Complete Genome Sequence of Sphingobacterium sp. Strain CZ-2T Isolated from Tobacco Leaves Infected with Wildfire Disease

Xunhui Cai
Hunan Agricultural University
Ruyi Wang
Hunan Agricultural University
Shengnan Hu
Hunan Agricultural University
Ying Li
Hunan Agricultural University
Tao Chen
Hunan Agricultural University
Jianyun He
Hunan Agricultural University
Siqiao Tan
Hunan Agricultural University
Wei Zhou (mengrzhou@163.com)
Hunan Agricultural University https://orcid.org/0000-0002-5715-6079

Research article

Keywords: Sphingobacterium sp., PacBio and Illumina sequencing, Evolution, Cluster of Orthologous, Gene Ontology, KEGG, Comparative genomics

Posted Date: September 27th, 2019

DOI: https://doi.org/10.21203/rs.2.15235/v1

Abstract

Background Sphingobacterium is a class of Gram-negative, non-fermentative bacilii, and rarely involved in human infections. It is characterized by a large number of cellular membrane sphingophospolipids. Due to its wide ecological distribution and oil degradation ability, environmental microbiologists have paid much attention to it.

Results A novel gram-negative bacterium, designated CZ-2 T , was isolated from a sample of tobacco leaves infected with wildfire disease in Guiyang County, Chenzhou City, Hunan Province, China, and its phylogenetic position was investigated by GC content determination, PCR amplification, sequencing and phylogenetic analysis. Growth occurred on TGY medium at 30 °C and pH 7.0. The GC content of the DNA of strain CZ-2 T is 40.68 mol%. Genome relatedness, rDNA phylogeny and chemotaxonomic characteristics all indicate that strain CZ-2 T represents a novel species of the genus Sphingobacterium . We propose the name Sphingobacterium. tobaci sp. nov., with CZ-2 T as the type strain. Third-generation sequencing (TGS) and next-generation sequencing (NGS) were used to derive a finished genome sequence for strain CZ-2 T , consisting of a circular chromosome 3,925,977 bp in size. The genome of strain CZ-2 T features 3,462 protein-encoding and 50 tRNA-encoding genes. Unigenes were annotated by matching against Clusters of Orthologous Groups of proteins (COG; 2,021 genes), Gene Ontology (GO; 1,952 genes) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (1,380 genes). Comparison of the predicted proteome of CZ-2 T with those of other Sphingobacterium identified 677 species-specific proteins that may contribute to the adaptation of CZ-2 T to its native environment.

Conclusions As the first report of a Sphingobacterium genome sequenced by NGS and TGS, our work will serve as a useful reference for subsequent sequencing and mapping efforts for additional strains and species within this genus.

Introduction

Sphingobacterium is a rod-shaped, non-spore-producing, gram-negative bacterium, and the GC content of its DNA ranges from 35 to 44 mol% (Lee, et al., 2013). The genus *Sphingobacterium* was established by Yabuuchi (Yabuuchi, et al., 1983) in 1983 and comprises bacterial species whose membranes contain high concentrations of shingolipids. At present, the genus *Sphingobacterium* encompasses 41 validly published species names. However, this number continues to increase as novel *Sphingobacterium* strains are isolated from various samples of soil (Teng, et al., 2015), compost (Yoo, et al., 2007), sludge (Zhang, et al., 2012) and even human clinical specimens (Holmes, et al., 1982). *Sphingobacterium* sp. have been reported to cause infection and sepsis in humans (Hibi and Kumano, 2017). *Sphingobacterium multivorum* and *Sphingobacterium thalpophilum* have the ability to degrade petroleum. Between the two, *Sphingobacterium multivorum* SWH-2 has a higher capacity to degrade petroleum.

PacBio and Oxford Nanopore developed PacBio RSII and MinION platforms for long-read sequencing (average: 10-15 kb) (Chaisson, et al., 2014; Schatz and Delcher ALSalzberg, 2010), named Third-

