A First Genome Survey and Microsatellite Motif Identification of Taihangia rupestris

Lin-lin Song (✉ cnusll@hist.edu.cn)

   Henan Institute of Science and Technology   https://orcid.org/0000-0001-9514-9286

Hong-zhi Chen

   Xinxiang Institute of Engineering

Bao-ying Zhou

   Henan Institute of Science and Technology

Li Meng

   Henan Institute of Science and Technology

Abstract

Background: *Taihangia rupestris* is a perennial herb on the China species red list that is growing on the cliffs of Taihang Mountain in China. However research on the genome of *T. rupestris* has not been carried out, which severely restricts further research on it. The aim of this study was to conduct a first genome survey of *T. rupestris* and to develop SSR molecular markers of it.

Methods: The genome size and characteristics of *T. rupestris* were estimated by Illumina Hi-SeqXTen and K-mer analysis. We designed SSR primers in batches with Misa and Primer3, and *T. rupestris* from different populations were used to verify the selected primers. Finally, datas were analysised by Cervus 3.0 and GenAlex 6.5 for genetic diversity.

Results: The genome size of *T. rupestris* was estimated to be 976.97 Mb with a heterozygosity rate of 0.726% and a sequence repetition rate of 56.93%. The clean reads were assembled into 100973 contigs with the max length of 26073 bp and an N50 value of 2607 bp. Based on the genome data of *T. rupestris*, a total of 805600 SSR markers were identified and 72769 pairs of primers were designed. In the present study, 100 primers were used to verify that 82 primers were successfully amplified.

Conclusion: In general, the genome of *T. rupestris* is difficult to assemble genome with micro-heterozygosity and high repetition. In this study, 15 pairs of primers with good polymorphism can effectively distinguish different populations of *T. rupestris*. These analyses laid a foundation for the subsequent whole genome sequencing of *T. rupestris*.

Introduction

Taihangia rupestris Yu et Li is a perennial herb native to the tribe Dryadeaeis that grows sporadically on cliff sides located in the eastern Taihang Mountains [1] (Fig. 1). This species grows at altitudes of 600 to 1500 meters and prefers U-shaped cliffs where there is little direct sunlight [2]. *Taihangia rupestris* is included on the China species red list as being threatened with extinction in the wild [1]. However, current research in *T. rupestris* has primarily focused on artificial cultivation and morphological differentiation [3], nucleotide diversity [4, 5], and the regulatory mechanisms of staminate and perfect flower [6–8], but genome research has not yet been carried out.

Biological genetic information is intimately associated with physiological traits, according to the available research report [9]. Research on the genome of *T. rupestris* has greatly aided the study of its genetic evolution and biodiversity conservation [10]. Genome survey sequencing was utilized in this research to assess the genome size, repeat sequence, heterozygosity, and other constituent properties of *T. rupestris*, laying the groundwork for genome deep sequencing. Simple sequence repeats (SSRs), which have been frequently used as genetic markers, are tandem repeats consisting of 1 to 6 nucleotides with high polymorphism and abundance [4, 10]. SSR markers were highly polymorphic and codominant, thus they are useful for genetic research and gene resource usage.

In this paper, we will obtain genome size, repetiton rate, GC content, and other related information about *T. rupestris* through a genome survey, and develop SSR primers and systematically classify *T. rupestris* in different populations.

Materials And Procedures

Materials from plants and genome DNA extraction

Fresh leaves of *T. rupestris* for genome survey were obtained from Guanshan National Geopark (35°33'46"N, 113°31'31"E). The CTAB (cetyl trimethyl ammonium bromide) approach was used to extract the genomic DNA [11]. The concentration of DNA samples was determined by ultraviolet spectrophotometer, and purity was measured by an absorption ratio of 260 nm to 280 nm. The *T.rupestris* samples of the two groups used to detect primers were from Guanshan National Geopark, which were named population A (35°33'46"N, 113°31'31"E, n=24) and population B (35°33'54"N, 113°31'40"E, n= 21), respectively.

Genome sequencing

DNA samples of the *T. rupestris* were sheared randomly by ultrasonicator (Covaris M200) to be about 350 bp in size. DNA fragments of the required size were obtained by electrophoresis before end-repair, and then added sequencing adapters and poly A-tail. Two paired-end DNA libraries were constructed using 350 bp fragments and then sequenced at a 2×150 bp read length [12]. All clean reads for further analysis were obtained after removing signals that could interfere with subsequent analysis, such as reads with N ratios greater than 10%, adaptor reads, amplification repeats from PCR reaction, and reads of low-quality. Reads of high-quality have been contributed to the Sequence Read Archive in NCBI (accession number: PRJNA797565).

Assessment of genome size, heterozygosity, and repetition rate

Before genome assembly, GENOMESCOPE software was used to assess the genome characteristics, including genome size, repeat content, and heterozygosity rate of *T. rupestris* based on K-mer analysis (k=17) [13]. GC content and sequencing depth were calculated along the assembled sequence using a 10 kb non-overlapping sliding window.

Design and analysis of SSRs characteristics

The MISA program was used to find SSR markers. The following were the criteria: four repeat units with Hexa-, penta- and tetranucleotides, five with trinucleotide, six with dinucleotide, and ten with mononucleotides. Primer pairs for SSRs flanking area were designed using the Primer 3 program.

Validation of SSR markers

To verify the primers, a total of 100 pairs of primers were synthesized and PCR was carried out for amplification in 6 individuals of *T. rupestris*. The number of individuals from A and B was 3 and 3 respectively. The optimized PCR was carried out in a mixture of 0.25 μM forward and reverse primers, 5 μl PCR Mixture, and 20 ng genomic DNA in a 10 μl volume. The procedure was set to denaturation at 94°C for 5 min, then go through 30 cycles with 50 s at 94°C, 45 s at 60°C and 1 min at 72°C, and final extension at 72°C for 10 min [12]. The amplification product resolves the silver stain from a 10% PAGE (polyacrylamide gel electrophoresis). The pBR322 DNA/MspI marker (Tiangen, Beijing, China) was used to compare each SSR-PCR product.

Principal coordinate analysis

Fifteen pairs of SSR primers with good polymorphism were selected to conduct the genetic analysis of the 45 individuals from the two populations (A and B). The amplified data were analysised with Cervus 3.0 and GenAlEx 6.5. The principal coordinate analysis (PCoA) was performed based on Nei's genetic distance of the individuals in different populations [14].

