339	245	KP101151

(a) Continued.

Species drainage		Locality		Catalogue number	Extraction number	GenBank number
St. Johns R.	Juniper Cr.	Marion	FL	UAIC 14340	246	KP101152
St. Johns R.	Juniper Cr.	Marion	FL	UAIC 14340	247	KP101153
St. Johns R.	Juniper Cr.	Marion	FL	UAIC 14340	248	KP101154
St. Johns R.	Alexander Spr.	Lake	FL	UAIC 14341	249	KP101155
St. Johns R.	Alexander Spr.	Lake	FL	UAIC 14341	250	KP101156
St. Johns R.	Alexander Spr.	Lake	FL	UAIC 14341	251	KP101157
Pteronotropis euryzonus						
Chattahoochee R.	Maringo Cr.	Russell	AL	UAIC 12229	22	KM363629
Chattahoochee R.	Snake Cr.	Russell	AL	UAIC 10493	51	KM363648
Chattahoochee R.	Snake Cr.	Russell	AL	UAIC 10493	52	KM363649
Chattahoochee R.	Snake Cr.	Russell	AL	UAIC 10493	53	KM363650
Chattahoochee R.	Snake Cr.	Russell	AL	UAIC 10493	54	KM363651
Chattahoochee R.	Uchee Cr.	Russell	AL	UAIC 14344	258	KM363759
Chattahoochee R.	Uchee Cr.	Russell	AL	UAIC 14344	259	KM363760
Chattahoochee R.	Uchee Cr.	Russell	AL	UAIC 14344	260	KM363761
Pteronotropis merlini						
Choctawhatchee R.	Claybank Cr.	Dale	AL	UAIC 12595	08	KM363617
Choctawhatchee R.	Claybank Cr.	Dale	AL	UAIC 12595	09	KM363619
Pea R.	Clearwater Cr.	Coffee	AL	No voucher	15	KM363623
Pea R.	Clearwater Cr.	Coffee	AL	No voucher	16	KM363624
Pea R.	Clearwater Cr.	Coffee	AL	No voucher	17	KM363625
Choctawhatchee R.	W. Fork	Barbour	AL	UAIC 12735	29	KM363633
Choctawhatchee R.	W. Fork	Barbour	AL	UAIC 12735	30	KM363634
Choctawhatchee R.	W. Fork	Barbour	AL	UAIC 12735	31	KM363635
Choctawhatchee R.	Unnamed trib.	Geneva	AL	UAIC 14342	252	KM363754
Choctawhatchee R.	Unnamed trib.	Geneva	AL	UAIC 14342	254	KM363755
Choctawhatchee R.	Unnamed trib.	Geneva	AL	UAIC 14342	261	KP101158
Pteronotropis metallicus						
Suwannee R.	Sampson R.	Bradfrod	FL	UF 158855	96	KM363673
Ochlockonee R.	Rocky Comf. Cr.	Gadsden	FL	No voucher	155	KM363693
Ochlockonee R.	Rocky Comf. Cr.	Gadsden	FL	No voucher	156	KM363694
Ochlockonee R.	Rocky Comf. Cr.	Gadsden	FL	No voucher	157	KM363695
St. Marks R.	St. Marks R.	Leon	FL	No voucher	159	KM363696
St. Marks R.	Chicken Br.	Leon	FL	UAIC 14334	231	KM363742
St. Marks R.	Chicken Br.	Leon	FL	UAIC 14334	232	KM363743
St. Marks R.	Chicken Br.	Leon	FL	UAIC 14334	233	KM363744
Suwannee R.	Hunter Cr.	Hamilton	FL	UAIC 14336	234	KM363745
Suwannee R.	Hunter Cr.	Hamilton	FL	UAIC 14336	235	KM363746
Suwannee R.	Hunter Cr.	Hamilton	FL	UAIC 14336	236	KM363747
St. Marys R.	Cedar Cr.	Baker	FL	UAIC 14337	237	KM363748
St. Marys R.	Cedar Cr.	Baker	FL	UAIC 14337	238	KM363749
St. Marys R.	Cedar Cr.	Baker	FL	UAIC 14337	239	KM363750
Suwannee R.	Santa-Fe R.	Gilchrist	FL	UAIC 14338	240	KM363751
Suwannee R.	Santa-Fe R.	Gilchrist	FL	UAIC 14338	241	KM363752
Suwannee R.	Santa-Fe R.	Gilchrist	FL	UAIC 14338	242	KM363753
Pteronotropis signipinnis						
Pascagoula R.	Beaverdam Cr.	Forest	MS	UAIC 13416.03	04	KM363615
Pascagoula R.	Beaverdam Cr.	Forest	MS	UAIC 13416.03	05	KM363616
Mobile R.	Cedar Cr.	Mobile	AL	UAIC 12730.15	23	KM363630
Mobile R.	Cedar Cr.	Mobile	AL	UAIC 12730.15	24	KM363631
Tensaw R.	Ferris Cr.	Baldwin	AL	UAIC 11056	35	KM363637
Tensaw R.	Ferris Cr.	Baldwin	AL	UAIC 11056	36	KM363638
Tensaw R.	Ferris Cr.	Baldwin	AL	UAIC 11056	37	KM363639
Pearl R.	Lawrence Cr.	Washington	LA	UAIC 12204	44	KM363645
Pearl R.	Lawrence Cr.	Washington	LA	UAIC 12204	45	KM363646
Pearl R.	Lawrence Cr.	Washington	LA	UAIC 12204	46	KM363647

(a) Continued.

Species drainage		Locality		Catalogue number	Extraction number	GenBank number
Biloxi R.	Saucier Cr.	Harrison	MS	STL 85563	63	KM363659
Fish R.	Unnamed trib.	Baldwin	AL	UAIC 14317.02	171	KM363705
Fish R.	Unnamed trib.	Baldwin	AL	UAIC 14317.02	172	KM363706
Pteronotropis sp. cf. *signipinnis*						
Escambia R.	Pritchett Mill	Escambia	FL	UAIC 684.02	64	KP101159
Perdido R.	Beartree Cr.	Baldwin	AL	UAIC 14319	177	KP101160
Perdido R.	Beartree Cr.	Baldwin	AL	UAIC 14319	178	KP101161
Yellow R.	Julian Mill Cr.	Santa-Rosa	FL	UAIC 14323.02	194	KP101162
Yellow R.	Mill Cr.	Okaloosa	FL	UAIC 14325.02	201	KP101163
Yellow R.	Mill Cr.	Okaloosa	FL	UAIC 14325.02	202	KP101164
Yellow R.	Mill Cr.	Okaloosa	FL	UAIC 14325.02	203	KP101165
Apalachicola R.	Fourmile Cr.	Calhoun	FL	UAIC 14333	228	KP101166
Apalachicola R.	Fourmile Cr.	Calhoun	FL	UAIC 14333	229	KP101167
Apalachicola R.	Fourmile Cr.	Calhoun	FL	UAIC 14333	230	KP101168
Pteronotropis stonei						
Santee R.	Jacks Cr.	Clarendon	SC	UAIC 12590	14	KM363622
Santee R.	Unnamed Cr.	Calhoun	SC	STL 1120.01	98	KM363674
Santee R.	Unnamed Cr.	Calhoun	SC	STL 1120.01	99	KM363675
N. Fork Edisto R.	Murphy Mill Cr.	Calhoun	SC	STL 1121.01	100	KM363676
N. Fork Edisto R.	Murphy Mill Cr.	Calhound	SC	STL 1121.01	101	KM363677
Combahee R.	Savannah Cr.	Colleton	SC	STL 1122.01	102	KM363678
Combahee R.	Savannah Cr.	Colleton	SC	STL 1122.01	103	KM363679
Combahee R.	Salkehatchie R.	Barnwell	SC	STL 1123.01	105	KM363680
Santee R.	Congaree Cr.	Lexington	SC	STL 1124.01	106	KM363681
Santee R.	Congaree Cr.	Lexington	SC	STL 1124.01	107	KM363682
N. Fork Edisto R.	Black Cr.	Lexington	SC	STL 1125.01	108	KM363683
N. Fork Edisto R.	Black Cr.	Lexington	SC	STL 1125.01	109	KP101169
Savannah R.	Cedar Cr.	Aiken	SC	STL 1126.01	110	KP101171
Savannah R.	Cedar Cr.	Aiken	SC	STL 1126.01	111	KM363684
S. Fork Edisto R.	Unnamed trib.	Aiken	SC	STL 1127.01	112	KM363685
S. Fork Edisto R.	Unnamed trib.	Aiken	SC	STL 1127.01	113	KM363686
Savannah R.	Boggy Gut Cr.	Richmond	GA	STL 1128.01	114	KM363687
Savannah R.	Boggy Gut Cr.	Richmond	GA	STL 1128.01	115	KM363688
Pee Dee R.	Beaver Dam Cr.	Kershaw	SC	STL 1130.01	118	KM363691
Pee Dee R.	Beaver Dam Cr.	Kershaw	SC	STL 1130.01	119	KP101171

(b)

Outgroup taxa			
Species	GenBank number	Species	GenBank number
Cyprinella labrosa	AF111258.1	*Cyprinella lutrensis*	AF111210 AF111210.11
Cyprinella monacha	AF111228.1	*Cyprinella zanema*	AF111230.1
Lythrurus roseipinnis	AF111231.1	*Notropis atherinoides*	AF111232.1
Notropis baileyi	EF613593.1	*Notropis stramineus*	NC 008110.1
Notropis texanus	EF613581.1		

[36]. Bayesian analyses used MrBayes [37] with four heated Markov chains and default temperature setting. Each analysis was run for 1 million generations with sampling every 250 generations. Log-likelihood scores were plotted against generation time to establish burn-in; trees before stationary were discarded. A 50% majority rule consensus tree was generated from the remaining trees.

Uncorrected and corrected genetic distances were calculated using MEGA 3 [36]. Statistical parsimony [38, 39], as implemented in TCS 1.18 [40, 41], was used to group haplotypes into a minimum-connecting networks to illustrate potential genealogical connections. Mismatch distributions of the number of differences between haplotypes [12] and genetic statistics were conducted using DnaSP version 3 [42]

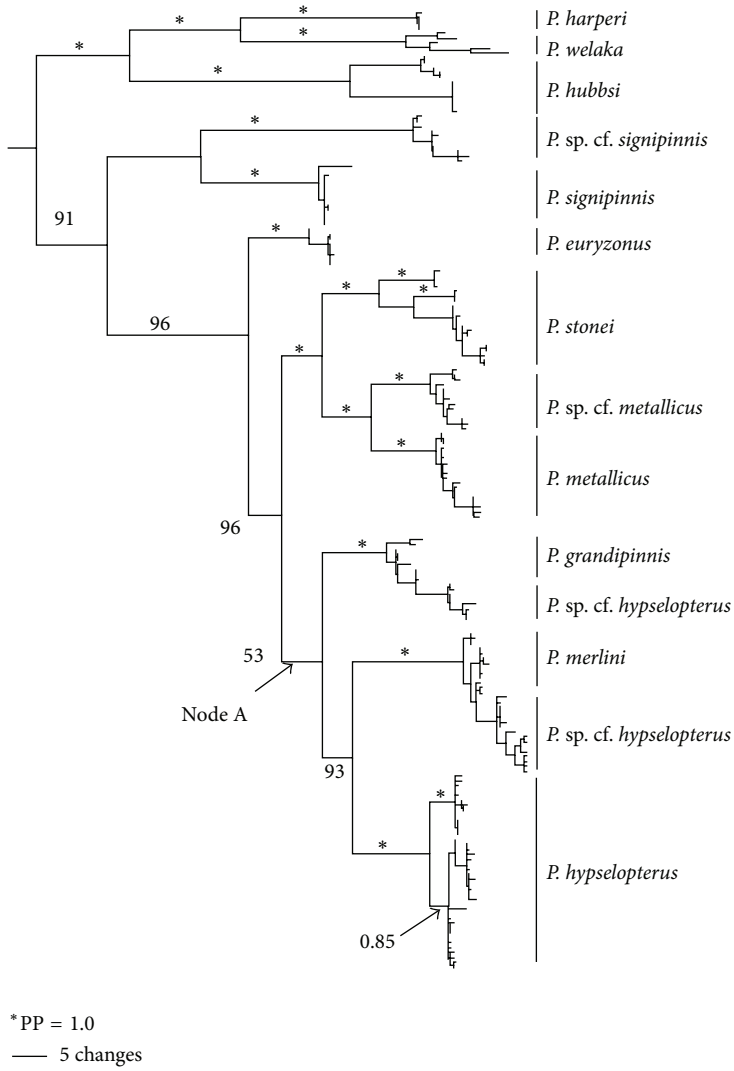


Figure 2: Relationships of species and populations of *Pteronotropis* as inferred from Bayesian analysis of ND2 gene sequences and summarized using a 50% majority rule consensus tree. ∗ = 100PP.

to examine demographic differences between clades within the trees.

4. Results

Sequences for the complete ND2 gene (1047 bp) yielded 172 haplotypes from among 188 individuals from 79 localities across the range of *Pteronotropis*. Of the 1047 characters identified, 517 were parsimony informative (49.5%). MPA resulted in 1867 equally parsimonious trees with 2437 steps (CI = 0.452, RI = 0.654). ML analyses resulted in a most likely tree log value of −41397.64. Bayesian analyses reached stationary after 100,000 generations; trees from the first 125,000 generations were discarded as burn-in, leaving 11,764 trees for phylogeny estimation.

MP, ML, and BA all recovered 10 strongly supported major clades within *Pteronotropis* with essentially identical topologies; only Bayesian trees are illustrated for further discussion. ND2 sequences failed to resolve *Pteronotropis* as a monophyletic group without the inclusion of *Notropis harperi*, a species now included in *Pteronotropis* by Mayden and Allen [8] from their analyses using two nuclear genes. The overall topology of the tree (Figure 2) reveals two reciprocally monophyletic groups, the *Pteronotropis harperi* clade sister to the *Pteronotropis signipinnis* clade. The *P. harperi* clade includes *P. welaka*, *P. hubbsi,* and *P. harperi*. Within this group *P. hubbsi* is sister to *P. welaka,* and *P. harperi* is the basal-most sister species. In the sister *P. signipinnis* clade, *Pteronotropis signipinnis* is resolved as the basal sister group to remaining species. *Pteronotropis euryzonus* is likewise resolved as sister to a monophyletic *P. hypselopterus* complex with two sister clades ((*P. stonei* plus *P. metallicus*) sister to (*P. hypselopterus* (*P. merlini, P. grandipinnis*))). These two clades represent eastern and western groups, respectively, as defined by the Apalachicola River drainage (Figure 2). All ten major clades in the genus received high bootstrap and posterior probabilities (PP = 0.98–1.00), as similarly observed with analyses of nuclear genes by Mayden and Allen [8].

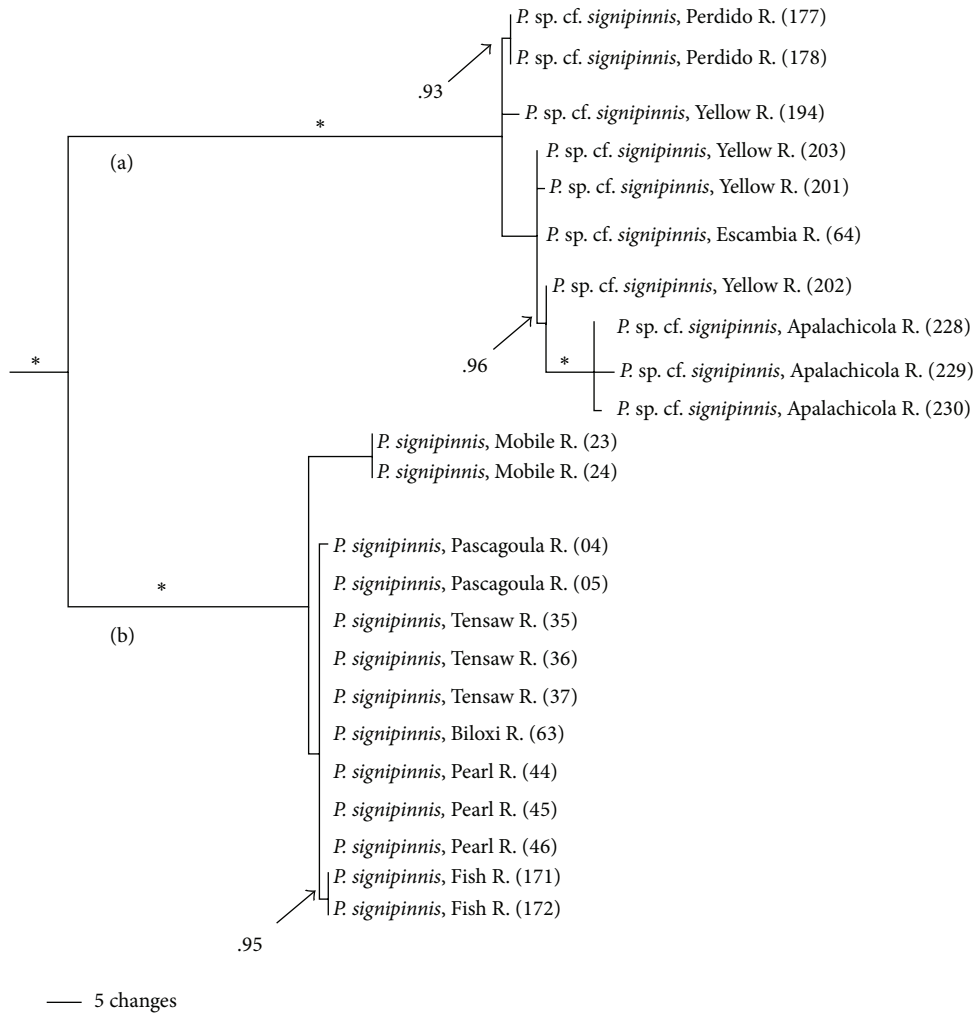


Figure 3: Relationships of populations in *Pteronotropis signipinnis* clade inferred from Bayesian analysis of ND2 sequences and summarized using a 50% majority rule consensus tree. ∗ = 100PP. (a) *Pteronotropis* sp. cf. *signipinnis* distributed east of the Mobile Basin. (b) *Pteronotropis signipinnis* distributed west of and from the Mobile Basin.

Considerable variation in ND2 sequence divergence exists across the range within the *P. signipinnis* clade with a mean sequence divergence of 10.7% (±2.6%). This is largely due to two subclades, each receiving maximum posterior probability support (PP = 1.0; Figure 3). The location of the Mobile Bay system geographically defined the boundaries of these two clades. Individuals from rivers west of and part of the Mobile Bay were recovered in one clade, herein referred to as the western *P. signipinnis* clade, whereas individuals from rivers east of the Mobile Bay formed the eastern *P. signipinnis* clade.

No structure was observed within the *P. euryzonus* clade, likely due to its highly restricted range and the long-term impacts on the habitats of this species and its shrinking range [21]. Almost all individuals possessed the same haplotype for ND2; sequence was 0.0%. The *P. stonei* clade is distributed from the Savannah River in the south to the Pee Dee River in the north (Figures 1 and 4). Populations from the Savannah River formed a clade sister to other populations (PP = 1.0; Figure 4). The Combahee River populations formed the sister group to a clade including individuals from the North and South Forks of the Edisto, Santee, and Pee Dee rivers; populations from these river systems had little genetic structure. The overall within sequence divergence for the *P. stonei* clade was 3.3% (±1.0%).

The *P. metallicus* clade includes two major subclades (PP = 1.0; Figure 5). One clade includes only an undescribed species (*P.* sp. cf. *metallicus*) from the Alafia and St. Johns rivers. The second clade includes only *P. metallicus*, and both clades received strong support (PP = 0.1). Additional, strongly supported genetic structuring exists within both *P.* sp. cf. *metallicus* and *P. metallicus*. Structure within *P.* sp. cf. *metallicus* was strongly supported divergences between and within the Alafia and St. John's rivers (PP = 0.96–1.0); some drainage structure occurred in *P. metallicus* but not along independent drainages. *Pteronotropis metallicus* populations from the Ochlocknee, St. Marks, Suwannee, and St. Marys rivers had a haplotype diversity of 0.989 and a nucleotide

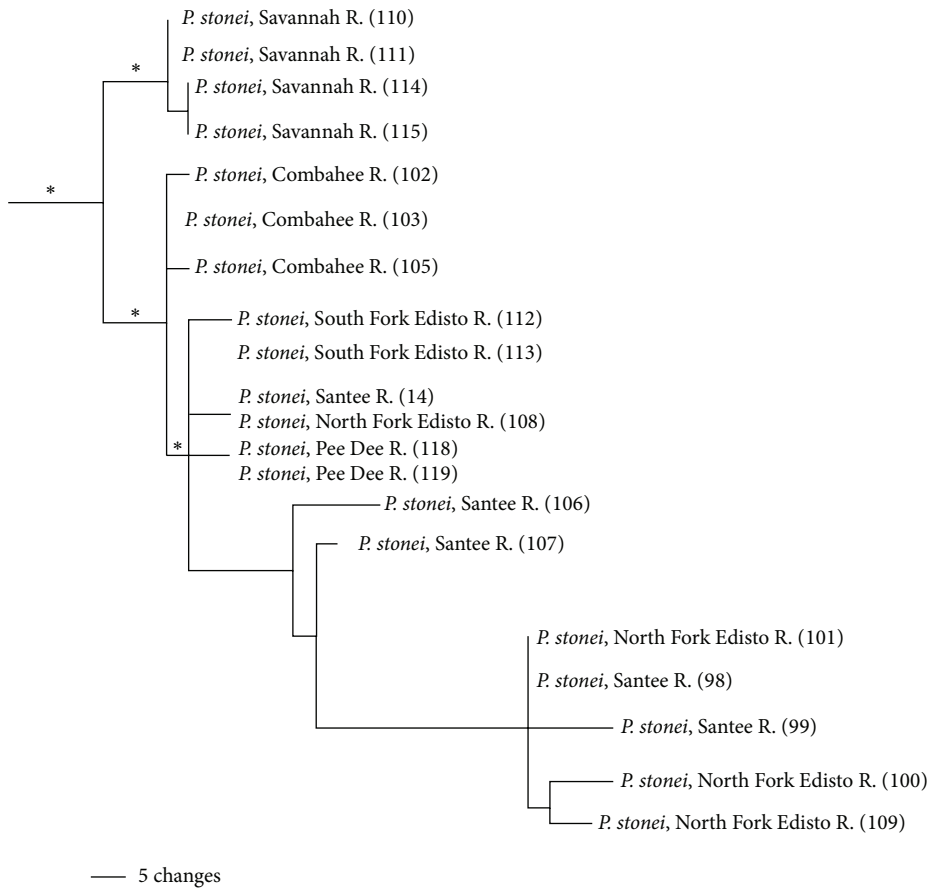


Figure 4: Relationships of populations in the *Pteronotropis stonei* clade inferred from Bayesian analysis of ND2 sequences and summarized using a 50% majority rule consensus tree. ∗ = 100PP.

Table 2: Statistical values for *Pteronotropis signipinnis*, *P. euryzonus*, and the *P. hypselopterus* complex.

Clade	N	H	H_D	π	D	P value
P. hypselopterus	40	29	0.858	0.042	−1.330	>.05
P. signipinnis	23	14	0.881	0.066	2.500	<.05*
P. merlini	30	16	0.874	0.008	0.465	>.05
P. grandipinnis	17	11	0.962	0.014	0.221	>.05
P. stonei	20	10	0.943	0.031	1.002	>.05
P. euryzonus	8	3	0.464	0.000	−1.448	>.05
P. metallicus	32	27	0.994	0.058	−0.483	>.05
P. signipinnis eastern	10	9	0.978	0.007	−0.071	>.05
P. signipinnis western	13	5	0.628	0.003	−1.882	<.05*
P. signipinnis Ochlockonee	19	16	0.989	0.059	−0.521	>.05
P. signipinnis St. Johns	13	11	0.987	0.010	−0.060	>.05
P. hypselopterus eastern	27	20	0.872	0.107	−1.23	>.05
P. hypselopterus western	13	9	0.723	0.002	−1.17	>.05

N: number of individuals, H: haplotype number, H_D: haplotype diversity, π: nucleotide diversity, D: Tajamas D statistic, and P value: the P value associated with Tajamas D statistic (* indicating a significant value).

diversity of 0.059 (Table 2). The St. Johns subclade includes populations from the St. Johns and Alafia Rivers and has a haplotype diversity of 0.987 and a nucleotide diversity of 0.010. The overall within group divergence for the *P. metallicus* clade was substantial (7.7% (±2.0%)), largely due to genetic differences between *P.* sp. cf. *metallicus* and *P. metallicus*.

Given the low support for the relationship between *P. grandipinnis* to *P. merlini* and *P. hypselopterus* these three species clades should be considered an unresolved trichotomy (Figure 2). Support for the monophyly of *P. merlini* and *P. hypselopterus* and the *P. grandipinnis* clade was strong (PP = 1.0 for each). However, as currently outlined in the evolution of haplotypes *P. grandipinnis* is paraphyletic with respect to populations of *P. hypselopterus* from the Choctawhatchee and St. Andrews bays (PP = 1.0) (Figure 6). The overall within group variation for the *P. grandipinnis* clade was low (0.6%, ±0.4%).

The *P. merlini* clade is strongly supported (PP = 1.0); however, similar to the *P. grandipinnis* clade, *P. merlini*, as currently outlined, is not monophyletic (Figure 7). Individuals of *P. hypselopterus* from the Choctawhatchee River drainage were resolved within the *P. merlini* clade. Resolution within

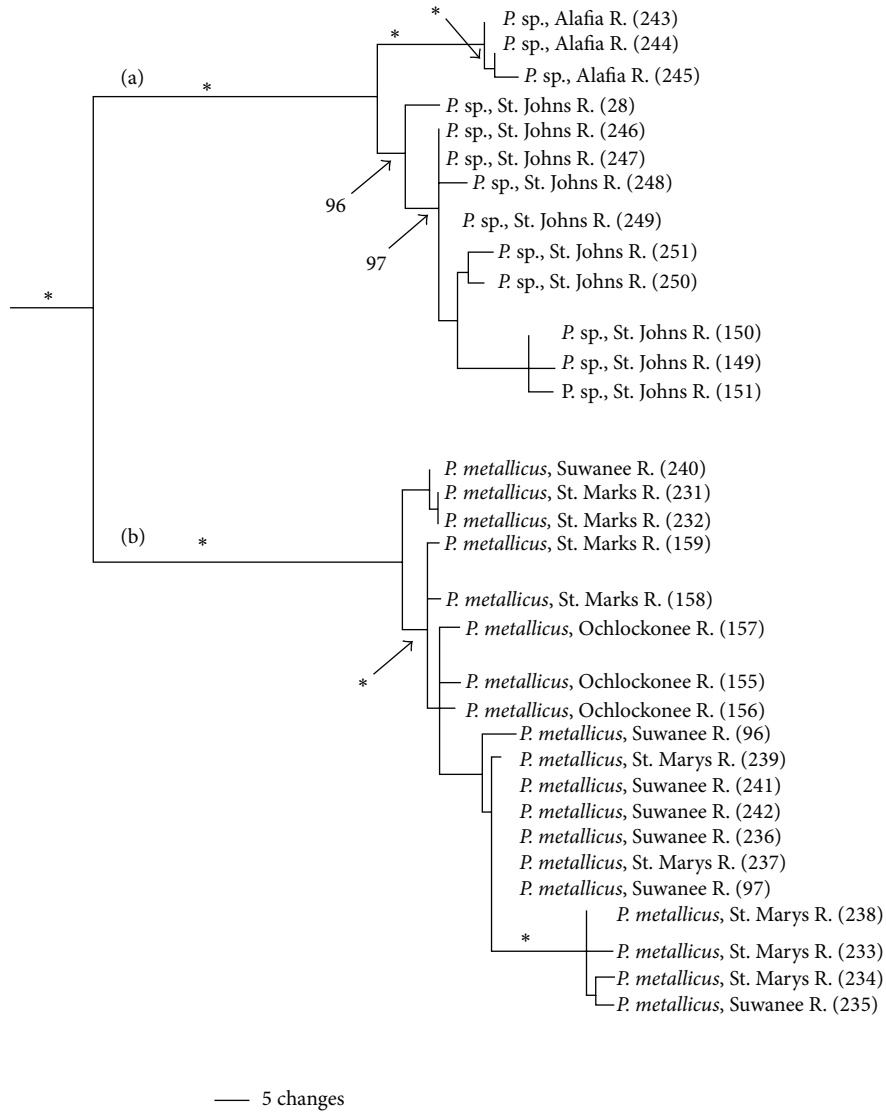


Figure 5: Relationships of populations in the *Pteronotropis metallicus* clade plus *P.* sp. cf. *metallicus* (a) inferred from Bayesian analysis of ND2 sequences and summarized using a 50% majority rule consensus tree. ∗ = 100PP.

this clade is limited and the within clade variation is low (1.6%, ±0.7%). There are, however, several well-supported subclades within this clade, one consisting of individuals of *P. merlini* from the Choctawhatchee (PP = 1.0), one *P. hypselopterus* from the Choctawhatchee River (PP = 0.99), and individuals of *P. hypselopterus* and *P. merlini*, also from within the Choctawhatchee (PP = 0.95). The *P. hypselopterus* clade is strongly supported (PP = 1.0) and includes two reciprocally monophyletic groups centered at the Mobile Bay, much like the *signipinnis* clade (Figure 8). Individuals from the Mobile Bay and associated rivers, the Alabama, Tombigbee, Fish, and Perdido rivers form one subclade (PP = 1.0) (herein referred to as the western *hypselopterus* clade). The other clade was also strongly supported and included individuals from the Yellow, Escambia, and Blackwater rivers (PP = 0.95) (herein referred to as the eastern *hypselopterus* clade). There was 4.8% (±1.2%) mean sequence divergence within the *hypselopterus* clade.

The lineages of *P. signipinnis*, *P. euryzonus*, *P. metallicus*, *P. stonei*, *P. grandipinnis*, *P. merlini*, and *P. hypselopterus* are examined in more detail in TCS analysis (Figures 9–16). The algorithm was unable to connect the eastern and western *signipinnis* clades, the *P. metallicus* subclades from the Ochlocknee and St. Johns, or the eastern and western *hypselopterus* clades at a 95% connection limit (21 steps; Figures 9–11). Within the eastern *signipinnis* clade nine haplotypes were recovered as well as strong geographic partitioning within this clade. Individuals from the Escambia, Yellow, and Perdido rivers clustered together and were nine mutational events apart from the three haplotypes found in individuals from the Apalachicola River. The western *signipinnis* clade included more individuals but included only five haplotypes. Many individuals from the Tensaw, Pascagoula, Pearl, and Biloxi rivers share a common haplotype with one individual from the Mobile Bay having a haplotype thirteen mutational steps from all others (Figure 9). For the Ochlocknee River

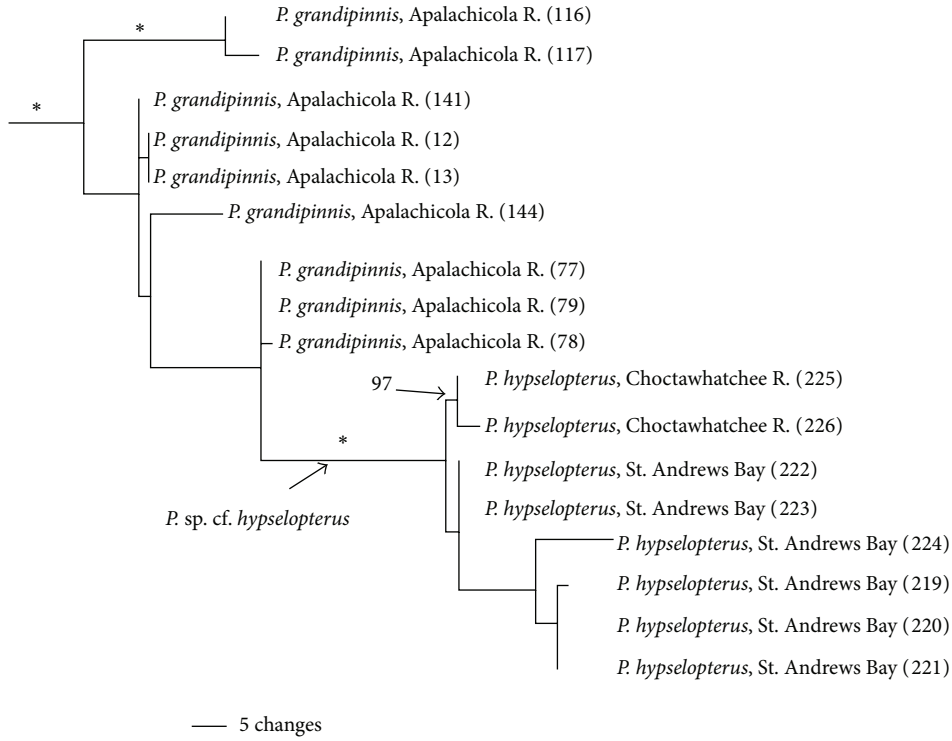


Figure 6: Relationships of populations in the *Pteronotropis grandipinnis* clade plus *P.* sp. cf. *hypselopterus* inferred from Bayesian analysis of ND2 sequences and summarized using a 50% majority rule consensus tree. ∗ = 100PP.

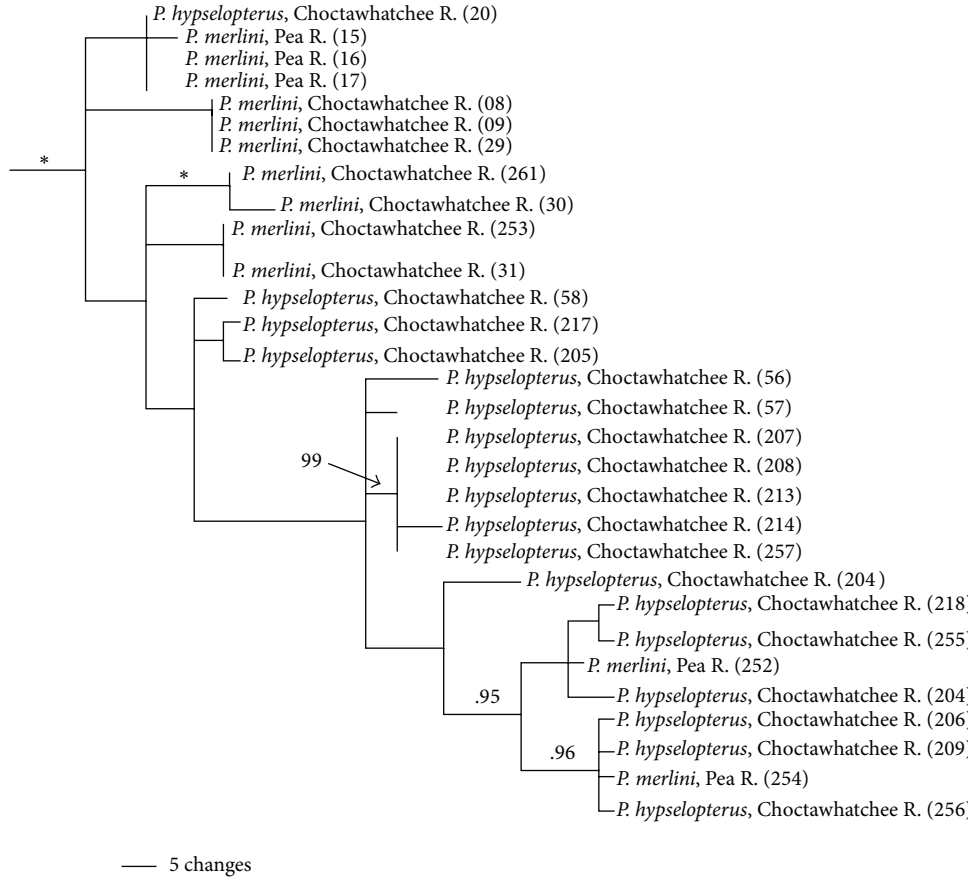


Figure 7: Relationships of populations in the *Pteronotropis merlini* clade plus *P.* sp. cf. *hypselopterus* inferred from Bayesian analysis of ND2 sequences and summarized using a 50% majority rule consensus tree. ∗ = 100PP.

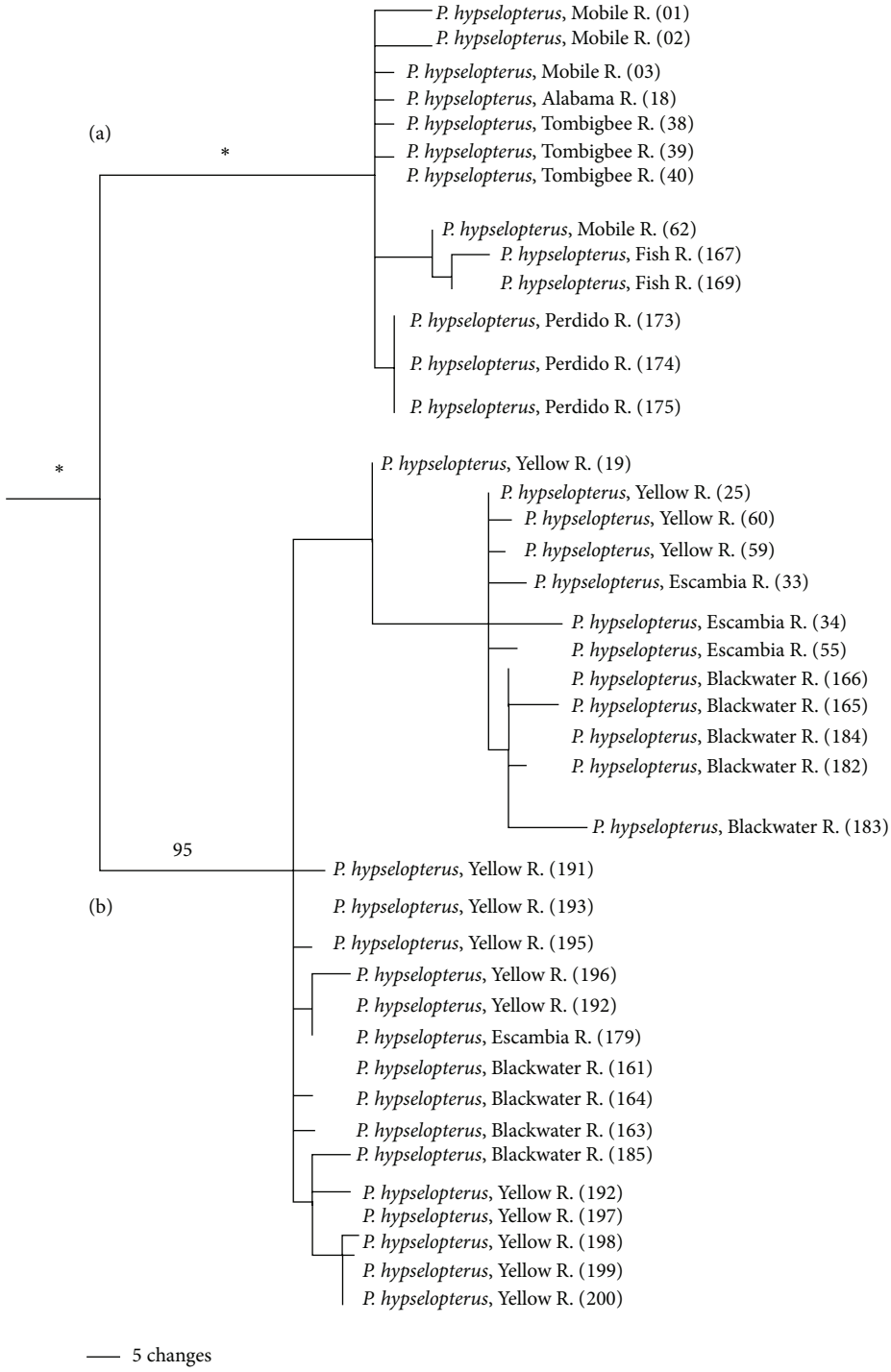


Figure 8: Bayesian 50% majority rule consensus tree for the relationships of populations in the *Pteronotropis hypselopterus* clade inferred from Bayesian analysis of ND2 sequences and summarized using a 50% majority rule consensus tree. ∗ = 100PP. (a) Clade of populations from rivers draining east and west of and from the Mobile Basin. (b) Clade of populations from rivers draining east of the Mobile Basin.

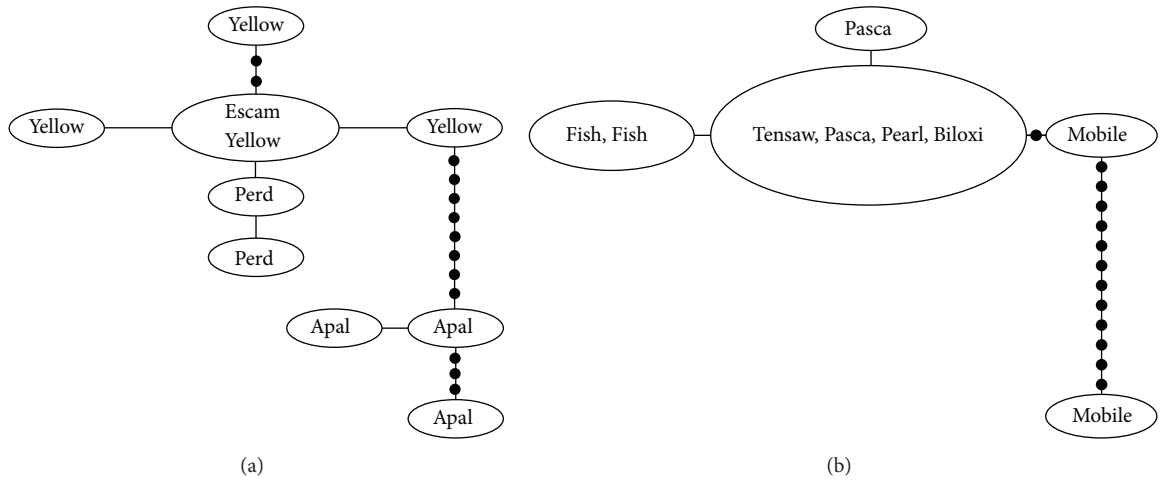


Figure 9: Haplotype network for *Pteronotropis signipinnis* clade. (a) *Pteronotropis* sp. cf. *signipinnis*. (b) *Pteronotropis signipinnis*. Apal = Apalachicola River system, Biloxi = Biloxi River system, Escam = Escambia River system, Fish = Fish River system, Pasc = Pascagoula River system, Pearl = Pearl River system, Perd = Perdido River system, Mobile = Mobile River system, Tensaw = Tensaw River system, and Yellow = Yellow River system.

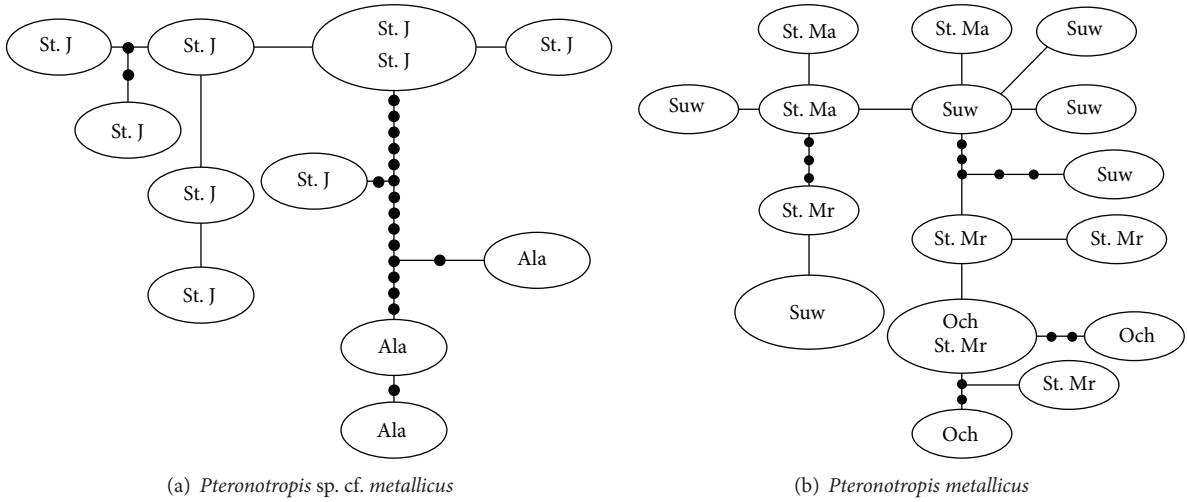


Figure 10: Haplotype networks for (a) *P.* sp. cf. *metallicus* and (b) *P. metallicus*. Ala = Alafia River system, Och = Ochlockonee, St. J = St. John River system, St. Ma = St. Marys River system, and Suw = Suwanee River system.

P. metallicus clade, 16 haplotypes were recovered from 30 individuals. In this clade there was little structure with most individuals having similar haplotypes (Figure 10(b)). This is in contrast to the St. Johns River clade that displayed significant structure (Figure 10(a)). Individuals from the St. Johns River clustered together, having 13–15 mutational steps from the three haplotypes in the Alafia River. Furthermore, within the St. Johns River clade, 11 haplotypes were recovered from 12 individuals. Both the eastern and western *P. hypselopterus* clades displayed limited geographic partitioning, with 20 recovered haplotypes from 39 individuals from the eastern *hypselopterus* clade and nine haplotypes from 13 individuals within the western *hypselopterus* clade (Figure 11).

Mismatch distribution plots for nearly all recovered clades had multimodal distributions, clearly fitting the model of nonexpanding populations or having populations at equilibrium (Figures 12 and 15–21). These results are supported by the lack of significance found in the Tajima's *D* test (Table 2). The one exception was the *signipinnis* clade (eastern and western clades combined). This clade had a bimodal distribution, indicating a stable population ([13, 14], Raggedness index = 0.288); however, Tajima's *D* test statistic was significant for a rapid population expansion (Table 2). One possible explanation for the conflicting results is that the *signipinnis* clade consisted of two distinct populations with likely independent demographic histories. To account

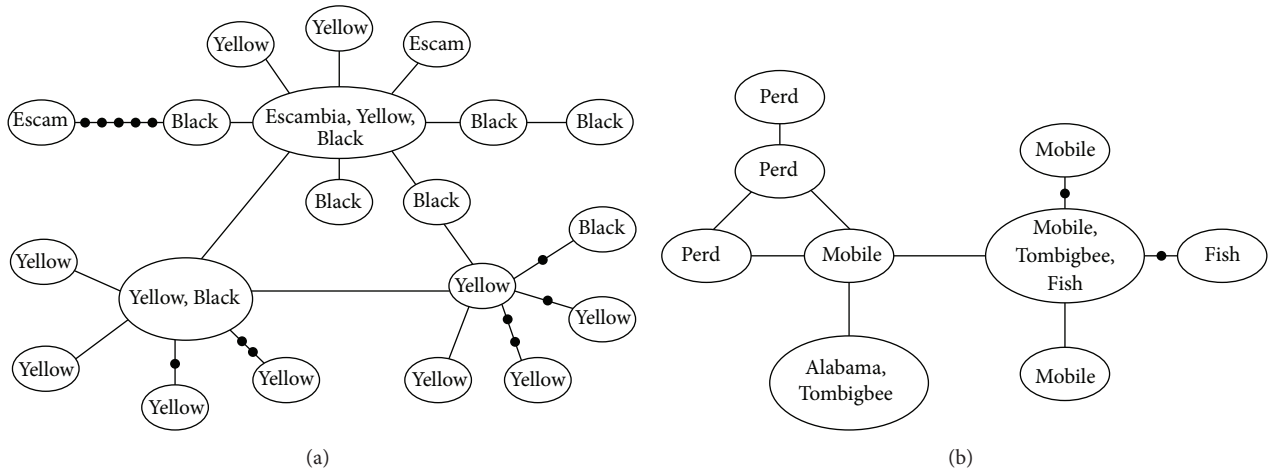


Figure 11: Haplotype networks for *Pteronotropis hypselopterus*. (a) Clade of populations from rivers draining east of Mobile Basin. (b) Clade of populations from rivers of the Mobile Basin and adjacent and more western river systems.

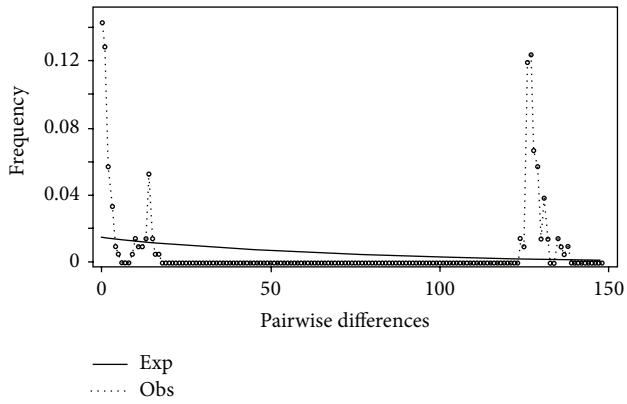


Figure 12: Mismatch distribution of *Pteronotropis signipinnis* clade.

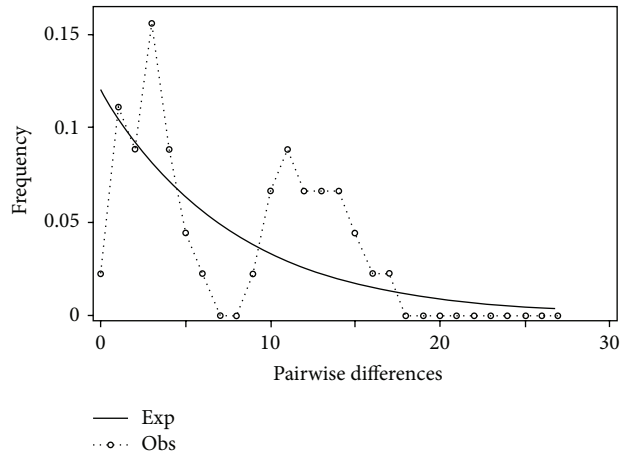


Figure 13: Mismatch distribution of eastern *Pteronotropis signipinnis* clade.

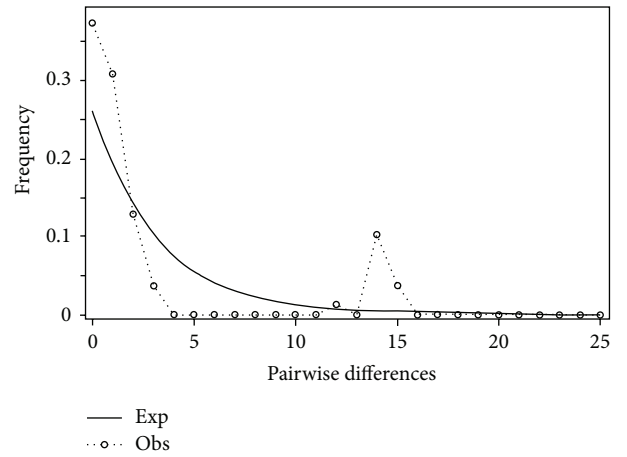


Figure 14: Mismatch distribution of western *Pteronotropis signipinnis* clade.

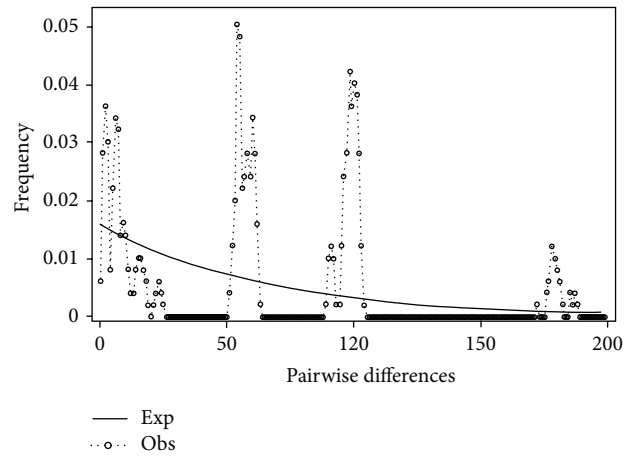


Figure 15: Mismatch distribution of *Pteronotropis metallicus* clade.

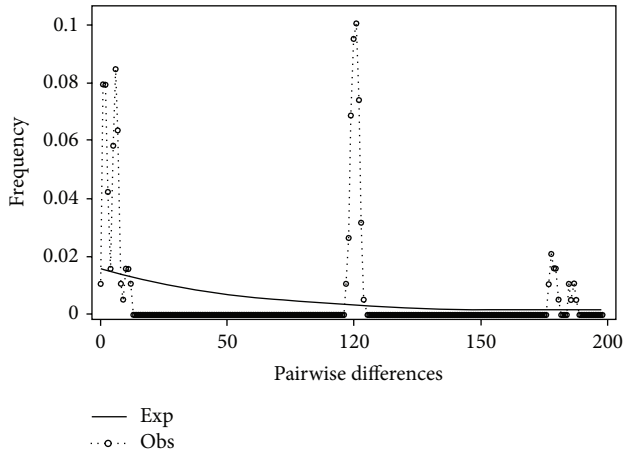


Figure 16: Mismatch distribution of the *Pteronotropis metallicus* clade excluding individuals from the St. Johns and Alafia Rivers.

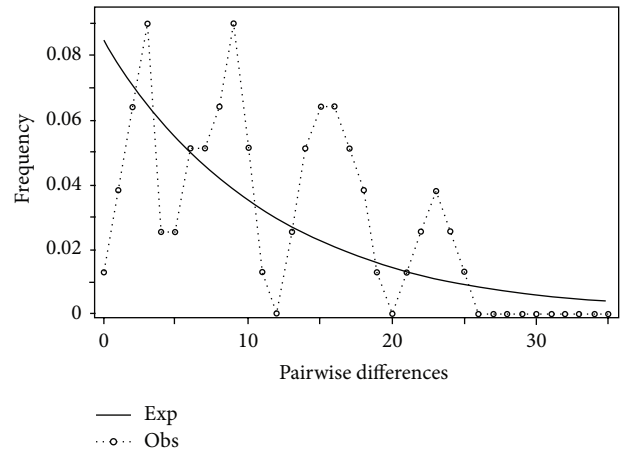


Figure 17: Mismatch distribution of the *P.* sp. cf. *metallicus* clade.

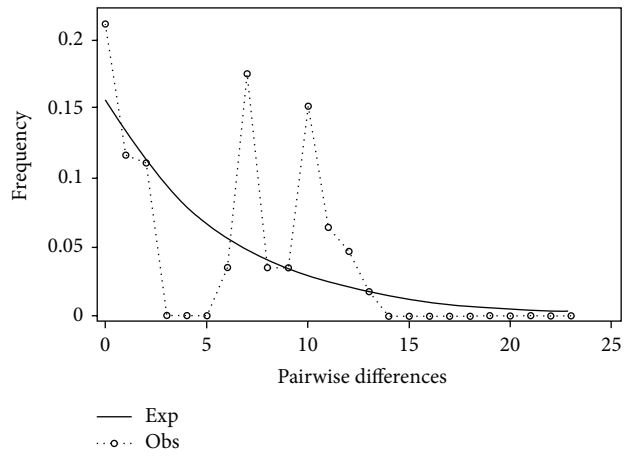


Figure 18: Mismatch distribution of the *Pteronotropis stonei* clade.

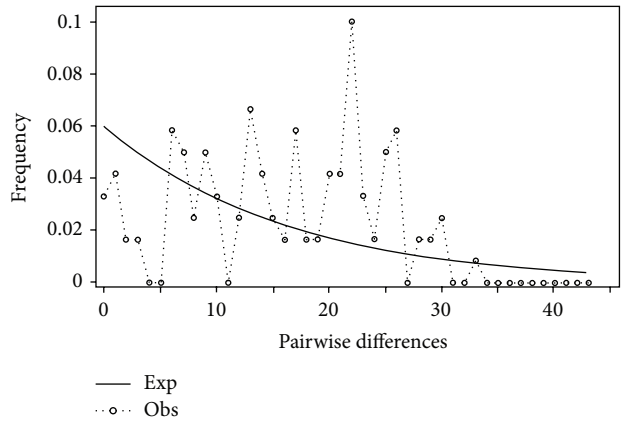


Figure 19: Mismatch distribution of the *Pteronotropis grandipinnis* clade.

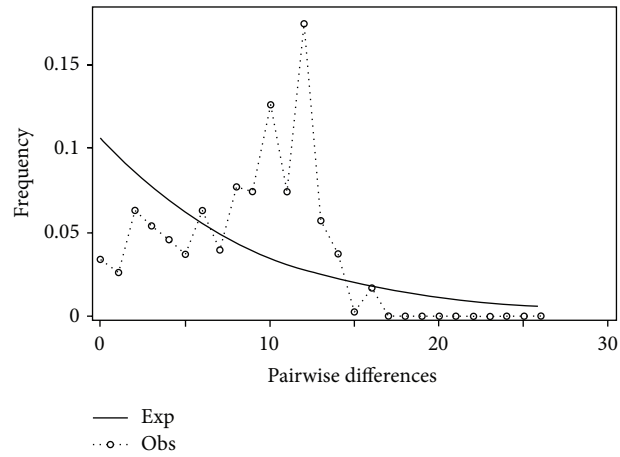


Figure 20: Mismatch distribution of the *Pteronotropis merlini* clade.

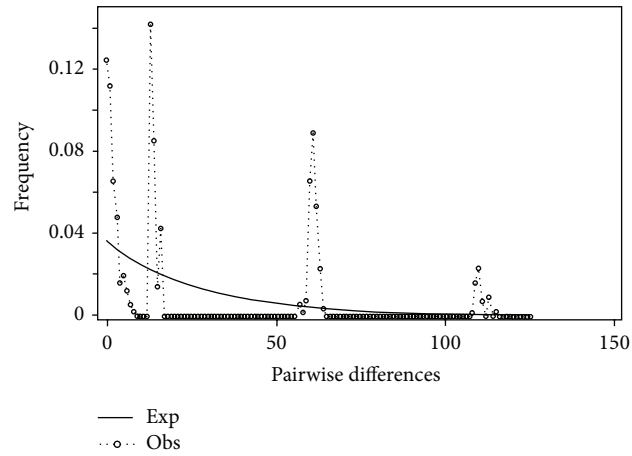


Figure 21: Mismatch distribution of the *Pteronotropis hypselopterus* clade.

for this, the clades that showed well-supported subclades (*P. signipinnis*, *P. metallicus,* and *P. hypselopterus*) were analyzed as partitioned data sets. Within this framework the western *signipinnis* clade displayed a unimodal distribution, corroborating population expansion (Figure 13), a hypothesis also supported by significance for Tajima's *D* test (Table 2).

5. Discussion

5.1. Overall Phylogeographic Structure. Findings in this study were consistent with previous analyses rejecting the monophyly of *Pteronotropis* if *P. harperi* continues to be placed in *Notropis* [8, 25–28]. In the BA tree, two major clades were resolved for the genus. The first included *P. hubbsi*, *P. welaka*, and *P. harperi*, the *Pteronotropis harperi* clade. The second contained *P. signipinnis* as the basal sister group *P. euryzonus* and members of the *P. hypselopterus* complex and *P. hypselopterus* (Figure 2), the *P. hypselopterus* clade. These relationships were well supported except for Node A (Figures 2 and 3); these three clades above this node, with this dataset, should be considered a trichotomy. These sister group relationships were also recovered using two nuclear gene loci by the authors [8]. Given the understanding of species relationships revealed by the BA tree and two nuclear gene loci, the discussion below focuses on phylogeographic patterns within *P. signipinnis*, *P. euryzonus*, and the *P. hypselopterus* complex.

One of the most important discoveries to come out of vicariance biogeography was the realization that a broad range of taxa, within a defined geographic area, will show similar distribution patterns, and likely similar sister group relationships, due to shared historical geological events isolating common ancestors across clades in the region. These geological events that impede gene flow and lead to strong intraspecific breaks among populations eventually lead to lineage splitting or cladogenesis. These replicated patterns are useful in comparative and evolutionary biology because they provide a null hypothesis or strong *a priori* predictions for unsampled taxa with similar distributions in a given area. Further, they can aid in conservation and management by delineating evolutionary significant units (ESUs) or cryptic species that might otherwise go unnoticed. The Bayesian tree structure (Figure 2) supports many phylogeographic patterns seen by other authors in the Coastal Plain of the southeastern United States, as well as some likely undescribed species.

Suttkus and Mettee [7] hypothesized that *P. signipinnis* had its origin in the upper Tombigbee system. The data presented here cannot refute this hypothesis. *Pteronotropis signipinnis* is resolved as sister to *P. euryzonus* and the *P. hypselopterus* complex (Figure 2) with high (17.4–19.5% uncorrected; 18.0–22.6% corrected) sequence divergence between it and the other species. Assuming the rate of evolution in the ND2 gene is similar between log perch darters (Percidae: *Percina*) and *Pteronotropis*, 2.0% per million years, as calculated by Near and Benard [43] and used in another study on minnows by Berendzen et al. [44], this would give a minimum divergence time for *P. signipinnis* of 22.6 million years before present (MYBP), roughly in the early Miocene. After an abrupt lowering of sea levels during the late Oligocene, sea levels rose rapidly (80–100 meters above present level) in the early Miocene and continued at this level, with some minor drops, for much of the Miocene [4]. Two of the rivers along the Gulf Slope that would have remained distinct and have separate outflows to the Gulf of Mexico at that time were the Tombigbee and Chattahoochee rivers. One scenario is that a contiguous ancestral population was split into two populations by the rise in sea level, leaving two disjunct populations, one in the then Tombigbee River and the other in the then Chattahoochee River. Presumably the ancestral populations in the Tombigbee River, through lineage splitting, eventually gave rise to *P. signipinnis* while the populations in the Chattahoochee River were ancestral to *P. euryzonus* plus the *hypselopterus* complex [7] and underwent subsequent isolation and divergence. Following another lowering of sea levels during this global cycle, dispersion of ancestral forms down these rivers that were now connected further out from the present coastline in the Gulf Slope would have provided for their movement to other river systems as habitat became available. This hypothesized historical biogeographic scenario involving both dispersal and vicariance is further supported by the fact that *P. euryzonus* is resolved, with high support, as the sister group to the diverse *P. hypselopterus* complex (Figure 2), and remains endemic to the middle and lower Chattahoochee River.

During the high sea stands of the Miocene, the Chattahoochee River is thought to have served as refugial habitat for populations of many freshwater fishes of the Gulf Slope resulting from inundation of waterways by the Gulf of Mexico and Atlantic Ocean [4]. Both the Chattahoochee and Flint rivers, as part of the same drainage system, flow in a southerly direction into Lake Seminole; the Apalachicola emerges from Lake Seminole and discharges into the Gulf of Mexico. This river system is known to have been and continues to be a major barrier to gene flow for populations of the same species with distributions to the east and west of the Apalachicola River. The distinctiveness of the diversity and divergence east and west of this drainage is renowned, so much so that Avise [1] devoted a large portion of his chapter on geological concordance to describing its history. Some of the organisms that show this break include reptiles [45], amphibians [20–46], fishes [3, 4, 47–50], macroinvertebrates [41], spiders [51], and trees [52]. The data presented here further corroborate these multilineage findings.

Within *Pteronotropis*, a clade containing *P. stonei*, *P. metallicus*, and *P.* sp. cf. *metallicus* is resolved as sister to clade wherein undescribed species may exist as separate lineages and where *P. merlini* + *P. hypselopterus* are sister to *P. grandipinnis*. All populations of *P. stonei* and *P. metallicus* occur to the east of the Apalachicola, *P. grandipinnis* occurs in the Apalachicola, and *P. merlini* and *P. hypselopterus* occur west of this drainage. The nodal support for the sister group relationship in this east-west split between *P. grandipinnis* and the *P. merlini* + *P. hypselopterus* clade is not strong (PP = 0.53; Node A, Figure 2); however, the nodes supporting the sister group relationship between P. grandipinnis plus *P.* sp. cf. *hypselopterus* as well as that for *P. merlini* + *P.* sp. cf. *hypselopterus* and *P. hypselopterus* are strongly supported by nuclear genes (PP = 1.0) (Figure 2 [8]).

Another evidence in support of *P. grandipinnis* being closely allied with the western group (*P. merlini* + *P.* sp. cf. *hypselopterus* and *P. hypselopterus*) comes from morphological data. Suttkus et al. [24] diagnosed *P. stonei* (eastern group) as having a dark lateral band continuous to the base of the caudal fin without any intensification at the base of the caudal fin, a ventral margin of the lateral band with a clearly defined border, and nuptial males of *P. stonei* lacking enlarged dorsal and anal fins. This is in contrast to *P. grandipinnis* and others of the western group which show an intensification in pigment in their lateral band at the caudal fin, a diffuse ventral margin of the lateral band, and slightly to greatly elevated dorsal and anal fins in nuptial males (especially in *P. grandipinnis*). As outlined by Avise [1] agreement between gene trees and other biogeographic data (aspect IV, genealogical concordance) provides assurance that gene trees can register these phylogeographic breaks.

5.2. Pteronotropis signipinnis Clade. All analyses of this group indicate that cryptic species diversity exists within *P. signipinnis*. Two reciprocally monophyletic clades are recovered (each with 100% bootstrap support and PP = 1.0) (Figures 2 and 3) within this species. Further, there is highly significant sequence divergence (10.7%), haplotype diversity, and nucleotide diversity (Table 2) within this clade. *Pteronotropis* sp. cf. *signipinnis* from the eastern *signipinnis* clade is a distinct taxon under both the Evolutionary Species Concept as the theoretical concept [53] and the Phylogenetic Species Criterion as the operational method for discovering evolutionary species as lineages [54–60]. The divergence between *P. signipinnis* and *P.* sp. cf. *signipinnis* is consistent with other taxa inhabiting the same area [3, 23, 61]. Bailey and Suttkus [62], in their description of *P. signipinnis*, observed differences in meristics between populations on either side of the Mobile Bay.

Higher levels of genetic variation in both haplotype and nucleotide diversity were also observed in *P.* sp. cf. *signipinnis* relative to *P. signipinnis* (Table 2). Most individuals in the *P. signipinnis* clade share a common haplotype, but those in *P.* sp. cf. *signipinnis* possess many unique haplotypes (Figure 9). At least three possible historical events may account for the lack of genetic variation within *P. signipinnis*. It could be the result of a genetic bottleneck or recent populations in and to the west of the Mobile Bay experienced a rapid population decline leading to the rapid fixation of only a few haplotypes. Given the current data these two alternative hypotheses cannot be differentiated or tested. Possibly a more plausible scenario, based upon the differences between populations east and west of the Mobile Bay, is that of founder affect. As shown in the mismatch distribution plot (Figure 14), *P. signipinnis* has undergone a recent range expansion, supported by a significant value for Tajima's *D* (Table 2). Genetic diversity in comparisons of individuals from the Mobile R. and Bay is notable with thirteen mutational steps (Figure 9). In this example, one possibility is that a haplotype yet to be found in the Mobile rivers was transported via dispersal, through stream captures or some other means, to the Escatawapa or Pascagoula rivers as habitat became available during a post-Pleistocene lowering of sea levels. This explanation is not without merit as many studies have demonstrated that mitochondrial genes are good markers for detecting range expansion of species or populations [15, 41, 63–67]. One interesting aspect of the data and analyses presented herein, however, is that no other clade within *Pteronotropis* shows evidence of a recent population expansion (Figures 12-13, 15–21). This seems to indicate that the western dispersal of *P. signipinnis* was much more recent than the possible migration of other members in the genus. Several studies of North American fish species have detected rapid range expansion in the examination of taxa from rivers of the Central Highlands, all presumably following glacial fronts and colonizing rivers in glaciated areas of the Central Lowlands as glaciers moved northward [15, 42, 63–69]. Conditions in Gulf Slope rivers, however, were quite different than those that once impacted the cold-water rivers of the Central Highlands. Swift et al. [4] predicted that if taxa from rivers of the Gulf Slope dispersed it would have been at the beginning (not the end, as in Central Highland taxa) of the Pleistocene, a difference of about 1.81 million years. A very real possibility exists that the mitochondrial ND2 gene marker does not possess adequate variation in *Pteronotropis* to detect early population expansions. To address this question would require examination of microsatellites amongst many populations.

5.3. Pteronotropis euryzonus Clade. Having strong lineage support (BS = 100%, PP = 1.0) *Pteronotropis euryzonus* is resolved as the basal sister group to this *P. hypselopterus* clade (Figure 2), a relationship corroborated in other studies using either morphology or molecular data [70–72], but not in the nuclear gene phylogeny of Mayden and Allen [8] where *P. euryzonus* is resolved as the sister species to a clade consisting of *P. metallicus* + *P. stonei*. In the description of *P. euryzonus*, Suttkus [73] described two morphological races within this species, a northern race termed the Uchee race and a southern Chattahoochee race from the lower portions of that river. Data presented here can neither corroborate nor refute this hypothesis because samples used herein were from the upper portions of the Chattahoochee River drainage. Warren et al. [74] listed *P. euryzonus* as a species of special concern, due to its limited range, and the results of this study indicate little genetic variation within the sampled populations. All individuals of the three populations examined from Snake, Maringo, and Uchee creeks have the same haplotype, and a within species sequence divergence of 0%, a haplotype diversity of 0.464, and a nucleotide diversity of 0.00 (Tables 2 and 3). These results emphasize the recommendations of Boschung and Mayden [21] that periodic population monitoring of known localities of this species should be a priority.

5.4. Pteronotropis stonei Clade. This clade is found in streams draining the Atlantic Slope, including the Pee Dee, Santee, North and South Fork Edisto, Combahee, and Savannah rivers. Suttkus et al. [24] elevated *P. stonei* from synonomy of *P. hypselopterus* and hypothesized that the species was closely related to *P. metallicus* but provided no phylogenetic evidence. Molecular variation and analyses presented here corroborate this hypothesis (Figure 2), because *P. stonei* and

Table 3: Corrected percent divergence values with standard errors (above diagonal) and pairwise differences with standard errors (below diagonal) for the *Pteronotropis hypselopterus* complex, *P. euryzonus,* and *P. signipinnis* for ND2.

Clade		1	2	3	4	5	6	7
1	*P. hypselopterus*	**4.8 ± 1.2**	22.6 ± 4.5	8.2 ± 2.2	9.4 ± 2.5	15.3 ± 3.6	10.4 ± 3.2	11.8 ± 2.6
2	*P. signipinnis*	17.4 ± 2.7	**10.7 ± 2.6**	18.4 ± 2.8	18.3 ± 2.8	19.5 ± 2.8	18.1 ± 2.8	18.0 ± 2.5
3	*P. merlini*	7.0 ± 1.6	18.4 ± 2.8	**1.6 ± 0.7**	8.5 ± 2.2	10.3 ± 2.2	8.5 ± 2.1	9.2 ± 2.5
4	*P. grandipinnis*	8.2 ± 2.1	18.3 ± 2.8	8.5 ± 2.2	**0.7 ± 0.4**	10.6 ± 2.3	6.4 ± 1.9	10.2 ± 2.3
5	*P. stonei*	12.4 ± 2.4	19.5 ± 2.8	10.3 ± 2.2	10.6 ± 2.3	**03.3 ± 1.0**	10.1 ± 2.2	9.6 ± 1.8
6	*P. euryzonus*	8.5 ± 2.0	18.1 ± 2.8	8.5 ± 2.1	6.4 ± 1.9	10.0 ± 2.2	**0.0 ± 0.0**	9.0 ± 1.9
7	*P. metallicus*	9.8 ± 1.9	18.0 ± 1.9	9.2 ± 1.9	10.2 ± 2.3	9.6 ± 1.8	9.0 ± 1.9	**7.7 ± 2.0**

Bold face values indicate within group variation.

P. metallicus are reciprocally monophyletic and sister species herein and with nuclear genes [8]. The *P. stonei* clade and some subclades of the gene tree were highly supported (BS = 100%, PP = 1.0; Figure 4).

Mitochondrial variation supports the hypothesis of range expansion and speciation within this complex originating and centered about the Chattahoochee and Apalachicola river systems. Using the logic of molecular divergence employed above, sequence divergence between *P. euryzonus* and *P. stonei* (10.1%, Table 3) would support a divergence time of about 10.1 MYA. Assuming a 2% molecular clock, this places this speciation event at mid-Miocene during a period of rising sea levels [4], potentially isolating populations in headwater streams and opportunity for divergence. The close level of divergence of *P. stonei* and *P. metallicus* (9.6%) from *P. euryzonus* (10.1%) would suggest that the ancestral populations had already spread eastward into the Ochlocknee and across the Atlantic coastal streams, thus creating multiple, refugial populations in the ancestor. The lowering of sea level during the late Miocene altered drainage patterns in their lower reaches and resulted in the connection of multiple formerly isolated basins on the continental shelf. This expanded the coastal plain, creating habitats identical to those currently inhabited by species of *Pteronotropis* and is herein hypothesized to have provided for a northern expansion of populations. This is consistent with mismatch analysis, and such an expansion may have occurred via the Tifton uplift in southern Georgia or the Ocala uplift in northern Florida, both thought to have been distinct since the Eocene [22]. Opportunities for other taxa to move into eastern streams may have existed at about this time. For example, populations of *Pteronotropis welaka*, *Lepisosteus oculatus*, and *Opsopoeodus emiliae* occur on both sides of the Apalachicola River, but their ancestral populations are thought to have existed west of this drainage [7]. For instance, *P. welaka* is the sister species of *P. hubbsi,* an endemic known in the Mississippi River valley from southern Illinois (now extirpated) and in the Little and Ouachita rivers in southern Arkansas and northern Louisiana, west of the Apalachicola. Studies of other freshwater fish species (Near et al. [75]; Roe et al. [61]) from these same drainages of the Gulf Slope have estimated similar speciation dates.

5.5. Pteronotropis metallicus Clade. The *P. metallicus* clade (Figure 5) contains two strongly supported major reciprocally monophyletic subclades (100% bootstrap, PP = 1.0; Figure 5). Overall within species sequence divergence in *P. metallicus* is 7.7% and is largely due to the presence of the two major subclades (Figure 5(b)). With the high support for these subclades and the large amount of sequence divergence between them, invoking the Phylogenetic Species criterion under the Evolutionary Species Concept as an overriding concept [76], we recognize this lineage as an undescribed species (Figure 5(a), St. Johns subclade). This divergence was also discussed by Suttkus [73] but the lineage was not officially named. The species can also be diagnosed using morphological traits [73] and was originally thought to occur in the Withlacoochee, Hillsborough, St. Marys, and St. Johns rivers. Subsequently, all populations in St. Marys River were referred to *P. metallicus* [24]. Suttkus [73] noted distinct populations of *P. hypselopterus* from the Alafia River (see Figure 5(a)) and recommended this as a distinct subspecies. The hypothesized diversity identified by Suttkus [73], herein recognized as species, is supported by current molecular data and analyses. Within the St. Johns subclade (Figure 5(a)) least two distinct genetic and morphologically diagnosable lineages exist, one in the Alafia River system (100% BS, PP = 1.0) and one in the St. Johns River system (PP = 0.96; Figures 5 and 10). TCS analysis (Figure 10) identifies populations in the Alafia River system as being fifteen mutational steps away from populations from the St. Johns River system. Separation of populations from the Alafia River from populations in the St. Johns River is predicted to have been fairly recent as TCS analysis can connect these two populations within a 95% connection limit.

5.6. Pteronotropis grandipinnis—"P. hypselopterus" Clade. Pteronotropis grandipinnis is sister to a clade inclusive of populations of *P. merlini* plus some populations of *P. hypselopterus* (Figure 2 node A and Figure 6). In this clade *P. hypselopterus* is not resolved as monophyletic as individuals of *P. hypselopterus* from St. Andrews Bay and Choctawhatchee Bay drainages were resolved as a subclade (100% bootstrap, PP = 1.0) within *Pteronotropis grandipinnis*. Constraining the gene tree of *P. hypselopterus* or *P.*

grandipinnis as monophyletic groups, respectively, resulted in significantly worse ML tree scores than the best trees (P = 0.0002) using the Shimodaira and Hasegawa [77] test. This clade of specimens of "*P. hypselopterus*" from these drainages recovered within *P. grandipinnis* may be either an instance where the gene tree does not accurately reflect the species tree or clade or these "*P. hypselopterus*" represent an undescribed species (*sensu* the Evolutionary Species Concept [53–55, 57–59, 76]). With regard to the first possibility, two potential explanations may account for the pattern seen in this clade, introgression between *P. hypselopterus* and *P. grandipinnis*, or incomplete lineage sorting of haplotypes within an ancestral species having an ancestral polymorphism [78]. Given the current data, it is impossible to distinguish between these alternatives. However, given that the *P. hypselopterus* clade has high support it likely indicates that this group of "*P.* sp. cf. *hypselopterus*" represents a different undescribed species. Currently the headwater tributaries of the St. Johns Bay river system and the Apalachicola River system are very close in air miles. For instance, the authors collected many individuals of *P. hypselopterus* at Bear Creek (tributary to St. Johns Bay drainage), which has its headwaters less than two air miles from the headwaters of Juniper Creek, a tributary of the Chipola-Apalachicola Rivers. Because both of these creeks flow through lowland cypress swamps, it is possible that these systems were connected one or more times in the evolution of this lineage (Figure 6). Given that the gene tree for the *P*. sp. cf. *hypselopterus* is highly divergent and monophyletic and the more basal specimens from the Apalachicola River have limited resolution, it is likely that this gene provides only some resolution and that other genetic markers are needed. It is possible that the populations from the Apalachicola River, *P. grandipinnis*, is a natural grouping with a monophyletic gene tree. Further analyses using alternative genes and finer scale markers such as microsatellites are needed for further resolution to aid in distinguishing between alternative explanations.

5.7. Pteronotropis merlini—"P. hypselopterus" Clade. As with the evolutionary relationships among populations of *P. grandipinnis,* the gene tree for *P. merlini* did not resolve all specimens of this species as closest relatives. Rather, some haplotypes of specimens of *P. merlini* were recovered as being more closely related to specimens of *P. hypselopterus* from the Choctawhatchee River drainage than to other *P. merlini* from the same drainage (Figure 7). While these relationships were resolved, there are no supporting values for basal nodes and some nodes between *P. merlini* and *P. hypselopterus* from the Choctawhatchee River; however, some nodes supporting monophyly of the gene tree for many individuals from the Choctawhatchee River are strongly supported. Constraining gene trees for *P. hypselopterus* or *P. merlini* as monophyletic resulted in significantly worse ML tree scores than the best trees (P = 0.0002) using the Shimodaira and Hasegawa [77] test. Some may question the validity of *P. merlini* due to its lack of genetic distinctiveness and its possible lack of evolutionarily independence from *P. hypselopterus* in the Choctawhatchee River Basin for the mitochondrial gene ND2. In this situation, unlike that in *P. grandipinnis,* the haplotypes of these *P. hypselopterus* are not clustered into a highly supported clade but are dispersed (Figure 7). Analyses do strongly support the monophyly of the gene tree uniting *P. merlini* and *P. hypselopterus* from the Choctawhatchee River, clearly indicating that these *P. hypselopterus* are not closely related to the others species occurring in different clades. Testing the relatedness of these populations and the monophyly of the gene tree for *P. merlini* requires additional genes and would benefit from microsatellite analyses. It is possible that the gene tree resolved in this pattern can be explained without invoking an active process being involved within the Choctawhatchee River drainages following the most recent common ancestor of the *P. grandipinnis* plus "*P. hypselopterus*" clade. Other than the simple lack of resolution using ND2 sequence variation, active process-free explanations following the divergence could be result of either incomplete lineage sorting in a shared ancestral population to both *P. merlini* and "*P. hypselopterus*" from the Choctawhatchee River or specimens/populations of *P. hypselopterus* from the geographic area in question retaining haplotype polymorphisms in their most recent common ancestor.

Pteronotropis merlini inhabits more upland habitats in this drainage, and *P. hypselopterus* occurs in more lowland habitats below the confluence with the Pea River [7], ecological and behavioral predispositions that may limit their geographic overlap. Further, morphological characteristics exist to distinguish the two species, including differences in body depth (*P. merlini* has a deeper body), orange coloration in the caudal fin of *P. merlini* versus olive-yellow coloration in *P. hypselopterus*, and the chevron-lunate shaped blotch on the caudal fin separated from the dark lateral band in *P. merlini* that is lacking in *P. hypselopterus*. These features argue for the independence of the two groups from this region and do serve as counter evidence for any ongoing gene flow, although morphology can be a poor surrogate for evidence of gene exchange [79–82]. Upon close examination no morphological intermediates have yet to be found between these two species *P. merlini* and *P.* sp. cf. *hypselopterus*. Further, no specimens morphologically identifiable as *P. merlini* have ever been taken downstream of the confluence of the Pea and Choctawhatchee rivers nor have any specimens morphologically identifiable as *P. hypselopterus* been taken upstream of this confluence. Additional morphological and more fine-scaled molecular data are needed in appropriate analyses to examine the possibility that populations of *P.* sp. cf. *hypselopterus* from the Choctawhatchee River drainage do not represent a distinct lineage. As multiple new species have been described or detected across the widespread distribution of the formerly recognized *P. hypselopterus*, additional study using different markers of varying degree of potential anagenesis and detailed morphological study remain as possible tests to the hypothesis of the two lineages in the Choctawhatchee River.

5.8. Pteronotropis hypselopterus Clade. In no gene tree of ND2 was a clade composed exclusively of currently recognized populations of *P. hypselopterus* resolved as monophyletic (Figures 2, 6, and 7). Some populations were found

to be more closely related to *P. grandipinnis* (Figure 6) or *P. merlini* (Figure 7). For most specimens of *P. hypselopterus* gene tree analysis identified a strongly supported (100%) monophyletic group with two well-supported and geographically defined independent groups with their distributions being east and west of the Mobile Bay (Figures 8 and 11), much like the pattern and relationships observed in *P. signipinnis*.

Other taxa have their distributional limits delineated east and west of the Mobile Bay [3, 25, 62]. The clear distinctiveness of taxa on either side of the Mobile Bay has been explained by elevated sea levels that isolated populations of species in the headwaters of rivers east and west of the Mobile Bay. After the subsequent lowering of sea levels during the mid- to late-Miocene, drainage flow of the Alabama-Tombigbee Rivers turned southward (from west or southwestward) further isolating populations on either side of the bay [3, 59]. However, these historical events do not fit the time signature seen in *P. hypselopterus*, if a constant molecular time divergence assumption hypothesis is valid. Assuming a molecular clock of 2% sequence divergence per million years [4, 44] the 4.8% within sequence divergence observed in *P. hypselopterus* would correspond to ~4.8 MYBP or roughly in the mid-Pliocene.

Haplotype diversity and structure differ between the eastern and western clades of *P. hypselopterus*. The eastern *P. hypselopterus* clade has high haplotype diversity (Table 2) but shows little genetic structuring relative to the hierarchical structuring of drainages (Figure 11). Many haplotypes are shared between the Escambia, Yellow, and Blackwater rivers. These river systems have few endemics and other freshwater fishes show similar distributional patterns in these systems [3]. The western clade of *P. hypselopterus* possesses two main haplotype clusters (Figure 11). One cluster includes haplotypes from the Perdido River group with those of the Mobile, Alabama, and Tombigbee rivers; the other cluster includes haplotypes from populations in the Fish and parts of the Mobile and Tombigbee rivers. Although two clusters of haplotypes occur in the western *P. hypselopterus* clade, they differ only by a single mutation, as indicated by low haplotype and nucleotide diversity (Table 2). Due to the low degree of within sequence divergence in the *P. hypselopterus* clade (compared with similar taxa in the region) and the apparent limited morphological distinctiveness between populations [7], we recommend no taxonomic changes. However, these populations warrant further study with additional more highly variable genetic markers and more detailed examination of both museum and live and breeding adults from all of the rivers to resolve potential lineage divergence or mixing within this clade. The limited divergence patterns observed in this clade and between the eastern and western clades may be due to recent divergences, a mismatch of appropriate genes and lack of detailed morphological studies of coloration of live and breeding adults, or simply a depressed rate of anagenesis in the *P. hypselopterus* lineage (excluding those populations that are more closely related to either *P. grandipinnis* or *P. merlini*).

6. Conclusions

Phylogenetic analysis of populations and species of *Pteronotropis* reveal multiple new hypotheses regarding the monophyly of genes, species diversity, potential undescribed species, and abiotic factors correlated with divergence events between and within species. These findings fully support those of earlier studies (Suttkus and Mettee [7], Suttkus et al. [24], Bailey and Suttkus [62], and Suttkus [73]) which were all based on morphological data. *Pteronotropis* is a monophyletic genus but only with the inclusion of *Pteronotropis harperi*, a species that has long had unresolved relationships. Findings and hypotheses herein also complement previous studies of fish diversity and biogeography in rivers of the Gulf and Atlantic slopes in the southeastern United States. As such, this group adds to the multiple other groups of aquatic organisms in this region for future comparative biogeographic analyses, but only if different clades that are being compared are of the same ages of divergences as determined by time-tree analyses or known abiotic factors. Comparisons of relationships in clades that diverged at different times conflate the comparative analysis and will likely lead to conflicting relationships with unknown reasons. Comparative time analyses are thus critical in future biogeographic studies of this group and others.

The phylogeographic patterns observed in species of *Pteronotropis* derived from phylogenetic analysis are largely consistent with previous studies of freshwater taxa inhabiting rivers occupied by members of *Pteronotropis*. The *P. hypselopterus* complex is widespread across the Gulf and Atlantic slopes and was once considered a single species. The status of this species changed with the elevation and descriptions of species (Suttkus and Mettee [7]) within the complex. Genetic data and analyses presented herein support the recognition of the species within this complex, as well as the need to recognize additional species, with the possible exception of the *P. merlini*, "*P. hypselopterus*" clade. These taxa will also likely possess additional diversity if examined more closely for morphological and genetic variation, as well as coloration of live breeding adults.

The southeastern fish fauna is second only to the Mississippi River drainage in terms of species diversity [21] and the challenge for taxonomists and systematists is to find and describe diversity before the extirpation of populations and species from these drainages that have been in isolation for millions of years. The detailed resolution and understanding of the phylogeography of species of *Pteronotropis* provide insights into the historical and contemporary events that were instrumental in the diversification of this group and offer insights into the importance of more dense sampling of any widespread taxa for clarity in diversification rates and patterns in a region.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank Brett Albanese, Bud Freeman, Nick Lang, Dave Neely, Larry Page, Brady Porter, Charles Ray, Chip Reinhart, Brian Skidmore, David Smith, and Dustin Smith who provided field assistance or valuable samples. The authors thank Susana Schönhuth and Anne Ilvarson for assistance with GenBank. This research was supported in part by NSF grants to Richard L. Mayden (EF 0431326 and DEB-1021840) and the William S. Barnickle Endowment at Saint Louis University.

References

[1] J. C. Avis, *Phylogeography: The History and Formation of Species*, Harvard University Press, Cambridge, Mass, USA, 2000.

[2] J. V. Crisci, L. Katinas, and P. Posadas, *Historical Biogeography: An Introduction*, Harvard University Press, Cambridge, Mass, USA, 2003.

[3] E. O. Wiley and R. L. Mayden, "Species and speciation in phylogenetic systematics, with examples from North American fish fauna," *Annals of the Missouri Botanical Garden*, vol. 72, no. 4, pp. 596–635, 1985.

[4] C. C. Swift, C. R. Gilbert, S. A. Bortone, and R. W. Yerger, "Zoogeography of the freshwater fishes of the southeastern United States: Savanna River to Lake Pontchartrain," in *The Zoogeography of North American Fishes*, C. H. Hocutt and E. O. Wiley, Eds., Wiley Interscience, New York, NY, USA, 1986.

[5] B. C. Nagle and A. M. Simons, "Rapid diversification in the North American minnow genus *Nocomis*," *Molecular Phylogenetics and Evolution*, vol. 63, no. 3, pp. 639–649, 2012.

[6] M. Sandel, *Evolutionary relationships and historical biogeography of pygmy sunfishes (Percomorphacea: Elassoma) [Ph.D. thesis]*, University of Alabama, Tuscaloosa, Ala, USA, 2012.

[7] R. D. Suttkus and M. F. Mettee, "Analysis of four species of *Notropis* included in the subgenus *Pteronotropis* fowler, with comments on relationships, origins, and dispersion," Geological Survey of Alabama Bulletin 170, Geological Survey of Alabama, 2001.

[8] R. L. Mayden and J. S. Allen, "Molecular systematics of the phoxinin genus *Pteronotropis* (Otophysi: Cypriniformes)," *BioMed Research International*. In press.

[9] A. R. Templeton, "Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history," *Molecular Ecology*, vol. 7, no. 4, pp. 381–397, 1998.

[10] L. Bernatchez and C. C. Wilson, "Comparative phylogeography of Nearctic and Palearctic fishes," *Molecular Ecology*, vol. 7, no. 4, pp. 431–452, 1998.

[11] L. Bernatchez, "The evolutionary history of brown trout (*Salmo trutta* L.) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation," *Evolution*, vol. 55, no. 2, pp. 351–379, 2001.

[12] A. R. Rogers and H. Harpending, "Population growth makes waves in the distribution of pairwise genetic differences," *Molecular Biology and Evolution*, vol. 9, no. 3, pp. 552–569, 1992.

[13] S. Schneider and L. Excoffier, "Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA," *Genetics*, vol. 152, no. 3, pp. 1079–1089, 1999.

[14] H. C. Harpending, M. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers, and S. T. Sherry, "Genetic traces of ancient demography," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 4, pp. 1961–1967, 1998.

[15] P. B. Berendzen, A. M. Simons, and R. M. Wood, "Phylogeography of the northern hogsucker, *Hypentelium nigricans* (Teleostei: Cypriniformes): genetic evidence for the existence of the ancient Teays River," *Journal of Biogeography*, vol. 30, no. 8, pp. 1139–1152, 2003.

[16] T. D. Kocher and K. L. Carleton, "Base substitution in fish mitochondrial DNA: patterns and rates," in *Molecular Systematics of Fishes*, T. D. Kocher and C. A. Stepien, Eds., Academic Press, San Diego, Calif, USA, 1997.

[17] W. Chen and R. L. Mayden, "Molecular systematics of the Cyprinoidea (Teleostei: Cypriniformes), the world's largest clade of freshwater fishes: further evidence from six nuclear genes," *Molecular Phylogenetics and Evolution*, vol. 52, no. 2, pp. 544–549, 2009.

[18] R. L. Mayden and W. J. Chen, "The world's smallest vertebrate species of the genus *Paedocypris*: a new family of freshwater fishes and the sister group to the world's most diverse clade of freshwater fishes (Teleostei: Cypriniformes)," *Molecular Phylogenetics and Evolution*, vol. 57, no. 1, pp. 152–175, 2010.

[19] E. Bermingham and J. C. Avise, "Molecular zoogeography of freshwater fishes in the southeastern United States," *Genetics*, vol. 113, no. 4, pp. 939–965, 1986.

[20] G. B. Pauly, O. Piskurek, and H. B. Shaffer, "Phylogeographic concordance in the southeastern United States: the flatwoods salamander, *Ambystoma cingulatum*, as a test case," *Molecular Ecology*, vol. 16, no. 2, pp. 415–429, 2007.

[21] H. T. Boschung and R. L. Mayden, *Fishes of Alabama*, Smithsonian Books, Washington, DC, USA, 2004.

[22] W. D. Thornbury, *Regional Geomorphology of the United States*, John Wiley and Sons, New York, NY, USA, 1965.

[23] M. E. Raley and R. M. Wood, "Molecular systematics of members of the *Notropis dorsalis* species group (Actinopterygii: Cyprinidae)," *Copeia*, vol. 2001, no. 3, pp. 638–645, 2001.

[24] R. D. Suttkus, B. A. Porter, and B. J. Freeman, "The status and infraspecific variation of *Notropis stonei* Fowleri," *Proceedings of the American Philosophical Society*, vol. 147, no. 4, pp. 354–376, 2003.

[25] A. M. Simons, P. B. Berendzen, and R. L. Mayden, "Molecular systematics of North American phoxinin genera (Actinopterygii: Cyprinidae) inferred from mitochondrial 12S and 16S ribosomal RNA sequences," *Zoological Journal of the Linnean Society*, vol. 139, no. 1, pp. 63–80, 2003.