Generation Sequencing (TGS). Compared with Next- Generation Sequencing (NGS), TGS generates long reads (more than 10,000 bp and some read lengths up to 100,000 bp or more)(Lee, et al., 2016), and non-GC biased (Chaisson, et al., 2014; Niu, et al., 2010) data from TGS has been used widely in studies of genome assembly (Gordon, et al., 2016; Jain, et al., 2018; Zheng, et al., 2016) and DNA 6mA methylation (Fang, et al., 2012; Greer, et al., 2015; Wu, et al., 2016; Zhang, et al., 2015). Importantly, longer read lengths span more repetitive elements and thus produce more contiguous reconstructions of the genome(Roberts, et al., 2013). With respect to structural variation analysis, long reads enable improved “split-read” analyses so that insertions, deletions, translocations and other structural changes can be recognized more readily(Chaisson, et al., 2015). Many important research results based on TGS have been published in the journals Cell, Nature, Science and so on. However, genome assembly from TGS data is time consuming due to the high sequencing error rate, whereas NGS data have the characteristics of a short-read length (50-200 bp) and low error rate (1%, mainly substitution)(John, et al., 2009; Langmead, et al., 2009; Li and Durbin, 2009). Therefore, it is necessary to correct TGS sequences by using NGS data.

CZ-2^T (T representing the type strain) is the first strain within the genus *Sphingobacterium* for which TGS and NGS technology were combined to produce a fully assembled and complete genome sequence; our work will serve as a high-quality reference for any future genomic studies of strains from this genus. Furthermore, comparison of the genome sequence of CZ-2^T with those of other *Sphingobacterium* strains allowed the identification of species-unique genes and pathways that may be important mediators of adaptation of this species to its environment.

Material And Methods

Bacterial strain and DNA extraction

The *Sphingobacterium* sp. strain CZ-2 was isolated from a tobacco leaf sample from Chenzhou city in China. The strain sample was dispersed in TGY medium (1.0% tryptone, 0.5% yeast extract and 0.1% glucose), and the culture was incubated for 2 days at 30 °C with shaking at 200 rpm before dilution plating on TGY agar plates (30 °C, 24 h) to isolate single colonies. Genomic DNA was isolated from the cell pellets with a Bacteria DNA Kit (OMEGA) according to the manufacturer’s instructions, and quality control was subsequently carried out on the purified DNA samples. Genomic DNA was quantified by using a TBS-380 fluorometer (Turner BioSystems Inc., Sunnyvale, CA). High-quality DNA (OD260/280 = 1.8~2.0, >6 ug) was used to construct the fragment library.

Illumina HiSeq sequencing

For Illumina pair-end sequencing of each strain, at least 3 μg of genomic DNA was used for sequencing library construction. Paired-end libraries with insert sizes of ~400 bp were prepared following Illumina’s standard genomic DNA library preparation procedure. Purified genomic DNA is sheared into smaller fragments of the desired size by Covaris, and blunt ends are generated by using T4 DNA polymerase. Following addition of an ‘A’ base to the 3' end of the blunt phosphorylated DNA fragments, adapters are

ligated to the ends of the DNA fragments. The desired fragments can be purified by gel electrophoresis then selectively enriched and amplified by PCR. The index tag is introduced into the adapter at the PCR stage, as appropriate, and a library quality test is performed. Finally, the qualified Illumina pair-end library is used for Illumina Hiseq sequencing (PE150 mode).

PacBio sequencing

For Pacific Biosciences sequencing, whole-genome shotgun libraries with 20-kb inserts were generated and sequenced on a Pacific Biosciences RS instrument using standard methods. An 8-μg aliquot of DNA was centrifuged in a Covaris g-TUBE (Covaris, MA) at 6,000 rpm for 60 seconds using an Eppendorf 5424 centrifuge (Eppendorf, NY). DNA fragments were then purified, end-repaired and ligated with SMRTbell sequencing adapters following the manufacturer's recommendations (Pacific Biosciences, CA). Resulting sequencing libraries were purified three times using 0.45 volumes of Agincourt AMPure XPbeads (Beckman Coulter Genomics, MA) following the manufacturer's recommendations.

Genome assembly

Raw sequencing data were generated by using Illumina base calling software CASAVA v1.8.2 (http://support.illumina.com/sequencing/sequencing_software/casava.ilmn) according to its user's guide. Contamination reads, such as those containing adaptors or primers, were identified by Trimmomatic (http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic) with default parameters. Clean data obtained from the above quality control processes were used in further analysis.

The CZ-2^T genome was sequenced using a combination of PacBio RS and Illumina sequencing platforms. The Illumina data were used to evaluate the complexity of the genome and correct the PacBio long reads. Firstly, we used ABySS (http://www.bcgsc.ca/platform/bioinfo/software/abyss) to perform genome assembly with multiple-Kmer parameters and obtained optimal results for the assembly (Jackman, et al., 2017). Secondly, canu (https://github.com/marbl/canu) was used to assemble the PacBio-corrected long reads. Finally, GapCloser software was subsequently applied to fill the remaining local inner gaps and correct the single nucleotide polymorphisms (https://sourceforge.net/projects/soapdenovo2/files/GapCloser/) for the final assembly results (Koren, et al., 2017).