Results And Discussion

Genome sequencing data statistics and genome estimation

The Illumina paired-end library was used to obtain a total of 108.97 Gb of raw data for *T. rupestris* in this study (Table 1). After filtering and rectification, there were 107.83 Gb of clean bases generated. Quality value is an important indicator of second-generation sequencing, which is obtained by sequencing every base in the genome. The higher it is, the more accurate the sequencing will be. The Q20 was 95.98% and the Q30 was 90.39% in this experiment. This suggests that the sequencing was of high quality. Twenty thousand clean reads were randomly selected and compared to NT (Nucleotide Sequence Database) database with BLAST in NCBI. The first five matched species were *Geum rupestre* (15.14%), *Rosa chinensis* (1.51%), *Fragaria vesca* (0.74%), *Rosa praelucens* (0.64%) and *Prunus pedunculata* (0.44%). The first number was

higher than the others because it was also Rosaceae the same as *T. rupestris.* The results showed no contamination from other species.

All clean data in this study was analyzed with a K-mer 17 (Fig. 2). The x-axis and y-axis in Fig. 2 represent the depth of k-mer and the number of corresponding depths, respectively. Then the expected genome size was obtained with the following formula: Genome size = K-mer num/K-mer depth [15]. The estimated genome size is 976.97 Mb. We determined that the gene heterozygosity rate of *T. rupestris* was 0.726% while the repetition rate was 56.93%. It belongs to the genome with micro-heterozygosity and high-repetition. The results showed that assembling a genome for *T. rupestris* is difficult.

Genome GC analysis and genome assembly

The assembled sequences of *T. rupestris* genome were subsequently studied for GC analysis (Fig. 3). The density points were mostly found in areas with a GC content of 20–50%, with an average of 33.75%. It is similar to *Apocynum venetum* (32.91%) [16] and *Macaranga indica* (33.83%) [17], but lower than *Cydonia oblonga* (38.66%) [18], *Rosa multiflora* (38.9%) [19], *Fragaria nilgerrensis* (39.22%) [20] and *Rosa rugosa* (39.30%) [21]. Above all, the genome of *T. rupestris* has a mid-GC content.

As shown in Table 1, the total number of contigs was 100973. In total, 91276077 bp contigs were generated from the *T. rupestris* genome, with the largest contig length of 26073 bp, a N50 contig length of 2607 bp and a N90 contig length of 1189 bp.

Table 1
Assembled sequencing results data statistics for
T. rupestris

Features	Number or ratio
Total number of bases (Gb)	108.97
Clean reads (Gb)	107.83
Clean read proportion (%)	98.95%
Q20	95.98%
Q30	90.39%
GC content	33.75%
heterozygosity rate	0.726%
repetition rate	56.93%
Total number of contigs	100973
Total length of contigs (bp)	91276077
Max length (bp)	26073
N50 (bp)	2607
N90 (bp)	1189

SSR identification and validation

Compared with the traditional design methods, using genomic data to develop SSR molecular markers is simple and effective [12]. Data showed that 805600 SSRs were discovered in this study, and the frequency of SSR repeats was

significantly different (Fig. 4). As shown in Fig. 5, the most common form of repeat was mononucleotide (63.70%), followed by dinucleotide (18.77%), tetranucleotide (11.63%), trinucleotide (3.25%), hexanucleotide (2.43%), and pentanucleotide (0.22%). SSR markers with both mononucleotide and dinucleotide sequences accounted for 82.47% of the total. A/T ratios were bigger than G/C ratios among mononucleotide repeats, accounting for 97.01% of the total (513158). The biggest number of dinucleotide was AG/CT (83998, 55.56%), followed by AT/AT (59954, 39.66%), AC/GT (7134, 4.72%) and CG/CG (94, 0.06%).The highest number among the trinucleotides was AAG/CTT repeats (7871, 30.03%), followed by ACC/GGT (4948, 18.88%), AAT/ATT (3679, 14.04%) and AAC/GTT (2889, 11.02%). The most dominant tetranucleotide is AGAG/CTCT among the 36 tetranucleotide types, which accounted for 58.22%, followed by ATAT/ATAT (33277, 35.51%), ACAC/GTGT (3096, 3.30%) and AAAT/ATTT (1210, 1.29%).

Validation of SSR markers

In the present study, 562246 SSR-containing sequences were used to construct primers, resulting in a total of 72769 primer pairs for use. We synthesized 100 primer pairs to detect these SSR markers, and the results showed there were 82 primer pairs successfully amplified achieving the desired size (Fig. 6). We selected 15 pairs of primers with good polymorphism to perform cluster analysis on 45 individuals from two populations, and the result was shown in Fig. 7.

Information about these 15 primer pairs is provided in Table 2. The results showed that the average allele (Na) of the 15 SSR loci was 3.67, the minimum observed heterozygosity (HO) and expected heterozygosity (HE) were 0.095 and 0.167, and the maximum HO and HE were 1.000 and 0.788, respectively. The average PIC (polymorphic information content) value was 0.44, with four highly polymorphic loci (TH17, TH25, TH31 and TH58) having PIC values higher than 0.5.

Table 2
Genetic characteristic statistics generated by 15 SSR markers on 45 *T. rupestris* individuals.