[26] R. L. Mayden, A. M. Simons, R. M. Wood, P. M. Harris, and B. R. Kuhajda, "Molecular Systematics and classification of North American Notropin shiners and minnows (Cypriniformes: Cyprinidae)," in *Studies of North American Desert Fishes in Honor of E.P.(Phil) Pister, Conservationist*, M. D. L. Lozano-Vilano and A. J. Contreras-Balderas, Eds., Universidad Autonoma de Nuevo, Leon Monterrey, Mexico, 2006.

[27] A. P. Bufalino and R. L. Mayden, "Phylogenetic relationships of North American phoxinins (Actinopterygii: Cypriniformes: Leuciscidae) as inferred from S7 nuclear DNA sequences," *Molecular Phylogenetics and Evolution*, vol. 55, no. 1, pp. 143–152, 2010.

[28] A. P. Bufalino and R. L. Mayden, "Molecular phylogenetics of North American phoxinins (Actinopterygii: Cypriniformes: Leuciscidae) based on RAG1 and S7 nuclear DNA sequence data," *Molecular Phylogenetics and Evolution*, vol. 55, no. 1, pp. 274–283, 2010.

[29] R. M. Bailey and H. W. Robison, "*Notropis hubbsi*, a new cyprinid fish from the Mississippi River Basin, with comments on *Notropis welaka*," University of Michigan Museum of Zoology Occasional Papers 683, University of Michigan Museum of Zoology, 1978.

[30] R. L. Mayden, *Phylogenetic Studies of North American Minnows: With Emphasis on the Genus Cyprinella (Teleostei, Cypriniformes)*, vol. 80 of *Miscellaneous Publications, Museum of Natural History*, University of Kansas, Lawrence, Kan, USA, 1989.

[31] T. A. Hall, *BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis, Version 5. 0. 9*, Department of Microbiology, North Carolina State University, Raleigh, NC, USA, 2001.

[32] D. S. Sikes and P. O. Lewis, *PAUPRat: PAUP* Implementation of the Parsimony Ratchet*, Beta Software, Version 1, Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Conn, USA, 2001.

[33] D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods), Version 4*, Sinauer Associates, Sunderland, Mass, USA, 2003.

[34] D. Zwickl, *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the Maximum Likelihood Criterion [Ph.D. thesis]*, University of Texas at Austin, Austin, Tex, USA, 2006.

[35] D. Zwickl, *Garli. Genetic Algorithm for Rapid Likelihood Inference. Version 0. 951*, 2006, http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html.

[36] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[37] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[38] S. Kumar, S. K. Tamura, and M. Nei, "MEGA3: integrated software for molecular evolutionary genetic analysis and sequence alignment," *Briefings in Bioinformatics*, vol. 5, no. 2, pp. 150–163, 2004.

[39] A. R. Templeton, E. Boerwinkle, and C. F. Sing, "A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*," *Genetics*, vol. 117, no. 2, pp. 343–351, 1987.

[40] A. R. Templeton and C. F. Sing, "A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination," *Genetics*, vol. 134, no. 2, pp. 659–669, 1993.

[41] M. Clement, D. Posada, and K. A. Crandall, "TCS: a computer program to estimate gene genealogies," *Molecular Ecology*, vol. 9, no. 10, pp. 1657–1659, 2000.

[42] J. Rozas and R. Rozas, "DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis," *Bioinformatics*, vol. 15, no. 2, pp. 174–175, 1999.

[43] T. J. Near and M. F. Benard, "Rapid allopatric speciation in logperch darters (Percidae: *Percina*)," *Evolution*, vol. 58, no. 12, pp. 2798–2808, 2004.

[44] P. B. Berendzen, T. Gamble, and A. M. Simons, "Phylogeography of the bigeye chub *Hybopsis amblops* (Teleostei: Cypriniformes): early Pleistocene diversification and post-glacial range expansion," *Journal of Fish Biology*, vol. 73, no. 8, pp. 2021–2039, 2008.

[45] R. M. Bailey, "An annotated checklist and biogeographic analysis of the insular herpetofauna of the Apalachicola region, Florida," *Herpetologica*, vol. 27, no. 4, pp. 406–430, 1971.

[46] D. B. Means, "Aspects of the significance to terrestrial vertebrates of the Apalachicola River Drainage Basin, Florida," *Florida Marine Research Publications*, vol. 26, pp. 37–67, 1977.

[47] R. E. Jenkins, *Systematic studies of the catostomid fish tribe Moxostomatini [Ph.D. thesis]*, Cornell University, New York, NY, USA, 1970.

[48] A. Y. Kristmundsdóttir and J. R. Gold, "Systematics of the blacktail shiner (*Cyprinella venusta*) inferred from analysis of mitochondrial DNA," *Copeia*, vol. 1996, no. 4, pp. 773–783, 1996.

[49] B. M. Burr and R. C. Cashner, "*Campostoma pauciradii*, a new cyprinid fish from southeastern United States, with a review of related forms," *Copeia*, vol. 1983, pp. 101–116, 1983.

[50] D. A. Neely, J. D. Williams, and R. L. Mayden, "Two new sculpins of the genus *Cottus* (Teleostei: Cottidae) from rivers of Eastern North America," *Copeia*, vol. 2007, no. 3, pp. 641–655, 2007.

[51] S. D. Marshall, W. R. Hoeh, and M. A. Deyrup, "Biogeography and conservation biology of Florida's *Geolycosa* wolf spiders: threatened spiders in endangered ecosystems," *Journal of Insect Conservation*, vol. 4, no. 1, pp. 11–21, 2000.

[52] N. G. Swenson and D. J. Howard, "Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America," *The American Naturalist*, vol. 166, no. 5, pp. 581–591, 2005.

[53] R. L. Mayden, "Consilience and a hierarchy of species concepts: advances toward closure on the species puzzle," *Journal of Nematology*, vol. 31, no. 2, pp. 95–116, 1999.

[54] R. L. Mayden, "On biological species, species concepts and individuation in the natural world," *Fish and Fisheries*, vol. 3, no. 3, pp. 171–196, 2002.

[55] R. L. Mayden, "Species, trees, characters, and concepts: ongoing issues, diverse ideologies, and a time for reflection and change," in *The Species Problem—Ongoing Issues*, I. Y. Pavlinov, Ed., pp. 171–191, InTech, 2013.

[56] Q. D. Wheeler and R. Meier, *Species Concepts and Phylogenetic Theory: A Debate*, Columbia University Press, New York, NY, USA, 2000.

[57] E. O. Wiley, *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*, Wiley-Interscience, New York, NY, USA, 1981.

[58] E. O. Wiley and R. L. Mayden, "The evolutionary species concept," in *Species Concepts and Phylogenetic Theory: A Debate*, Q. D. Wheeler and R. Meier, Eds., Columbia University Press, New York, NY, USA, 2000.

[59] E. O. Wiley and R. L. Mayden, "Comments on alternative species concepts," in *Species Concepts and Phylogenetic Theory: A Debate*, Q. D. Wheeler and R. Meier, Eds., Columbia University Press, New York, NY, USA, 2000.

[60] E. O. Wiley and R. L. Mayden, "A reply to our critics," in *Species Concepts and Phylogenetic Theory: A Debate*, Q. D. Wheeler and R. Meier, Eds., Columbia University Press, New York, NY, USA, 2000.

[61] K. J. Roe, R. L. Mayden, and P. M. Harris, "Systematics and zoogeography of the rock basses (Centrarchidae: Ambloplites)," *Copeia*, vol. 2008, no. 4, pp. 858–867, 2008.

[62] R. M. Bailey and R. D. Suttkus, "*Notropis signipinnis*, a new cyprinid fish from the southeastern United States," University of Michigan Museum of Zoology Occasional Papers 542, University of Michigan Museum of Zoology, 1952.

[63] T. J. Near, L. M. Page, and R. L. Mayden, "Intraspecific phylogeography of *Percina evides* (Percidae: Etheostomatinae): an additional test of the Central Highlands pre-Pleistocene vicariance hypothesis," *Molecular Ecology*, vol. 10, no. 9, pp. 2235–2240, 2001.

[64] M. E. Hardy, J. M. Grady, and E. J. Routman, "Intraspecific phylogeography of the slender madtom: the complex evolutionary history of the Central Highlands of the United States," *Molecular Ecology*, vol. 11, no. 11, pp. 2393–2403, 2002.

[65] J. F. Switzer, *Molecular systematics and phylogeography of the Etheostoma variatum species group (Actinopterygii: Percidae) [Ph.D. thesis]*, Saint Louis University, St. Louis, Mo, USA, 2004.

[66] T. J. Near and B. P. Keck, "Dispersal, vicariance, and timing of diversification in *Nothonotus* darters," *Molecular Ecology*, vol. 14, no. 11, pp. 3485–3496, 2005.

[67] J. M. Ray, R. M. Wood, and A. M. Simons, "Phylogeography and post-glacial colonization patterns of the rainbow darter, *Etheostoma caeruleum* (Teleostei: Percidae)," *Journal of Biogeography*, vol. 33, no. 9, pp. 1550–1558, 2006.

[68] P. B. Berendzen, A. M. Simons, R. M. Wood, T. E. Dowling, and C. L. Secor, "Recovering cryptic diversity and ancient drainage patterns in eastern North America: historical biogeography of the *Notropis rubellus* species group (Teleostei: Cypriniformes)," *Molecular Phylogenetics and Evolution*, vol. 46, no. 2, pp. 721–737, 2008.

[69] H. Dominik and A. M. Simons, "Cryptic speciation reversal in the *Etheostoma zonale* (Teleostei: Percidae) species group, with an examination of the effect of recombination and introgression on species tree inference," *Molecular Phylogenetic and Evolution*, vol. 70, no. 1, pp. 13–28, 2014.

[70] R. L. Mayden, *Phylogenetic studies of North American minnows, with emphasis on the genus Cyprinella (Teleostei: Cypriniformes) [Ph.D. thesis]*, University of Kansas, Lawrence, Kan, USA, 1985.

[71] W. W. Dimmick and R. Lawson, "Phylogenetic relationships of members of the genus *Pteronotropis* inferred from parsimony analysis of allozymic and morphological data (Cyprinidae: Cypriniformes)," *Biochemical Systematics and Ecology*, vol. 19, no. 5, pp. 413–419, 1991.

[72] A. M. Simons, K. E. Knott, and R. L. Mayden, "Assessment of monophyly of the minnow genus *Pteronotropis* (Teleostei: Cyprinidae)," *Copeia*, vol. 2000, no. 4, pp. 1068–1075, 2000.

[73] R. D. Suttkus, *A taxonomic study of five cyprinid fishes related to Notropis hypselopterus of southeastern United States [Ph.D. thesis]*, Cornell University, New York, NY, USA, 1950.

[74] M. L. Warren, B. M. Burr, S. J. Walsh et al., "Diversity, distribution, and conservation status of the native freshwater fishes of the southern United States," *Fisheries*, vol. 25, no. 10, pp. 7–31, 2000.

[75] T. J. Near, T. W. Kassler, J. B. Koppelman, C. B. Dillman, and D. P. Philipp, "Speciation in North American black basses, *Micropterus* (Actinopterygii: Centrarchidae)," *Evolution*, vol. 57, no. 7, pp. 1610–1621, 2003.

[76] R. L. Mayden, "A hierarchy of species concepts: the denouement in the saga of species problem," in *Species: The Units of Biodiversity*, M. F. Claridge, H. A. Dawah, and M. R. Wilson, Eds., Chapman and Hall, London, UK, 1997.

[77] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Molecular Biology and Evolution*, vol. 16, no. 8, pp. 1114–1116, 1999.

[78] W. P. Maddison, "Gene trees in species trees," *Systematic Biology*, vol. 46, no. 3, pp. 523–536, 1997.

[79] D. J. Funk and K. E. Omland, "Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA," *Annual Review of Ecology, Evolution, and Systematics*, vol. 34, pp. 397–423, 2003.

[80] J. C. Avise, *Molecular Markers, Natural History, and Evolution*, Sinauer Associates, Sunderland, Mass, USA, 2nd edition, 2004.

[81] J. M. Rhymer and D. Simberloff, "Extinction by hybridization and introgression," *Annual Review of Ecology and Systematics*, vol. 27, pp. 83–109, 1996.

[82] T. Sang and Y. Zhong, "Testing hybridization hypotheses based on incongruent gene trees," *Systematic Biology*, vol. 49, no. 3, pp. 422–434, 2000.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 315312, 8 pages
http://dx.doi.org/10.1155/2015/315312

Research Article

Structural and Population Polymorphism of RT-Like Sequences in Avian Schistosomes *Trichobilharzia szidati* (Platyhelminthes: Digenea: Schistosomatidae)

S. K. Semyenova, G. G. Chrisanfova, A. S. Guliaev, A. P. Yesakova, and A. P. Ryskov

Institute of Gene Biology, Russian Academy of Sciences, Vavilov Street 34/5, Moscow 119334, Russia

Correspondence should be addressed to S. K. Semyenova; seraphimas@mail.ru

Received 31 October 2014; Accepted 7 March 2015

Academic Editor: Peter F. Stadler

Recently we developed the genus-specific markers of the avian schistosomes of the genus *Trichobilharzia*, the causative agents of human cercarial dermatitis. The 7 novel genome sequences of *T. franki, T. regenti*, and *T. szidati* revealed similarity with genome repeat region of African schistosome *Schistosoma mansoni*. In the present work we analyzed the 37 new *T. szidati* sequences to study intragenome variability and host specificity for the parasite from three localities of East Europe. DNAs were isolated from cercariae or single sporocysts obtained from 6 lymnaeid snails *Lymnaea stagnalis* and *L. palustris* from Belarus and Russia. All sequences formed three diverged groups, one of which consists of the sequences with multiple deletions; other groups involved two paralogous copies with stop codons and frameshift mutations. Strong association between geographical distribution and snail host specificity cannot be established. All studied sequences have homology with the reverse transcriptase domain (RT) of Penelope-like elements (PLE) of *S. mansoni* and *S. japonicum* and new members of RT family were identified. We proposed that three diverged groups RT sequences of *T. szidati* are results of duplication or transposition of PLE during parasite evolution. Implications of the retroelement dynamics in the life history of avian schistosomes are discussed.

1. Introduction

Transposable elements (TEs) are an essential part of a moderately repetitive fraction of any eukaryotic genome. Incorporating into the various regions of genome, they play a significant role in increasing mutational variability and reorganization of the genome. TEs can also alter expression of individual genes and participate in formation of the new ones [1]. Genomic rearrangements induced by TEs are often associated with a variety of adaptations to the environment [2, 3] and thus promote reproductive isolation of organisms; that is, they are implicated in speciation events [4, 5]. TEs distributions can vary among isolates of single species, so TEs have been used as markers to distinguish genetically divergent populations and subpopulations [6].

Retrotransposons constitute a significant proportion of the TEs; their typical characteristic is the use of the reverse transcription mechanisms involving a reverse transcriptase (RT). Retrotransposons are major components of eukaryotic genomes and have been usually divided into four classes containing the long terminal repeat (LTR+), without LTR (non-LTR), Penelope-like elements (PLE), and DIRS. Non-LTRs are grouped into 11 different clades, based on the phylogeny of RT domains [7].

Several retrotransposons have been identified in blood flukes of the genus *Schistosoma*. Despite the fact that they are more than half (58.5%) of the repetitive elements [8, 9], detailed characteristics are known only for a few LTR retrotransposons (Boudicca, Gulliver for *S. mansoni* and Tiao for *S. japonicum*), as well as for some non-LTR retrotransposons (CR1 for *S. mansoni* and *S. japonicum* and SjR2 and Pido for *S. japonicum*) [10]. The third class of retrotransposons, PLE, is widespread among eukaryotes including schistosomes in which only two of PLEs were described [11, 12].

As compared with these *Schistosoma* species, the genome of our object belongs to a more ancient group of the blood flukes infecting waterfowl, namely, *Trichobilharzia*, and is still virtually unexplored.

Table 1: General data for 6 cercarial isolates from *L. stagnalis* and *L. palustris* and for 40 sequences obtained by the primers TR98F and TR98R. Distribution of clones examined between three groups indicated in Figure 1 is shown in square brackets [I, II, III].

Isolate	Host	Locality	Total number of sequences		**D*-distance (%), min–max (average)
			Cercariae	Sporocysts	
Sz31	*Ls*	Belarus	2 [2, 0, 0]		
Sz43	*Lp*	Belarus	2 [1, 0, 1]	spc 1 [1, 3, 3]	0.3–21.1 (10.6)
				spc **2 [0, 1, 6]**	**0–1.3 (0.5)**
				Σ = 14 [1, 4, 9]	0–20.9 (6.5)
Sz11	*Lp*	Karelia	5 [4, 0, 1]	—	0.5–19.6 (8.6)
Sz12	*Lp*	Karelia	—	spc 1 [2, 1, 1]	0–20.1 (13.7)
				spc **2 [4, 1, 1]**	**0.3–19.9 (8.1)**
				Σ = 10 [6, 2, 2]	0–20.5 (9.1)
Sz1	*Ls*	Moscow	2 [1, 1, 0]	—	—
Sz3*	*Ls*	Moscow	5 [0, 1, 4]	—	0–0.3 (0.2)
Σ	*Ls*		9 [3, 2, 4]	—	0–18.3 (10.5)
	Lp		7 [5, 0, 2]	24 [7, 6, 11]	0–21.2 (11.3)
	Ls + *Lp*		16 [8, 2, 6]	24 [7, 6, 11]	0–45.9 (24.1)

*This group includes three previously deposited sequences GU980751–GU980753. **The comparison was carried out only for 390 and 391 bp fragments; spc: sporocyst.

We converted RAPD amplicons into SCAR (Sequence Characterized Amplified Region) markers for three avian schistosome species *T. szidati*, *T. franki,* and *T. regenti* and found new genus-specific repetitive sequences which revealed 64% homology with the repeat region of *Schistosoma mansoni* [13]. For that reason, a pair of specific primers TR98F and TR98R was matched to the sequence of one *T. franki* RAPD spectrum amplicon. Following PCR allowed us to detect 391–393 bp fragments in the spectrum of each species and additional shorter fragment 274 bp was amplified only in *T. szidati* which parasitized one snail *Lymnaea stagnalis*. A few species-specific mutations and indels were found in seven nucleotide sequences from three schistosome genomes studied and confirm the suitability of these sequences for molecular diagnostics of species of genus *Trichobilharzia* [13].

In another study we assessed the overall representation of different types of repeats in a small RAPD library of *T. szidati* obtained from clonal offsprings, individual cercariae within daughter sporocysts. 50 polymorphic nonoverlapping DNA fragments ~300–1500 bp were revealed from RAPD patterns of 47 individual genomes of parasites infected 6 freshwater snails *L. stagnalis*. These sequences contained tandem, inverted, and dispersed repeats as well as regions homological to retroelements of two human parasites, *S. mansoni* and *S. japonicum*. Tandem and inverted repeats constituted 8.9% and 22.1%, respectively, while the percentage of dispersed repeats was 21.0%. About 40% of sequences of approximately 1000 bp included regions which displayed amino acid homology with open reading frame *pol* products of *S. mansoni* and *S. japonicum* retroelements: nonlong terminal repeat retrotransposons (nLTRs, 76%), long terminal repeat retrotransposons (LTRs, 14%), and Penelope-like elements (PLEs, 10%). Most of these regions (86.4%) contained frameshifts, gaps, and stop codons [14]. In the present study one of the SCARs is to provide detailed characteristics of *T. szidati* intragenome variability for the first time. Furthermore we examined the host specificity of the parasites from three geographic localities obtained from two freshwater snail species *L. stagnalis* and *L. palustris*. We present the results of structural, phylogenetic, and bioinformatic analyses to determine the distribution and possible functions of 37 newly identified genomic sequences belonging to the RT domain as a part of the PLE in *T. szidati*. We also demonstratedfor the first time that *T. szidati* genomes contain threediverged-groups ofRT sequences which are result of duplication or transposition of TEs during parasite evolution.

2. Material and Methods

2.1. Collection Sites and Sequence Generation. A total of 6 *T. szidati* isolates (infrapopulations) were collected from the freshwater snails *Lymnaea stagnalis (Ls, n* = 3) and *Lymnaea (Stagnicola) palustris (Lp, n* = 3). The snails were sampled from the three geographical localities, Moscow freshwater pond Altufyevo (in 2005), Lake Naroch (Belarus, 2008), and Lake Onega (Karelia, Russia, 2012) (Table 1). Total genomic DNA was extracted from 5–10 ethanols fixed free-swimming mature cercariae or fragments of individual sporocysts as described previously [14]. PCR with a specific primer pair, TR98F (CTCCGACTGATGATGACAAGAAGA) and TR98R (ATGAGTGGCGAACGGTATCCT), and cloning and sequencing of amplified products were carried out as described [13]. For each PCR fragment 2–5 clones were sequenced. In total, 37 newly generated sequences were analyzed, of which 30 clones contained inserts of 390 or 391 bp and 7 contained the shorter inserts of 274 bp in size. All sequences were deposited in GenBank under accession numbers KP889985-890021.

2.2. Data Analysis. Multiple alignments were made with CLUSTAL and MUSCLE algorithms implemented in

MEGA5.2 [15] software and were edited manually. Search of stop codons in alignments, AT/GC ratio, mean pairwise genetic distances (min-max, overall d-distance) [16], and codon based Z-test of neutrality (Nei-Gojobori method with Jukes-Cantor correction and 1000 bootstrap replications) were made using MEGA version 5.2. Phylogenetic analysis (Neighbour Joining and Bayesian Inference) was performed by MEGA version 5.2. and MrBayes version 3.2.2. software [17]. The best-fit nucleotide substitution model was selected using jModelTest version 2.1.6. [18]. We used HKY model for Bayesian analysis with two simultaneous runs of four chains for 5 000 000 generations, sampling trees every 500 generations. The first 25% trees sampled were discarded as "burn-in." For comparative analysis we used three sequences of *T. szidati* deposited in GenBank (Acc. numbers GU980751–GU980753, [13]).

Similarity searches of homology between our nucleotide sequences of *T. szidati* and previously known nucleotide and amino-acid sequences of mammalian schistosomes and other trematodes have been performed using BLAST (*blastn*, *blastx,* and *tblastx*) with the default parameters [19].

3. Results

3.1. Analysis of Intra- and among Population Variability. Six DNA patterns were obtained with the use of two specific primers TR98F and TR98R and DNA templates isolated from mature cercariae or individual sporocysts from six isolates infecting the three snails of *L. stagnalis* (*Ls*) and three snails of *L. palustris* (*Lp*) during the course of PCR amplification. Each of the patterns comprises two amplicons with the identical intensity of UV luminescence and approximate size of 400 and 300 bp. The size of the cloned sequences of the longer amplicon ($n = 30$) reached 390–391 bp, and the sequences of the shorter fragment ($n = 7$) contained 274 bp. Only 17 sequences out of 40 were unique, and the rest contained from two to four identical copies.

Estimates of genetic heterogeneity of each of the six *T. szidati* infrapopulations (isolates from single snails) are presented in Table 1. They were obtained by calculation of genetic distances for each pair of sequences of the size 390 and 391 bp of free-swimming mature cercariae (isolates Sz3 and Sz11) and fragments of single sporocysts (isolates Sz12 and Sz43). The maximum and minimum estimates of divergence between pairs ranged from 0 to 21.1% and depend mainly on the size of the sample. Sequence divergence was revealed to be up to 0.3% in a few samples of the parasite of snails *Ls*. The maximum differing copies (up to 21%) were found among the parasites that infect the snails *Lp* (isolates Sz11, Sz12, and Sz43). Despite this, the average levels of divergence between the copies do not differ much for cercariae isolated from the two different species of snails (10.5% and 11.3%). In the total sample, the divergence of copies reaches 24.1%.

The reason for such a high intraspecific heterogeneity becomes apparent in the construction of the dendrogram of genetic differences, demonstrating the distribution of 40 sequences in six infrapopulations from the three geographical localities (Figure 1). 15 sequences of 390 bp in size (Group I) and 17 sequences of 391 bp (Group III) are combined in two large clusters with high reliability (IB = 100%). Thus, the full-length amplicons form the two groups of significantly diverged sequences.

The intragroup differentiation is somewhat higher for Group III ($D = 3.9\%$) compared with Group I ($D = 0.6\%$), whereas the intergroup differences account for 24.1% (Table 2). Eight sequences of short copies of 274 bp form its own cluster (IB = 100%). It is composed of two distinct copies of the isolate Sz1 derived from the *L. stagnalis* Ls1 (Moscow). Apart from them, there are six sequences of the two schistosome isolates Sz12 and Sz43 from snails *L. palustris* (Karelia and Belarus) (Figure 1, Group II). The average value of D in the group II is 2.2%. In Group I, we found no clear subclusters neither characterizing geographic population identity nor belonging to either of two species of intermediate snail hosts. In the third group, four sequences of the parasite from one *L. stagnalis* from Moscow (isolate Sz3) and schistosome sequences from *L. palustris* from Belarus (isolate Sz43) comprised their own subclusters. The sequences of two isolates Sz11 and Sz12 from Karelian mollusks *L. palustris* either fall into one of the two subclusters or stand quite separately (e.g., variant Sz12_1_11 on Figure 1). Note that another sequence of isolate Sz11, namely, Sz11_5, holds an isolated position in Group I.

Occasionally, snails in natural populations can be infected with not one but two or more miracidia having different genotypes. This leads to biased estimates of variability in some infrapopulations. Therefore, we compared the variability not only of mature cercariae but also of individual sporocysts. For this purpose, two sporocysts (spc1 and spc2) were isolated from each of the two snails *L. palustris* from Belarus (Sz43) and Karelia (Sz12) and for each of them from four to seven sequences were obtained. Individual variability of sporocysts consisted of the presence of two or three differing copies in each of the three groups of sequences.

Sequences of Groups I and II were simultaneously identified in only three of sporocysts, where the average level of divergence was high and reached 8.1% (Sz12_2), 10.6% (Sz43_1), and 13.7% (Sz12_1). All sequences of the remaining sporocyst Sz43_2 belong to Group II and were almost identical ($D = 0.5\%$). All short copies were also almost the same, both within individual sporocysts and between sporocysts from the same mollusk (Figure 1), while the most divergent two copies from Group II (Sz12S_2_16 and Sz12S1_10) and Group III (Sz12_1_11 and Sz12_2_17), which define the highest level of infrapopulation variability in Sz12, are the part of the genomes of both sporocysts 1 and 2 of mollusk Sz12. Thus, comparing the genetic heterogeneity of cercariae from one sporocyst, we demonstrated that the composition of a bird schistosome *T. szidati* genome could simultaneously present three groups of copies of closely related sequences. Maximum intragenomic divergence is typical for the parasite infrapopulation from mollusk *L. palustris* (Karelia) and can reach 20%.

The distribution of copies in the six infrapopulations of snails indicates a lack of host specificity. However, there is a tendency to the formation of specific sets of copies of Groups II and III in geographically isolated parasite infrapopulations

Table 2: Characteristics of intra- and intergroup polymorphism and results of similarity search of studied *T. szidati* sequences.

Groups	Polymorphism	Similarity		
		BLASTN (+/+)	BLASTX (Frame +3)	TBLASTX (Frame +3/+3)
Group I	$N = 15$ $L = 390$ $V = 12$ Pinfo = 2 AT : GC = 62.3 : 37.7 Z test: dN-dS = −0.021 ($P = 0.983$) Dn = 0.6 Da = 1.1	No	Sm RT CAJ00247: score 49.3–52.4 bits, Exp 2*e* − 05–2*e* − 04, cover 51% (15–215 bp), **I = 39–42%** Sj RT CAX83715: score 48.1–50.8 bits, Exp 6*e* − 05–1*e* − 04, cover 54% (3–215 bp), **I = 35–38%** Cs RTGAA47523: score 45.8–46.2 bits, Exp 0.002–0.003, cover 40% (57–215 bp), **I = 42%**	Sj Penelope-like element retrotransposon Sj-penelope2 FN356226: score 65.4–70.2 bits, Exp 3*e* − 24–4*e* − 17, cover 98% (3–215 bp), **I = 39–42%** Sm Penelope-like retrotransposon Perere-10 BN000801: Score 54.2–58.8 bits, Exp 3*e* − 13–6*e* − 06, Cover 53-54% (15–221 bp), **I = 39–42%**
Group II	$N = 8$ $L = 274$ $V = 20$ Pinfo = 20 AT : GC = 57.8 : 42.2 Z test: dN-dS = −0.949 ($P = 0.344$) Dn = 2.2 Da = 2.0	No	Cs RTGAA47523: score 35.0, Exp 4.5–4.6, cover 43–48 (466–521 bp), **I = 38-39%**	No
Group III	$N = 17$ $L = 391$ $V = 24$ Pinfo = 85 AT : GC = 58.7 : 41.3 Z test: dN-dS = −1.597 ($P = 0.113$) Dn = 3.9 Da = 3.9	S j Anhui clone BAC C108_113I17 FN293021: score 55.4 bits Exp *e* − 0.4; cover 17% (308–375 bp); **I = 78%** Chr1, chr2, chr3, chr4, chr7, Wchr: score 48.2–46.4 bits Exp *e* − 0.26; cover 15–19% (116–220 bp); **I = 70–77%**	Sm RT CAJ00247: score 64.7–73.6 bits, Exp 9*e* − 13–2*e* − 07, cover 91–95% (15–386 bp), **I = 30–37%** Sj RT CAX83715: score 50.4–60.8 bits, Exp 9*e* − 05–2*e* − 07, cover 77–95% (3–374 bp), **I = 30–32%** Cs RTGAA47523: score 46.2–52.2 bits, Exp 5*e* − 05–0.002, cover 39–41% (3–374 bp), **I = 47–49%**	Sj Penelope-like element retrotransposon Sj-penelope2 FN356226: score 70.2–103.0 bits, Exp 1*e* − 26–2*e* − 19, cover 98% (3–386 bp), **I = 34–36%** Sm Penelope-like retrotransposon Perere-10 BN000801: score 82.6–83.1 bits, Exp 2*e* − 20–3*e* − 13, cover 95% (15–386 bp), **I = 35%** Sm Chr1, chr3, chr4, chr7, Wchr: score 118–48.2 bits, Exp *e* − 14–*e*28; cover 98-99% (3–386 bp), **I = 27–47%** S j Anhui clone BAC C108_113I17 FN293021: score 35–49.2 bits, Exp 1*e* − 06–4*e* − 04; cover 40–93 (81–230 bp); **I = 26–40%**
Groups I + III	$N = 32$ $L = 391$ $V = 203$ AT : GC = 60.4 : 39.6 Pinfo = 182 Z test: dN-dS = −2.640 ($P = 0.009$) Dn = 24.1 Da = 21.6	—	—	—

N: the sequence numbers, *L*: length (bp), *V*: the number of variable sites, Pinfo: the number of parsimonial sites, *D*: distance (%), Sm: *S. mansoni*, Sj: *S. japonicum*, Cs: *Clonorchis sinensis*, RT: reverse transcriptase, Dn: nucleotide divergence, Da: amino acid divergence.

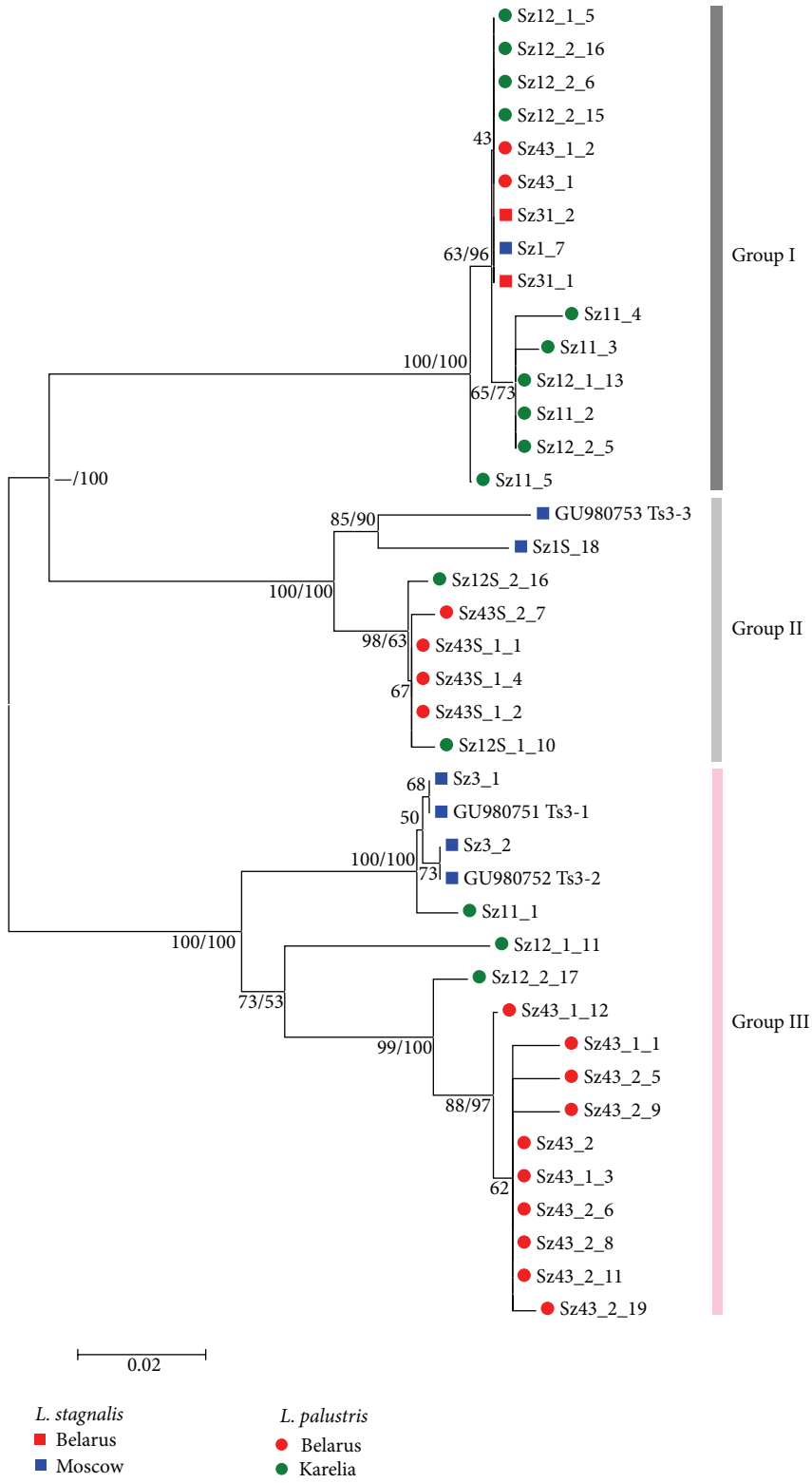


Figure 1: Phylogenetic tree of *T. szidati*, inferred from 40 RT-like sequences. Topology was inferred using MEGA 5.2 software (NJ, p-distance, 1000 bootstrap replications) and confirmed by MrBayes 3.2.2. Node support values are shown as follows: the first value is Bayesian posterior probability assessed using MrBayes software, and the second value is bootstrap support assessed by NJ method using MEGA 5.2 software. Sequences belonging to different localities and host snails are shown by differently colored figures (see the legend).

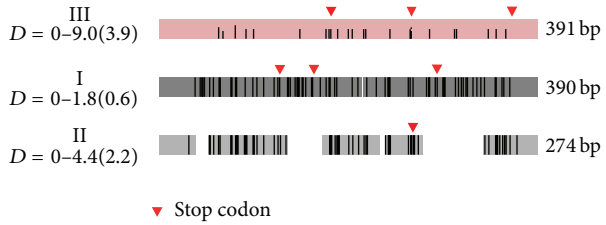


FIGURE 2: Schematic image of the distribution of polymorphic sites and deletions and the location of stop codons in each of the three RT-like sequences groups of *T. szidati*. D: min-max and average (in parentheses) pairwise genetic distances.

L. stagnalis from Moscow and in the pooled sample of parasites of *L. palustris* from Karelia and Belarus. However, to clarify these results, it is necessary to study more representative samples of *Trichobilharzia*–infected snails from wider geographical origins.

3.2. Analysis of the Sequence Structure and Search Similarity. Figure 2 gives schematically the distribution of polymorphic sites, deletions, and the location of stop codons in each of the three groups of sequences. All sequences of Group I have a length of 390 bp due to a single nucleotide deletion at position 211 compared with the sequences of Group III (391 bp). All short sequences of Group II with the size of 274 bp have four extended deletions of total length 117 bp (14 nt: 38–51 bp, 36 nt.: 133–168 bp 5 nt.: 229–233 bp, 62 nt.: 273–334 bp). Groups I and III sequences each contains three stop codons and Group II, one stop codon. Their position is indicated with triangle in Figure 2. The second stop codon in Group III is revealed only in a half of the sequences. An excess of AT-bases (AT : GC = 60.4 : 39.6) known in all trematodes is found in the nucleotide composition of all sequences (Table 2). Average estimates of divergence of nucleotide and amino acid sequences are the same for Groups II and III, and amino acid divergence is somewhat higher comparing to nucleotide divergence in Group I (Dn = 0.6, Da = 1.1). Codon-based z-test of neutrality showed that observed mutations are significantly neutral only within concatenated Groups I and III. Within each group separately results of neutrality test are insignificant, as well as results of both purifying and positive selection tests (Table 2).

Table 2 summarizes also the results of *T. szidati* sequence homology with nucleotide and amino acid sequences available throughout the NCBI. We found no significant similarity of *T. szidati* nucleotide sequences in Groups I and II with extended sequences of mammalian schistosome genomes (blastn). Only for Group III, a high similarity of short segments of the sequences (about 40 nt) with the intergenic regions of 1, 2, 3, 4, and 7 and W-chromosomes of *S. mansoni* as well as with more extensive (about 70 nt) unannotated region of the BAC-clone of *S. japonicum* (C108_113I17, FN293021) was revealed.

Translated nucleotide sequence similarity search (blastx) revealed a significant homology between the sequences of Groups I and III and the reverse transcriptase domain of two mammalian schistosome species, *S. mansoni* (Ass.n. SA00247) and *S. japonicum* (Ass.n. SAX83715). The length of these GenBank sequences is equal to 394 and 500 amino acid residues, respectively, and the regions of similarity are located almost at the 5′ terminus and comprise about 125 amino acid residues of the mammalian schistosome's RT. Moreover, a significant similarity (41–51%) with the two shorter regions (30–80 aa) of the RT sequence of the human fluke *Clonorchis sinensis* (532 amino acid residues, GAA47523) was found. Therefore, the results of similarity search with known *Schistosoma* amino acid sequences indicate that new sequences of *T. szidati* belong to the family of reverse transcriptase (RT) genes.

Furthermore, using another BLAST algorithm (blastx) the highest similarity our sequences (~36% for Group III, ~42% for Group I) was found to the 5′ terminus of the reverse transcriptase domain of two retrotransposons of *S. mansoni* (Perere-10, BN000801) and *S. japonicum* (Sj-penelope2, FN356226), belonging to a large class of Penelope-like transposable elements. In addition, the amino acid sequences of Group III demonstrate 27–47% similarity with the short regions on the chromosomes 1–4, 7 and W-chromosome of *S. mansoni*, included in the introns of hypothetical proteins or noncoding intergenic regions.

4. Discussion

Search for homology with known schistosome amino acid sequences of the Schistosoma genus indicates that new genome sequences of the avian schistosomes *T. szidati* belong to the RT domain, which is common to all retroretroelements. Reverse transcriptase is the most highly conserved protein encoded by retroviruses and retrotransposons. This peculiarity allows the use of RT sequences as recognizing phylogenetic signature of host taxa in a retrotransposon phylogeny, besides that, for studying the dynamics of retroposition in the life cycle to determine its life history [20, 21].

Relatively little is known about intraspecies and intragenomic variability of RT among invertebrates. Usually, it does not exceed 10% for the members of the same subfamily.

Thus, mean intraspecific divergence is 2.88% between reverse transcriptase sequences of SURL elements (from the *gypsy* group) in the closely related echinoid species *Strongylocentrotus purpuratus* and S. *droebachiensis* [22]. Several families of elements were found in African and Asian schistosomes, which were characterized by more than 80% of similarity in amino acid sequences of RT. It has been shown that a family can combine both copies of the same element and the closely related elements [23].

Significant variability in the composition of RT-like sequences (0–21.2%) of 390-391 bp in size was found when studying *T. szidati*, infecting three snails of *L. palustris*. These estimates do not depend on the geographical location of the snails (Belarus and Karelia), nor on the stage of the parasite life cycle (free-swimming mature cercariae or fragments of single sporocysts).

The main reason for the high heterogeneity of the RT-like sequences of *T. szidati* is a simultaneous occurrence

of significantly diverged copies of Groups I and III in one genome (Table 1, Figure 1).

The nucleotide and amino acid divergence between RT copies of these groups is 20% on average reaching the level of 45% for individual copies (Table 2). Given the lack of detailed annotation of the complete genome sequences of African and Asian mammalian schistosomes, we are able to conditionally include all detected RT copies to the members of the same family for the present. The average sequence divergence in a group is less than 4%; thus presumably we are dealing with representatives of several RT lines or subfamilies.

All detected RT-like copies probably are inactive copies as they contain either stop codons (Groups I and III) or a single nucleotide deletion (Group I), modifying the reading frame (frame shift mutation). Note that short copies of 274 bp in size of Group II pertain to inactive copies.

These copies with typical extended deletions are more degenerated comparing with the previously mentioned RT copies. Due to the fact that any detected changes in the structure of RT sequences in *T. szidati* are incapable of coding for a functional reverse transcriptase (breaking the reading frame of RT), we can refer the elements of each of the three groups to pseudogenes, derived from the RT protein-coding gene. Since the pseudogene evolved under neutrality (Z-tests, Table 2), they may show the higher level of diversity in some cases. For instance, there is a 45% of sequences divergence between the pairs Sz43_2_10 and Sz12_1_5 (Figure 1, Table 1). Thus, for the first time, we have discovered the three types of degenerated RT copies in the same genome of avian schistosomes probably belonging to a few closely related subfamilies of transposable elements.

To date, from all deposited RT-containing retrotransposons of mammalian schistosomes (see Introduction) new obtained sequences detected in the *T. szidati* genome show significant similarity with representatives of the Penelope-like elements (Perere-10 and Sj-penelope2). Therefore, we have reason to include currently all identified RT copies in the *T. szidati* genome to a class of PLE.

The absence of intact sequences among the discovered copies indicates their ancient origin, while the older group seems to be a group of highly degenerated and reduced in size sequences of Group II. Compared with them, paralogs of Group I are less degenerated, having only one reading frame shift mutation and several stop codons. Copies of Group III with two or three stop codons degenerated probably much later, and therefore only for copies of this group, small areas of similar genomic sequences on the five chromosomes of mammalian schistosome *S. mansoni* were found as well. Besides, their use for phylogenetic reconstruction demonstrates the presence of intraspecies structure in *T. szidati* (Figure 1).

We cannot yet reconstruct a detailed scenario for the origin and invading of discovered RT-like sequences in populations of *T. szidati* on the limited material. It is likely that the occurrence of paralogous RT copies is associated with transposition bursts that took place in the remote past of avian schistosomes. Apparently, two acts of transposition bursts could result in the three types of RT copies in the genome of modern *T. szidati*. The most compelling evidence of this assumption will be obtained from the analysis of the whole genome sequencing of different species of avian schistosomes. Currently, similar analysis was carried out for genomes of the schistosomes *S. mansoni* and *S. japonicum* [23]. Considerable differences in retrotransposon representation have been shown between the two species (22% and 13%, resp.). A large part of this difference can be attributed to higher representation of two previously described retrotransposon families SR2 and Perere-3/SR3 of *S. mansoni*. It was suggested that the *S. mansoni* SR2 families were the subject of recent bursts of transposition that were not paralleled by their *S. japonicum* counterparts. It was hypothesized that these bursts could be a consequence of the evolutionary pressure resulting from migration of *Schistosoma* from Asia to Africa and their establishment in this new environment, helping both speciation and adaptation [23].

Similar processes could occur during the life history of avian schistosomes. Their definitive hosts, ducks of the family Anatidae, are characterized by long-distance spatial and temporal migrations, changing of ecological niches, and multiple range expansions [24]. It is necessary to add that processes of snail-parasite interactions, occurring during the development or change of the intermediate snail host, also have a significant role in the genetic differentiation of schistosomes [25]. During the evolutionary radiation of mammalian schistosomes, Asian and African groups have adapted to parasitizing on the snails of different groups of Gastropoda. African schistosomes infected representatives of several families of pulmonate snails (Pulmonata) and Asian species feed on mollusks belonging to the Caenogastropoda. These processes resulted just in mitochondrial genome rearrangement. Thus, the ancestral gene order of mtDNA is conserved amongst East Asian *Schistosoma* spp. [26] and different amongst species sampled from Africa or India [27, 28]. In the evolution of a more ancient group of schistosomes, namely, avian schistosomes, multiple repeated changes of the definitive and intermediate hosts could also occur as well as generating of new molecular adaptations, and increasing the transposition activity of TEs may serve as markers of such events.

5. Conclusions

In the present work 37 new sequences obtained from genome of avian schistosome *Trichobilharzia szidati* parasitized 6 lymnaeid snails *L. stagnalis* and *L. palustris* from Belarus and Russia were revealed. Phylogenetic reconstructions and BLAST search results indicate that all studied sequences demonstrate homology with the reverse transcriptase domain (RT) of Penelope-like elements of African and Asian mammalian schistosomes *S. mansoni* and *S. japonicum*. Future whole genome sequencing and population-wide analysis of avian schistosomes will help to understand the features of the retrotransposon expansion during host-parasite coevolution.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

S. K. Semyenova and G. G. Chrisanfova equally contributed to the study.

Acknowledgments

The authors thank E. P. Ieshko, S. A. Be'er, and M. V. Voronin for help in collecting infected snails and A. A. Lopatkin for help with PCR amplification. This work was supported by the Russian Science Foundation no. 14-14-00832.

References

[1] M. G. Kidwell and D. Lisch, "Transposable elements as sources of variation in animals and plants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 15, pp. 7704–7711, 1997.

[2] A. L. Schmidt and L. M. Anderson, "Repetitive DNA elements as mediators of genomic change in response to environmental cues," *Biological Reviews of the Cambridge Philosophical Society*, vol. 81, no. 4, pp. 531–543, 2006.

[3] E. Casacuberta and J. González, "The impact of transposable elements in environmental adaptation," *Molecular Ecology*, vol. 22, no. 6, pp. 1503–1517, 2013.

[4] H. H. Kazazian Jr., "Mobile elements: drivers of genome evolution," *Science*, vol. 303, no. 5664, pp. 1626–1632, 2004.

[5] K. R. Oliver and W. K. Greene, "Transposable elements: powerful facilitators of evolution," *BioEssays*, vol. 31, no. 7, pp. 703–714, 2009.

[6] J. Jurka, W. Bao, and K. K. Kojima, "Families of transposable elements, population structure and the origin of species," *Biology Direct*, vol. 6, article 44, 16 pages, 2011.

[7] H. S. Malik, W. D. Burke, and T. H. Eickbush, "The age and evolution of non-LTR retrotransposable elements," *Molecular Biology and Evolution*, vol. 16, no. 6, pp. 793–805, 1999.

[8] M. Berriman, B. J. Haas, P. T. LoVerde et al., "The genome of the blood fluke *Schistosoma mansoni*," *Nature*, vol. 460, no. 7253, pp. 352–358, 2009.

[9] F. Liu, Y. Zhou, Z.-Q. Wang et al., "The *Schistosoma japonicum* genome reveals features of host–parasite interplay," *Nature*, vol. 460, no. 7253, pp. 345–351, 2009.

[10] P. J. Brindley, C. S. Copeland, and B. H. Kalinna, "Schistosome retrotransposon," in *Schistosomiasis*, W. E. Secor and D. G. Colley, Eds., pp. 13–26, 2005.

[11] I. R. Arkhipova, "Distribution and phylogeny of Penelope-like elements in eukaryotes," *Systematic Biology*, vol. 55, no. 6, pp. 875–885, 2006.

[12] R. DeMarco, A. T. Kowaltowski, A. A. Machado et al., "Saci-1, -2, and -3 and perere, four novel retrotransposons with high transcriptional activities from the human parasite *Schistosoma mansoni*," *Journal of Virology*, vol. 78, no. 6, pp. 2967–2978, 2004.

[13] A. V. Korsunenko, G. G. Chrisanfova, A. P. Ryskov, S. O. Movsessian, V. A. Vasilyev, and S. K. Semyenova, "Detection of European *Trichobilharzia schistosomes* (*T. franki, T. szidati, and T. regenti*) based on novel genome sequences," *Journal of Parasitology*, vol. 96, no. 4, pp. 802–806, 2010.

[14] A. Korsunenko, G. Chrisanfova, A. Arifov, A. Ryskov, and S. Semyenova, "Characterization of randomly amplified polymorphic DNA (RAPD) fragments revealing clonal variability in cercariae of avian schistosome *Trichobilharzia szidati* (Trematoda: Schistosomatidae)," *Open Journal of Genetics*, vol. 3, no. 3, pp. 141–158, 2013.

[15] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[16] F. Tajima and M. Nei, "Estimation of evolutionary distance between nucleotide sequences," *Molecular Biology and Evolution*, vol. 1, no. 3, pp. 269–285, 1984.

[17] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[18] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada, "JModelTest 2: more models, new heuristics and parallel computing," *Nature Methods*, vol. 9, no. 8, p. 772, 2012.

[19] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[20] Y. Xiong and T. H. Eickbush, "Origin and evolution of retroelements based upon their reverse transcriptase sequences," *The EMBO Journal*, vol. 9, no. 10, pp. 3353–3362, 1990.

[21] T. H. Eickbush, "Origin and evolutionary relationship of retroelements," in *Evolutionary Biology of Viruses*, S. S. Morse, Ed., pp. 121–157, Raven Press, 1994.

[22] M. S. Springer, N. A. Tusneem, E. H. Davidson, and R. J. Britten, "Phylogeny, rates of evolution, and patterns of codon usage among sea urchin retroviral-like elements, with implications for the recognition of horizontal transfer," *Molecular Biology and Evolution*, vol. 12, no. 2, pp. 219–230, 1995.

[23] T. M. Venancio, R. A. Wilson, S. Verjovski-Almeida, and R. DeMarco, "Bursts of transposition from non-long terminal repeat retrotransposon families of the RTE clade in *Schistosoma mansoni*," *International Journal for Parasitology*, vol. 40, no. 6, pp. 743–749, 2010.

[24] Y. Liu, I. Keller, and G. Heckel, "Breeding site fidelity and winter admixture in a long-distance migrant, the tufted duck (*Aythya fuligula*)," *Heredity*, vol. 109, no. 2, pp. 108–116, 2012.

[25] A. E. Lockyer, C. S. Jones, L. R. Noble, and D. Rollinson, "Trematodes and snails: an intimate association," *Canadian Journal of Zoology*, vol. 82, no. 2, pp. 251–269, 2004.

[26] T. H. Le, D. Blair, T. Agatsuma et al., "Phylogenies inferred from mitochondrial gene orders—a cautionary tale from the parasitic flatworms," *Molecular Biology and Evolution*, vol. 17, no. 7, pp. 1123–1125, 2000.

[27] A. E. Lockyer, P. D. Olson, P. Østergaard et al., "The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of *Schistosoma* Weinland, 1858," *Parasitology*, vol. 126, no. 3, pp. 203–224, 2003.

[28] B. L. Webster and D. T. J. Littlewood, "Mitochondrial gene order change in *Schistosoma* (Platyhelminthes: Digenea: Schistosomatidae)," *International Journal for Parasitology*, vol. 42, no. 3, pp. 313–321, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 202597, 16 pages
http://dx.doi.org/10.1155/2015/202597

Research Article

Biochemical and Molecular Phylogenetic Study of Agriculturally Useful Association of a Nitrogen-Fixing Cyanobacterium and Nodule *Sinorhizobium* with *Medicago sativa* L.

E. V. Karaushu,[1] I. V. Lazebnaya,[2] T. R. Kravzova,[3] N. A. Vorobey,[4] O. E. Lazebny,[5] D. A. Kiriziy,[4] O. P. Olkhovich,[1] N. Yu. Taran,[1] S. Ya. Kots,[4] A. A. Popova,[6] E. Omarova,[7] and O. A. Koksharova[6,7]

[1] *Educational and Scientific "Institute of Biology", Taras Shevchenko National University of Kyiv, 64/13 Volodymyrska Street, Kyiv 01601, Ukraine*
[2] *N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkin Street 3, Moscow 119333, Russia*
[3] *Lomonosov Moscow State University, Biocenter, Leninskie Gory 1-12, Moscow 119991, Russia*
[4] *Institute of Plant Physiology and Genetics, National Academy of Sciences of Ukraine, 31/17 Vasylkivska Street, Kyiv 03022, Ukraine*
[5] *N. K. Kol'tsov Institute of Developmental Biology, Russian Academy of Sciences, Vavilova Street 26, Moscow 119334, Russia*
[6] *Institute of Molecular Genetics, Russian Academy of Sciences, Moscow 123182, Russia*
[7] *Lomonosov Moscow State University, Belozersky Institute of Physical-Chemical Biology, Leninskie Gory 1-40, Moscow 119992, Russia*

Correspondence should be addressed to O. A. Koksharova; oa-koksharova@rambler.ru

Received 11 September 2014; Accepted 24 February 2015

Academic Editor: Peter F. Stadler

Seed inoculation with bacterial consortium was found to increase legume yield, providing a higher growth than the standard nitrogen treatment methods. Alfalfa plants were inoculated by mono- and binary compositions of nitrogen-fixing microorganisms. Their physiological and biochemical properties were estimated. Inoculation by microbial consortium of *Sinorhizobium meliloti* T17 together with a new cyanobacterial isolate *Nostoc* PTV was more efficient than the single-rhizobium strain inoculation. This treatment provides an intensification of the processes of biological nitrogen fixation by rhizobia bacteria in the root nodules and an intensification of plant photosynthesis. Inoculation by bacterial consortium stimulates growth of plant mass and rhizogenesis and leads to increased productivity of alfalfa and to improving the amino acid composition of plant leaves. The full nucleotide sequence of the rRNA gene cluster and partial sequence of the dinitrogenase reductase (*nifH*) gene of *Nostoc* PTV were deposited to GenBank (JQ259185.1, JQ259186.1). Comparison of these gene sequences of *Nostoc* PTV with all sequences present at the GenBank shows that this cyanobacterial strain does not have 100% identity with any organisms investigated previously. Phylogenetic analysis showed that this cyanobacterium clustered with high credibility values with *Nostoc muscorum*.

1. Introduction

Continuous anthropogenic impact on the environment of different chemicals, fertilizers, herbicides, plant protection from pests and diseases, plant growth regulators, and so forth that are used in agriculture, makes it necessary to develop an alternative to agricultural production, which would be based on the use of cost-effective and environmentally friendly systems for land application of fertilizers and plant protection. An important role in this respect is given to the maximum use of the soil microflora.

In many countries around the world, studies and implementation of the compositions consisting of symbiotic and free-living nitrogen-fixing microorganisms have started to increase productivity of crops. Among the wide range of diazotrophic microorganisms cyanobacteria are the most

versatile for biochemical potential, since they do not need to be provided with soil organic substances for nitrogen fixation unlike heterotrophic nitrogen-fixing microorganisms.

Positive ecological role of cyanobacteria in the soil as nitrogen-fixing bacteria which participate in deposition of organic matter is well known now, and besides they are the centers of microcosms as autotrophic organisms with amazing abilities for symbiotrophic relations [1, 2]. The last property of cyanobacteria is particularly interesting in case of using the consortia of microorganisms in biotechnology instead of monocultures [3]. In nature, cyanobacteria are never found in the form of cell populations of one species. They are in a close relationship with the microbial community, located in the mucus of the surrounding cells. Research in this field has shown that the composition of satellite cyanobacteria is very labile and it depends on changes in habitat conditions. Axenic cultures of cyanobacteria exist only in the laboratories. In nature, they form a community and, being the edificators of microbial communities, cyanobacteria can change the microbial composition [4]. It allows the constructing of artificial microbial consortium. Nitrogen-fixing activity (NFA) of soil compositions of diazotrophic microorganisms can be an effective way to supply the crop by environmentally friendly biological nitrogen. Use of this approach requires in-depth study of the relationship between bacteria, cyanobacteria, and plants, as well as compatibility of microorganisms-partners in created artificial associations. It is important to perform the screening of the most suitable strains of microorganisms, to create conditions for the effective functioning of these symbiotic consortia. It is necessary to select an optimal quantitative ratio of microorganisms and methods of their implantation into the rhizosphere.

The goals of this research were the study of the effect of artificial stable microbial consortium based on nitrogen-fixing cyanobacterium *Nostoc* PTV and Tn5-mutant of nodule bacteria *Sinorhizobium meliloti* T17, on the physiological and biochemical characteristics of growth and development of alfalfa, and, finally, on its yield and product quality and the molecular typing and phylogenetic analysis of this new cyanobacterial isolate *Nostoc* PTV.

2. Material and Methods

2.1. Organisms and Growth Conditions. Plant alfalfa *Medicago sativa* (L.) sort of Jaroslavna obtained from the NSC Institute of Agriculture of National Academy of Agrarian Sciences of Ukraine has been used in the experiments. For the inoculation of alfalfa seeds we used the strain of nodule bacteria *Sinorhizobium* (*Rhizobium*) *meliloti* T17 (patent of Ukraine number 55432) from the collection of nitrogen-fixing microorganisms of the Institute of Plant Physiology and Genetics, National Academy of Sciences of Ukraine (Kyiv) [6]. The strain of *S. meliloti* T17 was obtained as a result of intergeneric conjugation of *Escherichia coli* S17-1 (pSUP2021::Tn5) and *S. meliloti* 425a on agar medium TY (tryptone/yeast extract) as described in [7] and it was selected for improved symbiotic properties. To create the binary composition of nitrogen-fixing microorganisms the culture of cyanobacterium *Nostoc* PTV (from the collection of the Institute of Hydrobiology, National Academy of Sciences of Ukraine) was used. Cyanobacterium was grown on Fitzgerald medium with the modification by Zehnder and Gorham [8] in Erlenmeyer flask at 22°C ± 2°C and illumination of 2500 lux until the stationary growth phase. The concentration of chlorophyll (Chl) in cyanobacterial cells was determined by differential fluorometry (Fluorometer FL300 3M, Russia) [9]. The binary composition was prepared by mixing the bacterial suspensions consisting of nodule bacteria (1×10^9 cells/mL) and cyanobacteria (Chl, mg/L = 1506,6 ± 13,4, ΔF = 0,088) in the ratio 1 : 1. In parallel, the viability of cyanobacterial cells was determined by the difference of fluorescence intensity (ΔF) before and after the addition of simazine, the inhibitor of cells photosynthetic electron transport [10, 11].

Investigations were carried out in the model experiments in a growth area of Institute of Plant Physiology and Genetics with natural light and humidity of the substrate 60% of full capacity. Plastic containers with 10 kg of sand were used in experiments. 12 alfalfa plants were grown in each container. Containers were preliminarily sterilized with 20% solution of H_2O_2. Washed river sand with the mineral nutrient mixture of Gelrigel [12] containing the "start" of nitrogen (177 mg of $Ca(NO_3)_2 \times 4H_2O$ per 1 kg of sand) was used as a substrate. This amount of nitrogen represents one-quarter of the normal nitrogen supply. Before sowing the seeds were sterilized with concentrated sulfuric acid for 5 minutes, and then they were washed in running water for 1 h. The treatment of seeds by microorganism compositions was continuing during 1 h.

The controls in the experiments were the samples of seeds treated by monoculture of T17 *S. meliloti* or only by *N.* PTV. We used samples of alfalfa seeds moisturized with tap water as an additional "absolute" control. Experiments were performed in seven replications. Plants for analysis were selected in phases of stem (32nd day of emergence), budding (40th day), and flowering (50th day).

2.2. Measurements of Nitrogen Activity, Pigments Content, and Efficiency of Photosynthesis. Nitrogen-fixing (nitrogenase) activity was determined by the level of activity of root nodules by acetylene method and expressed as micromoles of ethylene formed by nodules per plant for 1 h [13]. The gas mixture was analyzed by gas chromatography of Agilent Technologies 6855 Network GC System (USA). The measurements were performed in five replications.

The content of the photosynthetic pigments in leaves of alfalfa plants was determined by the Wellburn method [14]. Pigments were extracted with dimethyl sulfoxide (0.1 g vegetable material was treated in 10 mL DMSO) of leaf cut for 3 h at +67°C until complete extraction. The absorbance of the solution was measured by spectrophotometer Smart Spec Plus (BioRad, USA) at 665 and 649 nm in a 1 cm cuvette. Leaves were collected from the middle tiers of the five randomized plants of the same version. Measurements were performed in triplicate.

The net assimilation rate of shoots was determined in controlled environment with installation built on base of the photoacoustic infrared gas analyzer GIAM-5 M (Russia),

which was connected by differential circuit. Container with plants was placed in sealed plexiglass chamber of 50 liters through which air was blown at rate of 15 L/min. At the outlet of chamber 1 L/min of air was taken to the gas analyzer, and the remaining air was discharged into atmosphere. The chamber was irradiated with light by the lamp CG-2000 through a water filter. The illumination on the substrate level was 250 W/m^2; temperature was 25° ± 2°C. After the adaptation of plants to the conditions of measurement (30–40 min after closing the chamber), the rate of absorption of CO_2 by plants was recorded (it is an apparent photosynthesis). After this, shoots of plants were cut at the substrate level and respiration of soil with roots were measured. Net assimilation rate was calculated as sum of apparent photosynthesis and respiration. Calculations of gas exchange parameters were performed according to the standard procedure [15].

The protein content was determined in leaves of alfalfa plants in the budding stage by Lowry method [16]. Qualitative and quantitative composition of amino acids was determined by liquid-ion exchange column chromatography with the use of automatic analyzer T339 (Czech) on the basis of ninhydrin detection method [17].

2.3. Plant Stress Resistance Determination. In order to study the effect of mono- and binary inoculation on plant resistance, the basic parameters of the stress state of alfalfa were determined. Plants in the budding stage were treated with herbicide diquat (100 pmol), which was used as a stress factor. Sampling was carried out after 30 minutes, 60 minutes, and 24 hours of diquat action on plants. Specific changes in the composition of the components of a lipid-pigment complex and antioxidant system were studied in photosynthetic tissues of alfalfa.

Intensity of lipid peroxidation (LPO) was evaluated by the number of end-products of lipid oxidation based on the reaction with 2-thiobarbituric acid (TBA) [18]. The activity of antioxidant systems is determined by the activity of superoxide dismutase (SOD) [19].

A statistical analysis of the experimental data was performed by standard methods, involving a package of special statistical functions of Microsoft Excel. Probability of differences between the variants was assessed by t-test and a significance level of $P < 0{,}05$.

2.4. Scanning Electron Microscopy (SEM). Cyanobacterial samples were fixed as described above and dehydrated through an ethanol series, with an overnight exposure in absolute acetone followed by critical-point drying in a Dryer HCP-2 (Hitachi, Japan), coated with Au-Pd alloy in an IB-3 Ion Coater (Eiko, Japan) and examined with a JSM-6380LA scanning electron microscope (JEOL, Japan).

2.5. DNA Isolation and PCR Amplification. For molecular typing cyanobacterial genomic DNA was isolated according to [20] and synthetic oligonucleotides ("Synthol," Moscow, Russia) have been used as cyanobacterial primers for 16S–23S rRNA PCR, according to [21]. As a second molecular marker the *nifH* gene has been used with corresponding PCR primers [22]. PCR for 16S–23S rRNA gene cluster was carried out on a Tercik DNA amplifier (DNA Technology, Russia) by using DreamTaq PCR Master Mix (Fermentas, EU), under the following conditions: 1 cycle at 94°C for 10 min, 25 cycles at 94°C for 45 sec, 54°C for 45 sec, 68°C for 2 min, 1 cycle at 68°C for 7 min, and a final soak step at 4°C. PCR for partial *nifH* gene was performed under the following conditions: 1 cycle at 94°C for 4 min, 25 cycles at 94°C for 30 sec, 54°C for 30 sec, 68°C for 30 sec, 1 cycle at 68°C for 7 min, and a final soak step at 4°C. PCR products were resolved in 1.5% agarose gel containing ethidium bromide at 5 microgram mL^{-1}.

2.6. Cloning and Sequencing of PCR Products. DNA fragments obtained during PCR were cloned with CloneJet PCR Cloning Kit # K1231 (Fermentas, EU). Transformation of competent XL-1 cells of *Escherichia coli* and plasmid purification were performed according to [23]. DNA sequencing was performed with ABI PRISM BigDye Terminator version 3.1 at the Applied Biosystems 3730 DNA Analyzer (Center for Collective Use "Genome"). Sequences were edited and assembled with Bioedit (Invitrogen, Carlsbad, CA). The full nucleotide sequence of the rRNA gene cluster of cyanobacterium *Nostoc* PTV and a part of the *nifH* gene were accomplished and deposited to GenBank under accession numbers JQ259185.1 and JQ259186.1.

2.7. Phylogenetic Analysis. Search of the nucleotide sequences in the database GenBank, homologous to the sequenced genes of studied species of cyanobacteria, was performed using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) with the option: the least degree of similarity (Minimum Identity). The sequences of selected species were aligned using the algorithm Muscle (MEGA 6.0) [24]. Phylogenetic reconstructions were performed using Bayesian inference (MrBayes version 3.1.2) [25] with the preselection of an adequate model of nucleotide substitutions (MEGA 6.0).

2.8. Mating and Conjugal Transfer of Plasmid DNA. Transformations of *Nostoc* PTV through triparental conjugations followed published protocols [26] with minor modifications. Standard bacterial mating involved the cyanobacterial strain *Nostoc* PTV and *E. coli* strains (DH10B) that harbored the following three plasmids: (i) the conjugal plasmid pRL443 [27], (ii) the "helper" plasmid pRL623, [27], and (iii) the cargo plasmid pRL692 that carries the mobile element Tn5-692 [28]. *E. coli* strains were grown in 3 mL LB with the appropriate antibiotic(s) and incubated at 37°C overnight. Cells of *E. coli* were diluted 1 : 20 and were grown for 1.5–2 h at 37°C. Then *E. coli* cells were harvested from 1 mL of each *E. coli* culture by centrifugation and resuspended in 1 mL fresh LB. This step was repeated twice to wash the cells. After the third centrifugation, the cells were resuspended in 200 mL BG-11. Five milliliters of a growing *Nostoc* PTV culture was harvested by centrifugation at low speed (4000 g) and resuspended in 1 mL BG-11. Then the filaments were fragmented in a water bath sonicator for 2 to 5 min so that more than half of the filaments were shorter than 5 cells.

Table 1: Dynamics of accumulation of vegetative mass of alfalfa inoculated by mono- and binary suspensions of diazotrophic microorganisms.

Inoculants	Phase of plant development					
	Stooling		Budding		Flowering	
	Above-ground mass (g/plant)	Mass of roots (g/plant)	Above-ground mass (g/plant)	Mass of roots (g/plant)	Above-ground mass (g/plant)	Mass of roots (g/plant)
Without inoculation (control)	0,42 ± 0,02	0,12 ± 0,01	1,17 ± 0,06	1,12 ± 0,1	1,25 ± 0,02	2,25 ± 0,16
N. PTV	0,62 ± 0,02	0,18 ± 0,01	1,64 ± 0,09	1,79 ± 0,09	1,70 ± 0,07	2,11 ± 0,15
S. meliloti T17	0,65 ± 0,04	0,15 ± 0,01	1,59 ± 0,13	1,22 ± 0,11	1,83 ± 0,04	2,59 ± 0,24
S. meliloti T17 + *N.* PTV	0,75 ± 0,08	0,22 ± 0,01	1,81 ± 0,18	1,82 ± 0,10	1,92 ± 0,03	2,95 ± 0,29

$P \leq 0{,}05$.

The cyanobacterial cells were collected by centrifugation for 2 min and resuspended in 1 mL BG-11. The cargo strain, the conjugal strain (for triparental mating), and *Nostoc* PTV were combined, pelleted by centrifugation, and finally resuspended in 200 mL BG-11. The conjugation mixture was incubated for about 1 h in low light at 28°C. Then the cells were spread on sterile nitrocellulose filters laid on BG11+ 5% (vol/vol) LB agar plates (mating plates). The mating plates were incubated without antibiotic selection for 18 to 24 h in low light at 28°C, and then the filters were transferred to BG-11 for 24 h and then to BG-11 agar with 10 μg/mL Spectinomycin (Sp10) and 2 μg/mL Streptomycin (Sm2). After incubation for 8 to 12 days, isolated antibiotic-resistant transconjugant colonies were patched on fresh selective BG-11 plates.

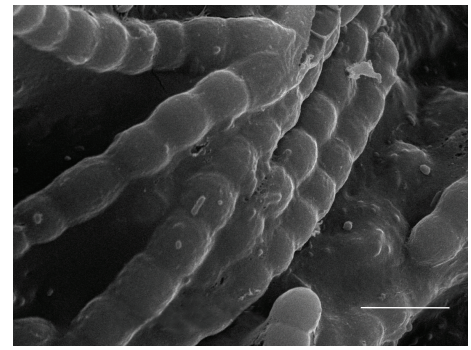


Figure 1: A SEM image of the *N.* PTV cells. Scale bar: 5 μm.

3. Results and Discussion

3.1. Effect of Microbial Inoculation on Plant Growth and Productivity. Earlier in the laboratory study of pure cultures of *N.* PTV and *S. meliloti* [29], we found stimulation of cell growth area of nodule bacteria around the colonies of cyanobacterium *N.* PTV on the surface of the agar medium. Our results are consistent with the literature data, since it is known that cyanobacteria are producers of a wide range of biologically active substances, which include a group of growth-stimulating compounds, analogues of phytohormone [30].

In our previous study we have tested different associations of nitrogen-fixing microorganisms in the rhizosphere of alfalfa [29]. The most effective bacterial consortium included cyanobacterium *Nostoc* PTV and Tn5-mutants of nodule bacteria *S. meliloti.* Usage of the optimal proportions of components in the inoculation mixtures promotes the absence of antagonism between microorganisms and provides the stimulating effect of these consortia on various physiological and biochemical features of alfalfa plants.

In this study the possibility of the formation of artificial stable microbial consortium based on nitrogen-fixing cyanobacterium *N.* PTV (Figure 1) and one Tn5-mutant of nodule bacteria *S. meliloti* T17 was investigated. In pot experiments it was revealed that inoculation of alfalfa by binary mixture of *S. meliloti* T17 + *N.* PTV has a stimulating effect on the growth of the vegetative mass of plants (Table 1). The increase of above-ground plant mass after application of the consortium *S. meliloti* T17 + *N.* PTV in the phase of stem was 15.4% compared with rhizobial T17 monoinoculation and 21% compared with *N.* PTV monoinoculation, respectively. The growth of above-ground alfalfa plant mass was 13.8% and 10.4%, after application of the *S. meliloti* T17 + *N.* PTV consortium in the budding stage, and 5% and 13% in the beginning of flowering, in comparison with monoinoculations, correspondingly. It is known that cells of nitrogen-fixing cyanobacteria produce polypeptides, amino acids, polysaccharides, and vitamins. Due to this diverse biochemical activity in the mucous environment of cyanobacterial cells favorable conditions for growth and reproduction of other microorganisms were created. Perhaps, it could promote a more active cell proliferation of nodule bacteria T17 associated with cyanobacteria in the root zone of alfalfa and contribute to formation of efficient *Rhizobium*-legume symbiosis.

Rhizogenesis was positively affected in plants, the seeds of which were treated with suspensions of microorganisms (Table 1). The largest increase of the plant root mass was detected after using the binary inoculation (*S. meliloti* T17 + *N.* PTV). Thus, in the phase of stem the mass of roots increased by 46.6% and 22.2%, in the budding stage by 49.2% and 13%, and in the early phase of flowering by 13.9% and 39.8%, compared with plants treated only by *S. meliloti* T17 or by only *N.* PTV, correspondingly.

Table 2: Number and mass of root nodules on alfalfa plants inoculated by mono- and binary suspensions of microorganisms.

Inoculants	Phase of plant development					
	Stooling		Budding		Flowering	
	Number of root nodules (pcs/plant)	Mass of root nodules (g/plant)	Number of root nodules (pcs/plant)	Mass of root nodules (g/plant)	Number of root nodules (pcs/plant)	Mass of root nodules (g/plant)
Without inoculation (control)	0	0	0	0	0	0
N. PTV	0	0	0	0	0	0
S. meliloti T17	12,0 ± 1,0	0,010 ± 0,00	30,0 ± 8,5	0,115 ± 0,002	45,0 ± 0,5	0,135 ± 0,02
S. meliloti T17 + *N.* PTV	14,0 ± 0,6	0,017 ± 0,04	57,0 ± 8,0	0,130 ± 0,001	70,0 ± 7,5	0,160 ± 0,02

Note. 15 plants of each variant of the experiment were analyzed for determination the average number of nodules on the roots of one plant.
$P \leq 0{,}05$.

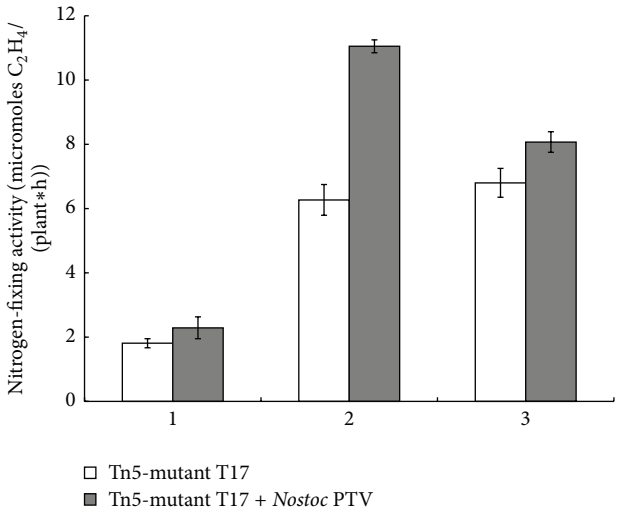


Figure 2: Dynamics of NFA of nodules of alfalfa plants inoculated by mono- and binary suspensions of microorganisms (micromoles of ethylene formed by nodules per plant per 1 h). 1: phase of stooling, 2: phase of budding, and 3: phase of flowering, $P \leq 0{,}05$.

Effective collaboration between all partners of symbiosis provides the activation of several metabolic processes and, above all, the fixation of atmospheric nitrogen. As a result of improved plant nutrition, their productivity increased and the quality of bioproducts improved.

In the phase of stem when the process of nitrogen fixation was still inactive, differences in NFA of nodules of alfalfa plants inoculated by mono- or binary bacterial complexes were not significant (Figure 2). However, in the budding stage and in the early flowering stage a nitrogen fixation in root nodules of plants infected with a mixture of *S. meliloti* T17 + *N.* PTV was more intensive. Positive regulatory role of cyanobacterium *N.* PTV is obvious according to the results presented in Table 2. Only in case of binary inoculation, plants demonstrate an increase in number and in weight of formed nodules (Table 2). Thus, the application of this microbial consortium provided increased NFA in nodules of alfalfa at the budding stage and maintained its relatively high level at the beginning of the early flowering stage (Figure 2). Therefore, this data proves that *N.* PTV has a stimulating effect on the functioning of root nodule bacteria *S. meliloti* T17.

It is known that the use of active strains of root nodule bacteria and their associations with other microorganisms affect the formation and functioning of the photosynthetic complexes through the nitrogen status of a host plant. Presence of nitrogen available to plants determines the efficiency of symbiotic systems.

Mono- and binary suspensions inoculations of seeds showed positive dynamics of accumulation of photosynthetic pigments in the leaves of alfalfa compared with the absolute control (Figure 3). The most significant differences were observed in plants whose seeds had been inoculated with nitrogen-fixing consortium of microorganisms (chlorophyll *a* and chlorophyll *b* increased by 114.6 and 82.9%) compared with the corresponding option of treatment only by strain T17. It is known that the content of chlorophyll in the leaves is directly proportional to the intensity of nitrogen fixation and depends on symbiotic properties of root nodule bacteria [31–33]. Increasing the number of pigments in the leaves of alfalfa inoculated with binary bacterial suspension indicates the ability of *N.* PTV to enhance the functional activity of rhizobia, which are directly interfaced with the intensity of nitrogen fixation.

Available forms of nitrogen, such as mineral and symbiotrophic, positively affect not only the formation of high grade, but also functional state of the plant photosynthetic apparatus. The net assimilation rate also demonstrates the effectiveness of the binary composition *S. meliloti* T17 + *N.* PTV. In particular, in the budding stage of these plants the net assimilation rate exceeded 12.7%, while in the phase of early flowering it increased by 43.7% of the corresponding rate during inoculation only by T17 (Figure 3).

The net assimilation rate of plant leaves typically is closely correlated with the content of nitrogen, and nitrogen is presented mainly in amino acids and proteins. Rubisco, the major cell photosynthetic enzyme of CO_2 assimilation, represents more than half of the soluble cell proteins in leaf. Obviously, the intensification of NFA in the binary composition was the main reason for the increase of plant assimilation

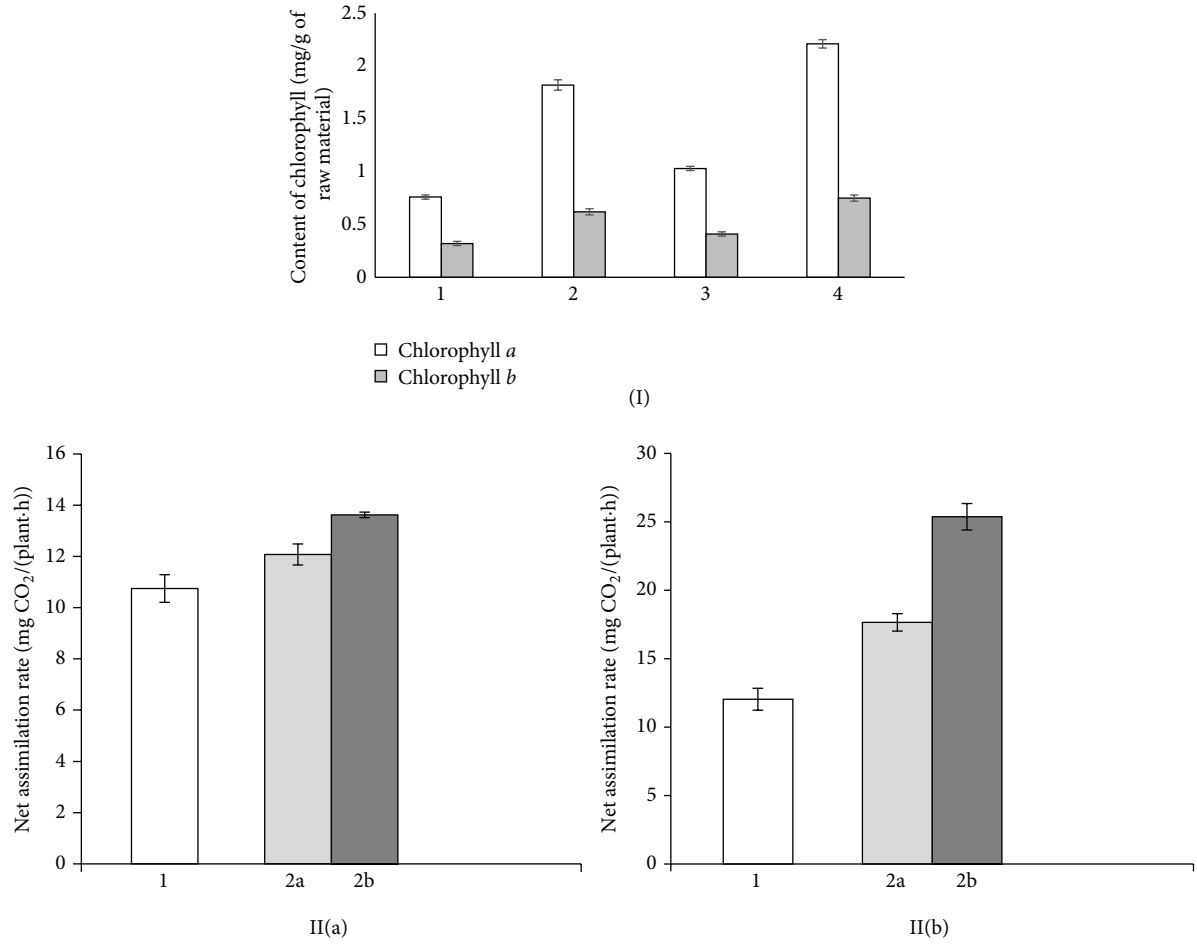


Figure 3: (I): content of chlorophyll (mg/g of raw material) in leaves of alfalfa inoculated by mono- and binary suspensions of microorganisms *S. meliloti* T17 and *N.* PTV. 1: control (without inoculation); 2: inoculation by *N.* PTV; 3: inoculation by *S. meliloti* T17; 4: inoculation by consortium of *S. meliloti* T17 + *N.* PTV. II: net assimilation rate (mg CO_2/(plant·hour)) of alfalfa inoculated by mono- and binary suspensions of microorganisms *S. meliloti* T17 and *N.* PTV. 1: inoculation by *N.* PTV; 2a: inoculation by *S. meliloti* T17; 2b: inoculation by consortium of *S. meliloti* T17 + *N.* PTV (II(a): phase of budding and II(b): phase of flowering).

rate. However, it is possible that more active symbiotic apparatus, which is formed on the roots of plants through binary inoculation, enhanced "request" on assimilates by the root system, thereby stimulating the photosynthetic activity of plant leaves. There is a gradient of transport forms of carbon, particularly sucrose, between roots and leaves in the conduction system and it accelerates the outflow of carbon from the leaves. This, in turn, eliminates restrictions by photosynthesis products imposed on the feedback principle and further accelerates photosynthetic carbon assimilation. Thus, the efficient operation of the symbiotic apparatus in inoculated plants greatly stimulated the accumulation of photosynthetic pigments and increased the net assimilation rate. The accumulation of organic matter contributes to the formation of the plant biomass, because the basis of the biological productivity of the plant organism, including those capable of symbiotic nitrogen fixation, is photosynthetic carbon assimilation [31].

In consequence of artificial inoculation of alfalfa seeds by consortium of nitrogen-fixing microorganisms *S. meliloti* T17 + *N.* PTV the yield of green mass of plants increased by 17.9% and the protein content in the leaves increased by 12.0% compared to monoinoculation by strain T17 (Table 3). This is an evidence of the effective interaction of test organisms in the cyano-*Rhizobium* associations and their positive impact on the growth and physiological characteristics of alfalfa plants (Table 3).

The amino acid composition is the main criterion of the biological value of proteins. An index of a total amino acid composition of vegetative mass of the experimental inoculated variants of alfalfa plants increased in comparison to the control (without bacterial inoculation). The maximum quantity of lysine, the most essential and deficient amino acid in humans and animals, was recorded in leaves of alfalfa (Table 4). As a result of using binary inoculation a total amino acid composition increased by 25.1%, compared with the case of inoculation only by T17. In particular, a quantity of essential amino acids increased by 33.9%, and a quantity of nonessential amino acids increased by 17.7% (Figure 4). At the same time, an increase of the content of methionine,

Table 3: Productivity and protein content in leaves of alfalfa, inoculated by mono- and binary suspensions of microorganisms.

Inoculants	Harvest of green mass of plant, g/vessel				Protein content in the leaves	
	I mowing	II mowing	Total harvest	% to monoinoculation by rhizobium		% to monoinoculation by rhizobium
Control	17,68 ± 0,51	19,70 ± 0,64	37,38		13,2	
N. PTV	21,40 ± 0,52	24,22 ± 0,76	45,62		14,6	
S. meliloti T17	21,81 ± 0,48	25,17 ± 0,30	46,98		18,32	
S. meliloti T17 + *N*. PTV	26,26 ± 0,37*	29,15 ± 0,17*	55,41	**117,9**	20,52	**112,0**

Table 4: Amino acid content in leaves of alfalfa, inoculated by mono- and binary suspensions of microorganisms.

Amino acid	Content of amino acids (mg/100 mg DW)			
	Control	*N. punctiforme*	T17	T17 + *N. punctiforme*
Gamma-aminobutyric acid	0,065	0,088	0,085	0,123
Lysine	**0,386**	**0,802**	**0,459**	**0,610**
Histidine	0,094	0,301	0,171	0,232
Arginine	0,297	0,905	0,439	0,585
Asparagine	0,675	0,882	0,748	0,779
Threonine	0,278	0,623	0,383	0,496
Serine	0,321	0,698	0,413	0,533
Glutamic acid	0,876	1,931	1,083	1,395
Proline	0,376	0,610	0,419	0,548
Glycine	0,423	0,790	0,470	0,537
Alanine	0,479	0,895	0,581	0,644
Cysteine	0,054	0,262	0,057	0,079
Valine	0,245	0,651	0,301	0,423
Methionine	0,111	0,298	0,149	0,200
Isoleucine	0,175	0,439	0,233	0,273
Leucine	0,523	1,234	0,685	0,938
Tyrosine	0,201	0,534	0,285	0,253
Phenylalanine	0,398	0,835	0,453	0,626
Total	**5,977**	**12,779**	**7,417**	**9,276**

histidine, arginine, and tyrosine was observed, which are present in small quantities in plant leaves, and this is one of the factors limiting the rate of biosynthesis of proteins, especially in generative organs. The results are a direct proof of the positive impact of cyanobacterial inoculation on the quality of agricultural products.

3.2. Stress Response of Plants Inoculated with Microbial Consortium. It is known that the plant productivity rate and resistance index are inversely dependent values. A positive effect of the cyanobacterial consortium T17 + *N*. PTV on the productivity of alfalfa is shown in our study. It was logical to study the effect of the binary inoculation on plant resistance to the adverse effects of certain environmental factors. It is an extremely important issue. We have used herbicide diquat as a model stress factor. For a short duration (30 minutes) of diquat treatment the content of TBA-reactive products in photosynthetic tissues of plants that were inoculated with the strain T17 was reduced by 18% compared to plants without inoculation (control 2). In the case when plants were inoculated with the consortium *S. meliloti* T17 + *N*. PTV the content of TBA-reactive products remained at the level of control. In the experiments with more prolonged action of the stress factor (60 min) the content of TBA-reactive products in plants inoculated only by the strain T17 decreased by 16.9% and in the case of binary inoculation the content of TBA-reactive products decreased by 25%. It should be noted that after 24 h of plants exposure with diquat, regardless of the inoculation agent used, reducing the amount of TBA-active products in photosynthetic tissues was not observed compared with the control. At the same time, in the experiment with the use of the consortium of microorganisms a difference (15.4%) with the inoculation only by strain T17 was marked (Figure 5). At short-term action of diquat (30 min) a rate of SOD activity in photosynthetic tissues of plants inoculated with *S. meliloti* T17 + *N*. PTV was 2.5 times higher than in controls and by 42.5% in plants, inoculated by strain T17. After herbicide treatment during 60 minutes a significant altering of SOD activity in inoculated plants (irrespective of whether a mono- or binary inoculation) was

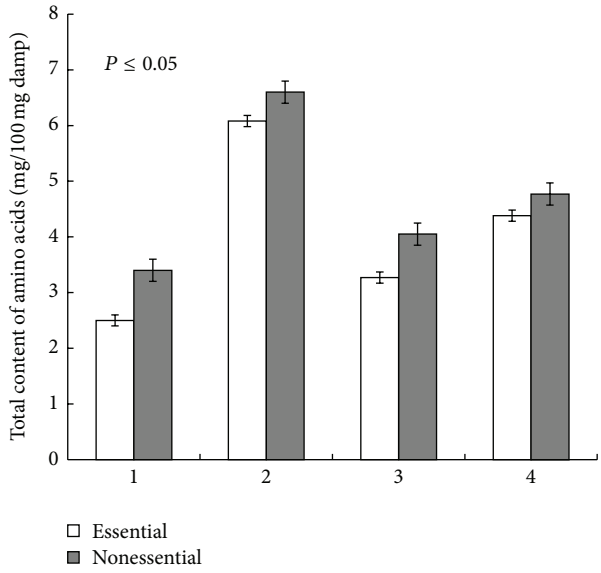


Figure 4: Total content of essential and nonessential amino acids in leaves of alfalfa grown after mono- and binary inoculation by cyanorhizobial compositions of microorganisms: 1: control (without inoculation), 2: monoinoculation of alfalfa seeds by cyanobacterium *N.* PTV, 3: inoculation of alfalfa seeds by Tn5-mutant strain of *S. meliloti* T17, and 4: binary inoculation of alfalfa seeds by Tn5-mutant strain of *S. meliloti* T17 + *N.* PTV.

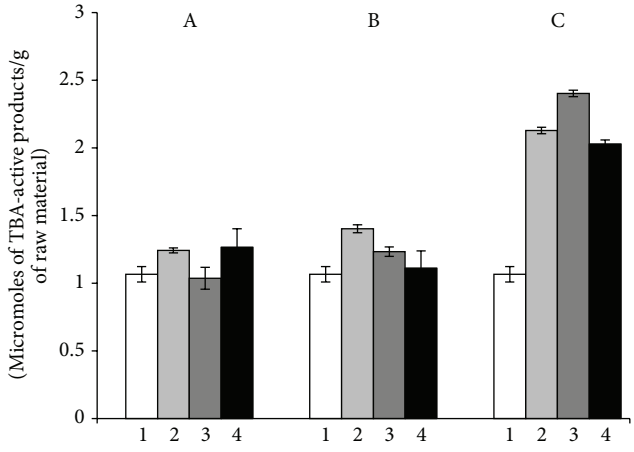


Figure 5: Content of TBA-active products in alfalfa leaves after herbicide diquat treatment: A: for 30 min, B: for 60 min, and C: for 24 h. 1: control (without inoculation and herbicide diquat treatment); 2: control (without inoculation, with herbicide diquat treatment); 3: inoculation by Tn5-mutant of *S. meliloti* T17, with herbicide diquat treatment; 4: inoculation by Tn5-mutant of *S. meliloti* T17+ *N.* PTV, with herbicide diquat treatment.

not observed. However, in comparison to the control, this difference was significant; the enzyme activity was increased by 54%. Under long-term stress (24 h) in plants inoculated with strain T17, SOD activity remained at the same level as for short-term exposure. In plants, the seeds were treated with consortium *S. meliloti* T17 + *N.* PTV, for the same conditions; this index decreased by 23.4% compared with

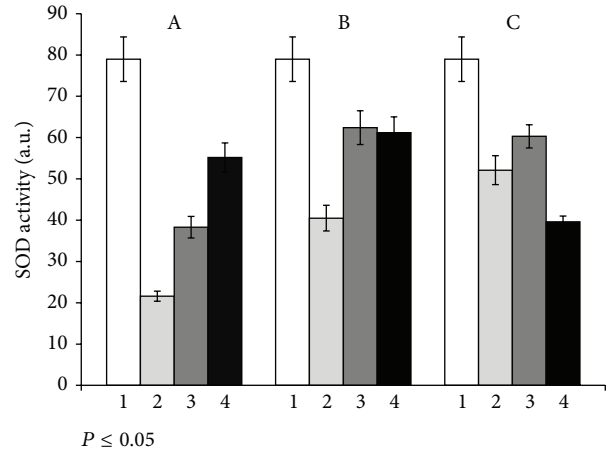


Figure 6: SOD activity in alfalfa leaves after herbicide diquat treatment: A: in 30 min, B: in 60 min, and C: in 24 h. 1: control (without inoculation and herbicide diquat treatment); 2: control (without inoculation, with herbicide diquat treatment); 3: inoculation by Tn5-mutant of *S. meliloti* T17, with herbicide diquat treatment; 4: inoculation by T17 + *N.* PTV, with herbicide diquat treatment.

the control and 34.4%, in comparison with the plants inoculated by strain T17 (Figure 6). Thus, the plants inoculated with algae-rhizobial composition proved to be more resistant to oxidative stress. It is possible due to the increased level of NFA of their symbiotic system and thus the increase in the number of available forms of nitrogen for alfalfa plants and the possible participation of NO in the defense reactions. In the literature, there are two hypotheses about the mechanisms of NO action under conditions of stress. First, NO may act as an antioxidant, directly linking to ROS, thereby protecting cells from damaging their actions [34]. Secondly, NO can act as a signaling molecule that triggers a cascade of reactions that lead to the expression of specific genes [35]. In their chemical and physical properties small molecule, rapid metabolism, lack of charging, and high diffusion coefficient of NO are well suited for the role of intracellular signaling mediator of plant stress responses.