Genome annotation

We used the *ab initio* prediction method to obtain gene models for strain CZ-2. Gene models were identified using Glimmer3 (Delcher, et al., 2007). Then, BLAST sequence alignment was performed with all gene models against the non-redundant (NR in NCBI) database, SwissProt (http://uniprot.org), KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/) (Minoru, et al., 2012) and COG (Clusters of Orthologous Groups of proteins, http://www.ncbi.nlm.nih.gov/COG) (Tatusov, et al., 2000) to carry out functional annotation by using the BLASTP module. In addition, tRNAs were identified using the tRNAscan-SE (v1.23, http://lowelab.ucsc.edu/tRNAscan-SE) (Lowe and Eddy, 1997), and rRNAs

(Lagesen, et al., 2007) were determined using the RNAmmer (v1.2, http://www.cbs.dtu.dk/services/RNAmmer/).

GC content, PCR amplification, sequencing and phylogenetic analysis

Genomic DNA from strain CZ-2^T was prepared using the TIANamp bacterial DNA isolation kit (Tiangen). The GC content of the DNA was determined according to the procedure of Zhou (Zhou, et al., 2012). PCR amplification and sequence analysis of the 16S rRNA gene has been described previously in detail (Zhou, et al., 2012). Phylogenetic dendrograms, which displayed substantially identical topologies, were constructed using the neighbour-joining method (Saitou and Nei, 1987), with bootstrap values calculated from 1,000 re-samplings.

Average nucleotide identity

Bacterial genome sequencing is rapidly emerging as the most important source of information for microbial taxonomy. For example, determination of the whole genome sequence of a newly isolated strain allows the calculation of average nucleotide identity (ANI) scores, providing for global comparisons of the new strain with previously isolated strains whose genome sequences are deposited in databanks. These ANI scores will probably serve as the next-generation gold standard for species delineation (Kim, et al., 2014). ANI was calculated by using the Chun Lab's online Average Nucleotide Identity calculator (Yoon, et al., 2017).

Results

Illumina paired-end sequencing and PacBio sequencing

An Illumina PE library (300-500 bp) and a PacBio library (15~20 kb) were constructed by Illumina Hiseq combined with the TGS technique. A total of 23,648,762 sequence reads were generated using the Illumina Hiseq platform (150 bp, paired-end run), and a total of 26473 sequence reads were generated using the PacBio platform (average length 6674 bp) (Additional file 1: Figure S1). The whole genome of the strain was mapped by bioinformatics analysis after the low-quality nucleotides were trimmed from the sequenced data (Table 1).

Genome properties

The genome is 3,925,977 bp long and comprises one circular chromosome with a 40.68% G+C content (Table 2 and Fig. 1). The genome of the CZ-2^T strain contains 50 tRNA genes (Additional file 2: Table S1)

and several rRNA gene clusters (Additional file 3: Table S2), including duplicate copies of the 16S rRNA gene, triplicate copies of the 23S rRNA gene and four copies of the 5S rRNA gene. An estimated 88.21% of the genome contains coding sequences (CDSs), and these CDSs are predicted to encode 3,462 putative proteins. The genome of CZ-2^T is approximately 3.92 Mb in size, which is much smaller than the estimated genome sizes of other *S. spiritivorum* strains (6.47 Mb in *S. spiritivorum*_ HMA12^T, GenBank ID: NZ_BEYR01000001; 5.33 Mb in *S. spiritivorum*_ML3W^T, GenBank ID: NZ_CP009278; 6.36 Mb in *S. spiritivorum*_B29^T, GenBank ID: NZ_CP019158; 6.23Mb in *S. spiritivorum*_21, GenBank ID: NC_015277). The possible reason is that new genomes assembled using short-read NGS data are often of lower quality in the genomic repeated region, resulting in longer repeated sequence and more repeat copy numbers. The results of NR, SWSS, KEGG, COG, and GO revealed that 3310, 1952, 1380, 2021, and 1952 unigenes were annotated in *S. spiritivorum* CZ-2^T strain, respectively (Additional file 4: Table S3; Additional file 5: Figure S2).

A genome map is a circular representation of one or several genomes that provides a rapid and simple method of identifying global patterns, large features or regions of interest based on various metrics. Visualizing all features at the genomic level can aid in understanding the organization of a genome or the similarities and differences across multiple genomes.