Locus	Motif	Primer sequence	k	HO	HE	PIC	HW	F(Null)
TH4	(ATTT)7	TGGGCTACGACGATTGAACA GCATGTACTAGCAAACTCGCA	3	0.289	0.461	0.378	ND	0.2225
TH17	(ATGAG)5	GTGTGCAAGTGGTTGGTTGG CTGCACCGTACCATCATGGA	5	0.444	0.672	**0.623**	NS	0.183
TH22	(GAT)9	TGTCTGATTCGGGCCCTAGA TGGCATGTGATTCGCCTTCT	7	0.289	0.508	0.481	ND	0.3385
TH25	(GAA)8	TGGCATAAAGAGTGGTCTGAGG AACGGTCTCTCCTCCTCCTC	6	0.409	0.788	**0.746**	ND	0.3146
TH28	(GCCA)5	TCTGTTTTGTTTAAGGCGTGCA ACACGTGTCATCTCGTCATTGA	3	0.182	0.581	0.484	***	0.5428
TH31	(AGAAGA)5(AGA)10	CGTGCGATTGGTGTACCTCT AGAGGTCATTACGATTTACAACCA	4	0.682	0.714	**0.653**	NS	0.0189
TH41	(T)13(TT)6	GCTAGCCAACACACCACTCT ACCCTAGGTGGCTACGAGTT	2	0.295	0.255	0.22	ND	-0.0764
TH44	(TTC)7	ACTCGATCCTCTCCCTAAAGGA CATAGGAGACGAGCAGAGGC	3	0.095	0.304	0.268	ND	0.5373
TH49	(AT)13(ATAT)6	GTTGTACTAGGTGGCTGCCC GCTGATGGCTAGGATTCACT	2	1.000	0.506	0.375	***	-0.333
TH50	(ACC)7	ACGACGTCACCTCCGTAAAC AGATTGAAGAGGCGGTGGTG	2	0.605	0.499	0.372	NS	-0.1015
TH53	(CTT)9	CCTACTGGCATCGAGACACC GGGATCTCCACTCCAACAGC	3	0.128	0.167	0.155	ND	0.1229
TH58	(CAAAA)5	TCATTCTCTGCACCAACCCC GGACGTGGAGGCATTCTTGA	6	0.622	0.773	**0.729**	NS	0.1122
TH67	(ATGGTG)5	CACAATCTTCCCTAAAAAGGCACA GAACCAAACCGCCCGAATTC	3	0.200	0.388	0.349	ND	0.3055
TH80	(GAAGAT)5	TTGTCATCTTCCGCGGTGAA TCCACACCCTCATGATCGGA	3	0.089	0.401	0.36	ND	0.6677
TH92	(TTTGT)5	TGAATGGGCAAAGGTCAACT ACCATACAAAGTTTTTGCATCCT	3	1.000	0.564	0.46	***	-0.3011
Mean			3.67	0.42	0.51	0.44		0.17

k: the number of alleles, HO: Observed heterozygosity, HE: Expected heterozygosity, PIC: Polymorphism information content, HW: Significance of deviation from Hardy-Weinberg equilibrium. NS = not significant, *** = significant at the 0.1% level, ND = not done. These significance levels include a Bonferroni correction if the Bonferroni correction option was selected. F (null): Null allele frequency.

The results in Fig. 7 showed that the SSR markers could distinguish genetic differences among populations based on different geographical locations. Since the two populations sampled were only one kilometer apart, there was gradual genetic penetration between the populations. Two coordinates explain 15.73% and 13.14% of the overall genetic variation, respectively.

Numerous genome-wide SSR loci have been identified, paving the way for the creation of high-density genetic maps and research into the regulatory mechanisms of *T. rupestris* therapeutic components. It also serves as a useful reference for future genomic in vestigations and molecular markers.

Discussion

T. rupestris belonging to ancient relic species is rosaceae herb fairy, wooden race's most original diploid plants (chromosome 2n=14). It has great significance that the flower structure changed from bisexual to unisexual by sexual evolution, which holds a special place in the fairy wooden race evolution for illuminating the origin and evolution of Rosaceae some taxa problems.

To measure the genome size of a species before genome sequencing can provide data support for the depth of genome sequencing. In this paper, after removing the K-mer mistake, the corrected genome size of *T. rupestris* was 976.97 Mb. These estimates were greater than the populations of *Fragaria vesca* (240 Mb) [22], *Pyrus bretschneideri* (527 Mb) [23], *Ficus erecta* (341 Mb) [24] and *Cydonia oblonga* (686 Mb) [18], smaller than *Arachis duranensis* (1.25 Gb) [25], *Lupinus angustifolius* (1.15 Gb) [26] and *Prunus fruticose* (1.2 Gb) [27], but are close to *Begonia fuchsioides* (935 Mb) [28] and *Quercus suber* (953 Mb) [29]. According to the heterozygosity and repetition, the genome was divided into low-heterozygosity (≥50%), micro-heterozygosity (0.5% 0.8%), high-heterozygosity (≥0.8%) and high-repetition (repeated ratio≥50 %), which directly reflects the difficulty of assembling the subsequent sequencing data [12,30]. The gene heterozygosity rate of *T. rupestris* was 0.726% while the repetition rate was 56.93%. It showed that assembling a genome for *T. rupestris* is difficult. The result has laid a solid platform for *T. rupestris* whole-genome sequencing.

  SSR markers have been widely used to analyze genetic characteristics in plants due to the advantages of co-dominant and stability. Wang *et al.* [31] obtained 10 SSR markers, which were extracted through microsatellite-enriched library construction, enzymolysis and probe retrieval. Wang *et al.* [32] amplified the nuclear genome DNA of 220 *T. rupestris* using 12 pairs of ISSR primers (from the University of British Columbia) to study the genetic diversity and genetic characteristics of this species. Sun *et al.* [2] claimed that DNA barcodes of two subspecies of Taihangia were constructed using five single loci from chloroplast, ITS and ITS2 from the nuclear genome. Cheng *et al.* [4] found that six chloroplast DNA loci and five mitochondrial DNA loci were used well to distinguish two subspecies of Taihangia and to study the genetic diversity among them. All of the design patterns as mentioned above were developed through enrichment libraries or molecular markers based on organelle DNA, but the number of molecular markers developed was so limited. With the development and widespread application of high-throughput sequencing technology, we developed SSR molecular markers of *T. rupestris* in large quantities. In this study, the total number of sequences examined was 4684186, the total number of identified SSRs was 805600, the number of SSR containing sequences was 562246, thus the occurrence frequency of SSR is 12%, the average distribution frequency was 824.59 SSRs per Mb, which was much higher than other plants such as *Schima superba* (644 SSRs per Mb) [33], *Fagopyrum tartaricum* (114.07 SSRs per Mb) [34], *Cunninghamia lanceolate* (1307.24 SSRs per Mb) [35], *Platostoma palustre* (12.8 SSRs per Mb) [36]. This result is inconsistent with the idea that genome size is negatively correlated with SSR distribution frequency [37]. Mononucleotides and dinucleotides were the majority of the SSRs. The percentage of A/T was higher than C/G. Among dinucleotides and trinucleotides, the proportion of AG/CT and