Thus, the inoculation of alfalfa seeds by a consortium of nitrogen-fixing microorganisms *S. meliloti* T17 + *N.* PTV increased the nitrogenase activity of root nodules, increased the net assimilation rate, and increased productivity and product quality, and also the stability of alfalfa plants under the influence of oxidative stress induced by herbicides.

3.3. Molecular Typing and Phylogenetic Analysis of Cyanobacterium Nostoc PTV. One of the main goals of this study was the molecular typing and phylogenetic analysis of a new cyanobacterial isolate *N.* PTV originated from the Institute of Hydrobiology of Academy of Science of Ukraine. As it was shown above, this cyanobacterium is effective for soil algalization. As a component of algae-rhizobium compositions, this cyanobacterium stimulates germinative energy, growth, and productivity of legumes.

To identify and to determine the phylogenetic positions of the new cyanobacterial isolate *N.* PTV we used a partial

sequence of the *nifH* gene (342 bp), encoding nitrogenase reductase, and 16S ribosomal RNA gene cluster (1765 bp) as molecular markers. Comparison of the *nifH* gene sequence and rRNA gene cluster sequence of cyanobacterium *N.* PTV with all the sequences present at the GenBank by using the program Blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) shows that this strain has no full similarity with any of early investigated organisms.

Comparison of rRNA gene cluster sequence (1765 bp) of the cyanobacterium *Nostoc* sp. PTV revealed that this cyanobacterium shows the highest similarity with several strains of *N. muscorum* and *N. commune* (Table 5). In general, support for branching in the tree, based on a fragment of 16S rRNA gene sequence (Figure 7), is worse in comparison with the reconstruction of the phylogeny of cyanobacteria based on the sequence of the gene *nifH* (Figure 8).

The group of *Nostoc* strains and species, which includes *Nostoc* PTV, forms a cluster with the minimum of allowable support, 0.95. Hierarchy of strains (HA 4355-MV2, PTV, and 8964) and *Nostoc* species (*N. muscorum* and *N. linckia*) cannot be evaluated because of the low topological support of this site of the tree: credibility values range from 0.56 to 0.94. It is difficult to discuss a relation of *Nostoc* PTV strain to strain HA 4355-MV2, due to the very low values of the other branches in the cluster, and very scarce information about the HA-MV2 4355 strain clearly does not help to solve the problem of their possible relationship.

N. muscorum is discussed below. *N. linckia* is also a freshwater strain and could be isolated in some terrestrial niches. Interestingly, the *Nostoc* strain UAM 307 is quite clearly differentiated from the other representatives of Nostocaceae (0.95). Being sufficiently close to strain of *Nostoc* PTV, this cyanobacterium has some significant features that provided the separation of this strain into the other branch of the common with *Nostoc* PTV cluster.

Even more interesting detail is that the last significant branch (credibility value is equal to unity) of the cluster is formed by the strain of *N. muscorum* Ind33, which is significantly aside not only from the desired strain of PTV, but also from the different strains of the same species (CCAP 1453-22). Thus, the strain of *Nostoc* PTV has teamed up with members of their own genus.

Outgroup of this dendrogram is represented by two strains of *Rivularia* (this kind of cyanobacteria forms heteropolar threads; their trichomes are densely agglomerated, covered with a total mucus). The genus is represented only by the species often living on calcareous substrates, but there are rare epilithic and epiphytic species.

Comparison of *nifH* sequence of cyanobacterium PTV revealed that this strain shows the highest similarity with several strains of *Nostoc* (Table 6). The closest relative is *Nostoc muscorum* UTAD N213, purified from rice paddy in Mondego River Basin (Portugal). Phylogenetic analysis (Figure 8) revealed that the cyanobacterium PTV forms a minicluster with *Nostoc muscorum* UTAD N213. *N. muscorum* is a free-living filamentous cyanobacterium, which inhabits both terrestrial and freshwater aquatic environments. They are phototrophic organisms performing photosynthesis and also fixing atmospheric nitrogen [36].

N. muscorum is the most common type of *Nostoc* in terrestrial ecosystems and is widely spread, due to the adaptability to many adverse conditions. It forms a symbiotic relationship with many types of terrestrial plants and fungi.

It is known that *N. muscorum* has great effect on soil biology and productivity which makes it an attractive soil inoculant. This cyanobacterium is able to obtain carbon and nitrogen from the air and has an advantage over heterotrophic soil inoculants, which are usually limited by carbon [37].

It also benefits plants and other soil bacteria by increasing soil organic matter in the form of carbohydrates and provides biological organic nitrogen that can be assimilated by plants [38]. *N. muscorum* helps to create the environment conditions to further colonization and growth by plants and other microorganisms [39].

Inoculation of the *N. muscorum* isolates caused a significant effect on growth of wheat and maize plants. Cyanobacterial inoculation positively affected pigment content, increased plant shoot and root dry weight, and increased leaf area [40].

In general, the topology of the dendrogram (Figure 8) has a good support; the main branch nodes are characterized by high credibility values.

Nostoc PTV is in the same cluster with *Nostoc* sp. UAM-362 and *Nostoc commune*. *Nostoc* sp. UAM-362 was isolated from the rock surface of calcareous river with brackish water in Spain: Muga, Girona (Northeast Spain). *N. commune* is a colonial species of cyanobacterium. As well as *N. punctiforme, N. commune* is able to survive in extreme conditions such as polar regions and arid areas.

Three more clusters are presented as the parts of the large one: two single clusters, represented by *Nostoc* sp. Baikal (nitrogen-fixing cyanobacteria from Lake Baikal) and by *Nodularia spumigena* from family Nostocaceae. *Nodularia* occurs mainly in brackish or saline waters. *Nodularia* cells occasionally can form heavy algal blooms. Some strains produce a toxin (nodularin), which is harmful to human health [41].

The third cluster is formed by four strains of *Tolypothrix* and by one strain of *Nostoc* sp. UAM-367 (isolated from rock surface of calcareous river with brackish water in Spain: Muga, Girona). Cyanobacteria *Tolypothrix* grow in unpolluted waters; several species are found in swamps, known aerophilic species growing on the bark of trees, in the wet sands, on wet rocks, and so forth.

Two species, and one strain (PCC 7120) of *Anabaena*, representing a family of filamentous cyanobacteria Nostocaceae, belong to this large cluster with a poor resolution (a credibility value of branching is 0.71). These cyanobacteria exist in the form of plankton; some species are symbionts of plants. *Anabaena* is one of the four genera of cyanobacteria that produce neurotoxins. *Anabaena* is a model for the study of cell differentiation and differential gene expression during nitrogen fixation [42]. *A. siamensis* and *A. sphaerica* are freshwater species. This mega cluster consisting of described species of cyanobacteria is well differentiated from the other

Table 5: BLAST results obtained by querying the 16S–23S rRNA gene cluster of *Nostoc* sp. PTV with GenBank and geographical and ecological origins of the hits.

Closest GenBank relative	GenBank access number	Query coverage, %	Score, %	Identity, %	*E* value	Origin of the strain and reference
Nostoc sp. HA4355-MV2 clone p9D	HQ847576	98	2872	97	0.0	Maniniholo Cave wall, near Haena USA: Kauai, Hawaii
Nostoc muscorum CCAP 1453/22	HF678509	98	2531	93	0.0	Scottish Association for Marine Science, Molecular and Microbial Biology, Dunstaffnage Marine Laboratory, Oban, PA37 1QA, United Kingdom
Anabaena variabilis ATCC 29413	CP000117	98	2457	92	0.0	It has been studied extensively for over 40 years and is the strain of choice for many laboratories throughout the world
Nostoc sp. PCC 7120	BA000019	98	2453	92	0.0	Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. (*Nostoc*) strain PCC 7120 is available
Nostoc commune NC1 clone 10	EU586726	98	2441	91	0.0	John Carroll University, 20700, North Park Boulevard, University Heights, OH 44118, USA
Nostoc commune NC1 clone M2	EU586725	98	2441	91	0.0	John Carroll University, 20700, North Park Boulevard, University Heights, OH 44118, USA
Nostoc commune NC5 clone 10	EU586727	98	2437	91	0.0	Biology, John Carroll University, 20700, North Park Boulevard, University Heights, OH 44118, USA
Nostoc commune NC1	EU784149	98	2428	91	0.0	Nostoc commune NC1 was isolated from soil (subaerophyt) in Třeboň/Czech republic in 2006.
Nostoc commune NC5 clone 11	EU586728	98	2423	91	0.0	Biology, John Carroll University, 20700, North Park Boulevard, University Heights, OH 44118, USA
Nostoc ellipsosporum CCAP 1453/15	HE975023	94	2399	92	0.0	Scottish Association for Marine Science, Molecular and Microbial Biology, Dunstaffnage Marine Laboratory, Oban PA37 1QA, United Kingdom
Nostoc cf. *punctiforme* Bashkir clone 6A	EU586732	96	2378	92	0.0	John Carroll University, 20700, North Park Boulevard, University Heights, OH 44118, USA
Calothrix sp. HA4356-MV2 clone p8i	JN385289	98	2361	90	0.0	Cave wall scraping, Maniniholo Cave near Haena, USA: Kauai, Hawaii
Calothrix sp. HA4340 LM2	KF417425	98	2361	90	0.0	Cave, USA: Kauai, Hawaii, Maniniholo Cave
Anabaena sp. CCAP 1403/4A	HE975015	93	2360	92	0.0	Scottish Association for Marine Science, Molecular and Microbial Biology, Dunstaffnage Marine Laboratory, Oban PA37 1QA, United Kingdom
Nostoc muscorum CCAP 1453/8	HF678508	93	2358	92	0.0	United Kingdom: Scotland
Calothrix sp. HA4356-MV2 clone p8i	JN385289	98	2356	90	0.0	Cave wall scraping, Maniniholo Cave near Haena, USA: Kauai, Hawaii
Cylindrospermum moravicum CCALA 993 clone operon 1	KF052607	97	2331	94	0.0	Cave sediment, Czech Republic: Amaterska Cave, South Moravia
Nostoc punctiforme PCC 73102	CP001037	97	2331	91	0.0	A symbiont from a cycad
Nostoc sp. *Peltigera malacea* DB3992 cyanobiont	JX219483	97	2322	91	0.0	Cyanobiont of lichenized fungi *Peltigera malacea*, Iceland
Nostoc muscorum CCAP 1453/20	HF678506	95	2318	91	0.0	Scottish Association for Marine Science, Molecular and Microbial Biology, Dunstaffnage Marine Laboratory, Oban PA37 1QA, United Kingdom

Table 5: Continued.

Closest GenBank relative	GenBank access number	Query coverage, %	Score, %	Identity, %	*E* value	Origin of the strain and reference
Nostoc sp. 10Dp66E	JQ259187	98	2318	90	0.0	From association with *Dynamena pumila* L., White Sea, Russia
Cylindrospermum catenatum CCALA 999 clone operon 1	KF052615	97	2309	94	0.0	Soil, Slovakia: forest above Stara Brzotinska Cave, Slovak Karst
Trichormus anomalus HA4352 LM2	KF417426	96	2309	92	0.0	Cave, USA: Kauai, Hawaii, Maniniholo Cave
Cylindrospermum sp. HA4236-MV2 clone p4	JN385290	98	2309	89	0.0	Taro field, Makiki Nature Center, USA: Oahu, Hawaii
Cylindrospermum catenatum CCALA 996 clone operon 1	KF052611	97	2307	94	0.0	Soil, Czech Republic: Amaterska Cave, South Moravia
Tolypothrix campylonemoides FI5-MK38 clone p10D	JQ083654	98	2305	90	0.0	Sand, USA: Fort Irwin NTC, San Bernardino Co., California
Cylindrospermum catenatum CCALA 990 clone operon 1	KF052601	95	2302	94	0.0	Soil, Czech Republic: Benešov nad Černou, South Bohemia
Spirirestis rafaelensis WJT-71-NPBG6 clone p1B	JQ083656	98	2300	90	0.0	Joshua Tree National Park, USA: Joshua Tree Forest, San Bernardino Co., California
Cylindrospermum badium CCALA 1000 clone operon 1	KF052616	97	2298	94	0.0	Reclaimed coal mine soil, USA: Pyramid State Recreation Area, Illinois
Cylindrospermum catenatum CCALA 991 clone operon 1	KF052603	97	2293	94	0.0	Soil, Czech Republic: Most Region, North Bohemia
Spirirestis rafaelensis WJT-71-NPBG6 clone p1A	JQ083655	98	2291	89	0.0	Joshua Tree National Park, USA: Joshua Tree Forest, San Bernardino Co., California
Nostoc sp. *Peltigera membranacea* cyanobiont N6	JX975209	97	2289	91	0.0	Symbiont of *Peltigera membranacea* lichen, Iceland
Cylindrospermum pellucidum CCALA 992 clone operon 1	KF052605	97	2287	94	0.0	Cave sediment, Slovakia: Dlha chodba in Domica Cave system, Slovak Karst
Hassallia sp. EM2-HA1 clone p4B	HQ847555	98	2282	89	0.0	Soil, Mojave National Preserve, USA: San Bernardino Co., California
Tolypothrix tenuis f. *terrestris* UFS-BI-NPMV-1A2-F06 clone p13E	JQ083651	98	2277	89	0.0	Arid soil after a burn, foothills of the Onaquee Mts. USA: Utah
Nostoc cf. *commune* 257-16	HQ877826	96	2271	90	0.0	Subaerial, on Bonampak's archeological building walls, Mexico: Chiapas
Hassallia sp. CNP3-B3-C04 clone p5D	HQ847556	98	2271	89	0.0	Soil, Needles District, Virginia Park, Canyonlands National Park, USA: San Bernardino Co., California
Uncultured cyanobacterium clone Emix3.12	JX887892	92	2269	91	0.0	Freshwater microbial mat Konstanz, Germany
Cylindrospermum muscicola SAG 44.79 clone operon 1	KF111150	96	2268	93	0.0	Soil France: Gif-Sur-Yvette, Ile-de-France Region
Tolypothrix tenuis f. terrestris UFS-BI-NPMV-1A2-F06 clone p13F	JQ083652	98	2268	89	0.0	Arid soil after a burn, foothills of the Onaquee Mts. USA: Utah
Hassallia sp. CM1-HA11 clone p8A	JQ083650	98	2268	89	0.0	Sandy loam near gypsum mine, USA: Clark Mountains, San Bernardino Co., California

Table 5: Continued.

Closest GenBank relative	GenBank access number	Query coverage, %	Score, %	Identity, %	*E* value	Origin of the strain and reference
Hassallia sp. CM1-HA08 clone p7B	JQ083648	98	2268	89	0.0	Sandy loam near gypsum mine, USA: Clark Mountains, San Bernardino Co., California
Nostoc cf. *commune* 257-20	HQ877827	94	2266	91	0.0	Biofilms of N. cf. commune were collected at Bonampak archeological area in 2008 from two sites on the building walls (Chiapas, Mexico).
Tolypothrix campylonemoides FI5-MK38 clone p10A	JQ083653	98	2266	89	0.0	Sand USA: Fort Irwin NTC, San Bernardino Co., California
Hassallia sp. CM1-HA08 clone p7F	HQ847554	98	2266	89	0.0	Soil, Clark Mountains, near gypsum mine USA: San Bernardino Co., California
Desmonostoc sp. HA7617 LM4	KF417429	96	2241	90	0.0	USA: Kauai, Hawaii, Waikapalae Cave
Camptylonemopsis sp. HA4241-MV5 clone B2-3 + p4	JN385292	96	2223	93	0.0	Moleka stream USA: Oahu, Hawaii
Anabaena circinalis 33-10 isolate	EF634474	98	2199	88	0.0	Ohau Channel, New Zealand
Tolypothrix sp. PCC 7504 isolate DBSU 18	FJ660999	90	2194	92	0.0	Freshwater Aquarium, Sweden
Nostoc sp. UAM 307	HM623782	70	2111	98	0.0	Rock surface of calcareous river Spain: Matarranya River, Teruel, East Spain
Rivularia sp. 1PA23	FJ660980	92	2021	88	0.0	Pozas Azules I, Mexico Microbialite freshwater
Nostoc sp. 8964:3	AM711541	66	2006	99	0.0	Host is *Gunneraprorepens* (Angiospermae), New Zealand
Rivularia sp. 1PA21	FJ660978	92	2004	88	0.0	Institute of Ecology, UNAM (Mexico)
Nostoc linckia var. *arvense* IAM M-30	AB325907	65	1999	99	0.0	Cultivated samples from the Institute of Molecular Biosciences at the University of Tokyo
Aphanizomenon ovalisporum ILC-164	JF768745	92	1988	93	0.0	Lake Kinneret, Israel; Banker et al., 1997 [5]
Nostoc muscorum Ind33	HM573462	65	1988	99	0.0	Paddy field, India: Agricultural Farms, Banaras Hindu University, Varanasi, Uttar Pradesh

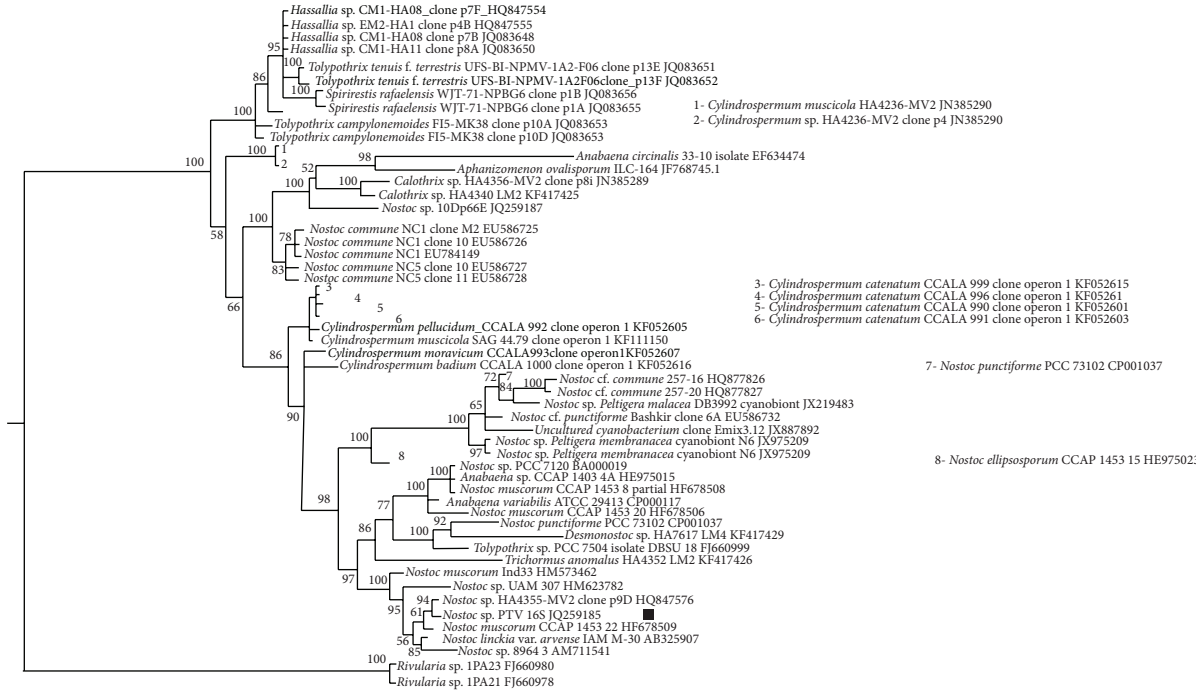


Figure 7: Phylogenetic relationships of *Nostoc* sp. PTV (designated by black square) inferred under the posterior probability criterion (MrBayes) from the gene for 16S rRNA, partial sequence information. Numbers at the nodes indicate the Bayesian statistical support values (posterior probabilities multiplied by 100); only values higher than 50% are given. The scale bar indicates the number of substitutions per nucleotide position.

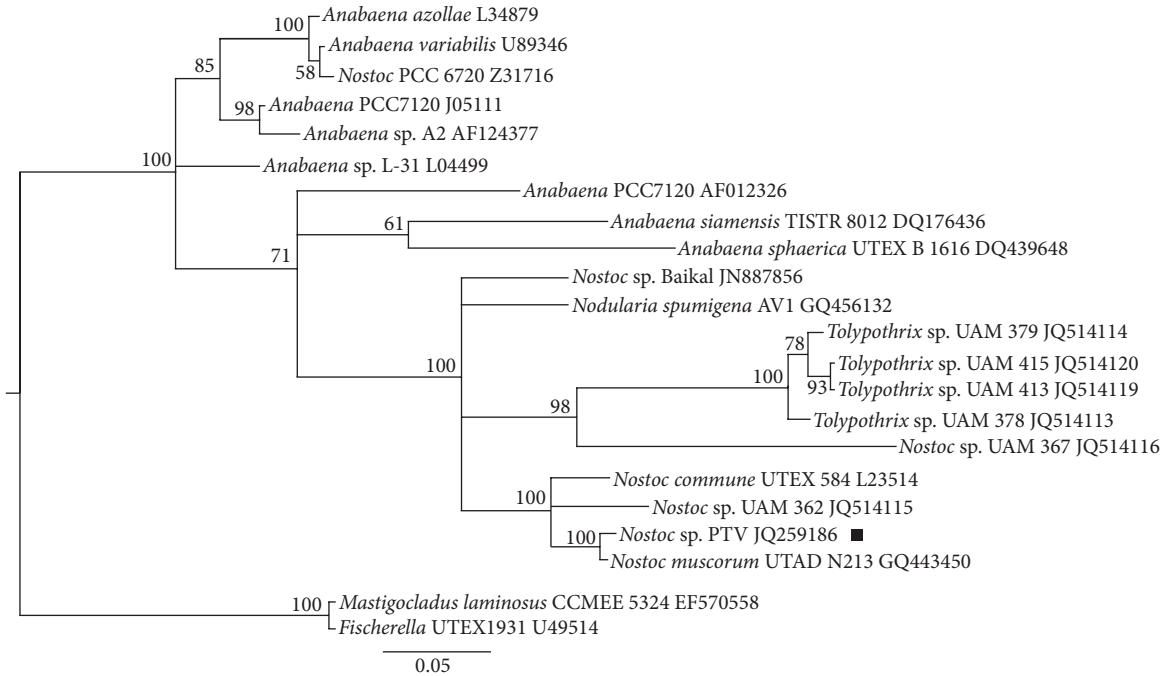


Figure 8: Phylogenetic relationships of *Nostoc* sp. PTV (designated by black square) inferred under the posterior probability criterion (MrBayes) from the gene for *nifH*, partial sequence information. Numbers at the nodes indicate the Bayesian statistical support values (posterior probabilities multiplied by 100); only values higher than 50% are given. The scale bar indicates the number of substitutions per nucleotide position.

Table 6: BLAST results obtained by querying the *nifH* gene of *Nostoc* sp. *PTV* with GenBank and geographical and ecological origins of the hits.

Closest GenBank relative	GenBank number	Query coverage %	Score %	Identity %	*E* value	Origin of the strain and reference
Nostoc muscorum UTAD N213	GQ443450.1	100	612	99	7*e* − 172	Rice paddy in Mondego River Basin, Portugal
Nostoc muscorum clone CC1090A1	AY221814.1	94	576	99	5*e* − 161	Ocean Sciences Department, University of California, Santa Cruz, CA 95064, USA
Nostoc commune (UTEX 584)	L23514.1	99	565	97	9*E* − 158	Scotland
Nostoc sp. UAM 362	JQ514115.1	99	553	96	5*e* − 154	Rock surface of calcareous river with brackish water, Spain: Muga, Girona (Northeast Spain)
Nostoc sp. Baikal	JN887856.1	99	517	94	4*e* − 143	Nitrogen-fixing cyanobacteria from Lake Baikal
Nodularia spumigena AV1	GQ456132.1	99	511	93	2*e* − 141	The surface waters of the Baltic Sea, Stockholm, Sweden
Tolypothrix sp. UAM 379	JQ514114.1	99	462	90	8*E* − 127	Rock surface of calcareous river with brackish water, Spain: Muga, Girona (Northeast Spain)
Tolypothrix sp. UAM 378	JQ514113.1	99	462	90	8*E* − 127	Rock surface of calcareous river with brackish water, Spain: Muga, Girona (Northeast Spain)
Tolypothrix sp. UAM 415	JQ514120.1	99	453	89	4*E* − 124	Rock surface of calcareous river with brackish water, Spain: Muga, Girona (Northeast Spain)
Tolypothrix sp. UAM 413	JQ514119.1	99	453	89	4*E* − 124	Rock surface of calcareous river with brackish water, Spain: Muga, Girona (Northeast Spain)
Nostoc PCC 6720	Z31716.1	99	430	88	4*E* − 117	*Nostoc* PCC 6720 was previously known as *Anabaenopsis circularis*. This is a freshwater species
Anabaena sp. L-31	L04499.1	99	426	88	5*E* − 116	The filamentous, heterocystous, nitrogen-fixing freshwater cyanobacterium
Anabaena PCC7120	J05111.1	99	426	88	5*E* − 116	http://wiki.annotation.jp/Kazusa:CyanoBase:Anabaena_sp._PCC_7120
Nostoc sp. UAM 367	JQ514116.1	100	423	87	7*E* − 115	Rock surface of calcareous river with brackish water, Spain: Muga, Girona (Northeast Spain)
Anabaena siamensis TISTR 8012	DQ176436.2	100	414	87	3*E* − 112	*Anabaena siamensis* is a filamentous heterocystous nitrogen-fixing cyanobacterium which originally was isolated from a rice paddy field in Thailand
Mastigocladus laminosus CCMEE 5324	EF570558.1	100	414	87	3*E* − 112	The cosmopolitan thermophilic cyanobacterium *Mastigocladus laminosus* from the University of Oregon's Culture Collection of Microorganisms from Extreme Environments (CCMEE)
Fischerella UTEX1931	U49514.1	100	414	87	3*E* − 112	Thermophilic cyanobacterium (synonym: *Mastigocladus laminosus*)
Anabaena sphaerica UTEX "B 1616"	DQ439648.1	99	408	87	1*E* − 110	Department of Chemistry and Chemical Engineering, University of Sheffield, Mappin Street, Sheffield, South Yorkshire S1 3JD, United Kingdom
Anabaena sp. A2	AF124377.1	99	408	87	1.*E* − 110	Molecular Evolution, BMC, Uppsala University, Husargatan 3, 751 24 Uppsala, Sweden
Anabaena azollae	L34879.1	99	435	88	1*E* − 118	*Anabaena azollae* 1a, a putative symbiont of *Azolla caroliniana*
Anabaena variabilis	U89346.1	99	435	88	1*E* − 118	*Anabaena variabilis ATCC* 29413 is a filamentous cyanobacterium that produces heterocysts and fixes nitrogen under a variety of environmental conditions

two clusters: a single one represented by strain *Anabaena* L-31 (freshwater cyanobacterium) and a poorly differentiated cluster, which consists of two strains of *Anabaena* (A2 and PCC 7120) and two well-differentiated species of *Anabaena, A. azollae* and *A. variabilis*. *A. azollae* forms symbiosis with water fern *Azolla*.

Outgroup for the described species of *Nostoc* and *Anabaena* is represented by *Mastigocladus laminosus* and by member of the genus *Fischerella* (strain UTEX 1931). The first organism is a typical representative of the genus *Mastigocladus*. *Fischerella* represents another squad, Stigonematales. Both types of reference are truly branching filamentous forms of thermophilic cyanobacteria.

3.4. Gene Transfer into Nostoc PTV Cells. *Nostoc* PTV cells were tested for their ability to conjugational DNA transfer. As a result of plasmid pRL692 transfer into cyanobacterial cells several hundred transconjugant colonies were selected on selective plates that contained solid BG-11 medium and antibiotics (Sp10 and Sm2). On control plates antibiotic-resistant colonies were absent (data not shown). In future experiments we plan to use this experimental approach for transposon mutagenesis of *Nostoc* PTV and selection of the new mutants with interesting characteristics.

4. Conclusions

The use of microbial consortium newly identified *Nostoc* PTV strain together with *Sinorhizobium meliloti* T17 was more efficient than the use of the single-rhizobium strain for alfalfa plant inoculation. This treatment provides an intensification of the processes of nitrogen fixation and photosynthesis and stimulates growth of above-ground plant mass and rhizogenesis and leads to increased productivity of *Medicago sativa* L. and improved amino acid composition of plant leaves. Phylogenetic analysis by using two different molecular markers showed that this new cyanobacterium belongs to a cluster of the genus *Nostoc*, with the closest relative of *Nostoc muscorum*. Gene transfer of transposon bearing plasmid DNA has been shown for cyanobacterium *Nostoc* PTV. It makes this strain very attractive model for future genetic and physiological experiments and biotechnological applications.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was partly supported by Grant no. 14-04-00656 of the Russian Foundation for Basic Research. Authors devote this study to memory of Professor Tatjana V. Parshikova, who initiated this research.

References

[1] A. N. Rai, E. Söderbäck, and B. Bergman, "Cyanobacterium—plant symbioses," *New Phytologist*, vol. 147, no. 3, pp. 449–481, 2000.

[2] C. Santi, D. Bogusz, and C. Franche, "Biological nitrogen fixation in non-legume plants," *Annals of Botany*, vol. 111, no. 5, pp. 743–767, 2013.

[3] S. R. Subashchandrabose, B. Ramakrishnan, M. Megharaj, K. Venkateswarlu, and R. Naidu, "Consortia of cyanobacteria/microalgae and bacteria: biotechnological potential," *Biotechnology Advances*, vol. 29, no. 6, pp. 896–907, 2011.

[4] L. J. Stal, "Physiological ecology of cyanobacteria in microbial mats and other communities," *New Phytologist*, vol. 131, no. 1, pp. 1–32, 1995.

[5] R. Banker, S. Carmeli, O. Hadas, B. Teltsch, R. Porat, and A. Sukenik, "Identification of cylindrospermopsin in *Aphanizomenon ovalisporum* (Cyanophyceae) isolated from lake kinneret, Israel," *Journal of Phycology*, vol. 33, no. 4, pp. 613–616, 1997.

[6] S. Y. Kots, N. A. Vorobey, S. M. Malichenko, and I. M. Butnitskiy, "Strain of bacteria *Sinorhizobium meliloti* T17 (IMB B-7282) to produce of the bacterial fertilizers for alfalfa," Patent for utility of the model N 55432. Bulletin no. 23, 2010.

[7] T. I. Novikova, L. A. Sharypova, and B. V. Simarov, "Transposon mutagenesis of the strain CXM1-105 *Rhizobium meliloti*," *Molecular Genetics Microbiology and Virology*, vol. 8, pp. 32–35, 1986.

[8] A. Zehnder and P. R. Gorham, "Factors influencing the growth of *Microcystis aeruginosa* Kutz Emend. Elenkin," *Canadian Journal of Microbiology*, vol. 6, no. 6, pp. 645–660, 1960.

[9] D. A. Kiriziy, N. A. Vorobey, and S. Y. Kots, "The relationship of nitrogen fixation and photosynthesis as the main components of the production process in alfalfa," *Russian Journal of Plant Physiology*, vol. 54, pp. 666–671, 2007.

[10] I. A. Kobbia, M. G. Battah, E. F. Shabana, and H. M. Eladel, "Chlorophyll *a* fluorescence and photosynthetic activity as tools for the evaluation of simazine toxicity to P*rotosiphon botryoides* and *Anabaena variabilis*," *Ecotoxicology and Environmental Safety*, vol. 49, no. 2, pp. 101–105, 2001.

[11] G. C. Papageorgiou, M. Tsimilli-Michael, and K. Stamatakis, "The fast and slow kinetics of chlorophyll a fluorescence induction in plants, algae and cyanobacteria: a viewpoint," *Photosynthesis Research*, vol. 94, no. 2-3, pp. 275–290, 2007.

[12] A. M. Grodzinskiy and D. M. Grodzinskiy, *Quick Reference Guide for Plant Physiology*, Naukova Dumka, 1964.

[13] R. W. F. Hardy, R. D. Holsten, E. K. Jackson, and R. C. Burns, "The acetylene-ethylene assay for N_2 fixation: laboratory and field evaluation," *Plant Physiology*, vol. 43, no. 8, pp. 1185–1207, 1968.

[14] A. R. Wellburn, "The spectral determination of chlorophylls *a* and *b*, as well as total carotenoids, using various solvents with spectrophotometers of different resolution," *Journal of Plant Physiology*, vol. 144, no. 3, pp. 307–313, 1994.

[15] J. Coombs, D. O. Hall, S. P. Long, and J. M. O. Scurlock, Eds., *Techniques in Bioproductivity and Photosynthesis*, Pergamon, 2nd edition, 1985.

[16] O. H. Lowry, N. Z. Rosenbrought, A. L. Farr, and R. Z. Randall, "Protein measurement with Folin fenol reagent," *The Journal of Biological Chemistry*, vol. 153, p. 265, 1951.

[17] G. Zubay, *Biochemistry*, The McGraw-Hill Companies, New York, NY, USA, 1998.

[18] L. I. Andreeva, L. A. Kozhemiakin, and A. A. Kishkun, "Modification of the method of determining lipid peroxidation in a test using thiobarbituric acid," *Laboratory Work*, pp. 41–43, 1988.

[19] R. L. Heath and L. Packer, "Photoperoxidation in isolated chloroplasts. I. Kinetics and stoichiometry of fatty acid peroxidation," *Archives of Biochemistry and Biophysics*, vol. 125, no. 1, pp. 189–198, 1968.

[20] O. A. Koksharova, M. Schubert, S. Shestakov, and R. Cerff, "Genetic and biochemical evidence for distinct key functions of two highly divergent GAPDH genes in catabolic and anabolic carbon flow of the cyanobacterium *Synechocystis* sp. PCC 6803," *Plant Molecular Biology*, vol. 36, no. 1, pp. 183–194, 1998.

[21] O. A. Koksharova, T. R. Kravzova, I. V. Lazebnaya et al., "Molecular identification and ultrastructural and phylogenetic studies of cyanobacteria from association with the white sea hydroid *Dynamena pumila* (L., 1758)," *BioMed Research International*, vol. 2013, Article ID 760681, 11 pages, 2013.

[22] U. Nübel, F. Garcia-Pichel, and G. Muyzer, "PCR primers to amplify 16S rRNA genes from cyanobacteria," *Applied and Environmental Microbiology*, vol. 63, no. 8, pp. 3327–3332, 1997.

[23] J. Sambrook, E. F. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 2nd edition, 1989.

[24] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[25] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[26] J. Elhai and C. P. Wolk, "Conjugal transfer of DNA to cyanobacteria," *Methods in Enzymology*, vol. 167, pp. 747–754, 1988.

[27] J. Elhai, A. Vepritskiy, A. M. Muro-Pastor, E. Flores, and C. P. Wolk, "Reduction of conjugal transfer efficiency by three restriction activities of *Anabaena* sp. strain PCC 7120," *Journal of Bacteriology*, vol. 179, no. 6, pp. 1998–2005, 1997.

[28] O. A. Koksharova and C. P. Wolk, "A novel gene that bears a DnaJ motif influences cyanobacterial cell division," *Journal of Bacteriology*, vol. 184, no. 19, pp. 5524–5528, 2002.

[29] N. A. Vorobey, O. V. Patsko, S. Y. Kots, and T. V. Parshikova, "Physiological traits of alfalfa plants development under the inoculation by mixed nitrogen fixing microorganisms cultures," *Physiology and Biochemistry of the Cultural Plants*, vol. 41, no. 4, pp. 344–352, 2009.

[30] E. Sergeeva, A. Liaimer, and B. Bergman, "Evidence for production of the phytohormone indole-3-acetic acid by cyanobacteria," *Planta*, vol. 215, no. 2, pp. 229–238, 2002.

[31] A. S. Voisin, C. Salon, C. Jeudy, and F. R. Warembourg, "Symbiotic N_2 fixation activity in relation to C economy of *Pisum sativum* L. as a function of plant phenology," *Journal of Experimental Botany*, vol. 54, no. 393, pp. 2733–2744, 2003.

[32] I. Berman-Frank, P. Lundgren, and P. Falkowski, "Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria," *Research in Microbiology*, vol. 154, no. 3, pp. 157–164, 2003.

[33] H. Küpper, N. Ferimazova, I. Šetlík, and I. Berman-Frank, "Traffic lights in *Trichodesmium*. Regulation of photosynthesis for nitrogen fixation studied by chlorophyll fluorescence kinetic microscopy," *Plant Physiology*, vol. 135, no. 4, pp. 2120–2133, 2004.

[34] M. V. Beligni and L. Lamattina, "Nitric oxide interferes with plant photo-oxidative stress by detoxifying reactive oxygen species," *Plant, Cell and Environment*, vol. 25, no. 6, pp. 737–748, 2002.

[35] D. Wendehenne, A. Pugin, D. F. Klessig, and J. Durner, "Nitric oxide: comparative synthesis and signaling in animal and plant cells," *Trends in Plant Science*, vol. 6, no. 4, pp. 177–183, 2001.

[36] E. Blumwald and E. Tel-Or, "Structural aspects of the adaptation of *Nostoc muscorum* to salt," *Archives of Microbiology*, vol. 132, no. 2, pp. 163–167, 1982.

[37] S. L. Rogers and R. G. Burns, "Changes in aggregate stability, nutrient status, indigenous microbial populations, and seedling emergence, following inoculation of soil with *Nostoc muscorum*," *Biology and Fertility of Soils*, vol. 18, no. 3, pp. 209–215, 1994.

[38] G. Z. de Caire, M. S. de Cano, M. C. Z. de Mulé, R. M. Palma, and K. Colombo, "Exopolysaccharide of *Nostoc muscorum* (Cyanobacteria) in the aggregation of soil particles," *Journal of Applied Phycology*, vol. 9, no. 3, pp. 249–253, 1997.

[39] W. K. Dodds, D. A. Gudder, and D. Mollenhauer, "The ecology of *Nostoc*," *Journal of Phycology*, vol. 31, no. 1, pp. 2–18, 1995.

[40] E. N. Sholkamy, H. El-Komy, A. A. Al-Arfaj, A. Abdel-Megeed, and A. A. Mostafa, "Potential role of *Nostoc muscoru*m and *Nostoc rivulare* as biofertilizers for the enhancement of maize growth under different doses of n-fertilizer," *African Journal of Microbiology Research*, vol. 6, no. 48, pp. 7435–7448, 2012.

[41] E. Dittmann, D. P. Fewer, and B. A. Neilan, "Cyanobacterial toxins: biosynthetic routes and evolutionary roots," *FEMS Microbiology Reviews*, vol. 37, no. 1, pp. 23–43, 2013.

[42] Y. Cai and C. P. Wolk, "*Anabaena* sp. strain PCC 7120 responds to nitrogen deprivation with a cascade-like sequence of transcriptional activations," *Journal of Bacteriology*, vol. 179, no. 1, pp. 267–271, 1997.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 680210, 9 pages
http://dx.doi.org/10.1155/2015/680210

Research Article

The Variability of the Order Burkholderiales Representatives in the Healthcare Units

Olga L. Voronina,[1] Marina S. Kunda,[1] Natalia N. Ryzhova,[1] Ekaterina I. Aksenova,[1] Andrey N. Semenov,[1] Anna V. Lasareva,[2] Elena L. Amelina,[3] Alexandr G. Chuchalin,[3] Vladimir G. Lunin,[1] and Alexandr L. Gintsburg[1]

[1]*N.F. Gamaleya Federal Research Center for Epidemiology and Microbiology, Ministry of Health of Russia, Gamaleya Street 18, 123098 Moscow, Russia*
[2]*Federal State Budgetary Institution "Scientific Centre of Children Health" RAMS, 119991 Moscow, Russia*
[3]*Research Institute of Pulmonology FMBA of Russia, 105077 Moscow, Russia*

Correspondence should be addressed to Olga L. Voronina; kirolg3@newmail.ru

Received 12 September 2014; Accepted 1 December 2014

Academic Editor: Vassily Lyubetsky

Background and Aim. The order Burkholderiales became more abundant in the healthcare units since the late 1970s; it is especially dangerous for intensive care unit patients and patients with chronic lung diseases. The goal of this investigation was to reveal the real variability of the order Burkholderiales representatives and to estimate their phylogenetic relationships. *Methods*. *16S rDNA* and genes of the *Burkholderia cenocepacia* complex (Bcc) Multi Locus Sequence Typing (MLST) scheme were used for the bacteria detection. *Results*. A huge diversity of genome size and organization was revealed in the order Burkholderiales that may prove the adaptability of this taxon's representatives. The following variability of the Burkholderiales in Russian healthcare units has been revealed: Burkholderiaceae (*Burkholderia*, *Pandoraea*, and *Lautropia*), Alcaligenaceae (*Achromobacter*), and Comamonadaceae (*Variovorax*). The *Burkholderia* genus was the most diverse and was represented by 5 species and 16 sequence types (ST). ST709 and 728 were transmissible and often encountered in cystic fibrosis patients and in hospitals. *A. xylosoxidans* was estimated by 15 genotypes. The strains of first and second ones were the most numerous. *Conclusions*. Phylogenetic position of the genus *Lautropia* with smaller genome is ambiguous. The Bcc MLST scheme is applicable for all Burkholderiales representatives for resolving the epidemiological problems.

1. Introduction

The antibiotic era resulted in the cardinal changes in the spectrum of the microorganisms, causing the healthcare-associated infections. Well-known bacterial pathogen *Staphylococcus aureus* was crowded by *Pseudomonas aeruginosa* [1]; then both were pressed by other Proteobacteria. The resistome of these bacteria has been enriched over the years of the nosocomial circulation, but most of them kept sensitive to at least one antibiotic.

The situation was complicated by the appearance of the order Burkholderiales bacteria after the late 1970s [2]. These bacteria are the common inhabitants of soil and water. They can be the plants' pathogens and have natural resistance to common antibiotics. They are especially dangerous for intensive care unit patients and patients with chronic lung diseases, particularly cystic fibrosis [3]. The taxonomy of this bacteria group has been developing since the 1980s and they were subdivided into different genera between 1981 and 2000 [4–8].

The infusion of nucleic acid sequencing technology in microbiology allowed Woese to start solving the bacterial phylogeny problem [9]. Proteobacteria, the most abundant and diverse bacterial phyla, were subdivided into classes on the base of the 16S rRNA gene sequences. First of all, 8–12 nucleotide signature sequences whose characteristic is unique to the species of Beta- and Gammaproteobacteria were identified (AAAAACCUUACC for Betaproteobacteria; AAACUCAAAUG for Gammaproteobacteria) [10, 11]. So the taxon *Pseudomonas*, according to Woese, was actually

a collection of at least five separate groups of bacteria [9]. It was subdivided into several genera, one of which was genus *Burkholderia* [5]. Later a new genus *Ralstonia* was separated from *Burkholderia* [7]. *Lautropia* [6]and *Pandoraea* [8] have appeared in the last few years. However the diversity, clinical and epidemiological significance of these taxa bacteria needs in detailed study. Continuing the investigation of *Bcc* role in nosocomial infections and using Multilocus Sequence Typing (MLST) as successful methodology in the epidemiology [12], we attempted to understand the variability of Burkholderiales in healthcare units.

2. Materials and Methods

2.1. Materials. Biological samples used for sequencing data are divided into two parts. The first part predominantly represented by nosocomial strains and strains from cystic fibrosis (CF) patients was described in [13] in detail. The second part contained some strains and mainly specimens of human sputum and aspirates from more than 300 CF patients.

2.2. DNA Isolation. DNA for PCR analysis was extracted from the bacterial cultures as described previously [13]. DNA from sputum and aspirate was isolated according to the protocol of the Maxwell 16 Tissue DNA Purification Kit for Maxwell MDX Instrument (Promega).

2.3. Species Identification. Identification of species was performed by amplification and sequencing of 16S ribosomal RNA gene (*16S rDNA*) fragment with primers [14, 15].

2.4. MLST. For Multilocus Sequence Typing, a modified scheme that allows differentiating 19 species of the *Burkholderia cepacia* complex (*Bcc*) was used [16]. The scheme includes the following targets for amplification: *atpD*, a β chain of ATP synthase; *gltB*, a large subunit of glutamate synthase; *gyrB*, a B subunit of DNA gyrase; *recA*, recombinase A; *lepA*, a GTP binding protein; *phaC*, acetyl CoA reductase; and *trpB*, a B subunit of tryptophane synthase. For DNA amplification, the following reagents were used: hot rescue DNA pol 5 units/μL, PCR buffer 10x (N.F. Gamaleya Institute for Epidemiology and Microbiology MoH), dNTP5 mM (Medigen), and primers (Evrogen). The modified amplification program was the same for all targets: 95°C—10 min (95°C—30 s, 63°C—40 s, 72°C—1 min) × 35, 72°C—5 min.

2.5. PCR Products Sequencing. PCR products were sequenced according to the protocol of BigDye Terminator 3.1 Cycle Sequencing kit for the Genetic Analyzer 3130 of Applied Biosystems/Hitachi. The electrophoretic DNA separation was performed in 50 cm capillaries with POP7 polymer.

2.6. Nucleotide Sequence Analysis. Analysis of sequences and alignment were made by the use of the program ClustalW2 [17]. Allele numbers for MLST genes were assigned with the help of the PubMLST website [18]. New alleles and STs were controlled and submitted by the curator of *Bcc* MLST database. Identification of *16S rDNA* sequences was carried out by BLAST search.

2.7. Nucleotide Sequence Polymorphism. The numbers of nucleotide/amino acid differences per site between concatenated sequences of 17 Bcc STs were obtained by pairwise distance calculation. Analyses were conducted in MEGA 4.0 [19].

Percent similarity and divergence coefficients of *gltB* gene nucleotide/amino acid sequences among analyzed representatives of the Burkholderiales were performed by the use of ClustalW2 [17], MEGA 6.0 [20], and MegAlign 5.05. For comparative sequence analysis and phylogenetic reconstruction 10 extra *gltB* gene sequences of the Burkholderiales order representatives (*Ralstonia solanacearum, Ralstonia pickettii, Acidovorax citrulli, Variovorax paradoxus, Bordetella bronchiseptica, Bordetella pertussis,* and *Lautropia mirabilis*) were retrieved from GenBank database (Table 3). The *gltB* sequences of *Pseudomonas aeruginosa* have been used as outgroup taxon (Table 3).

2.8. Phylogenetic Analysis. Phylogenetic analysis of *Bcc* was performed based on allelic profile data of *Bcc* STs and translated concatenated sequences of seven MLST loci. Phylogenetic tree of analyzed representatives of Burkholderiales order was constructed by the use of *gltB* sequences.

Analysis of profile data of *Bcc* STs was conducted using the software packages SplitsTree [21].

The phylogenetic tree of 17 *Bcc* STs based on translated concatenated sequences of seven MLST loci was obtained automatically by applying the neighbor-joining method [22]. The evolutionary distances between 17 *Bcc* STs were computed using the *p*-distance method [23] and were evaluated through the units of the amino acid differences' number per site. Evolutionary analyses were conducted in MEGA 6.0 [20]. Bootstrap analyses were performed with 500 replicates.

Phylogenetic tree of analyzed representatives of the Burkholderiales order was constructed by the use of neighborhood-joining, maximum likelihood, and maximum parsimony methods.

Genetic distances between microorganisms were evaluated by the use of Tamura 3-parameter model [24], which was chosen as an optimal evolution distance model derived from Modeltest based on the Akaike information criterion [25]. The evolutionary history was inferred by using the maximum likelihood method based on the general time reversible model GTR+G. Initial trees for the heuristic search were obtained automatically by applying neighbor-joining and BioNJ algorithms to a matrix of pairwise distances estimated by the use of the maximum composite likelihood approach and then selecting the topology with superior log likelihood value. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 0.8170). Maximum parsimony trees were constructed with an algorithm implemented in MEGA 6.0. Bootstrap analyses were performed with 1,000 replicates.

3. Results and Discussion

3.1. Common Characteristics of the Burkholderiales Genomes. The Burkholderiales is the dominating order among the

Table 1: The representatives of four families of the order Burkholderiales, which were detected in clinical specimens.

Class		Betaproteobacteria		
Order		Burkholderiales		
Family	Comamonadaceae	Alcaligenaceae	Burkholderiaceae	Ralstoniaceae
Genus	*Acidovorax*	*Bordetella*	*Burkholderia*	*Ralstonia*
Genus	*Variovorax*	*Achromobacter*	*Pandoraea*	
			Lautropia	

β-Proteobacteria available genomes, covering six families: Alcaligenaceae, Burkholderiaceae, Comamonadaceae, Oxalobacteraceae, Ralstoniaceae, and Sutterellaceae [26]. Four of them, demonstrated in Table 1, are more vital for the healthcare units. In the context of genome size, the order Burkholderiales is extraordinarily various (see S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/680210): from the smallest 0.070281 Mb of the Burkholderiales bacterium JGI 0001003-L21 (the rhizosphere and endosphere of *Arabidopsis thaliana*, INSDC AUNS00000000.1 [27]) to the biggest 11.2941 Mb of the *Burkholderia terrae* (the forest soil, INSDC AKAU00000000.1 [28]). The genomes of the Burkholderiales representatives are organized in different number of the chromosomes: 1, 2, or 3, without genome size correlation. So 7.35915 Mb genome of *Achromobacter xylosoxidans* has one chromosome, but 7.00881 Mb genome of *Burkholderia multivorans* has three chromosomes [27].

Analysis of the small genome group has demonstrated that all of them are the genomes of the host-restricted microbial symbionts: of plants, as abovementioned Burkholderiales bacterium JGI 0001003-L21 (INCDS AUNS00000000.1, genome size 0.07 Mb) [27], of sap-feeding insects, as *Candidatus Zinderia insecticola* (INCDS CP002161.1, genome size 0.208564 Mb) [29], or of human, as Burkholderiales bacterium1_1_47 (INCDS ADCQ00000000.1, genome size 2.61 Mb), isolated from feces in Human Microbiome Project [27].

But most of the bacteria of *Burkholderia cepacia* complex (*Bcc*) pathogenic for human keep big genome, providing for the genome plasticity and adaptability [30].

3.2. Bcc Diversity in the Healthcare Units of the Russian Federation. In our investigation of the microorganisms, causing the healthcare-associated infections, we drew attention to *Bcc* bacteria in departments both common and specialized for cystic fibrosis (CF) patients. Thirteen genotypes (sequence type, ST) were detected in the first phase of the analysis and nine of them (708, 709, 710, 711, 712, 714, 727, 728, and 729) were identified for the first time (Table 2). It was shown that strains causing nosocomial infections in most cases refer to genotypes 728 and 708. Genotype 709 detected in strains isolated from patients in seven federal regions of Russia should be recognized as epidemiologically significant for patients with cystic fibrosis [13].

The extension of the specimens' sampling in the second phase of the investigation demonstrated new *Bcc* genotypes in *B. cenocepacia* (ST862, 878) and *B. multivorans* species (ST835) and continued prevalence of the ST709. 79% of the CF patients, infected by *Bcc*, had ST709 strain. So, 16 STs were detected for *Bcc* isolated from patients in RF.

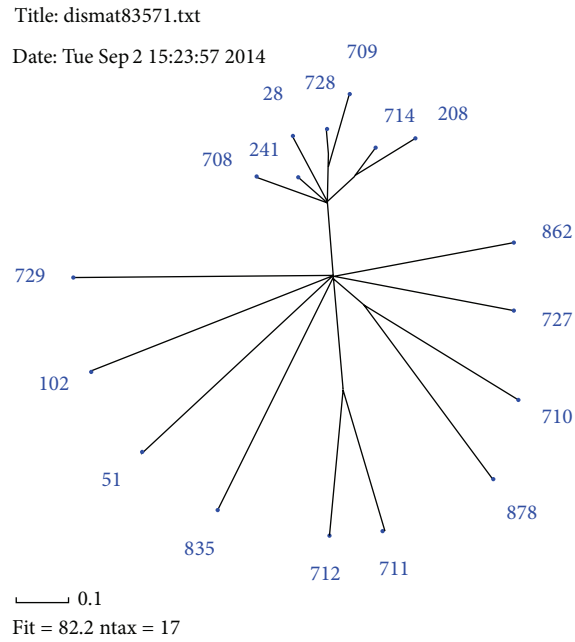


Figure 1: SplitsTree on the base of allelic profile data of the *Burkholderia cepacia* complex STs from RF.

To establish the relationship between different STs, we applied several methods of analysis. The first of them was SplitsTree analysis which was performed on the base of allelic profile data of *Bcc* STs (Figure 1). The most numerous group was formed by 6 STs of *B. cenocepacia* (708, 241, 728, 709, 714, and 208) closely related to the globally spread ST28. Next small groups were two STs containing first STs 710 and 878 that belonged to *B. cenocepacia* and second STs 711 and 712 related to *B. multivorans*. The other STs formed the separate branches.

To estimate the changes in the amino acid sequences, the concatenated sequences of MLST loci were translated. The bootstrap consensus tree using the neighbor-joining method was created (Figure 2). All groups of STs, represented different *Burkholderia* species, formed the separate branches with high bootstrap index (BI).

The most numerous group was formed by *B. cenocepacia*. It included 11 STs. Despite the fact that ST241, ST28, ST728, and ST709 had double locus variation (DLV) in the allelic profile, they formed the same branch with BI 63%. ST 708 which is also the DLV from ST241 was away from the group.

Table 2: The characteristics of the *Bcc* genotypes identified in RF.

Species	ST	ID PubMLST	Year	Source	Comments for ST distribution
B. cenocepacia	708	1149	2001	NON	Nosocomial strains in RF
B. cenocepacia	241	1258	2012	CF	Only CF strain in Far East of RF, but intercontinental spread strain in the world
B. cenocepacia	28	1268	1989	CF	Strains of multiple globally distinct locations, except for RF. Reference strain from Belgian collection
B. cenocepacia	728	1248	2004	NON	Nosocomial epidemic strains in RF, CF strain in all federal regions of RF
B. cenocepacia	709	1150	2008	CF	Epidemic strains for CF patients in all federal regions of RF, except for Far East
B. cenocepacia	714	1155	2003	NON	Strain from one hospital of the Southern Federal Region of RF
B. cenocepacia	208	1261	2012	CF	CF strains in Southern and Volga Federal Regions of RF and in USA
B. cenocepacia	862	1466	2014	CF	CF strain only in Far East of RF
B. cenocepacia	727	1246	2002	NON	Nosocomial strains in Northwestern Federal Region of RF
B. cenocepacia	710	1151	2012	CF	CF strains in RF
B. cenocepacia	878	1501	2014	CF	CF strain in RF
B. multivorans	711	1152	2012	CF	CF strains in RF
B. multivorans	712	1153	2011	CF	CF strains in RF and in Spain
B. multivorans	835	1443	2013	CF	CF strain in RF
B. stabilis	51	1267	1998	NON	Nosocomial strain in one hospital of RF, but intercontinental spread strain in the world
B. contaminans	102	1264	2000	CF	Nosocomial strain in one hospital and CF strain in Northwestern Federal Region of RF, but intercontinental spread strain in the world
B. vietnamiensis	729	1266	2012	CF	CF strain only in Far East of RF

CF: cystic fibrosis patient; NON: non-CF patient; RF: Russian Federation.

This can be explained by amino acid residues changes in the translated MLST sequences. In fact, between ST28, ST241, ST709, and ST728 there were no amino acid residues replacements in translated sequences, while ST708 had replacement V82A in large subunit of glutamate synthase. STs 714 and 208 with DLV in their allelic profiles formed a separate subgroup (BI = 100%). These STs differed from the other STs with the replacement E52D in ATP synthase beta chain; ST208 had additional replacement E7D in acetoacetyl-CoA reductase.

Another subgroup of *B. cenocepacia* with BI = 79% was represented by strains with STs 727, 862, 710, and 878.

Evolutionary divergence between sequences of 17 STs was received by pairwise distance calculation (Supplementary Material, S2 and S3). The less variability (0.002–0.005) was within the group including ST28, ST241, ST728, ST709, ST714, ST208, and ST708 related to *B. cenocepacia* (group 1). Group 2 was formed by *B. cenocepacia* strains too (ST727, ST862, ST710, and ST878). They had 0.019–0.023 base differences per site in comparison with group 1. So, in whole intraspecies *B. cenocepacia* STs variability was 0.002–0.023. The *B. multivorans* STs variability was almost the same—0.003–0.011. The most closely related species in the analyzed sample of *Bcc* were *B. cenocepacia* and *B. contaminans* (ST102) with variability from 0.037 to 0.040.

However, in most cases nucleotide rearrangement did not lead to changes in amino acid residues sequences and polymorphism within amino acid sequences was less than within nucleotides (Supplementary Material, S2 and S3). ST28, ST241, ST728, and ST709 had the same amino acid residues sequences. ST714, ST208, and ST708 differed from them with 0.001–0.003 amino acid residues per site. So, the number of amino acid residues' differences per site in group 1 was 0.000–0.003. More detectable changes were between STs group 1 and strains from group 2 (ST727, ST862, ST710, and ST878), 0.007–0.010 amino acid residues per site.

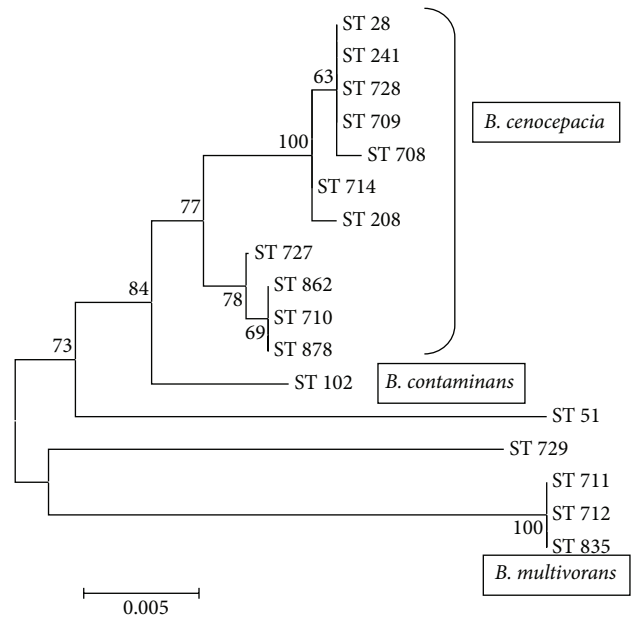


Figure 2: Neighbor-joining phylogenetic tree of seventeen *Burkholderia cepacia* complex STs based on translated concatenated sequences of seven MLST loci. ST51: *B. stabilis*; ST729: *B. vietnamiensis*.

Table 3: The sources of the *gltB* gene sequences for the phylogenetic analysis; *consecutive laboratory numeration of the registered alleles.

N	Species	Source or GenBank accession number	*gltB* allele or locus tag
1	*B. cenocepacia*	[18]	11
2	*B. cenocepacia*	[18]	16
3	*B. stabilis*	[18]	18
4	*B. multivorans*	[18]	50
5	*B. multivorans*	[18]	60
6	*B. contaminans*	[18]	80
7	*B. vietnamiensis*	[18]	103
8	*B. cenocepacia*	[18]	134
9	*B. cenocepacia*	[18]	136
10	*B. cenocepacia*	[18]	176
11	*B. cenocepacia*	[18]	352
12	*B. multivorans*	[18]	358
13	*A. xylosoxidans*	KC817498	1*
14	*A. xylosoxidans*	KC817500	2*
15	*A. xylosoxidans*	KF290958	3*
16	*A. xylosoxidans*	KF290959	4*
17	*A. xylosoxidans*	KJ941209	5*
18	*A. xylosoxidans*	KF297891	6*
19	*A. xylosoxidans*	KF963246	7*
20	*A. xylosoxidans*	KF963247	8*
21	*A. xylosoxidans*	KF963248	9*
22	*A. xylosoxidans*	KF963249	10*
23	*A. xylosoxidans*	KF963250	11*
24	*A. xylosoxidans*	KJ364657	12*
25	*A. xylosoxidans*	KJ439616	13*
26	*A. xylosoxidans*	KM262752	14*
27	*A. xylosoxidans*	KM262753	15*
28	*R. solanacearum*	AL646052.1	RSc2965
29	*R. pickettii*	CP006667.1	N234_19250
30	*A. citrulli*	CP000512.1	Aave_1008
31	*V. paradoxus*	CP001635.1	Vapar_1152
32	*B. bronchiseptica*	HE965806.1	BN112_3590
33	*B. pertussis*	BX640422.1	BP3753
34	*P. pnomenusa*	KM410934	KM410934
35	*P. pnomenusa*	CP007506.1	DA70_18115
36	*L. mirabilis*	KM410932	KM410932
37	*L. mirabilis*	KM410933	KM410933
38	*L. mirabilis*	AEQP01000020.1	EFV94423.1
39	*P. aeruginosa*	AE004091.2	PA5036
40	*P. aeruginosa*	FM209186.1	PLES_54261

So, *B. cenocepacia* ST241, ST728, ST709, ST714, ST208, and ST708 formed clonal complex, including ST28, which characterized the strains with global spread. Most of these STs were typical only for RF healthcare units (Table 2). Three STs (728, 708, and 709) were adaptive for epidemic spread.

3.3. The Potential of the Bcc MLST Scheme in the Burkholderiales Representatives Detection. During the second phase of the investigation we dealt not only with bacterial strains but also with a lot of samples of the sputum and aspirate. The *Bcc* MLST scheme adaptation to new conditions, amplification *Bcc* DNA in total DNA of the sample, suggested the apprehension of Spilker et al. [16] that degenerate primers, which allowed expansion of the modified *Bcc* MLST scheme, would not be specific only for *Burkholderia* species. First representative of Burkholderiales was *Achromobacter xylosoxidans*, in which *gltB* gene was amplified with the *Bcc* MLST scheme primers. After including this sequence in the analysis of the samples, we identified two different *gltB* alleles for this bacterium from the CF patients [13]. Then another thirteen *gltB* alleles were detected for *A. xylosoxidans*. The data analysis demonstrated the prevalence of the allele 1 and allele 2 among the CF patients from all federal regions of RF, except Far East, where only allele 3 was registered for *A. xylosoxidans*.

The increase of the number of *A. xylosoxidans* cases in the healthcare units, not only in CF patients, is according to the data of the new species registration. The data analysis of List of Prokaryotic Names with Standing in Nomenclature for the Burkholderiales order members demonstrated that over the last two years 18 new species of the *Burkholderia* genus have been registered [31], but only one was isolated from the human respiratory sample and others were environmental. On the other hand, species number of the *Achromobacter* genus increased two times during this period. All eight new species were clinical [32]. This data suggested *Achromobacter* significance as nosocomial bacterium.

Two targets of *Pandoraea pnomenusa* (*recA* and *gltB*) were amplified with the primers of *Bcc* MLST scheme too. But this dangerous and transmissible bacterium was isolated only from one CF patient. Some cases of *Lautropia mirabilis* were registered within one period of time. Only *gltB* gene was detectable by *Bcc* MLST primers. At last *Variovorax paradoxus*, detected in the group of the patients' samples, was amplified with *gyrB* primers. Detection of the seldom trace amount of *Ralstonia* spp. was possible with the *gltB* primers too.

So, we may conclude that the *Bcc* MLST scheme *gltB* primers are universal for most of the clinically significant Burkholderiales. The *gltB* gene sequences from our investigation and sequences avoided from GenBank were analyzed in the next step. *Bordetella* genus sequence included was explained by the importance of this genus as causative agent of human diseases.

3.4. The gltB Gene Sequence Polymorphism. 38 representatives of the order Burkholderiales and two of *Pseudomonas aeruginosa* (as an outgroup taxon) were used in analysis. A 414-base-pair alignment for the *gltB* gene region was obtained.

Totally 295 variable nucleotide sites have been detected; 240 of them characterized diversity of the orderBurkholderiales analyzed representatives. Differences between representatives of the Burkholderiales and the outgroup taxa, *Pseudomonas aeruginosa,* reached 48.8% (Pse-Ral,

Pse-Var-Aci); see Table 4. The differences among the investigated Burkholderiales bacteria varied from 0.2% to 32.5%. The *gltB* allele diversity in *Bcc* represented in analysis by five species was comparable to the diversity among *A. xylosoxidans* alleles and reached 6.8% and 5.1%, respectively (Supplementary Material, S4). These data suggested the close relatedness of the *Bcc* species. Percent of the differences in *gltB* gene sequence between representatives of the Alcaligenaceae family, *A. xylosoxidans* and *Bordetella bronchiseptica/Bordetella pertussis,* was 6.8–9.2%, indicating close relatedness of these taxa too. The *gltB* allele differences between other representatives of Burkholderiales fell into the range of 20–32.5%.

Surprisingly, the level of *gltB* gene sequence differences between the members of one family Burkholderiaceae (Bcc and *Lautropia mirabilis*) reached 27.9–31.6%, and that was more than differences between *Bcc* and the member of the other families: Ralstoniaceae (*Ralstonia*, 20.6–26.9%) and Comamonadaceae (*Acidovorax, Variovorax*, 25.2–28.6%) (Supplementary Material, S4). However, according to the data of Gerner-Smidt, based on variability of *16S rDNA* the differences between *Lautropia mirabilis* and *Burkholderia cepacia* were 7.7% [6], which can characterize the higher resolution features of *gltB* gene sequences.

Amino acid residues variability of the translated *gltB* gene fragments was also evaluated (Supplementary Material, S5). In the sequence, consisting of 136 amino acid residues, 111 residues were variable. So, out of 240 SNPs, characterizing the diversity of Burkholderiales, 81 SNPs resulted in amino acid residues substitutions. The interspecies diversity of *A. xylosoxidans* was characterized by six amino acid residues substitutions; the diversity of *Bcc*, by nine substitutions.

3.5. Phylogeny of the Analyzed Burkholderiales Representatives Based on gltB Sequences. ML phylogenetic tree based on *gltB* sequence is presented in Figure 3. *P. aeruginosa* well known as nosocomial bacterial pathogen, taken in this analysis as an outgroup taxon (Gammaproteobacteria, Pseudomonadales, and Pseudomonadaceae), formed the most divergent basal branch on the tree as was expected. The phylogenetic tree revealed two main groups of the Burkholderiales order representatives in this analysis. The first group (BI 78%) included only the members of the Alcaligenaceae family: fifteen alleles of *A. xylosoxidans* and reference *gltB* alleles from *Bordetella* genomes. It should be noted that *gltB* sequences allowed separating these taxa into distinct subclades.

The second group (BI 78%) was formed by the representatives of three families: Burkholderiaceae, Ralstoniaceae, and Comamonadaceae. Inside the second group twelve alleles of *Bcc* formed a large subgroup; two alleles of *Pandoraea pnomenusa* were closely related to this subgroup. However, three alleles of *Lautropia mirabilis* were more divergent from *gltB* alleles of Bcc and *Pandoraea pnomenusa* than representatives of two different families *Ralstoniacea* (*Ralstonia solanacearum, Ralstonia pickettii*) and *Alcaligenacea* (*Acidovorax citrulli, Variovorax paradoxus*).

A similar situation was described by phylogenetic cladogram, constructed for the order Burkholderiales representatives [33] by an automated pipeline of PATRIC genome database [34]. The construction of the phylogenetic tree on this server begins with amino acid sequence files for each genome. On this tree *Lautropia mirabilis* fell in one group (BI 79%) with two genomes of the host-restricted microbial symbionts: *Candidatus Zinderia insecticola* (INCDS CP002161.1) [29] and *Burkholderiales bacterium* 1_1_47 (INCDS ADCQ00000000.1) [27], and with the representatives of the genera *Parasutterella* and *Sutterella*, the member of the family Sutterellaceae. *Parasutterella* was isolated from human faeces [35], and *Sutterella* strains were

Table 4: Percent of *gltB* sequences variability among analyzed representatives of Burkholderiales.

Group of genotypes	Variability, %	
	DNA sequence	Amino acid residues sequence
Lau	4.1–4.4	0.0–0.7
Ach	0.2–5.1	0.0–3.7
Bcc	0.2–6.8	0.0–5.1
Ach-Bor	6.8–9.2	5.9–8.1
Bcc-Pan	21.1–23.1	24.3–25.7
Bcc-Bor	25.5–26	33.1–36
Bcc-Ach	23.8–26.7	30.9–35.3
Bcc-Ral	20.6–26.9	25–30.9
Ach-Pan	25.5–27.2	25–27.9
Bor-Pan	26.9–27.4	27.9–29.4
Ral-Pan	20.9–27.7	24.3–27.2
Bcc-Var-Aci	25.2–28.6	29.4–38.2
Var-Pan	28.2–29.1	33.1–34.6
Bor-Lau	27.7–29.4	32.4–33.1
Var-Bor	27.9–29.4	35.3–39.7
Ral-Var	25.2–29.6	36–38.2
Ach-Var	27.4–29.6	35.3–41.9
Ach-Lau	28.2–30.3	32.4–35.3
Pan-Lau	30.1–30.6	30.9–31.6
Ral-Bor	24.5–31.3	30.9–36.8
Var-Lau	29.4–31.3	33.1–36
Ach-Ral	24–31.6	28.7–38.2
Bcc-Lau	27.9–31.6	30.1–33.1
Ral-Lau	26–32.5	30.9–34.6
Pse-Bor	42.7–43	66.2–66.9
Ach-Pse	43.7–45.9	66.2–66.9
Pse-Lau	46.1–46.8	69.1–69.9
Bcc-Pse	44.7–47.1	65.4–66.2
Pse-Pan	47.6–47.8	66.2
Pse-Ral	46.1–48.8	66.9–69.9
Pse-Var	47.3–48.8	66.2–67.6

Bcc: *Burkholderia cepacia* complex; Ach: *Achromobacter xylosoxidans*; Lau: *Lautropia mirabilis*; Ral: *Ralstonia solanacearum/Ralstonia pickettii*; Aci: *Acidovorax citrulli*; Var: *Variovorax paradoxus*, Bor: *Bordetella bronchiseptica/Bordetella pertussis*; Pan: *Pandoraea pnomenusa*; Pse: *Pseudomonas aeruginosa*.

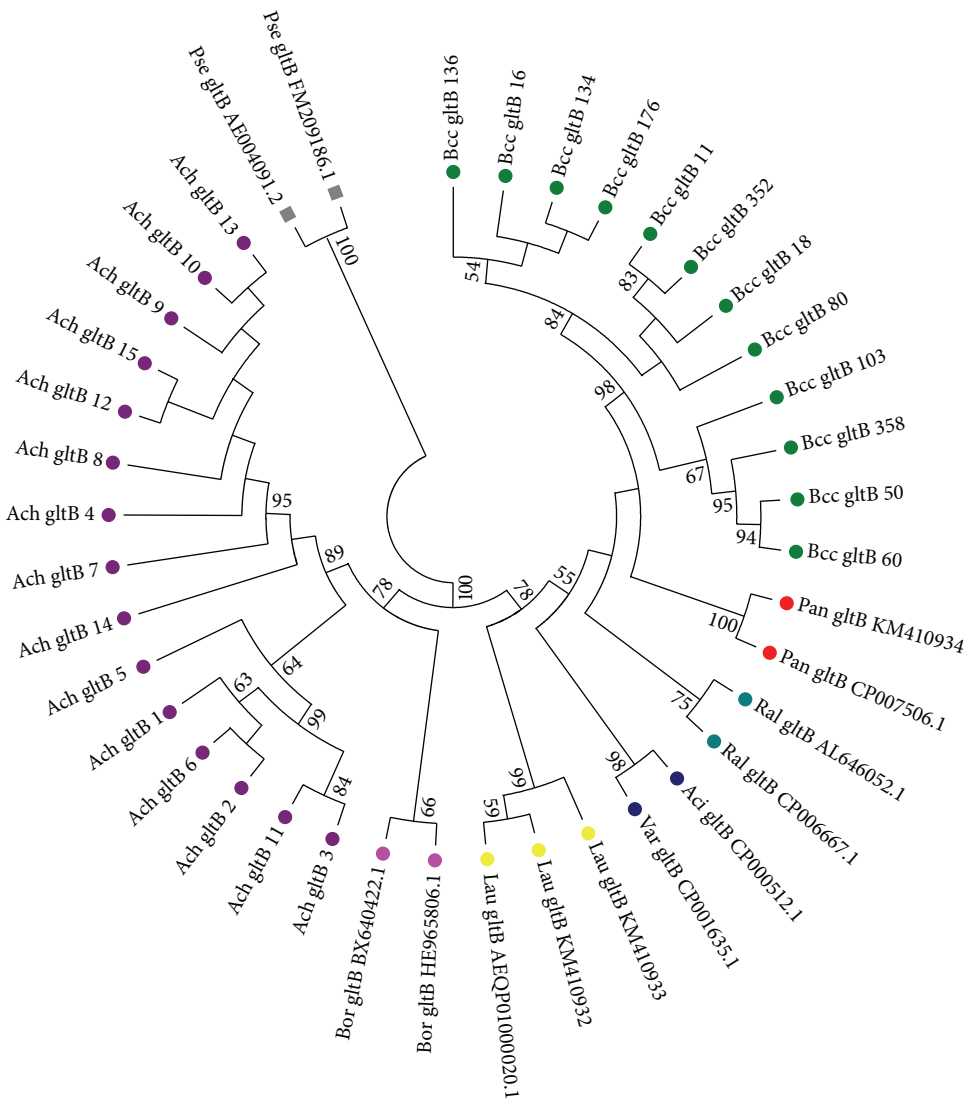


Figure 3: ML phylogenetic tree of analyzed representatives of Burkholderiales order based on *gltB* sequences.

isolated from infections that occurred below the diaphragm [36]. Both have small genome 2.3769–2.98833 Mb.

Lautropia mirabilis is the poorly investigated species of a Gram-negative motile coccus with the unusual morphology fairly recently isolated from the human mouth by Gerner-Smidt et al. [6]. The genome of this bacterium is surprisingly small (3.15192 Mb) as compared with *Bcc* and *Pandoraea*. The loss of some genes in the evolution of this bacterium may be suggested.

On the other hand, the *16S rDNA* sequencing data showed that *Lautropia mirabilis* belonged to a separate branch of the Betaproteobacteria and was most closely related to the genus *Burkholderia.* The isolated position together with the unique combination of chemotaxonomic and phenotypic properties allowed attributing of *Lautropia mirabilis* (strain AB2188) to the separate genus [6].

According to the last All-Species Living Tree (Release LTPs115, March 2014) [37] and also based on *16S rDNA* sequences, single *Lautropia mirabilis* (AEQP01000026) formed a separate basal branch more related to *Burkholderia* and *Pandoraea* in Burkholderiaceae clade, which is joined to Comamonadaceae clade.

Similar disagreements between two phylogenetic trees were revealed for *Ralstonia* genus too. *Ralstonia* is usually attributed to Ralstoniaceae on the base of the *gltB* sequences, but according to All-Species Living Tree [37] *Ralstonia* is joined to *Cupriavidus* and fell into Oxalobacteraceae clade.

Consequently, according to our results, the polymorphism of *gltB* gene sequences was high and allowed describing substantial diversity of the Burkholderiales order members, defined the main taxonomical groups represented by Burkholderiaceae (*Burkholderia, Pandoraea,* and *Lautropia*), Alcaligenaceae (*Achromobacter*), and Comamonadaceae (*Variovorax*), and revealed significant differences between *Lautropia* and the other Burkholderiaceae taxa.

In conclusion, we identified and characterized quite a wide range of the Burkholderiales order bacteria which are vital for the healthcare units at present in Russia. They

have been represented by five genera: *Burkholderia, Pandoraea, Lautropia* (Burkholderiaceae), *Achromobacter* (Alcaligenaceae), and *Variovorax* (Comamonadaceae). The most abundant were *Bcc* and *A. xylosoxidans* with prevalence of transmissible ST709 and ST728 of *Burkholderia cenocepacia* and the first and second genotypes of *A. xylosoxidans*. Also not common and unusual bacteria like *Pandoraea pnomenusa, Variovorax paradoxus, Lautropia mirabilis,* and *Ralstonia* spp. began to appear in the hospitals and were registered in the group of the patients' samples. These observations confirm profound changes in the spectrum of the microorganisms, causing the healthcare-associated infections over the past few years that can be associated with emergence and dissemination of novel antibiotic resistance from the natural reservoir to the clinical setting. So we may conclude that pathogenic potential of the Burkholderiales is on the increase. Clarification of some questions on bacteria phylogeny and future genomic analysis of Burkholderiales species will provide deeper large-scale insights into the evolution of virulence mechanisms. The timely identification of the Burkholderiales order representatives by genotyping is important to limit bacterial spread and so to resolve some epidemiological problems.

Abbreviations

MLST: Multilocus Sequence Typing
ST: Sequence type
DLV: Double locus variant
ML: Maximum likelihood
NJ: Neighbor-joining
CF: Cystic fibrosis and cystic fibrosis patient
NON: Non-CF patient
RF: Russian Federation
Bcc: *Burkholderia cepacia* complex
Ach: *Achromobacter xylosoxidans*
Lau: *Lautropia mirabilis*
Ral: *Ralstonia solanacearum/Ralstonia pickettii*
Aci: *Acidovorax citrulli*
Var: *Variovorax paradoxus*
Bor: *Bordetella bronchiseptica/Bordetella pertussis*
Pan: *Pandoraea pnomenusa*
Pse: *Pseudomonas aeruginosa.*

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] M. B. Mearns, G. H. Hunt, and R. Rushworth, "Bacterial flora of respiratory tract in patients with cystic fibrosis, 1950–71," *Archives of Disease in Childhood*, vol. 47, no. 256, pp. 902–907, 1972.

[2] R. Hobson, I. Gould, and J. Govan, "*Burkholderia (Pseudomonas) cepacia* as a cause of brain abscesses secondary to chronic suppurative otitis media," *European Journal of Clinical Microbiology and Infectious Diseases*, vol. 14, no. 10, pp. 908–911, 1995.

[3] A. R. Hauser, M. Jain, M. Bar-Meir, and S. A. McColley, "Clinical significance of microbial infection and adaptation in cystic fibrosis," *Clinical Microbiology Reviews*, vol. 24, no. 1, pp. 29–70, 2011.

[4] E. Yabuuchi and I. Yano, "*Achromobacter* gen. nov. and *Achromobacter xylosoxidans* (ex Yabuuchi and Ohyama 1971) nom. rev," *International Journal of Systematic and Evolutionary Microbiology*, vol. 31, no. 4, pp. 477–478, 1981.

[5] "Validation list No. 45," *International Journal of Systematic Bacteriology*, vol. 43, pp. 398–399, 1993.

[6] P. Gerner-Smidt, H. Keiser-Nielsen, M. Dorsch et al., "*Lautropia mirabilis* gen. nov., sp. nov., a Gram-negative motile coccus with unusual morphology isolated from the human mouth," *Microbiology*, vol. 140, no. 7, pp. 1787–1797, 1994.

[7] "Validation of the publication of new names and new combinations previously effectively published outside the IJSB, List No. 57," *International Journal of Systematic Bacteriology*, vol. 46, no. 2, pp. 625–626, 1996.

[8] T. Coenye, E. Falsen, B. Hoste et al., "Description of Pandoraea gen. nov. with Pandoraea apista sp. nov., Pandoraea pulmonicola sp. nov., Pandoraea pnomenusa sp. nov., Pandoraea sputorum sp. nov. and Pandoraea norimbergensis comb. nov," *International Journal of Systematic and Evolutionary Microbiology*, vol. 50, no. 2, pp. 887–899, 2000.

[9] C. R. Woese, "Bacterial evolution," *Microbiological Reviews*, vol. 51, no. 2, pp. 221–271, 1987.

[10] C. R. Woese, W. G. Weisburg, B. J. Paster et al., "The phylogeny of purple bacteria: the beta subdivision," *Systematic and Applied Microbiology*, vol. 5, no. 3, pp. 327–336, 1984.

[11] C. R. Woese, W. G. Weisburg, and C. M. Hahn, "The phylogeny of purple bacteria: the gamma subdivision," *Systematic and Applied Microbiology*, vol. 6, no. 1, pp. 25–33, 1985.

[12] A. B. Ibarz Pavón and M. C. J. Maiden, "Multilocus sequence typing," *Methods in Molecular Biology*, vol. 551, pp. 129–140, 2009.

[13] O. L. Voronina, M. Y. Chernukha, I. A. Shaginyan et al., "Characterization of genotypes for *Burkholderia cepacia* complex strains isolated from patients in hospitals of the Russian federation," *Molecular Genetics, Microbiology and Virology*, vol. 28, no. 2, pp. 64–73, 2013.

[14] S. M. Naser, K. E. Hagen, M. Vancanneyt, I. Cleenwerch, J. Swings, and T. A. Tompkins, "*Lactobacillus suntoryeus* Cachat and Priest 2005 is a later synonym of *Lactobacillus helveticus* (Orla-Jensen 1919) Bergey et al. 1925 (Approved Lists 1980)," *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, no. 2, pp. 355–360, 2006.

[15] L. L. Guan, K. E. Hagen, G. W. Tannock, D. R. Korver, G. M. Fasenko, and G. E. Allison, "Detection and identification of *Lactobacillus* species in crops of broilers of different ages by using PCR-denaturing gradient gel electrophoresis and amplified ribosomal DNA restriction analysis," *Applied and Environmental Microbiology*, vol. 69, no. 11, pp. 6750–6757, 2003.

[16] T. Spilker, A. Baldwin, A. Bumford, C. G. Dowson, E. Mahenthiralingam, and J. J. LiPuma, "Expanded multilocus sequence typing for *Burkholderia* species," *Journal of Clinical Microbiology*, vol. 47, no. 8, pp. 2607–2610, 2009.

[17] "The Main Site EMBL-EBI," European Bioinformatics Institute, 2014, http://www.ebi.ac.uk/Tools/msa/clustalw2.

[18] "The Main Site PubMLSR of the Faculty of Zoology, Oxford University, Great Britain," 2014, http://pubmlst.org.

[19] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[20] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[21] D. H. Huson, "SplitsTree: analyzing and visualizing evolutionary data," *Bioinformatics*, vol. 14, no. 1, pp. 68–73, 1998.

[22] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[23] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.

[24] K. Tamura, "Estimation of the number of nucleotide substitutions when there are strong transition- transversion and G + C-content biases," *Molecular Biology and Evolution*, vol. 9, no. 4, pp. 678–687, 1992.

[25] D. Posada and K. A. Crandall, "MODELTEST: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[26] K. A. Kormas, "Interpreting diversity of Proteobacteria based on 16S rRNA gene copy number," in *Proteobacteria: Phylogeny, Metabolic Diversity and Ecological Effects*, M. L. Sezenna, Ed., pp. 73–89, Nova Publishers, Hauppauge, NY, USA, 2011.

[27] NCBI, Genome List, http://www.ncbi.nlm.nih.gov/genome/browse/.

[28] H.-C. Yang, W.-T. Im, K. K. Kim, D.-S. An, and S.-T. Lee, "*Burkholderia terrae* sp. nov., isolated from a forest soil," *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, part 2, pp. 453–457, 2006.

[29] J. P. McCutcheon and N. A. Moran, "Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution," *Genome Biology and Evolution*, vol. 2, no. 1, pp. 708–718, 2010.

[30] P. Drevinek and E. Mahenthiralingam, "*Burkholderia cenocepacia* in cystic fibrosis: epidemiology and molecular mechanisms of virulence," *Clinical Microbiology and Infection*, vol. 16, no. 7, pp. 821–830, 2010.

[31] J. P. Euzéby, "The List of Prokaryotic names with Standing in Nomenclature," http://www.bacterio.net.

[32] P. Vandamme, E. R. B. Moore, M. Cnockaert et al., "Classification of *Achromobacter genogroups* 2, 5, 7 and 14 as *Achromobacter insuavis* sp. nov., *Achromobacter aegrifaciens* sp. nov., *Achromobacter anxifer* sp. nov. and *Achromobacter dolens* sp. nov., respectively," *Systematic and Applied Microbiology*, vol. 36, no. 7, pp. 474–482, 2013.

[33] The phylogentic tree for the order Burkholderiales, http://patricbrc.org/portal/portal/patric/Phylogeny?cType=taxon&cId=47671.

[34] A. R. Wattam, D. Abraham, O. Dalay et al., "PATRIC, the bacterial bioinformatics database and analysis resource," *Nucleic Acids Research*, vol. 42, no. 1, pp. D581–D591, 2014.

[35] F. Nagai, M. Morotomi, H. Sakon, and R. Tanaka, "*Parasutterella excrementihominis* gen. nov., sp. nov., a member of the family *Alcaligenaceae* isolated from human faeces," *International Journal of Systematic and Evolutionary Microbiology*, vol. 59, no. 7, pp. 1793–1797, 2009.

[36] H. M. Wexler, D. Reeves, P. H. Summanen et al., "*Sutterella wadsworthensis* gen. nov., sp. nov., bile-resistant microaerophilic *Campylobacter gracilis*-like clinical isolates," *International Journal of Systematic Bacteriology*, vol. 46, no. 1, pp. 252–258, 1996.

[37] "The All-Species Living Tree (Release LTPs115)," March 2014, http://www.arb-silva.de/fileadmin/silva_databases/living_tree/LTP_release_115/LTPs115_SSU_tree.pdf.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 387913, 8 pages
http://dx.doi.org/10.1155/2015/387913

Research Article

Signs of Selection in Synonymous Sites of the Mitochondrial Cytochrome b Gene of Baikal Oilfish (Comephoridae) by mRNA Secondary Structure Alterations

Veronika I. Teterina, Anatoliy M. Mamontov, Lyubov V. Sukhanova, and Sergei V. Kirilchik

Limnological Institute, Siberian Branch, Russian Academy of Sciences, P.O. Box 278, Irkutsk 664033, Russia

Correspondence should be addressed to Veronika I. Teterina; veronika@lin.irk.ru

Received 10 July 2014; Revised 10 November 2014; Accepted 10 November 2014

Academic Editor: Peter F. Stadler

Studies over the past decade have shown a significant role of synonymous mutations in posttranscriptional regulation of gene expression, which is particularly associated with messenger RNA (mRNA) secondary structure alterations. Most studies focused on prokaryote genomes and the nuclear genomes of eukaryotes while little is known about the regulation of mitochondrial DNA (mtDNA) gene expression. This paper reveals signs of selection in synonymous sites of the mitochondrial cytochrome b gene (Cytb) of Baikal oilfish or golomyankas (Comephoridae) directed towards altering the secondary structure of the mRNA and probably altering the character of mtDNA gene expression. Our findings are based on comparisons of intraspecific genetic variation patterns of small golomyanka (*Comephorus dybowski*) and two genetic groups of big golomyanka (*Comephorus dybowskii*). Two approaches were used: (i) analysis of the distribution of synonymous mutations between weak-AT (W) and strong-GC (S) nucleotides within species and groups in accordance with mutation directions from central to peripheral haplotypes and (ii) approaches based on the predicted mRNA secondary structure.

1. Introduction

Recent studies have shown a significant role of synonymous sites in the regulation of gene expression and fitness [1–5]. Synonymous mutations influence not only the rate of translation but also posttranslational modification of proteins [6]. One form of such regulation involves alterations of the secondary structure of messenger RNA (mRNA). The impact of synonymous codon sites on mRNA stability, folding, and, consequently, gene expression is greater than the impact of nonsynonymous codon sites [4, 7] and can therefore be exposed to various forms of natural selection more than previously believed. To date, most studies of the effects of synonymous mutations on gene expression have focused on prokaryote genomes and the nuclear genome of eukaryotes, while virtually little is known at the mitochondrial DNA (mtDNA) level. Meanwhile, the identification of signs of selection in synonymous sites of mitochondrial genes in natural objects can be an area of focus for understanding these processes. Baikal oilfish can be considered such an object.

Big golomyanka (BG) or big Baikal oilfish (*Comephorus baicalensis*) and small golomyanka (SG) or little Baikal oilfish (*Comephorus dybowskii*) are the only representatives of the endemic genus *Comephorus* of the Comephoridae family of Lake Baikal cottoid fish [8]. The habitats of these species largely overlap; representatives of BG and SG are distributed throughout Lake Baikal with the exception of the coastal zone and are found at all depths of the lake. Baikal oilfish are the most abundant fish in Lake Baikal. According to Starikov [9], the numbers of SG and BG were approximately 22.2–41.2 and 7.1–10.8 billion specimens, respectively, for the period of 1969 to 1972. These species are very adapted to the pelagic lifestyle of an open lake. This is reflected in the fact that golomyankas are extremely sensitive to changes in environmental conditions. The optimum water temperature for these species is ~3.6°C. During experiments when the temperature rose above 8.0–8.5°C, BG instantly fell into

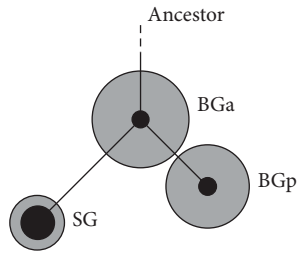


Figure 1: Schematic representation of the network of Baikal oilfish obtained by Teterina et al. [15]. Dark circle diameters are proportional to the frequencies of the main haplotypes of Cytb within the groups and grey circle diameters are proportional to the respective values of nucleotide diversity (π).

a stupor and began to die [10]. Both species are viviparous. BG breeds throughout the year, with two peaks of spawning in August, September, and December [11–13].

As previously demonstrated, there are two genetic groups of BG that differ by two nucleotide substitutions in the mitochondrial cytochrome b gene (Cytb) [15]. One group is represented by the line of ancestral haplotypes (BGa) while another is paraphyletically related to BGa (BGp); BGp and SG are derived from the central haplotype of BGa (Figure 1). Nuclear DNA analysis and morphological data revealed no differences between BGa and BGp. It was suggested that these groups correspond to the two peaks of BG spawning. It was also assumed that the differences in mitochondrial DNA are the result of differences in female breeding timing, and nuclear genome panmixia is the result of independent mating of males with females from both genetic groups.

This paper showed signs of selection in synonymous sites of Cytb of Baikal golomyankas directed towards altering the secondary structure of the mRNA and probably altering the character of mtDNA gene expression. We examined the patterns of Cytb molecular variation with the methods of prediction of the secondary structure of nucleic acids and used the data on the fishes' life history.

2. Materials and Methods

2.1. Sampling and DNA Extraction. Adult specimens were sampled using a beam trawl between 2000 and 2010 in three Lake Baikal basins: southern, central, and northern. Larvae were collected with a large DJOM plankton closing net (2 m diameter, 1 vv mesh) in July, September, and February 2005 to 2011. To determine which of the two spawning groups the BG larvae belonged to, the body lengths of the larvae were measured, which allowed approximate determination of their time of birth. Only individuals with a body length <20 mm were used. Larvae age was approximately assessed as follows: individuals with a body length up to 10 mm were graded as being 1 to 1.5 months old and specimens with a length of 10 to 20 mm were considered to be 1.5 to 3 months old (Elena V. Dzyuba, LIN SB RAS, personal communication). The larvae collected in July and September were attributed to the "summer" spawning peak whereas the larvae collected in February were attributed to the "winter" maximum.

Total DNA was isolated from fish muscle or from whole larvae using phenol-chloroform extraction [18]. An allele-specific real-time polymerase chain reaction (PCR) protocol was developed to identify individuals belonging to a particular genetic group of BG. A similar system was developed to identify BG and SG individuals since species affiliation of larvae and juveniles cannot always be accurately determined based solely on external morphology.

2.2. Single Nucleotide Polymorphism (SNP) Analysis. For the rapid and reliable separation of the oilfish individuals by species and groups, SNP markers were designed based on the same DNA sequences as in [15]. The substitutions by which the species and the genetic groups were differentiated are shown in Table 1. To identify the representatives of BG and SG, we used the BG-specific forward primer FBG (5′-ACTACGGATGACTTATCCGTAACC-3′) and the SG-specific forward primer FSG (5′-ACTACGGATGACTTATCCGTAACAC-3′). The reverse primer was common in both cases: RBGSG (5′-TACCCTACGAAAGCGGTTATTATTACAA-3′). To identify the representatives of BGp and BGa, we employed the BGp-specific forward primer FBGp (5′-GCCTGAGTGGTACTTCCTGTTC-3′) and the (BGa + SG) specific forward primer FBGa (5′-GCCTGAGTGGTACTTCTTGTTT-3′). Again, the reverse primer was the same for both: RBGaBGp (5′-CTCCAAGTTTGTTGGGGAT-3′).