The single genome map presents a pseudogenome comprised of all assembled contigs (Fig 1), including plasmids found in the bacterium. The coloured bands in Section 1 represent contigs, each of which can be clicked on to reveal only the features within that contig. Section 2 represents the annotated reference genes (specifically, CDSs) found on the forward strand, and Section 3 represents the same information on the reverse strand. Different colours were assigned to indicate COG functions (more details). Section 4 displays only rRNA and tRNA found in this genome. Section 5 displays the GC skew metric, which can be used as an indicator for identifying replication loci and leading/lagging strands. The genomic mean GC-skew value is used as the baseline, relative to which higher-than-average values are displayed in red, whereas lower-than-average values are displayed in blue. Finally, Section 6 displays the GC ratio metric, which can be used to profile the genome, identify isochores or observe co-variations with other data. The GC ratio also uses the genomic mean GC ratio value as its baseline, with higher-than-average values in pink and lower-than-average values in sky blue.

Fig. 1 Genome map of *Sphingobacterium* sp. CZ-2^T.

Rings from the outermost to the centre: (1) scale marks; (2) protein-coding genes on the forward strand; (3) protein-coding genes on the reverse strand (color-coded by functional category); (4) rRNA (red) and tRNA genes (purple); (5) GC content; and (6) GC skew. Protein-coding genes are colour coded according to their COG categories.

Evolution of CZ-2^T

The 16S rRNA gene sequence of strain CZ-2^T is 1,432 bp in length. BLAST searches in the GenBank database and the EzTaxon server (Kim, et al., 2012) (http://www.ezbiocloud.net/eztaxon) indicated that strain CZ-2^T belongs to the genus *Sphingobacterium* of the phylum Bacteroidetes. The 16S rRNA gene of strain CZ-2^T exhibits the highest similarity to sequences from *Sphingobacterium lactis* DSM 22361^T (96.90%), *Sphingobacterium kyonggiense* KEMC 2241-005^T (96.80%), *Sphingobacterium soli* YIM X0211^T (96.73%), *Sphingobacterium cellulitidis* R-53603^T (96.62%), *Sphingobacterium daejeonense* TR6-04^T (96.55%), *Flavobacterium mizutaii* NCTC 12149^T (95.99%), *Sphingobacterium hotanense* XH4^T (95.97%) and *Sphingobacterium humi* D1^T (95.64%). Phylogenetic analysis confirmed that strain CZ-2^T forms a coherent cluster with members of the genus *Sphingobacterium*, and an intra-genus clade with *S. lactis* DSM 22361^T, *S. daejeonense* TR6-04^T and *S. kyonggiense* KEMC 2241-005^T (Fig. 2). The average nucleotide identity (ANI) of the genome sequence of strain CZ-2^T against the four other *Sphingobacterium* species for which genome sequences are publicly available ranged from 68.87% (with strain *S. spiritivorum* _21) to 70.76% (with strain *S. spiritivorum*_ML3W^T). These ANI values are also considerably lower than the 95% to 96% threshold used to identify isolates as belonging to the same bacterial species (Goris, et al., 2007; Richter and Rossello-Mora, 2009). Thus, rDNA phylogeny, genome relatedness and chemotaxonomic characteristics all indicate that strain CZ-2^T represents a novel species within the genus *Sphingobacterium*. We propose the name *S. tobaci* sp. nov. (isolated from tobacco), with CZ-2^T (In Chenzhou city) as the type strain.

Fig. 2 Neighbour-joining phylogenetic analysis based on 16S rRNA gene sequences from the EZ BioCloud database (accession numbers are given in parentheses) and depicting the position of strain CZ-2 among members of the genus *Sphingobacterium*.

Multiple alignment, distance calculations (according to the Kimura 2-parameter model) and clustering were performed using the MEGA-7 software package. Bootstrap values (≥50) based on 1000 resamplings are shown at the nodes. Filled circles indicate that the corresponding nodes were also recovered in trees reconstructed with maximum parsimony and minimum evolution algorithms. Bar, 0.005 nucleotide substitutions per nucleotide position.

Annotation

The Clusters of Orthologous Groups (COG) database is an effective tool for the annotation of functional proteins (Galperin, et al., 2015). In this investigation, COG categorization according to the HMM profiles of Gammaproteobacteria was found in the popular EggNOG database (Huertacepas, et al., 2016; Sean, et al., 2012) with an E-value cut-off of 1e-5.