AAG/CTT were the highest that it was consistent with the results of most plants, such as *E.gunnii* [38], *E.camaldulensis* [39], *Pyrus pyrifolia* [40] and *Taxus fauna* [41]. In this study, 15 SSR markers were used for detection, which could reflect the genetic diversity of *T. rupestris* population to a certain extent. The expected heterozygosity range of these microsatellite loci was 0.167 0.788, and the mean value was 0.51, indicating that the genomic microsatellite of *T. rupestris* had moderate heterozygosity and genetic diversity [14]. Polymorphic information content (PIC) is the probability that an offspring acquired an allele marker from the same allele in its parent value, and PIC>0.5, it is a highly polymorphic site; 0.25<PIC<0.5, it was a moderately polymorphic site; PIC>0.25, it is a low degree polymorphic site. In this study, the PIC values of the 15 microsatellite loci ranged from 0.155 to 0.746, and the average value was 0.44, indicating that these microsatellite loci contained moderately genetic information in the genome of *T. rupestris*, which were suitable for genetic analysis and application of *T. rupestris*.

Conclusion

Prior to this study, there were no reports to evaluate the genomic associated characteristics of *T. rupestris.* Results showed that *T. rupestris* had a genome size of 976.97 Mb, with gene heterozygosity and duplication rates of 0.726% and 56.93% respectively. The average percentage of GC in the sample was 33.75%. A total of 805600 SSR markers were created, along with three sets of primers. Through the synthesis and verification of 100 pairs of primers in 45 *T. rupestris* individuals, it is provided that the SSRs obtained in this experiment can effectively distinguish different populations. The findings pave the way for further genomic research and molecular breeding of *T. rupestris*.

Abbreviations

SSR, simple sequence repeat; NCBI, National Center for Biotechnology Information;

MISA, MicroSAtellite identification tool; Blast, The Basic Local Alignment Search Tool

Declarations

Competing Interests

The authors state that the work does not include any competing interests.

Author Contribution

L.L.S. and H.Z.C. was in charge of the experiments, data analysis, and manuscript writing. The sample was prepared by B.Y.Z. The manuscript was revised by L.M. The final paper was approved by all authors.

Funding

Data availability statement

The *T. rupestris* genome project has been registered in NCBI under the BioProject number PRJNA797565. The whole-genome sequence has been deposited in the Sequence Read Archive (SRA) database under accession numbers SRR17639490 and SRR17775415. https://www.ncbi.nlm.nih.gov/sra/PRJNA797565

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Duan, N., Liu, S., and Liu, B.B. (2018) Complete chloroplast genome of *Taihangia rupestris* var. *rupestris* (Rosaceae), a rare cliff flower endemic to China. *Conservation Genetics Resources* 10, 809-811, http://doi.org/10.1007/s12686-017-0936-5

2. Sun, X., Wang, Y.P., Liu, C., and Huang, L.F. (2019) Molecular identification of *Taihangia rupestris* Yu et Li, an endangered species endemic to China. *South African Journal of Botany* 124, 173-177, http://doi.org/10.1016/j.sajb.2019.04.026

3. Wang, L.M., Zhu, M.X., Han, G.Y., Wang, Y. J. (2012) Rooting culture and Acclimation and transplanting of tissue culture plantlets of the endangered plant *Taigangia rupestris* var. ciliata. *Jiangsu Agricultural Sciences* 40,64-65,http://doi.org/10.15889/j.issn.1002-1302.2012.07.122

4. Cheng, Y., Duan, J., Jiao, Z., Wang, G.G., Yan, F., and Wang, H. (2016) Cytoplasmic DNA disclose high nucleotide diversity and different phylogenetic pattern in *Taihangia rupestris* Yu et Li. *Biochemical Systematics and Ecology* 66, 201-208, https://doi.org/10.1016/j.bse.2016.04.009

5. Li, W., Liu, S., Jiang, S., Li, X., and Li, G. (2018) Development of 30 SNP markers for the endangered plant *Taihangia rupestris* based on transcriptome database and high resolution melting analysis. *Conservation Genetics Resources* 10, 775-778, http://doi.org/ 10.1007/s12686-017-0927-6

6. Li, W., Zhang, L., Ding, Z., Wang, G., Zhang, Y., Gong, H. et al. (2017a) De novo sequencing and comparative transcriptome analysis of the male and hermaphroditic flowers provide insights into the regulation of flower formation in andromonoecious *Taihangia rupestris*. *BMC Plant Biology* 17, 4-19, http://doi.org/HYPERLINK "http://doi.org/10.1186/s12870-017-0990-x"10.1186/s12870-017-0990-x

7. Li, W., Zhang, L., Zhang, Y., Wang, G., Song, D., and Zhang, Y. (2017b) Selection and validation of appropriate reference genes for quantitative real-time PCR normalization in staminate and perfect flowers of andromonoecious *Taihangia rupestris*. *Frontiers in plant science* 8, 729, http://doi.org/HYPERLINK "http://doi.org/10.3389/fpls.2017.00729"10.3389/fpls.2017.00729

8. Li, W., Ma, Y., Zheng, C., and Li, G. (2022) Variations of cytosine methylation patterns between staminate and perfect flowers within andromonoecious *Taihangia rupestris* (Rosaceae) revealed by methylation-sensitive amplification polymorphism. *Journal of Plant Growth Regulation* 41, 351-363, http://doi.org/HYPERLINK "http://doi.org/10.1007/s00344-021-10308-3"10.1007/s00344-021-10308-3

9. Liu, L.S. and Guo, Y.D. (2011) Analysis of genome content in some Cruciferous vegetables. *Journal of Plant Genetic Resources* 12, 103-106, http://doi.org/10.13430/j.cnki.jpgr.2011.01.027

10. Zhou, X.J., Liu, M.X., Lu, X.Y., Sun, S.S., Cheng, Y.W., and Ya, H.Y. (2020) Genome survey sequencing and identification of genomic SSR markers for *Rhododendron micranthum*. *Bioscience reports* 40, 1-7, http://doi.org/HYPERLINK "http://doi.org/10.1042/BSR20200988"10.1042/BSR20200988