Real-time PCR was performed on a Rotor-Gene Q (Qiagen) in a 15 μL reaction mixture containing 1.5 U Encyclo polymerase (Evrogen, Russia), 1.5 μL Encyclo buffer, 1x SYBR Green I dye (BioDye, Russia), 200 μM of each dNTP, 5 pmol of each primer, and 5–50 ng template DNA. The PCR temperature profile involved a 3 min denaturation at 95°C followed by 35 cycles of 10 s at 95°C and 10 s at 60°C. Each reaction was conducted at least in triplicate, and all runs were completed with a melt curve analysis.

2.3. Distribution of Mutations. Cytb analyses were performed using largely the same DNA sequences as in [15] supplemented by nine haplotypes of BGp (GenBank Accession Numbers KC571825–KC571833). Mutation calculations were performed according to direction in haplotype genealogy, as evaluated by an unrooted network and a statistical parsimony criterion with NETWORK 4.6 [19] and using the option "statistics." The distribution of nucleotide diversity (π) along Cytb was explored using the Sliding Window Option (SWA) of the program VariScan version 1 [20]. The width of the window was 100 bp, and the window slide was 10 bp.

2.4. Prediction of mRNA Secondary Structure. Prediction of mRNA secondary structure (RSSP) was performed using programs UNAFold version 3.8 [16] and RNAfold of the Vienna RNA Package version 1.8.5 [17]. Both synonymous and nonsynonymous mutations were used. A temperature option equal to 3.5°C (which is close to the natural environmental temperature of golomyankas) was chosen to calculate the minimum free energy. To calculate distances between RNA secondary structures, 31 runs of RNAfold and RNAdistance programs [17] were performed with temperatures ranging

Table 1: Substitutions, segregating SG, BGa, and BGp, and their location in the Cytb.

Sympatric groups	Substitution positions*					
	15627	15966	16004	16206	16211	16502
BGp	**C**	A	C	**C**	**C**	A
BGa	**C**	A	C	**T**	**T**	A
SG	**A**	C	T	**T**	**T**	G

*The substitution positions are given relative to the mitochondrial sequence of *Salmo salar* [14] (GenBank Accession Number U12143).
Substitutions used for SNP analysis in bold.

Table 2: Number of BGa and BGp adults collected from the south, middle, and north basins of the lake.

Sympatric groups	Basins of the lake			Total number
	South	Middle	North	
BGa	114	19	32	165
BGp	108	18	32	158

from 1 to 4°C and steps equal to 0.1°C. The chosen temperature interval roughly covers the range of golomyanka temperature environments. The Fitch-Margoliash method implemented in the Fitch program (PHYLIP package, version 3.6) was used to generate the series of trees based on the distance matrices, and the Consense program (PHYLIP package) was then used to generate a majority rule consensus tree [21]. The values obtained in the internal edges were interpreted as branch support measurements. Sequence analysis, tree drawings, and tree manipulations were performed using MEGA version 6 [22].

3. Results

3.1. SNP Analysis. The allele-specific primers were tested using DNA samples with known Cytb sequences [15]. In all cases, there was strict conformity between the test results and the nucleotide sequence. We used this approach to determine the ratio of individuals belonging to different genetic groups of BG collected from different regions of Lake Baikal (Table 2). We also analyzed the larvae born in different seasons (Table 3). It was evident that the ratio of individuals belonging to different genetic groups in different basins and in the lake as a whole was close to 1 : 1. Larvae analysis did not reveal any clear patterns between the time of birth and affiliation with a certain genetic group. Representatives of BGa and BGp of the same larval size range were present in each sample.

3.2. Mutations Bias. As can be seen in Table 1, the BGa and BGp haplotypes groups differed in two substitutions located at four nucleotides from one another. These were the key substitutions that formed the main haplotypes of the groups. Considering the direction from BGa to BGp (Figure 1), both substitutions were from T to C (T→C). The T→C type polymorphism was also found at position 15500 (data not shown), where the C nucleotide was present at a very high frequency in BGp and absent in BGa. All substitutions were synonymous. Furthermore, we analyzed the amount and distribution of mutations affecting the GC composition of Cytb, which were mutations of weak-AT (W) or strong-GC (S) nucleotides according to their direction from the central to peripheral haplotypes (Test of Centrifugal Substitution Bias, TCSB) within each group (Figure 1, as well as Figure 1 in [15]). To reveal more detailed patterns of mutation distributions within the Comephoridae family, a group of SG haplotypes was also included in the analysis. Nonsynonymous mutations were excluded from the calculations in five sites within BGp and SG and one site within BGa.

TCSB revealed some differences among the studied groups (Figure 2). The direction of mutations within BGa showed strong deviation toward the W→S type. Within BGp, the predominance of W→S mutations was much weaker. The SG group demonstrated the opposite pattern: the direction to W was prevalent. Fisher's exact test failed to show any significant differences between BGa and BGp but we found statistically significant differences between BGp and SG and very statistically significant differences between BGa and SG (two-tailed P values of 0.3073, 0.0147, and 0.0006, resp.). Sliding Window Analysis (SWA) of Cytb showed that, within BGa, W→S mutations dominated over S→W for almost the entire Cytb gene, with two maxima located at ~400 and 600 bp (Figure 3). Thus, the dominance of W→S mutations in this group did not appear to be site- or region-specific. Within SG, the opposite picture was observed; with a few exceptions, S→W mutations dominated. No clear patterns were found within BGp.

3.3. RNA Secondary Structure. As noted above, the main haplotypes of BGa and BGp were separated by the two synonymous substitutions T→C located in close proximity to each other. This probably led to alterations of the local thermodynamic stability of DNA and corresponding mRNA. To test how these substitutions could affect mRNA folding, the secondary structures of mRNA for the main haplotypes of BGa, BGp, and SG were predicted. The mRNA structures predicted by UNAFold and RNAfold were different for the same Cytb sequences. However, both programs showed that the main haplotypes of BGa and SG had very similar structures while the structure for BGp was considerably different (Figure 4).

According to the tree based on mRNA structures (Figure 5), BGa and SG haplotypes were largely mixed while the representatives of BGp with few exceptions were arranged in a more compact manner: 15 representatives formed one monophyletic group. Although the reliability of the tree was

Table 3: Number of BGa and BGp larvae sampled in the summer (July, September) and winter (February) at height of spawning.

Sympatric groups	July	September		Total number	February		Total number
		Length, mm			Length, mm		
	<10	<10	11–15		11–15	16–20	
BGa	6	8	1	15	3	7	10
BGp	5	7	5	17	6	2	8

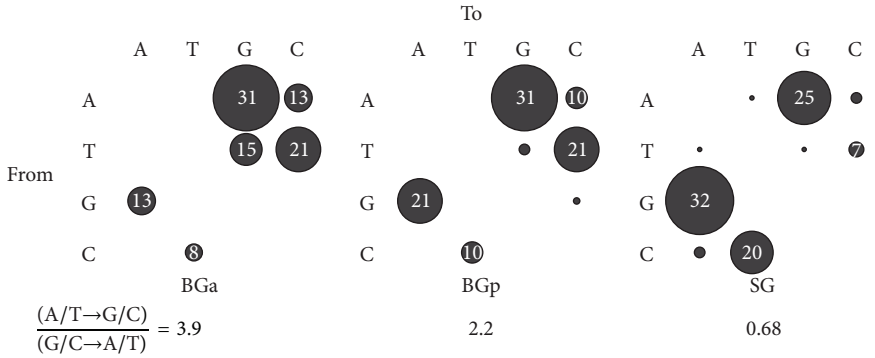


Figure 2: Synonymous mutations in Cytb mapped according to their directions from the central to peripheral haplotypes. Circle sizes and numbers indicate the total number of the mutations as percentages.

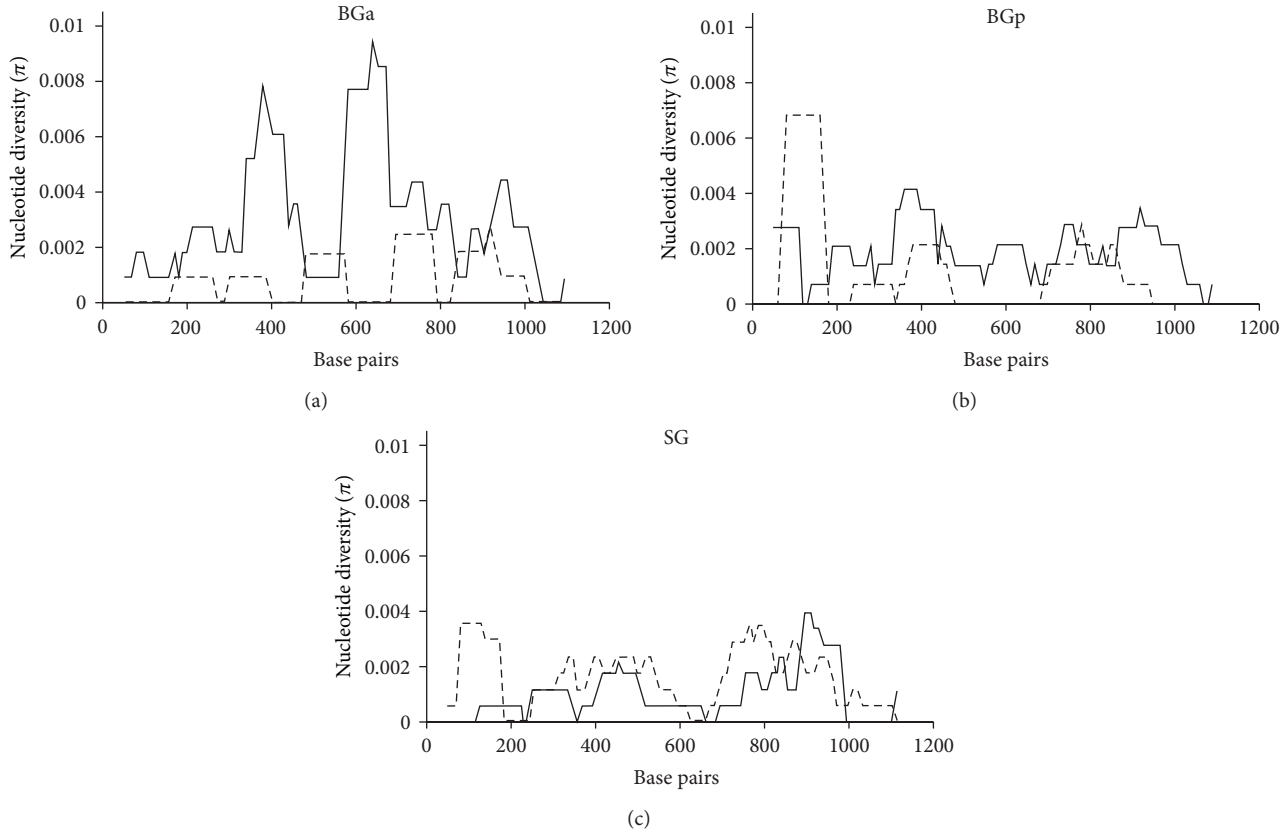


Figure 3: Sliding Window Analysis of nucleotide diversity (π) along the Cytb. Bold solid lines indicate W→S mutations and thin dashed lines indicate S→W mutations.

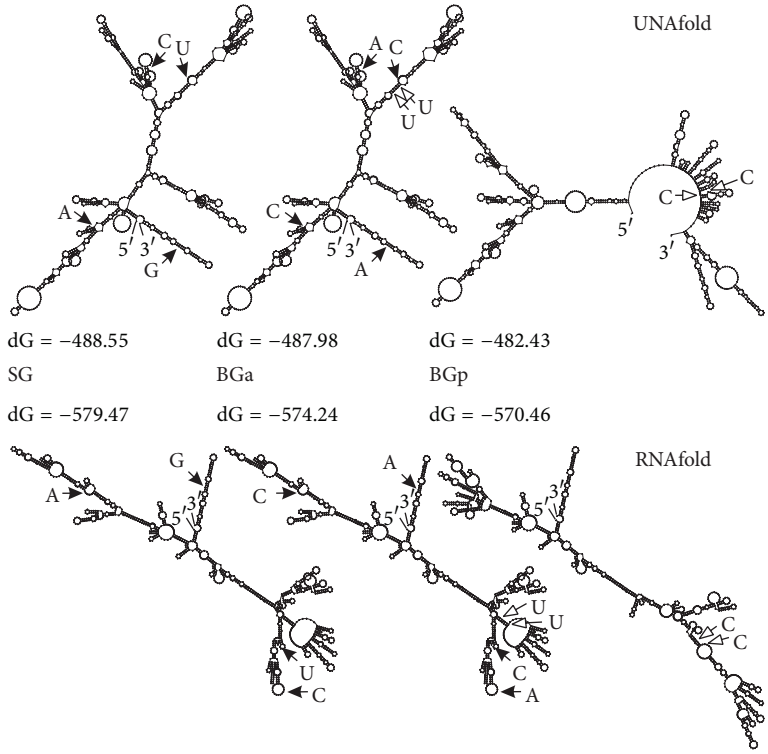


Figure 4: mRNA secondary structure of the main haplotypes of Cytb of BGa, BGp, and SG predicted by UNAFold [16] and RNAfold [17] programs. Black and white arrows indicate substitutions between the main haplotypes of BGa↔SG and BGa↔BGp, respectively.

not evaluated in a traditional way and branch support values were not high, its pattern could serve as an additional indirect indicator of a nonrandom distribution of substitutions during the formation of genetic polymorphism of the groups (i.e., the action of selection).

4. Discussion

Mutational biases and differences in the nucleotide composition of homologous mtDNA sequences in animals and plants are widely known. Several analyses have been performed and various hypotheses have been proposed to explain these phenomena, but the causes remain unclear [23–25]. The MtDNA of Baikal oilfish can serve as a good example of mutational biases caused by the action of translational selection.

As noted above, very little is known about the post-transcriptional regulation of gene expression in synonymous sites of mtDNA. Many tests and methods used to study these processes in the nuclear genome have been ineffective for mtDNA. Jia and Higgs suggested tests for context-dependent mutation and translational selection in mtDNA [26]. However, these tests could hardly be used for phylogenetically closely related species and groups (e.g., Baikal oilfish), especially when the analysis is restricted to a small DNA fragment (Cytb). The approach used here (TCSB + SWA + RSSP + phylogenetic) was definitely not universal and did not show absolute proof of translational selection in the mtDNA of golomyankas. Nevertheless, there were serious grounds for such assumptions for the following reasons: the distinguishing feature of the studied objects was a close phylogenetic proximity of sympatric species and groups that had similar nucleotide sequence compositions of mtDNA [15, 27] on the one hand and had different vectors of Cytb mutational processes in the different groups on the other hand (Figure 2). One of our results suggested that the distribution of larvae was not consistent with the previously proposed hypothesis of the origin of BGa and BGp representatives from different spawning stock [15] and that the representatives of both BG groups were present in equal amounts in any place within the lake at any time and any age. Taking into account the sympatric character of the formation of intraspecific BG and SG genetic structure and similar demographic characteristics of BGa and BGp [15], the TCSB and SWA differences obtained could not be fully explained by the forces of nonadaptive nature, that is, forces that did not affect the expression of Cytb.

There are several ways of regulating gene expression in synonymous sites associated with mRNA folding, mRNA stability, and mRNA splicing and with preferred synonymous codons [2, 5, 7, 28]. The latter two are apparently not the case in the mtDNA of most animals [26]. The assumption of translational selection was indirectly confirmed by the results of the four tests used here: TCSB, SWA, RSSP, and phylogenetic analysis.

Identified regularities could be associated with the extreme sensitivity of golomyankas to temperature variations

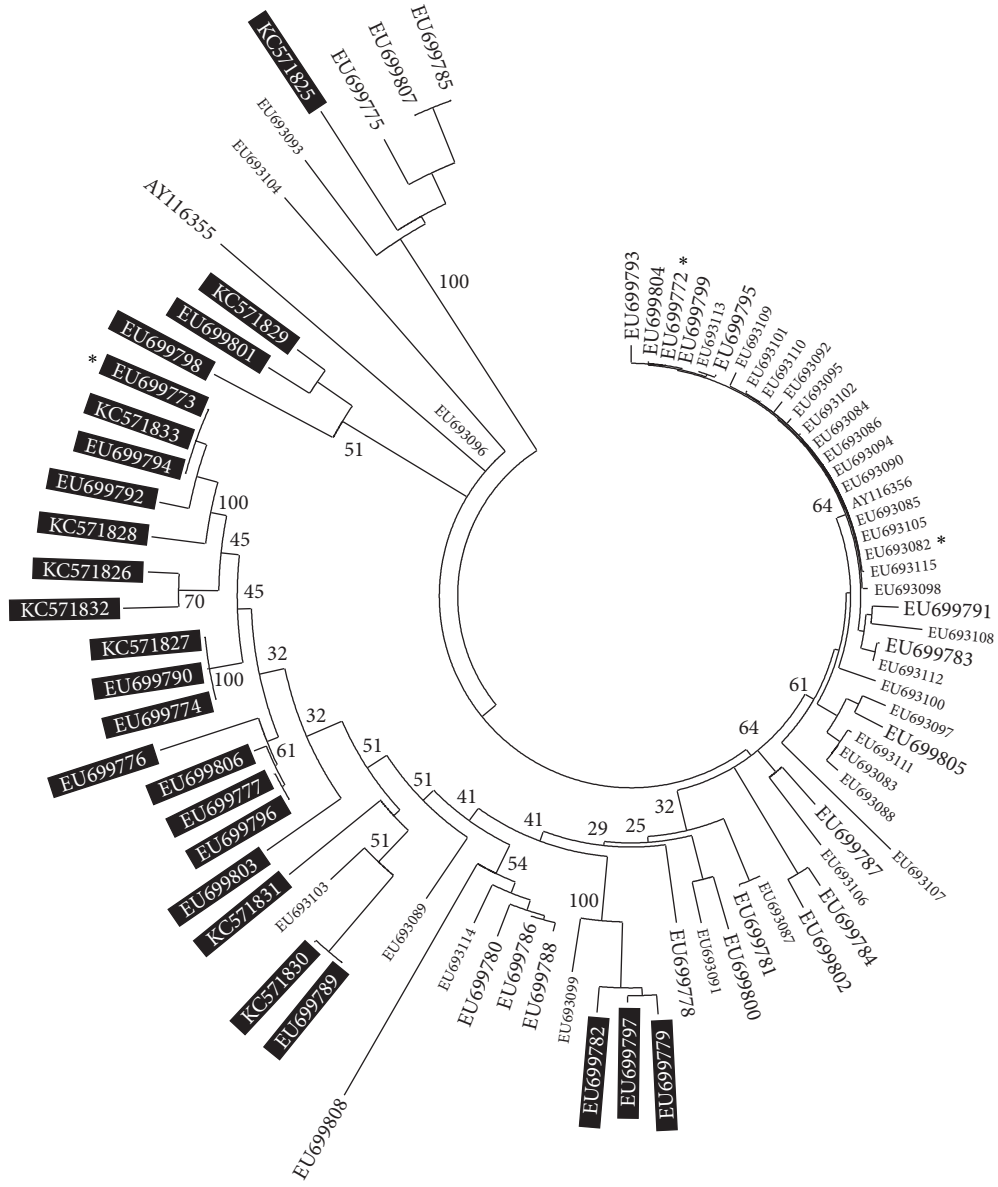


Figure 5: Fitch-Margoliash consensus tree of the Cytb haplotypes of BGa, BGp, and SG based on genetic distances between mRNA secondary structures. Haplotypes are shown as GenBank Accession Numbers. Small letters represent SG; large black letters represent BGa; large white letters in black rectangles represent BGp. Asterisks indicate the main haplotypes. Numbers indicate percentage support values.

[10]. It is known, for example, that the initial stages of thermal alterations in the intensity of oxidative phosphorylation in cardiac muscle cells of Atlantic wolffish occur in enzymatic complex III, which includes the cytochrome b protein [29]. It can be assumed that the synonymous diversity of Cytb within the species and groups of golomyankas could be formed to some extent as a result of the selection pressure of different intensities and directions. We assumed that the genetic diversity within BGa was apparently formed under the action of selection that was directional with respect to the type of mutations against W nucleotides and diversifying with respect to mutation localization (Figure 3). It is possible that selection was associated with the expression of electron transport chain genes by altering the thermodynamic characteristics of the mRNA and consequent refolding. The compactness of BGp on the tree (Figure 5) could indicate an increasing role of a stabilizing selection, which supports a specific mRNA structure. It can be assumed from the two substitutions that separate BGa and BGp made it possible for the representatives of BGp to obtain a partial selective advantage, acting, for example, only under certain environmental conditions so that the total domination of BGp haplotypes within BG did not occur. In that case, the formation of BGp could have happened very quickly in geological time scales and much earlier than the emergence of SG. This could explain some of the discrepancies between the levels of nucleotide diversity within the groups and the number of substitutions between them. For SG and BGp, which have

passed through the bottleneck stage and currently have a large population size [9, 15], one would expect a positive correlation to some extent between the distance from the ancestral line (BGa ↔ BGp, BGa ↔ SG) and the magnitude of nucleotide diversity. However, we actually observed an inverse relationship (Figure 1).

One more important result of this work was that golomyankas could serve as a valuable data source for further study of the regulation of mtDNA gene expression and the role of synonymous substitutions in these processes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors are grateful to I. V. Khanaev for his help with sampling and to E. V. Dzyuba for valuable discussion. This work was supported by Budget Research Projects no. VI.50.1.4 and in part by the Russian Foundation for Basic Research (Grant nos. 08-04-01434-a and 08-04-10123-k). Language support was provided by The Charlesworth Group Author Services (http://www.charlesworthauthorservices.com/).

References

[1] J. V. Chamary, J. L. Parmley, and L. D. Hurst, "Hearing silence: non-neutral evolution at synonymous sites in mammals," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 98–108, 2006.

[2] A. G. Nackley, S. A. Shabalina, I. E. Tchivileva et al., "Human catechol-*O*-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure," *Science*, vol. 314, no. 5807, pp. 1930–1933, 2006.

[3] A. A. Komar, "SNPs, silent but not invisible," *Science*, vol. 315, no. 5811, pp. 466–467, 2007.

[4] J. L. Parmley and L. D. Hurst, "How do synonymous mutations affect fitness?" *BioEssays*, vol. 29, no. 6, pp. 515–519, 2007.

[5] S. A. Shabalina, N. A. Spiridonov, and A. Kashina, "Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity," *Nucleic Acids Research*, vol. 41, no. 4, pp. 2073–2094, 2013.

[6] C. Kimchi-Sarfaty, J. M. Oh, I.-W. Kim et al., "A "silent" polymorphism in the *MDR1* gene changes substrate specificity," *Science*, vol. 315, no. 5811, pp. 525–528, 2007.

[7] S. A. Shabalina, A. Y. Ogurtsov, and N. A. Spiridonov, "A periodic pattern of mRNA secondary structure created by the genetic code," *Nucleic Acids Research*, vol. 34, no. 8, pp. 2428–2437, 2006.

[8] V. G. Sideleva, *The Endemic Fishes of Lake Baikal*, Backhuys Publishers, Leiden, The Netherlands, 2003.

[9] G. V. Starikov, *Lake Baikal Golomyanka*, Nauka, Novosibirsk, Russia, 1977, (Russian).

[10] D. N. Taliev, *Lake Baikal Sculpins (Cottoidei)*, USSR Academy of Sciences, Moscow, Russia, 1955, (Russian).

[11] Z. A. Chernyaev, "Morphoecological peculiarities of reproduction and growth of big golomyanka *Comephorus baicalensis* (Pallas)," *Journal of Ichthyology*, vol. 14, pp. 990–1003, 1974.

[12] L. V. Zubina, L. V. Dzyuba, and A. N. Zaitseva, "Are we real witnesses of life cycle transformation of hydrobionts (case study: Lake Baikal golomyanka)?" in *Recent Problems of Hydrobiology in Siberia*, V. I. Romanov, Ed., pp. 40–41, Tomsk State University, Tomsk, Russia, 2001, (Russian).

[13] E. V. Dzyuba, "Two coexisting species of Baikal golomyankas, *Comephorus baicalensis* and *C. dybowski*: Seasonal dynamics of juveniles and their feeding," *Hydrobiologia*, vol. 568, no. 1, pp. 111–114, 2006.

[14] C. D. Hurst, S. E. Bartlett, W. S. Davidson, and I. J. Bruce, "The complete mitochondrial DNA sequence of the Atlantic salmon, *Salmo salar*," *Gene*, vol. 239, no. 2, pp. 237–242, 1999.

[15] V. I. Teterina, L. V. Sukhanova, and S. V. Kirilchik, "Molecular divergence and speciation of Baikal oilfish (Comephoridae): facts and hypotheses," *Molecular Phylogenetics and Evolution*, vol. 56, no. 1, pp. 336–342, 2010.

[16] N. R. Markham and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization," *Methods in Molecular Biology*, vol. 453, pp. 3–31, 2008.

[17] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatshefte für Chemie Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1989.

[18] J. Sambrook, E. P. Fritsch, and T. Maniatis, *Molecular Cloning: Laboratory Manual*, Cold Spring Harbor Laboratory Press, New York, NY, USA, 1989.

[19] H.-J. Bandelt, P. Forster, and A. Röhl, "Median-joining networks for inferring intraspecific phylogenies," *Molecular Biology and Evolution*, vol. 16, no. 1, pp. 37–48, 1999.

[20] A. J. Vilella, A. Blanco-Garcia, S. Hutter, and J. Rozas, "VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data," *Bioinformatics*, vol. 21, no. 11, pp. 2791–2793, 2005.

[21] J. Felsenstein, "PHYLIP—phylogeny inference package (version 3.2)," *Cladistics*, vol. 5, pp. 164–166, 1989.

[22] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[23] K. Tamura, "The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA," *Molecular Biology and Evolution*, vol. 9, no. 5, pp. 814–825, 1992.

[24] J. W. O. Ballard, "Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup," *Journal of Molecular Evolution*, vol. 51, no. 1, pp. 48–63, 2000.

[25] C. Haag-Liautard, N. Coffey, D. Houle, M. Lynch, B. Charlesworth, and P. D. Keightley, "Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*," *PLoS Biology*, vol. 6, no. 8, article e204, 2008.

[26] W. Jia and P. G. Higgs, "Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection," *Molecular Biology and Evolution*, vol. 25, no. 2, pp. 339–351, 2008.

[27] T. Kontula, S. V. Kirilchik, and R. Väinölä, "Endemic diversification of the monophyletic cottoid fish species flock in Lake Baikal

explored with mtDNA sequencing," *Molecular Phylogenetics and Evolution*, vol. 27, no. 1, pp. 143–155, 2003.

[28] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin, "Coding-sequence determinants of expression in *Escherichia coli*," *Science*, vol. 324, no. 5924, pp. 255–258, 2009.

[29] H. Lemieux, J.-C. Tardif, J.-D. Dutil, and P. U. Blier, "Thermal sensitivity of cardiac mitochondrial metabolism in an ectothermic species from a cold environment, Atlantic wolffish (*Anarhichas lupus*)," *Journal of Experimental Marine Biology and Ecology*, vol. 384, no. 1-2, pp. 113–118, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 298658, 8 pages
http://dx.doi.org/10.1155/2015/298658

Research Article

Molecular Systematics of the Phoxinin Genus *Pteronotropis* (Otophysi: Cypriniformes)

Richard L. Mayden[1] and Jason S. Allen[2]

[1]*Department of Biology, Saint Louis University, 3507 Laclede Avenue, St. Louis, MO 63103, USA*
[2]*Department of Biology, Saint Louis Community College, Meramec Campus, 11333 Big Bend Road, St. Louis, MO 63122, USA*

Correspondence should be addressed to Richard L. Mayden; cypriniformes@gmail.com

Received 22 June 2014; Revised 13 December 2014; Accepted 13 December 2014

Academic Editor: Vassily Lyubetsky

The genus *Pteronotropis* is widely distributed along the gulf slope of eastern North America from Louisiana to Florida and rivers in South Carolina along the Atlantic slope. *Pteronotropis* have very distinctive, flamboyant coloration. The habitats most frequently associated with these species include heavily vegetated backwater bayous to small sluggish or flowing tannin-stained streams. Although *Pteronotropis* is recognized as a valid genus, no phylogenetic analysis of all the species has corroborated its monophyly. In recent years, four additional species have been either described or elevated from synonymy: *P. merlini*, *P. grandipinnis*, *P. stonei*, and *P. metallicus*, with the wide-ranging *P. hypselopterus* complex. To examine relationships within this genus and test its monophyly, phylogenetic analyses were conducted using two nuclear genes, recombination activating gene 1, RAG1, and the first intron of S7 ribosomal protein gene in both maximum parsimony and Bayesian analyses. In no analysis was *Pteronotropis*, as currently recognized, recovered as monophyletic without the inclusion of the currently recognized *Notropis harperi*, herein referred to as *Pteronotropis*. Two major clades are supported: one inclusive of *P. hubbsi*, *P. welaka*, and *P. harperi* and the second inclusive of *P. signipinnis*, *P. grandipinnis*, *P. hypselopterus* plus *P. merlini* sister to *P. euryzonus*, and *P. metallicus* plus *P. stonei*.

1. Introduction

The subfamily (or family) Leuciscinae includes all cyprinid species in North America, except *Notemigonus*, and species across Eurasia. Many of the species of this North American fauna have been examined in different phylogenetic studies at varying degrees of universality using both morphological and molecular data. Initial morphological studies by Mayden [1] and Coburn and Cavender [2] revealed exciting new relationships and a reclassification of the North American fauna. These studies were followed with several molecular analyses of different major lineages, genera, and species groups that supported many, but not all, of the monophyletic groups previously identified in one or both of the above studies [3–7]. However, not all proposed genera have been examined for species relationships using molecular markers.

One such genus in North America with an increasing and intriguing diversity, biology, and geographic distribution, as well as complex taxonomic history, is *Pteronotropis*. This genus contains one of North America's most colorful shiners. *Pteronotropis hubbsi* and *P. welaka* are relatively slender-body species that, in breeding males, possess enlarged dorsal fins, whereas the remaining species, *P. euryzonus*, *P. hypselopterus*, *P. merlini*, *P. grandipinnis*, *P. stonei*, *P. metallicus*, and *P. signipinnis*, are more deep-bodied and lack enlarged dorsal fins in breeding males. The most frequently associated habitat of these species across their ranges includes deep, backwater bayous, small sluggish tannin-stained streams, and flowing tannin-stained streams, all with ample aquatic vegetation. However, despite several studies on shiners and relatives to date, *Pteronotropis* has received essentially no recent attention as to their relationships and has been proposed to be an unnatural grouping. Herein, we provide the first examination of phylogenetic relationships of all species in the genus (formerly subgenus of *Notropis* [1]) and a test of the monophyly of this purported lineage. Two nuclear genes are used in this analysis because of their previously demonstrated genetic distances and resulting ability to resolve nodes deeper than at

the crown of trees. These genes have been used successfully for resolution of more basal lineages of North American cyprinids by several recent papers [3–10].

Resulting phylogenetic inferences of species of this group and their eventual placement relative to other North American cyprinids are critical as they largely facilitate more process-level questions as to the evolution of the biology of the species and other lineages to better understand the processes of anagenesis and speciation. While multiple papers listed above have made groundbreaking strides in providing a phylogenetic framework where one previously did not exist for North American cyprinids, Hollingsworth et al. [10] provide an excellent evaluation of a subset of the fauna and a novel hypothesis as to habitat shifts for clades with differing rates of speciation. Given that no study has examined all of the species of *Pteronotropis*, we provide a review of the history of the genus and molecular phylogenetic analyses of the species using two nuclear genes that result in identical species relationships based on mitochondrial genes in Mayden and Allen [11].

Taxonomic History. *Pteronotropis* currently includes nine species in rivers and streams distributed along the gulf slope from Louisiana to Florida and along the Atlantic slope as far north as South Carolina. One species, *P. hubbsi*, currently occurs only in southern Arkansas and northern Louisiana but was likely to be more widely distributed in lowland habitats; the conservation status of this species is of concern, as it has not been found in some locations (including southern Illinois) for several decades.

In a study focusing on 566 morphological traits of a large number of cyprinids, Mayden [1] elevated *Pteronotropis* to generic level and included *P. welaka*, *P. signipinnis*, *P. hypselopterus*, and *P. euryzonus* within the genus but left *P. hubbsi* in *Notropis*. Recently, Suttkus and Mettee [12], with no characters, phylogenetic analysis, or substantive phylogenetic argument, maintained that *Pteronotropis* was a subgenus within the genus *Notropis* (as classified before Mayden's [1] analysis) and that this subgenus contained only *P. euryzonus* and the *P. hypselopterus* complex (*P. hypselopterus*, *P. grandipinnis*, *P. stonei*, *P. metallicus*, and *P. merlini*).

The phylogenetic relationships of *Pteronotropis* have been somewhat enigmatic over the years. Species share derived and distinctive color patterns that include bright red-orange to yellow striped dorsal, caudal, and anal fins and a broad dark lateral band extending from the head to the caudal peduncle. The genus was divided into two groups based on morphological and molecular characters [1, 4]. *Pteronotropis signipinnis* was described by Bailey and Suttkus [13] and was considered a member of the genus *Notropis* (subgenus *Pteronotropis* by Fowler [14]), along with *P. hypselopterus*. *Pteronotropis euryzonus* [15] was later added to this subgenus and was considered a close relative to *P. hypselopterus*; however, neither of the above two studies included *P. hubbsi* or *P. welaka* and they were conducted in a prephylogenetic era. *Pteronotropis hubbsi* was described by Bailey and Robison [16] and was thought to be closely related to *P. welaka*; at that time, neither species was allocated to the subgenus *Pteronotropis*. In a study utilizing twenty-one allozyme loci, Dimmick [17] examined nine species (mostly *Pteronotropis*). This allozyme analysis revealed *Pteronotropis* as nonmonophyletic, with *P. hubbsi* and *P. welaka* as distantly related and *N. signipinnis* and *N. hypselopterus* as sister species. Consequently, Dimmick [17] argued that all of the morphological characters of Bailey and Robison [16], thought to indicate a close relationship between *P. hubbsi* and *P. welaka*, were the result of convergent evolution.

In the first sequence analysis of this group, Simons et al. [18] used mitochondrial cytochrome *b* gene and failed to corroborate *Pteronotropis* as a monophyletic group. With both parsimony and likelihood analyses, *P. euryzonus* was sister to *P. hypselopterus* and an unrelated clade included *P. signipinnis* sister to *P. hubbsi* plus *P. welaka*. Later, in a subsample of *Pteronotropis* species, Simons et al. [4], using two mitochondrial genes (12S, 16S), and Bufalino and Mayden [5, 6], using two nuclear loci (RAG1, S7), found *Pteronotropis* as monophyletic but, again, only with the inclusion of "*Notropis*" *harperi*; however, neither of these analyses included all species of the genus. Other early molecular data and analyses also failed to resolve the phylogenetic relationships of the above species that were generally phenetically similar. Most recently a study by Hollingsworth et al. [10], using one mtDNA gene and nDNA genes, corroborated the monophyly of a subsample of species of *Pteronotropis* that also included *N. harperi*.

While there have been several efforts testing the monophyly of *Pteronotropis*, its composition, and at resolving the phylogenetic relationships of species since its elevation to genus, no single study has included all of the species in the genus and appropriate outgroups based on earlier studies and some did not include the morphologically similar *Notropis harperi*. With the elevation of species from synonymy with *P. hypselopterus* and the description of a new species [12], the complexity involved in testing the monophyly of the genus and species relationships have become even more biologically interesting. While Suttkus and Mettee [12] did provide dialogue invoking phylogenetic terminology as to species relationships, their study contained no phylogenetic analyses, no discussions of character homology, or any morphological or molecular synapomorphies. To date, no investigation has been completed for this group inclusive of all of the purported species of *Pteronotropis*. Thus, the objectives of the current study are twofold: (1) testing the monophyly of the genus and (2) examining relationships of all of the purported species of the genus using two nuclear genes.

2. Materials and Methods

2.1. Specimens and DNA Extraction/Amplification and Alignment. Museum catalogue numbers for vouchers in this study include UAIC (University of Alabama Ichthyological Collection) and SLUM (Saint Louis University Museum). Specimens examined in this study were either frozen at Saint Louis University, preserved in 95% ethanol, or captured alive and transported to Saint Louis University (Table 1). Outgroup taxa included species from the genera *Cyprinella*, *Lythrurus*, and *Notropis*. Species of *Cyprinella* were included

Table 1: Species, localities, and GenBank numbers of specimens used for sequencing and analyses of S7 and RAG1.

(a)

Species and drainage	Stream, county, state	Catalogue number	S7	RAG1	Extraction
Pteronotropis euryzonus					
Chattahoochee R.	Maringo Cr., Russell, AL	UAIC 12229	KM048270	KJ634252	22
Chattahoochee R.	Snake Cr., Russell, AL	UAIC 10493	KM048276	KJ634258	51
Chattahoochee R.	Snake Cr., Russell, AL	UAIC 10493	KM048277	KJ634259	52
Pteronotropis grandipinnis					
Apalachicola R.	Irwin Mill Cr., Houston, AL	No voucher	KM048265	KJ634247	12
Apalachicola R.	Irwin Mill Cr., Houston, AL	No voucher	KM048266	KJ634248	13
Pteronotropis hypselopterus					
Mobile R.	Cedar Cr., Mobile, AL	UAIC 12730	KM048256	KJ634238	01
Mobile R.	Cedar Cr., Mobile, AL	UAIC 12730	KM048257	KJ634239	02
Mobile R.	Cedar Cr., Mobile, AL	UAIC 12730	KM048258	KJ634240	03
Alabama R.	Little Reedy Cr., AL	UAIC 14326	KM048269	KJ634251	18
Pteronotropis hubbsi					
Ouachita R.	Backwater pond, Ouachita, LA	UAIC 11928	KM048261	KJ634243	06
Ouachita R.	Backwater pond, Ouachita, LA	UAIC 11928	KM048262	KJ634244	07
Little R.	Little R., McCurtain, OK	UAIC 12053	KM048273	KJ634255	41
Pteronotropis merlini					
Pea R.	Clearwater Cr., Coffee, AL	No voucher	KM048267	KJ634249	16
Pea R.	Clearwater Cr., Coffee, AL	No voucher	KM048268	KJ634250	17
Pteronotropis metallicus					
Suwannee R.	Sampson R., Bradford, FL	UF 158855	KM048278	KJ634260	96
Suwannee R.	Sampson R., Bradford, FL	UF 158855	KM048279	KJ634261	97
Pteronotropis signipinnis					
Pascagoula R.	Beaverdam Cr., Forest, MS	UAIC 13416	KM048259	KJ634241	04
Pascagoula R.	Beaverdam Cr., Forest, MS	UAIC 13416	KM048260	KJ634242	05
Mobile R.	Cedar Cr., Mobile, AL	UAIC 12730	KM048271	KJ634253	23
Mobile R.	Cedar Cr., Mobile, AL	UAIC 12730	KM048272	KJ634254	24
Pteronotropis stonei					
N. Fork Edisto R.	Murphy Mill Cr., Calhoun, SC	SLUM 1121	KM048281	KJ634263	101
N. Fork Edisto R.	Murphy Mill Cr., Calhoun, SC	SLUM 1121	KM048280	KJ634262	100
Combahee R.	Savannah Cr., Colleton, SC	SLUM 1122	KM048282	KJ634264	102
Pteronotropis welaka					
Cahaba R.	Lightsey pond, Bibb, AL	UAIC 10391	KM048263	KJ634245	10
Cahaba R.	Lightsey pond, Bibb, AL	UAIC 10391	KM048264	KJ634246	11
Pearl R.	Lees Cr., Washington, LA	UAIC 12205	KM048274	KJ634256	48
Mobile Bay	Lees Cr., Washington, LA	UAIC 12205	KM048275	KJ634257	49
Pteronotropis harperi			GU134235	GU136332	

(b)

Outgroup taxa (note that *Pteronotropis harperi* was also originally an outgroup species)		
Species	S7	RAG1
Cyprinella formosa	GU 134192	GU136293
Lythrurus fumeus	GU134222	GU136231
Lythrurus umbratilis	GU134223	GU136322
Nocomis leptocephalus	GU134236	GU136333
Notropis asperifrons	GU134231	GU136330

(b) Continued.

Outgroup taxa (note that *Pteronotropis harperi* was also originally an outgroup species)		
Species	S7	RAG1
Notropis atherinoides	GU134232	EF452832
Notropis blennius	GU134234	GU136331
Notropis leuciodus	GU134237	GU136334
Notropis maculatus	GU134238	GU136335
Notropis ortenburgeri	GU134240	GU136337
Notropis nazas	GU134239	GU136336
Notropis stilbius	GU134241	GU136338
Notropis volucellus 1	GU134242	GU136339
Notropis volucellus 2	GU134243	GU136340

as outgroup taxa due to previous studies indicating their close relationships to *Pteronotropis*. Because this analysis focuses on nuclear gene variation as it contributes to phylogenetic relationships and the inadequate sampling of all relevant taxa in previous studies, cytochrome *b* sequences of previous mitochondrial analyses are not included. Genomic DNA was extracted using the QIAGEN QIAamp tissue kit according to the manufacturer's recommendations (QIAGEN, Valencia, CA). The two nuclear genes included recombination activating gene 1, RAG1, and the first intron of S7 ribosomal protein gene. Both genes were amplified, via PCR, and internal primers amplification and sequencing were developed for S7. These include the forward primers 5′-GCCACTGCAGCCGCCATAAT-3′ and 5′-GCCCCA-GCTTTCCACCCATTAC-3′ and reverse primers 5′-CCC-GAGGGCTGTGAGGAGTAA-3′ and 5′-CCCCCTCAG-CCGCCGACTA-3′. Universal primers for RAG1 and S7 were detailed in López et al. [19] and Chow and Hazama [20], respectively. In addition, both forward and reverse internal primers were developed for S7. For RAG1, each 25 μL PCR reaction consisted of 2 μL of DNTPs, 2.5 μL of 10X Taq buffer, 3 μL of both forward and reverse primers, 10.375 μL of dH_2O, 1 μL of Taq polymerase, or .125 μL of HotStart *Taq* Polymerase (QIAGEN, Valencia, CA). Amplifications consisted of 35 cycles of an initial denaturation of 95°C for 15 minutes with an additional denaturation of 94°C for 40 seconds. This was followed by an annealing temperature of 55°C for 1 minute, an initial extension of 72°C for 90 seconds, and a final extension of 72°C for 5 minutes. Conditions for S7 were identical except the annealing temperature was set at 59°C. For the S7 intron, products that failed to amplify using the universal primers were reamplified using nested PCR reactions with the same conditions except for specific annealing temperatures as specified by the chemistry for the internal primers. Taxa failing to amplify with internal primers were cloned using the pGEM-T Easy Vector System kit (PROMEGA, Madison, WI) as outlined in Lang and Mayden [9]. PCR products were purified using QIAGEN gel extraction kits (QIAGEN, Valencia, CA). Sequencing was performed using a BigDye labeled dideoxy sequencing kit (BigDye) and visualized on an ABI 377 automated sequencer (Auburn University Molecular Genetics Instrumentation Facility, Auburn, AL) or an ABI 3700 (Macrogen Sequencing Facility, Seoul, South Korea). Both the heavy and light strands were sequenced for all samples and the sequences were aligned with Clustal X [21] with reference to the accompanying electropherograms. Some individuals contained heterozygote peaks in the RAG1 data and these heterozygote base pair positions were coded using standard degeneracy codes.

2.2. Phylogenetic Analyses. An incongruence-length difference analysis (ILD [22]) was performed with 1000 replicates to test for incongruence between the RAG1 and S7 data sets. Maximum parsimony (MP) analyses (MPA) were conducted in PAUP*4.0b10 [23]. All analyses consisted of a heuristic search model with 1000 random addition sequence replicates and TBR. Support for the parsimony analyses was generated using bootstrap analysis (BS) with 1000 bootstrap pseudoreplicates [18]. Bayesian analyses (BA) were conducted in Mr. Bayes 3.0b4 [24]. S7 intron all gaps were treated as missing data. The model of sequence evolution was determined using Modeltest v3.04 [25] with single partitions for each marker; the best-fit model for S7 was HKY + G and that for Rag1 was TrN + I + G. BA included four heated Markov chains using the default temperature setting. Log-likelihood scores were plotted against generation time to establish burn-in; trees prior to stationarity were discarded. Post-burn-in trees were used to develop the 50% majority rule consensus tree. Posterior probabilities (PP) were used as an indication of nodal support in BA.

3. Results and Discussion

As the ILD test was nonsignificant for heterogeneity between RAG1 and S7, the gene sequences were analyzed both individually and as a concatenated data set. MP analysis of the aligned 1001 bp of S7 (aligned sequence lengths ranged from 839 to 919 bp) yielded 245 bp parsimony informative sites (12.9%). Analyses of these data resulted in 90 equally parsimonious trees (Figure 1; length = 697, CI = 0.803, and RI = 0.875). The more conservative RAG1 sequences included 1521 bp with 151 bp sites (9.9%) being parsimony

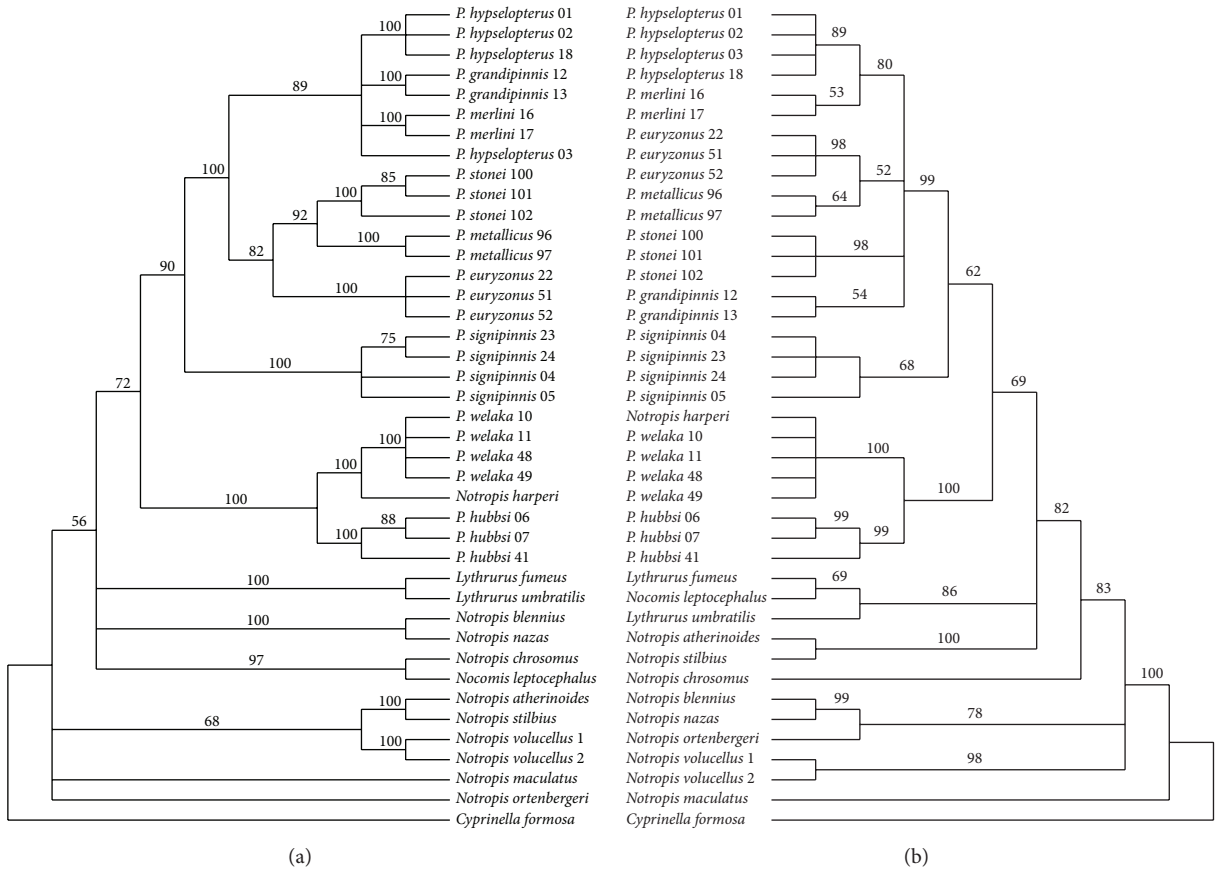


Figure 1: Inferred species relationships of species of *Pteronotropis* based on maximum parsimony analyses of RAG1 (a) and S7 (b). Nodal values indicate bootstrap support.

informative. MP analyses of RAG1 resulted in 46,668 equally parsimonious trees (Figure 1; length = 371 steps, CI = 0.658, and RI = 0.866). Individual BA analyses for each gene resulted in some variations in sister-group relationships but all were consistent and supported the monophyly of *Pteronotropis* (Figure 2). Both MPA and BA of the combined S7 + RAG1 data recovered identical topologies (Figure 3).

As in previous studies involving species of *Pteronotropis*, nuclear sequence variation, neither individual nor combined [5, 6], resolved *Pteronotropis* as a monophyletic group if *Notropis harperi* is excluded from the genus. Constraining *Pteronotropis* to be monophyletic in the S7 + RAG1 data set without *N. harperi* resulted in a significantly worse tree (1246 steps). In both BA and MPA, *Notropis harperi* is resolved as sister to *P. welaka* within the ingroup, a sister-group relationship with strong PP and BS support (Figures 1 and 2). *Pteronotropis hubbsi* is resolved as sister to this clade, also with strong PP and BS support. All three of these taxa (*P. hubbsi* (*P. welaka* + *N. harperi*)) are resolved as monophyletic and sister to the remaining species traditionally referred to as *Pteronotropis* (PP 95, bootstrap 75; Figure 2). The strong support for the monophyly of the (*P. hubbsi* (*P. welaka* + *N. harperi*)) clade (Figures 1 and 2) is logical as the three species are phenetically and ecologically similar. They possess aspects of similar body coloration in life when not in breeding condition and have similar habitat associations [5, 26, 27]. They are found in deep pools with ample aquatic vegetation and in areas where *P. welaka* and *N. harperi* are sympatric they are often taken syntopically in a sample (pers. obs.). The authors are unaware of any studies corroborating nest association in *N. harperi*, as observed in *P. welaka* and *P. hubbsi* [28–30]. In light of the relationships presented here and in Bufalino and Mayden [4, 5] and Hollingsworth et al. [10], studies of *N. harperi* may reveal ecological and behavioral synapomorphies.

In all analyses, *P. signipinnis* is resolved as sister to a clade of remaining species of *Pteronotropis* (Figures 1 and 2). In analyses of S7 and S7 + Rag1 data sets, the latter clade formed two clades: one inclusive of *P. hypselopterus*, *P. grandipinnis*, and *P. merlini* and the other inclusive of *P. euryzonus*, *P. stonei*, and *P. metallicus*. Resolution of the former clade was not complete in either Rag1 or S7 analyses, but both are fully consistent with the phylogeny recovered with the Rag1 + S7 data set. These relationships are in contrast to those hypothesized by Simons et al. [4] based on 12S and 16S ribosomal RNA sequences wherein *P. signipinnis* was resolved as sister to

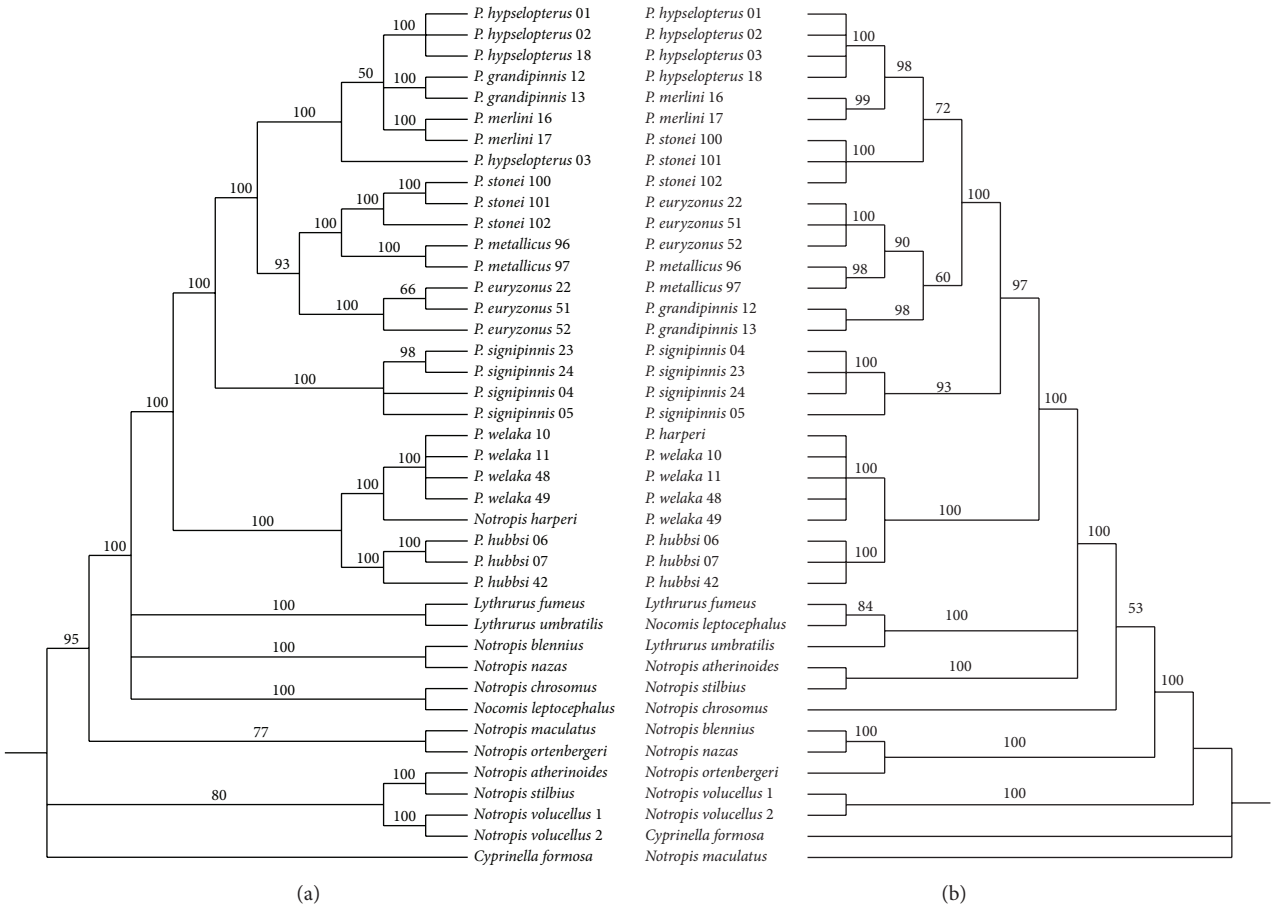


Figure 2: Inferred species relationships of species *Pteronotropis* based on Bayesian analyses of S7 (a) and Rag1 (b). Nodal values indicate posterior probabilities.

P. welaka + *P. hubbsi*. However, this latter study did not include all of the then or currently known species of *Pteronotropis*.

4. Conclusions

Given the consistent sister-group relationship between formerly recognized *Notropis harperi* and *Pteronotropis welaka*, the former species is herein referred to as *Pteronotropis*. Nuclear genes RAG1 and S7 support the long-standing question/hypothesis regarding the monophyly of *Pteronotropis* and provide new insight into the phylogenetic placement of *Pteronotropis harperi* and the basal-most relationships between the species groups (*P. hubbsi*, *P. welaka*, and *P. harperi*) relative to the remaining species of *Pteronotropis*. These relationships are also consistent with those presented by Bailey and Suttkus [13] using mitochondrial gene ND2. In recent years, the general trend in phylogenetics has been to place greater emphasis on the use of nuclear genes, largely because of issues associated with hybridization, intergradation, lineage sorting, and disagreement between gene and species trees [13]. While these nuclear genes have shown a greater ability to resolve relationships at supraspecific levels for this group with greater consistency and stronger branch support, the results presented herein illustrate the benefit in using nuclear genes. However, it is also true that mitochondrial genes have been extremely useful in phylogenetic resolutions [26, 27], and like nuclear genes they also vary in their degree of anagenesis and abilities to resolve trees at different levels of universality. While these and other nuclear genes used in the above-cited papers for Cypriniformes clearly display a reduced phylogenetic signal and are more limited in phylogenetic resolution for relationships of populations and species, they are essential for resolution of deeper nodes. This is to be expected as rates of mutation of many nuclear genes (especially protein coding) are generally not as high as that typically found in most mitochondrial genes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

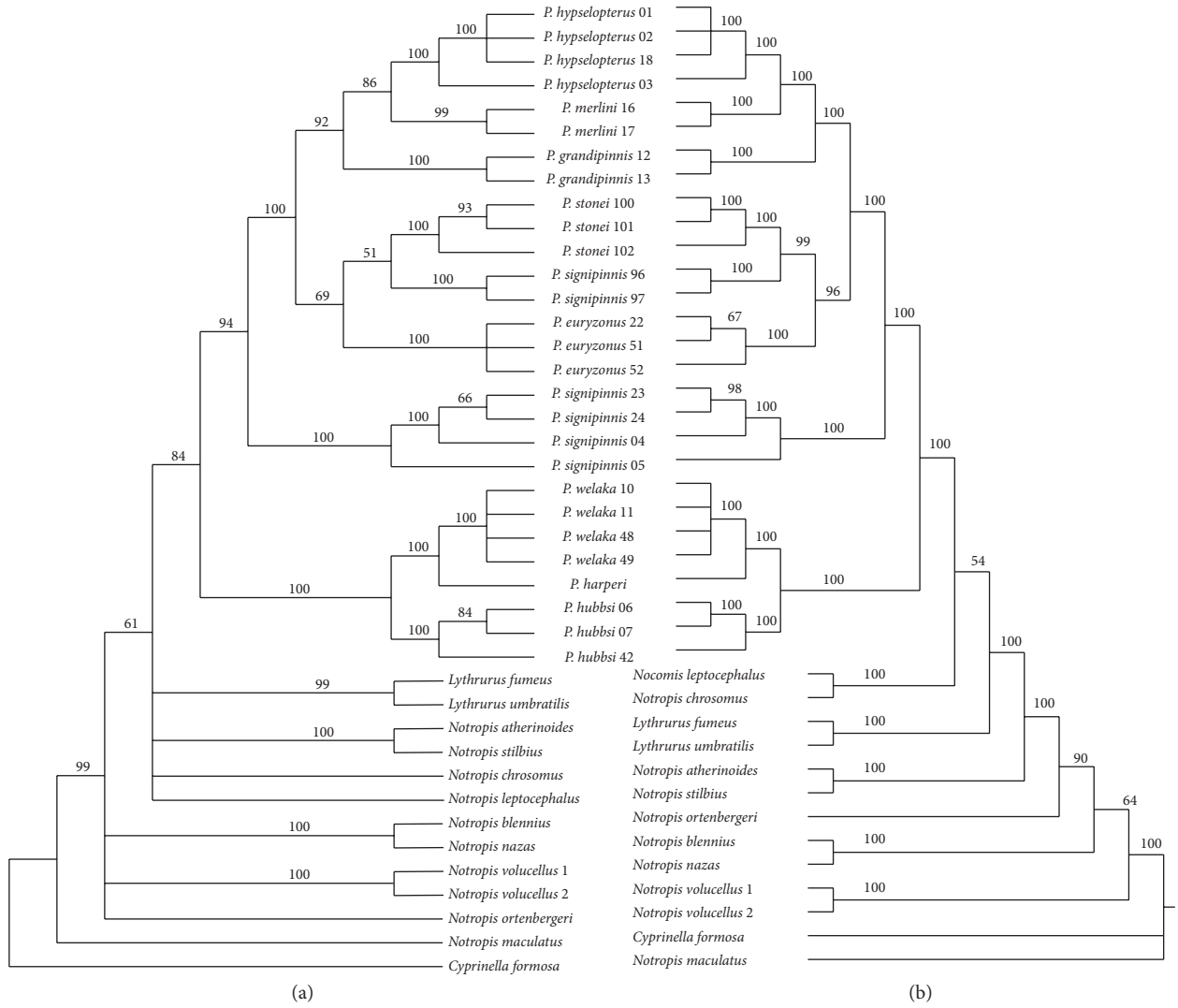


Figure 3: Inferred species relationships of species of *Pteronotropis* based on maximum parsimony and Bayesian analyses of combined Rag1 + S7 (a) and Rag1 + S7 (b), respectively. Nodal values indicate posterior probabilities.

Acknowledgments

The authors thank Brett Albanese, Bud Freeman, Nick Lang, Dave Neely, Larry Page, Brady Porter, Charles Ray, Chip Reinhart, Brian Skidmore, David Smith, and Dustin Smith for providing field assistance or valuable samples. They also thank Lei Yang and Susana Schönhuth for their assistance in the laboratory, Susana Schönhuth for running some analyses presented herein, and Qui Ren, Susana Schönhuth, and Anne Ilvarson for assistance with GenBank. This research was supported in part by NSF grants to Richard L. Mayden (EF-0431326) and the William S. Barnickle Endowment at Saint Louis University.

References

[1] R. L. Mayden, *Phylogenetic Studies of North American Minnows: With Emphasis on the Genus* Cyprinella *(Teleostei, Cypriniformes)*, vol. 80 of *Miscellaneous Publications, Museum of Natural History*, University of Kansas, Lawrence, Kan, USA, 1989.

[2] M. M. Coburn and T. M. Cavender, "Interrelationships of North American cyprinid fishes," in *Systematics, Historical Ecology and North American Freshwater Fishes*, R. L. Mayden, Ed., Stanford University Press, Stanford, Calif, USA, 1992.

[3] S. Perea, M. Böhme, P. Zupančič et al., "Phylogenetic relationships and biogeographical patterns in Circum-Mediterranean subfamily Leuciscinae (Teleostei, Cyprinidae) inferred from both mitochondrial and nuclear data," *BMC Evolutionary Biology*, vol. 10, no. 1, article 265, 2010.

[4] A. M. Simons, P. B. Berendzen, and R. L. Mayden, "Molecular systematics of North American phoxinin genera (Actinopterygii: Cyprinidae) inferred from mitochondrial 12S and 16S ribosomal RNA sequences," *Zoological Journal of the Linnean Society*, vol. 139, no. 1, pp. 63–80, 2003.

[5] A. Bufalino and R. L. Mayden, "Molecular phylogenetics of North American phoxinins (Actinopterygii: Leuciscidae) based on RAG1 and S7 nuclear DNA sequence data," *Molecular Phylogenetics and Evolution*, vol. 55, pp. 274–283, 2010.

[6] A. P. Bufalino and R. L. Mayden, "Phylogenetic relationships of North American phoxinins (Actinopterygii: Cypriniformes: Leuciscidae) as inferred from S7 nuclear DNA sequences," *Molecular Phylogenetics and Evolution*, vol. 55, no. 1, pp. 143–152, 2010.

[7] S. Schönhuth, I. Doadrio, O. Dominguez-Dominguez, D. M. Hillis, and R. L. Mayden, "Molecular evolution of southern North American Cyprinidae (Actinopterygii), with the description of the new genus *Tampichthys* from central Mexico," *Molecular Phylogenetics and Evolution*, vol. 47, no. 2, pp. 729–756, 2008.

[8] S. Schönhuth, A. Perdices, L. Lozano-Vilano, F. J. García-de-León, H. Espinosa, and R. L. Mayden, "Phylogenetic relationships of North American western chubs of the genus *Gila* (Cyprinidae, Teleostei), with emphasis on southern species," *Molecular Phylogenetics and Evolution*, vol. 70, no. 1, pp. 210–230, 2014.

[9] N. J. Lang and R. L. Mayden, "Systematics of the subgenus *Oligocephalus* (Teleostei: Percidae: *Etheostoma*) with complete subgeneric sampling of the genus *Etheostoma*," *Molecular Phylogenetics and Evolution*, vol. 43, no. 2, pp. 605–615, 2007.

[10] P. R. Hollingsworth, A. M. Simons, J. A. Fordyce, and C. D. Hulsey, "Explosive diversification following a benthic to pelagic shift in freshwater fishes," *BMC Evolutionary Biology*, vol. 13, no. 1, article 272, 2013.

[11] R. L. Mayden and J. S. Allen, "Phylogeography of *Pteronotropis signipinnis*, *P. euryzonus*, and the *P. hypselopterus* species complex, with comments on the history of streams draining into the Gulf of Mexico and the southern Atlantic," *Molecular Phylogenetics*. In press.

[12] R. D. Suttkus and M. F. Mettee, "Analysis of four species of *Notropis* included in the subgenus *Pteronotropis* Fowler, with comments on relationships, origins, and dispersion," *Geological Survey of Alabama*, vol. 170, pp. 1–50, 2001.

[13] R. M. Bailey and R. D. Suttkus, "*Notropis signipinnis*, a new cyprinid fish from the southeastern United States," *University of Michigan Museum of Zoology Occasional Paper*, vol. 542, pp. 1–15, 1952.

[14] H. W. Fowler, "Notes on South Carolina fresh-water fishes," in *Contributions of the Charleston Museum*, vol. 7, pp. 1–28, 1935.

[15] R. D. Suttkus, *A taxonomic study of five cyprinid fishes related to Notropis hypselopterus of southeastern United States [Ph.D. thesis]*, Cornell University, New York, NY, USA, 1950.

[16] R. M. Bailey and H. W. Robison, "*Notropis hubbsi*, a new cyprinid fish from the Mississippi River basin with comments on *Notropis welaka*," *Occasional Papers of the Museum of Zoology University Michigan*, vol. 638, pp. 1–21, 1978.

[17] W. W. Dimmick, "Phylogenetic relationships of *Notropis hubbsi*, *N. welaka*, and *N. emiliae* (Cypriniformes: Cyprinidae)," *Copeia*, vol. 1987, no. 2, pp. 316–325, 1987.

[18] A. M. Simons, E. K. Knott, and R. L. Mayden, "Assessment of monophyly of the minnow genus *Pteronotropis* (Teleostei: Cyprinidae)," *Copeia*, no. 4, pp. 1068–1075, 2000.

[19] J. A. López, W. J. Chen, and G. Ortí, "Esociform phylogeny," *Copeia*, vol. 2004, no. 3, pp. 449–464, 2004.

[20] S. Chow and K. Hazama, "Universal PCR primers for S7 ribosomal protein gene introns in fish," *Molecular Ecology*, vol. 7, no. 9, pp. 1255–1256, 1998.

[21] T. A. Hall, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, *BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis*, ver. 5.0.9, Department of Microbiology, North Carolina State University, Raleigh, NC, USA, 2001.

[22] J. S. Farris, M. Källersjö, A. G. Kluge, and C. Bult, "Testing significance of incongruence," *Cladistics*, vol. 10, no. 3, pp. 315–319, 1994.

[23] D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (*and other Methods)*, Verison 4, Sinauer Associates, Sunderland, Mass, USA, 2003.

[24] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[25] D. Posada and K. A. Crandall, "MODELTEST: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[26] W. S. Moore, "Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees," *Evolution*, vol. 49, no. 4, pp. 718–726, 1995.

[27] M. Miya and M. Nishida, "The mitogenomic contributions to molecular phylogenetics and evolution of fishes: a 15-year retrospect," *Ichthyological Research*, 2014.

[28] D. E. Fletcher and B. M. Burr, "Reproductive biology, larval description, and diet of the North American bluehead shiner, *Pteronotropis hubbsi* (Cypriniformes: Cyprinidae), with comments on conservation status," *Ichthyological Explorations of Freshwaters*, vol. 3, pp. 219–218, 1992.

[29] C. E. Johnston and C. L. Knight, "Life-history traits of the bluenose shiner, *Pteronotropis welaka* (Cypriniformes: Cyprinidae)," *Copeia*, no. 1, pp. 200–205, 1999.

[30] D. E. Fletcher, "Male ontogeny and size-related variation in mass allocation of bluenose shiners (*Pteronotropis welaka*)," *Copeia*, vol. 2, pp. 479–486, 1999.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 649034, 15 pages
http://dx.doi.org/10.1155/2014/649034

Research Article

The Characteristics of Ubiquitous and Unique *Leptospira* Strains from the Collection of Russian Centre for Leptospirosis

Olga L. Voronina, Marina S. Kunda, Ekaterina I. Aksenova, Natalia N. Ryzhova, Andrey N. Semenov, Evgeny M. Petrov, Lubov V. Didenko, Vladimir G. Lunin, Yuliya V. Ananyina, and Alexandr L. Gintsburg

N.F. Gamaleya Institute for Epidemiology and Microbiology, Ministry of Health of Russia, Gamaleya Street 18, Moscow 123098, Russia

Correspondence should be addressed to Olga L. Voronina; kirolg3@newmail.ru

Received 22 May 2014; Revised 29 July 2014; Accepted 5 August 2014; Published 2 September 2014

Academic Editor: Vassily Lyubetsky

Background and Aim. *Leptospira*, the causal agent of leptospirosis, has been isolated from the environment, patients, and wide spectrum of animals in Russia. However, the genetic diversity of *Leptospira* in natural and anthropurgic foci was not clearly defined. *Methods*. The recent MLST scheme was used for the analysis of seven pathogenic species. 454 pyrosequencing technology was the base of the whole genome sequencing (WGS). *Results*. The most wide spread and prevalent *Leptospira* species in Russia were *L. interrogans, L. kirschneri,* and *L. borgpetersenii*. Five STs, common for Russian strains: 37, 17, 199, 110, and 146, were identified as having a longtime and ubiquitous distribution in various geographic areas. Unexpected properties were revealed for the environmental *Leptospira* strain Bairam-Ali. WGS of this strain genome suggested that it combined the features of the pathogenic and nonpathogenic strains and may be a reservoir of the natural resistance genes. Results of the comparative analysis of *rrs* and *rpoB* genes and MLST loci for different *Leptospira* species strains and phenotypic and serological properties of the strain Bairam-Ali suggested that it represented separate *Leptospira* species. *Conclusions*. Thus, the natural and anthropurgic foci supported ubiquitous *Leptospira* species and the pool of genes important for bacterial adaptivity to various conditions.

1. Introduction

Leptospira is a genus of bacteria that is encountered in all geographical areas, except arctic and arid regions. Some *Leptospira* are responsible for leptospirosis, a natural-focus disease. According to the List of Prokaryotic names with Standing in Nomenclature, the *Leptospira* genus has 21 species [1]. Seven of the species have been established as pathogenic: *L. interrogans*, *L. borgpetersenii*, *L. kirschneri*, *L. noguchii*, *L. santarosai*, *L. weilii*, and *L. alexanderi*. Two species, *L. alstonii* and *L. kmetyi*, are candidates for assignment to the pathogenic group; they are rarely isolated, and the sources of their isolation are not ill humans. *L. alstonii* was isolated from a frog and *L. kmetyi* was isolated from soil. Phylogenetic analysis with 16S rRNA gene sequences showed that *L. alstonii* and *L. kmetyi* clustered with the pathogenic *Leptospira* species [2, 3]. Five species, *L. broomii, L. fainei, L. inadai, L. licerasiae,* and *L. wolffii,* are classified as intermediate, and six species, *L. biflexa, L. meyeri, L. terpstrae, L. vanthielii, L. wolbachii,* and *L. yanagawae,* are nonpathogenic [4]. One new species, *L. idonii*, isolated from the environmental water, is candidate for assignment to the nonpathogenic group. This species is placed within the clade of the known saprophytic species of the genus *Leptospira* on the 16S rRNA gene-based phylogenetic analysis [5].

According to WHO guidance [6], the incidence of leptospirosis ranges from about 0.1–1 per 100 000 persons per year in temperate climates to 10–100 per 100 000 in the humid tropics. In the Russian Federation, only 0.01 cases per 100 000 were reported in 2013. During 2012-2013, 506 cases were registered in Russia according to the report of the Federal Service for Supervision of Consumer Rights Protection and Human Welfare (http://rospotrebnadzor.ru/). Long-term control of the multiple natural and anthropurgic

foci in the USSR has been organized with the participation of the MoH Centre for Leptospirosis, an action which may be responsible for the decrease in incidence. During the registration period, strains of *Leptospira* species were isolated, from the animals (maintenance and supplementary hosts of the *Leptospira*), human patients, and environment, and saved in the Gamaleya Institute Microbial Collection (GIMC).

Among the strains of *Leptospira* species in GIMC one strain was mysterious and not closely related to any *Leptospira*, either pathogenic or saprophytic; it was named *Leptospira* spp. strain Bairam-Ali. Bairam-Ali was isolated from the water of a drainage canal in Turkmenia in 1971. On the basis of phenotypic tests, it was classified as a saprophytic species, although its resistance to the leptospiracidal activity of normal mammal sera and some other features made it closer to the pathogenic leptospires. Thus, *Leptospira* sp. strain Bairam-Ali was used in a diagnostic system on the basis of macroagglutination reaction. This diagnostic system was different from the original genus-specific microagglutination test, which requires multiple serovars of live *Leptospira* and involves a risk of infection. Strain Bairam-Ali is a natural substitute for the microcapsule of synthetic polymer [7] as carrier of antigens similar to pathogenic strains, and it is safe for humans.

Only whole genome sequence could help resolve the mystery of the strain Bairam-Ali and clarify its phylogenetic position in the *Leptospira* genus and relationships with the pathogenic *Leptospira* species.

2. Materials and Methods

2.1. Bacterial Strains. *Leptospira* strains were cultured by the Russian MoH Centre for Leptospirosis laboratory at the N.F. Gamaleya Institute for Epidemiology and Microbiology, Moscow, according to WHO guidance [6]. Fifty-eight strains, including 29 reference strains and 29 isolates from various sources and geographical regions, were analyzed. Twenty-six reference strains were members of seven pathogenic species. By the start of this study, seven reference strains were absent from the *Leptospira* MLST database [8].

2.2. Phenotypic and Serological Characterization of Strain Bairam-Ali. Methods for differentiation of pathogenic versus saprophytic strains and for cross-agglutination-absorption reactions were performed according to WHO guidance [6].

2.3. Scanning Electron Microscopy. Samples were prepared as described in detail [9] and analyzed with dual-beam focused ion beam/scanning electron microscope, Quanta 200 3D (FEI Company, USA), in both high and low vacuum, mostly at 5 kV electron beam acceleration [10].

2.4. DNA Isolation. DNA for PCR analysis was extracted from the bacterial cultures as described previously [11]. Preparation of genomic DNA for the whole genome sequencing was performed according to [12].

2.5. Leptospira Species Identification. The species of isolates were identified by amplification and sequencing of the *rpoB* gene (coding β subunit of bacterial RNA polymerase), according to La Scola et al. [13]. Sequence data for *rpoB* has been deposited in GenBank, with accession numbers KJ701730–KJ701749.

2.6. MLST. MLST for the strains of *L. interrogans* and *L. kirschneri* was performed by use of the original scheme of Thaipadungpanit et al. [14]. After publication of the modified MLST scheme [15], we completed earlier results by using *caiB* gene sequences. For the new isolates and strains of five other species, we used only the modified MLST scheme. Some modifications were inserted in the published protocol. The conditions of the amplification were modified for the *glmU*, *pntA*, and *sucA* genes by raising the melting temperature to 50°C. Also, the $MgCl_2$ concentration was changed to 3.5 mM for the *glmU*, *pntA*, and *sucA* genes. Reference collection strains were used for adaptation of the method to our laboratory and for control of the reproducibility of the results.

2.7. PCR Products Sequencing. PCR products were sequenced according to the protocol of BigDye Terminator 3.1 Cycle Sequencing kit for the Genetic Analyzer 3130 of Applied Biosystems/Hitachi.

2.8. Nucleotide Sequence Analysis. The alignment of *rpoB* and MLST gene sequences was made by use of ClustalW2 [16]. BLAST search was used for species identification; similarity of *rpoB* gene sequences was more than 98%.

Allele numbers for MLST genes were assigned with the help of website MLST Home. Allelic profiles, in the order glmU-pntA-sucA-tpiA-pfkB-mreA-caiB, were used to assign sequence types (STs) to strains. New alleles and ST were controlled and submitted by the curator of *Leptospira* spp. MLST database [8]. All new *Leptospira* strains were submitted in the *Leptospira* spp. MLST database under the identification numbers indicated in Table 1.

Relatedness between STs on the base of allelic profiles was analyzed by use of BURST version 1.00 [17, 18].

2.9. Nucleotide Sequence Polymorphism. BLAST search was used for retrieving homologues *rrs* (16S rRNA-coding), *rpoB*, and MLST gene sequences from GenBank (http://www.ncbi.nlm.nih.gov/genome/browse/). For comparative sequence analysis and phylogenetic reconstruction, nucleotide sequences of 92 additional *Leptospira* strains were retrieved. Seventy-five nucleotide sequences represented either complete or partial cds of *rrs* gene and *rpoB* gene, and 15 sequences represented whole genome sequencing data and genome drafts (Table 1). *Turneriella parva* have been included in the analysis as an out-group taxon from *Leptospiraceae*, for which whole genome sequence data are available (Table 1). Sequences of seven concatenated MLST loci for 201 ST available at the time of analysis were retrieved [8]. The alignments of *rrs*, *rpoB*, and MLST gene sequences and analysis nucleotide similarity (in %) were performed by use of ClustalW2 [16].

Table 1: Strains used in genotyping and phylogenetic analysis.

Number	Strain	Species	ST	Our submission ID or GenBank accession number for MLST genes	rpoB GenBank accession number	16S rDNA GenBank accession number
1	*GIMC2029:Li130	*L. alexanderi*, serogroup Manhao, serovar lichuan	207	337	KJ701742	
2	*GIMC2051:Veldrat Bataviae 46	*L. borgpetersenii*, serogroup Javanica, serovar javanica	143	In MLST Home data base	KJ701733	
3	*GIMC2052:Sari	*L. borgpetersenii*, serogroup Mini, serovar mini	142	In MLST Home data base	KJ701736	
4	GIMC2002:101PJ	*L. borgpetersenii*, serogroup Javanica, serovar poi	146	339		
5	GIMC2003:29PJ	*L. borgpetersenii*, serogroup Javanica, serovar poi	146	340	KJ701734	
6	GIMC2004:Yaroslavl 7	*L. borgpetersenii*, serogroup Javanica, serovar poi	146	341		
7	GIMC2005:1217PJ	*L. borgpetersenii*, serogroup Javanica, serovar javanica	146	342		
8	GIMC2006:1622PJ	*L. borgpetersenii*, serogroup Javanica	146	343		
9	GIMC2007:5-I	*L. borgpetersenii*, serogroup Javanica	146	344		
10	*GIMC2049:Castellon 3	*L. borgpetersenii*, *serogroup Ballum, serovar castellonis*	149	In MLST Home data base	KJ701737	
11	*GIMC2050:Perepelitsin	*L. borgpetersenii*, serogroup Tarassovi, serovar tarassovi	153	In MLST Home data base	KJ701735	
12	*GIMC2008:Mus 24	*L. borgpetersenii*, serogroup Sejroe, serovar saxkoebing	155	345		
13	*GIMC2048:Naam	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar naam	23	In MLST Home data base	KJ701731	
14	GIMC2014:AV4	*L. interrogans*, serovar no data	23	351		
15	*GIMC2047:Djasiman	*L. interrogans*, serogroup Djasiman, serovar djasiman	11	In MLST Home data base		
16	*GIMC2046:RGA	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar icterohaemorrhagiae	17	In MLST Home data base		
17	GIMC2010:Sv-19	*L. interrogans*, serogroup Icterohaemorrhagiae	17	347		
18	GIMC2009:Rn-2010	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	17	346		
19	GIMC2011:Rn-493	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	17	348		

Table 1: Continued.

Number	Strain	Species	ST	Our submission ID or GenBank accession number for MLST genes	rpoB GenBank accession number	16S rDNA GenBank accession number
20	GIMC2012:Rn-16	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	17	349		
21	GIMC2013:Rn-77	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	17	350		
22	*GIMC2042:Ezh-1 = Jez Bratislava	*L. interrogans*, serogroup Australis, serovar bratislava	24	In MLST Home data base		
23	*GIMC2043:Akiyami A	*L. interrogans*, serogroup Autumnalis, serovar autumnalis	27	In MLST Home data base		
24	*GIMC2044:Zanoni	*L. interrogans*, serogroup Pyrogenes, serovar zanoni	31	In MLST Home data base	KJ701732	
25	*GIMC2045:Hebdomadis	*L. interrogans*, serogroup Hebdomadis, serovar hebdomadis	36	In MLST Home data base		
26	GIMC2015:Kashirsky	*L. interrogans*, serogroup Canicola, serovar canicola	37	352		
27	GIMC2016:Mitronov	*L. interrogans*, serogroup Canicola, serovar canicola	37	353		
28	GIMC2017:Bugay	*L. interrogans*, serogroup Canicola, serovar canicola	37	354		
29	GIMC2018:Sobaka 2000	*L. interrogans*, serogroup Canicola, serovar canicola	37	355		
30	GIMC2019:Udalov	*L. interrogans*, serogroup Canicola, serovar canicola	37	356		
31	*GIMC2038:Ballico	*L. interrogans*, serogroup Australis, serovar australis	51	In MLST Home data base		
32	*GIMC2039:3705	*L. interrogans*, serogroup Sejroe, serovar woffi	58	In MLST Home data base		
33	*GIMC2040:Szwajizak	*L. interrogans*, serogroup Mini	73	In MLST Home data base		
34	*GIMC2041:Pomona	*L. interrogans*, serogroup Pomona, serovar pomona	140	In MLST Home data base		
35	*GIMC2020:M-20R	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	199	357	KJ701730	
36	GIMC2021:Abduloev	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	199	358		
37	GIMC2022:CL-11	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	199	359		
38	GIMC2030:CL-17	*L. interrogans*, serogroup Icterohaemorrhagiae, serovar copenhageni	206	336		

Table 1: Continued.

Number	Strain	Species	ST	Our submission ID or GenBank accession number for MLST genes	rpoB GenBank accession number	16S rDNA GenBank accession number
39	*GIMC2023:Vleermuis 3868	*L. kirschneri*, serogroup Cynopteri, serovar cynopteri	70	360		
40	*GIMC2024:Moskva V	*L. kirschneri*, serogroup Grippotyphosa, serovar grippotyphosa	110	361		
41	GIMC2025:181PG	*L. kirschneri*, serogroup Grippotyphosa, serovar grippotyphosa	110	362	KJ701738	
42	GIMC2026:617PG	*L. kirschneri*, serogroup Grippotyphosa, serovar grippotyphosa	110	363		
43	GIMC2027:859PG	*L. kirschneri*, serogroup Grippotyphosa, serovar grippotyphosa	110	364		
44	GIMC2028:1106PG	*L. kirschneri*, serogroup Grippotyphosa, serovar grippotyphosa	110	365		
45	*GIMC2037:5621	*L. kirschneri*, serogroup Pomona, serovar mozdok	117	In MLST Home data base		
46	*GIMC2031:HS 26R	*L. kirschneri*, serogroup Bataviae, serovar djatsi	204	334	KJ701739	
47	*GIMC2033:LSU 1945	*L. noguchii*, serogroup Louisiana, serovar panama	169	In MLST Home data base	KJ701740	
48	*GIMC2034:CZ 214 K	*L. noguchii*, serogroup Panama, serovar panama	171	In MLST Home data base	KJ701741	
49	*GIMC2035:Celledoni	*L. weilii*, serogroup Celledoni, serovar celledoni	185	In MLST Home data base	KJ701746	
50	*GIMC2036:Sarmin	*L. weilii*, serogroup Sarmin, serovar sarmin	191	In MLST Home data base		
51	*GIMC2032:CZ299U	*L. santarosai*, serogroup Pomona, serovar tropica	208	338		
52	GIMC2001:Bairam-Ali	*L.* sp., serogroup Bairam-Ali, serovar bairam-ali		KJ676852-KJ676858	KJ701604	KJ701750
53	*GIMC2055:Lyme	*L. inadai*, serogroup Lyme, serovar lyme			KJ701743	
54	GIMC2056:EMJH 86	*L. inadai*, serovar lyme			KJ701744	
55	GIMC2057:Enr 88	*L. inadai*, serogroup Lyme-Detroit, serovar lyme-detroit			KJ701745	
56	*GIMC2060:Sao Paulo	*L. yanagawae*, serogroup Semaranga, serovar sao paulo			KJ701747	
57	GIMC2058:LT-8	*L. biflexa*, serovar patoc			KJ701748	
58	GIMC2059:GR	*L. biflexa*, serovar andamana			KJ701749	
59	DSM 21527	*Turneriella parva*		NC_018020	CP002959	CP002959
60	80–412	*L. alstoni*, serovar pingchang			NZ_AOHD02000041.1	NZ_AOHD02000066.1
61	79601	*L. alstoni*, serovar sichuan				AY631881

Table 1: Continued.

Number	Strain	Species	ST	Our submission ID or GenBank accession number for MLST genes	rpoB GenBank accession number	16S rDNA GenBank accession number
62	L 60	*L. alexanderi*, serovar manhao 3			NZ_AHMT02000060.1	
63	A23	*L. alexanderi*, serovar manzhuang				AY996803.1
64	A85	*L. alexanderi*, serovar mengla				DQ991481.1
65	M 6901	*L. alexanderi*, serovar nanding				AY996804
66	Mus 127	*L. borgpetersenii*, serovar ballum			EU747302	
67	Lely 607	*L. borgpetersenii*, serogroup Sejroe, serovar hardjo				FJ154586
68	L550	*L. borgpetersenii*, serovar hardjo-bovis			CP000348.1	
69	JB197	*L. borgpetersenii*, serovar hardjo-bovis			CP000350	
70	Lely 607	*L. borgpetersenii*, serovar hardjo-bovis			EU747305	
71	Veldrat Batavia 46	*L. borgpetersenii*, serovar javanica				AY887899.1
72	M84	*L. borgpetersenii*, serovar sejroe			EU747311	
73	Perepelitsin	*L. borgpetersenii*, serovar tarassovi			EU747307	JQ988861.1
74	Whitticombi	*L. borgpetersenii*, serovar whitticombi			EU747314	
75	Mini-CTG	*L. borgpetersenii*, serovar no data				JQ765635.1
76	Ballico	*L. interrogans*, serovar australis				FJ154556.1
77	Akiyami A	*L. interrogans*, serovar autumnalis				AM050580.1
78	Jez-bratislava	*L. interrogans*, serovar bratislava			EU747300	
79	Mallika	*L. interrogans*, serovar bulgarica				AY996792.1
80	Hond Utrecht IV	*L. interrogans*, serovar canicola				FJ154561.1
81	Fiocruz L1-130	*L. interrogans*, serovar copenhageni	17	AE016823.1	AE016823.1	
82	Djasiman	*L. interrogans*, serovar djasiman			EU747312	FJ154550.1
83	Hardjo_DB33	*L. interrogans*, serovar hardjo				JQ988854.1
84	RGA	*L. interrogans*, serovar icterohaemorrhagiae				NR_029361.1
85	Kremastos	*L. interrogans*, serovar kremastos				AY461868.1
86	IPAV	*L. interrogans*, serovar lai			CP001221	
87	LT 398	*L. interrogans*, serovar manilae				FJ154545.1
88	Hond HC	*L. interrogans*, serovar medanesis				DQ991471.1
89	Pomona	*L. interrogans*, serovar pomona			EU747306	AY996800.1
90	Salinem	*L. interrogans*, serovar pyrogenes				FJ154552.1
91	Vleermuis	*L. interrogans*, serovar schueffneri			EU747313	
92	Szwajizak	*L. interrogans*, serovar szwajizak				DQ991466.1
93	3705	*L. interrogans*, serovar wolffi			EU747308	
94	Zanoni	*L. interrogans*, serovar zanoni				DQ991473
95	Bataviae_DB59	*L. interrogans*, serovar no data				JQ988841.1

Table 1: Continued.

Number	Strain	Species	ST	Our submission ID or GenBank accession number for MLST genes	rpoB GenBank accession number	16S rDNA GenBank accession number
96	Agogo	*L. kirschneri*, serovar agogo				DQ991476.1
97	Bafani	*L. kirschneri*, serovar bafani				DQ991477.1
98	1051	*L. kirschneri*, serovar bim				AY996802.1
99	Butembo	*L. kirschneri*, serovar butembo				Q991478.1
100	3522 C	*L. kirschneri*, serovar cynopteri				FJ154546.1
101	Moskva V	*L. kirschneri*, serovar grippotyphosa			EU747301	
102	Kambale	*L. kirschneri*, serovar kambale				AY461878.1
103	5621	*L. kirschneri*, serovar pomona				FJ154559.1
104	Wumalasena	*L. kirschneri*, serovar ratnapura				DQ991479.1
105	Bejo-Iso9	*L. kmetyi*, serovar malaysia		AHMP02000003	AHMP02000003	NR_041544.1
106	Hook	*L. noguchii*, serogroup Australis			EU349504	
107	Bonito	*L. noguchii*, serogroup Autumnalis			EU349501	
108	Caco	*L. noguchii*, serogroup Autumnalis			EU349498	
109	Cascata	*L. noguchii*, serogroup Bataviae			EU349502	
110	LSU 1945	*L. noguchii*, serogroup Louisiana			EU349505	
111	1348U	*L. noguchii*, serovar claytoni				DQ991498.1
112	1996K	*L. noguchii*, serovar cristobali				DQ991497.1
113	M7	*L. noguchii*, serovar huallaga				DQ991499.1
114	1011	*L. noguchii*, serovar nicaragua			EU349499	
115	LSU 2580	*L. noguchii*, serovar orleans			EU349500	
116	CZ 214 K	*L. noguchii*, serovar panama			EU349497	NR_043050.1
117	DB57	*L. noguchii*, serovar panama			EU349501	JQ988837.1
118	HS 616	*L. santarosai*, serovar alexi				FJ154585.1
119	Alice	*L. santarosai*, serovar alice				DQ991493.1
120	MAVJ 401	*L. santarosai*, serovar arenal			AHMU02000055	
121	LT 79	*L. santarosai*, serovar bakeri				FJ154589.1
122	LT 117	*L. santarosai*, serovar georgia				AY996805.1
123	CZ320	*L. santarosai*, serovar kobbe				DQ991495.1

Table 1: Continued.

Number	Strain	Species	ST	Our submission ID or GenBank accession number for MLST genes	rpoB GenBank accession number	16S rDNA GenBank accession number
124	1342	*L. santarosai*, serovar shermani				FJ154576.1
125	CZ390	*L. santarosai*, serovar weaveri				DQ991496.1
126	CBC613	*L. santarosai*, serovar no data			ANIH01000040	
127	ST188	*L. santarosai*, serovar no data			AOHA02000081	
128	MOR084	*L. santarosai*, serovar no data			AHON02000051	
129	2000030832	*L. santarosai*, serovar no data			AFJN02000030	
130	Celledoni	*L. weilii*, serovar celledoni				DQ991486.1
131	H27	*L. weilii*, serovar hekou				DQ991487.1
132	M39090	*L. weilii*, serovar langati				DQ991488.1
133	WB46	*L. weilii*, serovar sarmin				U12673.1
134	LT 89-68	*L. weilii*, serovar vughia				FJ154590.1
135	5399	*L. broomii*, serovar hurstbridge		AHMO02000008	AHMO02000008	AHMO02000008.1
136	BUT 6	*L. fainei*, serovar hurstbridge		AKWZ02000010	AKWZ02000010	NR_043049.1
137	BKID 6	*L. fainei*, serovar hurstbridge				AY996789.1
138	10	*L. inadai*, serovar lyme		AHMM02000025	AHMM02000015	AHMM02000015.1
139	VAR 010	*L. licerasiae*, serovar varillal		AHOO02000005	AHOO02000005	NR_044310.1
140	Khorat-H2	*L. wolffii*, serovar khorat		AKWX02000020		NZ_AKWX02000004.1
141	Patoc Ames	*L. biflexa*, serovar patoc		NC_010842	NC_010842	NC_010842
142	Patoc Paris	*L. biflexa*, serovar patoc		NC_010602	NC_010602	
143	CH 11	*L. biflexa*, serovar andamana				FJ154577.1
144	Veldrat Semarang 173	*L. meyeri*, serovar semaranga				AF157089.1
145	Went 5	*L. meyeri*, serovar hardjo		AKXE01000002	AKXE01000001	NZ_AKXE01000007.1
146	LT 11-33 ATCC 700639	*L. terpstrae*, serovar hualin		AOGW02000006	AOGW02000010	NZ_AOGW02000008.1
147	Waz Holland ATCC 700522	*L. vanthielii*, serovar holland		AOGY02000070	AOGY02000051	NZ_AOGY02000072.1
148	CDC; ATCC 43284	*L. wolbachii*, serovar codice		AOGZ02000014	AOGZ02000008	NR_043046.1
149	Sao Paulo ATCC 700523	*L. yanagawae*, serovar saopaulo		AOGX02000015	AOGX02000024	Z_AOGX02000022.1
150	Eri-1(T) DSM 26084(T) = JCM 18486(T)	*L. idonii*, serogroup Hebdomadis				AB721966

*The reference strains from GIMC.