11. Porebski, S., Bailey, L.G., and Baum, B.R. (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant molecular biology reporter* 15, 8-15, http://doi.org/10.1007/bf02772108

12. Zhou, X.J., Wang, Y.Y., Xu, Y.N., Yan, R.S., Zhao, P., and Liu, W.Z. (2015) De novo characterization of flower bud transcriptomes and the development of EST-SSR markers for the endangered tree *Tapiscia sinensis*. *International journal of molecular sciences* 16, 12855-12870, http://doi.org/10.3390/ijms160612855

13. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. et al. (2017) GenomeScope: fast reference-free genome profilingfrom short reads. *Bioinformatics* 33, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153

14. Karbstein K., Tomasello S., Prinz K. (2019) Desert-like badlands and surrounding (semi-) dry grasslands of Central Germany promote small-scale phenotypic and genetic differentiation in *Thymus praecox*. *Ecology and Evolution* 9:

14066–14084, https://doi.org/HYPERLINK "https://doi.org/10.1002/ece3.5844"10.1002/ece3.5844

15. Marcais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770, https://doi.org/10.1093/bioinformatics/btr011

16. Li, G.Q., Song, L.X., Jin, C.Q., Li, M., Gong, S.P., and Wang, Y.F. (2019) Genome survey and SSR analysis of *Apocynum venetum*. *Bioscience reports* 39, https://doi.org/10.1042/BSR20190146

17. Li, J.Y., Lu, T.Q., Yang, J.B., and Tian, B. (2021) Genome survey and analysis of SSR molecular on *Macarange indica* and *M.denticulata*. *Guihaia* 41, 1897-1904, http://HYPERLINK "http://kns.cnki.net/kcms/detail/"kns.cnki.net/kcms/detail/45.1134.Q.20200608.1940.014.html

18. Soyturk, A., Sen, F., Uncu, A.T., Celik, I., and Uncu, A.O. (2021) De novo assembly and characterization of the first draft genome of quince (*Cydonia oblonga* Mill.). *Scientific reports* 11, 1-11, https://doi.org/10.1038/s41598-021-83113-3

19. Nakamura, N., Hirakawa, H., Sato, S., Otagaki, S., Matsumoto, S., Tabata, S. et al. (2018) Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses. *Dna Research* 25, 113-121, https://doi.org/10.1093/dnares/dsx042

20. Zhang, J., Lei, Y., Wang, B., Li, S., Yu, S., Wang, Y. et al. (2020) The high-quality genome of diploid strawberry (*Fragaria nilgerrensis*) provides new insights into anthocyanin accumulation. *Plant biotechnology journal* 18, 1908-1924, https://doi.org/ 10.1111/pbi.13351

21. Chen, F., Su, L., Hu, S., Xue, J.Y., Liu, H., Liu, G. et al. (2021) A chromosome-level genome assembly of rugged rose (*Rosa rugosa*) provides insights into its evolution, ecology, and floral characteristics. *Horticulture Research* 8, 1-13, https://doi.org/10.1038/s41438-021-00594-z

22. Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L. et al. (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nature genetics* 43, 109-116, https://doi.org/HYPERLINK "https://doi.org/10.1038/ng.740"10.1038/ng.740

23. Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S.et al. (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome research* 23, 396-408, https://doi.org/10.1101/gr.144311.112

24. Shirasawa, K., Yakushiji, H., Nishimura, R., Morita, T., Jikumaru, S., Ikegami, H. et al. (2020) The Ficus erecta genome aids *Ceratocystis* canker resistance breeding in common fig (*F. carica*). *The Plant Journal* 102, 1313-1322, https://doi.org/HYPERLINK "https://doi.org/10.1111/tpj.14703"10.1111/tpj.14703

25. Bertioli, D.J., Cannon, S.B., Froenicke, L., Huang, G., Farmer, A.D., Cannon, E.K. et al. (2016) The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature genetics* 48, 438-446, https://doi.org/HYPERLINK "https://doi.org/10.1038/ng.3517"10.1038/ng.3517

26. Yang, H., Tao, Y., Zheng, Z., Zhang, Q., Zhou, G., Sweetingham, M. W.et al. (2013) Draft genome sequence, and a sequence-defined genetic linkage map of the legume crop species *Lupinus angustifolius* L. *PloS one* 8, e64799, https://doi.org/HYPERLINK "https://doi.org/10.1371/journal.pone.0064799"10.1371/journal.pone.0064799

27. Woehner, T.W., Emeriewen, O.F., Wittenberg, A., Schneiders, H., Vrijenhoek, I., Halasz, J. et al. (2021) The draft chromosome-level genome assembly of tetraploid ground cherry (*Prunus fruticosa* Pall.) from long reads. *Genomics* 113, 4173-4183, https://doi.org/10.1101/2021.06.01.446499

28. Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M.B. et al. (2018) Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* 361, 1-16, http://dx.doi.org/10.1126/science.aat1743

29. Ramos, A.M., Usié, A., Barbosa, P., Barros, P. M., Capote, T., Chaves, I. et al. (2018) The draft genome sequence of cork oak. *Scientific data* 5, 1-12, https://doi.org/10.1038/sdata.2018.69

30. Wu, Y.F., Xiao, F.M., Xu, H.N., Zhang, T. and Jiang, X.M. (2014) Genome survey in *Cinnamomum camphora* L. Presl. *J. Plant Genet. Resour* 15, 149–152, https://doi.org/HYPERLINK "https://doi.org/10.13430%20/j.cnki.jpgr.2014.01.020"10.13430 /j.cnki.jpgr.2014.01.020

31. Wang, H.W., Bing, Z., Wang, Z.S., Cheng, Y. Q., and Ye, Y. Z. (2010) Development and characterization of microsatellite loci in *Taihangia rupestris* (Rosaceae), a rare cliff herb. *American Journal of Botany* 97(12), e136-8, https://doi.org/

10.3732/ajb.1000334

32. Wang, H.W., Fang, X.M., Ye, Y.Z., Cheng, Y.Q., and Wang, Z.S. (2011) High genetic diversity in *Taihangia rupestris* yu et li, a rare cliff herb endemic to china, based on inter-simple sequence repeat markers. *Biochemical Systematics and Ecology* 39(4-6), 553-561, https://doi.org/ 10.1016/j.bse.2011.08.004