2.10. Phylogenetic Analysis. Phylogenetic analysis of *Leptospira* species was performed based on the *rrs* gene fragment, *rpoB* gene fragment, and seven concatenated sequences of MLST loci. Phylogenetic trees were constructed by use of neighbor-joining, maximum likelihood, and maximum parsimony methods.

Genetic distances between *Leptospira* genotypes were evaluated by use of Kimura's two-parameter model [19], which was chosen as an optimal evolution distance model derived from model test based on the Akaike information criterion [20]. The evolutionary history was inferred by using the Maximum Likelihood method based on the general time reversible model [21]. Initial tree(s) for the heuristic search were obtained automatically by applying neighbor-joining and BioNJ algorithms to a matrix of pairwise distances estimated by use of the maximum composite likelihood approach and then selecting the topology with superior log likelihood value. A discrete gamma distribution was used to model evolutionary rate differences among sites (six categories (+G, parameter = 0.4818)). Maximum parsimony trees were constructed with an algorithm implemented in MEGA version 6.0 [22]. Bootstrap analyses were performed with 1,000 replicates.

2.11. Whole Genome Sequencing. Whole genome sequencing of *Leptospira* spp. strain GIMC2001:Bairam-Ali was performed according to the manufacturer's guidelines (Roche) for the next generation sequencing (NGS). Two protocols were used for a shotgun-sequencing library preparation: rapid library and pair-end library. The rapid library was made according to the Rapid Library Preparation Method Manual (Roche). The pair-end library was performed according to the 3 kb protocol provided by the manufacturer to aid in scaffold building. The paired-end library insert size was from 1347 to 4364 bp, with an average of 2695 bp.

Assembly was performed with 454 Sequencing System Software v.2.7 (Roche), yielding 14 scaffolds, with the largest size being 3 342 467 bp. Gap closure was performed by use of Contig Graph result file generated by GS De Novo Assembler program (Roche). For the oriC region search, Ori-Finder program was used [23].

2.12. Gene Annotation. The software Rapid Annotations Subsystems Technology and SEED [24, 25] were used for annotating the genome of *Leptospira* spp. strain GIMC2001:Bairam-Ali.

3. Results and Discussion

A sample of the collection (GIMC) of the Russian MoH Centre for Leptospirosis was used for the verification the species and strains based on the currently recommended molecular-genetic methods. The sample included 29 reference strains and 29 isolates of *Leptospira* from various geographical regions (Table 1, *—marker of the reference strains from GIMC).

The analysis of *rpoB* gene sequences demonstrated that 29 reference strains belonged to seven pathogenic species, one nonpathogenic species (*L. yanagawae*), and one intermediate species (*L. inadai*). Among 29 isolates of *Leptospira* twenty-four field isolates were related to the three pathogenic species (*L. interrogans*, *L. borgpetersenii,* and *L. kirschneri*). Four cultures were isolated from the commercially available Ellinghausen-McCullough-Johnson-Harris (EMJH) medium: two isolates from this group belonged to the nonpathogenic species *L. biflexa* and the other two isolates belonged to the intermediate species *L. inadai*. Thus, the prevailing species among the isolates collected on the territory of Russia comprised *L. interrogans*, *L. borgpetersenii,* and *L. kirschneri*.

The modified MLST scheme was applied to the isolates and reference strains representing 7 pathogenic *Leptospira* species. Twenty-four isolates of pathogenic species belonged to 8 different STs (Table 1). Five STs, common for Russian strains of *Leptospira* (*L. interrogans* ST37, 17, and 199; *L. kirschneri* ST110; and *L. borgpetersenii* ST146), were identified as having a longtime and ubiquitous distribution in various geographic areas.

Among the strains of *Leptospira* species available in GIMC, one strain seemed to be mysterious and not closely related to any *Leptospira*, either pathogenic or saprophytic, or intermediate. It was named *Leptospira* spp. strain Bairam-Ali, because it was isolated in 1971 in Turkmenia from the water of a drainage canal.

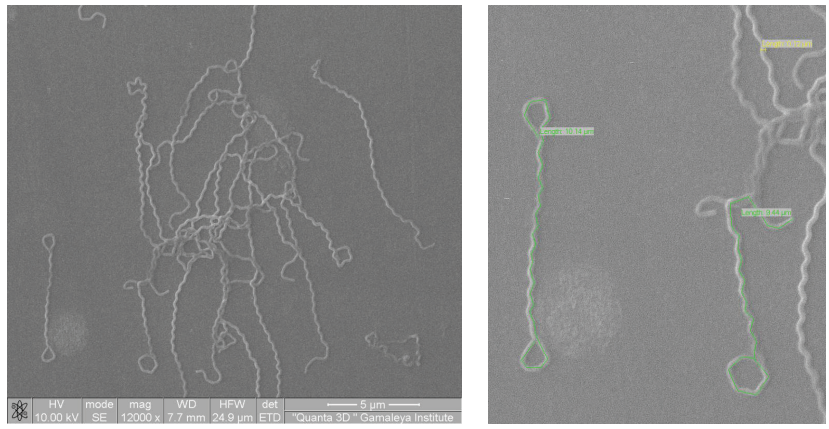
3.1. Morphology of the Leptospira spp. Strain Bairam-Ali. The morphology of the mysterious strain Bairam-Ali is typical of that of the *Leptospira* genus (Figure 1). Electron microscopy demonstrated that its cells are corkscrew-shaped with end hooks. They are thin and helical, like the cells of all known leptospires. Also, the cells have a diameter (d) of 0.12 μm and length (l) from 9.44 to 10.14 μm, like that of known leptospires (d = 0.15–0.3 μm and l = 6,00–20,00 μm) [26].

3.2. Phenotypic Characterization of the Leptospira spp. Strain Bairam-Ali. Physiological characteristics of Bairam-Ali, demonstrated in Table 2, suggest that strain Bairam-Ali can be classified as saprophytic *Leptospira*.

For pathogenicity experiment, six four-week-old male golden Syrian hamsters were inoculated subcutaneously with 10^8 cells of strain Bairam-Ali in 1 mL of PBS. No hamster died from infection even at such a high bacterial dose. *Leptospira* cells were not detected during following bacterioscopic and bacteriological examination of the hamsters' viscera. This is an additional evidence of saprophytic quality of Bairam-Ali.

On the other hand, Bairam-Ali was resistant to the bactericidal (leptospiracidal) activity of the normal serum of human and of some other mammal animals. The most of pathogenic *Leptospira* species were resistant too, whereas the saprophytic species *L. biflexa* was sensitive.

3.3. Serological Characterization of the Leptospira spp. Strain Bairam-Ali. The strain Bairam-Ali had no antigenic affinity with 18 different serovars represented by reference strains. So the conclusion about the original serotype of Bairam-Ali was



(a) (b)

Figure 1: Dual-beam scanning electron microscopy image of *L.* species Bairam-Ali cells. (a) Original image; (b) with measuring the object in the program "Scandium" (green, length 10,14 μm; yellow, length (diameter) 0,12 μm; and green, length 9,44 μm).

Table 2: Physiological characteristics of the *Leptospira* spp. strain Bairam-Ali.

Strain	Growth at temp (°C) of			Growth in the presence of 8-Azaguanine 225 pg/mL	Lipase activity	Hemolytic activity against sheep erythrocytes
	11	30	37			
Bairam-Ali	+	+	+	+	+	−

made. The serovar of Bairam-Ali was named bairam-ali. It formed the separated serogroup Bairam-Ali.

3.4. The Whole Genome Sequence Analysis. The whole genome sequence could help to resolve the mystery of the strain Bairam-Ali and clarify its phylogenetic position in the *Leptospira* genus.

According to the NGS data the whole Bairam-Ali genome is 4,4 Mb, with GC % 34,74 and two chromosomes. The first chromosome is 4 059 463 bp and the second is 325 649 bp. Plasmid, typical of nonpathogenic *Leptospira* species, was absent from the Bairam-Ali genome. On the base of genome size and GC content, demonstrated in Table 3, Bairam-Ali is more similar to the pathogenic species *L. interrogans* and *L. noguchii*.

Based on *rpoB* gene sequence, we established that Bairam-Ali is more similar to pathogenic strain *L. interrogans* Fiocruz-L1-130 (but the level of similarity is only 70.00%). However, by sequence of *rDNA* genes it is similar to nonpathogenic strains: for gene *rrs*, coding 16S rRNA, *L. meyeri* was most closely related (92.74%); for gene *rrl,* coding 23S rDNA, *L. biflexa* was more similar (95.00% fragment 1 and 93.00% fragment 2). Sequences of *rDNA* genes of Bairam-Ali were deposited in GenBank, accession numbers KJ701750, KJ701751, and KJ701752.

These data suggest dual phenotypic characteristics of the strain Bairam-Ali. In spite of having differences from both pathogenic and nonpathogenic species, many of the functional categories that are involved in essential housekeeping functions are represented in its core gene [27, 28]. Thus, basic groups of proteins involved in cell metabolism, survival, environmental adaptability, and potential pathogenic factors were present. Enzyme complexes participating in implementation of genetic information, in particular in DNA replication, were identified in strain Bairam-Ali: chromosomal replication initiator protein DNA, single-stranded DNA-binding protein; all subunits of DNA polymerase III, DNA polymerase I, and DNA polymerase IV; DNA gyrase; ligase; and helicase [29]. Also, the large groups of enzymes that take part in DNA reparation [29], for example, excinuclease ABC subunits A–C and exodeoxyribonuclease (III, V, and VII), and proteins of postreplicative mismatch repair system (Mutator S and Mutator L) were present. Archaeal DNA polymerase I gene was detected in the Bairam-Ali genome; this gene is a member of Family B and bacterial DNA polymerase II. The same gene was present in nonpathogenic strain *L. biflexa* Patoc genome but not in pathogenic *L. interrogans* Fiocruz-L1-130 and *L. borgpetersenii* Hardjo-bovis-L550 genomes [27, 30]. The sequences of DNA metabolome have been deposited in GenBank, with accession numbers KJ701710–KJ701729.

Table 3: Comparison L. spp. Bairam-Ali strain genome characteristics with most closely related *Leptospira* species.

Leptospira strain or species	Genome size, Mbp	GC%
L-BA*	4,4	34,7
L. interrogans	4,56 ± 0,32	35,22 ± 0,27
L. noguchii	4,76 ± 0,16	35,63 ± 0,27
L. terpstrae	4,09	38,2
L. meyeri	4,15	38,05
L. vanthielii	4,23	38,9
L. biflexa	3,95	38,9

*L-BA: *Leptospira* spp. Bairam-Ali.

According to Bourret et al. [31], enteric bacteria usually have about 50 genes coding for structural and functional proteins involved in motility. Motility is a distinctive feature of *Leptospira*; forty-seven different proteins provide bacterial locomotion [32]. In the Bairam-Ali genome, the genes for proteins of the basic parts of periplasmic flagella were detected: basal-body (Flg B–D, FlgF, and FlgG), hook (FlgE, FlgK, FlgL, FliE, FliD, and fliK), and four copies of FlaA (sheath protein of filament). The genes for proteins of flagellar rings L (FlgH), MS (FliF), and P (FlgI); biosynthesis protein (FlhA, FlhB (2), FlhF, FliL, and Fli Q-S); motors (MotB (3), MotA (2), FliG (3), FliM, and FliN); flagellar assembly factor (FliW and FliH), cell division (BolA (2), FtsA, FtsH (2), FtsI, FtsK, FtsW, and FtsZ), and gliding motility (GldF and GldG) proteins also were registered. Strain Bairam-Ali was found to have more than ten genes for the regulation of flagellum genes transcription and for signal transduction to flagella motor. Sequences of the flagellum genes have been deposited in GenBank, accession numbers KJ701653–KJ701709.

It should be noted that the important structural element, that is, numerous groups of predicted lipoproteins (Lip), which may be either surface-exposed or located in the periplasm, is present in both saprophytic and pathogenic *Leptospira* species. In the strain Bairam-Ali genome, as in the genome of nonpathogenic species *L. biflexa* [27], genes or orthologs of the major lipoproteins LipL32 and LipL41, which are important immunodominant antigens in pathogenic *Leptospira* species [33, 34], were not identified. Nevertheless, in the genome of Bairam-Ali, we detected six genes of LipL45 (GenBank accession numbers KJ701647–KJ701652), which are processed to a peripheral membrane associated with the outer membrane complex. According to Matsunaga et al. [35] expression of P31, derived from the carboxy terminus of LipL45, is upregulated in stationary phase cultures; thus it may have a membrane-stabilizing function. Also, in a hamster model infected with pathogenic *L. kirschneri*, antibodies to LipL45 [35] were produced, suggesting its involvement in pathogenesis.

Further analysis of the genome of Bairam-Ali revealed the presence of genes for more than 40 different classes of proteins, ensuring its natural resistance to a wide range of antibiotics (β-lactams, tetracyclines, glycopeptides (vancomycin), and polymyxin); efflux system and resistance to heavy metals (Czc A–C and arsenical-resistance proteins); and possible abortive infection phage-resistance protein. The broad resistance characteristics of strain Bairam-Ali are consistent with it being a natural reservoir for storage and possible transmission of these properties to other free-living *Leptospira* species. Sequences of the resistome genes and their proteins have been deposited in GenBank, accession numbers KJ701605–KJ701646.

3.5. Sequence Polymorphism. To determine the place of the original strain Bairam-Ali in the *Leptospira* genus, we undertook phylogenetic analysis of the *Leptospira* groups with different pathogenicity.

Based on performed alignments, the percent nucleotide similarities of *rrs, rpoB,* and MLST gene sequences of analyzed *Leptospira* genotypes were determined; see Supplementary Material available online at http://dx.doi.org/10.1155/2014/649034 (S1, S2, and S3). According to the obtained data, aligned sequences of 16S rRNA-coding *rrs* gene were 1305 bp, aligned sequences of gene *rpoB* were 493 bp, and aligned sequences of seven MLST tags (*glmU-pntA-sucA-tpiA-pfkB-mreA-caiB*) were 3105 bp.

Among the investigated *Leptospira* sequences, gene *rpoB* showed the highest variability, reaching 40.61%. 16S rRNA-coding gene was more conservative, and general sequence variability of the *rrs* gene across all strains and species reached 11.52%. Sequence variability of seven MLST tags reached 34.88%.

Detectable intraspecific sequence polymorphism for studied loci also was different. Intraspecific differences detected by *rrs* were less than 1%. Intraspecific differences detected by most polymorphic *rpoB* genes ranged from 0.61% (*L. kirschneri*) to 11.63% (*L. borgpetersenii*), with, in some cases, overlapping of intraspecific and interspecific values (S1, S2, and S3).

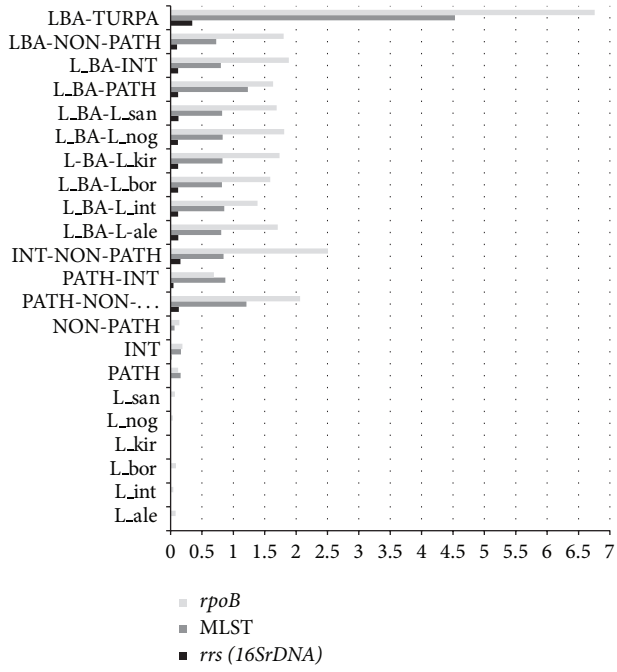
Notably, the intraspecific differences detected by seven MLST tags resolved all of the *Leptospira* genotypes and matched 3% accepted threshold values for prokaryote species divergence [36, 37]. The exception was *L. weilii*; interspecies differences of *L. weilii* strains revealed by seven MLST tags reached 5.77% (S1, S2, and S3).

Despite the differences in resolution ability of studied loci, all of them confirmed the unique nature of *Leptospira* spp. strain Bairam-Ali. *rrs*, *rpoB*, and MLST tags obtained for Bairam-Ali had low similarity with other sequences. Maximum similarity of Bairam-Ali MLST tags (69.7%) was detected for *L. interrogans* (Table 3). At the same time, related levels of sequence similarity also characterized nonpathogenicand pathogenic *Leptospira* species (67.29–68.95%) and intermediate and pathogenic *Leptospira* species (70.0–72.8%). Sequence similarity in pairs of Bairam-Ali, to intermediate *Leptospira* species, to nonpathogenic *Leptospira* species, and to pathogenic *Leptospira* species, was 65.8–67.31%, 67.6–69.18%, and 66.27–69.7%, respectively. The level of sequence similarity between Bairam-Ali and *Turneriella parva* was the lowest (55.3%), which is evidence of Bairam-Ali belonging to the genus *Leptospira*.

3.6. Leptospira Phylogeny and L. spp. Bairam-Ali Location in the Leptospira Genus. The most prominent phylogenetic information was obtained through concatenated sequences of seven MLST tags. All genotypes were divided into three species groups: pathogenic, nonpathogenic, and intermediate. *T. parva* formed a distinct basal out-group branch.

The genetic diversity of pathogenic *Leptospira* reached 0.159 (MLST data), 0.006 (*rrs* data), and 0.121 (*rpoB* data). The most variable species were *L. borgpetersenii* (0.086 *rpoB*) and *L. noguchii* (0.069). Despite the fact that *L. interrogans* was the most representative species in the analysis (120 MLST genotypes), intraspecies genetic diversity for *L. interrogans* was only 0.002 (Figure 2).

The genetic diversity of nonpathogenic *Leptospira* was similar to that of the pathogenic *Leptospira* species and



Species or group name	Genes		
	rrs	MLST	*rpoB*
LBA-TURPA	0.344	4.532	6.757
LBA-NON-PATH	0.104	0.727	1.800
L_BA-INT	0.121	0.803	1.886
L_BA-PATH	0.121	1.232	1.635
L_BA-L_san	0.125	0.821	1.690
L_BA-L_nog	0.117	0.830	1.811
L-BA-L_kir	0.121	0.827	1.737
L_BA-L_bor	0.121	0.819	1.589
L_BA-L_int	0.120	0.856	1.385
L_BA-L-ale	0.124	0.807	1.707
INT-NON-PATH	0.157	0.845	2.509
PATH-INT	0.047	0.873	0.691
PATH-NON-	0.133	1.208	2.063
NON-PATH	0.0072	0.0633	0.1395
INT	0.013	0.167	0.188
PATH	0.006	0.159	0.121
L_san	0.0004	0.0011	0.0684
L_nog	0	0.0069	0.0356
L_kir	0.0002	0.0023	0.0042
L_bor	0.0008	0.0014	0.0865
L_int	0	0.002	0.043
L_ale	0.002	$7,00E-04$	0.081

Figure 2: The number of base substitutions per site from averaging over all sequence pairs within each group is shown. Analyses were conducted using the Kimura 2-parameter model.

reached 0.063 (MLST data), 0.007 (*rrs* data), and 0.139 (*rpoB* data).

Interestingly, the intermediate group *Leptospira* had broader genetic diversity than did the nonpathogenic or pathogenic *Leptospira* (0.167, MLST data; 0.013, *rrs* data; and 0.188 *rpoB* data) (Figure 2).

In pathogenic group accessions of each species, excluding *L. weilii,* clearly distinct phylogenetic clusters, with high bootstrap supporting values, were present.In some data sets, significant subdivision of *L. weilii* genotypes, combined with a close relationship between *L. weilii* and *L. alexanderi,* was found. Possible polyphyletic nature of *L. weilii* has been described previously, particularly in the highly divergent *tpiA* locus [15]. According to our data, excluding from analysis *glmU* and *pfkB* loci, *L. weilii* and *L. alexanderi* genotypes fall into two different but closely related subclusters. This differentiation was found also with *rrs* gene phylogeny (S4).

Three species in the pathogenic group, *L. interrogans, L. kirschneri,* and *L. noguchii,* formed one close genetic subgroup (BI = 99%, bootstrap index). The strains of these species are often the cause of the human leptospirosis. Four species, *L. weilii, L. alexanderi, L. borgpetersenii, and L. santarosai,* formed another phylogenetic subgroup (BI = 98%) (Figure 3). The strains of these species are usually isolated in natural foci. The candidate pathogenic species, *L. alstoni* and *L. kmetyi,* fell in the common big cluster of pathogenic *Leptospira* species. On the MLST tree, *L. alstoni* formed a sister clade to the *L. weilii/L. alexanderi/L. borgpetersenii/L. santarosai* group of pathogenic *Leptospira* species (BI = 70%), whereas *L. kmetyi* formed a basal branch (BI = 99) (Figure 3). The levels of genetic similarity of *L. alstoni* and *L. kmetyi* to pathogenic *Leptospira* species (0.913 and 0.926) were much higher than to nonpathogenic (0.13 and 0.09) or intermediate species (0.58 and 0.54). Our data confirm the results of previous characterization for *L. alstoni* on the base of *16SrDNA* gene and for *L. kmetyi* on the base of *16SrDNA, gyrB*, and *rpoB* genes analysis [2, 3].

The single *Leptospira* spp. strain, Bairam-Ali, formed a separate distinct branch on dendrogram, with genetic distances comparable to those of distantly related species of different *Leptospira* groups. For example, differences between pathogenic and nonpathogenic, pathogenic and intermediate, and intermediate and nonpathogenic groups were 1.208, 0.873, and 0.845, respectively (Figure 2), whereas the differences between *Leptospira* spp. Bairam-Ali and pathogenic, intermediate, and nonpathogenic species were 1.232, 0.803, and 0.727, respectively; this result points to significant differences between Bairam-Ali and other *Leptospira* species, a result that is consistent with whole genome sequencing data.

It should be noted that the comparative studies of different *Leptospira* groups based on the *rrs* gene variability, which added to the analysis the newly described saprophytic *L. idonii* strain Eri-1T [5] and recently approved 16S rDNA sequences of leptospires from the Peruvian Amazon, previously termed "clade C" [38, 39], did not change separate phylogenetic position of *Leptospira* spp.strain Bairam-Ali.

Thus, according to our results the *Leptospira* strain Bairam-Ali forms a separate phylogenetic branch of the *Leptospira* genus and cannot be attributed to any group of

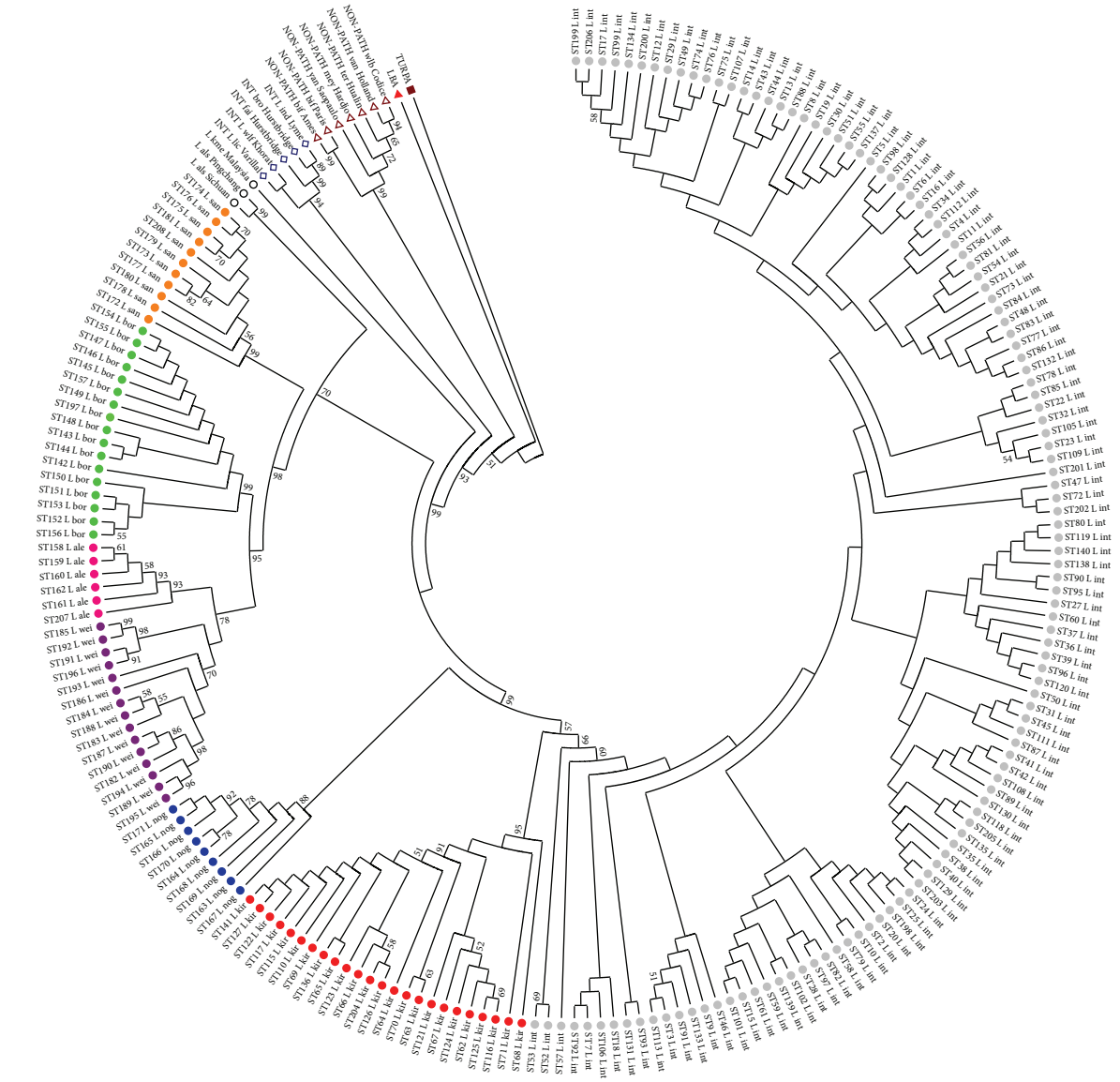


Figure 3: Phylogenetic tree of *Leptospira* species based on concatenated sequences of seven MLST loci.

pathogenic, intermediate, or saprophytic *Leptospira*, illustrating its uniqueness.

In conclusion, we found that the natural and anthropurgic foci supported ubiquitous pathogenic, intermediate, and nonpathogenic *Leptospira* species. They have circulated for a long time, interacting with maintenance and supplementary animal hosts. The relationships between the different *Leptospira* strains provide for horizontal gene transfer and create in the bacterial population the pool of genes important for adaptivity to various conditions. Modern molecular genetic methods are suitable for investigation of *Leptospira* and control of the pathogenic species, which are the causal agent of leptospirosis. Also, the taxonomic position of the new strains with unexpected properties can be founded on the base of these methods. Further investigation of the genetic diversity in the natural and anthropurgic foci is required for the identification of the *Leptospira* genotypes, control of the strain's transmission, and better understanding of the origin and evolution of the *Leptospira* species.

Abbreviations

L int: *L. interrogans*
L kir: *L. kirschneri*
L nog: *L. noguchii*
L wei: *L. weilii*
L ale: *L. alexanderi*
L bor: *L. borgpetersenii*
L san: *L. santarosai*
L als: *L. alstoni*
L kme: *L. kmetyi*
L lic: *L. licerasiae*

L wlf: *L. wolffii*
fai: *L. fainei*
bro: *L. broomii*
L ind: *L. inadai*
bif: *L. biflexa*
yan: *L. yanagawae*
mey: *L. meyeri*
ter: *L. terpstrae*
van: *L. vanthielii*
wlb: *L. wolbachii*
LBA: *L.* species Bairam-Ali
TURPA: *Turneriella parva*
INT: Intermediate species
NON-PATH: Nonpathogenic species.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] The List of Prokaryotic names with Standing in Nomenclature, Site founded by: J. P. Euzéby, http://www.bacterio.net/.

[2] L. Smythe, B. Adler, R. A. Hartskeerl, R. L. Galloway, C. Y. Turenne, and P. N. Levett, "Classification of *Leptospira* genomospecies 1, 3, 4 and 5 as *Leptospira alstonii* sp. nov., *Leptospira vanthielii* sp. nov., *Leptospira terpstrae* sp. nov. and *Leptospira yanagawae* sp. nov., respectively," *International Journal of Systematic and Evolutionary Microbiology*, vol. 63, no. 5, pp. 1859–1862, 2013.

[3] A. T. Slack, S. Khairani-Bejo, M. L. Symonds et al., "*Leptospira kmetyi* sp. nov., isolated from an environmental source in Malaysia," *International Journal of Systematic and Evolutionary Microbiology*, vol. 59, no. 4, pp. 705–708, 2009.

[4] G. M. Cerqueira and M. Picardeau, "A century of Leptospira strain typing," *Infection, Genetics and Evolution*, vol. 9, no. 5, pp. 760–768, 2009.

[5] M. Saito, S. Y. Villanueva, Y. Kawamura et al., "Leptospira idonii sp. nov., isolated from environmental water," *International Journal of Systematic and Evolutionary Microbiology*, vol. 63, part 7, pp. 2457–2462, 2013.

[6] *Human Leptospirosis: Guidance for Diagnosis, Surveillance and Control*, World Health Organization, Geneva, Switzerland, 2003, http://www.who.int/csr/don/en/WHO_CDS CDS_CSR_EPH_2002.23.pdf.

[7] Y. Arimitsu, E. Kmety, Y. Ananyina et al., "Evaluation of the one-point microcapsule agglutination test for diagnosis of leptospirosis," *Bulletin of the World Health Organization*, vol. 72, no. 3, pp. 395–399, 1994.

[8] Leptospira MLST database which is located at Imperial College London and is funded by the Wellcome Trust, http://leptospira.mlst.net/.

[9] L. V. Didenko, G. A. Avtandilov, N. V. Shevlyagina et al., "Biodestruction of polyurethane by Staphylococcus aureus (an investigation by SEM, TEM and FIB)," in *Current Microscopy Contributions to Advances in Science and Technology*, A. Méndez-Vilas, Ed., vol. 1, pp. 323–334, Formatex Research Center, Badajoz, Spain, 2012.

[10] M. Milani, D. Drobne, and F. Tatti, *Atlas of FIB/SEM in Soft Materials and Life Sciences*, Aracne, Rome, Italy, 2006.

[11] O. L. Voronina, M. Y. Chernukha, I. A. Shaginyan et al., "Characterization of genotypes for Burkholderia cepacia complex strains isolated from patients in hospitals of the Russian federation," *Molecular Genetics, Microbiology and Virology*, vol. 28, no. 2, pp. 64–73, 2013.

[12] K. Wilson, "Unit 2.4. Preparation of genomic DNA from bacteria," in *Current Protocols in Molecular Biology*, John Wiley and Sons, New York, NY, USA, 2001.

[13] B. La Scola, G. Baranton, A. Khamis, and D. Raoult, "Partial rpoB gene sequencing for identification of Leptospira species," *FEMS Microbiology Letters*, vol. 263, no. 2, pp. 142–147, 2006.

[14] J. Thaipadungpanit, V. Wuthiekanun, W. Chierakul et al., "A dominant clone of Leptospira interrogans associated with an outbreak of human leptospirosis in Thailand," *PLoS Neglected Tropical Diseases*, vol. 1, no. 1, article e56, 2007.

[15] S. Boonsilp, J. Thaipadungpanit, P. Amornchai et al., "A Single Multilocus Sequence Typing (MLST) scheme for seven pathogenic Leptospira species," *PLoS Neglected Tropical Diseases*, vol. 7, no. 1, Article ID e1954, 2013.

[16] The Main Site EMBL-EBI, European Bioinformatics Institute, http://www.ebi.ac.uk/Tools/msa/clustalw2.

[17] E. J. Feil, E. C. Holmes, D. E. Bessen et al., "Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 1, pp. 182–187, 2001.

[18] K. A. Jolley, E. J. Feil, M. S. Chan, and M. C. J. Maiden, "Sequence type analysis and recombinational tests (START)," *Bioinformatics*, vol. 17, no. 12, pp. 1230–1231, 2002.

[19] M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," *Journal of Molecular Evolution*, vol. 16, no. 2, pp. 111–120, 1980.

[20] D. Posada and K. A. Crandall, "MODELTEST: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[21] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.

[22] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[23] F. Gao and C.-T. Zhang, "Ori-finder: a web-based system for finding oriCs in unannotated bacterial genomes," *BMC Bioinformatics*, vol. 9, article 79, 2008.

[24] R. K. Aziz, D. Bartels, A. Best et al., "The RAST Server: rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, article 75, 2008.

[25] R. Overbeek, T. Begley, R. M. Butler et al., "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes," *Nucleic Acids Research*, vol. 33, pp. 5691–5702, 2005.

[26] M. Picardeau, "Diagnosis and epidemiology of leptospirosis," *Médecine et Maladies Infectieuses*, vol. 43, no. 1, pp. 1–9, 2013.

[27] M. Picardeau, D. M. Bulach, C. Bouchier et al., "Genome sequence of the saprophyte Leptospira biflexa provides insights into the evolution of *Leptospira* and the pathogenesis of leptospirosis," *PLoS ONE*, vol. 3, no. 2, Article ID e1607, 2008.

[28] S.-X. Ren, G. Fu, X.-G. Jiangk et al., "Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing," *Nature*, vol. 422, no. 6934, pp. 888–893, 2003.

[29] M. Singer and P. Berg, *Genes & Genomes: A Changing Perspective*, University Science Books, 1991.

[30] A. L. T. O. Nascimento, A. I. Ko, E. A. L. Martins et al., "Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis," *Journal of Bacteriology*, vol. 186, no. 7, pp. 2164–2172, 2004.

[31] R. B. Bourret, N. W. Charon, A. M. Stock, and A. H. West, "Bright lights, abundant operons—fluorescence and genomic technologies advance studies of bacterial locomotion and signal transduction: review of the BLAST meeting, Cuernavaca, Mexico, 14 to 19 January 2001," *Journal of Bacteriology*, vol. 184, no. 1, pp. 1–17, 2002.

[32] C. Li, C. W. Wolgemuth, M. Marko, D. G. Morgan, and N. W. Charon, "Genetic analysis of spirochete flagellin proteins and their involvement in motility, filament assembly, and flagellar morphology," *Journal of Bacteriology*, vol. 190, no. 16, pp. 5607–5615, 2008.

[33] A. L. T. O. Nascimento, S. Verjovski-Almeida, M. A. van Sluys et al., "Genome features of *Leptospira interrogans* serovar Copenhageni," *Brazilian Journal of Medical and Biological Research*, vol. 37, no. 4, pp. 459–478, 2004.

[34] P. A. Cullen, S. J. Cordwell, D. M. Bulach, D. A. Haake, and B. Adler, "Global analysis of outer membrane proteins from *Leptospira interrogans* serovar Lai," *Infection and Immunity*, vol. 70, no. 5, pp. 2311–2318, 2002.

[35] J. Matsunaga, T. A. Young, J. K. Barnett, D. Barnett, C. A. Bolin, and D. A. Haake, "Novel 45-kilodalton leptospiral protein that is processed to a 31-kilodalton growth-phase-regulated peripheral membrane protein," *Infection and Immunity*, vol. 70, no. 1, pp. 323–334, 2002.

[36] E. Stackebrandt and J. Ebers, "Taxonomic parameters revisited: tarnished gold standards," *Microbiology Today*, vol. 33, pp. 152–155, 2006.

[37] P. Vandamme and P. Dawyndt, "Classification and identification of the Burkholderia cepacia complex: past, present and future," *Systematic and Applied Microbiology*, vol. 34, no. 2, pp. 87–95, 2011.

[38] C. A. Ganoza, M. A. Matthias, D. Collins-Richards et al., "Determining risk for severe leptospirosis by molecular analysis of environmental surface waters for pathogenic *Leptospira*," *PLoS Medicine*, vol. 3, no. 8, article e308, 2006.

[39] J. S. Lehmann, M. A. Matthias, J. M. Vinetz, and D. E. Fouts, "Leptospiral pathogenomics," *Pathogens*, vol. 3, no. 2, pp. 280–308, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 648326, 9 pages
http://dx.doi.org/10.1155/2014/648326

Research Article

Phytoliths in Taxonomy of Phylogenetic Domains of Plants

Kirill S. Golokhvast,[1] Ivan V. Seryodkin,[2] Vladimir V. Chaika,[1] Alexander M. Zakharenko,[3] and Igor E. Pamirsky[4]

[1] *Scientific Educational Center of Nanotechnology, Far Eastern Federal University, 10 Pushkinskaya Street, Vladivostok 690990, Russia*
[2] *Laboratory of Ecology and Protection Animals, Pacific Institute of Geography FEB RAS, 7 Radio Street, Vladivostok 690041, Russia*
[3] *Laboratory of Enzyme Chemistry, Pacific Institute of Bioorganic Chemistry FEB RAS, 159 Prospect 100 Let Vladivostoku, Vladivostok 690022, Russia*
[4] *Laboratory of Molecular Biology, Blagoveshchensk State Pedagogical University, 104 Lenina Street, Blagoveshchensk 675000, Russia*

Correspondence should be addressed to Kirill S. Golokhvast; droopy@mail.ru and Igor E. Pamirsky; parimski@mail.ru

Received 17 April 2014; Accepted 3 July 2014; Published 27 August 2014

Academic Editor: Vassily Lyubetsky

We discuss, from the aspect of phylogeny, the interrelationships of the phytolith types in plants from the main taxonomical groups (algae, lichens, horsetails, gymnosperms, and floral plants) with homologues of known proteins of biomineralization. Phytolith morphotypes in various phylogenetic plant domains have different shapes. We found that, in ancient types of plants (algae, horsetails, and gymnosperms), there are fewer different phytolith morphotypes compared to more modern plants (floral plants). The phytolith morphotypes in primitive plants are generally larger than the morphotypes in more highly organized plants. We found that the irregular ruminate and irregular smooth morphotypes are the two most frequently encountered phytolith morphotypes in the tested plants (from algae to floral plants). These two morphotypes probably have a universal role. Silacidins, silicon transporters, silicateins, silaffins, and silicase homologues are often found in the major taxonomic groups of plants. Red algae had the smallest number of homologues of the biomineralization proteins (70–80), Monocotyledonous: 142, Coniferous: 166, Mosses: 227, and Dicotyledones: 336.

1. Introduction

Silicon is an extremely important element in all plants [1–5]. Phytoliths, which are specific biomineral silicon formations, are found in many groups of plants [6–8], and the roles and functions of phytoliths are varied [9, 10]. They provide defense from predators [11, 12], carry out optical functions in plants [13], and provide structural and frame elements, which contribute to stalk and leaf stiffening [10].

The role of phytoliths in the life of plants is extremely important, and the biomineralization process and phylogenetic history are evolutionarily fixed in plants.

Red algae and lichens are among the most ancient forms of plants [14–20].

Red algae appeared approximately 1 billion years ago [18, 20–22].

Lichens, according to paleontological findings, emerged in the early Devonian Period (approximately 400 million years ago) [23] or slightly earlier [24–27]. The first lichens might have been aqueous [28].

Horsetails arose in the top Devonian Period and developed from currently extinct rhyniophytes (Rhyniales) [18, 29, 30].

Gymnospermous plants appeared at the end of the Devonian Period, approximately 370 million years ago [29, 30].

Brown algae comprises a relatively young group of organisms and date approximately from 150 million years ago [31] to 200 to million years ago [32]. Evolutionarily, brown algae

comprise a unique group of live organisms because they developed from eukaryotes and have a multicellular body structure; they have survived to the present.

For studying the development and regularities of biomineral processes in a phylogeny of plants, it is necessary to study the most ancient forms (algae and lichens) and then track their metamorphoses into modern (floral) plants.

This work aims to analyze the morphogenesis of plant phytoliths from the aspect of phylogenetics.

2. Materials and Methods

2.1. Phytolith Analysis. The samples of land plants were collected near the Sikhote-Alin Biosphere Reserve (Primorsky region, Russia), and the mature algae originated in the Sea of Japan (Peter the Great Bay). The samples of phytoliths were prepared within 2 days after the collection of the plants. The phytoliths were collected according to the technique of Piperno [7]. The requisite part of the plant was heated for 4 hours and then washed with a 10% HCl solution and was followed by washing with concentrated nitric acid and distilled water. The sediment was centrifuged by an OPN-8 centrifuge ("Dastan," Bishkek, Kyrgyzstan) for 10 minutes at 1000 g and then selected for microscopic research using an AxioScope A1 light microscope with an AxioCam 3 digital video camera (Zeiss, Oberkochen, Germany). The definition of the phytolith morphotypes was carried out using the International Code for Phytolith Nomenclature 1.0. [8]. For the analysis, 150 phytoliths of each plant were selected. The statistical analysis was performed using Biostat software with an assessment of the statistical importance of the indicators.

2.2. In Silico Analysis. The search for homologues of typical representatives of silicon transporters (SITs), aquaporins, silaffins, silicateins, silacidins, and silicase (peptides and proteins of diatoms, sponges, rice, and corn were chosen), in the nucleotide sequence bases, was conducted using BLAST (http://blast.ncbi.nlm.nih.gov) as described in the paper [33]. From the Uniprot base (March 2014), we obtained the amino acid sequences of SIT (ID O81199 and C7G3B4), aquaporin (ID Q6Z2T3), silaffins (ID Q9SE35, Q5Y2C0 and Q5Y2C2), and silicateins (ID B5B2Z1, B1GSK9, and B5LT52). The sequences of silacidins were obtained from the work [34], and those from silicase were taken from [35].

3. Results

We hypothesized that there might be communication between known biosilicification proteins and/or their homologues and phytolith morphotypes in plants of different phylogenetic origin.

There are no studies regarding representative silaffins, silacidins, silicateins, and silicases (except research concerning aquaporins) in plants. We attempted a computer search (in proteome and genome databases) of the genes and homologues of the proteins identified above (Table 1).

The genomes of all of the studied species of plants contained genes of aquaporins (SIT) and/or their homologues (a high degree of identification of primary structures). Some of these aquaporins have been identified previously (database information). No silaffins, silacidins, silicateins, or silicases were found. However, many short and long fragments of various proteins (some almost whole proteins) (Figures 1, 2, and 3) were homologous to these proteins.

It should be noted that, for all of the studied plants, homologues with existence conserved based pairs of length 3–6 amino acids were observed most often, and longer stretches were rarely observed.

The most characteristic homologues for silicateins were various cathepsins, and carbohydrases were the most characteristic homologues for silicase. Structural homologues of the silacidins presented with very short parts of the amino acid chains of various proteins: unknown proteins, histone deacetylase proteins, xylosidase, photosystem proteins, synthetase, gigantean, DNA binding proteins, oxidase, transcription factors, RNA polymerase, floral homeotic proteins, synthase, and gamma-gliadin.

Only one protein from among the homologues of silacidins participates in a biomineralization: dentin sialophosphoprotein-like of soy bean. The silaffin homologues are represented by unknown proteins, ATPases, transcription factors, kinases, dehydrogenases, synthases, seed maturation proteins, and glucosidases.

The low degree of reliability or no reliability, which is reflected in high E values, is caused by the small length of the homologous chains in most cases. Low level or no reliability was characteristic of silacidins, in which the length of the homologous parts does not exceed 2–20 base pairs. The obtained data are of interest, in particular to matrix peptides and proteins of a biomineralization, silaffins (Figure 1) and silacidins.

The existence of a certain zwitterionic structure [36, 37] in a polypeptide chain (by analogy with silaffins and silacidins) of most of the identified proteins indirectly points to the possibility of their participation in the biosilicification process as catalysts of the sedimentation of biosilicon dioxide. This finding requires further study.

The results of the percentage of the phytolith morphotypes of all of the studied plants are given in Table 2.

The most interesting specific types of phytoliths of algae are presented in Figure 4.

The specific phytolithic profile of the horsetail plant, which is one of the most ancient plants to have survived, is shown in the almost full silicification of the plant surface (Figures 5(a) and 5(b)), which apparently serves as protection against drying and is analogous to the wax layer on floral plants.

The phytolithic profile of a horsetail plant predominantly consists of a polylobate, tracheid, and separately located siliceous stomates (Figures 6(a) and 6(b)).

Another specific phytolith of the horsetail plant is the polylobate, which is a structural component of the silicon "armor" that is fastened with "teeth-" like components [38].

In floral plants, we observed a wide polymorphism and ratio of morphotypes (Figures 7(a)–7(h)).

Apparently, more specific types of phytoliths are observed in floral plants than in algae, lichens, and gymnosperms.

```
Query  272  PTGYPTSSPSHSPTKSPSKSPSSSPTTSPSASPTETPTETPTESPTESPTESPTLSPTES  331
            P G P + P+ +P ++PS +P  +   +P  +PTETPTE P  +PTE+PTE+P L+PTE+
Sbjct  118  PAGAPLNWPNEAPNQAPSNAPMGASLEAPELAPTETPTEAPELAPTETPTEAPELAPTET  177

Query  332  PT----LSPTESPSLSPTLSTTWSPTGYPTLAPSPSPSISSAPSVSSAPSSPPSISSAPS  387
            PT    L+PT++P L+PT + T +P    T AP+ +P +  AP+ ++AP   P++  A
Sbjct  178  PTEAPELAPTDAPELTPTDAPTEAPESTSTNAPAKTPEL--APT-TNAPMQSPTVEEADE  234

Query  388  VSSAPSK 394
               AP+
Sbjct  235  DELAPTN 241
```

(a)

```
Query  24   RTLEGHGGDHSISMSMHSSKAEKQAIEAAVEEDVAGPAKAAKLFKPKASKAGSMP-DEAG  82
            R L  HG      +S      EKQ IE     D+   A   +         AG +  DEA
Sbjct  269  RWLRAHGRPKQPELS------EKQKIELRECFDIIDSAGTGE------GVAGVITTDEAS  316

Query  83   AKSAKMSMDTKSGKSEDAAAVDAKASKESHMSISGDMSMAKSHKAEAEDVTEMSMAKAGK  142
               A ++MD  S + E A     +  K     +   + ++K HK  A+   E + ++AGK
Sbjct  317  LAFAVLNMDMGSPEIEAAIFQINRNKKSMEFPLFQRLILSKLHKWPAD---EYAPSEAGK  373

Query  143  DEASTEDMCMPFA---------KSDKEMSVKSKQGKTEMSVADAKASKE   182
            D     +  +PFA         K  + + V  KQ +  MS    KA +E
Sbjct  374  DAPC--EYTLPFALLAQAYRRKKLIEAVMVGDKQTQDRMSHQSDKAFRE   420
```

(b)

```
Query  276  PTSSPSHSPTKSPSKSP-SSSPTTSPSASPTETPTETPTESPTESPTESPTLSPTESPTL  334
            P  SP  +P  SP  +P   SP  SP  SP E+P ++P ESP ESP ESP  SP ESP
Sbjct  333  PVESPMEAPMSSPMMAPPIESPMESPMESPMESPMKSPMESPMESPMESPMESPMESPVE  392

Query  335  SPTESPSLSP  344
            SP E+P + P
Sbjct  393  SPAETPVMMP  402
```

(c)

Figure 1: Examples of the alignment of the polypeptide chains of silaffins and the proteins of plants. (a) silaffin of diatom *T. pseudonana* (ID Q5Y2C2, Query) and an unknown protein of a soy bean *G. Max* (Sbjct); (b) silaffin of diatom *T. pseudonana* (ID Q5Y2C0, Query) and an unknown protein of the moss *P. Patens* (Sbjct); (c) silaffin of diatom *T. pseudonana* (ID Q5Y2C2, Query) and an unknown protein the moss *P. patens* (Sbjct).

Table 1: The presence of functional and structural homologues of proteins and biosilicification peptides in plants of various types of taxonomy.

	Parameters	Silaffins	Silacidins	Silicase	Silicateins	SIT
Chondrus crispus[1]	*	30 F	2 F	6 F and P	12 F and P	24 F and P
	E value	0.046–9.9	2.4–6.3	$7e^{-04}$–7.2	0.78–6.5	$2e^{-11}$–9.8
	%	26–60	35–48	24–48	26–42	23–79
Physcomitrella patens[2]	*	12 F and P	>100 F	5 F and P	54 F and P	56 F and P
	E value	$1e^{-09}$–2.4	0.002–3.4	$4e^{-13}$–8.3	$1e^{-65}$–5.6	$1e^{-80}$–9.3
	%	37	32–100	26–35	23–45	23–58
Pinus pinaster[3]	*	10 F and P	>100 F	11 F	18 F and P	27 F and P
	E value	0.24–8.3	0.71–1116	0.6–9.2	$5e^{-28}$–10	0.51–69
	%	24–50	26–100	26–40	20–37	19–75
Triticum turgidum[4]	*	13 F	>100 F	2 F	3 F	24 F and P
	E value	0.92–9.5	0.002–390	5.1–6.5	0.078–4.2	$1e^{-20}$–9.8
	%	26–31	32–100	31	23–45	18–56
Glycine max[5]	*	11 F	>100 F	25 F and P	>100 F and P	>100 F
	E value	$1e^{-12}$–5	0.002–1.80	$4e^{-16}$–9.2	$1e^{-70}$–9	$2e^{-9}$–8.7
	%	24–40	32–100	19–37	22–52	23–67

*Number of homologues, fragment (F), protein (P), [1]red algae, [2]mosses, [3]coniferous, [4]monocotyledonous, and [5]dicotyledonous.

```
Query  35   HQKSYQNDLEELDRHTVWLSNKKYIEAHNQNSHVFGFTLAMNHFADLTDQEWTE----KF  90
            ++K Y+   E   R   +L + K +E HN+  H   ++LA+N FAD+T +E+ +    K
Sbjct  36   YKKEYKTVEELKHRFVTFLESVKLVETHNKGQH--SYSLAVNEFADMTFEEFRDSRLMKG  93

Query  91   VTHVSDTAGNYTKYYEPNQFKSYPDTVDWRTKDAVTKVKDQSQCGASYAFSAVGALEGAN  150
              + S T GN+    E     S P T DWR +  V++VK+Q+ CG+ + FS  GALE A+
Sbjct  94   EQNCSATVGNHVLTGE-----SLPKTKDWREEGIVSQVKNQASCGSCWTFSTTGALEAAH  148

Query  151  ALATGSLSVLSEQNIIDCSVPYGNHGCKGGNMLYAFKYIIANDGLDVAKSYPFQGKQQSC  210
            A ATG + +LSEQ ++DC+  + N GC GG    AF+YI  N G+D   SYP+  K   C
Sbjct  149  AQATGKMVLLSEQQLVDCAGEFNNFGCGGGLPSQAFEYIRYNGGIDTEDSYPYNAKDSQC  208

Query  211  VYDDQDTGGKISGMVRIKQGSESDLIGAVANVGPVSVAIDGSSNAFRFYASGVYDSSRCS  270
             +     G ++  +V I +G+E+ L  A+A + PVSVA +   + FR Y  GVY S  C
Sbjct  209  RFHKNTIGAQVWDVVNITEGAETQLKHAIATMRPVSVAFE-VVHDFRLYNGGVYTSLNCH  267

Query  271  S--SKLNHAMVVTGYGT-YGGKDYWLVKNSWGTNWGQSGYIMMARGKYNQCGIASDACYP  327
            +    +NHA++  GYG    G  YW++KNSWG +WG +GY  M  GK N CG+A+ A YP
Sbjct  268  TGPQTVNHAVLAVGYGEDENGVPYWIIKNSWGADWGMNGYFNMEMGK-NMCGVATCASYP  326

Query  328  TL 329
             +
Sbjct  327  VV 328
```

Figure 2: Example of the alignment of the polypeptide chains of silicatein of a sponge, *L. oparinae* (ID B5LT52, Query), and an unknown protein of the moss *P. patens* (Sbjct).

```
Query  1    MALASVAKVVLGSIAFCVFWVLAVFPSVPFMPIGRTAGALLSAVLMIVFHVISPDDAYAS  60
            MALASV KVV G IAF +FWVLAVFP++PF+PIGRTAG+LL A+LM++F VI+PD+AY +
Sbjct  1    MALASVPKVVFGLIAFAIFWVLAVFPAIPFLPIGRTAGSLLGAMLMVIFQVITPDEAYDT  60

Query  61   VDLPILGLLFATMVVGSYLKNAGMFKHLGTLLAWRSQGGRDLLCRVCVVTALASALFTND  120
            +DLPILGLLF TMVV +YL+ A MFK++G LLAW+S+G +DLLCR+CV++A++SALFTND
Sbjct  61   IDLPILGLLFGTMVVSTYLERADMFKYIGKLLAWKSRGAKDLLCRICVISAISSALFTND  120

Query  121  TCCVVLTEFVLELAAERNLPAKPFLLALASSANIGSSATPIGNPQNLVIAFNSKISFPKF  180
            T CVVLTEF+L++A + NLP  PFLLALASSANIGSSATPIGNPQNLVIA  SKISF  F
Sbjct  121  TSCVVLTEFILKIARQHNLPPTPFLLALASSANIGSSATPIGNPQNLVIAVQSKISFGNF  180

Query  181  LLGILPAMLAGMAVNMVMLLCMYW------KDLDGSGSG---MDLDGKR-----------  220
            L+GILPAM+AG+  N ++LL M+W      KD + +G+     + D  R
Sbjct  181  LIGILPAMVAGVVANAIILLIMFWKLLSVHKDEEDAGAEDVVEEYDSHRFSPATMSHYSS  240

Query  221  ---MEAVEEGAVVVVEPSPKQQ--------------------------QQQLGGSNGYM  250
                E       + V+ SP+ Q                                 G+NG
Sbjct  241  LNSQEWSSHLDAITVQNSPQVQILRNRSIANASESNGISSNTFDTARISSVSRDGTNGVA  300

Query  251  SPLMTE----NISTKHPWFMQCTEQR-------RKLFIKSFAYVVTVGMVIAYMVGLNMS  299
            S    E    N S      +  +E++       +++  KS  Y++TVGM++A ++GLNMS
Sbjct  301  SMAKEETSPSNSSAGVDTLIPPSERKTNFIIKWKRVLWKSCVYIITVGMLVALLLGLNMS  360

Query  300  WTAITTAIALVVVDFRDAEPCLNTVSYSLLVFFSGMFITVSGFNKTGLPAAIWNFMAPYS  359
            WTAIT A+AL+V+DF+DA PCL  VSYSLL+FF GMFITV G NKTG+P+A+W+ M PYS
Sbjct  361  WTAITAALALIVLDFKDATPCLEKVSYSLLIFFCGMFITVDGLNKTGIPSALWDIMEPYS  420

Query  360  KVNSVGGISVLSVIILLLSNLASNVPTVLLMGGEVASAAALISPAAVVRSWLLLAWVSTV  419
             V+   GI++L+++IL+LSNLASNVPTVLL+GG VA++AA IS A   ++WL+LAW ST+
Sbjct  421  HVDRASGIAILAIVILVLSNLASNVPTVLLLGGRVAASAAAISKADEKKAWLILAWASTI  480

Query  420  AGNLSLLGSAANLIVCEQARRAQRNAYDLTFWNHIVFGVPSTLIVTAIGIPLI  472
            +GNLSLLGSAANLIVCEQA RA    Y LTFW+H+ FG+PST+IVTAIG+  I
Sbjct  481  SGNLSLLGSAANLIVCEQAIRAPNLPYTLTFWSHLKFGLPSTIIVTAIGLTFI  533
```

Figure 3: Example of the alignment of the polypeptide chains of the SIT of corn, *Z. mays* (ID C7G3B4, Query), and an unknown protein of soy bean *G. Max* (Sbjct).

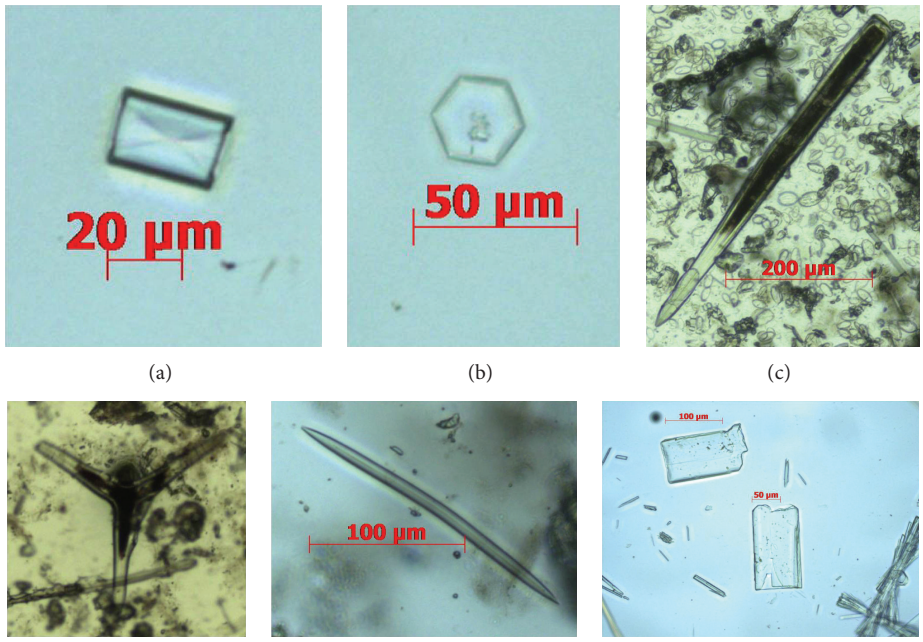


Figure 4: The most characteristic forms of the phytoliths of algae: (a-b) *Tichocarpus crinitus:* (a) pyramid; (b) hexagon, (c) *Laurencio tropica*: (c) hollow needle, (d-e) *Amphiroa anceps*: (d) triple needle; (e) needle, (f) *Fucus evanescens*: (f) long smooth rectangular.

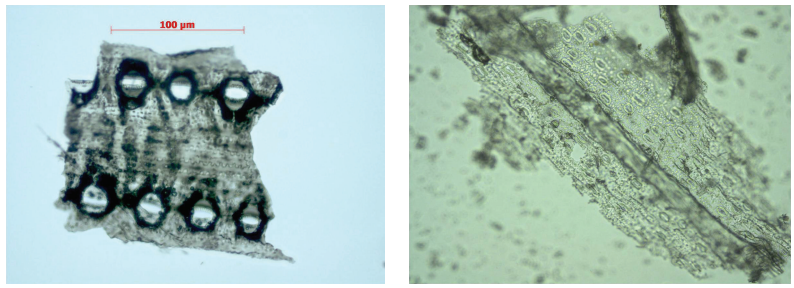


Figure 5: Parts of the silicon armor of *Equisetum hyemale* (a) and *Equisetum fluviatile* (b).

4. Discussion

The results of the comparison of the total number of homologues allowed us to draw some conclusions. Figure 8 shows the plants in which the homologues of silacidins are observed most often. This group is extremely widespread across the phylogenetic tree of plants. We hypothesized that matrix peptides and proteins (silaffins and silacidins) could be among the most ancient matrix substances of biomineralization [33].

Based on the degree of frequency of the occurrence of homologues, SIT is first, followed by silicateins, silaffins, and silicase. Figure 8 shows that lines of siliffins and silicateins are similar and demonstrate their joint participation in plant processes.

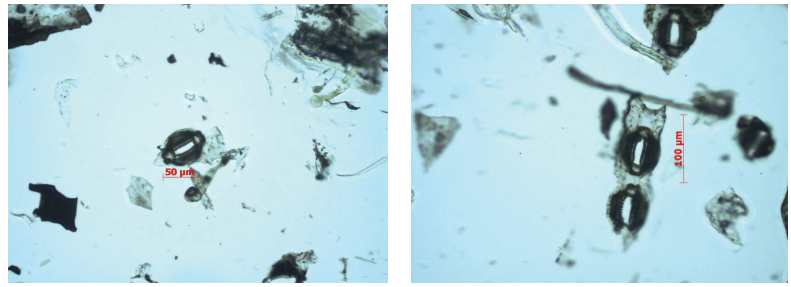
In the analysis, primitive red algae had the smallest number (70–80) of homologues of the biomineralization proteins of the 4 types. The number of phytoliths morphotypes in red and brown algae (4–6) was insignificant.

In Monocotyledonous, 142 homologues were found. Coniferous contained 166 homologues and from 3 to 5 morphotypes, including a specific type in the shape of a circle. More homologues of biomineralization were observed more in the Mosses than in the other groups (more than 227). Mosses are an ancient deadlocked branch of the highest plants, which appeared at the beginning from algae of the Cambrian Period (approximately 600 million years ago) [39].

More than 336 homologues and from 8 to 12 phytolith morphotypes were observed in Dicotyledones.

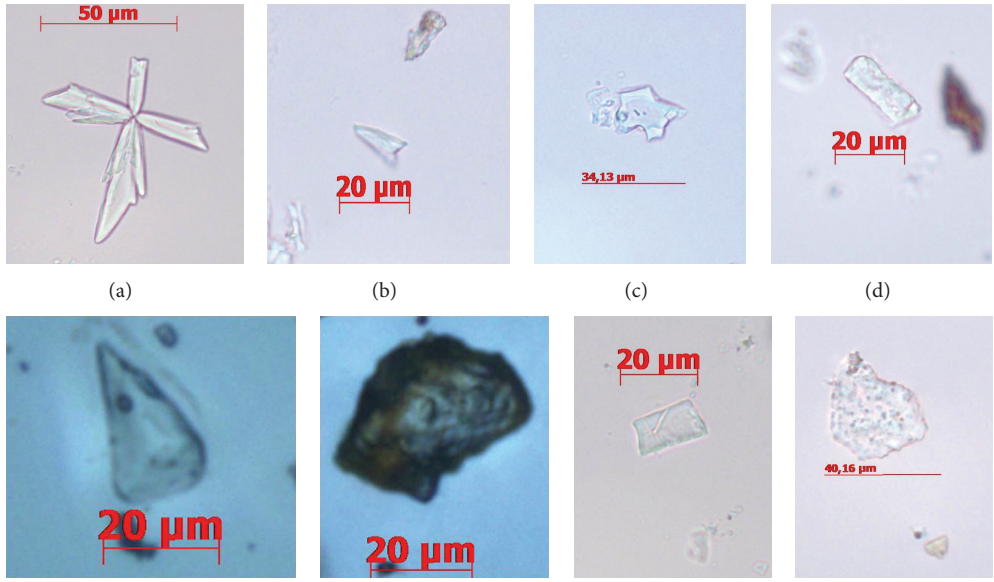
Across all of the plants (from primitive algae to floral) a variety of forms (12 morphotypes in certain case) and a large quantity of biomineralization protein homologues were observed.

The morphometric analysis provided additional data indicating that the sizes of the phytoliths of floral plants differ from each other by no more than 10-fold. In primitive plants, particularly in algae and lichens, the difference in the sizes



(a) (b)

Figure 6: Separately located stomates from the silicon armor of *Equisetum hyemale*.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7: The most characteristic phytolith forms of floral plants were as follows. (a) *Berberis amurensis*: (a) stellatus (meet 4-, 5-, 6-, 7-, and 8-beam differences); (b–d) *Schisandra chinensis*: (b) bark, cuneiform; (c) bark, polyhedron smooth; and (d) leaves, parallelepiped extended rough; (e-f) *Panax ginseng*: (e) leaves, cuneiform; (f) root, irregular ruminate; (g) *Bergenia pacifica*: (g) leaves, parallelepiped smooth; and (h) *Eleutherococcus senticosus*: (h) leaves, irregular ruminate.

of the phytoliths of similar morphotypes is 2 or 3 and even 4-fold.

The phytoliths in these primitive plants considerably exceed the size of the phytoliths in more complexly organized plants. In *Laurencio tropica* red algae, the surface area of the needle phytolith morphotype is 19489,32 ± 875,75 mkm^2, which is 10 times larger than the average area of the phytoliths of similar morphotypes in floral plants, and, in particular, in *Eleutherococcus senticosus* (1673,64 ± 1108,22 mkm^2). The identical difference in the indicators in different plants is observed for other morphometric parameters, including the perimeter, length, and width.

Identical phytolith morphotypes in plants in different ecological niches carry out various functions.

In most cases, in all of the studied plants (from algae to floral plants), there are two main types of phytoliths, irregular smooth (shapeless) and irregular ruminate. These two types have a likely universal role. The structural mechanisms must be universal for all of the groups, as in the case of the implementation of stability and stalk strength. In the structural phytolith group, it is worth considering the extended morphotypes. The synthesis mechanism of the shapeless phytoliths is similar across the entire phylogenetic tree and is possibly connected with silacidins, as their homologues occur most often. Silacidins are small peptides that catalyze the formation and regulate the size of the nanospheres of silicon dioxide from silicon acid and play a central role in the formation of the cellular wall of diatom algae [34].

It is logical to assume that these phytoliths have a general ancient function in all types of plants, which is confirmed by our data. In all primitive plants, shapeless and extended phytoliths (Table 3) prevail.

Specific types of phytoliths unambiguously show specific and probably unique functions in groups of plants that

Table 2: Phytolith morphotypes from the plants from different taxons.

	Number of morphotypes	Prevailing morphotypes, %	Specific morphotypes, %	Figure
		Red algae		
Tichocarpus crinitus	6	Irregular ruminate (*41.3*)	Pyramid (*36*) Hexagon (*11.3*)	Figure 4(a) Figure 4(b)
Mastocarpus stellatus	4	Irregular smooth (*38*)	Cylinder smooth (*8.7*)	
Laurencio tropica	5	Irregular ruminate (*13.3*)	Hollow needle (*8.7*)	Figure 4(c)
Amphiroa anceps	5	Irregular ruminate (*46.7*)	Needle (*8.7*) Triple needle (1%)	Figure 4(d) Figure 4(e)
		Brown algae		
Fucus evanescens	4	Irregular ruminate (*33.5*) Irregular smooth (*25.5*)	Elongate smooth (*5.3*) Elongate rectangle smooth (*4.7*)	Figure 4(f)
Saccharina latissima	6	Irregular smooth (25.7) Irregular ruminate (23.7)	Elongate (*19.7*)	
		Lichens		
Cladonia spp.	4	Irregular ruminate (*56*) Irregular smooth (*28*)	Oval (*4.7*)	
Usnea spp.	5	Irregular smooth (*43.3*) Irregular ruminate (*24.6*)	Circle (*16*)	
		Horsetails		
Equisetum fluviatile	3	Silicification of a surface of a plant	Polylobate Tracheid Stomata	Figures 5(a) and 5(b) Figures 6(a) and 6(b)
Equisetum hyemale	3	Silicification of a surface of a plant	Polylobate Tracheid Stomata	Figures 5(a) and 5(b) Figures 6(a) and 6(b)
		Coniferous		
Pinus koraiensis	4	Circle (*50.67*) Irregular ruminate (*30*)	Circle (*50.7*)	
Abies squamata	3	Irregular ruminate (*68*) Irregular smooth (*28.7*)	Circle (*3*)	
Juniperus sibirica	3	Irregular ruminate (*59.33*)	Circle (*2.7*)	
Larix cajanderi	5	Irregular ruminate (*62.3*)	Circle (*24.7*)	
		Floral		
Berberis amurensis	8	Stellatus (*48*) Irregular ruminate (*17*)	Stellatus (*48*)	Figure 7(a)
Schisandra chinensis	12	Irregular ruminate (*43*)	Cylinder smooth curved (*10*) Epidermal short cell (*7*)	
Panax ginseng	7	Irregular smooth (*35.3*) Cuneiform (*10.7*)	Stomata	
Bergenia pacifica	8	Irregular ruminate (*37*) Globular (*20*)	Oblong (*13*) parallel epipedal with the rounded-off edges (9)	
Eleutherococcus senticosus	5	Irregular ruminate (*85*)		

Table 3: Evolutionary aspects of the morphotypes of phytoliths.

Age	Morphotypes	Plant taxons
Ancient	Irregular smooth, irregular ruminate, needles, elongated, elongate, and geometrically correct (pyramids, hexagons, etc.)	Red algae, lichens
Median	Circular and polylobate, silicification of a surface of a plant	Horsetails, Coniferous plants
Modern	Rectangular, elongated, and polygonal	Brown algae, floral

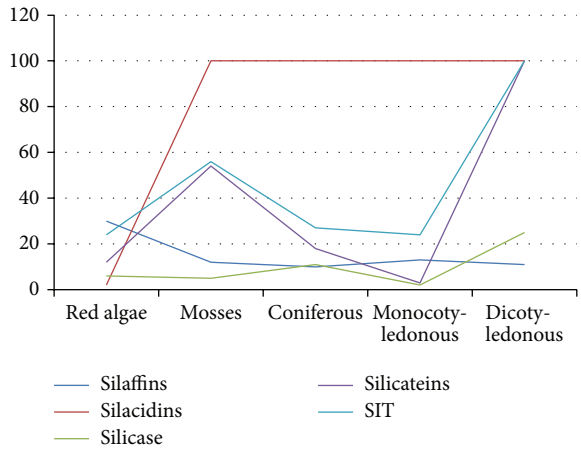


Figure 8: Chart of the quantity of the homologues of the biomineralization proteins in different taxonomical groups of plants.

appeared later with evolutionary selection. Klančnik et al. found that the circle- and needle-shaped phytoliths found in some algae could carry out optical functions and focus and redirect sunlight [13].

We demonstrated previously [40] that micrometer silicate aluminosilicate objects could concentrate charges comparable to the membrane potential on the surface (according to our own data, the charge of the particles of quartz of 1–10 microns in size reaches more than −27 mV and that of volcanic glass reaches more than −36 mV). Some phytolith morphotypes in plants could be catalysts of various chemical reactions in the intercellular environment and of supramolecular interactions in the course of ontogenesis.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the Scientific Fund of the Far Eastern Federal University, the Government Task Force of the Ministry of Education and Science of the Russian Federation, the Russian Scientific Fund, and the Presidential Grant for Young Science Candidates (MK-1547.2013.5).

References

[1] J. A. Raven, "Cycling silicon—the role of accumulation in plants," *New Phytologist*, vol. 158, no. 3, pp. 419–430, 2003.

[2] Z. Li, P. Lin, J. He, Z. Yang, and Y. Lin, "Silicon's organic pool and biological cycle in moso bamboo community of Wuyishan Biosphere Reserve," *Journal of Zhejiang University: Science B*, vol. 7, no. 11, pp. 849–857, 2006.

[3] Y. Liang, W. Sun, Y. Zhu, and P. Christie, "Mechanisms of silicon-mediated alleviation of abiotic stresses in higher plants: a review," *Environmental Pollution*, vol. 147, no. 2, pp. 422–428, 2007.

[4] H. Mizuta and H. Yasui, "Protective function of silicon deposition in *Saccharina japonica* sporophytes (Phaeophyceae)," *Journal of Applied Phycology*, vol. 24, no. 5, pp. 1177–1182, 2012.

[5] I. Sivanesan and B. R. Jeong, "Silicon promotes adventitious shoot regeneration and enhances salinity tolerance of *Ajuga multiflora* bunge by altering activity of antioxidant enzyme," *The Scientific World Journal*, vol. 2014, Article ID 521703, 10 pages, 2014.

[6] D. R. Piperno, *Phytolith Analysis, An Archaeological and Geological Perspective*, Academic Press, San Diego, Calif, USA, 1988.

[7] D. R. Piperno, *Phytoliths: A Comprehensive Guide for Archaeologists and Paleoecologists*, AltaMira Press, Lanham, Md, USA, 2006.

[8] M. Madella, A. Alexandre, and T. Ball, "International code for phytolith nomenclature 1.0," *Annals of Botany*, vol. 96, no. 2, pp. 253–260, 2005.

[9] S. Agarie, W. Agata, H. Uchida, F. Kubota, and P. B. Kaufman, "Function of silica bodies in the epidermal system of rice (*Oryza sativa* L.): testing the window hypothesis," *Journal of Experimental Botany*, vol. 47, no. 298, pp. 655–660, 1996.

[10] P. Bauer, R. Elbaum, and I. M. Weiss, "Calcium and silicon mineralization in land plants: transport, structure and function," *Plant Science*, vol. 180, no. 6, pp. 746–756, 2011.

[11] C. A. E. Strömberg, "Evolution of grasses and grassland ecosystems," *Annual Review of Earth and Planetary Sciences*, vol. 39, pp. 517–544, 2011.

[12] C. A.E. Strömberg, R. E. Dunn, R. H. Madden, M. J. Kohn, and A. A. Carlini, "Decoupling the spread of grasslands from the evolution of grazer-type herbivores in South America," *Nature Communications*, vol. 4, article 1478, 2013.

[13] K. Klančnik, K. Vogel-Mikuš, and A. Gaberščik, "Silicified structures affect leaf optical properties in grasses and sedge," *Journal of Photochemistry and Photobiology B*, vol. 130, pp. 1–10, 2014.

[14] D. Moreira, H. Le Guyader, and H. Philippe, "The origin of red algae and the evolution of chloroplasts," *Nature*, vol. 405, no. 6782, pp. 69–72, 2000.

[15] G. I. McFadden and G. G. van Dooren, "Evolution: red algal genome affirms a common origin of all plastids," *Current Biology*, vol. 14, no. 13, pp. R514–R516, 2004.

[16] L. A. Lewis and R. M. McCourt, "Green algae and the origin of land plants," *American Journal of Botany*, vol. 91, no. 10, pp. 1535–1556, 2004.

[17] H. S. Yoon, K. M. Müller, R. G. Sheath, F. D. Ott, and D. Bhattacharya, "Defining the major lineages of red algae (Rhodophyta)," *Journal of Phycology*, vol. 42, no. 2, pp. 482–492, 2006.

[18] J. T. Clarke, R. C. M. Warnock, and P. C. J. Donoghue, "Establishing a time-scale for plant evolution," *New Phytologist*, vol. 192, no. 1, pp. 266–301, 2011.

[19] M. D. Guiry and G. M. Guiry, *AlgaeBase*, World-wide electronic publication, National University of Ireland, Galway, Ireland, 2013, http://www.algaebase.org.

[20] N. J. Butterfield, "*Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes," *Paleobiology*, vol. 26, no. 3, pp. 386–404, 2000.

[21] H. Nozaki, M. Matsuzaki, M. Takahara et al., "The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids," *Journal of Molecular Evolution*, vol. 56, no. 4, pp. 485–497, 2003.

[22] H. S. Yoon, J. D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya, "A molecular timeline for the origin of photosynthetic eukaryotes," *Molecular Biology and Evolution*, vol. 21, no. 5, pp. 809–818, 2004.

[23] T. N. Taylor, H. Hass, W. Remy, and H. Kerp, "The oldest fossil lichen," *Nature*, vol. 378, no. 6554, p. 244, 1995.

[24] G. J. Retallack, "Were the Ediacaran fossils lichens?" *Paleobiology*, vol. 20, no. 4, pp. 523–544, 1994.

[25] A. H. Jahren, S. Porter, and J. J. Kuglitsch, "Lichen metabolism identified in Early Devonian terrestrial organisms," *Geology*, vol. 31, no. 2, pp. 99–102, 2003.

[26] B. J. Fletcher, D. J. Beerling, and W. G. Chaloner, "Stable carbon isotopes and the metabolism of the terrestrial Devonian organism *Spongiophyton*," *Geobiology*, vol. 2, no. 2, pp. 107–119, 2004.

[27] G. J. Retallack, "Growth, decay and burial compaction of Dickinsonia, an iconic Ediacaran fossil," *Alcheringa*, vol. 31, no. 3, pp. 215–240, 2007.

[28] X. Yuan, S. Xiao, and T. N. Taylor, "Lichen-like symbiosis 600 million years ago," *Science*, vol. 308, no. 5724, pp. 1017–1020, 2005.

[29] W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, Cambridge University Press, Cambridge, UK, 2nd edition, 1993.

[30] A. Bennici, "Origin and early evolution of land plants. Problems and considerations," *Communicative & Integrative Biology*, vol. 1, no. 2, pp. 212–218, 2008.

[31] L. Medlin, W. H. C. F. Kooistra, D. Potter, G. Saanders, and R. A. Wandersen, "Phylogenetic relationships of the 'golden algae' (haptophytes, heterokont chromophytes) and their plastids," in *The Origin of the Algae and their Plastids*, D. Bhattacharya, Ed., vol. 11 of *Plants Systematics and Evolution*, pp. 187–219, 1997.

[32] B.-L. Lim, "Molecular evolution of 5S ribosomal RNA from red and brown algae," *Japanese Journal of Genetics*, vol. 61, no. 2, pp. 169–176, 1986.

[33] I. E. Pamirsky and K. S. Golokhvast, "Origin and status of homologous proteins of biomineralization (Biosilicification) in the taxonomy of phylogenetic domains," *BioMed Research International*, vol. 2013, Article ID 397278, 7 pages, 2013.

[34] P. Richthammer, M. Börmel, E. Brunner, and K. Van Pée, "Biomineralization in diatoms: the role of silacidins," *ChemBioChem*, vol. 12, no. 9, pp. 1362–1366, 2011.

[35] H. C. Schröder, A. Krasko, D. Brandt et al., "Silicateins, silicase and spicule-associated proteins: synthesis of demosponge silica skeleton and nanobiotechnological applications," in *Porifera Research: Biodiversity, Innovation and Sustainability*, pp. 581–592, Museu Nacional, Rio de Janeiro, Brazil, 2007.

[36] N. Kröger, R. Deutzmann, and M. Sumper, "Silica-precipitating peptides from diatoms: the chemical structure of silaffin-1A from *Cylindrotheca fusiformis*," *The Journal of Biological Chemistry*, vol. 276, no. 28, pp. 26066–26070, 2001.

[37] S. V. Patwardhan, K. Shiba, H. C. Schröder, W. E. G. Müller, S. J. Clarson, and C. C. Perry, "The interaction of silicon with proteins. Part 2. The rold of bioinspired peptide and recombinant proteins in silica polymerization," *ACS Symposium Series*, vol. 964, pp. 328–347, 2007.

[38] M. Bonomo, A. F. Zucol, B. Gutiérrez Téllez, A. Coradeghini, and M. S. Vigna, "Late holocene palaeoenvironments of the Nutria Mansa 1 archaeological site, Argentina," *Journal of Paleolimnology*, vol. 41, no. 2, pp. 273–296, 2009.

[39] B. Goffinet and A. J. Shaw, Eds., *Bryophyte Biology*, Cambridge University Press, Cambridge, UK, 2nd edition, 2008.

[40] I. E. Pamirsky, K. S. Golokhvast, A. M. Panichev et al., "Influence of nano- and microparticles of natural minerals on human thrombocytes aggregation," *Reports of the Samara Scientific Center RAS*, vol. 4, no. 3, pp. 723–728, 2010 (Russian).

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 231589, 14 pages
http://dx.doi.org/10.1155/2014/231589

Research Article

Retrotransposon-Based Molecular Markers for Analysis of Genetic Diversity within the Genus *Linum*

Nataliya V. Melnikova,[1] Anna V. Kudryavtseva,[1] Alexander V. Zelenin,[1] Valentina A. Lakunina,[1] Olga Yu. Yurkevich,[1] Anna S. Speranskaya,[1,2] Alexey A. Dmitriev,[1] Anastasia A. Krinitsina,[2] Maxim S. Belenikin,[1,3] Leonid A. Uroshlev,[1] Anastasiya V. Snezhkina,[1] Asiya F. Sadritdinova,[1] Nadezda V. Koroban,[1] Alexandra V. Amosova,[1] Tatiana E. Samatadze,[1] Elena V. Guzenko,[4] Valentina A. Lemesh,[4] Anastasya M. Savilova,[5] Olga A. Rachinskaia,[1] Natalya V. Kishlyan,[6] Tatiana A. Rozhmina,[6] Nadezhda L. Bolsheva,[1] and Olga V. Muravenko[1]

[1] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia*
[2] *Department of Higher Plants, Lomonosov Moscow State University, Moscow 119991, Russia*
[3] *Research Institute of Physico-Chemical Medicine, Moscow 119435, Russia*
[4] *Institute of Genetics and Cytology, National Academy of Science of Belarus, 220072 Minsk, Belarus*
[5] *Research Center for Obstetrics, Gynecology and Perinatology, Moscow 117997, Russia*
[6] *All-Russian Research Institute for Flax of the Russian Academy of Agricultural Sciences, Torzhok 172002, Russia*

Correspondence should be addressed to Nadezhda L. Bolsheva; nlbolsheva@mail.ru

Received 25 April 2014; Revised 18 July 2014; Accepted 1 August 2014; Published 27 August 2014

Academic Editor: Peter F. Stadler

SSAP method was used to study the genetic diversity of 22 *Linum* species from sections *Linum*, *Adenolinum, Dasylinum, Stellerolinum*, and 46 flax cultivars. All the studied flax varieties were distinguished using SSAP for retrotransposons *FL9* and *FL11*. Thus, the validity of SSAP method was demonstrated for flax marking, identification of accessions in genebank collections, and control during propagation of flax varieties. Polymorphism of *Fl1a, Fl1b*, and *Cassandra* insertions were very low in flax varieties, but these retrotransposons were successfully used for the investigation of *Linum* species. Species clusterization based on SSAP markers was in concordance with their taxonomic division into sections *Dasylinum, Stellerolinum, Adenolinum*, and *Linum*. All species of sect. *Adenolinum* clustered apart from species of sect. *Linum*. The data confirmed the accuracy of the separation in these sections. Members of section *Linum* are not as closely related as members of other sections, so taxonomic revision of this section is desirable. *L. usitatissimum* accessions genetically distant from modern flax cultivars were revealed in our work. These accessions are of utmost interest for flax breeding and introduction of new useful traits into flax cultivars. The chromosome localization of *Cassandra* retrotransposon in *Linum* species was determined.

1. Introduction

The genus *Linum* comprises about 200 species which are distributed throughout the temperate and subtropical regions of the world. The genus is subdivided by Ockendon and Walters into five sections: *Linum, Dasylinum* (Planch.) Juz., *Linastrum* (Planchon), Bentham, *Syllinum* Griseb., and *Cathartolinum* (Reichenb.) Griseb. [1]. Some taxonomists classified the members of the *L. perenne* group from section *Linum* to an independent section *Adenolinum* (Reichenb.) Juz. [2, 3]. The species *L. stelleroides* (Planch.), distributed in Far East and China, was classified by Yuzepchuk [2] to a monotype section *Stellerolinum* Juz. ex Prob. The phylogenetic analyses based on chloroplast (*ndhF*, *trnL-F*, and *trnK* 3′ intron) and

nuclear ITS (internal transcribed spacer) DNA sequences revealed that genus *Linum* was not monophyletic. It contains two major lineages: a yellow-flowered clade (sections *Linopsis*, *Syllinum*, and *Cathartolinum*) and a blue-flowered clade (sections *Linum*, *Dasylinum*, and *Stellerolinum*) [4]. The cultivated flax (*L. usitatissimum* L.) belongs to sec. *Linum* from a blue-flowered clade. *L. usitatissimum* is believed to have originated as a result of domestication of wild species *L. angustifolium* Huds. approximately 8000 years ago [5–8]. For a long time flax has been cultivating as a dual-purpose crop grown for its fiber and linseed oil.

According to morphological and qualitative traits, cultivated flax was divided into five main types: (1) fiber flax (*L. usitatissimum* subsp. *usitatissimum*); (2) oil flax (*L. usitatissimum* L. subsp. *humile* Czernom.); (3) dual-purpose flax (*L. usitatissimum* L. subsp. *intermedium* Czemom.) that was an intermediate form between the first two ones cultivated for fiber and oil; (4) large seeded flax (*L. usitatissimum* L. subsp. *latifolium* Snankev.) which is characterized by a set of specific morphological features and cultivated for oil in the Mediterranean region and North Africa; (5) winter flax (*L. usitatissimum* L. subsp. *bienne* Mill. Snankev.) cultivated for fiber and oil in the Caucasus, Turkey, Balkans, and some other south regions of Europe [9, 10]. In addition, collections of flax germplasm maintain accessions of primitive flax forms with dehiscent capsules (*L. usitatissimum* convar. *crepitans* [Boenningh.] Kulpa et Danert) [11].

The taxonomy of the genus cannot be considered as finally established one because the phylogenetic linkages between the individual taxa have not been sufficiently investigated. The phylogeny of species of the genus *Linum* was previously studied by the use of molecular and cytogenetic approaches [4, 12–17], but there are problems that still remain to be solved.

Transposon-based molecular markers are successfully used in phylogenic studies. Transposable elements were shown to influence changing in genomic structure as well as transcriptional regulation occurring during the evolution [18, 19]. The presence of transposons in various species of plants, their high integration activity, conservative sequences, and a large number of copies encouraged the use of transposons in the studies of genetic diversity and profiling of plant varieties [20–22]. Several molecular marker systems based on the information available for the transposable elements sequences were developed for plants [20, 22–27]. SSAP (sequence-specific amplified polymorphism) method was shown to have a number of advantages as compared to other marker systems. SSAP method produces many polymorphic fragments and allows differentiation of most samples using only a single combination of specific primers [23, 28–30]. Different plants were successfully studied by SSAP analysis, but the method has not been applied for the investigation of species of the genus *Linum* yet. Only recently flax sequences have appeared in databases [31–34], and development of a marker system based on flax transposable elements for the investigation of cultivated and wild species of the genus *Linum* has become possible.

In this study the SSAP method was used for assessment of genetic diversity. Besides, the possibilities of application of marker-based profiling for identification of *L. usitatissimum* varieties were analyzed. We studied 46 varieties of *L. usitatissimum* mainly bred in Russia and a number of varieties which were grown in geographically close or distant regions. We also analyzed different types of cultivated flax (fiber, oilseed, large seeded, winter, and dehiscent flax) together with 21 wild species and subspecies from sections *Linum*, *Adenolinum*, *Dasylinum*, and *Stellerolinum* to estimate the possibility of using the SSAP method for the investigation of flax domestication history and phylogenic linkages between different taxa of the genus *Linum*.

2. Materials and Methods

2.1. Plant Materials. For the investigation of genetic diversity of cultivated flax, 46 *L. usitatissimum* varieties, mainly of Russian origin, were obtained from the All-Russian Research Institute for Flax (VNIIL) (Table 1). For analyzing the reproducibility of the SSAP analysis, flax variety "Stormont cirrus" (IPK genbank, Gatersleben, accession number: LIN 261) was used.

For the investigation of genetic diversity, 47 flax accessions belonging to 22 species from sections *Linum*, *Adenolinum*, *Dasylinum,* and *Stellerolinum* were used (Table 2). Most of these accessions were obtained from genebank of Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) (Gatersleben, Germany), seed collections of All-Russian Flax Institute (VNIIL) (Torzok, Russian Federation), N.I. Vavilov Research Institute of Plant Industry (VIR) (St. Petersburg, Russian Federation), and Institute of Genetics and Cytology of NAS Belarus (IGC) (Minsk, Belarus). The accessions of *L. amurense* and *L. stelleroides* were kindly provided by Dr. L. N. Mironova, Botanic Garden Institute of the Far-Eastern Branch of the Russian Academy of Sciences (BGI) (Vladivostok, Russian Federation). Some accessions were collected in the wild by Dr. A. A. Svetlova, Komarov Botanical Institute RAS (St. Petersburg, Russian Federation), by Dr. N. L. Bolsheva, Engelhardt Institute of Molecular Biology RAS (Moscow, Russian Federation), and by Dr. M. Pavelka, Euroseeds (Novy Jicin, Chech Rep.).

2.2. Confirmation of Species Determination. Species determination of some accessions of wild *Linum* species was done during the course of our earlier cytogenetic investigations [14, 17, 37]. To confirm the species determination, the rest of the accessions were planted in the ground, and, additionally, chromosome analysis (determination of chromosome number) using acetocarmine staining according to previously developed approach was performed [14]. Chromosome numbers of all the studied accessions of wild flax are represented in Table 2.

2.3. SSAP Analysis. Genomic polymorphism of different *Linum* species was studied using SSAP method [23] with modifications described earlier through genomic studies of wheat [38] and strawberry [39]. Total genomic DNA was extracted from young flax leaves according to Edwards et al. [40] with minor modifications. 30 ng of genomic DNA

Table 1: Studied flax varieties.

Number	Variety	Type	Originator	Variety pedigree
1	Alfa	Fiber	VNIIL	L-1120, Tomskij 4, T-10, Torzhokskij-4, G-4887
2	Gorizont	Fiber	VNIIL	L-1120, T-5, T-10, I-2, I-17, Tvertsa. Zarya, Marit Smolenskij, Severyanin, Silva, Natasha, Viera
3	Lavina	Fiber	Smolenskaja GOSHOS	Sadko × S-108 [X-5 (selected from L-1120) × I-7]
4	Vasilyok	Fiber	Belarus	I-7, L-1120, X-5 (selected from L-1120), 221-84-4
5	Soyuz	Fiber	Smolenskaja GOSHOS	
6	Lada	Fiber	VNIIL	Viking, Fibra
7	Slavnyj 82	Fiber	VNIIL	from Shokinskij variety, by single plant selection
8	Krom	Fiber	Pskovskij NIISH	
9	Tverskoj	Fiber	VNIIL	L-1120, T-4, T-6, T-10, Torzhokskij-4, I-2, G-2782
10	Rosinka	Fiber	VNIIL	L-1120, T-5, I-17, Fibra, 1288/12
11	Lenok	Fiber	VNIIL	L-1120, M-34 (L-1120 × T-4), T-6, Lazurnyj
12	A-93	Fiber	VNIIL	L-1120, T-4, T-6, T-10, Torzhokskij-4
13	Tost 3	Fiber	Sibirskij NIISH	
14	Mogilevskij-2	Fiber	Belarus	Stahanovets, L-1120, T-5, T-9
15	Belochka	Fiber	Vjatskaja GSHA	L-1120, Tvertsa
16	Torzhokskij-4	Fiber	VNIIL	M-34 (L-1120 × T-4) × T-10 (G-360 × G-354)
17	Tvertsa	Fiber	VNIIL	T-5 × L-1120
18	Smolich	Fiber	Smolenskaja GOSHOS	Zarya (X-5 × T-5), mutant A-710, 806/3, severyanin
19	A-29	Fiber	VNIIL	L-1120, T-4, T-5, T-10
20	Antey	Fiber	Pskovskij NIISH	
21	Rusich	Fiber	Pskovskij NIISH	Pskovskij 83 (Pryadilshhik, L-1120, T-5, Pobeditel) × Rodnik (T-9, Progress, I-9, L-1120, VNIIL11, Spartak)
22	Veralin	Fiber	Netherlands	Torzhokskij-4 × Lidiya
23	Merilin	Fiber	Netherlands	
24	Smolenskij	Fiber	Smolenskaja GOSHOS	Tvertsa (T-5 × L-1120) × Zarya (X-5—selected from L-1120 × T-5)
25	Tost-5	Fiber	Sibirskij NIISH	
26	S-108	Fiber	Smolenskaja GOSHOS	X-5 (selected from L-1120) × I-7
27	Impuls	Fiber	Smolenskaja GOSHOS	S-108, X-5, I-7, Zarya
28	Praleska	Fiber	Belarus	
29	Lider	Fiber	Smolenskaja GOSHOS	Tayga (France selection), mutagen treatment, selection
30	Voronezhskij 1308/138	Oil	VNIIMK	
31	Svetoch	Fiber	VNIIL	Selected from Cherskij kryazh
32	Ki-5	Oil	Ukraine	
33	LM-98	Oil	VNIIL	
34	Novotorzhskij	Fiber	VNIIL	L-1120, I-2, I-17, selected from Bogotolskij kryazh, G-2307, G-4523-6-13
35	TMP 1919 china 1	Fiber	china	
36	k-1147, local form	Oil	Ethiopia	
37	v-29	Oil	China	
38	Aleksim	Fiber	VNIIL	L-1120, T-4, T-5, T-10, Tekstilshhik, M-34, Pryadilshhik, Tvertsa
39	Ocean k-4497	Oil	France	
40	Orshanskij-2	Fiber	Belarus	I-16 × L-1120
41	K-6	Fiber	Pskovskij NIISH	L 1120 × T-5
42	Vih oil v-2	Oil	France	
43	Donskoj-95	Oil	Donskaja op.st.	
44	Diplomat	Fiber	VNIIL	Viking × Fibra (with subsequent selection)
45	Ford	Fiber	Belarus	
46	Tomskij-16	Fiber	Sibirskij NIISH and T	T-9 × G-1077

Table 2: The accessions of studied species of genus *Linum*.

Number	Accession name	Source	Genebank number	Origin	Chromosome number (2n)
			Sect. *Adenolinum*		
1	*L. perenne* L.	IPK	LIN 1807	Russian Federation	2n = 18 [14]
2	*L.* perenne L. subsp. *extraaxillare* (Kit.) Nyman	IPK	LIN 1651	Poland	2n = 36 [17]
3	*L. altaicum* Ledeb. ex Juz.	IPK	LIN 1632	Unknown	2n = 18 [17]
4	*L. komarovii* Juz.	IPK	LIN 1716	Unknown	2n = 18 [17]
5	*L. perenne* L. subsp. *perenne*	IPK	LIN 1521	Slovakia	2n = 18 [17]
6	*L.* perenne L. subsp. *alpinum* (Jacq.) Stoj. and Stef.	IPK	LIN 1905	Austria	2n = 18 [17]
7	*L. perenne* L. subsp. *anglicum* (Mill.) Ockendon	IPK	LIN 1524	GB	2n = 36 [17]
8	*L. perenne* L.	VNIIL	K 5500	Unknown	2n = 36 [17]
9	*L. leonii* F. W. Schultz	IPK	LIN 1672	Germany	2n = 18 [14]
10	*L. pallescens* Bunge	IPK	LIN 1645	Tajikistan	2n = 18 [17]
11	*L. pallescens* Bunge	Wild population, collected by N. L. Bolsheva		Russian Federation, Altai	2n = 18 (present study)
12	*L. mesostylum* Juz.	IPK	LIN 1774	Tajikistan	2n = 18 [17]
13	*L. mesostylum* Juz.	IPK	LIN 1662	Tajikistan	2n = 18 [17]
14	*L. lewisii* Pursh	IPK	LIN 1648	USA	2n = 18 [17]
15	*L. lewisii* Pursh	IPK	LIN 1550	USA	2n = 18 [17]
16	*L. austriacum* L.	IPK	LIN 1608	Germany	2n = 18 [17]
17	*L. austriacum* L.	Wild population, collected by A. A. Svetlova		Russian Federation, Rostov	2n = 18 (present study)
18	*L. austriacum* L. subsp. *euxinum* (Juz.) Ockendon	IPK	LIN 1546	Ukraine	2n = 18 [17]
19	*L. austriacum* L.	Wild population, collected by A. A. Svetlova		Crimea	2n = 18 (present study)
20	*L. austriacum* L.	Wild population, collected by A. A. Svetlova		Ukraine	2n = 18 (present study)
21	*L. amurense* Alef.	BGI	outdoors cultivated collection	Russian Federation, Far East, near Pokrovka settlement	2n= 18 [17]
			Sect. *Dasylinum*		
22	*L. hirsutum* L. subsp. *hirsutum* L.	IPK	LIN 1676	Hungary	2n = 16 [35]
23	*L. hirsutum* subsp. *hirsutum* L.	IPK	LIN1649	Romania	2n = 16 [36]
24	*L. hirsutum* L. subsp. *pseudoanatolicum* P.H.Davis	Wild population, collected by M. Pavelka		Turkey, Karaman	2n = 32 (present study)
25	*L. hirsutum* L. subsp. *anatolicum* (Boiss.) Hayek	Wild population, collected by M. Pavelka		Turkey, Aksaray	2n = 32 (present study)
			Sect. *Linum*		
26	*L. marginale* A.Cunn. ex Planch	IPK	LIN 1920	Australia	2n = 84 (present study)
27	*L. narbonense* L.	IPK	LIN 2002	Unknown	2n = 28 [14]
28	*L. narbonense* L.	IPK	LIN1653	France	2n = 28 [14]
29	*L. decumbens* Desf.	IPK	LIN 1754	Italy	2n = 16 [14]
30	*L. decumbens* Desf.	IPK	LIN1913	Italy	2n = 16 [14]

Table 2: Continued.

Number	Accession name	Source	Genebank number	Origin		Chromosome number (2n)
31	*L. grandiflorum* Desf.	IPK	LIN 2000	Unknown		2n = 16 [14]
32	*L. grandiflorum* Desf.	IPK	LIN 974	Unknown		2n = 16 [14]
33	*L. angustifolium* Huds.	IPK	LIN 1692	France		2n = 30 (present study)
34	*L. angustifolium* Huds.	VNIIL	K 5695	Unknown		2n = 30 (present study)
35	*L. angustifolium* Huds.	VNIIL	K 3108	Belgium		2n = 30 (present study)
36	*L. angustifolium* Huds	IGC	15	Belarus		2n = 30 [14]
37	*L. angustifolium* Huds.	VNIIL	K 4731	East Germany		2n = 30 (present study)
38	*L. bienne* Mill. (syn. *L. angustifolium* Huds.)	IGC	14	Hungary		2n = 30 [14]
39	*L. usitatissimum* L. subsp. *bienne* Mill. Snankev.	VIR	u-303794	Unknown	winter flax	2n = 30 (present study)
40	*L. usitatissimum convar.crepitans* [Boenningh.] Kulpa et Danert	IGC	260	Belarus	dehiscent flax	2n = 30 (present study)
41	*L. usitatissimum convar.crepitans* [Boenningh.] Kulpa et Danert	IPK	LIN 119	Portugal	dehiscent flax	2n = 30 (present study)
42	*L. usitatissimum* L. variety Giza	IPK	LIN 277	Egypt	large seeded flax	2n = 30 (present study)
43	*L. usitatissimum* L. G. 12 Ruzokvety	IPK	LIN 633	Czechoslovakia	dual-purpose flax	2n = 30 (present study)
44	*L. usitatissimum* L.,variety Colchidsky 1	VIR	u 099845.	Abkhazia	winter flax	2n = 30 (present study)
45	*L. usitatissimum* L.	VNIIL	k 7131	Morocco	large seeded flax	2n = 30 (present study)
			Sect. *Stellerolinum*			
46	*L. stelleroides* Planchon	BGI	Outdoors cultivated collection	Russian Federation, Far East, Telyakovsky Inlet		2n = 20 (present study)
47	*L. stelleroides* Planchon	BGI	Outdoors cultivated collection	Russian Federation, Far East, near Kraskino settlement		2n = 20 (present study)

was treated with 10 U of *TaqI* restriction enzyme (Thermo Scientific, USA) for 3 h at 65°C. Ligation was performed by adding 2.5 U of T4 DNA-ligase (Thermo Scientific, USA), 5 mM ATP, and 50 pmol of double-strand adapter [5′-ACTCGATTCTCAACCCGAAAGTATAGATCCCA; 5′-PO_4 -CGTGGGATCTATACTT-(C6linker)-NH_2] and incubation for 7 h at 37°C. The product was diluted twice with deionized water. The DNA sequence between the LTR (long terminal repeat) region of retrotransposons and the $Taq^{\alpha}I$ restriction site was amplified using the adapter primer (5′-GTTTACTCGATTCTCAACCCGAAAG 3′) and primers to the LTR regions of *FL1a*, *FL1b*, *FL4*, *Cassandra*, *FL10*, *FL8*, *FL7*, *FL12*, and *FL9* retrotransposons [32]:

1826 5′-ACCCCTTGAGCTAACTTTTGGGGTAAG-3′ (*FL1a, FL1b*)

1833 5′-CTTGCTGGAAAGTGTGTGAGAGG-3′ (*FL4*)

1838 5′-TGTTAATCGCGCTCGGGTGGGAGCA-3′ (*FL1a, FL1b, Cassandra*)

1845 5′-AGCCTGAAAGTGTTGGGTTGTCG-3′ (*FL11*)

1846 5′-CTGGCATTTCCATTGTCGTCGATGC-3′ (FL10)

1854 5′-GCATCAGCCTGGACCAGTCCTCGTCC-3′ (*FL8*)

1868 5′-CACTTCAAATTTTGGCAGCAGCGGATC-3′ (*FL1a, FL1b*)

1881 5′-TCGAGGTACACCTCGACTCAGG-3′ (*FL7*)

1886 5′-ATTCTCGTCCGCTGCGCCCCTACA-3′ (*FL12*)

1899 5′-TGAGTTGCAGGTCCAGGCATCA-3′ (*FL9*)

Amplification was performed in two stages. At the first stage, only the primer to LTR of the retrotransposon was used. Amplification was carried out in 25 μL of PCR mix containing 5 μL of ligation mix, 1 U of TrueStart Hot Start *Taq* DNA polymerase (Thermo Scientific, USA), TrueStart *Taq* DNA polymerase buffer, 0.5 mM $MgCl_2$, 20 μM dNTP (Thermo Scientific), and 5 pmol of the LTR primer. The program for amplification for the first stage was 95°C for 15 min, 30 cycles (95°C for 30 s, 62°C for 1 min, 72°C for 2 min) and 72°C for 10 min. At the second stage of amplification, the adapter primer 5′-GTTTACTCGATTCTCAACCCGA-3′ and one of the primers to the LTR region of retrotransposons were used. Amplification was carried out in 25 μL of PCR mix containing 12 μL of the first-stage PCR product, 1 U of *Taq* DNA polymerase (Thermo Scientific, USA), *Taq* DNA polymerase buffer, 1.5 mM $MgCl_2$, 200 μM dNTP (Thermo Scientific, USA), 25 pmol of LTR primer, and 25 pmol of the adapter primer. The program for amplification for the second stage was 95°C for 15 min, 35 cycles (95°C for 30 s, 62°C for 1 min, 72°C for 2 min), and 72°C for 10 min. The PCR products were separated in 2.5% agarose gel using TBE buffer and then stained with ethidium bromide. Ten PCR products were excised from agarose gel and characterized by sequencing on Applied Biosystems 3730 DNA Analyzer to confirm the specificity of SSAP reaction. The Bio-Rad Gel Doc system was used for gel documentation and photography as well as for visual detection of presence or absence of polymorphic fragments in the samples from different accessions. These data were recorded in the form of a binary matrix in which the presence of a fragment was coded as 1 and its absence as 0.

The genetic distances between varieties were calculated based on the binary matrix of amplified fragments using Dice's formula [41]. The dendrograms were constructed using SplitsTree 4.10 software [42]. Cluster analysis was performed using neighbor-joining method [43] and bootstrap values were determined based on 5000 permutations.

2.4. Preparation of Cassandra Retrotransposon DNA Probe. PCR primers were designed for amplification of *Cassandra* retrotransposon. We amplified an internal domain (primers IntDom-F AGTGGTATCCGAGCCTCT and IntDom-R CCCATAGGACTCAACGTC) and the LTR with the exception of the 5S rDNA region of this retrotransposon (primers LTR-1-86-F TGTAATGTAACACGTTAGGCA and LTR-1-86-R TTAGTTAGGGACGGATTGTT; LTR-206-279-F AAATAAATCTGTGAGGGATTAGT and LTR-206-279-R ACTTGTAACACCCCGTACT). The amplification was carried out in 20 μL of PCR mixture that contained 1 U of *TaqF* DNA polymerase (Amplisens, Russia), 1x *TaqF* buffer, 25 pmol of the forward and reverse primer, 200 μM dNTP (Amplisens, Russia), and 10 ng of genomic DNA. The program for amplification was 95°C for 15 min, 40 cycles (95°C for 10 s, 62°C for 20 s, 72°C for 30 s), and 72°C for 10 min. The amplicons were analyzed in 2% agarose gel and then used as a template for biotin PCR labeling to obtain biotin-labeled probes for FISH. PCR labeling was carried out using Biotin PCR Labeling Core Kit (Jena Bioscience, Germany) according to the manufacture's protocol. Labeled PCR products were precipitated with ethanol.

2.5. FISH with Cassandra Retrotransposon DNA Probe. Chromosome preparation was carried out according to the technique developed earlier for plants having small-sized chromosomes [14]. The hybridization mixture contained 2x SSC, 50% formamide, 10% dextran sulphate, and 2 ng/μL of a biotinylated DNA probe of *Cassandra* retrotransposon. The probe was hybridized overnight at 31°C. After hybridization the slides were washed twice with 0.1x SSC at 38°C for 10 min, followed by two washes with 2x SSC at 44°C for 5 min and a final 5 min wash in 2x SSC at room temperature. The biotin-labeled DNA probe was detected using a highly sensitive Alexa Fluor 488, Tyramide Signal Amplification system (Invitrogen) according to manufacturer's instructions.

3. Results

3.1. Analysis of SSAP Fingerprints. For analyzing of the reproducibility of the SSAP method, DNA of flax variety

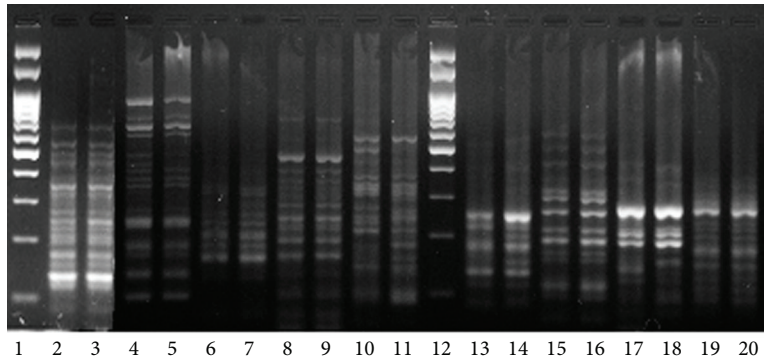


Figure 1: The test of reproducibility of SSAP markers obtained for "Stormont cirrus" flax variety. Lanes 2 and 3, primer 1899; lanes 4 and 5, primer 1826; lanes 6 and 7, primer 1838; lanes 8 and 9, primer 1845; lanes 10 and 11, primer 1846; lanes 13 and 14, primer 1854; lanes 15 and 16, primer 1868; lanes 17 and 18, primer 1881; lanes 19 and 20, primer 1886; lanes 1 and 12, 100 bp DNA ladder.

"Stormont cirrus" was restricted twice, ligated, and amplified with the primers. The obtained PCR products were visualized in acrylamide and agarose gels (Figure 1). Electrophoretic spectra of PCR products obtained with primers 1826, 1838, 1845, 1868, 1886, and 1899 coincided completely demonstrating high reproducibility of the SSAP results. The fingerprints obtained with primers 1846, 1854, and 1881 varied in individual amplified fragments, so the use of primers 1846, 1854, and 1881 for SSAP analysis of flax varieties will require further optimization of restriction, ligation, or PCR conditions. We selected primers with high reproducibility (1838, 1845, 1868, and 1899) as they yielded PCR products that were easily discernible in an agarose gel. These primers were used for analysis of 46 flax varieties by the SSAP method.

All the examined varieties produced identical or very similar fingerprints with primers 1838 and 1868 (*FL1a*, *FL1b*, and *Cassandra*). At the same time, the PCR products obtained with primers 1845 and 1899 were unique for different flax varieties. So, primers 1845 and 1899 were chosen for analyzing of genetic diversity of flax varieties. Several of PCR products have been sequenced (Supplementary Material available online at http://dx.doi.org/10.1155/2014/231589), the majority of obtained sequences are significantly similar to the sequences of corresponding retrotransposons.

3.2. Analysis of Genetic Diversity of Flax Varieties. Visual analysis of SSAP fingerprints based on retrotransposons *FL11* and *FL9* revealed 44 polymorphic retrotransposon insertions (23 fragments for primer 1845 and 21 fragments for primer 1899) in 46 flax varieties. Each of the 46 varieties had their own unique spectrum of retrotransposon insertions. So, we could differentiate all the 46 varieties using only two SSAP primers. In order to analyze genetic diversity of these varieties, we compiled a binary matrix of the presence/absence of polymorphic insertions of the above-mentioned retrotransposons, calculated the genetic distances between the varieties using Dice's formula [41], and constructed a dendrogram by using the neighbor-joining method (Figure 2). The obtained tree branching pattern revealed no distinct clusters among examined varieties.

3.3. Genomic Diversity of Species of the Genus Linum. For investigation of species of the genus *Linum*, we chose retrotransposons *FL1a*, *FL1b*, and *Cassandra* which did not show high insertion polymorphism within cultivated flax varieties. Primers 1838 and 1868 were used for SSAP analysis. As the result, 95 bands that originated with primer 1838 and 128 bands with primer 1868 (Figure 3) were scored. All the bands were polymorphic. Based on the SSAP fingerprint similarity, nine groups of closely related species (A-I) were distinguished. Group A included different species of sect. *Adenolinum* (syn. *L. perenne* group); group B consisted of *L. hirsutum* subsp. *hirsutum* accessions; groupC'' included *L. hirsutum* subsp. *pseudoanatolicum* and *L. hirsutum* subsp. *anatolicum*. Groups D, E, F, and G comprised species accessions of sect. *Linum* (*L. marginale; L. narbonense, L. decumbens, and L. grandiflorum*, resp.). Group H included accessions of *L. angustifolium* and *L usitatissimum* (sect. *Linum*); and group I consisted of *L. stelleroides* accessions (sect. *Stellerolinum*).

All the groups contained at least one group-specific marker. The results of phylogenetic analysis of *Linum* species are shown in the dendrogram on Figure 4. As the dendrogram shows nine clearly distinguished groups of species supported by high bootstrap values can be observed.

3.4. FISH with Cassandra Retrotransposon DNA Probe. The highly sensitive tyramide FISH method was applied for the investigation of abundanceof *Cassandra* retrotransposons as well as their distribution along chromosomes in three species of sect. *Linum* (*L. usitatissimum*, *L. grandiflorum,* and *L. narbonense*) and *L. amurense* (sect. *Adenolinum*). FISH revealed that *Cassandra* dispersed along the whole length of chromosomes in karyotypes of four studied species, but its distribution along the chromosomes was nonrandom (Figure 5). In species having small-sized chromosomes (*L. usitatissimum*, *L. grandiflorum,* and *L. amurense*), *Cassandra* was mainly localized in pericentromeric and subtelomeric chromosome regions. The patterns of *Cassandra* distribution were chromosome specific and were similar in homologous pairs of chromosomes (Figure 5(e)). In *L. narbonense*, which

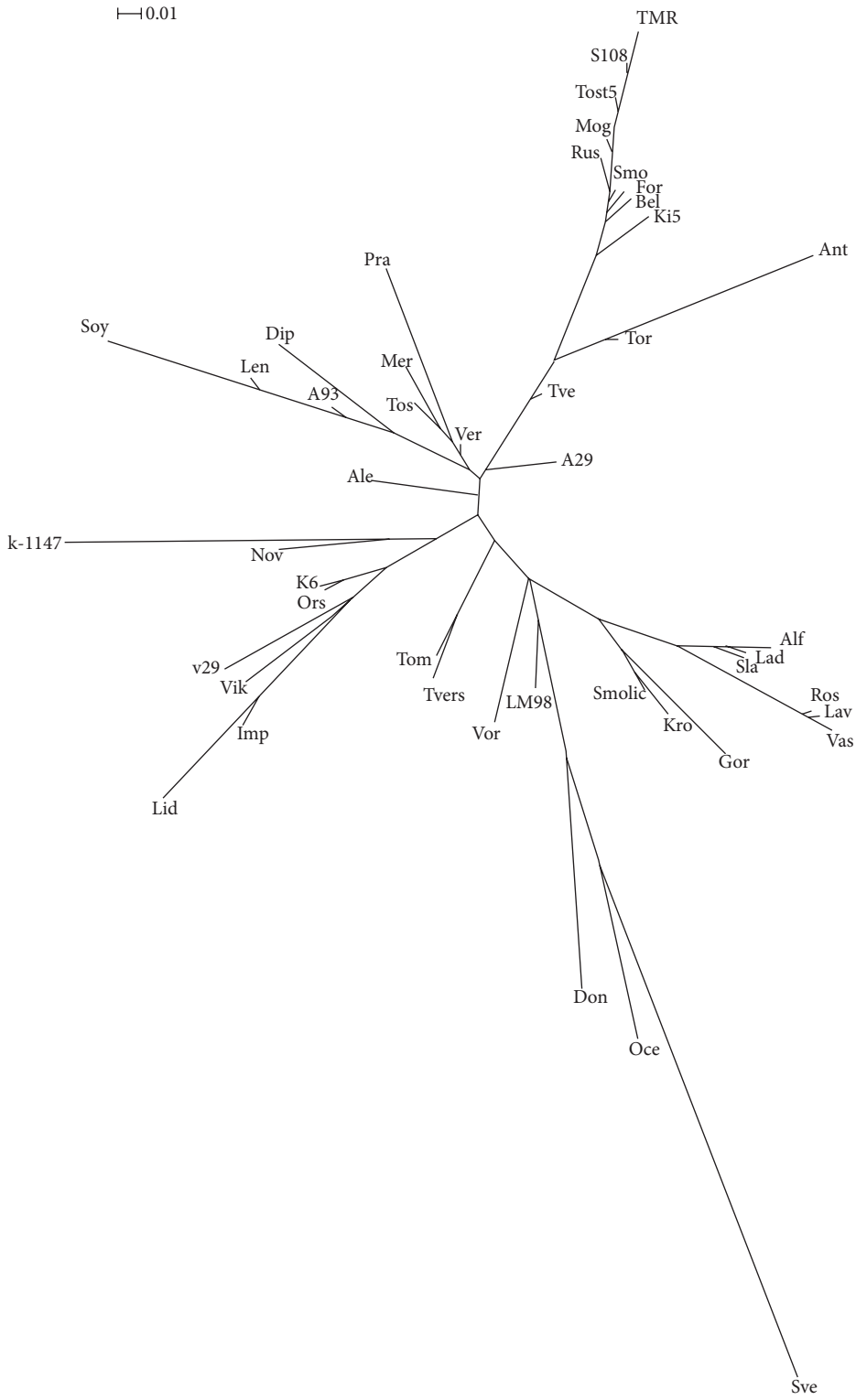


Figure 2: A neighbor-joining dendrogram for SSAP markers. Flax varieties: Alf-Alfa; Gor: Gorizont; Lav: Lavina; Vas: Vasilyok; Soy: Soyuz; Lad: Lada; Sla: Slavnyj 82; Kro: Krom; Tvers: Tverskoj; Ros: Rosinka; Len: Lenok; A93: A-93; Tos: Tost 3; Mog: Mogilevskij 2; Bel: Belochka; Tor: Torzhokskij 4; Tve: Tvertsa; Smolic: Smolich; A29: A-29; Ant: Antey; Rus: Rusich; Ver: Veralin; Mer: Merilin; Smo: Smolenskij; Tost5: Tost 5; S108: S 108; Imp: Impuls; Pra: Praleska; Lid: Lider; Vor: Voronezhskij; Sve: Svetoch; Ki5: Ki-5; Lm98: LM-98; Nov: Novotorzhskij; TMR: TMR-1919; k-1147: k-1147; v29: v-29; Ale: Aleksim; Oce: Ocean k-4497; Ors: Orshanskij 2; K6: K-6; Vic: Viksoil V-2; Don: Donskoj 95; Dip: Diplomat; For: Ford; Tom: Tomskij 16.

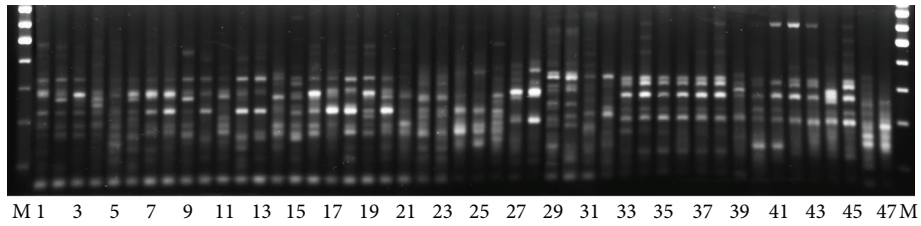


Figure 3: SSAP markers generated using 1868 primer for Linum species: 1: *L. perenne* LIN 1807; 2: *L. perenne* subsp. *extraaxilare*, LIN 1651; 3: *L. altaicum*, LIN 1632; 4: *L. komarovii* LIN 1716; 5: *L. perenne*, LIN 1521; 6: *L. perenne* susp. *alpinum*, LIN 1905; 7: *L. perenne* susp. *anglicum*, LIN 1524; 8: *L. perenne*, K 5500; 9: *L. leonii*, LIN 1672; 10: *L. pallescens*, LIN 1645; 11: *L. pallescens*, Altai; 12: *L. mesostylum*, LIN 1774; 13: *L. mesostylum*, LIN 1662; 14: *L. lewisii*, LIN 1648; 15: *L. lewisii*, LIN 1550; 16: *L. austriacum*, LIN 1608; 17: *L. austriacum*, Rostov; 18: *L. austriacum* subsp. *euxinum*, LIN 1546; 19: *L. austriacum*, Crimea; 20: *L. austriacum*, Ukraine; 21: *L. amurense*; 22: *L. hirsutum*, LIN 1676; 23: *L. hirsutum*, LIN1649; 24: *L. hirsutum* subsp. *pseudoanatolicum*; 25: *L. hirsutum* subsp. *anatolicum*; 26: *L. marginale*; 27: *L. narbonense*, LIN 2002; 28: *L. narbonense*, LIN1653; 29: *L. decumbens*, LIN 1754; 30: *L. decumbens*, LIN1913; 31: *L. grandiflorum*, LIN 2000; 32: *L. grandiflorum*, LIN 974; 33: *L. angustifolium*, LIN 1692; 34: *L. angustifolium*, K 5695; 35: *L. angustifolium*, K 3108; 36: *L. angustifolium,* Belarus; 37: *L. angustifolium*, K 4731; 38: *L. biene* (syn. *L. angustifolium*); 39: winter flax (*L. usitatissimum* subsp. *biene*); 40: dehiscent flax (*L. usitatissimum* convar. *crepitans*); 260; 41: dehiscent flax (*L. usitatissimum* convar. *crepitans*), LIN 119; 42: large seeded flax(*L. usitatissimum*), LIN 277; 43: dual-purpose flax (*L. usitatissimum*), LIN 633; 44: winter flax (*L. usitatissimum*), u 099845; 45: large seeded flax (*L. usitatissimum*), к 7131; 46: *L. stelleroides*, Telyakovsky Inlet; 47: *L. stelleroides*, Kraskino settlement. M—100 bp DNA ladder.

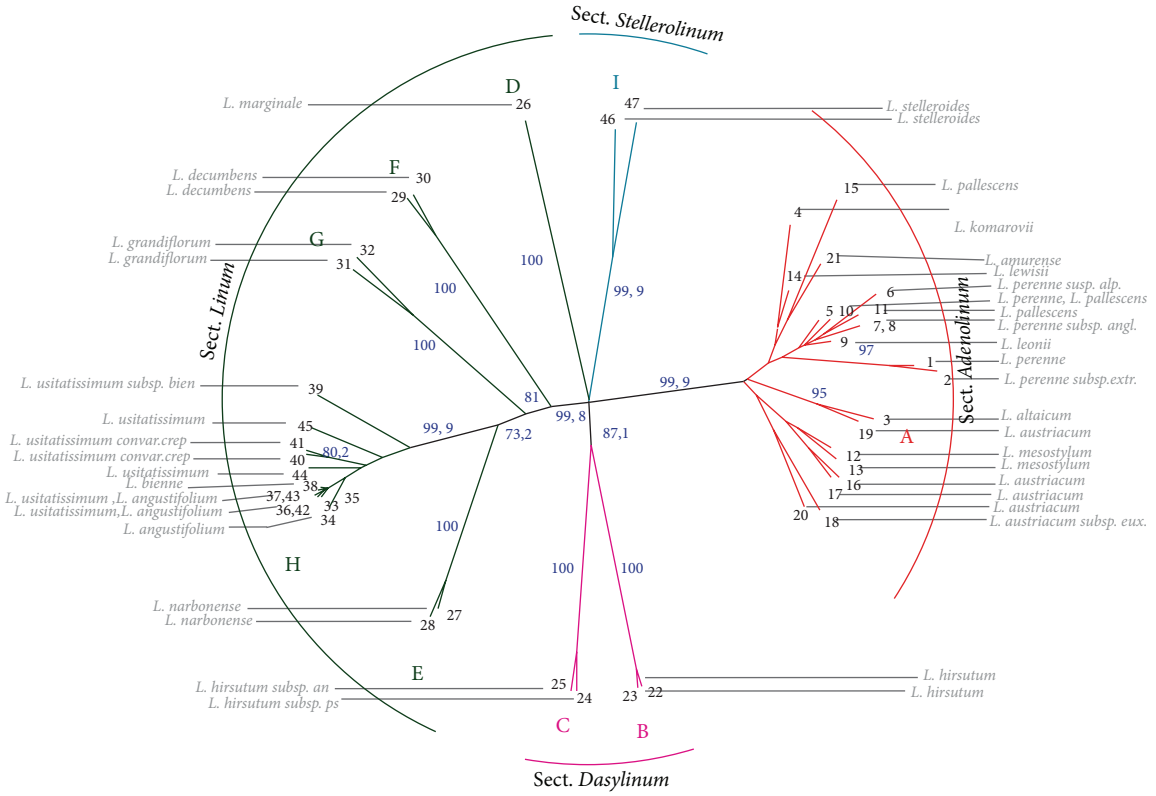


Figure 4: Neighbor-joining dendrogram for SSAP markers. Flax species: 1: *L. perenne* LIN 1807; 2: *L. perenne* subsp. *extraaxilare*, LIN 1651; 3: *L. altaicum*, LIN 1632; 4: *L. komarovii* LIN 1716; 5: *L. perenne*, LIN 1521; 6: *L. perenne* susp. *alpinum*, LIN 1905; 7: *L. perenne* susp. *anglicum*, LIN 1524; 8: *L. perenne*, K 5500; 9: *L. leonii*, LIN 1672; 10: *L. pallescens*, LIN 1645; 11: *L. pallescens*, Altai; 12: *L. mesostylum*, LIN 1774; 13: *L. mesostylum*, LIN 1662; 14: *L. lewisii*, LIN 1648; 15: *L. lewisii*, LIN 1550; 16: *L. austriacum*, LIN 1608; 17: *L. austriacum*, Rostov; 18: *L. austriacum* subsp. *euxinum*, LIN 1546; 19: *L. austriacum*, Crimea; 20: *L. austriacum*, Ukraine; 21: *L. amurense*; 22: *L. hirsutum*, LIN 1676; 23: *L. hirsutum*, LIN1649; 24: *L. hirsutum* subsp. *pseudoanatolicum*; 25: *L. hirsutum* subsp. *anatolicum*; 26: *L. marginale*; 27: *L. narbonense*, LIN 2002; 28: *L. narbonense*, LIN1653; 29: *L. decumbens*, LIN 1754; 30: *L. decumbens*, LIN1913; 31: *L. grandiflorum*, LIN 2000; 32: *L. grandiflorum*, LIN 974; 33: *L. angustifolium*, LIN 1692; 34: *L. angustifolium*, K 5695; 35: *L. angustifolium*, K 3108; 36: *L. angustifolium,* Belarus; 37: *L. angustifolium*, K 4731; 38: *L. biene* (syn. *L. angustifolium*); 39: winter flax (*L. usitatissimum* subsp. *biene*); 40: dehiscent flax (*L. usitatissimum* convar. *crepitans*); 260; 41: dehiscent flax (*L. usitatissimum* convar. *crepitans*), LIN 119; 42: large seeded flax (*L usitatissimum*), LIN 277; 43: dual-purpose flax (*L usitatissimum*), LIN 633; 44: winter flax (*L usitatissimum*), u 099845; 45: large seeded flax (*L usitatissimum*), к 7131; 46: *L. stelleroides*, Telyakovsky Inlet; 47: *L. stelleroides*, Kraskino settlement. Bootstrap values that exceeded 70% are shown in italic.

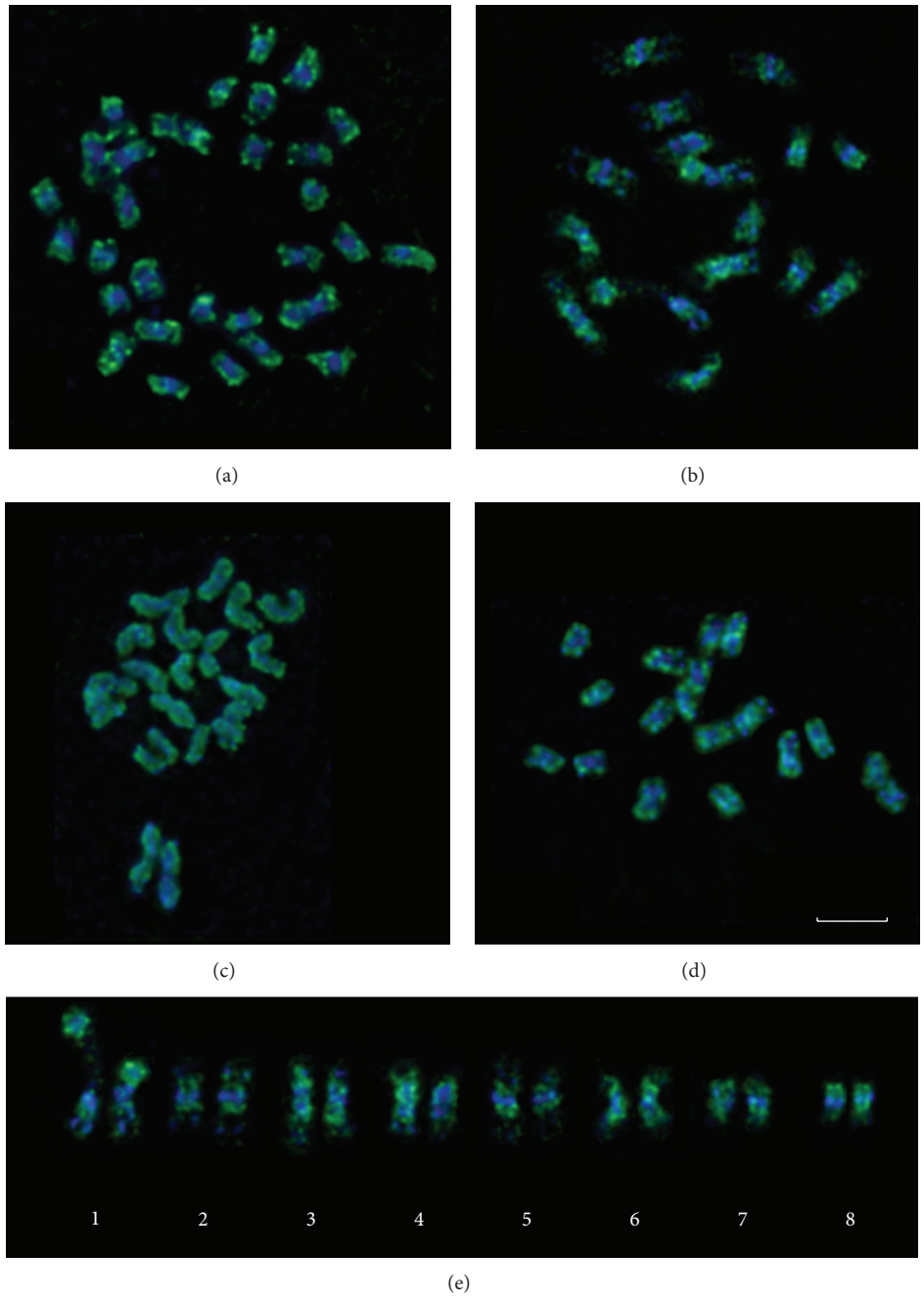


Figure 5: FISH with DNA probe of Cassandra retrotransposons (green). (a) *L. usitatissimum* (sect. *Linum*); (b) *L. grandiflorum* (sect. *Linum*); (c) *L. narbonense* (sect. *Linum*); (d) *L. amurense* (sect. *Adenolinum*); (e) a karyotype of *L. grandiflorum*. Chromosomes were stained with DAPI (blue). Bar—5 μm.

possessed large chromosomes, the patterns of *Cassandra* distribution resembled the patterns observed in karyotypes of the above-mentioned species (having small-sized chromosomes) though they were more regular.

4. Discussion

4.1. Use of SSAP Analysis for Identification of Flax Varieties and Estimation of Genetic Diversity. To estimate genetic polymorphism and to characterize *L. usitatissimum* varieties, the SSAP method was used, and also high reproducibility of the method was shown. The validity of the SSAP method for molecular genetic studies of flax varieties using *FL11* and *FL9* retrotransposons was demonstrated. Obtained with primers 1838 and 1868 (retrotransposons *FL1a*, *FL1b,* and *Cassandra*) PCR products were very similar in all the studied varieties. Such low diversity of cultivated flax might be a result of low transposition activity and/or creation of bottleneck effect during flax selection.

In analyzed flax varieties, 44 polymorphic insertions for *FL11* (primer 1845) and *FL9* (primer 1899) retrotransposons were revealed. Every studied variety possessed a unique set of SSAP markers. Therefore, the SSAP method can be used to mark the genotypes, to identify varieties of *L. usitatissimum* in genebank collections, to exercise control during of flax variety growth, and to obtain high quality seed material, when the varietal identity is particularly important.

The genetic similarity of 46 flax varieties was characterized by genetic distances calculated based on the SSAP data. The dendrogram (Figure 3) did not contain clearly isolated clusters of varieties. Thus, the studied flax varieties could not be subdivided into distinct groups. Our results were in agreement with earlier obtained data shown that flax accessions examined by IRAP analysis did not form distinct clusters in studies of their origin or the type of commercial use (fiber or oil). These data indicated an overlap in genetic diversity despite of disruptive selection for fiber or seed oil types [32]. In our study, the SSAP method also failed to distinguish fiber or oil seed flax varieties. Since varieties with the best characteristics are commonly used as parents in breeding practice, some valuable flax forms present in the genealogy of most modern varieties. Besides, the lines selected for crossing are usually characterized by low genetic diversity. So, the commercial flax varieties were shown to be less diverse than wild flax species and landraces [32].

Although the examined flax varieties could not be clustered into different groups by SSAP method, it might be used for estimation of their genetic similarity based on polymorphic insertions of retrotransposons. The estimation can be used for choosing the parents in breeding practice and also for creation of core collections which should include genetically diverse accessions.

4.2. Diversity and Phylogeny of Linum Species. In the present study, 20 accessions from sect. *Linum*, 21 accessions from sect. *Adenolinum*, 4 accessions from sect. *Dasylinum*, and 2 accessions from sect. *Stellerolinum* were analyzed by using SSAP method. All the examined species were clustered into 9 groups mainly according to common taxonomic division of the genus *Linum* into sections (Figure 4).

4.3. Section Dasylinum. Species from sect. *Dasylinum* clustered together and formed two related groups B and C. Group B included *L. hirsutum* subsp. *pseudoanatolicum* and *L. hirsutum* subsp. *anatolicum* and group C included *L. hirsutum* subsp. *hirsutum*. Thus, SSAP analysis singled out sect. *Dasylinum* as a well-supported clade. Our results were in agreement with the AFLP and ITS data as well as chloroplast phylogenies, chromosome studies, and transcriptome analysis of *Linum* species [4, 13, 37, 44]. It should be mentioned that the subdivision of accessions of sect. *Dasylinum* into two related clusters correlated with their difference in chromosome numbers and the origin of accessions. Thus, the accessions of *L. hirsutum* subsp. *hirsutum* (cluster C) from Europe was characterized by chromosome number of 2n = 16, while accessions from Turkey, *L. hirsutum* subsp. *pseudoanatolicum* and *L. hirsutum* subsp. *anatolicum* (cluster B) have chromosome number 2n = 32. Chromosome numbers for *L. hirsutum* subsp. *pseudoanatolicum* and *L. hirsutum* subsp. *anatolicum* were firstly determined in the present study.

4.4. Section Stellerolinum. Two accessions of *L. stelleroides* have rather similar SSAP fingerprints which were differed significantly from all the others *Linum* species and formed a separated clade. The similar results were obtained by phylogenetic analyses of chloroplast and ITS DNA sequences [4]. Moreover, *L. stelleroides* was shown to have chromosome number 2n = 20 which was unique for blue-flowered flaxes [45].

4.5. Section Adenolinum. All the members of sect. *Adenolinum* formed an independent group clustered separately from species of sect. *Linum* and other sections. Distinct isolation of this species group was also revealed in several molecular and karyological investigations [4, 12–14, 17]. The data were in good agreement with the opinion of Yuzepchuk [2] and Egorova [3] who isolated the group from sect. *Linum* into an independent section *Adenolinum*.

SSAP fingerprints of the accessions within sect. *Adenolinum* were highly polymorphic, but SSAP markers did not allow us to reveal any species subclusters supported by a high bootstrap value. Thus, SSAP analysis used in the present study as well as AFLP and RAPD analyses [13, 17] separated individual accessions but did not identify individual species inside sect. *Adenolinum*.

4.6. Section Linum. Sect. *Linum* was subdivided into 5 groups by neighbor-joining clustering. Accessions of *L. marginale*, *L. grandiflorum*, *L. decumben*, and *L. narbonense* formed four independent single species groups, while the fifth group combined accessions *L. angustifolium* and *L. usitatissimum*. Similar results had been obtained earlier by AFLP, RAPD, molecular phylogeny based on chloroplast *RbsL* sequence, and molecular cytogenetic methods (C/DAPI-banding patterns and localization of rRNA genes on chromosomes) [4, 12–14].

Within a subgroup consisted of *L. usitatissimum* and *L. angustifolium*, the accessions of large seeded flax (breeding cultivar), dual-purpose flax, and *L. angustifolium* were rather similar. Their fingerprints did not differ significantly from fingerprints of studied 46 flax varieties. This data were in a good agreement with the suggestion that *L. angustifolium* was the progenitor of *L. usitatissimum* [5, 7].

The accession of large seeded flax landrace, the accessions of winter flax, and the accessions of dehiscent flax differed significantly from the other members of cluster H. Both accessions of dehiscent flax grouped together (supported by a high bootstrap value) and had species-specific SSAP markers.

It should be noted that flax accessions, which are genetically distant from modern flax cultivars, are particular important for flax breeding. The genetic diversity of cultivated flax decreased significantly during the last decades. It might lead to the lack of useful alleles in genomes of modern cultivars [13]. Therefore, introduction of new useful traits

from the ancient primitive forms of cultivated flax and wild species could increase the polymorphism of modern flax varieties. SSAP markers allowed us to identify the unique accessions which are important for the investigation of the history of flax domestication.

L. marginale, the last member of sect. *Linum*, is a wild flax native to Australia. We found that it had the maximal chromosome number (2n = 84) in the genus *Linum*. The number indicated a high level of ploidy of *L. marginale* genome. SSAP patterns of the species were significantly different compared with the other species of sect. *Linum*. Therefore, *L. marginale* clustered apart from the other species. The obtained results were in contradiction with ITS and chloroplast topologies which clustered the species together with *L. bienne* and *L. usitatissimum* [4]. Rogers [46] assumed that Australian and New Zealand species *L. marginale* Cunn. and *L. monogynum* Forst. were related to European species *L. hologynum* Reichenb. (sec. *Linum*). This assumption based on the fact that diploid chromosome number of *L. hologynum* (2n = 42) corresponded to haploid chromosome number of *L. monogynum* and *L. marginale*. Moreover, all the three species had fused styles and pantoporate pollen grains that were unusual for blue-flowered flaxes. Thus, all the above-mentioned data indicated that the phylogenic lineages of *L. marginale* need further investigation.

The data obtained in the present study, as well as the results of other molecular phylogenetic and chromosomal investigations, indicated that members of section *Linum* were not as closely related as members of other sections. Therefore, taxonomic revision of this section is desirable.

4.7. Chromosome Location of Cassandra Retrotransposon. As differences in SSAP fingerprints for several flax species were found, we decided to analyze the distribution of *Cassandra* retrotransposon along the chromosomes of *Linum* species. *Cassandra* is a terminal-repeat retrotransposon in miniature (TRIM) that carries conserved 5S rDNA sequences in its LTRs. *Cassandra* was found in a number of vascular plants [31]. In our work, we investigated the distribution of this retrotransposon along the chromosomes of *Linum* species using tyramide FISH. We revealed that *Cassandra* localized in pericentromeric and subtelomeric regions of chromosomes that was typical for transposable elements [47]. A more uniform distribution of *Cassandra* retrotransposon was found in *L. narbonense* in comparison with *L. usitatissimum*, *L. grandiflorum*, and *L. amurense*. It was probably due to a higher content of transposable elements correlated with a larger size of its chromosomes (therefore its genome).

5. Conclusions

The availability of LTR sequences of flax retrotransposons and high polymorphism of SSAP markers offer a promising potential for SSAP analysis of genus *Linum*. Applications of SSAP analysis, for example, evolutionary and phylogenetic studies, assessment of genetic diversity, accession identification, and search for exotic genepools of cultivated flax, could be applied to *L. usitatissimum* and other *Linum* species. SSAP analysis was shown to be very useful for characterization of flax varieties and identification of accession belonging to different species or sections and provided new information about of phylogenetic relationships within the genus *Linum*.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Nataliya V. Melnikova, Anna V. Kudryavtseva, and Alexander V. Zelenin contributed equally to this work.

Acknowledgments

The work was financially supported by the Russian Foundation for Basic Research (Grants 12-04-01469-a, 13-04-01770-a, and 14-08-01167); by Fundamental Research Program of the Russian Academy of Sciences "Dynamics of Plant, Animal and Human Genofonds"; and by the Ministry of Education and Science of the Russian Federation under state Contract 14.621.21.0001. Part of this work was performed at the EIMB RAS "Genome" center.

References

[1] D. J. Ockendon and S. M. Walters, *Linaceae*, Cambridge University Press, Cambridge, Mass, USA, 1968.

[2] S. A. Yuzepchuk, "Genus Linum—Linaceae Dumort," in *Flora SSSR (Flora of the Soviet Union)*, B. K. Shishkin, Ed., vol. 14, pp. 84–146, Izdatel'stvo Akademii Nauk, Leningrad, Russia, 1949.

[3] T. V. Egorova, *Genus Linum, Linaceae*, St Petersburg Publishing, 1996.

[4] J. McDill, M. Repplinger, B. B. Simpson, and J. W. Kadereit, "The phylogeny of Linum and Linaceae subfamily Linoideae, with implications for their systematics, biogeography, and evolution of heterostyly," *Systematic Botany*, vol. 34, no. 2, pp. 386–405, 2009.

[5] H. Helbaek, "Domestication of food plants in the old world," *Science*, vol. 130, no. 3372, pp. 365–372, 1959.

[6] W. van Zeist and J. A. H. Bakker-Heeres, "Evidence for linseed cultivation before 6000 bc," *Journal of Archaeological Science*, vol. 2, no. 3, pp. 215–219, 1975.

[7] A. Diederichsen and K. Hammer, "Variation of cultivated flax (Linum usitatissimum L. subsp. usitatissimum) and its wild progenitor pale flax (subsp. angustifolium (Huds.) Thell.)," *Genetic Resources and Crop Evolution*, vol. 42, no. 3, pp. 263–272, 1995.

[8] D. Zohary and M. Hopf, *Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley*, Oxford University Press, Oxford, UK, 2000.

[9] I. A. Sisov, *Flax*, Selhosgiz, Leningrad, Russia, 1955.

[10] N. M. Cernomorskaja and A. K. Stankevic, "K voprosu o vnutryvidovoj klassifikacii lna obyknovennogo (Linum usitatissimum L.). To the problem of intraspecific classification of common flax (Linum usitatissimum L.)," *Selekcija I Genetika Tehnicheskih Kultur*, vol. 113, pp. 53–63, 1987.

[11] A. Diederichsen, "Ex situ collections of cultivated flax (*Linum usitatissimum* L.) and other species of the genus *Linum* L.," *Genetic Resources and Crop Evolution*, vol. 54, no. 3, pp. 661–678, 2007.

[12] Y. Fu, G. Peterson, A. Diederichsen, and K. W. Richards, "RAPD analysis of genetic relationships of seven flax species in the genus *Linum* L," *Genetic Resources and Crop Evolution*, vol. 49, no. 3, pp. 253–259, 2002.

[13] J. Vromans, *Molecular genetic studies in flax (Linum usitatissimum L.) [Ph.D. thesis]*, Wageningen University, Wageningen, The Netherlands, 2006.

[14] O. V. Muravenko, O. Y. Yurkevich, N. L. Bolsheva et al., "Comparison of genomes of eight species of sections Linum and Adenolinum from the genus Linum based on chromosome banding, molecular markers and RAPD analysis," *Genetica*, vol. 135, no. 2, pp. 245–255, 2009.

[15] Y. Fu and R. G. Allaby, "Phylogenetic network of Linum species as revealed by non-coding chloroplast DNA sequences," *Genetic Resources and Crop Evolution*, vol. 57, no. 5, pp. 667–677, 2010.

[16] B. J. Soto-Cerda, H. U. Saavedra, C. N. Navarro, and P. M. Ortega, "Characterization of novel genic SSR markers in *Linum usitatissimum* (L.) and their transferability across eleven *Linum* species," *Electronic Journal of Biotechnology*, vol. 14, no. 2, 2011.

[17] O. Y. Yurkevich, A. A. Naumenko-Svetlova, N. L. Bolsheva et al., "Investigation of genome polymorphism and seed coat anatomy of species of section Adenolinum from the genus Linum," *Genetic Resources and Crop Evolution*, vol. 60, no. 2, pp. 661–676, 2013.

[18] J. L. Bennetzen, "Transposable element contributions to plant gene and genome evolution," *Plant Molecular Biology*, vol. 42, no. 1, pp. 251–269, 2000.

[19] C. Feschotte and E. J. Pritham, "DNA transposons and the evolution of eukaryotic genomes," *Annual Review of Genetics*, vol. 41, pp. 331–368, 2007.

[20] C. Feschotte, N. Jiang, and S. R. Wessler, "Plant transposable elements: where genetics meets genomics," *Nature Reviews Genetics*, vol. 3, no. 5, pp. 329–341, 2002.

[21] P. S. Schnable, D. Ware, R. S. Fulton et al., "The B73 maize genome: complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, 2009.

[22] R. Kalendar, A. J. Flavell, T. H. N. Ellis, T. Sjakste, C. Moisy, and A. H. Schulman, "Analysis of plant diversity with retrotransposon-based molecular markers," *Heredity*, vol. 106, no. 4, pp. 520–530, 2011.

[23] R. Waugh, K. McLean, A. J. Flavell et al., "Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP)," *Molecular and General Genetics*, vol. 253, no. 6, pp. 687–694, 1997.

[24] T. H. N. Ellis, S. J. Poyser, M. R. Knox, A. V. Vershinin, and M. J. Ambrose, "Polymorphism of insertion sites of Ty1-copia class retrotransposons and its use for linkage and diversity analysis in pea," *Molecular and General Genetics*, vol. 260, no. 1, pp. 9–19, 1998.

[25] R. A. Queen, B. M. Gribbon, C. James, P. Jack, and A. J. Flavell, "Retrotransposon-based molecular markers for linkage and genetic diversity analysis in wheat," *Molecular Genetics and Genomics*, vol. 271, no. 1, pp. 91–97, 2004.

[26] N. V. Melnikova, F. A. Konovalov, and A. M. Kudryavtsev, "Long terminal repeat retrotransposon Jeli provides multiple genetic markers for common wheat (Triticum aestivum)," *Plant Genetic Resources: Characterisation and Utilisation*, vol. 9, no. 2, pp. 163–165, 2011.

[27] H. Kim, S. Terakami, C. Nishitani et al., "Development of cultivar-specific DNA markers based on retrotransposon-based insertional polymorphism in Japanese pear," *Breeding Science*, vol. 62, no. 1, pp. 53–62, 2012.

[28] S. R. Pearce, M. Knox, T. H. N. Ellis, A. J. Flavell, and A. Kumar, "Pea Ty1-copia group retrotransposons: transpositional activity and use as markers to study genetic diversity in Pisum," *Molecular and General Genetics*, vol. 263, no. 6, pp. 898–907, 2000.

[29] A. V. Vershinin, T. R. Allnutt, M. R. Knox, M. J. Ambrose, and T. H. N. Ellis, "Transposable elements reveal the impact of introgression, rather than transposition, in Pisum diversity, evolution, and domestication," *Molecular Biology and Evolution*, vol. 20, no. 12, pp. 2067–2075, 2003.

[30] A. M. Sanz, S. G. Gonzalez, N. H. Syed, M. J. Suso, C. C. Saldaña, and A. J. Flavell, "Genetic diversity analysis in Vicia species using retrotransposon-based SSAP markers," *Molecular Genetics and Genomics*, vol. 278, no. 4, pp. 433–441, 2007.

[31] R. Kalendar, J. Tanskanen, W. Chang et al., "Cassandra retrotransposons carry independently transcribed 5S RNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 15, pp. 5833–5838, 2008.

[32] P. Smýkal, N. Bačová-Kerteszováč, R. Kalendar, J. Corander, A. H. Schulman, and M. Pavelek, "Genetic diversity of cultivated flax (Linum usitatissimum L.) germplasm assessed by retrotransposon-based markers," *Theoretical and Applied Genetics*, vol. 122, no. 7, pp. 1385–1397, 2011.

[33] R. Ragupathy, R. Rathinavelu, and S. Cloutier, "Physical mapping and BAC-end sequence analysis provide initial insights into the flax (Linum usitatissimum L.) genome," *BMC Genomics*, vol. 12, article 217, 2011.

[34] L. G. González and M. K. Deyholos, "Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome," *BMC Genomics*, vol. 13, no. 1, article 644, 2012.

[35] I. V. Nosova, O. Y. Semenova, T. E. Samatadze et al., "Investigation of karyotype structure and mapping of ribosomal genes on chromosomes of wild linum species by FISH," *Biologicheskie Membrany*, vol. 22, no. 3, pp. 244–248, 2005.

[36] N. L. Bolsheva, O. Y. Semenova, O. V. Muravenko, I. V. Nosova, K. V. Popov, and A. V. Zelenin, "Localization of telomere sequences in chromosomes of two flax species," *Biologicheskie Membrany*, vol. 22, no. 3, pp. 227–231, 2005.

[37] O. V. Muravenko, N. L. Bol'sheva, O. I. Iurkevich et al., "Karyogenomics of species of the genus *Linum* L," *Genetika*, vol. 46, no. 10, pp. 1339–1342, 2010.

[38] F. A. Konovalov, N. P. Goncharov, S. Goryunova, A. Shaturova, T. Proshlyakova, and A. Kudryavtsev, "Molecular markers based on LTR retrotransposons BARE-1 and Jeli uncover different strata of evolutionary relationships in diploid wheats," *Molecular Genetics and Genomics*, vol. 283, no. 6, pp. 551–563, 2010.

[39] N. V. Melnikova, A. V. Kudryavtseva, A. S. Speranskaya et al., "The FaRE1 LTR-retrotransposon based SSAP markers reveal genetic polymorphism of strawberry (Fragaria x ananassa) cultivars," *Journal of Agricultural Science*, vol. 4, no. 11, pp. 111–118, 2012.

[40] K. Edwards, C. Johnstone, and C. Thompson, "A simple and rapid method for the preparation of plant genomic DNA for PCR analysis," *Nucleic Acids Research*, vol. 19, no. 6, p. 1349, 1991.

[41] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[42] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, 2006.

[43] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[44] S. Sveinsson, J. McDill, G. K. Wong et al., "Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (Linum) using transcriptomics," *Annals of Botany*, vol. 113, no. 5, pp. 753–761, 2014.

[45] A. P. Sokolovskaya and N. S. Probatova, "Chromosome numbers in the vascular plants from the Primorye territory, Kamchatka, region, Amur valley and Sakhalin," *Botanicheskii Zhurnal SSSR*, vol. 70, no. 4, pp. 997–999, 1985.

[46] C. M. Rogers, "A further note on the relationships of the European Linum hologynum and the Australian species of Linum (Linaceae)," *Plant Systematics and Evolution*, vol. 147, no. 3-4, pp. 327–328, 1984.

[47] J. S. P. Heslop-Harrison and T. Schwarzacher, "Organisation of the plant genome in chromosomes," *Plant Journal*, vol. 66, no. 1, pp. 18–33, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 620530, 11 pages
http://dx.doi.org/10.1155/2014/620530

Research Article

Binding Sites of miR-1273 Family on the mRNA of Target Genes

Anatoly Ivashchenko, Olga Berillo, Anna Pyrkova, and Raigul Niyazova

National Nanotechnology Laboratory, Al-Farabi Kazakh National University, Al-Farabi 71, Almaty 050038, Kazakhstan

Correspondence should be addressed to Anatoly Ivashchenko; a_ivashchenko@mail.ru

Received 14 April 2014; Revised 11 July 2014; Accepted 23 July 2014; Published 26 August 2014

Academic Editor: Vassily Lyubetsky

This study examined binding sites of 2,578 miRNAs in the mRNAs of 12,175 human genes using the MirTarget program. It found that the miRNAs of miR-1273 family have between 33 and 1,074 mRNA target genes, with a free hybridization energy of 90% or more of its maximum value. The miR-1273 family consists of miR-1273a, miR-1273c, miR-1273d, miR-1273e, miR-1273f, miR-1273g-3p, miR-1273g-5p, miR-1273h-3p, and miR-1273h-5p. Unique miRNAs (miR-1273e, miR-1273f, and miR-1273g-3p) have more than 400 target genes. We established 99 mRNA nucleotide sequences that contain arranged binding sites for the miR-1273 family. High conservation of each miRNA binding site in the mRNA of the target genes was found. The arranged binding sites of the miR-1273 family are located in the 5′UTR, CDS, or 3′UTR of many mRNAs. Five repeating sites containing some of the miR-1273 family's binding sites were found in the 3′UTR of several target genes. The oligonucleotide sequences of miR-1273 binding sites located in CDSs code for homologous amino acid sequences in the proteins of target genes. The biological role of unique miRNAs was also discussed.

1. Introduction

Once a microRNA (miRNA) has been discovered, the number of publications devoted to clarifying its biological role increases constantly and quickly [1]. Researchers are interested in miRNAs because they participate in the posttranscription regulation of gene expression [2]. These nanoscale molecules participate, directly or indirectly, in almost all key organism processes [1–3]. Identifying the target genes of a miRNA is an imperfect process, and some programs predict a large number of false-positive binding sites. Additionally, some papers have discussed the existence of miRNA binding sites only in the 3′-untranslated region (3′UTR) and the obligatory presence of a "seed" in the 5′ end of the miRNA, but these statements and others are poorly substantiated [4, 5]. The binding sites located in coding domain sequences (CDSs) of mRNAs appeared recently [6]. The process of establishing a miRNA's precise biological function is slow because they are poorly understood, despite the large number of publications devoted to them. Because miRNAs regulate gene expression, they participate in many pathological processes [7–17]. Changes in the miRNA concentration have been shown to occur during the development of breast [7], lung [8], esophageal [9], stomach [10], intestine [11], prostate [12], and other cancers [13–15]. Changes in the interactions between the miRNAs and mRNAs of oncogenes [16] and genes suppressors [17] have been shown to cause malignant diseases. Thus, it is necessary to clarify the role of miRNAs in disease development.

In this work, we studied the binding of 2,578 miRNAs with 12,175 mRNAs for genes. The majority of these genes participate in the development of lung cancer, breast cancer, gastrointestinal cancer, and others. First, it is necessary to determine the features of miRNA binding sites. One miRNA can bind to one or more mRNAs, and some mRNAs have multiple binding sites for different miRNAs that are within the same family. The expression of most human protein-coding genes depends directly or indirectly on more than 2,500 miRNAs. We must also establish whether the connections between the miRNAs and mRNAs are minor and only affect individual genes or whether they are organized to regulate system-wide gene expression. Specifically, the relationships between the binding sites of one family of miRNAs and all of the mRNA sites must be elucidated.

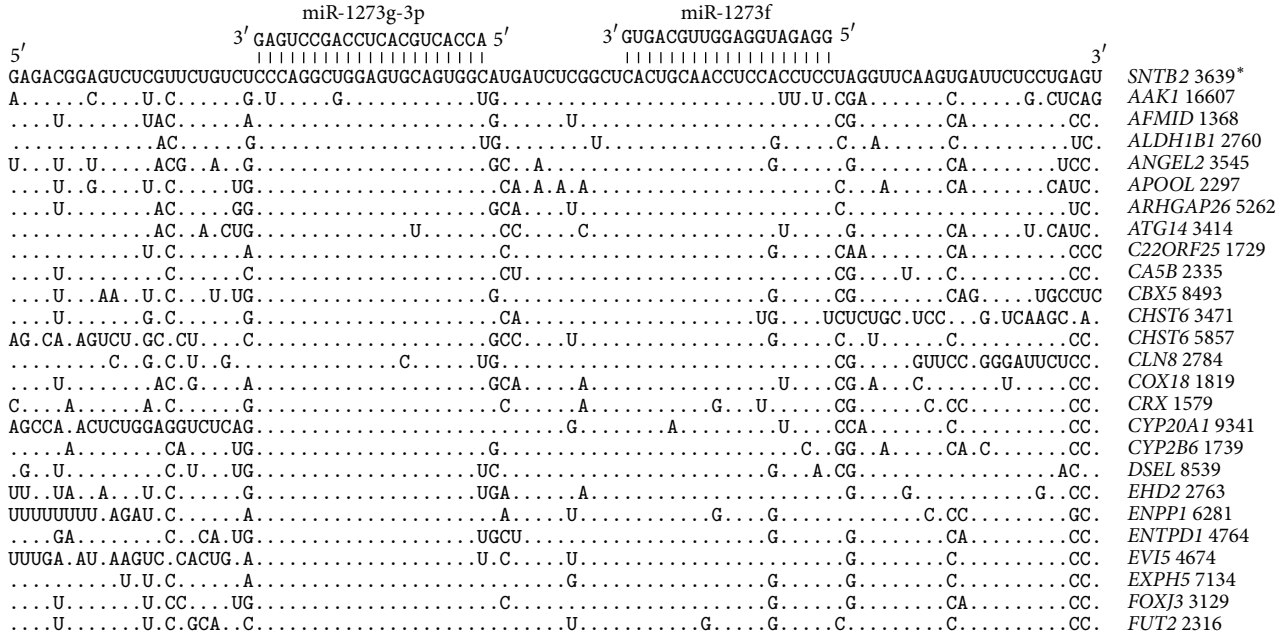


Figure 1: Arranged binding sites miR-1273g-3p and miR-1273f in 3′UTR mRNA target genes. Note Figures 1–11. Symbol | is hydrogen bonds between nucleotides miRNA and mRNA; ∗ is position of binding sites miR-1273g-3p on mRNA; (.) equals nucleotide.

2. Materials and Methods

Human miRNAs (hsa-miRNAs) were taken from the miRBase site (http://mirbase.org). The mRNAs for human genes were taken from the GenBank database (http://www.ncbi.nlm.nih.gov) using Lextractor002 script (http://sites.google.com/site/malaheenee/software). The target genes for the tested miRNAs were revealed using the MirTarget program, which was developed in our laboratory. This program defines the following features of binding: (a) the beginning of a miRNA binding with mRNAs; (b) the localization of miRNA binding sites in the 5′-untranslated regions (5′UTRs), CDSs and 3′UTRs of the mRNAs; (c) the free energy of hybridization (ΔG, kJ/mole); and (d) the schemes of nucleotide interactions between the miRNAs and the mRNAs. The ratio $\Delta G/\Delta G_m$ (%) was counted for each site, where ΔG_m equaled the free energy of a miRNA binding with its perfect complementary nucleotide sequence. The miRNA binding sites located on the mRNAs had $\Delta G/\Delta G_m$ ratios of 90% and more. We note the position of the binding sites on the mRNA, beginning from the first nucleotide of the mRNA's 5′UTR. It found bonds between adenine (A) and uracil (U), guanine (G) and cytosine (C), and G and U, as well as between A and C via a hydrogen bond [18]. The distance between A and C was equal to the G-C, A-U, and G-U distances [19]. The numbers of hydrogen bonds in the G-C, A-U, G-U, and A-C interactions were taken to be 3, 2, 1, and 1, respectively. The free binding energies of these nucleotide pairs were accepted as the same values (3 : 2 : 1 : 1).

3. Results and Discussion

3.1. Features of the miR-1273 Family. The binding powers between the 2,578 tested hsa-miRNAs and the mRNAs from 12,175 human genes were calculated. Some members of the miR-1273 family have a greater number of target genes than others. For example, miR-1273g-3p and miR-1273f can bind to 1,074 and 766 genes, respectively, with $\Delta G/\Delta G_m$ ratios of 90% and more. Other miRNAs have some target genes. For example, 1271-5p and 1271-3p have only six and nine target genes, respectively. The miRNAs with over 400 target genes were called unique miRNAs (umiRNAs). In addition, the binding sites for these unique miRNAs are unusually located in the mRNAs. Members of the miR-1273 family have different origins, lengths, quantities, and properties of the miRNA binding sites, among other features. Some characteristics of the miR-1273 family are outlined below.

With a length of 25 nt, miR-1273a is coded in an intron of the regulator of G-protein signaling 22 gene (*RGS22*), located on chromosome 8. We found that miR-1273a has 154 binding sites on 148 target mRNAs; thus, some of the mRNAs have two binding sites. Of those, 146 miR-1273a binding sites are located in 3′UTRs, six sites are located in 5′UTRs, and two sites are located in CDSs.

With a length of 22 nt, miR-1273c is coded in an intron of the T cell lymphoma invasion and metastasis 2 gene (*TIAM2*), located on chromosome 6. We found that 84 target gene mRNAs have one binding site for miR-1273c, while *GOLGA3* has 2 sites, for a total of 86 miR-1273c sites. Seven of those are located in 5′UTRs, two sites are located in CDSs, and 76 sites are located in 3′UTRs.

With a length of 25 nt, miR-1273d is coded in an intron of the Kinesin family member 1B gene (*KIF1B*), located on chromosome 1. We found that 114 target gene mRNAs have one binding site, while *ARGFX* mRNA has two sites, for a total of 116 miR-1273d sites. Six of those are located in 5′UTRs, five sites are located in CDSs, and 104 sites are located in 3′UTRs.

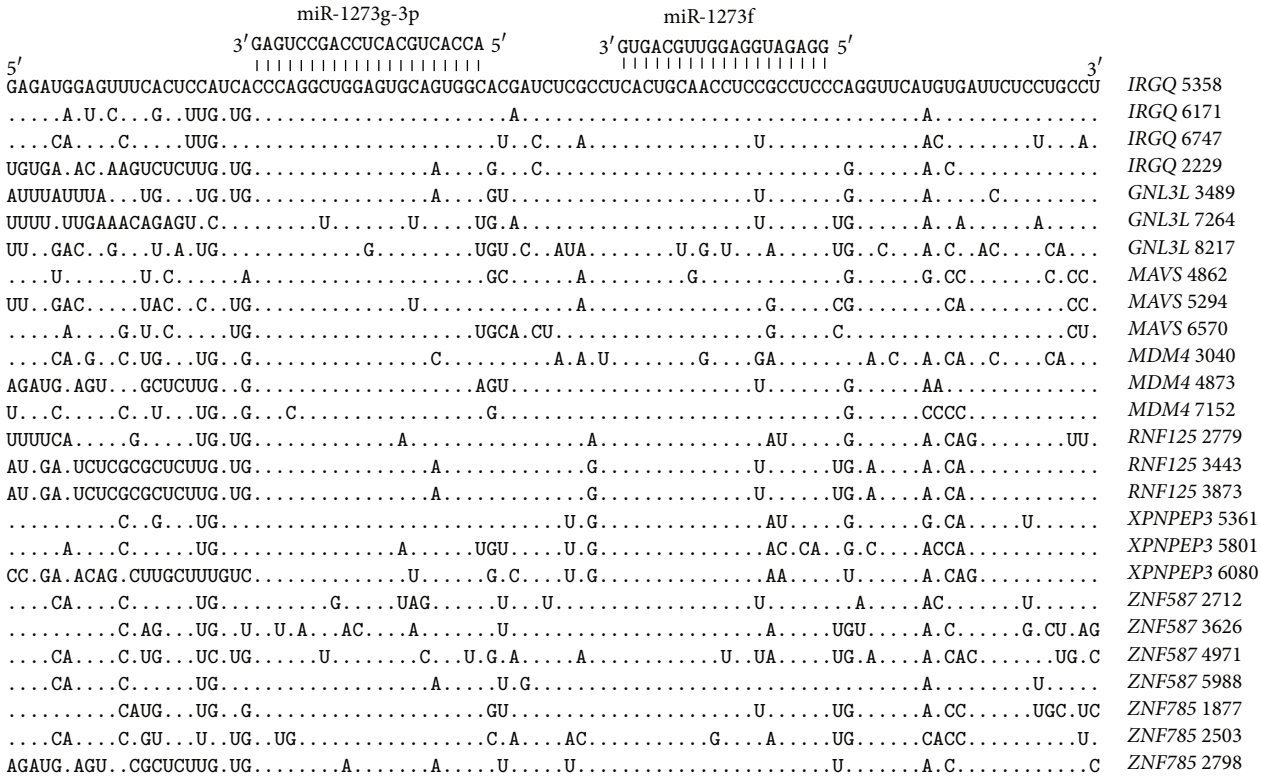


Figure 2: Arranged binding sites miR-1273g-3p and miR-1273f in 3′UTR mRNA that contain three and four pair binding sites.

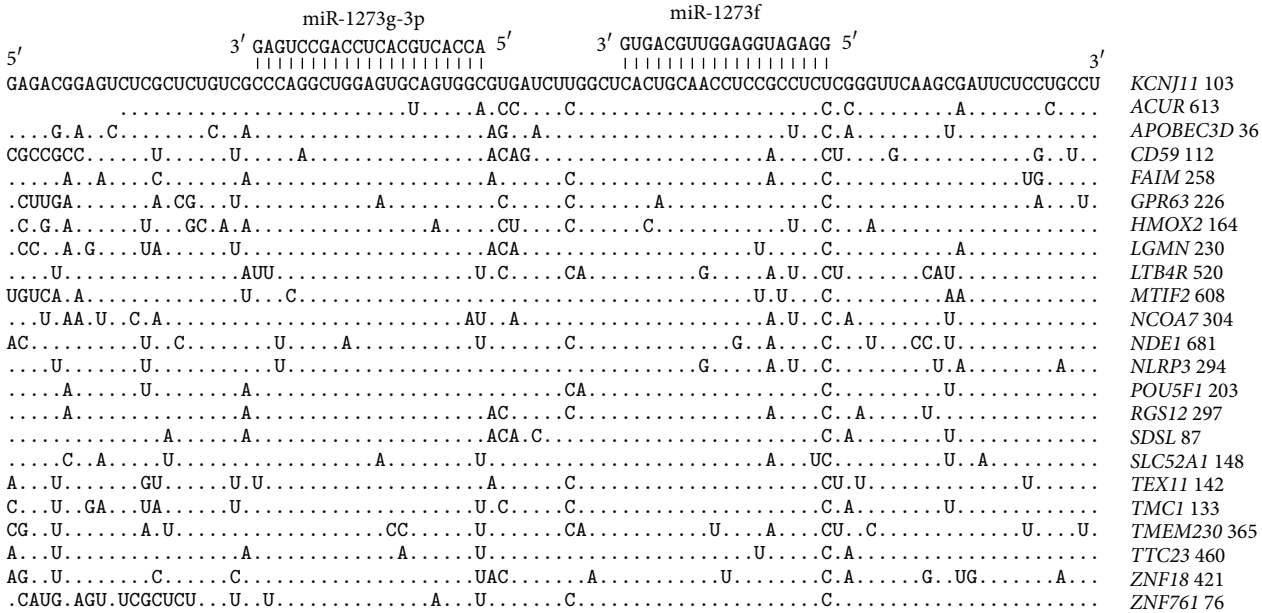


Figure 3: Arranged binding sites miR-1273g-3p and miR-1273f in 5′UTR mRNA target genes.

With a length of 22 nt, miR-1273e's origin was not established. We found 449 miR-1273e binding sites on the mRNAs of 413 target genes. Of those, 19 binding sites are located in 5′UTRs, nine sites are located in CDSs, and 421 sites are located in 3′UTRs.

With a length of 19 nt, miR-1273f is coded in an intron of the sterol carrier protein 2 (*SCP2*) gene, located on chromosome 1. We found that the mRNAs of 766 genes contain 886 miR-1273f binding sites. Of those, 45 sites are located in 5′UTRs, 40 sites are located in CDSs, and 801

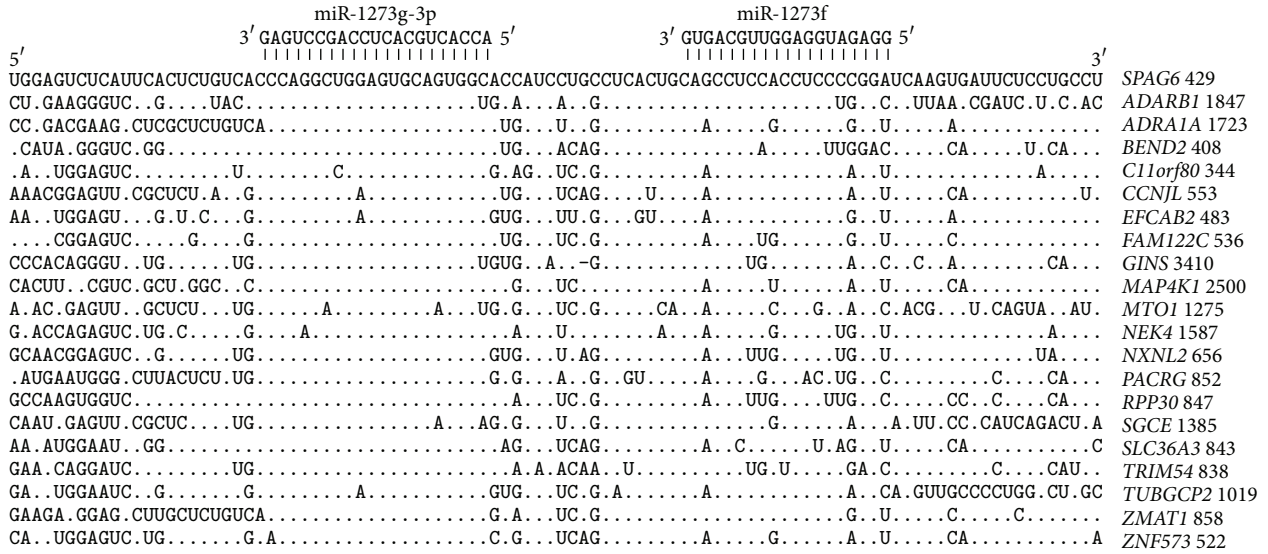


Figure 4: Arranged binding sites miR-1273g-3p and miR-1273f in CDS mRNA target genes.

QSLALSPKLECSGTILAHSNLRLLGSSDSPASASR *NEK4*
VSFTLSPRLECSGTILAHCSLHLPGSSDSPASVSQ *SPAG6*
QDLTLLPRLECSGTNTTYCSLDLPSSSDPPILASQ *TRIM54*
FRLLGSPRLECSGTISPHCNLLLPGSSNSPASASR *MAP4K1*
NGSYSLPRLECSGAIMARCNLDHLGSSDPPTSASQ *PACRG*
WNLALSPRLECSGKISAHCNPHLQGSSNSPAQASR *SLC36A3*
GVSLLSPRLKCSGMISAHCNLHLPGSSNSPASAPH *CCNJL*
RSLALLPRLECSGVILAHCNLCLLGSSDSLALASR *NXNL2*
RSLTVSPRLECSGMISAHCNLCLPGSSDSPASDSR *FAM122C*
TKSRSVTRLECSGMILAHCNLRLPGSRDSPASASQ *ADRA1A*
GVLLLLPRMECNGAISAHHNLPLPGYGVQYDYLDP *MTO1*
QGFALLPRLECSGVIWLTAALTSQAPEILPPQPPM *GINS3*
WSLTLLPRPECSGAVSAHCNLHLPGSSDSHASVPR *C11orf80*
MESCSVTRLECSGAISAHCSLHLPGSSDSPASASQ *ZMAT1*
MESCSVAQAGVQWPDLSSLQPPPPRFKQFSCHSLQ *ZNF573*
WSFAPVAQAGVQWSDLGSLQPPPPRNLPHQTQIPQ *SGCE*
YGSGSVTQAGVQWHDHSSLQPQPLGLKQFFHLSLP *BEND2*
KWSHSVTQAGVQWHNLGSLQPLPLGLKPSSHLSLP *RPP30*
EGSRSYTQAGVQWCNHGSLQPRPPGLLSDPSTSTF *ADARB1*

Figure 5: Amino acid sequences are coded by the segment of mRNA that corresponds to miR-1273g-3p and miR-1273f binding sites.

sites are located in 3′UTRs. The mRNAs of ten genes have completely complementary binding sites for miR-1273f. Each mRNA of the *GNL3L, IRGQ, ORAI2,* and *PLCXD1* genes has four miR-1273f binding sites that are located in 3′UTRs.

With a length of 21 nt, miR-1273g-3p is coded in an intron of the *SCP2* gene, located on chromosome 1. We found that miR-1273g-3p has 1,330 binding sites on 1,074 mRNAs. Of those, 69 miR-1273g-3p binding sites are located in 5′UTRs, 38 sites are located in CDSs, and 1,223 sites are located in 3′UTRs. The mRNAs of seven genes have completely complementary binding sites for miR-1273g-3p. The mRNAs of the *NOL9, PLCXD1, ZNF490, CYP20A1, GNL3L, PPM1K, RBMS2, SAR1B,* and *SLC35E2* genes have four binding sites. The *IRCQ* and *ZNF850* genes have five binding sites, and the mRNA of the *MDM4* gene has six miR-1273g-3p binding sites. All of these sites are located in 3′UTRs.

With a length of 22 nt, miR-1273g-5p is coded in an intron of the *SCP2* gene, located on chromosome 1. The mRNAs of 33 target genes have one miR-1273g-5p binding site. Two of those sites are located in 5′UTRs, five sites are located in CDSs, and 26 sites are located in 3′UTRs.

With a length of 21 nt, miR-1273h-3p is coded in the intergenic nucleotide sequence of chromosome 16. We found that miR-1273h-3p has 38 target genes. The mRNA of these target genes have only one miR-1273h-3p binding site. Three sites are located in 5′UTRs and 35 sites are located in 3′UTRs, but no sites were found in CDSs.

With a length of 21 nt, miR-1273h-5p is coded in the intergenic sequence of chromosome 16. We found that miR-1273h-5p has 127 binding sites on 126 target gene mRNAs. Eleven sites are located in 5′UTRs, 14 sites are located in CDSs, and 102 sites are located in 3′UTRs.

3.2. Arrangement of the miR-1273 Family's Binding Sites in the mRNA of Target Genes. This study revealed that several hundred mRNAs have homologous nucleotide sequences containing binding sites for members of the miR-1273 family. Two miRNA binding sites located on one mRNA were termed pair sites. Specifically, we examined pairs composed of miR-1273g-3p with another member of the miR-1273 family. Data about the localization of these pair sites are presented in the text below. These arranged pair sites are located in mRNA segments that have a length of just 99 nucleotides.

The mRNAs of 582 general target genes have pair sites for both miR-1273g-3p and miR-1273f. Of those, 24 mRNAs are located in 5′UTRs, 18 are located in CDSs, and 540 are located in 3′UTRs. The locations of the miR-1273g-3p and miR-1273f binding sites in the 3′UTRs of mRNAs are presented in Figure 1. The nucleotide sequence in the 3′UTR of the *SNTB2* gene that contained this pair binding site is chosen for

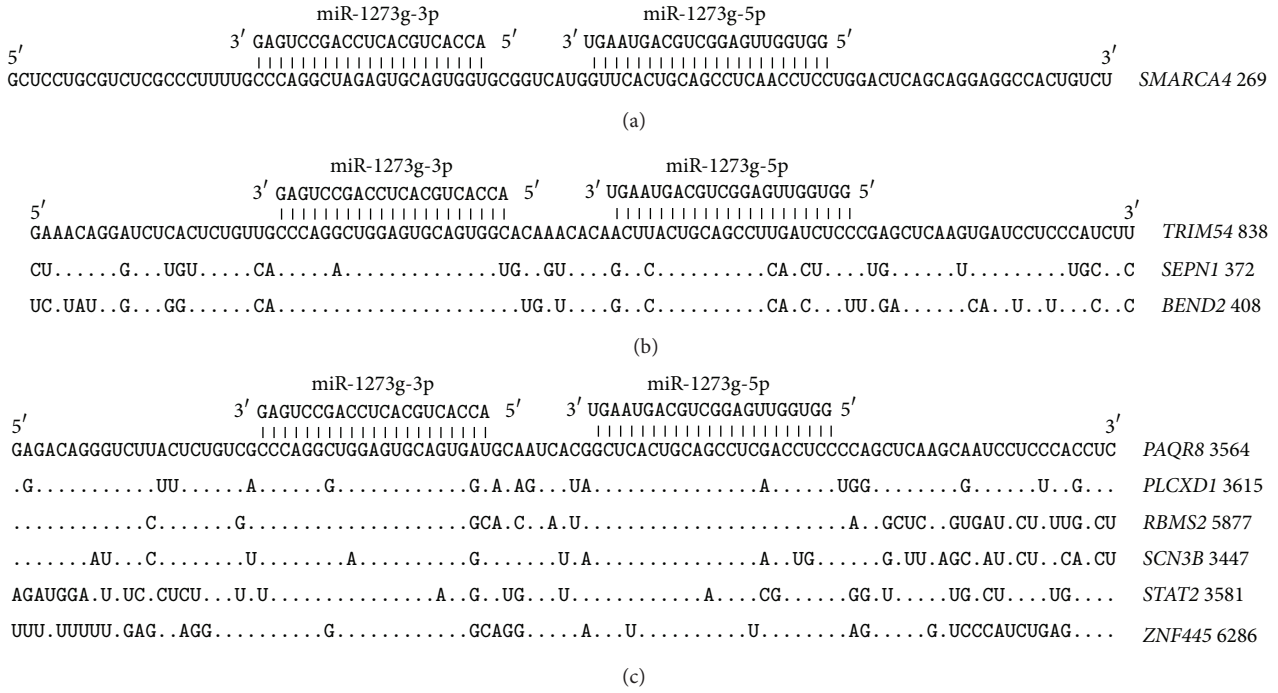


Figure 6: Arranged binding sites miR-1273g-3p and miR-1273g-5p in 5′UTR (a), CDS (b), and 3′UTR (c) mRNA target genes.

comparison with the pair sites of other mRNAs. Most binding sites have nucleotide replacements (purine to purine and pyrimidine to pyrimidine) to retain their hydrogen bonds. Figure 1 shows that the miR-1273g-3p and miR-1273f binding sites in all of the tested mRNAs are located at distance of 12 nucleotides. The nucleotide sequences of these revealed that pair sites are highly homologous, indicating that their origins are not casual.

The mRNAs of many genes contain two or more pair sites for miR-1273g-3p and miR-1273f. The nucleotide sequences of sites in mRNA 3′UTRs that contain three and four arranged pairs of sites for these two miRNAs are shown in Figure 2. The 3′UTR of the *IRGQ* gene, for example, has six pair sites. The nucleotide sequences of the repeating pair binding sites have a high degree of homology, again testifying that the origin of these sites in the 3′UTR is not random. The distance between the binding sites is still 12 nucleotides.

The 5′UTRs of 24 genes also have pair binding sites for miR-1273g-3p and miR-1273f (Figure 3). The nucleotide sequences of the sites in the 5′UTRs also have a high degree of homology. The distance between the binding sites is 12 nucleotides, indicating that both the 5′UTR and 3′UTR binding sites have a common origin.

The miR-1273g-3p and miR-1273f pair binding sites are present in the CDSs of 12 genes, and their locations are presented in Figure 4. The distance between the binding sites is again 12 nucleotides. The nucleotides of the miR-1273g-3p and miR-1273f binding sites in CDSs are less homologous than those located in the 5′UTRs and 3′UTRs. However, it is still possible to suppose a general origin for all of the pair sites located in the CDSs, 5′UTRs, and 3′UTRs.

The nucleotide sequences of the binding sites in CDSs are translated into corresponding amino acid sequences that create proteins. If the nucleotides of the miR-1273g-3p binding sites are read in different open reading frames (ORFs), three different oligopeptides can be produced. The oligonucleotide 5′-CUCAGGCUGGAGUGCAGUGGU-3′ of miR-1273g-3p's binding site can code the LRLECSG, SGWSAVV, and QAGVQW oligopeptides. The mRNAs of 14 genes have ORF oligopeptides that are homologous to RLECSG (Figure 5). Six mRNAs have other ORF and code oligopeptides that are homologous to QAGVQW. The third ORF was found only in the *NOP2* gene. The amino-acid sequences adjoining the studied oligopeptides are also homologous in some proteins. For example, in the ZNF573 and ZMAT1 proteins, the MESCSV hexapeptide is located near the TRLECSG and AQAGVQW oligopeptides, which corresponds to the nucleotides of the miR-1273g-3p binding sites. The oligonucleotide 5′-CACUGCAACCUCCAUCUCC-3′, in the miR-1273f binding site, can code the HCNLHL, TATSIS, and SLQPPS oligopeptides. In 5 genes that contain the miR-1273f binding site in their CDSs, the oligonucleotides code homologous oligopeptides in all three ORFs (Figure 5).

The homology of the nucleotide sequences adjacent to the miR-1273f binding sites causes the homology of the corresponding oligopeptides. The mRNA part between the miR-1273g-3p and miR-1273f binding sites codes homologous tripeptides (DLG and ILA) and tetrapeptides (AISA in both the MTO1 and ZMAT1 proteins). The nucleotide sequences of the mRNA segments adjacent to the miR-1273f site code homologous oligopeptides in some proteins. For example, the PGSSDS hexapeptide is located in both the ZMAT1 and C11orf80 proteins, the GSSNSPA heptapeptide is located in

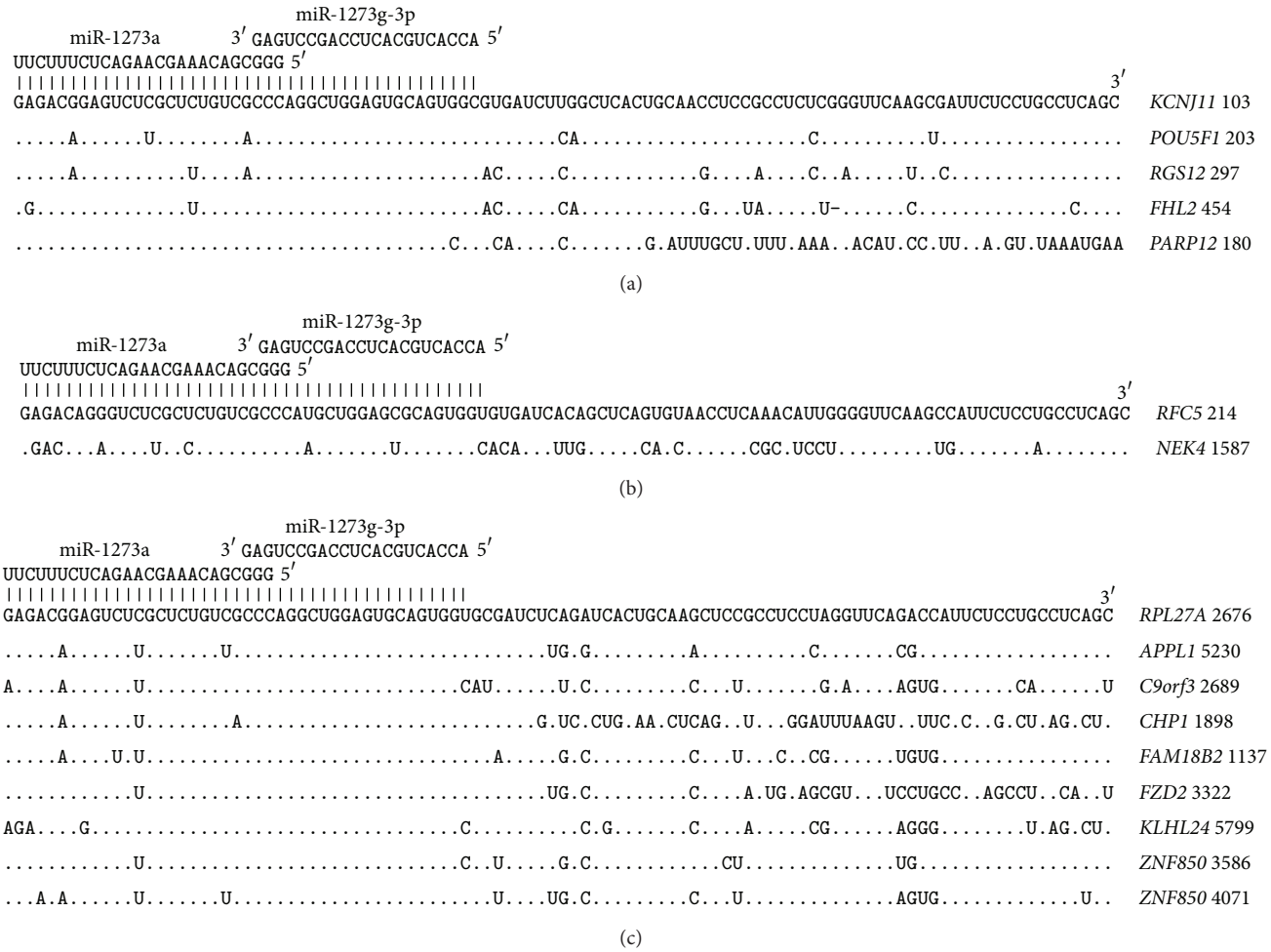


Figure 7: Arranged binding sites miR-1273g-3p and miR-1273a in 5′UTR (a), CDS (b), and 3′UTR (c) mRNA target genes.

the MAP4K1 and SLC36A3 proteins, and the GSSDSPAS nonapeptide is located in the NEK4, SPAG6, FAM122C, and ZMAT1 proteins.

The 3′UTR of 16 genes have pair binding sites for miR-1273g-3p and miR-1273g-5p. The mRNA of the *PAQR8* gene is chosen to compare with sites from other mRNAs (Figure 6). This mRNA can form hydrogen bonds with all of the nucleotides in both the miR-1273g-3p and miR-1273g-5p binding sites. The miR-1273g-3p and miR-1273g-5p binding sites in the 3′UTR have a high degree of homology. The distance between the binding sites is 9 nucleotides, indicating a general origin of these pair binding sites in the 3′UTR of the studied genes. The 5′UTR of *SMARCA4* has paired miR-1273g-3p and miR-1273g-5p binding sites (Figure 6). All of the nucleotides in the binding sites of these miRNAs form hydrogen bonds. The CDSs of 4 genes have paired miR-1273g-3p and miR-1273g-5p binding sites (Figure 6). Homologous oligonucleotides in the miR-1273g-3p binding sites coded the homologous oligopeptides PRLECSG and QAGVQW through two ORFs (Figure 6).

Both miR-1273g-3p and miR-1273a have pair binding sites in the mRNA of 113 genes. Five pair binding sites are located in mRNA 5′UTRs (Figure 7). The nucleotide sequences of these binding sites have three common nucleotides that are identical in five mRNAs. A high degree of homology was found in 99 nucleotide segments of the 5′UTR of the *KCNJ11, POU5F1, RGS12,* and *FHL2* genes. Only half of the binding sites located in the 5′UTRs of the *PARP12* gene are highly homologous. The CDSs of two genes contain pair binding sites for miR-1273g-3p and miR-1273a (Figure 7). Both of these gene sites are highly homologous and have three overlapped nucleotides. These sites can also code homologous polypeptides.