33. Lin, Y., He, Z.D., Mao, J.P., Jiang, K.B., Wang J.B., and Huang, S.W. (2018) Development and preliminary analysis of SSR locus in *Schima superba* genome. *Chinese Journal of Tropical Crops* 2018, 39(9): 1766-1771

34. Hou, S.Y., Sun, Z.X., Linghu, B., Xu, D.M., Wu, B., Zhang, B. et al. (2016) Genetic diversity of buckwheat cultivars (*Fagopyrum tartaricum* Gaertn.) assessed with ssr markers developed from genome survey sequences. *Plant Molecular Biology Reporter* 34(1), 233-241, http://dx.doi.org/ 10.1007/s11105-015-0907-5

35. Xu, Y., Chen, J.H., Li, Y., Hong, Z., Wang, Y., Zhao, Y.Q. et al. (2014) Development of EST-SSR and genomic-SSR in Chinese fir. *Journal of Nanjing Forestry University (Natural Sciences Edition)* 38(1), 9-14, https://doi.org/ 10.3969/j.issn.1000-2006.2014.01.002

36. Zheng, Z., Zhang, N., Huang, Z., Zeng, Q., Huang, Y., Qi, Y. (2022) Genome sequencing and characterization of simple sequence repeat (SSR) markers in *Platostoma palustre* (Blume) A.J.Paton (Chinese mesona). *Scientific Reports* 12(1): 3501-3469, https://doi.org/ 10.1038/s41598-021-04264-x

37. Morgante, M., Hanafey, M., and Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30(2), 194-200, https://doi.org/10.1038/ng822

38. Rengel, D., San, C.H., Servant, F., Ladouce, N., Paux, E., Wincker, P., Couloux, A., et al. (2009) A new genomic resource dedicated to wood formation in Eucalyptus. *BMC Plant Biology* 9: 36-36, http://doi.org/10.1186/1471-2229-9-36

39. Hirakawa, H., Nakamura, Y., Kaneko, T., Isobe, S., Sakai, H., Kato, T., et al. (2011) Survey of the genetic information carried in the genome of *Eucalyptus camaldulensis*. *Plant Biotechnology* 28(5): 471-480, http://doi.org/10.5511/plantbiotechnol-ogy.11.1027b

40. Yue, X.Y., Liu, G.Q., Zong, Y., Teng, Y.W., Cai, D.Y. (2014). Development of genic SSR markers from transcriptome sequencing of pear buds. *Journal of Zhejiang University SCIENCE B* 15(4), 303–312. https://doi.org/ 10.1631/jzus.B1300240

41. Shen, X.B., Zhu, Y.J., Xu, G.B. (2021) Distribution characteristics of SSR loci and development of molecular markers in *Taxus fauna*. *Journal of central south university of forestry and technology* 41: 139-146, https://doi.org/10.14067/j.cnki.1673-923x.2021.04.016

Figures

Figure 1

Leaves and inflorescences of *T. rupestris* in the Guanshan National Geopark (35°33'46"N,113°31'31"E)
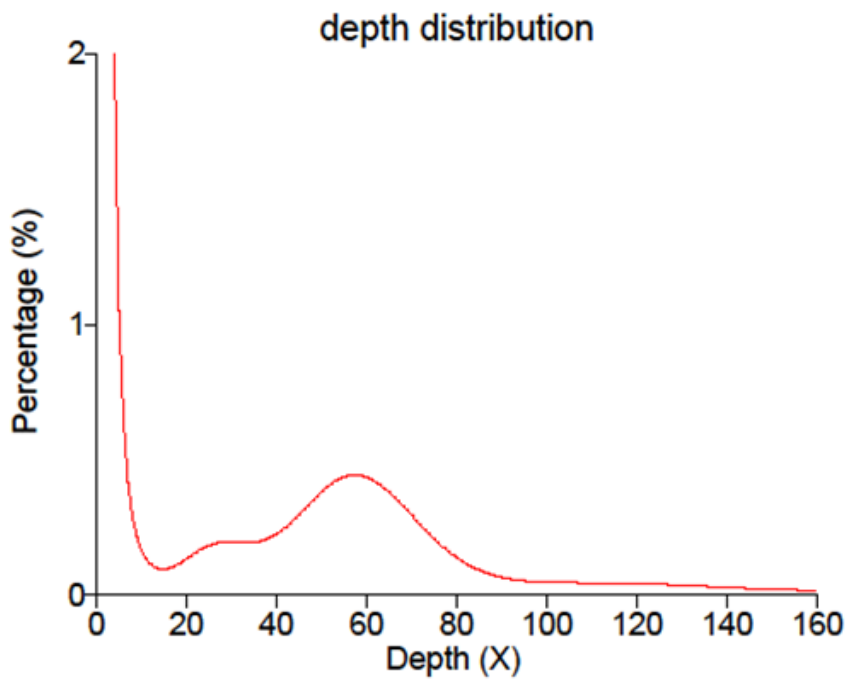

Figure 2

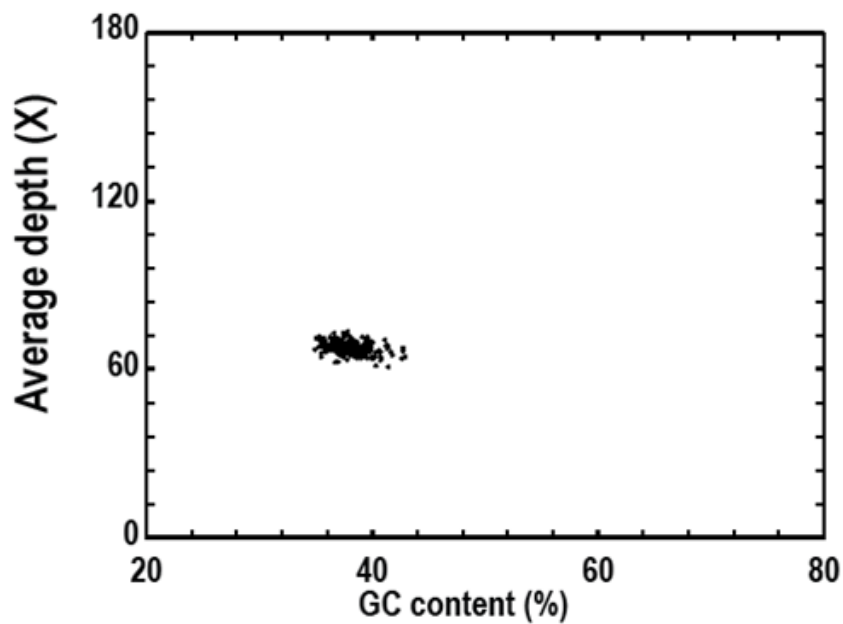K-mer distribution curve (k=17) of *T. rupestris*

Figure 3

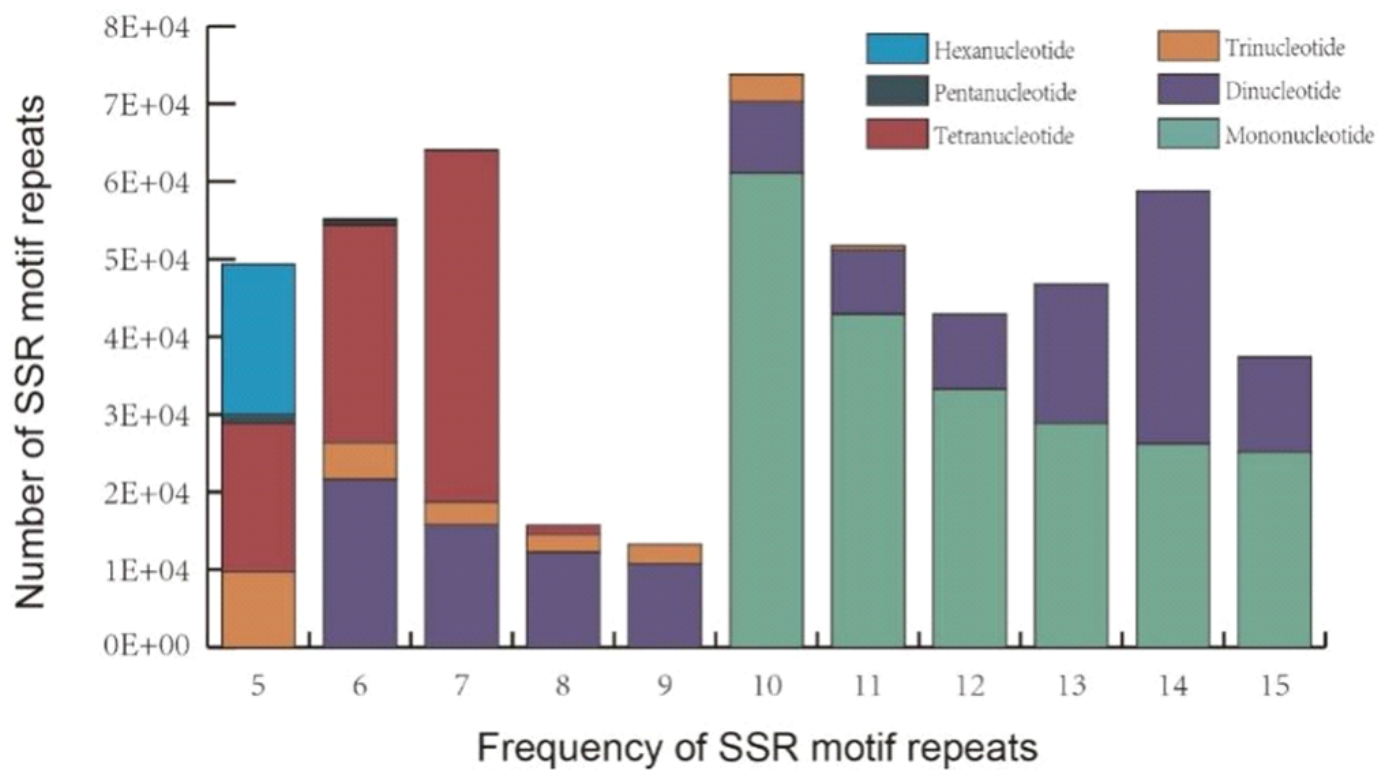GC analysis of *T. rupestris* genome data
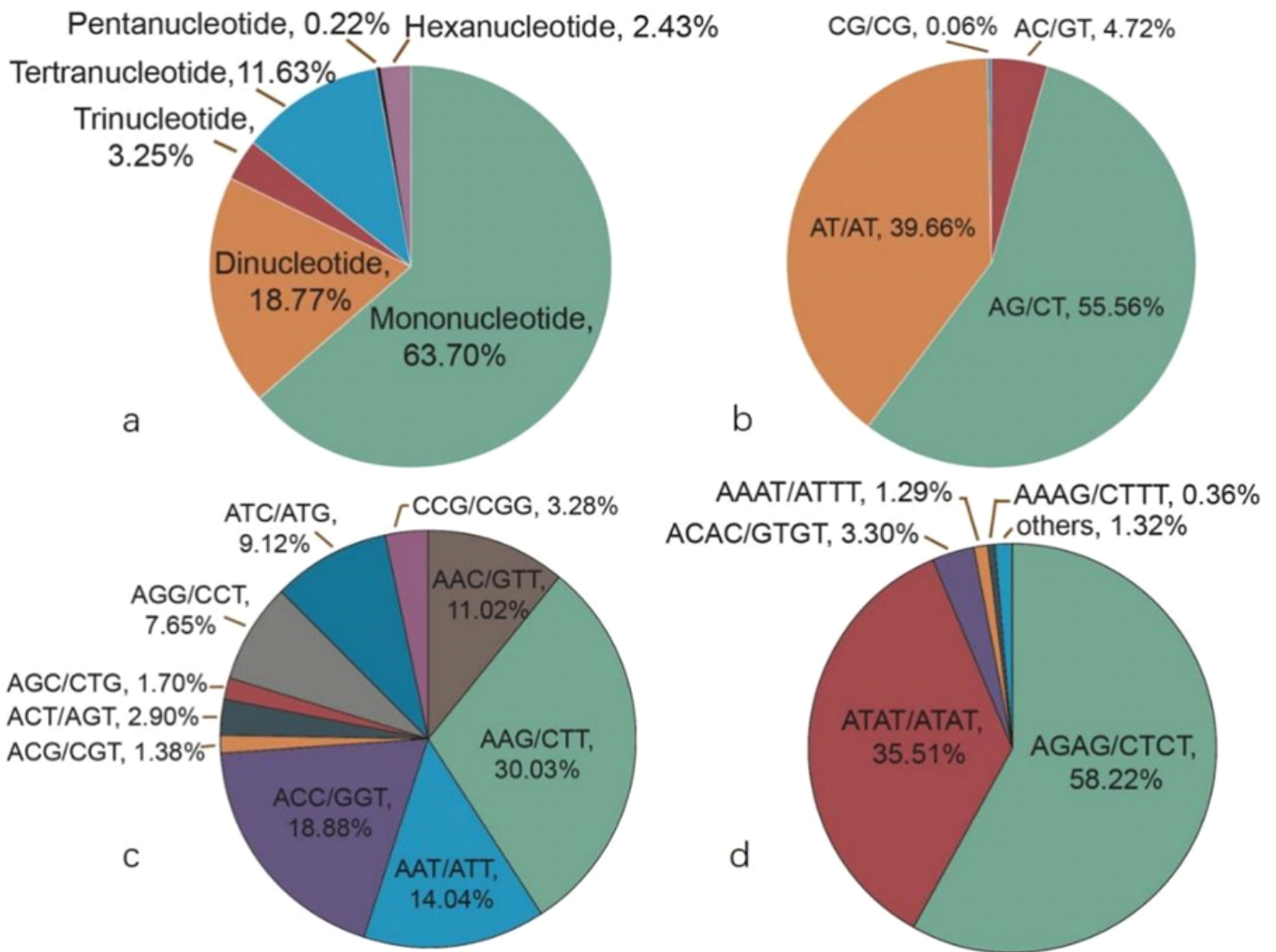