The 3′UTR of target genes have paired miR-1273g-3p and miR-1273a binding sites that are also located in the 5′UTR, with three overlapped nucleotides. The miR-1273g-3p and miR-1273a sites in the 3′UTR are highly homologous. The 3′-end sites also have homology with the nucleotides in the mRNA of many genes. The mRNAs of four genes have paired miR-1273g-3p and miR-1273c binding sites located in their 5′UTRs; two nucleotides are common to two sites (Figure 8). The nucleotide sequences of the binding sites are identical in the target genes' mRNAs. Other portions of the mRNA also have homologous nucleotide sequences. The location of

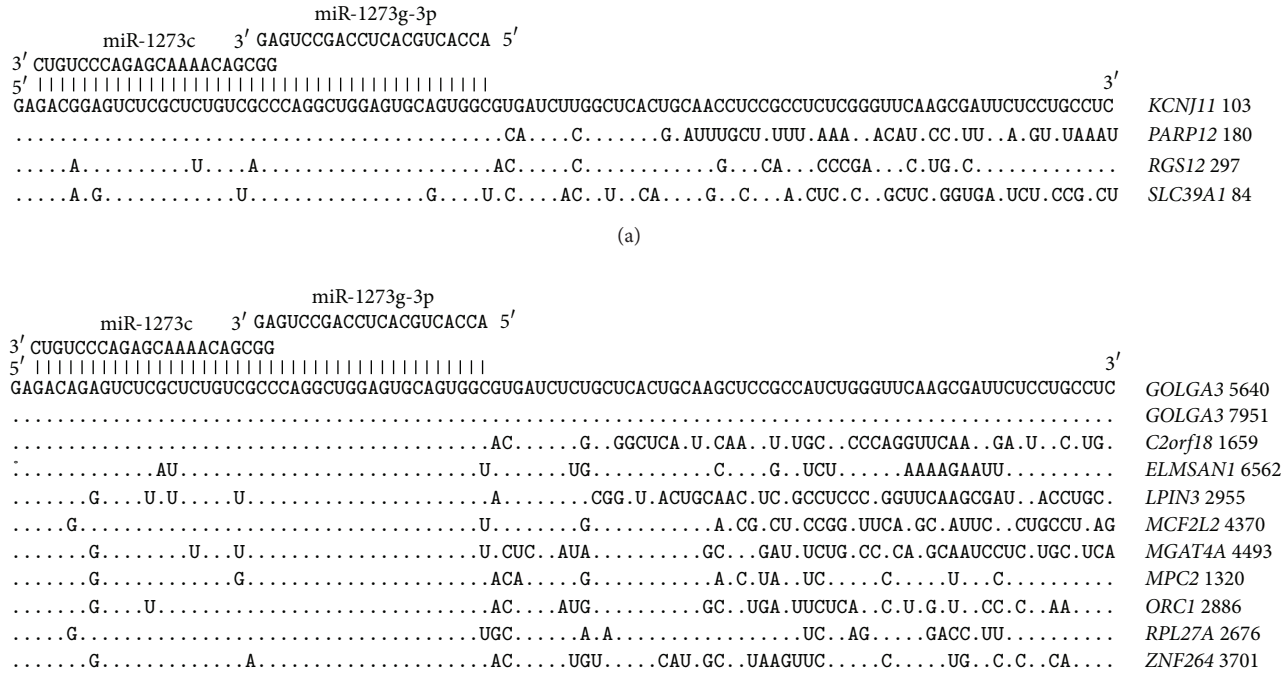


Figure 8: Arranged binding sites miR-1273g-3p and miR-1273c in 5′UTR (a) and 3′UTR (b) mRNA target genes.

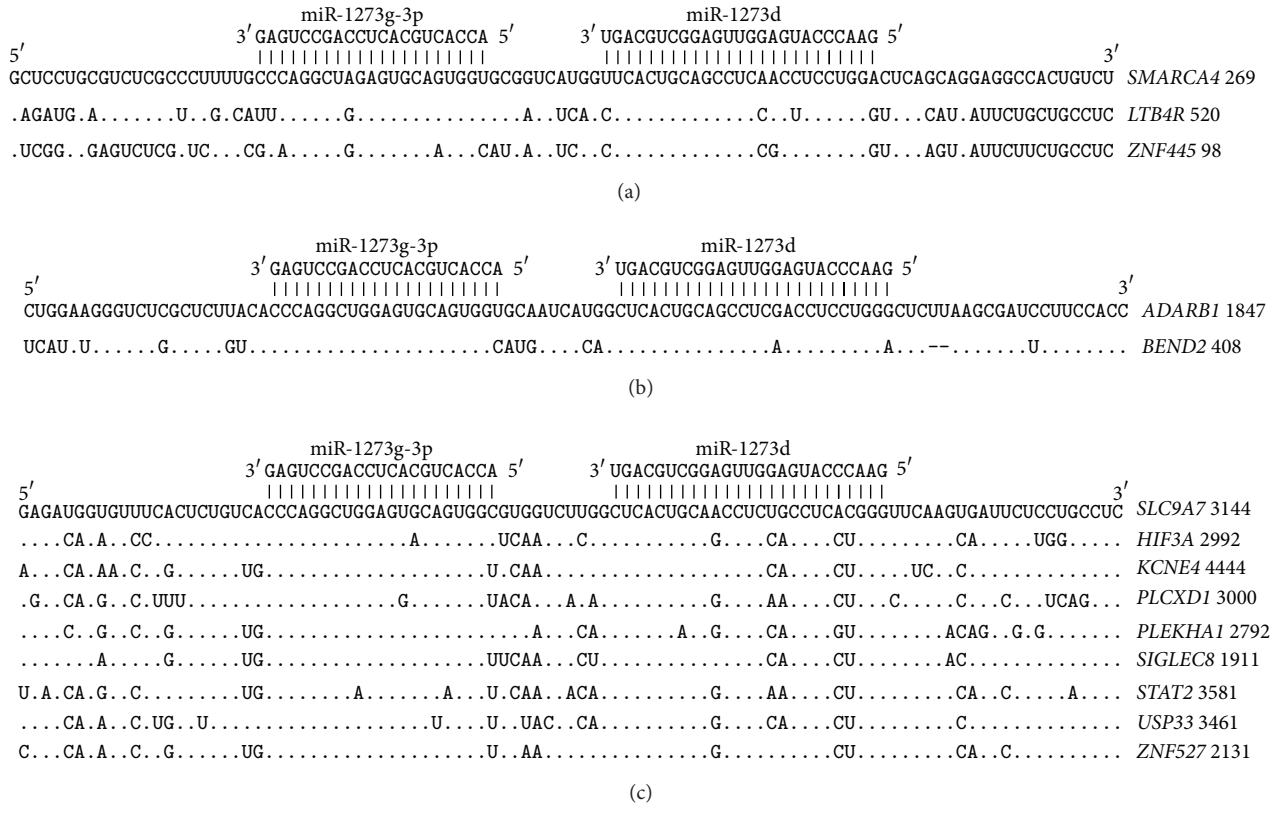
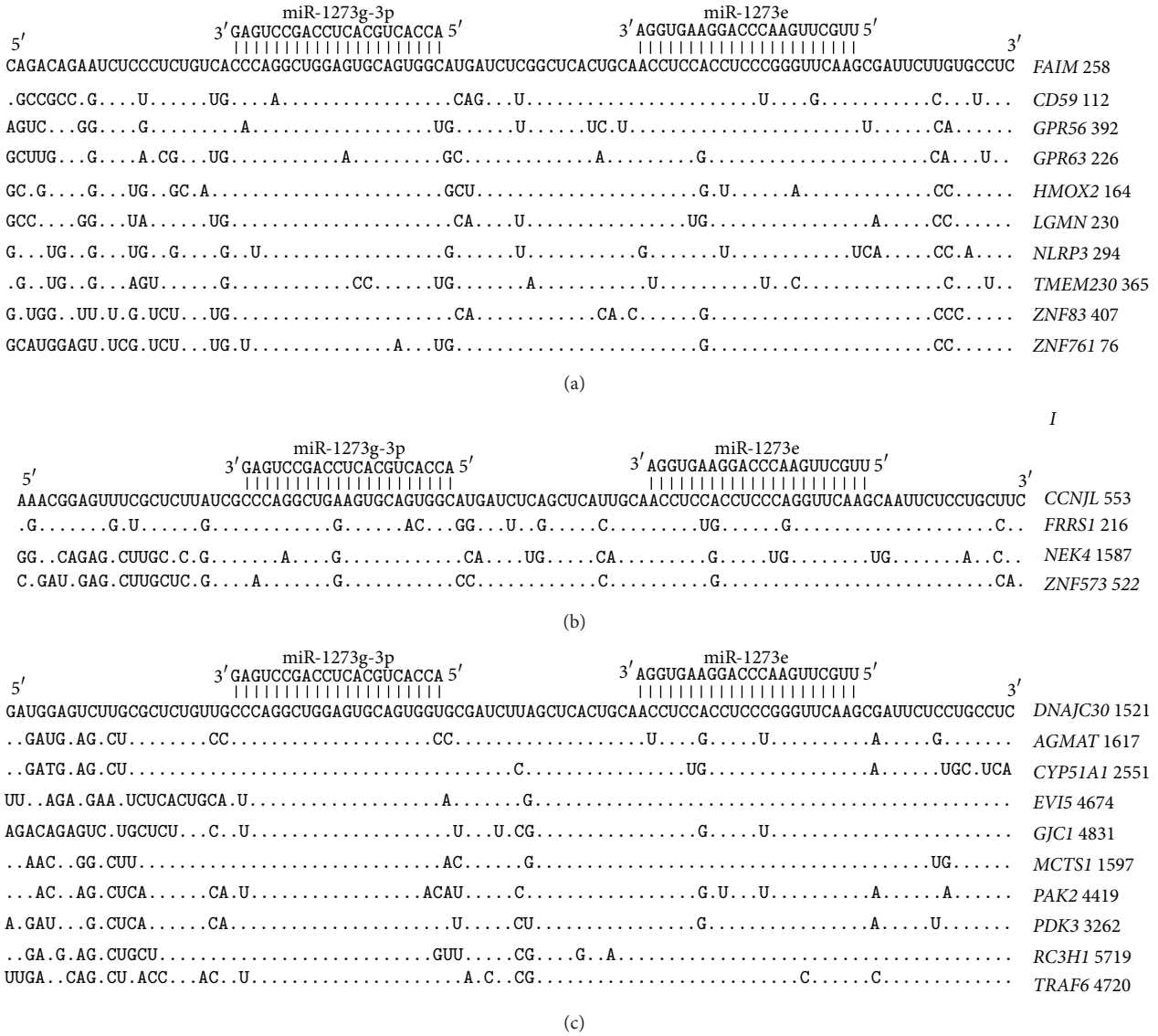


Figure 9: Arranged binding sites miR-1273g-3p and miR-1273d in 5′UTR (a), CDS (b), and 3′UTR (c) mRNA target genes.



(a)

(b)

(c)

Figure 10: Arranged binding sites miR-1273g-3p and miR-1273e in 5′UTR (a), CDS (b), and 3′UTR (c) mRNA target genes.

paired miR-1273g-3p and miR-1273c binding sites is identical in both the mRNA 5′UTRs and the 3′UTRs (Figure 8). The homology of the nucleotide sequences in the binding sites is high. The nucleotide sequences adjacent to the miR-1273g-3p binding site are also very homologous.

The paired miR-1273g-3p and miR-1273d binding sites are located in the 5′UTR at a distance of 13 nucleotides (Figure 9). The homology of the nucleotide sequences in the binding sites is high. The segments of mRNA at the 5′-end of the miR-1273d binding site, consisting of 10 nucleotides to one side and 18 nucleotides to the other, have only three different nucleotides. We assume that there is a common origin for the 5′UTR sites because of their high similarity. The paired miR-1273g-3p and miR-1273d binding sites are located in the 3′UTR at a distance of 13 nucleotides (Figure 9). The homology level of the nucleotide sequences is high not only in the miR-1273g-3p and miR-1273d binding sites but also in the mRNA regions adjacent to these sites.

The paired miR-1273g-3p and miR-1273d binding sites in the CDSs of *ADARB1* and *BEND2* mRNA are shown in Figure 9. The distance between these two binding sites is 13 nucleotides. The nucleotide sequences in the mRNA of the miR-1273g-3p and miR-1273d binding sites are very homologous. Taking into account a deletion of two nucleotides in the 3′-end of the site in *BEND2* mRNA, the homology of the adjacent parts of *ADARB1* and *BEND2* is high (Figure 9). Polypeptides correspond to these sites according to ORFs.

The nucleotide sequences of paired miR-1273g-3p and miR-1273e binding sites located in the 5′UTR and their adjacent parts have a high homology level (Figure 10). The mRNA segments in CDSs containing paired miR-1273g-3p and miR-1273e binding sites also have a high degree of homology (Figure 10). The paired miR-1273g-3p and miR-1273e binding sites are found in the 3′UTR of 300 genes, and they have a high degree of homology, as well (Figure 10).

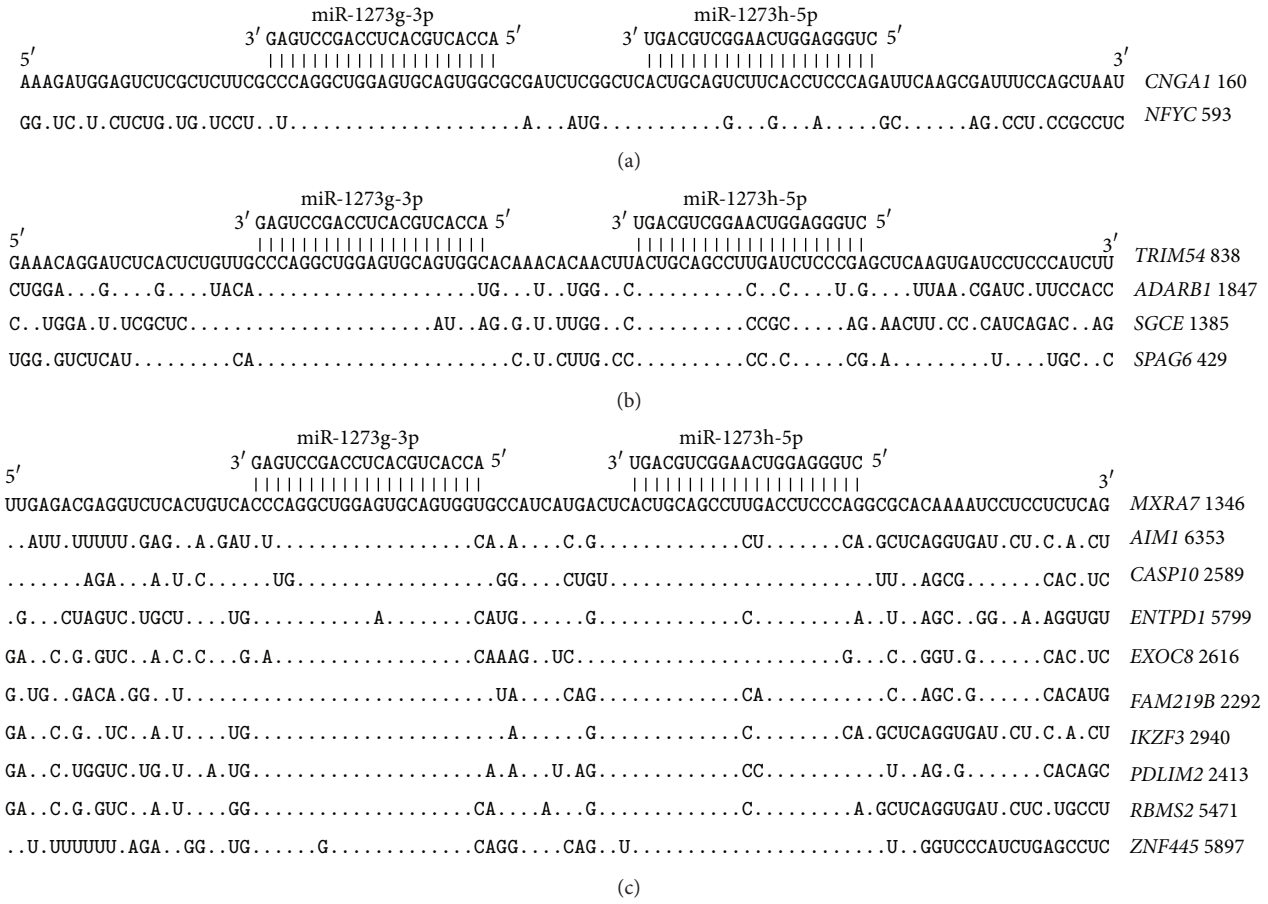


Figure 11: Arranged binding sites miR-1273g-3p and miR-1273h-5p in 5′UTR (a), CDS (b), and 3′UTR (c) mRNA target genes.

The segments of the 5′UTR in the *CNGA1* and *NFYC* genes that contain the miR-1273g-3p and miR-1273h-5p binding sites are shown in Figure 11. All of the nucleotides of these miRNAs form hydrogen bonds in the binding sites, and the degree of their homology is high. The distance between the miR-1273g-3p and miR-1273h-5p binding sites is 12 nucleotides.

The distance between the miR-1273g-3p and miR-1273h-5p binding sites located in the CDSs of four genes is 12 nucleotides. The nucleotide sequences of these binding sites and some adjacent segments have a high degree of homology. The nucleotides of the miR-1273g-3p and miR-1273h-5p binding sites code polypeptides of different ORFs. The LRLECSG and HCNLHL polypeptides are homologous in proteins SPAG6 and TRIM54, while the QAGVQW and LQPPSP polypeptides are homologous in proteins ADARB1 and SGCE (Figure 11). The nucleotide sequences of the 3′UTRs indicate that paired miR-1273g-3p and miR-1273h-5p binding sites are located similarly to those in 5′UTRs and CDSs, with a separation distance of 12 nucleotides (Figure 11). This part of the binding site mRNA is highly conserved, and the adjacent mRNAs are similarly homologous. No paired binding sites are found for miR-1273g-3p and miR-1273h-3p in any of the mRNA locations described above.

3.3. Arrangement of the Binding Sites of the miR-1273 Family in mRNA. This analysis of the localization of paired miR-1273 binding sites in the mRNA of target genes leads to the conclusion that they evolved from a common ancestor. Most of these binding sites are located in mRNA segments 99 nucleotides long (Figures 1–11). Such compactness in the binding site location of the miR-1273 family could be a result of embedding one general nucleotide sequence into the target genes. This work showed that pair binding sites have a monophyletic origin. The complementary nucleotide sequence to pre-miR-1273h includes binding sites for the miR-1273 family, and it is the most probable precursor for these segments (Figure 12). The adaptation of miRNA binding sites to each member of the miR-1273 family or to their combinations could also be due to the evolution of target gene mRNA and their varying functions.

The nucleotide sequences of miR-1273g-3p and miR-1273a have three overlapped nucleotides, as well as pair binding sites (Figure 12). Both miR-1273g-3p and miR-1273c have two overlapped nucleotides and pair binding sites, whose schemes are shown in Figure 8. The nucleotide sequences of miR-1273g-3p and miR-1273h have 16 overlapped nucleotides (Figure 12) that correspond to overlapping of their binding sites, shown in Figure 11. The distance between miR-1273g-3p and miR-1273g-5p is nine nucleotides (Figure 12), which

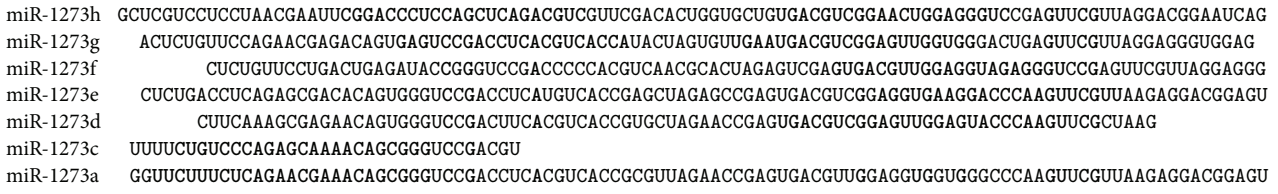


Figure 12: A scheme showing the homology of the pre-miR-1273 family.

correspond to the interval between the miR-1273g-3p and miR-1273g-5p binding sites, per their schemes (Figure 6). The distances between miR-1273g-3p and miR-1273h-5p (Figure 12) and between their pair binding sites (Figure 11) are each 13 nucleotides. The distance between the nucleotides of miR-1273g-3p and miR-1273d is 13 nucleotides, matching the distances between the pair binding sites of these miRNAs in the schemes of Figure 9. The interval between miR-1273g-3p and miR-1273e is 22 nucleotides (Figure 10), again matching the distance between their pair sites, shown in Figure 12. However, the distance between miR-1273g-3p and miR-1273f is 18 nucleotides (Figure 12) while the distance between their pair sites is only 12 nucleotides (Figures 3 and 4). It is possible that the deletion of six nucleotides occurred in the primary site at an early stage of this pair's formation.

The distances described above between the pair binding sites of the miR-1273 family are nearly always matched in the target gene mRNA. However, all of the pair binding sites of the miR-1273 family have deviations of one-two nucleotides between them. Thus, the average distance between the miR-1273g-3p and miR-1273f mRNA binding sites is 12.1 ± 2.2 nucleotides.

A feature of the miR-1273 family that this study discovered is the presence of pair binding sites in mRNA segments of 100 nucleotides. Figure 12 shows that the miRNA binding sites locate in the mRNAs of target genes occur in a certain order, using different combinations of miR-1273g-3p binding sites and those of other members of this miRNA family.

Increases or decreases in miRNA synthesis, particularly umiRNAs, can lead to an imbalance of gene expressions across the genome. Thus, changes to miRNA expression can lead to disturbances in metabolic processes, the achievement of an organism's development program, an organism's response to different impacts, or ultimately the development of various pathologies. The role of umiRNAs and other miRNAs is assumed to be vast because they circulate in the blood, and almost all of the cells in an organism are available to them [20, 21].

Highly conserved binding sites of miR-1273 family in a large number of genes testify about their emergence in the early stages in human evolution. Arranged localization of these binding sites suggests an interconnected development of evolution of miRNAs and their target genes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank A. Moldagaliyeva and S. Sagaydak for their help in preparing the materials for analysis. The authors would also like to thank Dr. V. Khaylenko for writing the Lextractor002 script. This study was supported by a grant from the Ministry of Education and Science, Kazakhstan.

References

[1] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, "Identification of novel genes coding for small expressed RNAs," *Science*, vol. 294, no. 5543, pp. 853–858, 2001.

[2] Y. C. Po and G. Meister, "microRNA-guided posttranscriptional gene regulation," *Biological Chemistry*, vol. 386, no. 12, pp. 1205–1218, 2005.

[3] J. T. Mendell, "MicroRNAs: critical regulators of development, cellular physiology and malignancy," *Cell Cycle*, vol. 4, no. 9, pp. 1179–1184, 2005.

[4] N. Elefant, Y. Altuvia, and H. Margalit, "A wide repertoire of miRNA binding sites: prediction and functional implications," *Bioinformatics*, vol. 27, no. 22, Article ID btr534, pp. 3093–3101, 2011.

[5] D. Didiano and O. Hobert, "Molecular architecture of a miRNA-regulated 3′ UTR," *RNA*, vol. 14, no. 7, pp. 1297–1317, 2008.

[6] A. T. Ivashchenko, A. S. Issabekova, and O. A. Berillo, "MiR-1279, miR-548j, miR-548m, and miR-548d-5p binding sites in CDSs of paralogous and orthologous PTPN12, MSH6, and ZEB1 genes," *BioMed Research International*, vol. 2013, Article ID 902467, 10 pages, 2013.

[7] Y. Sun, M. Wang, G. Lin et al., "Serum microRNA-155 as a potential biomarker to track disease in breast cancer," *PLoS ONE*, vol. 7, no. 10, Article ID e47003, 2012.

[8] J. Kang, S. Y. Lee, S. Y. Lee et al., "MicroRNA-99b acts as a tumor suppressor in non-small cell lung cancer by directly targeting fibroblast growth factor receptor 3," *Experimental and Therapeutic Medicine*, vol. 3, no. 1, pp. 149–153, 2012.

[9] S. G. Liu, X. G. Qin, B. S. Zhao et al., "Differential expression of miRNAs in esophageal cancer tissue," *Oncology Letters*, vol. 5, no. 5, pp. 1639–1642, 2013.

[10] X. Zhao, X. Li, and H. Yuan, "MicroRNAs in gastric cancer invasion and metastasis," *Frontiers in Bioscience*, vol. 18, no. 1, pp. 803–810, 2013.

[11] X. Luo, C. Stock, B. Burwinkel, and H. Brenner, "Identification and evaluation of plasma microRNAs for early detection of colorectal cancer," *PLoS ONE*, vol. 8, no. 5, Article ID e62880, 2013.

[12] B. A. Walter, V. A. Valera, P. A. Pinto, and M. J. Merino, "Comprehensive microRNA profiling of prostate cancer," *Journal of Cancer*, vol. 4, no. 5, pp. 350–357, 2013.

[13] H. Yang, W. Zheng, W. Zhao, C. Guan, and J. An, "Roles of miR-590-5p and miR-590-3p in the development of hepatocellular carcinoma," *Nan Fang Yi Ke Da Xue Xue Bao*, vol. 33, no. 6, pp. 804–811, 2013 (Chinese).

[14] W. Wang, T. Li, G. Han, Y. Li, L. Shi, and H. Li, "Expression and role of miR-34a in bladder cancer," *Indian Journal of Biochemistry and Biophysics*, vol. 50, no. 2, pp. 87–92, 2013.

[15] S. Vang, H. Wu, A. Fischer et al., "Identification of Ovarian Cancer Metastatic miRNAs," *PLoS ONE*, vol. 8, no. 3, Article ID e58226, 2013.

[16] A. E. Frampton, T. M. Gall, E. Giovannetti et al., "Distinct miRNA profiles are associated with malignant transformation of pancreatic cystic tumors revealing potential biomarkers for clinical use," *Expert Review of Molecular Diagnostics*, vol. 13, no. 4, pp. 325–329, 2013.

[17] O. A. Berillo, G. K. Baidildinova, and A. T. Ivashchenko, "miRNAs as regulators of tumour suppressor expression," *World Academy of Science Engeneering and Technology*, vol. 73, no. 1, pp. 82–86, 2013.

[18] E. T. Kool, "Hydrogen bonding, base stacking, and steric effects in DNA replication," *Annual Review of Biophysics and Biomolecular Structure*, vol. 30, pp. 1–22, 2001.

[19] N. B. Leontis, J. Stombaugh, and E. Westhof, "The non-Watson-Crick base pairs and their associated isostericity matrices," *Nucleic Acids Research*, vol. 30, no. 16, pp. 3497–3531, 2002.

[20] P. S. Mitchell, R. K. Parkin, E. M. Kroh et al., "Circulating microRNAs as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 30, pp. 10513–10518, 2008.

[21] F. Russo, S. di Bella, G. Nigita et al., "miRandola: Extracellular Circulating MicroRNAs Database," *PLoS ONE*, vol. 7, no. 10, Article ID e47786, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 926342, 15 pages
http://dx.doi.org/10.1155/2014/926342

Research Article

Phylogenetic Information Content of Copepoda Ribosomal DNA Repeat Units: ITS1 and ITS2 Impact

Maxim V. Zagoskin,[1] Valentina I. Lazareva,[2] Andrey K. Grishanin,[2,3] and Dmitry V. Mukha[1]

[1] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkin Street. 3, Moscow 119991, Russia*
[2] *Papanin Institute for Biology of Inland Waters, Russian Academy of Sciences, Borok 152742, Russia*
[3] *Dubna International University for Nature, Society and Man, Universitetskaya Street 19, Dubna 141980, Russia*

Correspondence should be addressed to Maxim V. Zagoskin; zagoskinmv@gmail.com, Andrey K. Grishanin; andreygrishanin@mail.ru and Dmitry V. Mukha; dmitryVmukha@gmail.com

Received 23 April 2014; Revised 8 July 2014; Accepted 8 July 2014; Published 18 August 2014

Academic Editor: Peter F. Stadler

The utility of various regions of the ribosomal repeat unit for phylogenetic analysis was examined in 16 species representing four families, nine genera, and two orders of the subclass Copepoda (Crustacea). Fragments approximately 2000 bp in length containing the ribosomal DNA (rDNA) 18S and 28S gene fragments, the 5.8S gene, and the internal transcribed spacer regions I and II (ITS1 and ITS2) were amplified and analyzed. The DAMBE (Data Analysis in Molecular Biology and Evolution) software was used to analyze the saturation of nucleotide substitutions; this test revealed the suitability of both the 28S gene fragment and the ITS1/ITS2 rDNA regions for the reconstruction of phylogenetic trees. Distance (minimum evolution) and probabilistic (maximum likelihood, Bayesian) analyses of the data revealed that the 28S rDNA and the ITS1 and ITS2 regions are informative markers for inferring phylogenetic relationships among families of copepods and within the Cyclopidae family and associated genera. Split-graph analysis of concatenated ITS1/ITS2 rDNA regions of cyclopoid copepods suggested that the *Mesocyclops*, *Thermocyclops,* and *Macrocyclops* genera share complex evolutionary relationships. This study revealed that the ITS1 and ITS2 regions potentially represent different phylogenetic signals.

1. Introduction

Copepods are important components of zooplankton and the food chain in marine and freshwater ecosystems. The subclass Copepoda is believed to contain approximately 13,000 morphospecies; however, the actual number of species in this subclass might be much greater [1]. The majority of the freshwater copepod species belong to the order Cyclopoida, which includes the free-living species (approximately 800) in the family Cyclopidae. The other two free-living families (Oithonidae and Cyclopinidae) contain mainly marine species except for a few species in Oithonidae [2].

Systematic analyses of cyclopoid copepods (order Cyclopoida) have primarily focused on morphological characteristics [2–8], and the majority of molecular studies have targeted marine copepods [9–41]. The phylogenetic history of freshwater cyclopoid copepods is not well understood. A few studies on Cyclopidae have used molecular and morphological analyses on the *Mesocyclops* genus (Crustacea: Cyclopidae) [42], *E. serrulatus* group [43], *Acanthocyclops vernalis-robustus* species complex [44, 45], and 11 populations of *Macrocyclops albidus* [46]; other phylogenetic analyses have focused only on molecular markers in *Diacyclops* spp., which are found in western Australia [47] and Lake Baikal [48].

Molecular markers such as genomic DNA fragments are used for phylogenetic analyses to elucidate the evolutionary history of living organisms, and the region of genomic DNA analyzed is critical. Mitochondrial DNA fragments (genes encoding the cytochrome c oxidase subunit I (COI), 16S rRNA, and cytochrome b) [9–21, 39, 40, 49–52] and/or nuclear rDNA regions have been used for the phylogenetic analysis of cyclopoid copepods [22–37, 41, 53–58]. Mitochondrial DNA fragments might be less useful for the analysis of copepod phylogeny compared to the phylogeny of other

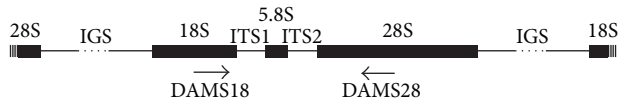


Figure 1: Organization of eukaryotic tandemly repeated rDNA clusters. 18S, 5.8S, and 28S ribosomal RNA genes; ITS1 and ITS2 internal transcribed spacers; IGS intergenic spacer. Arrows indicate the locations of the DAMS18 and DAMS28 primers.

taxa; furthermore, amplification of COI is difficult in some copepods [59–61]. However, these DNA fragments might be informative for analyses of population differentiation or cryptic speciation [9, 11–15, 38–40, 50, 52]. Therefore, comparison of nuclear rDNA regions might be informative for the phylogenetic analysis of copepods.

In most eukaryotes, rRNA genes are located in a multigene family of genomic clusters of repeated sequences. Within these clusters, the 18S, 5.8S, and 28S rRNA genes are separated by internal transcribed spacers (ITS1 and ITS2) and an intergenic spacer [62] (Figure 1). Ribosomal DNA (rDNA) is a reliable and informative phylogenetic marker [63] that contains sequences with different rates of evolutionary variability. In most eukaryotes, the most evolutionarily conserved genes are the rRNA genes; comparison of their sequences allows estimation of the evolutionary distances at intergeneric and higher taxonomic levels. Comparison of the more evolutionarily variable spacer sequences enables the study of phylogenetic relationships at the species and population levels [63–65]. Therefore, comparison of different regions of rDNA enables the phylogenetic analysis of organisms over extended evolutionary distances.

Phylogenetic relationships among cyclopoid copepods at higher systematic levels (ordinal, familial, and generic) have been resolved using the 18S and 28S nuclear rRNA genes [28, 29, 55–57], and the relationships at the lower taxonomic levels (species and populations) have been resolved using the ITS2 of the nuclear rDNA gene cluster [30, 40, 42, 52, 58].

Notably, analysis of the evolutionary history of living organisms based on only one molecular marker can uncover bifurcating phylogenetic trees, revealing branched evolution. However, it recently becomes evident that evolution is not always tree-like. Comparisons of gene trees based on different genetic loci often reveal conflicting tree topologies. These discrepancies are not always due to the problems with the sampling and the gene tree reconstruction methods. Reticulation events such as horizontal gene transfer (HGT) and hybridization may be responsible for contradictions in lineages. During an HGT event, a DNA segment is transferred from one organism to another which is not its offspring, whereas hybridization describes the origin of a new species through an interspecies mating. Both processes yield genomes that are mixtures of DNA regions derived from different species. Consequently, evolutionary relationships between species whose past includes reticulation can often be better represented by using phylogenetic networks rather than trees [65–67].

In view of the above, comparing phylogenetic trees based on different molecular markers may be used for the analysis of evolutionary events caused by reticulate evolution. Phylogenetic signals from various molecular markers are potentially divergent during reticulate evolution, resulting in phylogenetic trees with alternative positions for the individual branches [68–71]. Comparative analysis of the rDNA ITS1 and ITS2 sequences is suitable for studying phylogenetic relationships in terms of branching and reticulate evolution [63, 68, 70, 72–82].

Reticulate evolution is primarily driven by hybrid speciation, which is common among plants [83] but also occurs among animals, particularly including fish [84], amphibians, and several invertebrates [85–87]. In both mammals [88] and arthropods [89, 90], a single instance of hybrid speciation has been well described. Interspecies hybridization typically results in complicated relationships within species complexes, characterized by indistinct species borders. Reticulate evolution among crustaceans has been observed only within species complexes of daphnids [91–94].

In this study, we analyzed the phylogenetic relationships within a small group of cyclopoid copepods representing several genera of freshwater (*Cyclops*, *Thermocyclops*, *Diacyclops*, *Megacyclops*, *Macrocyclops*, and *Mesocyclops*) and marine (*Oithona* and *Paracyclopina*) organisms. Specific freshwater species were selected for analysis because these species are important for the maintenance of food chains in Russian freshwater ecosystems. The aim of this study was to analyze the sequence characteristics of the rDNA 28S gene, ITS1, and ITS2 regions as phylogenetic markers for the selected group of organisms.

2. Materials and Methods

2.1. Samples Collection. Nine freshwater species of the Cyclopidae family were collected near the Borok settlement in the Yaroslavskaya region of Russia: *Mesocyclops leuckarti* (Claus, 1857), *Cyclops strenuus* (Fischer, 1851), and *Cyclops insignis* (Claus, 1857) (population no. 1) from the Barskiy Pond (58°3′59.35″N; 38°15′10.16″E); *Thermocyclops oithonoides* (Sars, 1863) from the Sunoga pond (58°2′34.66″N; 38°14′41.29″E); *Thermocyclops crassus* (Fischer, 1853), *Macrocyclops distinctus* (Richard, 1887), *Macrocyclops albidus* (Jurine, 1820), *Diacyclops bicuspidatus* (Claus, 1857), and *Megacyclops viridis* (Jurine, 1820) (population no. 1) from the Ikhteologichesky Canal (58°3′55.62″N; 38°15′21.05″E); and *Megacyclops viridis* (Jurine, 1820) (population no. 2) from a pond in the flood zone of the Rybinsk Reservoir (58°4′4.70″N; 38°15′39.88″E).

Cyclops kolensis (Lilljeborg, 1901) and *Cyclops insignis* (Claus, 1857) (population no. 2) were collected from the Andreevsky small pond in Vorob'evy Gory, Moscow, Russia (55°42′35.40″N; 37°34′6.61″E). Two marine species, *Oncaea* sp. (Claus) and *Oithona similis* (Claus, 1866), were collected from the Norwegian Sea (68°52′36.67″N; 3°8′21.91″E).

Individuals of each species were collected for further analysis at the specified locations over a 0.5- to 1-hour period.

No specific permission was required to collect samples at these locations. None of the studied species is endangered or protected.

A fragment of the 28S gene from each of the four marine species *Paracyclopina nana* (Smirnov, 1935) (GenBank accession number FJ214952), *Oithona nana* (Giesbrecht, 1893) (GenBank accession number FM991727), *Oithona simplex* (Farran, 1913) (GenBank accession number AF385458), and *Oithona helgolandica* (Claus, 1863) (GenBank accession number FM991724.1) was also used for the analysis. The DNA was extracted from either samples preserved in 70% ethanol or raw materials (Moscow populations).

2.2. DNA Extraction, PCR Amplification, and Sequencing. The genomic DNA was isolated from 10–20 individuals of each collected species using the DNeasy Blood & Tissue kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions and was frozen at −20°C. The rDNA region (approximately 2000 bp) was amplified from the genomic DNA by polymerase chain reaction (PCR) using the universal eukaryotic rDNA primers DAMS18 and DAMS28 [95–97] (Figure 1). The amplified rDNA regions contained the ITS1 (261–388 bp) and ITS2 (188–262 bp) regions, the 5.8S gene (157 bp), and approximately 200 and 1000 bp of the 18S and 28s genes, respectively. The amplification was performed in 50 μL reactions using a PCR Master Mix (2X) (Fermentas, Vilnius, Lithuania) according to the manufacturer's instructions; the reactions were performed in a Primus 25 advanced Thermocycler (PEQLAB, Erlangen, Germany) using previously published rDNA-specific parameters [98]. The PCR products were resolved on 1.0% agarose gels, and DNA was extracted from the observed unique bands using the QIAquick Gel Extraction Kit (QIAGEN, Hilden, Germany). The extracted products were cloned into the pGEM-T Easy vector (Promega, USA), and the resulting plasmids were used to transform *Escherichia coli* JM109 competent cells (Promega, USA) according to the manufacturer's instructions. For each species, the amplified product and five clones were sequenced. Automated sequences were generated on an ABI PRISM 310 Genetic Analyzer according to Sanger et al. [99] with a BigDye Termination kit (Applied Biosystems, USA). The sequences generated in this study were deposited in GenBank under the accession numbers KF153689–KF153701.

2.3. Phylogenetic Analyses. The rDNA sequences were aligned using ClustalW 2.1 [100, 101] with some manual adjustments. The boundaries of the ITS1 and ITS2 regions and the 28S gene were identified by comparing the primer-delimited sequences against sequences in the GenBank database using BLAST analysis. The boundaries of the conserved sequences were considered to represent the 5.8S, 18S, and 28S gene flanking regions if they were 100% similar to the boundaries of rDNA sequences in the GenBank database. The initial sequence alignment flanked by DAMS18/DAMS28 primers was divided into ITS1 and ITS2 alignments. The rDNA genetic distances were estimated using the MEGA V5.2 software [102]. DAMBE (Data Analysis in Molecular Biology and Evolution) software was used to analyze substitution saturation [103–105]. This method computes the entropy-based index of substitution saturation and its critical value. If the index of substitution saturation (Iss) approaches 1 or if the Iss is not smaller than the critical Iss value (Iss.c), then sequences are considered to contain substantial saturation. As is known, the substitution saturation decreases phylogenetic information contained in sequences and has plagued the phylogenetic analysis involving deep branches. In the extreme case when sequences have experienced full substitution saturation, the similarity between the sequences will depend entirely on the similarity in nucleotide frequencies, which often does not reflect phylogenetic relationships [106].

The rDNA-based phylogenetic trees were estimated using probabilistic (maximum likelihood (ML), Bayesian) and distance (minimum evolution (ME)) methods [107–110]. ML and ME analyses of ITS1, ITS2, and 28S data were performed using the program MEGA V5.2. Branch support was assessed using the bootstrap method [111] (1,000 replicates) with the close-neighbor-interchange (CNI) algorithm at a search level of 1 for ME analysis and heuristic search for ML analysis. The Bayesian information criterion (BIC), as implemented in MEGA V5.2, was used to identify the best-fit model of sequence evolution for the trees estimated using ML. The evolutionary history was inferred using the ML method based on the general time reversible with the gamma distribution shape parameter (GTR+G) model for 28S and the Hasegawa-Kishino-Yano with gamma distribution shape parameter (HKY+G) model [112] for the ITS1, ITS2, and concatenated ITS1/ITS2 alignments. In addition to these methods, ITS1 and ITS2 alignments were constructed using the MAFFT version 7 (http://mafft.cbrc.jp/alignment/server/) [113] and Gblocks version 0.91b (http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=gblocks) software programs [114–116] to eliminate poorly aligned and highly divergent regions. Default parameters were used for both of these methods. The Tamura 3-parameter model (T92) [117] and HKY with evolutionary invariable (HKY+I) for Gblocks-treated MAFFT ITS1 and ITS2 data, respectively, were used to infer evolutionary history inference using the ML method.

The Bayesian analysis was performed using MrBayes version 3.1.2 software [118, 119]. Two replicate analyses of 1 million generations each were performed for each dataset, with sampling every 10 generations. The hierarchical likelihood ratio test (hLRT) implemented in MrModeltest version 2.3 software [120] was used to identify the model of best fit (Hasegawa-Kishino-Yano with invariant sites and gamma distribution shape parameter (HKY+I+G) [112] for ITS1 and the HKY+G model for ITS2). Trees from the first 53,000 and 118,000 generations were discarded as burn-in for ITS1 and ITS2, respectively. The Bayesian tree was estimated from the majority-rule consensus of the post-burn-in trees.

A reticulogram [121] was constructed using the T-REX version 4.01a software [122] with the distance matrix computed using the Kimura 2-parameters model (ignoring missing bases); the weighted least-squares method was used for tree reconstruction [123], and addition of reticulation branches stopped when $K = 1$ branches were added.

Network reconstruction was performed using Splits Tree 4 version 4.11.3 software [65]. The neighbor-net network

method and uncorrected p-distances were used to analyze and visualize reticulate relationships. All gaps were excluded for analysis. Network robustness was tested using 1,000 bootstrap replicates.

3. Results and Discussion

3.1. Characteristics and Analysis of an rDNA Sequence Dataset. In each species, the nucleotide sequences of the amplified rDNA region and five clones were not significantly different from each other. The frequency of variable nucleotides did not exceed the average rate of nucleotide substitutions caused by DNA polymerase errors, which is approximately one substitution per 1,000 nucleotides. The compared sequences contained both relatively evolutionarily conserved (fragments of 18S and 28S rDNA and the complete 5.8S rDNA) and evolutionary variable genomic regions (ITS1 and ITS2). For different taxa, the ITS1 and ITS2 sequences vary significantly among individuals at the inter- and intrapopulation levels; furthermore, these sequences can exhibit intragenomic variability [25, 41, 53, 54]. Recently, a high level of intrapopulation polymorphism of the 28S rDNA sequences was observed within *Oithona* spp. [22]. However, there are instances of strong evolutionary conservation of the 28S and ITS sequences [15, 23, 30, 34]. Notably, the *M. leuckarti* ITS2 sequence obtained in this study did not exhibit any nucleotide substitutions compared to the *M. leuckarti* ITS2 sequence described previously (GenBank accession number GQ848499) [42]. Therefore, the strong evolutionary conservation of ITS1 and ITS2 sequences is a characteristic feature of the copepod species analyzed in this study.

In this study, the applicability of different segments of rDNA containing the ITS1 and ITS2 regions, the 5.8S RNA gene, and fragments of the 18S and 28S rRNA genes was examined for reconstruction of the phylogenetic relationships among freshwater cyclopoid copepods. The 5.8S gene and the analyzed fragment of the 18S gene were not considered for phylogenetic reconstruction due to their short length and strong evolutionary conservation: only a few nucleotide substitutions were detected by comparing these sequences with evolutionary distant species (data not shown).

For 15 specimens of Cyclopoida species (including the two marine species), the average length of the 28S gene fragment sequenced was 1051 bp. We trimmed these sequences to 703 bp and compared them with the 28S gene sequences of marine species available in GenBank. These 703 bp of 28S rDNA sequences were aligned, and 342 variable sites were observed. *Oncaea* sp. (Oncaeidae family) was used as the out group.

The ITS1 sequence lengths varied from 267 to 388 bp among the 13 Cyclopidae specimens. The ITS1 sequence alignments possessed 442 characters, and among them, 283 were variable. The ITS2 sequence lengths varied from 188 to 262 bp among the 13 Cyclopidae specimens. ITS2 sequence alignment possessed 302 characters, and 190 were variable.

All alignment sets were examined for homogeneity of base frequencies and substitution saturation. The average base frequencies of the 28S gene fragment ($A = 20.52$, $C = 25.33$, $G = 32.75$, and $T = 21.40\%$) differed from the ITS1 ($A = 14.27$, $C = 30.68$, $G = 27.55$, and $T = 27.50\%$) and ITS2 ($A = 13.08$, $C = 29.64$, $G = 30.08$, and $T = 27.19\%$) regions. Gaps were excluded while estimating the average base frequencies of the ITS sequences. Using the chi-squared test, no significant differences were observed in the base compositions of the 28S ($\chi^2 = 39.77$, $df = 54$, and $P = 0.93$), ITS1 ($\chi^2 = 22.04$, $df = 36$, and $P = 0.97$), and ITS2 ($\chi^2 = 22.65$, $df = 36$, and $P = 0.96$) sequences among different taxa.

Table 1: False test of substitution saturation.

Alignment	Iss	Iss.c	Std. error
28S	0.200	0.739	0.019
ITS1	0.427	0.679	0.044
ITS2	0.376	0.665	0.043

Testing whether the observed Iss is significantly ($P < 0.001$, two-tailed t-test) lower than the Iss.c for a symmetrical tree.

To analyze whether the divergence of 28S, ITS1, and ITS2 rDNA fragments among species was saturated, we performed a substitution saturation test and generated saturation plots. Using DAMBE, the substitution saturation test revealed an Iss value that was significantly ($P < 0.001$) lower than the Iss.c in all cases (Table 1). This result indicated the suitability of the data for phylogenetic analysis. The total numbers of transition and transversion substitutions were plotted individually against model-corrected maximum-likelihood pairwise distances for the 28S, ITS1, and ITS2 sequences (see Supplementary Figure 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/926342). Using linear regression analysis on the 28S, ITS1, and ITS2 saturation graphs, the coefficients of determination (R^2) were calculated for both classes of substitutions: for transitions, the R^2 values were 0.79, 0.68, and 0.74 for the 28S, ITS1, and ITS2 sequences, respectively; for transversions, the R^2 values were 0.95, 0.93, and 0.91 for the 28S, ITS1, and ITS2 sequences, respectively. The R^2 values indicated that no less than 70% of the total variation in pairwise transitions and transversions could be explained by the linear relationship between pairwise distances and the total number of transitions and transversions. All saturation plots showed significant linear correlations (Supplementary Figure 1). Therefore, both transitions and transversions steadily accumulated as the corrected pairwise divergence increased, indicating that saturation was not reached.

3.2. Distance Analyses and Phylogenetic Tree Reconstruction. Phylogenetic analysis of the cyclopoid copepods species based on rDNA showed that the 28S rDNA sequences are informative for the phylogeny of both higher-level and closely related Copepoda species, whereas the ITS1 and ITS2 sequences are highly informative for reconstruction of the evolutionary history of closely related species. The ITS1 and ITS2 sequences are known to evolve more rapidly than the ribosomal RNA genes. Consistent with this observation, in

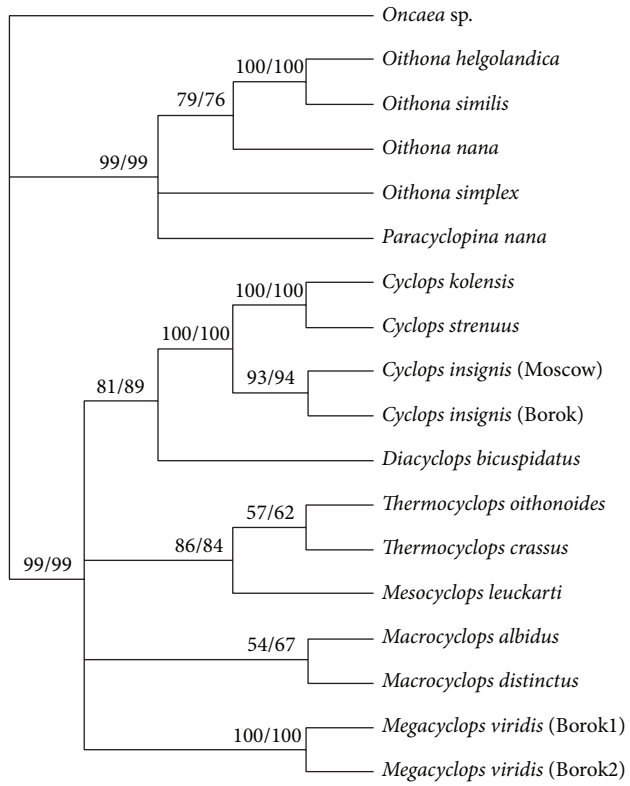


Figure 2: Phylogenetic relationships of Cyclopoida based on ~700 bp of the 28S rRNA gene. The consensus cladogram inferred from the 28S ribosomal DNA fragment sequence data of 16 Podoplea superorder species using maximum likelihood (ML) analysis under the HKY+G model and minimum evolution (ME) analysis. The numbers above branches indicate bootstrap percentages. The values are listed for ML/ME.

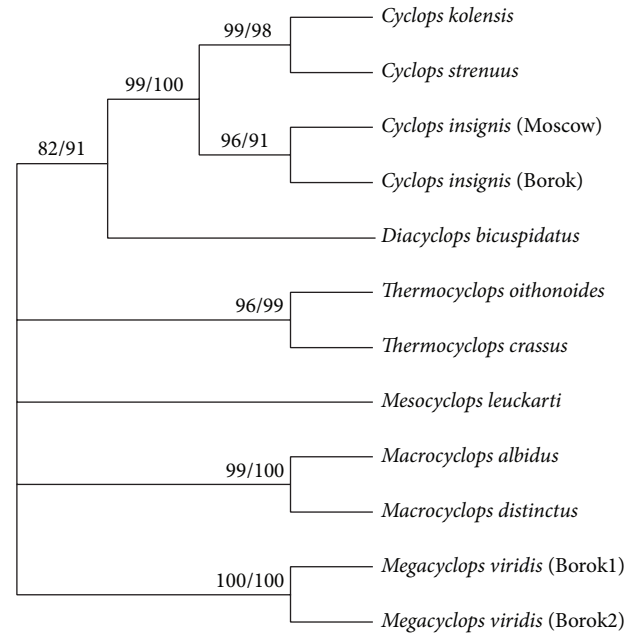


Figure 3: Phylogenetic relationships of Cyclopoida based on ~500 bp of concatenated ITS1/ITS2 rDNA sequences. The consensus cladogram inferred from the ITS1-ITS2 ribosomal DNA fragment sequence data of 10 species of the Cyclopidae family using maximum likelihood analysis under the HKY+G model and minimum evolution (ME) analysis. The numbers above branches indicate bootstrap percentages. The values are listed for ML/ME.

this study, the pairwise ITS1/ITS2 *p*-distances were significantly higher thanthe 28S *p*-distances (compare Tables 2 and 3). These data are consistent with other studies showing considerable variation in ITS1 and ITS2 divergence levels among different groups of copepods [40, 42, 52, 124, 125]. In this study, fragments of 28S rDNA sequences were used for the analysis of marine and freshwater cyclopoid copepods species, whereas ITS1 and ITS2 sequences were used exclusively for the analysis of freshwater cyclopoid copepods species.

The cladogram based on comparison of the 28S rDNA sequences reflected the evolutionary history of the analyzed species (Figure 2). *Oncaea* sp. was used as the out group. Similar topologies and levels of support at most nodes were obtained for all 28S phylogenetic trees constructed using the ML and ME methods. The specimens belonging to the order Cyclopoida with high bootstrap support (ML/ME 99) formed two major clades on the tree (Figure 2). One clade combined the marine cyclopoid copepods species, whereas the freshwater species specimens formed the second clade. The *p*-distance between these two clades varied in the range of 0.171–0.245 (Table 2). The 28S phylogenetic tree revealed detailed relationships among the *Oithona* spp. with high bootstrap support (ML 79, 100 and ME 76, 100). However, *P. nana* (Cyclopettidae family) and *Oithona* spp. (Oithonidae family) were poorly resolved. Notably, this study is the second on the molecular phylogenetics of the *Oithona* spp.; the previous study described the phylogenetic relationships between three *Oithona* spp.: *O. similis*, *O. atlantica,* and *O. nana* [22].

The cladogram based on the comparison of the concatenated ITS1/ITS2 sequences is shown in Figure 3. Notably, the 28S and ITS1/ITS2 cladograms had several common features, reflecting the evolutionary history of the analyzed freshwater cyclopoid copepods species. Both cladograms revealed that *D. bicuspidatus* and specimens of the *Cyclops* genus with high bootstrap values (>80) are separated from other studied freshwater copepods in a distinct clade. The *p*-distance between *D. bicuspidatus* and *Cyclops* spp. calculated based on ITS1/ITS2 analysis varied in the range of 0.232–0.250, whereas the *p*-distance between *D. bicuspidatus* and *Thermocyclops* spp. varied in the range of 0.298–0.333, and the *p*-distance between *D. bicuspidatus* and other analyzed freshwater species varied in the range of 0.310–0.405. This result is consistent with a previous phylogenetic study based on 18S rDNA sequence analysis [48]. Notably, the systematic position of this species, based solely on the analysis of morphological characteristics, remained unclear. *Diacyclops bicuspidatus* is considered to be evolutionarily closer to *Thermocyclops* spp. [8].

Another important conclusion from the analysis of 28S and ITS1/ITS2 cladograms relates to the systematic position of *C. strenuus*. The *Cyclops* genera subclade was divided into

Table 2: Pairwise uncorrected genetic distance among 18 sequences of Cyclopoida based on comparison of 28S rRNA gene.

#	Species name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	*Oncaea* sp.																	
2	*Paracyclopina nana*	0.229																
3	*Oithona similis*	0.230	0.148															
4	*Oithona helgolandica*	0.227	0.143	0.011														
5	*Oithona nana*	0.233	0.181	0.155	0.146													
6	*Oithona simplex*	0.244	0.152	0.175	0.166	0.177												
7	*Cyclops kolensis*	0.209	0.223	0.197	0.195	0.244	0.212											
8	*Cyclops strenuus*	0.207	0.220	0.197	0.195	0.242	0.207	0.005										
9	*Cyclops insignis* Moscow	0.216	0.223	0.204	0.203	0.245	0.201	0.037	0.032									
10	*C. insignis* Borok	0.209	0.216	0.198	0.197	0.241	0.197	0.027	0.023	0.014								
11	*Diacyclops bicuspidatus*	0.186	0.200	0.183	0.180	0.224	0.194	0.090	0.085	0.098	0.087							
12	*Thermocyclops oithonoides*	0.220	0.203	0.191	0.186	0.216	0.184	0.128	0.123	0.120	0.116	0.098						
13	*Thermocyclops crassus*	0.218	0.203	0.181	0.178	0.220	0.204	0.127	0.125	0.127	0.125	0.093	0.084					
14	*Mesocyclops leuckarti*	0.198	0.183	0.174	0.171	0.215	0.181	0.133	0.128	0.123	0.122	0.090	0.087	0.079				
15	*Macrocyclops albidus*	0.221	0.198	0.197	0.191	0.215	0.178	0.145	0.143	0.137	0.134	0.099	0.131	0.114	0.105			
16	*Macrocyclops distinctus*	0.230	0.216	0.192	0.192	0.218	0.191	0.152	0.148	0.143	0.142	0.113	0.127	0.143	0.131	0.123		
17	*Megacyclops viridis* Borok1	0.213	0.207	0.195	0.192	0.230	0.191	0.137	0.134	0.134	0.123	0.105	0.116	0.117	0.114	0.130	0.139	
18	*M. viridis* Borok2	0.207	0.210	0.206	0.203	0.236	0.198	0.123	0.122	0.127	0.119	0.085	0.114	0.110	0.105	0.125	0.130	0.050

All positions containing gaps and missing data were eliminated (complete deletion option). Final dataset, 657 positions.

Table 3: Pairwise uncorrected genetic distance among 12 Cyclopidae sequences based on comparison of concatenated ITS1/ITS2 rRNA sequences.

#	Species name	1	2	3	4	5	6	7	8	9	10	11
1	*Cyclops kolensis*											
2	*Cyclops strenuus*	0.006										
3	*Cyclops insignis* Moscow	0.054	0.048									
4	*C. insignis* Borok	0.048	0.042	0.006								
5	*Diacyclops bicuspidatus*	0.250	0.244	0.232	0.232							
6	*Thermocyclops oithonoides*	0.333	0.330	0.330	0.330	0.313						
7	*Thermocyclops crassus*	0.301	0.298	0.295	0.295	0.262	0.188					
8	***Mesocyclops leuckarti***	0.324	0.321	0.310	0.310	0.286	0.280	0.241				
9	*Macrocyclops distinctus*[*]	0.414	0.411	0.408	0.405	0.393	0.429	0.408	0.369[*]			
10	*Macrocyclops albidus*	0.393	0.390	0.381	0.378	0.357	0.369	0.366	0.336	0.321		
11	*Megacyclops viridis* Borok1[*]	0.381	0.384	0.396	0.390	0.387	0.387	0.387	0.390[*]	0.455	0.438	
12	*M. viridis* Borok2[*]	0.375	0.378	0.384	0.378	0.363	0.366	0.357	0.381[*]	0.438	0.429	0.149

All positions containing gaps and missing data were eliminated (complete deletion option). Final datasets, 336 positions for concatenated ITS1/ITS2. The names of the three taxa most evolutionarily distant from *Mesocyclops* and the values of the corresponding genetic distances are shown with asterisk.

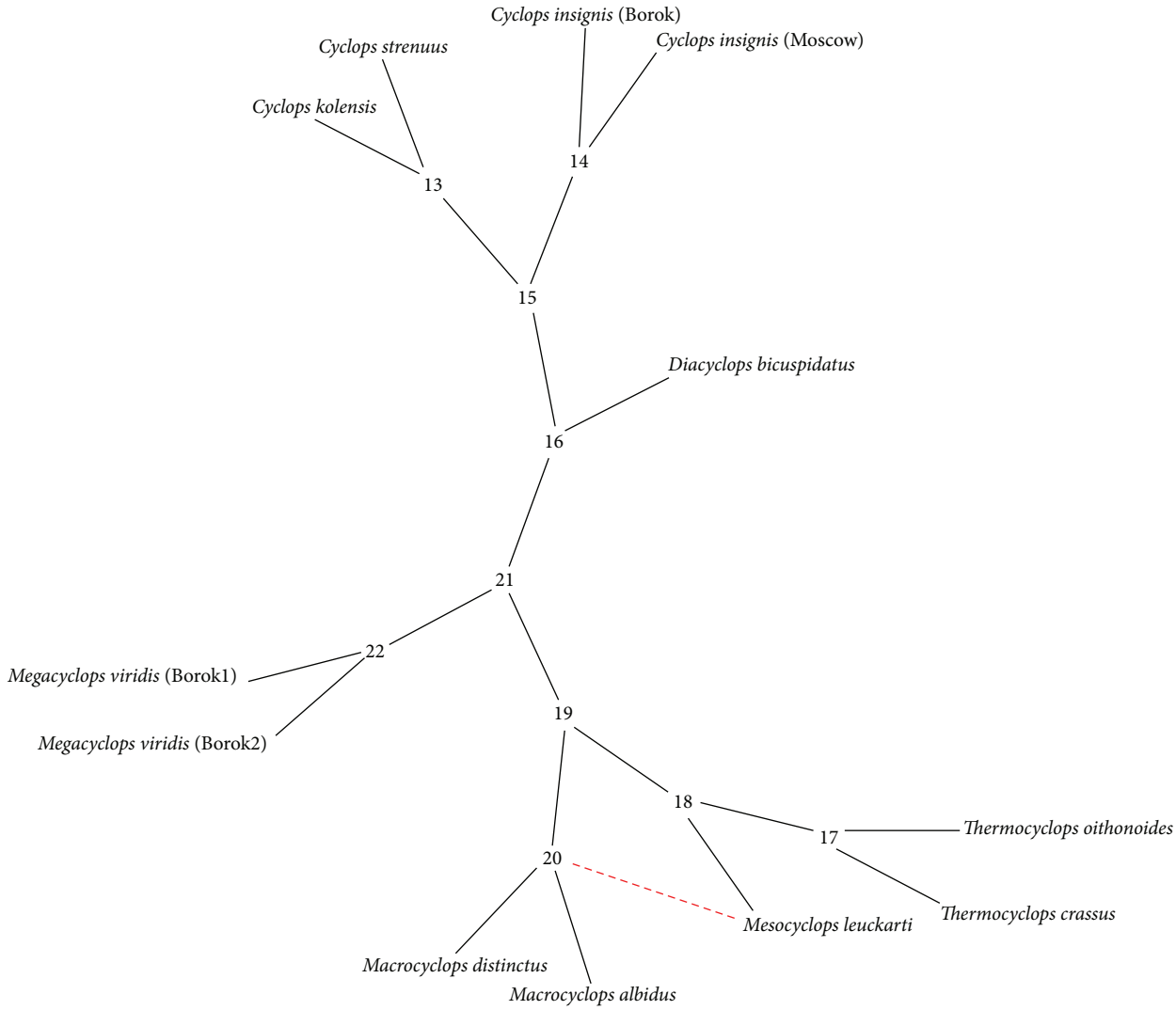


Figure 4: Reticulogram for concatenated ITS1/ITS2 sequences of 10 species of the Cyclopidae family. The red dashed line indicates the reticulation event connecting *M. leuckarti* to the *Macrocyclops* clade node. The number of internal vertices begins with $n + 1$, where n is the number of leaves. The order of internal vertices distribution corresponds to the increasing lengths of the 22 reticulogram edges.

the *C. kolensis*: *C. strenuus* subsubclade and the *C. insignis* subsubclade (Figures 2 and 3). The *p*-distance between *C. strenuus* and *C. kolensis* calculated based on ITS1/ITS2 analysis is 0.006, whereas the *p*-distance between *C. strenuus* and *C. insignis* varied in the range of 0.042–0.054. Therefore, *C. strenuus* is more closely related to *C. kolensis* than to *C. insignis*. Notably, the phylogenetic relationships between the studied *Cyclops* species could not be elucidated solely on the basis of morphological characteristics.

The only difference between the 28S and ITS1/ITS2 cladograms within freshwater copepods was the position of *M. leuckarti* (Figures 2 and 3). The cladogram based on comparison of the 28S rDNA sequences showed that the *M. leuckarti* and *Thermocyclops* cluster together to form a separate subclade (Figure 2). This result is consistent with the previous observation that the *Mesocyclops* and *Thermocyclops* genera are phylogenetically closely related, which was confirmed by the similarity of morphological characteristics and using molecular data [42]. ITS1/ITS2 analysis revealed that *M. leuckarti* is located separately from *Thermocyclops* and other clades (Figure 3). Using phylogenetic networks, we analyzed whether the *M. leuckarti* position in the ITS1/ITS2 cladogram was caused by different contributions of ITS1 and ITS2 sequences to the phylogenetic signal.

3.3. Phylogenetic Networks. A reticulogram-based phylogenetic network inference approach was used to verify the reticulate evolution of the studied copepods. Concatenated ITS1/ITS2 sequences of 10 species from the Cyclopidae family were used for reticulogram reconstruction. The reticulogram revealed a network with *Mesocyclops* and *Thermocyclops* clustered together and a reticulation (lateral branch) connecting *M. leuckarti* to the *Macrocyclops* clade node (Figure 4). Therefore, the reticulogram indicated the reticulation in *Mesocyclops* evolution.

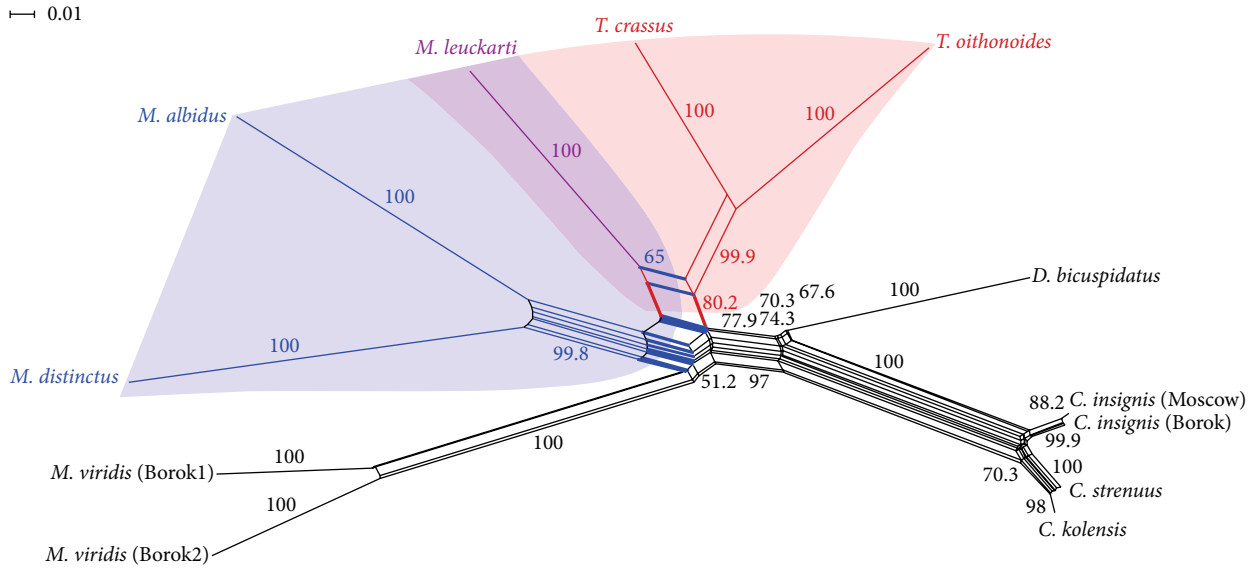


FIGURE 5: Split networks for concatenated ITS1/ITS2 sequences of 10 species of the Cyclopidae family. Split network based on concatenated ITS1/ITS2 sequences; the split separating *M. leuckarti*, *T. oithonoides,* and *T. crassus* is indicated in bold red; the split separating *M. leuckarti*, *M. distinctus,* and *M. albidus* is indicated in bold blue; purple indicates the *M. leuckarti* reticulate relationship with *Thermocyclops* and *Macrocyclops*. The values on the branches indicate bootstrap percentages.

A split network represents incompatible edges of trees as a band of parallel edges. Parallel edges split a network into two sets of nodes. Split-graph analysis of concatenated ITS1/ITS2 sequences of 10 species from the Cyclopidae family revealed a reticulate relationship between *Mesocyclops*, *Thermocyclops,* and *Macrocyclops* with high reliability (Figure 5). All principal splits were well supported. Two splits were observed in the ITS1/ITS2 split network. The first split (parallel edges highlighted with bold red) separated *M. leuckarti*, *T. oithonoides,* and *T. crassus* with 80.2% bootstrap support. The second split (parallel edges highlighted with bold blue) separated *M. leuckarti*, *M. distinctus*, and *M. albidus* with 65.0% bootstrap support.

In addition to the network data, we performed phylogenetic reconstruction based on independent ITS1 and ITS2 analyses using probabilistic and distance methods. Irrespective of the method used, the main difference between the topologies of the ITS1 and ITS2 phylogenetic trees was as follows: based on the ITS1 analysis, *M. leuckarti* is clustered with *Thermocyclops*, whereas the ITS2 analysis revealed that *M. leuckarti* clustered with *Macrocyclops* (Figures 6(a)–6(d)).

The impact of the chosen DNA sequence on the clustering of *M. leuckarti* might reflect the different evolutionary histories of ITS1 and ITS2, which indicates the potential hybrid origin of *M. leuckarti*. However, the values of bootstrap support for the clustering of *Mesocyclops* and *Thermocyclops* and of *Mesocyclops* and *Macrocyclops* depended on the method used for phylogenetic tree reconstruction and varied over a wide range (Figures 6(a)–6(d)).

Phylogenetic trees can be inconsistent due to the so-called long-branch attraction (LBA) phenomenon, which occurs when two nonadjacent taxa share many homoplastic character states along long branches and/or from uncorrected sequence alignments. Interpretation of the observed similarity depends on the method used for phylogenetic analysis, and this similarity can often be interpreted as homology. Model-based methods are most resistant to LBA, but these methods can exhibit LBA if their assumptions are seriously violated or if there are insufficient taxa in the analysis to accurately estimate the parameters of the evolutionary model [126]. Taxon sampling is a crucial factor for avoiding LBA in phylogenetic analysis [127]. The inclusion of additional taxa in phylogenetic analysis increases the accuracy of the inferred topology by dispersing homoplasty across the tree and reducing the effect of LBA. The LBA effect might also be revealed by exclusion of the long-branched taxon from the analysis [127].

To reduce the possible effects of LBA and correctness of the sequence alignment, we used the following approaches: (1) three taxa the most evolutionarily distant from *M. leuckarti* (*M. distinctus*, *M. viridis* Borok1, and *M. viridis* Borok2) were removed from the list of species used for ITS1 and ITS2 phylogenetic tree reconstruction (the taxa selection was based on the data presented in Table 3) and (2) ITS1 and ITS2 sequences were aligned using new multiple sequence alignment programs to eliminate poorly aligned and highly divergent regions (see Section 2). The final ITS1 and ITS2 alignments are shown in Supplementary Figures 2(a) and 2(b), and the resulting phylogenetic trees are shown in Figures 6(e) and 6(f). Based on the ITS1 analysis, *M. leuckarti* clustered with *Thermocyclops*, whereas the ITS2 analysis revealed that *M. leuckarti* clustered with *Macrocyclops* (Figures 6(e) and 6(f)). Notably, the topology of the new phylogenetic trees had high bootstrap support: 84 (ML)/77 (ME) for ITS1 and 77 (ML)/72 (ME) for ITS2 (Figures 6(e) and 6(f)).

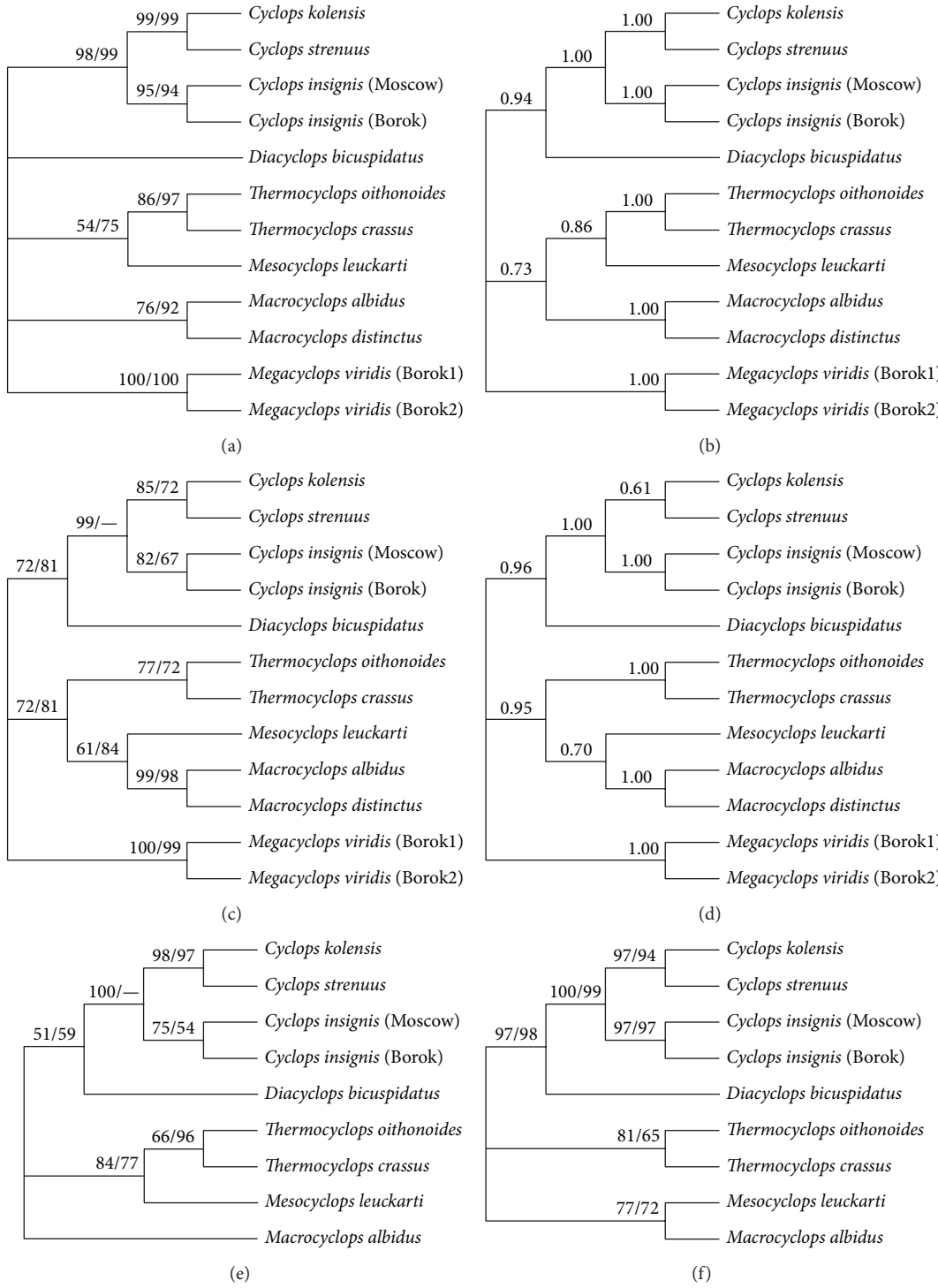


Figure 6: Phylogenetic relationships of Cyclopidae based on ITS1 and ITS2 sequences. The ITS1 consensus Clustal tree of ten Cyclopidae species constructed using (a) maximum likelihood and minimum evolution and (b) Bayesian inference. The ITS2 consensus Clustal tree of ten Cyclopidae species constructed using (c) maximum likelihood and minimum evolution and (d) Bayesian inference. The consensus Gblocks-treated MAFFT trees of eight Cyclopidae species constructed using maximum likelihood and minimum evolution: (e) ITS1, (f) ITS2. The numbers above the branches indicate bootstrap percentages and Bayesian posterior probabilities. The values are listed for ML/ME. Missing or weakly supported nodes (<50% or 0.5) are denoted by a "—".

We think that one of the most intriguing explanations for the observed differences in the clustering of *M. leuckarti* is the interspecific hybridization between extinct taxa (presumably closely related) that were ancestral to both *Mesocyclops*, *Macrocyclops*, and *Thermocyclops*. However, a rigorous proof of this hypothesis requires further analysis of a larger number of species. This will be the subject of our further research.

4. Conclusion

We evaluated the utility of a ~2000 bp fragment of rDNA (easily amplified by universal primers) for the phylogenetic reconstruction of the relationships of Copepoda species. Our data showed that the 28S rDNA and the ITS1 and ITS2 regions are highly informative for the phylogeny of both higher-level and closely related Copepoda species. Comparative analysis of the ITS1 and ITS2 nucleotide sequences among closely related Copepoda species revealed an unusual evolutionary history of these spacer sequences; therefore, the ITS1 and ITS2 regions might contain different phylogenetic signals.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank Grace A. Wyngaard for criticisms and comments on earlier versions of this paper. This project was supported in part by a Russian Foundation for Basic Research Grant no. 12-04-32032-a to Maxim V. Zagoskin, Grants nos. 06-04-48474-a and 10-04-01376-a to Andrey K. Grishanin, and a Russian Academy of Sciences grant "Wild life: current status and development problems" (subprogram "Dynamics and Conservation of Gene Pools") to Dmitry V. Mukha.

References

[1] G. A. Boxshall and D. Defaye, "Global diversity of copepods (Crustacea: Copepoda) in freshwater," *Hydrobiologia*, vol. 595, no. 1, pp. 195–207, 2008.

[2] B. Dussart and D. Defaye, *World Directory of Crustacea Copepoda of Inland Waters. II—Cyclopiformes*, Backhuys, Leiden, The Netherlands, 2006.

[3] F. Kiefer, "Versuch eines Systems der Cyclopiden," *Zoologischer Anzeiger*, vol. 73, no. 11-12, pp. 302–308, 1927.

[4] R. Gurney, *British Fresh-Water Copepoda. III. Cyclopoida*, Ray Society, London, UK, 1933.

[5] V. M. Rylov, *Cyclopoida presnykh vod. Freshwater Cyclopoida V.3*, Leningrad, Moscow, Russia, 1948.

[6] H. C. Yeatman, "Free-living Copepoda. Cyclopoida," in *Freshwater Biology*, W. T. Edmondson, Ed., pp. 795–815, John Wiley & Sons, New York, NY, USA, 1959.

[7] B. Dussart, *Les Copépodes des Eaux Continentales de Europe Occidentals, Vol. 2. Cyclopoides et Biologie*, N. Boubée, Paris, France, 1969.

[8] V. I. Monchenko, "Gnathostomata cyclopoida: cyclopidae," *The Fauna of Ukraina Naukova Dumka, Kiev, USSR*, vol. 27, no. 3, pp. 1–452, 1974, http://scholar.google.ru/citations?view_op=view_citation&hl=ru&user=b97TEWUAAAAJ&citation_for_view=b97TEWUAAAAJ:hqOjcs7Dif8C.

[9] A. Bucklin, B. W. Frost, and T. D. Kocher, "DNA sequence variation of the mitochondrial 16S rRNA in *Calanus* (Copepoda: Calanoida): intraspecific and interspecific patterns," *Molecular Marine Biology and Biotechnology*, vol. 1, no. 6, pp. 397–407, 1992.

[10] A. Bucklin, T. C. LaJeunesse, E. Curry, J. Wallinga, and K. Garrison, "Molecular diversity of the copepod, *Nannocalanus minor*: genetic evidence of species and population structure in the North Atlantic Ocean," *Journal of Marine Research*, vol. 54, no. 2, pp. 285–310, 1996.

[11] R. S. Burton, "Intraspecific phylogeography across the point conception biogeographic boundary," *Evolution*, vol. 52, no. 3, pp. 734–745, 1998.

[12] C. C. Caudill and A. Bucklin, "Molecular phylogeography and evolutionary history of the estuarine copepod, *Acartia tonsa*, on the Northwest Atlantic coast," *Hydrobiologia*, vol. 511, pp. 91–102, 2004.

[13] S. Edmands, "Phylogeography of the intertidal copepod *Tigriopus californicus* reveals substantially reduced population differentiation at northern latitudes," *Molecular Ecology*, vol. 10, no. 7, pp. 1743–1750, 2001.

[14] S. Eyun, Y. Lee, H. Suh, S. Kim, and Y. S. Ho, "Genetic identification and molecular phylogeny of *Pseudodiaptomus* species (Calanoida, Pseudodiaptomidae) in Korean waters," *Zoological Science*, vol. 24, no. 3, pp. 265–271, 2007.

[15] E. Goetze, "Population differentiation in the open sea: insights from the pelagic copepod pleuromamma xiphias," *Integrative and Comparative Biology*, vol. 51, no. 4, pp. 580–597, 2011.

[16] S. Laakmann and S. Holst, "Emphasizing the diversity of North Sea hydromedusae by combined morphological and molecular methods," *Journal of Plankton Research*, vol. 36, no. 1, pp. 64–76, 2014.

[17] P. K. Lindeque, R. P. Harris, M. B. Jones, and G. R. Smerdon, "Distribution of *Calanus* spp. as determined using a genetic identification system," *Scientia Marina*, vol. 68, supplement 1, pp. 121–128, 2007.

[18] W. Minxiao, S. Song, L. Chaolun, and S. Xin, "Distinctive mitochondrial genome of Calanoid copepod *Calanus sinicus* with multiple large non-coding regions and reshuffled gene order: useful molecular markers for phylogenetic and population studies," *BMC Genomics*, vol. 12, no. 1, article 73, 2011.

[19] P. D. Rawson and R. S. Burton, "Molecular evolution at the cytochrome oxidase subunit 2 gene among divergent populations of the intertidal copepod, *Tigriopus californicus*," *Journal of Molecular Evolution*, vol. 62, no. 6, pp. 753–764, 2006.

[20] H. Y. Soh, E. Ok Park, B. A. Venmathi Maran, and S. Yong Moon, "A new species of Acartia subgenus Euacartia (Copepoda: Calanoida: Acartiidae) from Korean estuaries based on morphological and molecular evidence," *Journal of Crustacean Biology*, vol. 33, no. 5, pp. 718–729, 2013.

[21] C. S. Willett and J. T. Ladner, "Investigations of fine-scale phylogeography in *Tigriopus californicus* reveal historical patterns of population divergence," *BMC Evolutionary Biology*, vol. 9, no. 1, article 139, 2009.

[22] G. D. Cepeda, L. Blanco-Bercial, A. Bucklin, C. M. Berón, and M. D. Viñas, "Molecular systematic of three species of Oithona

(Copepoda, Cyclopoida) from the Atlantic ocean: comparative analysis using 28S rDNA," *PLoS ONE*, vol. 7, no. 4, Article ID e35861, 2012.

[23] J. Hirai, S. Shimode, and A. Tsuda, "Evaluation of ITS2-28S as a molecular marker for identification of calanoid copepods in the subtropical western North Pacific," *Journal of Plankton Research*, vol. 35, no. 3, pp. 644–656, 2013.

[24] J. Kim and W. Kim, "Molecular phylogeny of poecilostome copepods based on the 18S rDNA sequences," *Korean Journal of Biological Sciences*, vol. 4, no. 3, pp. 257–261, 2000.

[25] A. P. Shinn, B. A. Banks, N. Tange et al., "Utility of 18S rDNA and ITS sequences as population markers for *Lepeophtheirus salmonis* (Copepoda: Caligidae) parasitising *Atlantic salmon* (Salmo salar) in Scotland," *Contributions to Zoology*, vol. 69, no. 1-2, pp. 89–98, 2000.

[26] S. J. Adamowicz, S. Menu-Marque, S. A. Halse et al., "The evolutionary diversification of the Centropagidae (Crustacea, Calanoida): a history of habitat shifts," *Molecular Phylogenetics and Evolution*, vol. 55, no. 2, pp. 418–430, 2010.

[27] L. Blanco-Bercial, J. Bradford-Grieve, and A. Bucklin, "Molecular phylogeny of the Calanoida (Crustacea: Copepoda)," *Molecular Phylogenetics and Evolution*, vol. 59, no. 1, pp. 103–113, 2011.

[28] E. Braga, R. Zardoya, A. Meyer, and J. Yen, "Mitochondrial and nuclear rRNA based copepod phylogeny with emphasis on the Euchaetidae (Calanoida)," *Marine Biology*, vol. 133, no. 1, pp. 79–90, 1999.

[29] A. Bucklin, B. W. Frost, J. Bradford-Grieve, L. D. Allen, and N. J. Copley, "Molecular systematic and phylogenetic assessment of 34 calanoid copepod species of the Calanidae and Clausocalanidae," *Marine Biology*, vol. 142, no. 2, pp. 333–343, 2003.

[30] A. Bucklin and B. W. Frost, "Morphological and molecular Phylogenetic analysis of evolutionary lineages within *Clausocalanus* (Copepoda: Calanoida)," *Journal of Crustacean Biology*, vol. 29, no. 1, pp. 111–120, 2009.

[31] A. Cornils and L. Blanco-Bercial, "Phylogeny of the paracalanidae giesbrecht, 1888 (Crustacea: Copepoda: Calanoida)," *Molecular Phylogenetics and Evolution*, vol. 69, no. 3, pp. 861–872, 2013.

[32] S. M. Dippenaar, "Estimated molecular phylogenetic relationships of six siphonostomatoid families (Copepoda) symbiotic on elasmobranchs," *Crustaceana*, vol. 82, no. 12, pp. 1547–1567, 2009.

[33] D. F. Figueroa, "Phylogenetic analysis of *Ridgewayia* (Copepoda: Calanoida) from the Galapagos and of a new species from the Florida keys with a reevaluation of the phylogeny of Calanoida," *Journal of Crustacean Biology*, vol. 31, no. 1, pp. 153–165, 2011.

[34] E. Goetze, "Cryptic speciation on the high seas; global phylogenetics of the copepod family Eucalanidae," *Proceedings of the Royal Society B: Biological Sciences*, vol. 270, no. 1531, pp. 2321–2331, 2003.

[35] S. Laakmann, G. Gerdts, R. Erler, T. Knebelsberger, P. Martínez Arbizu, and M. J. Raupach, "Comparison of molecular species identification for North Sea calanoid copepods (Crustacea) using proteome fingerprints and DNA sequences," *Molecular Ecology Resources*, vol. 13, no. 5, pp. 862–876, 2013.

[36] R. J. Machida, M. U. Miya, M. Nishida, and S. Nishida, "Molecular phylogeny and evolution of the pelagic copepod genus *Neocalanus* (Crustacea: Copepoda)," *Marine Biology*, vol. 148, no. 5, pp. 1071–1079, 2006.

[37] M. Taniguchi, "Molecular phylogeny of *Neocalanus* copepods in the subarctic Pacific Ocean, with notes on non-geographical genetic variations for *Neocalanus cristatus*," *Journal of Plankton Research*, vol. 26, no. 10, pp. 1249–1255, 2004.

[38] C. E. Lee, "Global phylogeography of a cryptic copepod species complex and reproductive isolation between genetically proximate 'populations'," *Evolution*, vol. 54, no. 6, pp. 2014–2027, 2000.

[39] S. J. Adamowicz, S. Menu-Marque, P. D. N. Hebert, and A. Purvis, "Molecular systematics and patterns of morphological evolution in the Centropagidae (Copepoda: Calanoida) of Argentina," *Biological Journal of the Linnean Society*, vol. 90, no. 2, pp. 279–292, 2007.

[40] E. Goetze, "Global population genetic structure and biogeography of the oceanic copepods *Eucalanus hyalinus* and *E. spinifer*," *Evolution*, vol. 59, no. 11, pp. 2378–2398, 2005.

[41] R. J. Machida and A. Tsuda, "Dissimilarity of species and forms of planktonic *Neocalanus* copepods using mitochondrial COI, 12S, nuclear ITS, and 28S gene sequences," *PLoS ONE*, vol. 5, no. 4, Article ID e10278, 2010.

[42] G. A. Wyngaard, M. Hołyńska, and J. A. Schulte, "Phylogeny of the freshwater copepod *Mesocyclops* (Crustacea: Cyclopidae) based on combined molecular and morphological data, with notes on biogeography," *Molecular Phylogenetics and Evolution*, vol. 55, no. 3, pp. 753–764, 2010.

[43] V. Alekseev, H. J. Dumont, J. Pensaert, D. Baribwegure, and J. R. Vanfleteren, "A redescription of *Eucyclops serrulatus* (Fischer, 1851) (Crustacea: Copepoda: Cyclopoida) and some related taxa, with a phylogeny of the *E. serrulatus*-group," *Zoologica Scripta*, vol. 35, no. 2, pp. 123–147, 2006.

[44] M. R. Miracle, V. Alekseev, V. Monchenko, V. Sentandreu, and E. Vicente, "Molecular-genetic-based contribution to the taxonomy of the *Acanthocyclops robustus* group," *Journal of Natural History*, vol. 47, no. 5–12, pp. 863–888, 2013.

[45] M. Bláha, M. Hulák, J. Slouková, and J. Těšitel, "Molecular and morphological patterns across *Acanthocyclops vernalis-robustus* species complex (Copepoda, Cyclopoida)," *Zoologica Scripta*, vol. 39, no. 3, pp. 259–268, 2010.

[46] T. Karanovic and M. Krajicek, "When anthropogenic translocation meets cryptic speciation globalized bouillon originates; molecular variability of the cosmopolitan freshwater cyclopoid *Macrocyclops albidus* (Crustacea: Copepoda)," *Annales de Limnologie*, vol. 48, no. 1, pp. 63–80, 2012.

[47] T. Karanovic and M. Krajicek, "First molecular data on the Western Australian *Diacyclops* (Copepoda, Cyclopoida) confirm morpho-species but question size differentiation and monophyly of the *Alticola*-group," *Crustaceana*, vol. 85, no. 12-13, pp. 1549–1569, 2012.

[48] T. Y. Mayor, N. G. Sheveleva, L. V. Sukhanova, O. A. Timoshkin, and S. V. Kiril'chik, "Molecular-phylogenetic analysis of cyclopoids (Copepoda: Cyclopoida) from Lake Baikal and its water catchment basin," *Russian Journal of Genetics*, vol. 46, no. 11, pp. 1373–1380, 2010.

[49] T. Karanovic and S. J. B. Cooper, "Molecular and morphological evidence for short range endemism in the *Kinnecaris solitaria* complex (Copepoda: Parastenocarididae), with descriptions of seven new species," *Zootaxa*, no. 3026, pp. 1–64, 2011.

[50] F. Marrone, S. L. Brutto, and M. Arculeo, "Molecular evidence for the presence of cryptic evolutionary lineages in the freshwater copepod genus *Hemidiaptomus* G.O. Sars, 1903 (Calanoida, Diaptomidae)," *Hydrobiologia*, vol. 644, no. 1, pp. 115–125, 2010.

[51] R. Scheihing, L. Cardenas, R. F. Nespolo et al., "Morphological and molecular analysis of centropagids from the high Andean

plateau (Copepoda: Calanoidea)," *Hydrobiologia*, vol. 637, no. 1, pp. 45–52, 2010.

[52] R. A. Thum and R. G. Harrison, "Deep genetic divergences among morphologically similar and parapatric *Skistodiaptomus* (Copepoda: Calanoida: Diaptomidae) challenge the hypothesis of Pleistocene speciation," *Biological Journal of the Linnean Society*, vol. 96, no. 1, pp. 150–165, 2009.

[53] H. Y. Soh, S. W. Kwon, W. Lee, and Y. H. Yoon, "A new *Pseudodiaptomus* (Copepoda, Calanoida) from Korea supported by molecular data," *Zootaxa*, no. 3368, pp. 229–244, 2012.

[54] M. A. Marszalek, S. Dayanandan, and E. J. Maly, "Phylogeny of the genus *Hesperodiaptomus* (Copepoda) based on nucleotide sequence data of the nuclear ribosomal gene," *Hydrobiologia*, vol. 624, no. 1, pp. 61–69, 2009.

[55] R. A. Thum, "Using 18S rDNA to resolve diaptomid copepod (Copepoda: Calanoida: Diaptomidae) phylogeny: an example with the North American genera," *Hydrobiologia*, vol. 519, no. 1–3, pp. 135–141, 2004.

[56] R. Huys, J. Llewellyn-Hughes, P. D. Olson, and K. Nagasawa, "Small subunit rDNA and Bayesian inference reveal *Pectenophilus ornatus* (Copepoda *incertae sedis*) as highly transformed Mytilicolidae, and support assignment of Chondracanthidae and Xarifiidae to Lichomolgoidea (Cyclopoida)," *Biological Journal of the Linnean Society*, vol. 87, no. 3, pp. 403–425, 2006.

[57] R. Huys, J. Llewellyn-Hughes, S. Conroy-Dalton, P. D. Olson, J. N. Spinks, and D. A. Johnston, "Extraordinary host switching in siphonostomatoid copepods and the demise of the Monstrilloida: integrating molecular data, ontogeny and antennulary morphology," *Molecular Phylogenetics and Evolution*, vol. 43, no. 2, pp. 368–378, 2007.

[58] J. Ki, K. Lee, H. G. Park, S. Chullasorn, H. Dahms, and J. Lee, "Phylogeography of the copepod *Tigriopus japonicus* along the Northwest Pacific rim," *Journal of Plankton Research*, vol. 31, no. 2, pp. 209–221, 2009.