Figure 4

Distribution of SSR motif repeats.

Figure 5

Identification and characteristics of SSR motif. a Percentage of different SSR types. b Percentage of major dinucleotide SSR motifs. c Percentage of major trinucleotide SSR motifs. d Percentage of major pertranucletide SSR motifs
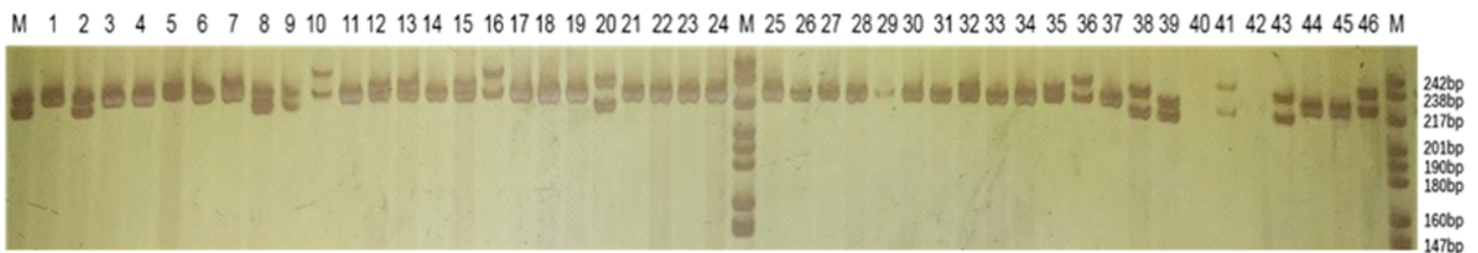

Figure 6

SSR-PCR amplication profile of marker 22 for 45 *T. rupestris* individuals. 1-24 individuals from the first population (POP1) and 25-46 individuals from the second population (POP2), M is the DNA marker pBR322 DNA/MspI.
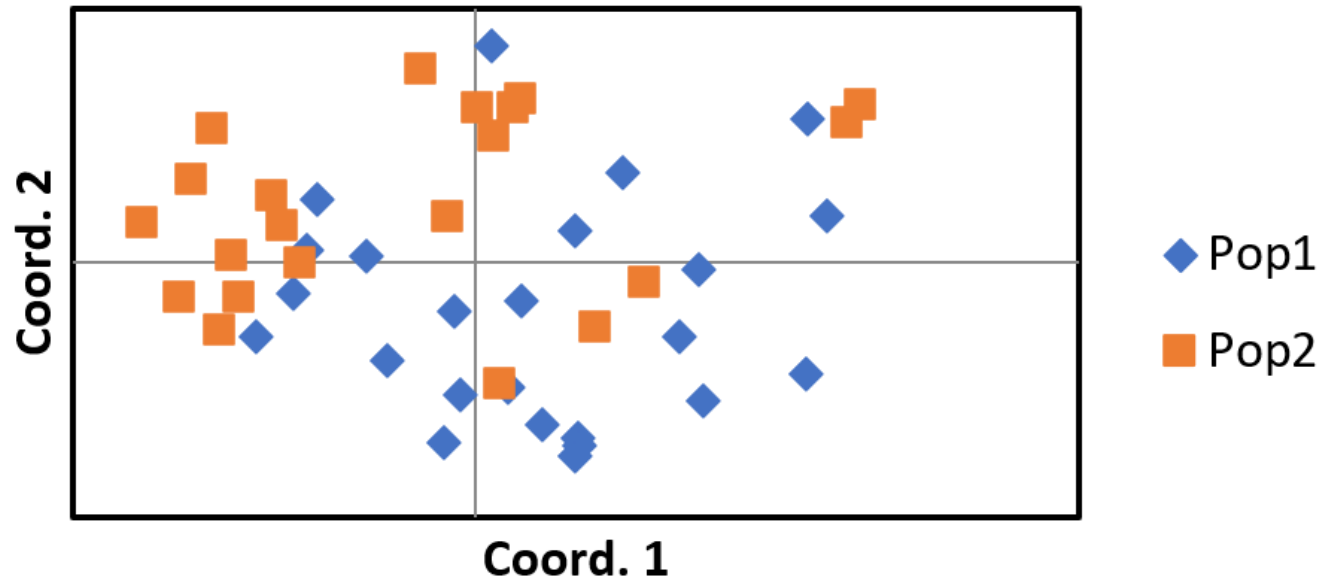
Figure 7

PCoA of genetic distance for 45 *T. rupestris* individuals from populations A (pop 1) and B (pop2)