[59] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. DeWaard, "Biological identifications through DNA barcodes," *Proceedings of the Royal Society B*, vol. 270, no. 1512, pp. 313–321, 2003.

[60] R. J. Machida, M. U. Miya, M. Nishida, and S. Nishida, "Complete mitochondrial DNA sequence of *Tigriopus japonicus* (Crustacea: Copepoda)," *Marine Biotechnology*, vol. 4, no. 4, pp. 406–417, 2002.

[61] R. S. Burton, R. J. Byrne, and P. D. Rawson, "Three divergent mitochondrial genomes from California populations of the copepod *Tigriopus californicus*," *Gene*, vol. 403, no. 1-2, pp. 53–59, 2007.

[62] S. A. Gerbi, "Evolution of ribosomal DNA," in *Molecular Evolutionary Genetics*, pp. 419–517, Plenum, New York, NY, USA, 1985.

[63] U. W. Hwang and W. Kim, "General properties and phylogenetic utilities of nuclear ribosomal DNA and mitochondrial DNA commonly used in molecular systematics.," *Korean Journal of Parasitology*, vol. 37, no. 4, pp. 215–228, 1999.

[64] D. M. Hillis and M. T. Dixon, "Ribosomal DNA: molecular evolution and phylogenetic inference," *Quarterly Review of Biology*, vol. 66, no. 4, pp. 411–453, 1991.

[65] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, 2006.

[66] D. M. Hillis, C. Moritz, and B. K. Mable, Eds., *Molecular Systematics*, Sinauer Associates, Sunderland, Mass, USA, 1996.

[67] E. V. Koonin, *The Logic of Chance: The Nature and Origin of Biological Evolution*, Pearson Education, FT Press, 2012.

[68] T. Sang, D. J. Crawford, and T. F. Stuessy, "Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 15, pp. 6813–6817, 1995.

[69] S. Giessler and C. C. Englbrecht, "Dynamic reticulate evolution in a *Daphnia* multispecies complex," *Journal of Experimental Zoology A, Ecological Genetics and Physiology*, vol. 311, no. 7, pp. 530–548, 2009.

[70] H. Kauserud and T. Schumacher, "Ribosomal DNA variation, recombination and inheritance in the basidiomycete *Trichaptum abietinum*: implications for reticulate evolution," *Heredity*, vol. 91, no. 2, pp. 163–172, 2003.

[71] P. M. W. Wyatt, C. S. Pitts, and R. K. Butlin, "A molecular approach to detect hybridization between bream *Abramis brama*, roach *Rutlius rutilus* and rudd *Scardinius erythrophthalmus*," *Journal of Fish Biology*, vol. 69, pp. 52–71, 2006.

[72] J. Fuertes Aguilar and G. Nieto Feliner, "Additive polymorphisms and reticulation in an ITS phylogeny of thrifts (*Armeria*, Plumbaginaceae)," *Molecular Phylogenetics and Evolution*, vol. 28, no. 3, pp. 430–447, 2003.

[73] H. Yamaji, T. Fukuda, J. Yokoyama et al., "Reticulate evolution and phylogeography in *Asarum*sect. *Asiasarum* (Aristolochiaceae) documented in internal transcribed spacer sequences (ITS) of nuclear ribosomal DNA," *Molecular Phylogenetics and Evolution*, vol. 44, no. 2, pp. 863–884, 2007.

[74] M. A. Hershkovitz, C. C. Hernández-Pellicer, and M. T. K. Arroyo, "Ribosomal DNA evidence for the diversification of *Tropaeolum* sect. *Chilensia* (Tropaeolaceae)," *Plant Systematics and Evolution*, vol. 260, no. 1, pp. 1–24, 2006.

[75] A. Hugall, J. Stanton, and C. Moritz, "Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic *Meloidogyne*," *Molecular Biology and Evolution*, vol. 16, no. 2, pp. 157–164, 1999.

[76] J. F. Wendel, A. Schnabel, and T. Seelanan, "Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 1, pp. 280–284, 1995.

[77] K. O'Donnell and E. Cigelnik, "Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus *Fusarium* are nonorthologous," *Molecular Phylogenetics and Evolution*, vol. 7, no. 1, pp. 103–116, 1997.

[78] K. O'Donnell, E. Cigelnik, and H. I. Nirenberg, "Molecular systematics and phylogeography of the *Gibberella fujikuroi* species complex," *Mycologia*, vol. 90, no. 3, pp. 465–493, 1998.

[79] C. M. Brasier, D. E. L. Cooke, and J. M. Duncan, "Origin of a new *Phytophthora* pathogen through interspecific hybridization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 10, pp. 5878–5883, 1999.

[80] K. W. Hughes and R. H. Petersen, "Apparent recombination or gene conversion in the ribosomal ITS region of a *Flammulina* (Fungi, Agaricales) hybrid," *Molecular Biology and Evolution*, vol. 18, no. 1, pp. 94–96, 2001.

[81] G. Newcombe, B. Stirling, S. McDonald, and H. D. Bradshaw Jr., "*Melampsora* × *columbiana*, a natural hybrid of *M. medusae* and *M. occidentalis*," *Mycological Research*, vol. 104, no. 3, pp. 261–274, 2000.

[82] K. H. Chu, C. P. Li, and H. Y. Ho, "The first internal transcribed spacer (ITS-1) of ribosomal DNA as a molecular marker for phylogenetic and population analyses in crustacea," *Marine Biotechnology*, vol. 3, no. 4, pp. 355–361, 2001.

[83] L. H. Rieseberg, "Hybrid origins of plant species," *Annual Review of Ecology and Systematics*, vol. 28, no. 1, pp. 359–389, 1997.

[84] R. Cui, M. Schumer, K. Kruesi, R. Walter, P. Andolfatto, and G. G. Rosenthal, "Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes," *Evolution*, vol. 67, no. 8, pp. 2166–2179, 2013.

[85] L. Bullini, "Origin and evolution of animal hybrid species," *Trends in Ecology and Evolution*, vol. 9, no. 11, pp. 422–426, 1994.

[86] T. E. Dowling and C. L. Secor, "The role of hybridization and introgression in the diversification of animals," *Annual Review of Ecology and Systematics*, vol. 28, pp. 593–619, 1997.

[87] J. Mallet, "Hybridization as an invasion of the genome," *Trends in Ecology and Evolution*, vol. 20, no. 5, pp. 229–237, 2005.

[88] U. Arnason, R. Spilliaert, A. Pálsdóttir, and A. Arnason, "Molecular identification of hybrids between the two largest whale species, the blue whale (*Balaenoptera musculus*) and the fin whale (*B. physalus*)," *Hereditas*, vol. 115, no. 2, pp. 183–189, 1991.

[89] A. V. Z. Brower, "Introgression of wing pattern alleles and speciation via homoploid hybridization in *Heliconius butterflies*: a review of evidence from the genome.," *Proceedings Biological Sciences*, vol. 280, no. 1752, Article ID 20122302, 2013.

[90] H. L. Carson, K. Y. Kaneshiro, and F. C. Val, "Natural hybridization between the sympatric Hawaiian species *Drosophila silvestris* and *Drosophila heteroneura*," *Evolution*, vol. 43, no. 1, pp. 190–203, 1989.

[91] S. Gießler and C. C. Englbrecht, "Dynamic reticulate evolution in a *Daphnia* multispecies complex," *Journal of Experimental Zoology A: Ecological Genetics and Physiology*, vol. 311, no. 7, pp. 531–549, 2009.

[92] S. Gießler, E. Mader, and K. Schwenk, "Morphological evolution and genetic differentiation in *Daphnia* species complexes," *Journal of Evolutionary Biology*, vol. 12, no. 4, pp. 710–723, 1999.

[93] D. J. Taylor, P. D. N. Hebert, and J. K. Colbourne, "Phylogenetics and evolution of the *Daphnia longispina* group (Crustacea) based on 12S rDNA sequence and allozyme variation," *Molecular Phylogenetics and Evolution*, vol. 5, no. 3, pp. 495–510, 1996.

[94] R. Vergilino, S. Markova, M. Ventura, M. Manca, and F. Dufresne, "Reticulate evolution of the *Daphnia pulex* complex as revealed by nuclear markers," *Molecular Ecology*, vol. 20, no. 6, pp. 1191–1207, 2011.

[95] D. V. Mukha and A. P. Sidorenko, "Detection and analysis of *Tetrahymena pyriformis* 26S ribosomal DNA domain sequences, differing in degree of evolutionary conservation," *Molekulyarnaya Biologiya*, vol. 29, no. 3, pp. 529–537, 1995.

[96] D. V. Mukha and A. P. Sidorenko, "Identification of highly conservative domains within the 17s ribosomal dna sequence from *Tetrahymena pyriformis*," *Genetika*, vol. 32, no. 11, pp. 1494–1497, 1996.

[97] D. V. Mukha, A. P. Sidorenko, I. V. Lazebnaya, B. M. Wiegmann, and C. Schal, "Analysis of intraspecies polymorphism in the ribosomal DNA cluster of the cockroach *Blattella germanica*," *Insect Molecular Biology*, vol. 9, no. 2, pp. 217–222, 2000.

[98] D. Mukha, B. M. Wiegmann, and C. Schal, "Evolution and phylogenetic information content of the ribosomal DNA repeat unit in the Blattodea (Insecta)," *Insect Biochemistry and Molecular Biology*, vol. 32, no. 9, pp. 951–960, 2002.

[99] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463–5467, 1977.

[100] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[101] M. Goujon, H. McWilliam, W. Li et al., "A new bioinformatics analysis tools framework at EMBL-EBI," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq313, pp. W695–W699, 2010.

[102] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[103] X. Xia and Z. Xie, "DAMBE: software package for data analysis in molecular biology and evolution," *Journal of Heredity*, vol. 92, no. 4, pp. 371–373, 2001.

[104] X. Xia, *Data Analysis in Molecular Biology and Evolution*, Kluwer Academic Publishers, Boston, Mass, USA, 2000.

[105] X. Xia, Z. Xie, M. Salemi, L. Chen, and Y. Wang, "An index of substitution saturation and its application," *Molecular Phylogenetics and Evolution*, vol. 26, no. 1, pp. 1–7, 2003.

[106] P. Lemey, M. Salemi, and A.-M. Vandamme, Eds., *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, Cambridge University Press, Cambridge, UK, 2009.

[107] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specific tree topology," *Systematic Biology*, vol. 20, no. 4, pp. 406–416, 1971.

[108] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376, 1981.

[109] A. Rzhetsky and M. Nei, "A simple method for estimating and testing minimum-evolution trees," *Molecular Biology and Evolution*, vol. 9, no. 5, pp. 945–967, 1992.

[110] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback, "Bayesian inference of phylogeny and its impact on evolutionary biology," *Science*, vol. 294, no. 5550, pp. 2310–2314, 2001.

[111] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[112] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, vol. 22, no. 2, pp. 160–174, 1985.

[113] K. Katoh, K. Kuma, H. Toh, and T. Miyata, "MAFFT version 5: improvement in accuracy of multiple sequence alignment," *Nucleic Acids Research*, vol. 33, no. 2, pp. 511–518, 2005.

[114] A. Dereeper, S. Audic, J. Claverie, and G. Blanc, "BLAST-EXPLORER helps you building datasets for phylogenetic analysis," *BMC Evolutionary Biology*, vol. 10, no. 1, article 8, 2010.

[115] A. Dereeper, V. Guignon, G. Blanc et al., "Phylogeny.fr: robust phylogenetic analysis for the non-specialist," *Nucleic Acids Research*, vol. 36, pp. W465–W469, 2008.

[116] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 540–552, 2000.

[117] K. Tamura, "Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases," *Molecular Biology and Evolution*, vol. 9, no. 4, pp. 678–687, 1992.

[118] J. P. Huelsenbeck and F. Ronquist, "MRBAYES: bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.

[119] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[120] J. A. A. Nylander, *MrModeltest v2. Program Distributed by the Author*, Evolutionary Biology Centre, Uppsala University, 2004.

[121] P. Legendre and V. Makarenkov, "Reconstruction of biogeographic and evolutionary networks using reticulograms," *Systematic Biology*, vol. 51, no. 2, pp. 199–216, 2002.

[122] V. Makarenkov, "T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks," *Bioinformatics*, vol. 17, no. 7, pp. 664–668, 2001.

[123] V. Makarenkov, "An algorithm for the fitting of a tree metric according to a weighted least-squares criterion," *Journal of Classification*, vol. 16, no. 1, pp. 3–26, 1999.

[124] A. Rocha-Olivares, J. W. Fleeger, and D. W. Foltz, "Decoupling of molecular and morphological evolution in deep lineages of a meiobenthic harpacticoid copepod," *Molecular Biology and Evolution*, vol. 18, no. 6, pp. 1088–1102, 2001.

[125] B. W. Frost and A. Bucklin, "Morphological and molecular phylogenetic analysis of evolutionary lineages within *Clausocalanus* (Copepoda: Calanoida)," *Journal of Crustacean Biology*, vol. 29, no. 1, pp. 111–120, 2009.

[126] T. A. Heath, S. M. Hedtke, and D. M. Hillis, "Taxon sampling and the accuracy of phylogenetic analyses," *Journal of Systematics and Evolution*, vol. 46, no. 3, pp. 239–257, 2008.

[127] J. Bergsten, "A review of long-branch attraction," *Cladistics*, vol. 21, no. 2, pp. 163–193, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 720715, 8 pages
http://dx.doi.org/10.1155/2014/720715

Research Article

The Properties of Binding Sites of miR-619-5p, miR-5095, miR-5096, and miR-5585-3p in the mRNAs of Human Genes

Anatoly Ivashchenko, Olga Berillo, Anna Pyrkova, Raigul Niyazova, and Shara Atambayeva

National Nanotechnology Laboratory, Al-Farabi Kazakh National University, Almaty 050038, Kazakhstan

Correspondence should be addressed to Anatoly Ivashchenko; a_ivashchenko@mail.ru

Received 14 April 2014; Revised 3 July 2014; Accepted 3 July 2014; Published 4 August 2014

Academic Editor: Vassily Lyubetsky

The binding of 2,578 human miRNAs with the mRNAs of 12,175 human genes was studied. It was established that miR-619-5p, miR-5095, miR-5096, and miR-5585-3p bind with high affinity to the mRNAs of the 1215, 832, 725, and 655 genes, respectively. These unique miRNAs have binding sites in the coding sequences and untranslated regions of mRNAs. The mRNAs of many genes have multiple miR-619-5p, miR-5095, miR-5096, and miR-5585-3p binding sites. Groups of mRNAs in which the ordering of the miR-619-5p, miR-5095, miR-5096, and miR-5585-3p binding sites differ were established. The possible functional and evolutional properties of unique miRNAs are discussed.

1. Introduction

MicroRNAs (miRNAs) participate in the regulation of the expression of protein-coding genes at the posttranscriptional stage [1]. miRNAs, as a part of the RNA-induced silencing complex, bind to mRNAs and interfere with translation or promote mRNA destruction [2]. The study of the properties of miRNAs and their influences on the expression of the genes that participate in all key cellular processes of cells was established in the last 20 years. The actions of miRNAs on the cell cycle [3], apoptosis [4], differentiation [5], and growth and development in plants [6] and animals [7] have been shown. Connections between miRNA expression and the development of various diseases have been established. miRNA concentrations change in cancer [8] and cardiovascular diseases [9]. Metabolic disturbances necessarily change miRNA concentrations in cells [10]. It is possible to normalize some processes using miRNAs [11]. The aforementioned roles do not encompass the full list of the biological processes in which miRNAs participate, which proves the importance of their biological functions.

Despite the appreciable successes in the study of miRNA properties, there are obstacles to establishing the target genes of miRNAs. Normally, one miRNA interacts with the mRNA of one gene. However, there are miRNAs that bind to many mRNAs, and one mRNA can be the target of many miRNAs. These circumstances significantly complicate the study of the properties of miRNAs and their diagnostic and medical applications. There are more than 2,500 miRNAs in the human genome, and they are thought to act on 50% or more of genes. It will be difficult to draw unique conclusions about the participation of miRNAs in specific biological processes, and until those conclusions can be drawn, the connections between the majority of miRNAs and their target genes will remain unknown.

Recently, we found a set of unique miRNAs that have hundreds of target genes and bind to mRNAs with high affinity. The binding sites unique to miRNAs are located in the 5′-untranslated regions (5′UTRs), the coding domain sequences (CDSs) and the 3′-untranslated regions (3′UTRs) of mRNAs. In present work, we studied some unique miRNAs that bind to the mRNAs of several hundred human genes.

2. Materials and Methods

The human gene mRNAs were taken from GenBank (http://www.ncbi.nlm.nih.gov/) using Lextractor002 script (http://sites.google.com/site/malaheenee/software). The nucleotide sequences of human miR-619-5p, miR-5095, miR-5096,

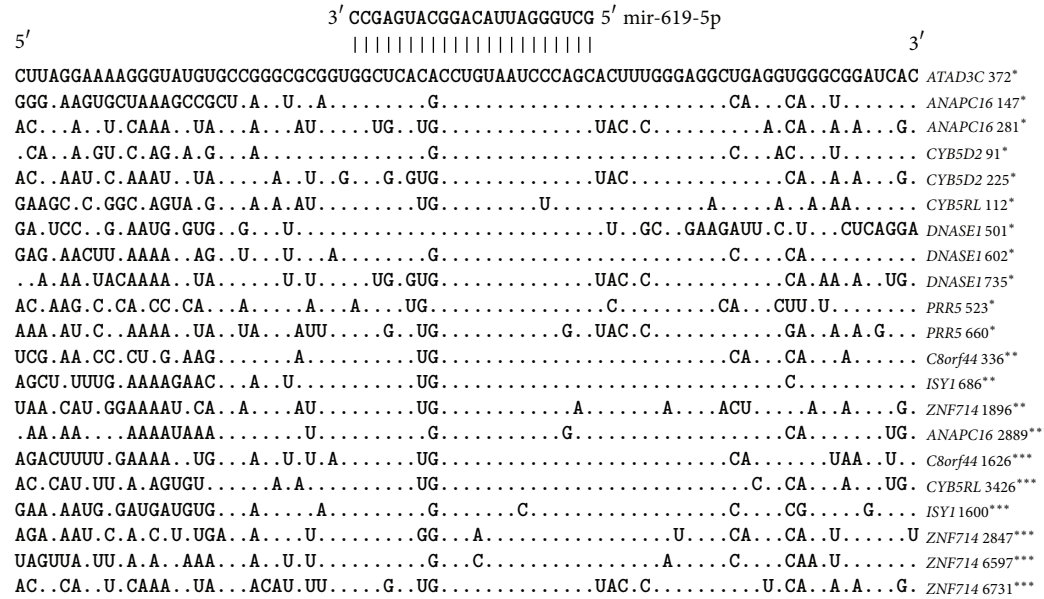


Figure 1: miR-619-5p binding sites in the 5′UTRs, CDSs, and 3′UTRs of human genes. Note: for Figures 1 and 2, the symbols $*$, $**$ and $***$ indicate the position of the origin of the miR-619-5p binding site from the first nucleotide of the 5′UTR in the 5′UTRs, CDSs, and 3′UTRs, respectively.

and miR-5585-3p were taken from the miRBase site (http://mirbase.org/).

The target genes for the tested miRNAs were revealed using the MirTarget program, which was developed in our laboratory. This program defines the following features of binding: (a) the origin of the initiation of miRNA binding to mRNAs; (b) the localization of miRNA binding sites in the 5′-untranslated regions (5′UTRs), the coding domain sequences (CDSs), and the 3′-untranslated regions (3′UTRs) of the mRNAs; (c) the free energy of hybridization (ΔG, kJ/mole); and (d) the schemes of nucleotide interactions between the miRNAs and the mRNAs. The ratio $\Delta G/\Delta G_m$ (%) was determined for each site (ΔG_m equals the free energy of an miRNA binding with its perfect complementary nucleotide sequence). The miRNA binding sites located on the mRNAs had $\Delta G/\Delta G_m$ ratios of 90% or more. We also noted the positions of the binding sites on the mRNA, beginning from the first nucleotide of the mRNA's 5′UTR. This program found hydrogen bonds between adenine (A) and uracil (U), guanine (G) and cytosine (C), G and U, and A and C. The distances between A and C were equal to those between G and C, A and U, and G and U. The numbers of hydrogen bonds in the G-C, A-U, G-U, and A-C interactions were found to be 3, 2, 1, and 1, respectively. The free binding energies of these nucleotide pairs were taken as the same values (i.e., 3, 2, 1, and 1, resp.) [12, 13].

3. Results and Discussion

3.1. Features of miR-619-5p, miR-5096, miR-5585-3p, and miR-5095. The binding powers between the 2,578 tested hsa-miRNAs and the mRNAs of 12,175 human genes were calculated. Some of these miRNAs have greater numbers of target genes than others. For example, miR-619-5p, miR-5095, miR-5096, and miR-5585-3p were found to be capable of binding more 600 genes each with value $\Delta G/\Delta G_m$ ratios of 90% or more. These miRNAs were termed unique miRNAs (umiRNAs). Additionally, the binding sites for these unique miRNAs were unusually located in the mRNAs. miR-619-5p, miR-5095, miR-5096, and miR-5585-3p have different miRNA binding site origins, lengths, quantities, and miRNA binding site properties, among other features. Some characteristics of these unique miRNAs are outlined below.

With a length of 22 nucleotides (nt), miR-619-5p is coded in an intron of the slingshot protein phosphatase 1 host gene (*SSH1*), which is located on chromosome 12. We found that miR-619-5p has 1811 binding sites on 1215 target mRNAs. Of those, 1772 miR-619-5p binding sites are located in 3′UTRs, 26 sites are located in 5′UTRs, and 13 sites are located in CDSs. The mRNAs of 197 genes have completely complementary binding sites for miR-619-5p. The mRNAs of 27 genes have four binding sites. Seven genes have five binding sites, and the mRNAs of the *CATAD1, ICA1L, GK5, POLH,* and *PRR11* genes have six miR-619-5p binding sites. The mRNAs of the *OPA3* and *CYP20A1* genes have eight and ten binding sites, respectively. All of these sites are located in 3′UTRs.

With a length of 21 nt, miR-5096 is coded in an intron of the BMP2 inducible kinase host gene (*BMP2K*), which is located on chromosome 4. We found that miR-5096 has 997 binding sites on 832 target mRNAs. Of these, 984 miR-5096 binding sites are located in 3′UTRs, nine sites are located in 5′UTRs, and four sites are located in CDSs. The mRNAs of 42 genes have completely complementary binding sites for miR-5096. The mRNAs of the *IP09* gene have four binding sites. The *PRR11* gene has five binding sites. The mRNAs of the *OPA3* and *CYP20A1* genes have six and eleven miR-5096

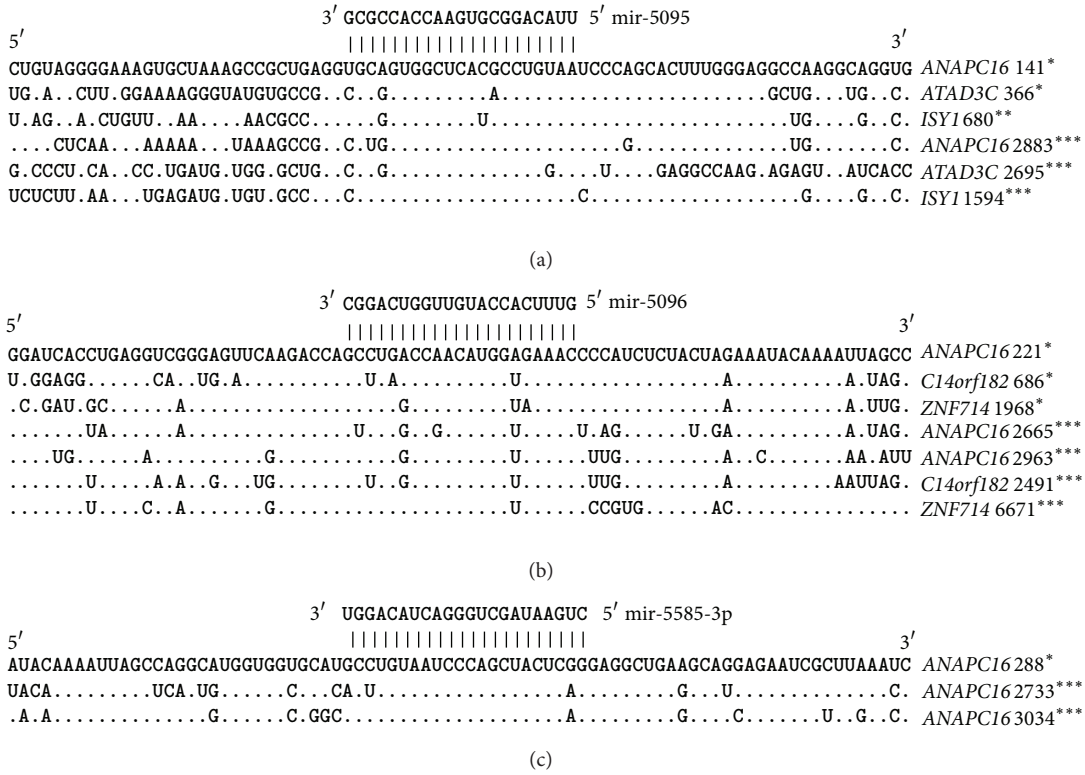


Figure 2: miR-5095, miR-5096, and miR-5585 binding sites in the 5′UTRs, CDSs, and 3′UTRs of human genes.

binding sites, respectively. All of these sites are located in 3′UTRs.

With a length of 22 nt, miR-5585-3p is coded in an intron of the transmembrane protein 39b host gene (*TMEM39B*), which is located on chromosome 4. We found that 725 target gene mRNAs have 844 binding sites for miR-5585-3p. Nine of these binding sites are located in 5′UTRs, two sites are located in CDSs, and 833 sites are located in 3′UTRs. The mRNAs of the *CYP20A1* and *GPR155* genes each have four binding sites.

With a length of 21 nt, miR-5095 is coded in an intron of the sterol carrier protein 2 host gene (*SCP2*), which is located on chromosome 1. We found that 655 target gene mRNAs have 734 binding sites. Fourteen of these binding sites are located in 5′UTRs, eight sites are located in CDSs, and 712 sites are located in 3′UTRs. The mRNAs of two genes have completely complementary binding sites for miR-5095. The mRNAs of the *OPA3* and *SPN* genes each have four binding sites.

3.2. miRNA Binding Sites in 5′UTRs, CDSs, and 3′UTRs. The miR-619-5p, miR-5095, miR-5096, and miR-5585-3p binding sites in the 5′UTRs, CDSs, and 3′UTRs of several genes were predicted using the MirTarget program. Multiple miRNA binding sites were revealed to be in the 5′UTRs of several genes. For example, miR-619-5p has two binding sites in each of the 5′UTRs of the *ANAPC16, CYB5D2*, and *PRR5* mRNAs and three binding sites in the *DNASE1* mRNA (Figure 1).

The mRNAs of some genes have binding sites for miR-619-5p, miR-5095, miR-5096, and miR-5585-3p within their 5′UTRs and 3′UTRs or CDSs and 3′UTRs. For example, the 5′UTRs and 3′UTRs of the *ATAD3C* and *CYB5RL* genes have miR-619-5p binding sites. The CDSs and 3′UTRs of the *C8orf44, ISY1*, and *ZNF714* genes have miR-619-5p binding sites.

The 5′UTR and 3′UTR of the *ANAPC16* gene have miR-5095, miR-5096, and miR-5585-3p binding sites (Figure 2). The 5′UTR and 3′UTR of the *ATAD3C* gene have miR-5095 and miR-619-5p binding sites. The 5′UTRs and 3′UTRs of the *C14orf182* and *CYB5RL* genes have miR-5096 and miR-619-5p binding sites, respectively.

miR-5095 and miR-619-5p binding sites were found in the CDS and 3′UTR of the *ISY1* gene. The CDS and 3′UTR of the *ZNF714* gene have binding sites for miR-5096 and miR-619-5p, and the *C8orf44* mRNA has only an miR-619-5p binding site.

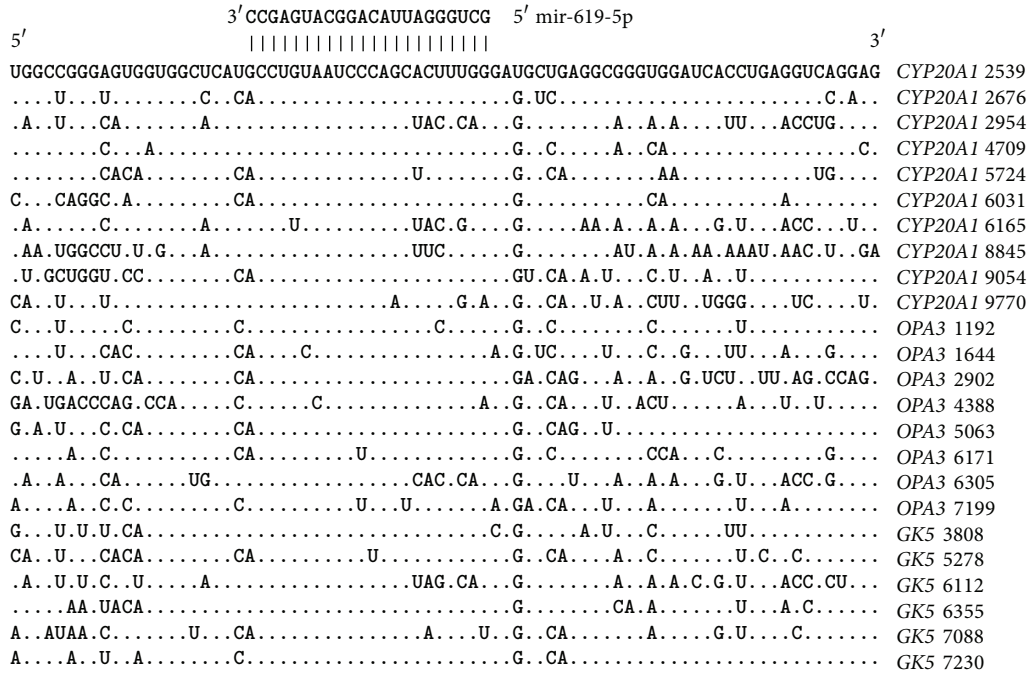
The nucleotide sequences of the miR-619-5p binding sites mRNAs of the OPA3 and SPN genes each have four binding locate in the CDSs of the *C8orf44, ISY1*, and *ZNF714* genes. The sites code for the following oligopeptides

ENHWKGRA**RWLMPVIPALWEA**K**AG**R**S** *C8orf44,*

LFEKERQV**RWLMPVIPALWEA**E**AG**G**S** *ISY1,*

KHRKIQQGMV**AHACNPN**TLRGLGEQI *ZNF714.*

The first two oligopeptides are coded in one open reading frame (ORF), and the amino acids the miR-619-5p binding site codes for are highly conserved (highlighted in bold). The homologous oligonucleotide of the miR-619-5p binding site in *ZNF714* mRNA codes for an oligopeptide in other ORF. The presence of miR-619-5p binding sites in the CDSs



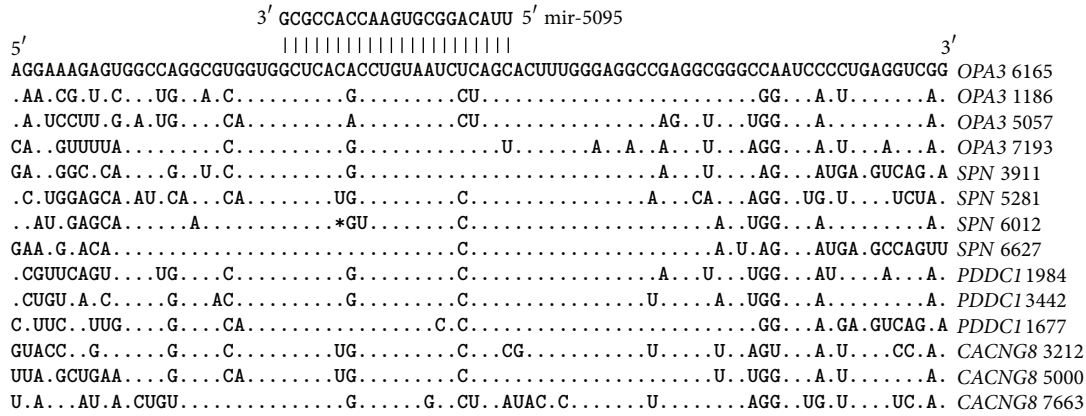
(a)

(b)

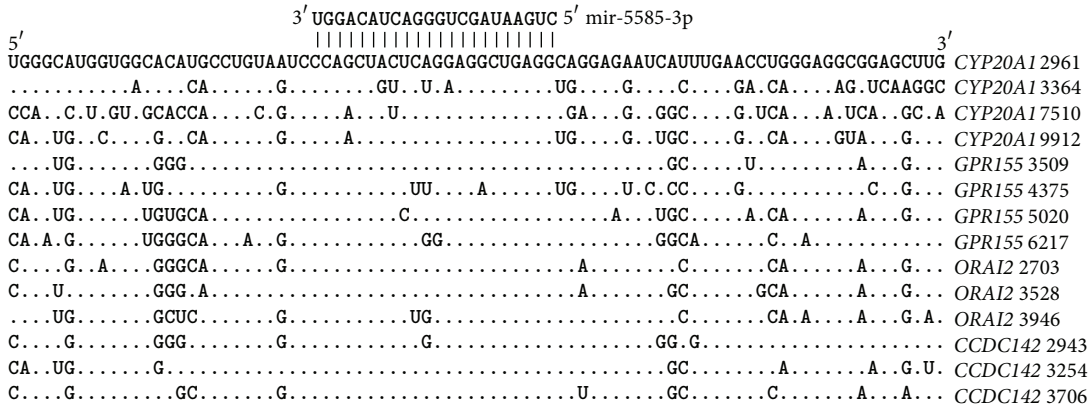
Figure 3: The nucleotide parts of 3′UTRs having multiple miR-5095 (a) and miR-5585-3p (b) binding sites. Note: ∗—one adenine was deleted before cytosine.

of three genes with different functions and the evolutionary conservation of these sites testify to the importance of the role of miRNA in the regulation of the expression of these genes. The nucleotide sequences of the parts of the *C8orf44, ISY1*, and *ZNF714* mRNAs that contain the miR-619-5p binding sites in the CDSs are homologous among themselves and homologous with the binding sites located in the 5′UTRs and 3′UTRs.

In mRNA oligonucleotides, the miR-619-5p, miR-5095, miR-5096, and miR-5585-3p binding sites and the nucleotide sequences immediately downstream of the binding sites are highly homologous. It is possible that these downstream sequences are binding sites for other miRNAs. The sequences of the 5′UTRs, CDSs, and 3′UTRs upstream of the miR-5096 and miR-5585-3p binding sites are also homologous, likely for the same reason. The sequences of the mRNAs upstream of



(a)



(b)

Figure 4: The nucleotide parts of 3′UTRs having multiple miR-619-5p (a) and miR-5096 (b) binding sites.

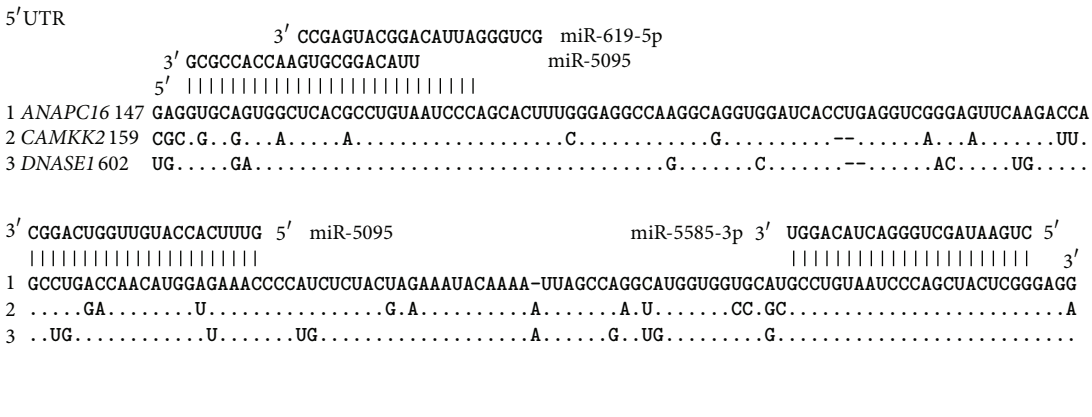


Figure 5: miR-619-5p, miR-5095, miR-5096, and miR-5585-3p binding sites located in 5′UTRs.

the miR-5095 binding sites are not homologous, and no other miRNA binding sites are identified in these sequences. The 10-nucleotide sequences upstream of the miR-619-5p binding sites contain homologous nucleotide sequences because they are the miR-5095 binding sites.

3.3. Multiple miRNA Binding Sites in the mRNAs of Target Genes. The mRNAs of some genes have multiple umiRNA binding sites. The nucleotide sequences with lengths of 95 nt that contain multiple miR-619-5p, miR-5096, miR-5095, and miR-5585-3p binding sites are given in Figures 3 and 4. These results testify to the high degree of homology between the umiRNA binding sites in the mRNAs of different genes. In addition to these binding sites, many other nucleotide sequences of the mRNAs are also homologous. It is possible that the nucleotide sequences adjacent to the binding sites are binding sites for other miRNAs.

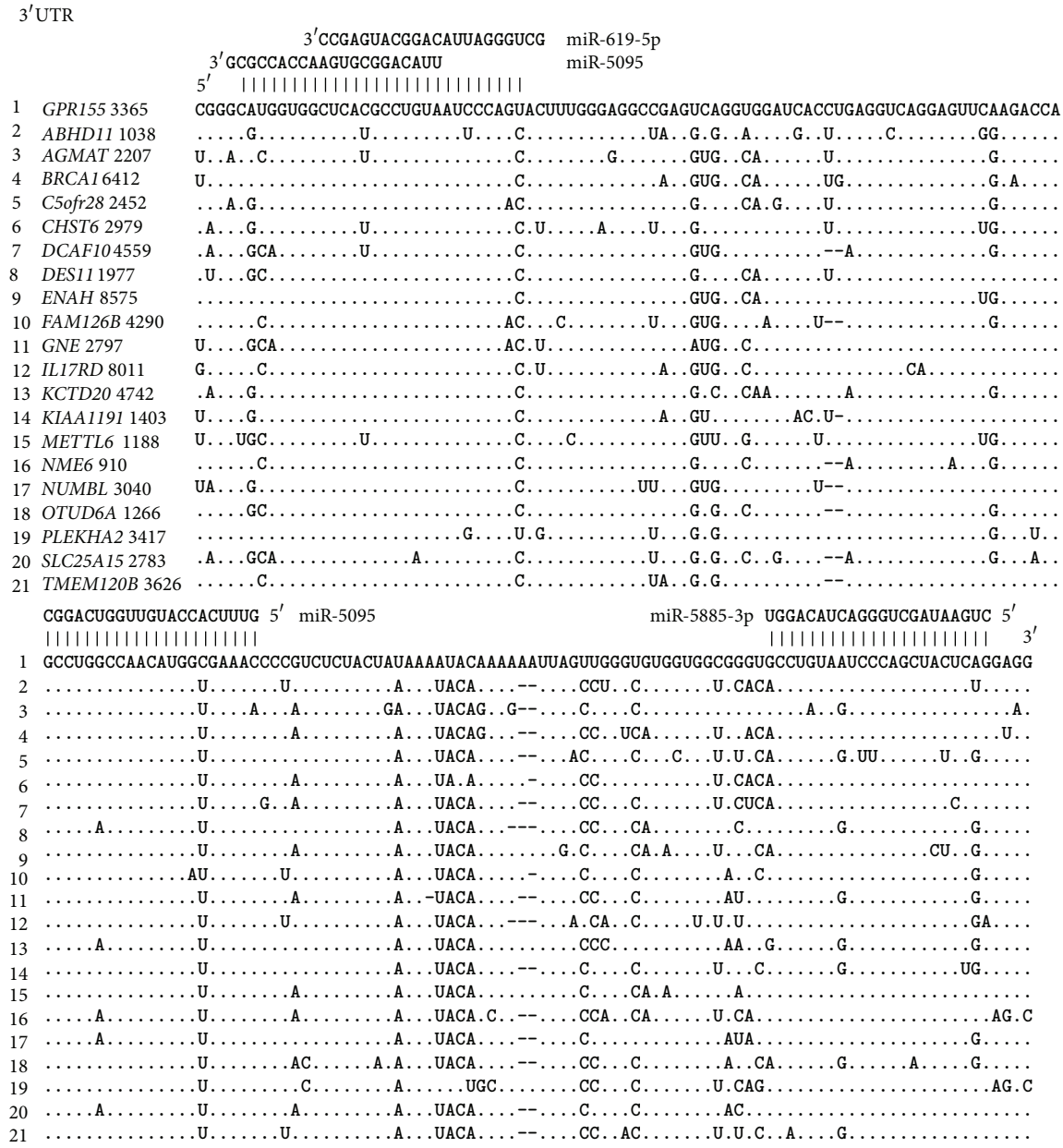


Figure 6: miR-619-5p, miR-5095, miR-5096, and miR-5585-3p binding sites located in the 3′UTRs.

3.4. The Arrangements of the Locations of umi-RNA Binding Sites. The mRNAs that were targeted by miR-619-5p, miR-5096, miR-5095, and miR-5585-3p were established. The 5′UTRs of three target genes contain these miRNA-binding sites (Figure 5). The degree of homology of the nucleotide sequences in these genes is high not only in the binding sites of the studied miRNAs but also across all mRNA 150 nt sequences. The distance between the miR-5095 and miR-5096 binding sites is 57–59 nt and that between the miR-5096 and miR-5585-3p binding sites is 46-47 nt. The miR-5095 and miR-619-5p binding sites partially overlap.

The greatest numbers of miR-619-5p, miR-5096, miR-5095, and miR-5585-3p binding sites are located in the 3′UTRs, and it is, therefore, possible that many target genes have umiRNAs binding sites. The data about the locations of the miR-619-5p, miR-5096, miR-5095, and miR-5585-3p binding sites and the degrees of homology of the corresponding nucleotide sequences in the mRNAs of 21 genes are presented in Figure 6. The distances between the miR-5095 and miR-5096 binding sites are all 57–60 nt. The distances between the miR-5095 and miR-5096 binding sites in the mRNAs of 78 genes averaged 58.6 ± 0.9 nt. Thus, the distances between miR-5095 and miR-5096 binding sites are highly conserve. The distances between the miR-5096 and miR-5585-3p binding sites are all 46–49 nt. The distances between the miR-5096 and miR-5585-3p binding sites in the mRNAs of 325 genes average 47.3 ± 1.1.

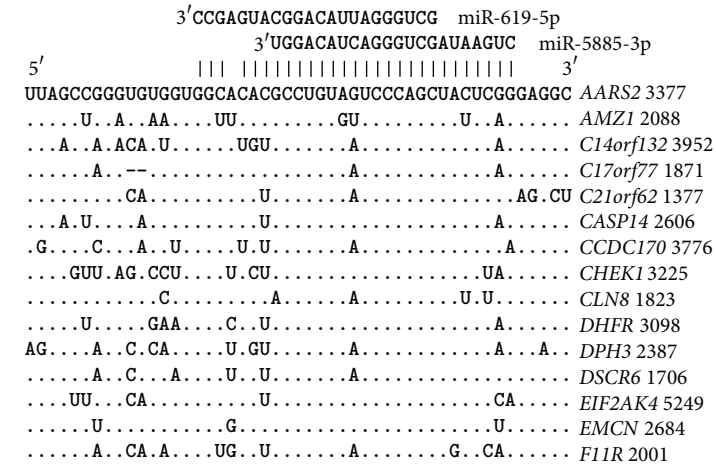


Figure 7: 3′UTRs with miR-619-5p binding sites upstream of the miR-5585-3p binding sites. Note: Symbol "≈" indicates nucleotide sequence equaled to 84 nt, which is not shown here.

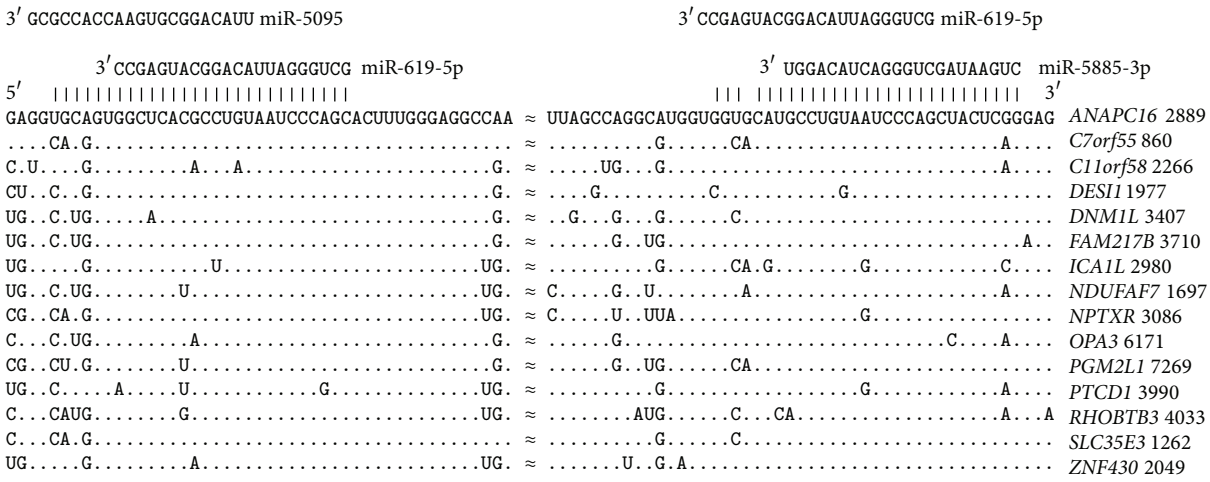


Figure 8: 3′UTRs in which the miR-619-5p binding site was downstream of the miR-5095 binding site and upstream of the miR-5585-3p binding site.

The degree of homology of the nucleotide sequences containing the miR-619-5p, miR-5096, miR-5095, and miR-5585-3p binding sites is high. These areas contain binding sites for miRNAs other than the studied umiRNAs. Other miRNA binding sites are not present in all genes, and these binding sites have lower affinities (data not shown). It is possible that there are conserved domains in the nucleotide sequences of mRNAs.

3.5. Variability in the Arrangement of umiRNA Binding Site Locations. The miR-619-5p binding site is located at a distance of six nucleotides downstream of the beginning of the miR-5095 site in the majority of genes containing arranged umiRNA sites. However, in another group of mRNAs, the beginnings of the miR-619-5p binding sites are located at distances of seven nucleotides upstream of the miR-5585-3p binding sites (Figure 7).

There is another group of genes in which the miR-619-5p binding sites are downstream of the miR-5095 sites and upstream of the miR-5585-3p sites (Figure 8). The distances between the positions of the two miR-619-5p binding sites in the mRNAs of these genes are constant at 112 nt. The nucleotide sequences of the mRNAs with miR-619-5p, miR-5095, and miR-5585-3p binding sites are highly homologous, which testifies to the strength of the selection pressure on these nucleotide sequences.

4. Conclusion

The detection of a large number of binding sites of miR-619-5p, miR-5095, miR-5096, and miR-5585 in the mRNAs of the genes studied here presumably indicates new functional opportunities. It is possible that these umiRNAs are coordinators of gene expression that participate in many major biological processes. The influences of miRNAs on

the expression of genes that code for transcription factors [14, 15] and proteins that participate in the cellular cycle [3], apoptosis [4], stress responses [16], and so forth, have previously been shown.

This study established that the binding sites of umiRNAs and miRNAs are in 3′UTRs, CDSs, and 5′UTRs. Highly conserved miR-619-5p, miR-5095, miR-5096, and miR-5585 binding sites in a large number of genes indicate the emergence of these sites in the early stages of human evolution. We have shown previously that miRNA binding sites located in CDS are conserved in target orthologous genes of organisms that diverged some hundred million years ago [17, 18]. *SSH1, BMP2K, TMEM39B*, and *SCP2* host genes and target genes of miR-619-5p, miR-5095, miR-5096, and miR-5585 have independent evolution origin. Revealed miRNA binding sites have arranged localization in mRNA of multiple target genes participating in different metabolic processes. Host genes often coexpress with their intragenic miRNAs; therefore, an expression of *SSH1, BMP2K, TMEM39B*, and *SCP2* genes can be connected with target genes via their miRNAs. Thus, it predicted that host genes and target genes are interconnected between themselves. Arranged localization of binding sites suggests an interconnected evolution of miRNAs and their target genes. Conservatism of the arranged localization of miRNA binding sites in 3′UTRs, 5′UTRs, and CDSs also demonstrates the common origin of the binding sites and the evolution of relationship of miRNAs with their target genes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank A. Moldagaliyeva and S. Sagaydak for their help in preparing the materials for analysis. The authors would also like to thank Dr. V. Khaylenko for writing the Lextractor002 script. This study was supported by a grant from the Ministry of Education and Science, Kazakhstan Republic.

References

[1] E. Doxakis, "Principles of miRNA—target regulation in metazoan models," *International Journal of Molecular Sciences*, vol. 14, no. 8, pp. 16280–16302, 2013.

[2] G. Tang, "siRNA and miRNA: an insight into RISCs," *Trends in Biochemical Sciences*, vol. 30, no. 2, pp. 106–114, 2005.

[3] Q. Luo, X. Li, J. Li et al., "MiR-15a is underexpressed and inhibits the cell cycle by targeting *CCNE*1 in breast cancer," *International Journal of Oncology*, vol. 43, no. 4, pp. 1212–1218, 2013.

[4] X. Li, Y.-T. Chen, S. Josson et al., "MicroRNA-185 and 342 inhibit tumorigenicity and induce apoptosis through blockade of the SREBP metabolic pathway in prostate cancer cells," *PLoS ONE*, vol. 8, no. 8, Article ID e70987, 2013.

[5] N. Qian, W. Liu, W. Lv et al., "Upregulated microRNA-92b regulates the differentiation and proliferation of EpCAM-positive fetal liver cells by targeting C/EBPβ," *PLoS ONE*, vol. 8, no. 8, Article ID e68004, 2013.

[6] K. Rogers and X. Chen, "Biogenesis, turnover, and mode of action of plant microRNAs," *Plant Cell*, vol. 25, no. 7, pp. 2383–2399, 2013.

[7] Y. H. Ling, J. P. Ding, X. D. Zhang et al., "Characterization of microRNAs from goat (*Capra hircus*) by Solexa deep-sequencing technology," *Genetics and Molecular Research*, vol. 12, no. 2, pp. 1951–1961, 2013.

[8] Q. J. Guo, J. N. Mills, S. G. Bandurraga et al., "MicroRNA-510 promotes cell and tumor growth by targeting peroxiredoxin1 in breast cancer," *Breast Cancer Research*, vol. 15, no. 4, article R70, 2013.

[9] A. Magenta, S. Greco, C. Gaetano, and F. Martelli, "Oxidative stress and microRNAs in vascular diseases," *International Journal of Molecular Sciences*, vol. 14, no. 9, pp. 17319–17346, 2013.

[10] S. Swaminathan, K. Suzuki, N. Seddiki et al., "Differential regulation of the Let-7 family of microRNAs in $CD4^+$ T cells alters IL-10 expression," *Journal of Immunology*, vol. 188, no. 12, pp. 6238–6246, 2012.

[11] D. V. Glazkova, A. S. Vetchinova, E. V. Bogoslovskaia, I. A. Zhogina, M. L. Markelov, and G. A. Shipulin, "Downregulation of human CCR5 receptor gene expression using artificial microRNAs," *Molecular Biology*, vol. 47, no. 3, pp. 475–485, 2013.

[12] E. T. Kool, "Hydrogen bonding, base stacking, and steric effects in DNA replication," *Annual Review of Biophysics and Biomolecular Structure*, vol. 30, pp. 1–22, 2001.

[13] N. B. Leontis, J. Stombaugh, and E. Westhof, "The non-Watson-Crick base pairs and their associated isostericity matrices," *Nucleic Acids Research*, vol. 30, no. 16, pp. 3497–3531, 2002.

[14] Q. Cui, Z. Yu, Y. Pan, E. O. Purisima, and E. Wang, "MicroRNAs preferentially target the genes with high transcriptional regulation complexity," *Biochemical and Biophysical Research Communications*, vol. 352, no. 3, pp. 733–738, 2007.

[15] L. Yan, M. Kang, Z. Qin, W. Zhang, Y. Li, and H. Ou, "An intronic miRNA regulates expression of the human endothelial nitric oxide synthase gene and proliferation of endothelial cells by a mechanism related to the transcription factor SP-1," *PLoS ONE*, vol. 8, no. 8, Article ID e70658, 2013.

[16] K. Cawley, S. E. Logue, A. M. Gorman et al., "Disruption of microRNA biogenesis confers resistance to ER stress-induced cell death upstream of the mitochondrion," *PLoS ONE*, vol. 8, no. 8, Article ID e73870, 2013.

[17] A. Bari, S. Orazova, and A. Ivashchenko, "Mir156- and mir171-binding sites in the protein-coding sequences of several plant genes," *BioMed Research International*, vol. 2013, Article ID 307145, 7 pages, 2013.

[18] A. T. Ivashchenko, A. S. Issabekova, and O. A. Berillo, "MiR-1279, miR-548j, miR-548m, and miR-548d-5p binding sites in CDSs of paralogous and orthologous PTPN12, MSH6, and ZEB1 genes," *BioMed Research International*, vol. 2013, Article ID 902467, 10 pages, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 147145, 8 pages
http://dx.doi.org/10.1155/2014/147145

Research Article

Extended Genetic Diversity of Bovine Viral Diarrhea Virus and Frequency of Genotypes and Subtypes in Cattle in Italy between 1995 and 2013

Camilla Luzzago,[1] Stefania Lauzi,[1] Erika Ebranati,[2] Monica Giammarioli,[3] Ana Moreno,[4] Vincenza Cannella,[5] Loretta Masoero,[6] Elena Canelli,[4] Annalisa Guercio,[5] Claudio Caruso,[6] Massimo Ciccozzi,[7,8] Gian Mario De Mia,[3] Pier Luigi Acutis,[6] Gianguglielmo Zehender,[2] and Simone Peletto[6]

[1] *Department of Veterinary Science and Public Health, University of Milan, Via Celoria 10, 20133 Milan, Italy*
[2] *Department of Clinical Sciences "Luigi Sacco", Section of Infectious Diseases, University of Milan, Via G.B. Grassi 74, 20157 Milan, Italy*
[3] *Istituto Zooprofilattico Sperimentale dell'Umbria e delle Marche, Via G. Salvemini 1, 06126 Perugia, Italy*
[4] *Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, Via Bianchi 9, 25124 Brescia, Italy*
[5] *Istituto Zooprofilattico Sperimentale della Sicilia "A. Mirri", Via G.Marinuzzi 3, 90129 Palermo, Italy*
[6] *Istituto Zooprofilattico Sperimentale Piemonte, Liguria e Valle D'Aosta, Via Bologna 148, 10154 Torino, Italy*
[7] *Department of Infectious Parasitic and Immunomediated Diseases, National Institute of Health, Viale Regina Elena 299, 00161 Rome, Italy*
[8] *University of Biomedical Campus, Via Álvaro del Portillo, 200, 00128 Rome, Italy*

Correspondence should be addressed to Camilla Luzzago; camilla.luzzago@unimi.it

Received 21 March 2014; Accepted 29 May 2014; Published 22 June 2014

Academic Editor: William H. Piel

Genetic typing of bovine viral diarrhea virus (BVDV) has distinguished BVDV-1 and BVDV-2 species and an emerging putative third species (HoBi-like virus), recently detected in southern Italy, signaling the occurrence of natural infection in Europe. Recognizing the need to update the data on BVDV genetic variability in Italy for mounting local and European alerts, a wide collection of 5′ UTR sequences ($n = 371$) was selected to identify the frequency of genotypes and subtypes at the herd level. BVDV-1 had the highest frequency, followed by sporadic BVDV-2. No novel HoBi-like viruses were identified. Four distribution patterns of BVDV-1 subtypes were observed: highly prevalent subtypes with a wide temporal-spatial distribution (1b and 1e), low prevalent subtypes with a widespread geographic distribution (1a, 1d, 1g, 1h, and 1k) or a restricted geographic distribution (1f), and sporadic subtypes detected only in single herds (1c, 1j, and 1l). BVDV-1c, k, and l are reported for the first time in Italy. A unique genetic variant was detected in the majority of herds, but cocirculation of genetic variants was also observed. Northern Italy ranked first for BVDV introduction, prevalence, and dispersion. Nevertheless, the presence of sporadic variants in other restricted areas suggests the risk of different routes of BVDV introduction.

1. Introduction

Bovine viral diarrhea virus (BVDV), a widespread pathogen of cattle, was first described in North America in 1946 [1, 2]. Seroprevalence rates between 36% and 88% have subsequently been reported since the 1960s in North America, Europe, Australia, and East Africa. The endemic diffusion of BVDV persists in geographic areas where no systematic control measures have been implemented and it reflects the pathogenic mechanisms of BVDV through which the virus can establish both transient and persistent infections. Persistent infected (PI) animals, originating from a transient infection of pregnant cows or born from PI cows, shed large amounts of virus throughout their lives and ensure

viral persistence in the host population. Besides the previous maintenance strategy, a peculiar biological characteristic of BVDV contributes to iatrogenic diffusion of infection. In fact, two different biotypes coexist and are differentiated by their effect on cell culture in cytopathic (cp) and noncytopathic biotypes (ncp). Contamination of fetal bovine serum (FBS) by the ncp biotype has long been known and remains a recognized risk factor for the worldwide distribution of BVDV [3, 4]. Because FBS is used in the production of vaccines and other biological products, the global trade of infected FBS products is a potential source of transboundary spread of BVDV.

BVDV belongs to the *Pestivirus* genus of the Flaviviridae family. Genetic typing of BVDV isolates distinguishes two recognized species: BVDV-1 and BVDV-2. To date, 17 potential BVDV-1 subtypes have been identified [5–10] and BVDV-2 strains can be clustered into at least three subtypes [5, 11, 12]: BVDV-1a to BVDV-1q and BVDV-2a to BVDV-2c, respectively. A putative third bovine species, referred to as HoBi-like virus or BVDV-3 [13], comprises viral strains recently detected in FBS batches originating mainly from South America, as well as from Australia, Canada, Mexico, and the United States [14]. Unlike the widespread diffusion of HoBi-like viruses via FBS, natural infection in cattle has been reported in Southeast Asia [15], in South America [16, 17], and in two dairy herds in southern Italy [18, 19]. The Italian cases represent the first identification of HoBi-like virus in cattle in Europe.

In addition to the risk of transboundary spread of BVDV through potentially infected FBS, the genetic diversity of the virus in a given geographic area has been largely influenced by animal movement within countries and/or introduction from other countries, as recently observed in Switzerland [27], Italy [21], and England and Wales [28], showing that the introduction and spatial distribution of BVDV can also be influenced by livestock management practices.

Nucleotide sequencing is a rapid and inexpensive diagnostic tool for unambiguous typing of all bovine pestiviruses. Broad systematic surveillance of BVDV genetic variability is advisable for updating data on the distribution and frequency of emerging genotypes and subtypes in BVDV endemic countries. The aim of this study was therefore to analyze a representative and epidemiologically well-characterized collection of BVDV sequences from Italian cattle. The previous genotyping studies were carried out on a limited number of isolates [20, 22–24, 29, 30]. The present study represents a comprehensive collection of Italian isolates obtained from all cattle breeding areas over a time period spanning nearly two decades (1995–2013). Genetic variability was determined to identify the frequency of genotypes and subtypes with resolution at the herd level.

2. Materials and Methods

2.1. Samples and Data Set. The material comprised samples sent to laboratories between 1995 and 2013 for routine testing because of suspected BVDV infection and for screening in voluntary herd control programs. The BVDV positive samples and the derived sequences were selected on the basis of the following criteria: (1) sequences obtained from different animals, (2) known herd and locality where the strain was isolated and sampling dates, and (3) sequences representative of all Italian BVDV subtypes and genotypes. A total of 371 BVDV 5′UTR sequences were included, 272 of which were novel sequences and 99 were BVDV Italian sequences retrieved from published peer-reviewed journals. Samples were obtained from 357 cattle and 14 bulk milk specimens collected from 259 Italian cattle herds from around the country. Samples were obtained from 164 dairy herds, 40 beef herds, and 11 mixed production systems, and 44 were undetermined. Only one sequence from each herd was available for the majority of the herds ($n = 210$); 2 to 20 sequences were included for the 49 remaining herds. The localities of origin of the strains were grouped into four macroareas: the North, the Center, the South, and the Islands (Sicily and Sardinia). In detail, the North macroarea comprised the regions of Emilia Romagna ($n = 41$), Lombardy ($n = 108$), Piedmont ($n = 130$), Valle d'Aosta ($n = 1$), and Veneto ($n = 18$); the Center, Latium ($n = 6$), Marches ($n = 4$), Tuscany ($n = 2$), and Umbria ($n = 15$); the South, Basilicata ($n = 4$), Calabria ($n = 5$), Campania ($n = 4$), and Puglia ($n = 2$); the Islands, Sicily ($n = 29$) and Sardinia ($n = 2$).

2.2. RT-PCR and Sequencing. Viral RNA of the 272 novel sequences was extracted from original biological samples ($n = 261$) and growth medium after passage in cell culture ($n = 11$). Reverse transcription and PCR assays targeting a 288 bp region of 5′UTR of BVDV were performed using previously described primers by Letellier et al. [31], with the exception of strains collected in Sicily, which were tested using primers by Vilcek et al. [32]. The samples collected in the Center and the South macroareas were also tested by primers for atypical *Pestivirus* [14]. For the BVDV-1 subtypes identified for the first time in Italy, a 428 bp region encoding autoprotease N^{pro} was amplified using nested PCR, as previously described [20, 29].

For each sample, the amplicons of the expected sizes were purified and sequenced using forward and reverse primers by cycle sequencing using a Big Dye Terminator version 1.1 Cycle Sequencing kit (Applied Biosystems, CA, USA) and an ABI PRISM 3130 sequencing device or sent for outsource sequencing (Primm).

2.3. Phylogenetic Analysis. All the sequences were aligned with BVDV reference strains retrieved from GenBank representative of BVDV-1, BVDV-2, and HoBi-like virus using Clustal X; manual editing was performed with Bioedit software version 7.0 (freely available at http://www.mbio.ncsu.edu/bioedit/bioedit.html). Phylogeny was estimated by the neighbor-joining algorithm (NJ) and the maximum likelihood (ML) method.

The evolutionary model that best fitted the data (GTR + I + G) was selected using an information criterion implemented in JmodelTest [33] (freely available at http://darwin.uvigo.es/software/jmodeltest.html).

Table 1: Frequency of BVDV genotypes and subtypes in Italy.

Genotype/subtype	Total sequences number (%)	Herd* number (%)	Years	Geographic origin**	Published/ total sequences (*n*)	References
BVDV-1a	9 (2.4)	9 (3.4)	2000–2009	NCI	3/9	[20]
BVDV-1b	149 (40.2)	115 (43.9)	1995–2013	NCSI	52/149	[20–24]
BVDV-1c	1 (0.3)	1 (0.4)	2010	S	0/1	Not published
BVDV-1d	20 (5.4)	16 (6.1)	1995–2010	NCS	5/20	[22, 23]
BVDV-1e	126 (34)	72 (27.5)	1996–2013	NCSI	22/126	[20–24]
BVDV-1f	28 (7.5)	17 (6.5)	1999–2012	N	1/28	[23]
BVDV-1g	4 (1.1)	4 (1.5)	2002–2010	NS	1/4	[20]
BVDV-1h	15 (4.0)	14 (5.3)	1996–2011	NCI	3/15	[21–23]
BVDV-1j	1 (0.3)	1 (0.4)	1995	N	1/1	[22]
BVDV-1k	3 (0.8)	3 (1.1)	2001–2011	NS	0/3	Not published
BVDV-1l	1 (0.3)	1 (0.4)	2007	C	0/1	Not published
BVDV-2	10 (2.7)	7 (2.7)	1995–2004	NS	7/10	[20, 22]
HoBi-like	4 (1.1)	2 (0.8)	2007–2011	S	4/4	[18, 19, 25, 26]

*Herds with different BVDV genotypes or subtypes are counted for each virus type.
**N: northern Italy, C: central Italy, S: southern Italy, and I: Islands.

NJ analysis of the 5′UTR was performed using molecular evolutionary genetics analysis (MEGA version 5) [34], with 1000 bootstrap replicates; ML tree was reconstructed with Phyml version 3.0 (http://www.atgc-montpellier.fr/phyml) with 1000 bootstrap replicates.

The sequences within the same genotypes and subtypes from the same herd and from different herds were compared, and the percentage of nucleotide similarity of pairwise evolutionary distances was calculated using MEGA version 5.

3. Results

A total of 269 out of 272 sequences obtained in the present study were typed as BVDV-1 upon analysis of 5′UTR to reference strains and three were classified as BVDV-2, using both NJ and ML methods. No HoBi-like viruses were identified in the Center and South macroareas using primers described by Xia et al. [14] and Letellier et al. [31], in northern Italy and Sardinia using primers by Letellier et al. [31], and in Sicily using primers described by Vilcek et al. [32]. A selection of Italian BVDV sequences representative of all BVDV genotypes and subtypes detected were reported in the NJ phylogenetic tree of the 5′UTR region, together with reference strains (Figure 1). A phylogenetic tree of all the sequences was also represented (see Supplementary Figure 1 available online at http://dx.doi.org/10.1155/2014/147145).

With regard to both the novel and the published sequences, 357 (96.2%) sequences belonged to BVDV-1, ten (2.7%) to BVDV-2, and four (1.1%) to HoBi-like virus. The BVDV-1 sequences belonged to 11 different subtypes (a, b, c, d, e, f, g, h, j, k, and l). The frequency of genotypes and subtypes is summarized in Table 1.

Among the subtypes identified, BVDV-1c, k, and l are reported here for the first time in Italy; typing of these subtypes was confirmed by N^{pro} region analysis (data not shown). Phylogenetic analysis showed that the BVDV-2 strains clustered with the subtypes a and c according to [12].

BVDV-1b and BVDV-1e showed the highest frequency at both the animal and the herd levels, being detected in 43.9% and 27.5% of herds, respectively. The frequency of BVDV-1a, d, f, g, h, and k was less than 10%; BVDV-1c, j, and l were sporadically obtained from single herds. The frequency of BVDV-2 was low at both the animal and the herd levels (2.7% at the herd level, Table 1).

The North macroarea, where cattle population density is highest, accounted for 298 (80.3%) BVDV sequences, the Center for 27 (7.3%), the South for 15 (4%), and the Islands (Sicily and Sardinia) for 31 (8.4%). When differentiated by geographic distribution, BVDV-1 and BVDV-2 were present in the North, BVDV-1 in the Center and the Islands, and BVDV-1, BVDV-2, and HoBi-like virus in the South. Nine different BVDV-1 subtypes were detected in the North, six in the Center and the South, and four in the Islands. The more frequent subtypes (BVDV-1b and 1e) were distributed across the entire country, whereas the less frequent subtypes (BVDV-1a, d, g, h, and k) were present in two or three macroareas, except for BVDV-1f, which was limited to the North (Table 1 and Figure 2). The most frequent subtypes were also distributed across all the years, whereas the lower prevalence subtypes and genotypes were detected sporadically (Supplementary Table 1).

The percentage of sequence similarity of pairwise evolutionary distances within BVDV genotypes and subtypes ranged from 83.1% to 100%. Forty of the 49 herds with more than one sequence had the same genotype or subtype (37 BVDV-1, two BVDV-2, and one HoBi-like). In addition, 100% sequence identity of BVDV-1 within herds was observed in 19/37 herds within a range of sampling between 1 day and 14 months. In 18/37 herds, the nucleotide percentage similarity was ≤ 99.4% within a range of sampling between 2 months

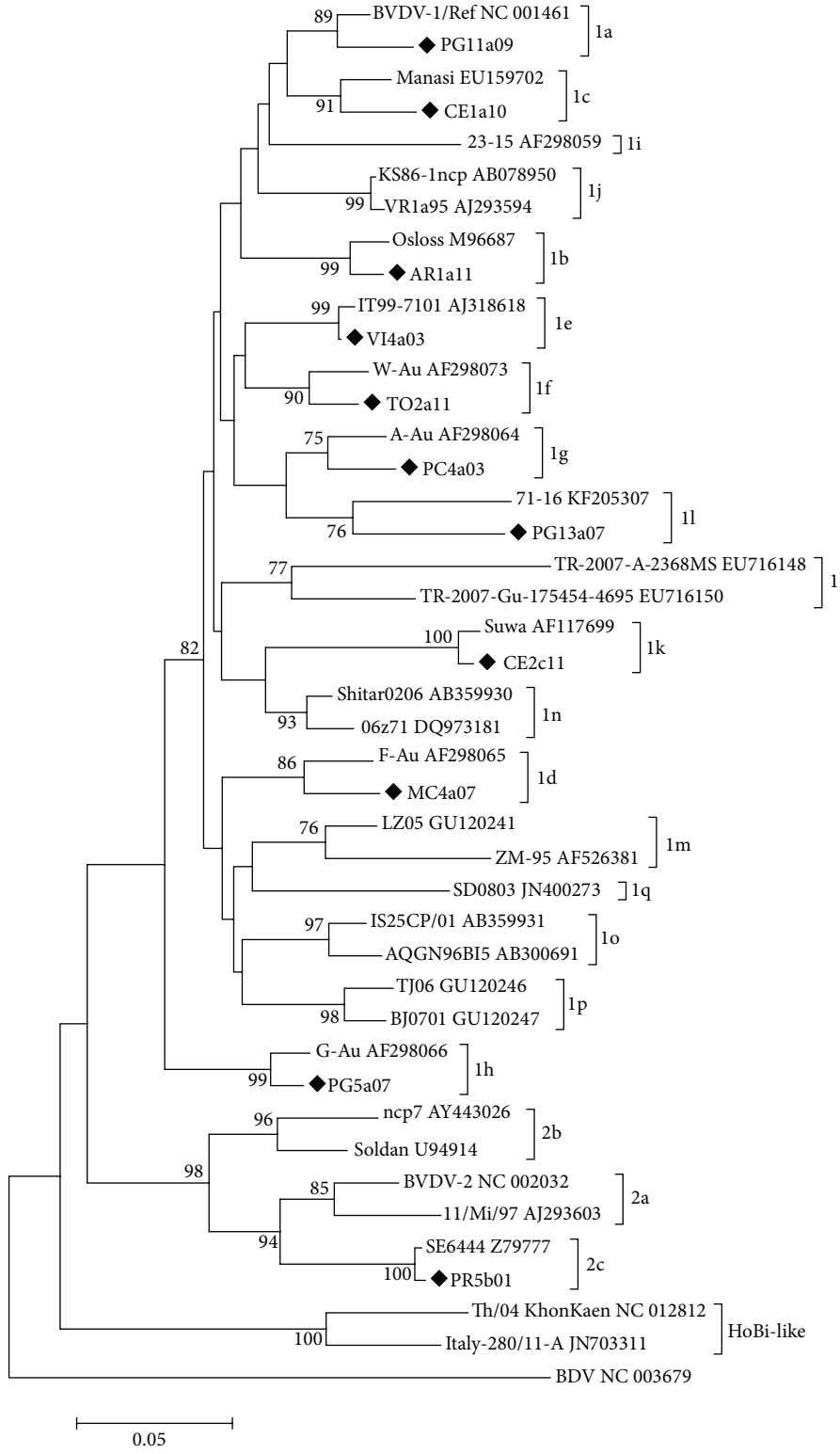


Figure 1: Phylogenetic tree based on the 5′-UTR of selected Italian sequences representative of all BVDV genotypes and subtypes detected between 1995 and 2013 and reference BVDV-1, BVDV-2, and HoBi-like strains. Molecular evolutionary genetics analyses were performed with MEGA5 using the NJ method. Distances were computed using the Kimura 2-parameter model. Bootstrap values > 70% are shown. Published sequences and references are identified by GenBank accession number (available at http://www.ncbi.nlm.nih.gov/pubmed/). The symbol "◆" indicates selected novel nucleotide sequences of BVDV Italian strains.

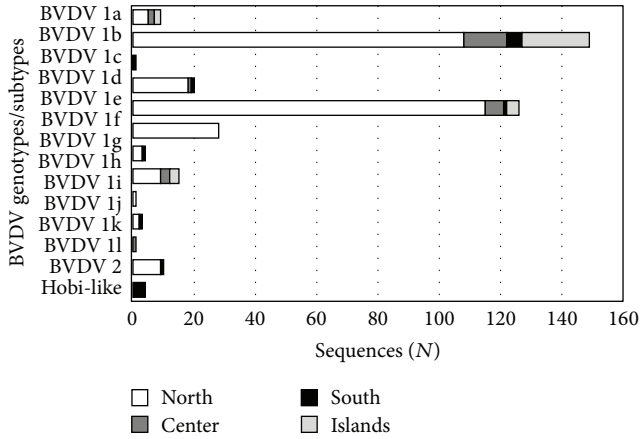


Figure 2: Frequency of BVDV genotypes and subtypes in the four Italian macroareas.

and 12 years (Table 2). The remaining nine herds had different genotypes or subtypes, without showing any relationship between the number of sequences analyzed and the time of sampling (Table 3).

4. Discussion

Molecular typing studies have demonstrated a wide genetic diversity of BVDV in Italy [20, 22–24, 29, 30], where its geographic distribution is influenced by animal movement within the country and/or importation from other countries [21]. Besides the two known BVDV species, a third species, referred to as HoBi-like or BVDV-3, has recently been detected in two herds in Italy [18, 19]. Surveillance of Italian BVDV isolates is therefore advisable to update data on the national genetic variability for mounting local and European-wide alerts.

A comprehensive collection of new BVDV sequences from all cattle breeding areas in Italy was investigated by phylogenetic analysis and compared against a selection of reference sequences of known genotypes and subtypes retrieved from public genetic databases. A total of 371 sequences have been included over a time span of 18 years (1995–2013) and were analyzed using the neighbor-joining and the maximum likelihood methods. At the genotype level, BVDV-1 had the highest frequency, followed by sporadic BVDV-2 in a limited area of the North (Lombardy and Emilia Romagna) and only once in the South. In the present study, the most recent detection of BVDV-2 dates back to 2004, encompassing nearly a decade during which the potential risk of transmission by BVDV-2 contaminated biological products had been reported [35, 36]. No novel HoBi-like virus was identified using panpestivirus primers described by Letellier et al. [31] that are able to detect HoBi-like virus [37, 38]; moreover, in central and southern Italy, a protocol for atypical *Pestivirus* detection [14] has been also applied. Regarding Sicily, we cannot exclude failure of HoBi-like viruses detection especially in samples with a low viral load, as previously reported for the primer pair used [4, 37]. The sporadic frequency of the HoBi-like virus in Italy was also confirmed by the National Reference Center for *Pestivirus* that did not detect any other HoBi-like strains on testing an additional collection of 450 BVDV field samples (Giammarioli personal communication, 2014).

The genetic diversity of BVDV-1 in Italy is increasing, as compared to previously published findings [20, 30]. Here we report the circulation of three additional subtypes, namely, BVDV-1c, 1k, and 1l, the last classified as the French 1l subtype described by [6], accounting for 1l out of the 17 BVDV-1 subtypes recognized worldwide [9, 10]. The increased phylogenetic diversity of BVDV-1 and the presence of new viral subtypes during the years were also reported in other European countries, especially due to the analysis of broader collection of BVDV isolates and/or introduction and movement of cattle from Europe [6, 28]. Different BVDV-1 subtypes are predominant in European countries; concerning the Italian neighboring countries, the most prevalent subtypes are BVDV-1e in France, BVDV-1e and 1h in Switzerland, BVDV-1h in Austria, and BVDV-1d and 1f in Slovenia [6, 27, 39, 40]. Interestingly, BVDV-1e is predominant in Piedmont, the Italian region close to the French border that is also characterized by the major commercial introduction of cattle from France.

Four frequency and distribution patterns of BVDV-1 subtypes were identified in Italy: high prevalent subtypes with a wide temporal-spatial distribution (BVDV-1b and 1e), low prevalent subtypes with a widespread geographic distribution (BVDV-1a, 1d, 1g, 1h, and 1k), low prevalent subtypes in restricted geographic areas (BVDV-1f in the North), and sporadic subtypes detected only in single herds (BVDV-1c, 1j, and 1l). The North macroarea showed the highest genetic variability, with nine out of 11 BVDV-1 subtypes and the cocirculation of BVDV-2, confirming the predominant role of this area in BVDV introduction into Italy from other European countries [21]. Nevertheless, the identification of some sporadic genetic variants restricted to the Center (BVDV-1l) or the South (BVDV-1c and HoBi-like) and the presence of eight BVDV genotypes and subtypes in the South, despite the low frequency of total sequences, suggest that BVDV has been likely introduced in Italy also through different commercial livestock flows or the use of contaminated biological products.

The genetic variability among BVDV isolates of the same subtype in the same herd was investigated in all the herds where more than one sequence was available. Two major genetic patterns were observed: the presence of herd-specific strains, also for prolonged periods (up to 14 months), and the presence of genetic variability of the same BVDV subtype within single herds, particularly within several months after the first date of sampling, likely indicating a new infection with a different strain. In this respect, further molecular analysis and investigation of epidemiological links among farms are needed to assess and gain insight into the frequency of reinfection and/or the molecular evolution of BVDV strains detected in the same herd and between herds, as recently applied by [41]. Moreover, a third less frequently observed genetic pattern was the presence of different subtypes or genotypes within the same herd at the same date of sampling,

Table 2: Sequence similarity (%) of pairwise distances within herds with ≥2 animals/sequences belonging to the same BVDV-1 subtypes.

Sequence similarity (%)	BVDV-1 subtype	Herd identification	Sequences in each herd (*n*)	Temporal range of collection (months)	Herd production system*
100	1b	RG7	2	2	nd
100	1b	BG23	2	12[a]	D
100	1b	PD3	2	12[a]	nd
100	1b	CO2	2	1[b]	D
100	1b	CN29	2	1[b]	nd
100	1b	AT3	2	1[b]	B
100	1b	CN8	2	1	B
100	1b	TO3	2	1[b]	D
100	1b	TO17	3	13	M
100	1d	CN20	2	1[b]	B
100	1e	RG2	2	8	nd
100	1e	CR22	2	1[b]	D
100	1e	AL2	2	1[b]	M
100	1e	CN24	2	6	D
100	1e	TO8	5	1[b]	D
100	1e	TO14	8	12	B
100	1e	TO19	7	1[b]	D
100	1f	CN2	3	14	M
100	1h	NO1	2	1	D
99.4–100	1e	VI7	4	12[a]	nd
99.4	1b	RG6	2	1[b]	nd
99.4	1b	TO23	2	3	D
99.4	1d	CN19	2	1[b]	D
99.4	1f	AT2	2	1[b]	B
98.8–100	1b	LO2	4	3	D
98.8–100	1f	CN1	7	14	D
98.8	1f	CN5	2	2	D
98.2–100	1b	RG4	4	2	nd
98.2	1b	RM1	2	12[a]	D
98.2	1b	TR1	2	12[a]	B
97.1	1b	CR13	2	12[a]	D
96.5	1b	RM2	2	12[a]	D
94.7–100	1b	LC1	5	129	D
93.6–100	1e	CN16	3	10	B
93–100	1b	RG3	3	3	nd
90.1	1e	CN17	2	13	D
88.9–100	1e	TO12	7	19	M

*D: dairy herd, B: beef herd, M: mixed farm, and nd: not determined.
[a]Data available only for year of sample collection.
[b]Samples collected on the same day.

indicating BVDV cocirculation, possibly through exposure to multiple viral sources. Cocirculation of different BVDV subtypes was detected in both milk and beef production systems, confirming that the diversity of viral strains in the Italian cattle population influences the variability also at the herd level.

5. Conclusion

This comprehensive overview of the genetic variability of BVDV strains circulating in Italy highlights a marked genetic diversity. The temporal-spatial distribution of BVDV variants suggests the risk of different routes of BVDV introduction and dispersion, through different commercial livestock flows

Table 3: Herds with ≥2 animals/sequences belonging to different genotypes and subtypes.

Genotype	Herd identification	Total sequences number	Temporal range of collection (months)	Herd production system*
BVDV-1b, 1e	TO9	2	1**	D
BVDV-1b, 1e	CN7	20	1	B
BVDV-1b, 1f	CN6	2	3	B
BVDV-1b, 1k	CS2	3	12	D
BVDV-1b, 2	BS33	2	1**	D
BVDV-1d, 1e	NO2	3	21	M
BVDV-1d, 1e	MN4	2	1**	D
BVDV-1e, 1f	CN3	5	1**	D
BVDV-1e, 1h	LC2	2	19	D

*D: dairy herd, B: beef herd, and M: mixed farm.
**Samples collected on the same day.

or the use of contaminated biological products, likely related to the lack of coordinated control measures. Also, it highlights the importance of phylogenetic studies for genetic characterization and for reconstruction of the evolutionary relationships between strains through which the ecological and epidemiological mechanisms driving such genetic heterogeneity may be elucidated. Advanced phylogenetic analysis of the evolutionary dynamics of BVDV strains present in a population can aid in tracing transmission chains and prevents and controls infections and sources of reinfections.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was funded by Italian Ministry of Health in the framework of the Research Project "Filodinamica, filogeografia e caratterizzazione molecolare full genome di Bovine Viral Diarrhoea Virus (BVDV)" (Project IZS PLV 16/11RC).

References

[1] T. Childs, "X disease of Cattle—Saskatchewan," *Canadian Journal of Comparative Medicine and Veterinary Science*, vol. 10, no. 11, pp. 316–319, 1946.

[2] P. Olafson, C. A. Mac, and F. H. Fox, "An apparently new transmissible disease of cattle," *The Cornell Veterinarian*, vol. 36, pp. 205–213, 1946.

[3] B. Makoschey, P. T. J. A. V. Gelder, V. Keijsers, and D. Goovaerts, "Bovine viral diarrhoea virus antigen in foetal calf serum batches and consequences of such contamination for vaccine production," *Biologicals*, vol. 31, no. 3, pp. 203–208, 2003.

[4] F. V. Bauermann, J. F. Ridpath, R. Weiblen, and E. F. Flores, "HoBi-like viruses: an emerging group of pestiviruses," *Journal of Veterinary Diagnostic Investigation*, vol. 25, no. 1, pp. 6–15, 2013.

[5] Š. Vilček, B. Ďurkovič, M. Kolesárová, I. Greiser-Wilke, and D. Paton, "Genetic diversity of international bovine viral diarrhoea virus (BVDV) isolates: identification of a new BVDV-1 genetic group," *Veterinary Research*, vol. 35, no. 5, pp. 609–615, 2004.

[6] A. Jackova, M. Novackova, C. Pelletier et al., "The extended genetic diversity of BVDV-1: typing of BVDV isolates from France," *Veterinary Research Communications*, vol. 32, no. 1, pp. 7–11, 2008.

[7] Y. Kadir, F. Christine, B.-W. Barbara et al., "Genetic heterogeneity of bovine viral diarrhoea virus (BVDV) isolates from Turkey: identification of a new subgroup in BVDV-1," *Veterinary Microbiology*, vol. 130, no. 3-4, pp. 258–267, 2008.

[8] M. Nagai, M. Hayashi, M. Itou et al., "Identification of new genetic subtypes of bovine viral diarrhea virus genotype 1 isolated in Japan," *Virus Genes*, vol. 36, no. 1, pp. 135–139, 2008.

[9] F. Xue, Y.-M. Zhu, J. Li et al., "Genotyping of bovine viral diarrhea viruses from cattle in China between 2005 and 2008," *Veterinary Microbiology*, vol. 143, no. 2–4, pp. 379–383, 2010.

[10] S. Gao, J. Luo, J. Du et al., "Serological and molecular evidence for natural infection of Bactrian camels with multiple subgenotypes of bovine viral diarrhea virus in Western China," *Veterinary Microbiology*, vol. 163, no. 1-2, pp. 172–176, 2013.

[11] M. Tajima, H.-R. Frey, O. Yamato et al., "Prevalence of genotypes 1 and 2 of bovine viral diarrhea virus in Lower Saxony, Germany," *Virus Research*, vol. 76, no. 1, pp. 31–42, 2001.

[12] M. Beer, G. Wolf, and O. R. Kaaden, "Phylogenetic analysis of the 5′-untranslated region of german BVDV type II isolates," *Journal of Veterinary Medicine B*, vol. 49, no. 1, pp. 43–47, 2002.

[13] L. Liu, H. Xia, C. Baule, and S. Belák, "Maximum likelihood and Bayesian analyses of a combined nucleotide sequence dataset for genetic characterization of a novel pestivirus, SVA/cont-08," *Archives of Virology*, vol. 154, no. 7, pp. 1111–1116, 2009.

[14] H. Xia, B. Vijayaraghavan, S. Belák, and L. Liu, "Detection and identification of the atypical bovine pestiviruses in commercial foetal bovine serum batches," *PLoS ONE*, vol. 6, no. 12, Article ID e28553, 2011.

[15] K. Ståhl, J. Kampa, S. Alenius et al., "Natural infection of cattle with an atypical "HoBi"-like pestivirus—implications for BVD control and for the safety of biological products," *Veterinary Research*, vol. 38, no. 3, pp. 517–523, 2007.

[16] A. Cortez, M. B. Heinemann, A. M. M. G. De Castro et al., "Genetic characterization of Brazilian bovine viral diarrhea virus isolates by partial nucleotide sequencing of the 5′-UTR

region," *Pesquisa Veterinaria Brasileira*, vol. 26, no. 4, pp. 211–216, 2006.

[17] E. Bianchi, M. Martins, R. Weiblen, and E. F. Flores, "Perfil genotípico e antigênico de amostras do vírus da diarréia viral bovina isoladas no Rio Grande do Sul (2000–2010)," *Pesquisa Veterinaria Brasileira*, vol. 31, no. 8, pp. 649–655, 2011.

[18] N. Decaro, M. S. Lucente, V. Mari et al., "Atypical pestivirus and severe respiratory disease in calves, Europe," *Emerging Infectious Diseases*, vol. 17, no. 8, pp. 1549–1552, 2011.

[19] N. Decaro, V. Mari, M. S. Lucente et al., "Detection of a Hobi-like virus in archival samples suggests circulation of this emerging pestivirus species in Europe prior to 2007," *Veterinary Microbiology*, vol. 167, no. 3-4, pp. 307–313, 2013.

[20] M. Giammarioli, C. Pellegrini, C. Casciari, E. Rossi, and G. M. De Mia, "Genetic diversity of Bovine viral diarrhea virus 1: Italian isolates clustered in at least seven subgenotypes," *Journal of Veterinary Diagnostic Investigation*, vol. 20, no. 6, pp. 783–788, 2008.

[21] C. Luzzago, E. Ebranati, D. Sassera et al., "Spatial and temporal reconstruction of bovine viral diarrhea virus genotype 1 dispersion in Italy," *Infection, Genetics and Evolution*, vol. 12, no. 2, pp. 324–331, 2012.

[22] C. Luzzago, C. Bandi, V. Bronzo, G. Ruffo, and A. Zecconi, "Distribution pattern of bovine viral diarrhoea virus strains in intensive cattle herds in Italy," *Veterinary Microbiology*, vol. 83, no. 3, pp. 265–274, 2001.

[23] E. Falcone, P. Cordioli, M. Tarantino, M. Muscillo, G. La Rosa, and M. Tollis, "Genetic heterogeneity of bovine viral diarrhoea virus in Italy," *Veterinary Research Communications*, vol. 27, no. 6, pp. 485–494, 2003.

[24] V. Cannella, E. Giudice, S. Ciulli et al., "Genotyping of bovine viral diarrhea viruses (BVDV) isolated from cattle in Sicily," *Comparative Clinical Pathology*, vol. 21, no. 6, pp. 1733–1738, 2012.

[25] N. Decaro, V. Mari, P. Pinto et al., "Hobi-like pestivirus: both biotypes isolated from a diseased animal," *Journal of General Virology*, vol. 93, no. 9, pp. 1976–1983, 2012.

[26] N. Decaro, M. S. Lucente, V. Mari et al., "Hobi-like pestivirus in aborted bovine fetuses," *Journal of Clinical Microbiology*, vol. 50, no. 2, pp. 509–512, 2012.

[27] C. Bachofen, H. Stalder, U. Braun, M. Hilbe, F. Ehrensperger, and E. Peterhans, "Co-existence of genetically and antigenically diverse bovine viral diarrhoea viruses in an endemic situation," *Veterinary Microbiology*, vol. 131, no. 1-2, pp. 93–102, 2008.

[28] R. Strong, J. Errington, R. Cook, N. Ross-Smith, P. Wakeley, and F. Steinbach, "Increased phylogenetic diversity of bovine viral diarrhoea virus type 1 isolates in England and Wales since 2001," *Veterinary Microbiology*, vol. 162, no. 2–4, pp. 315–320, 2013.

[29] Š. Vilček, D. J. Paton, B. Durkovic et al., "Bovine viral diarrhoea virus genotype 1 can be separated into at least eleven genetic groups," *Archives of Virology*, vol. 146, no. 1, pp. 99–115, 2001.

[30] S. Ciulli, E. Galletti, M. Battilani et al., "Genetic typing of bovine viral diarrhoea virus: evidence of an increasing number of variants in Italy," *New Microbiologica*, vol. 31, no. 2, pp. 263–271, 2008.

[31] C. Letellier, P. Kerkhofs, G. Wellemans, and E. Vanopdenbosch, "Detection and genotyping of bovine diarrhea virus by reverse transcription-polymerase chain amplification of the 5′ untranslated region," *Veterinary Microbiology*, vol. 64, no. 2-3, pp. 155–167, 1999.

[32] S. Vilcek, A. J. Herring, J. A. Herring, P. F. Nettleton, J. P. Lowings, and D. J. Paton, "Pestiviruses isolated from pigs, cattle and sheep can be allocated into at least three genogroups using polymerase chain reaction and restriction endonuclease analysis," *Archives of Virology*, vol. 136, no. 3-4, pp. 309–323, 1994.

[33] D. Posada, "jModelTest: phylogenetic model averaging," *Molecular Biology and Evolution*, vol. 25, no. 7, pp. 1253–1256, 2008.

[34] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[35] E. Falcone, M. Tollis, and G. Conti, "Bovine viral diarrhea disease associated with a contaminated vaccine," *Vaccine*, vol. 18, no. 5-6, pp. 387–388, 1999.

[36] H. W. Barkema, C. J. M. Bartels, L. Van Wuijckhuise et al., "Outbreak of bovine virus diarrhoea on Dutch dairy farms induced by a bovine herpesvirus 1 marker vaccine contaminated with bovine virus diarrhoea virus type 2," *Tijdschrift voor Diergeneeskunde*, vol. 126, no. 6, pp. 158–165, 2001.

[37] N. Decaro, R. Sciarretta, M. S. Lucente et al., "A nested pcr approach for unambiguous typing of pestiviruses infecting cattle," *Molecular and Cellular Probes*, vol. 26, no. 1, pp. 42–46, 2012.

[38] S. Peletto, F. Zuccon, M. Pitti et al., "Detection and phylogenetic analysis of an atypical pestivirus, strain IZSPLV_To," *Research in Veterinary Science*, vol. 92, no. 1, pp. 147–150, 2012.

[39] I. Toplak, T. Sandvik, D. Barlič-Maganja, J. Grom, and D. J. Paton, "Genetic typing of bovine viral diarrhoea virus: most Slovenian isolates are of genotypes 1d and 1f," *Veterinary Microbiology*, vol. 99, no. 3-4, pp. 175–185, 2004.

[40] A. Hornberg, S. R. Fernández, C. Vogl et al., "Genetic diversity of pestivirus isolates in cattle from Western Austria," *Veterinary Microbiology*, vol. 135, no. 3-4, pp. 205–213, 2009.

[41] R. E. Booth, C. J. Thomas, L. M. El-Attar, G. Gunn, and J. Brownlie, "A phylogenetic analysis of bovine viral diarrhoea virus (BVDV) isolates from six different regions of the UK and links to animal movement data," *Veterinary Research*, vol. 44, no. 1, article 43, pp. 1–14, 2013.