To determine the functional proteins encoded by 3,462 genes, COG results were examined, revealing 2,021 unigenes that were annotated in the CZ-2^T strain of *S. spiritivorum*. COG assigned functional proteins to 21 classifications, which are summarized in Table S4 (Additional file 6). Among these annotated unigenes, 45.98% were related to metabolism, 18.32% to cellular processes and signalling and 19.27% to information storage and processing. However, 16.43% of the unigenes were poorly classified by COG category because their features and functions remain obscure. The classification of COG protein-encoding genes found in the genome of *S. spiritivorum* CZ-2^T is plotted in Fig. 3.

Fig. 3 COG functional categories

KEGG pathway annotation was helpful for assigning biological functions to genes via interpretation of enzymes and other proteins in biochemical processes (Kanehisa and Goto, 2000). Metabolic pathway annotation was carried out using predicted protein sequences and the KEGG Automatic Annotation Server (KAAS) (Yuki, et al., 2007). In the present study, a total of 1,380 unigene sequences were assigned to 173 pathways of KEGG annotated protein sequences of *S. spiritivorum* CZ-2^T (Additional file 7: Table S5). Most (455) of the unigenes were assigned to “metabolic pathways”, 211 to “biosynthesis of secondary metabolites” and 111 to “microbial metabolism in diverse environments”, forming the dominant categories. The top 20 representative pathways from KEGG pathway annotation are depicted in Fig. 4.

Fig. 4 Number of genes associated with the KEGG pathways (top 20).

KEGG pathway is a collection of annually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development

In the present study, 1,952 unigenes were annotated using the GO database (Table 3; Additional file 8: Figure S3). The GO-annotated unigenes were assigned to the biological process (573 unigenes), cellular component (259 unigenes) and molecular function (344 unigenes) categories. Table 3 represents the distribution of unigenes regarding the number of GO terms. For the biological process category, the predominant GO terms included “cellular process” (144 unigenes), “single-organism process” (99 unigenes) and “metabolic process” (181 unigenes). In the cellular component category, “cell part” (62 unigenes), “membrane” (53 unigenes) and “cell” (62 unigenes) were the predominant GO terms for the classification of unigenes. Meanwhile, the unigenes were assigned to nine different functions in the “molecular function process” category, the predominant one being “binding” (130 unigenes) and “catalytic activity” (121 unigenes).

Discussion

Evolution and comparative genomics

We compared the predicted proteome of strain CZ-2^T with those of four other *Sphingobacterium* strains for which genome sequence data are available at NCBI (*S. spiritivorum*_ML3W^T, *S. spiritivorum*_B29^T, *S. spiritivorum* _21 and *S. spiritivorum*_ HMA12^T) (Fig. 5). These four sequenced *Sphingobacterium* strains share 1,706 orthologous protein groups. These common orthologous protein groups encompass the enzymes for central carbon metabolism and include the pentose phosphate pathway, the tricarboxylic acid cycle (TCA), amino acid biosynthesis and the assembly of purine and pyrimidine nucleotides. Genes for this set of predicted pathways are well conserved in the genomes of the sequenced *Sphingobacterium* strains. Orthologous protein groups shared by CZ-2^T with only one of the other *Sphingobacterium* strains were also identified and contained 97 (*S. spiritivorum*_ML3W^T), 626 (*S. spiritivorum*_B29^T), 95 (*S. spiritivorum*_21) and 62 (*S. spiritivorum*_ HMA12^T) groups, respectively. Importantly, we detected 677 species-unique predicted proteins belonging to 677 orthologous groups in strain CZ-2^T. Many of these CZ-2^T-specific genes are involved in transport systems, DNA repair and the biosynthesis of small molecules. These and other species-unique proteins may facilitate the adaptation of *S. spiritivorum* to tobacco leaf environments.

Fig. 5 Venn diagram depicting orthologous groups of predicted proteins encoded in five *Sphingobacterium* genomes

Sphingobacterium strains are presumably well adapted to living in diverse environments, and transport systems are key components of bacterial tolerance to extreme conditions (Konings, et al., 2002). For example, most of the known transport systems of extremophiles are sugar uptake systems which belong to the ABC family of transporters (Konings, et al., 2002). To discover the mechanisms of *Sphingobacterium* strain adaptation to various environments, we compared the transport system of strain CZ-2 (isolated in 2018 from tobacco leaves) with those of four other *Sphingobacterium* strains (Strain HMA12^T was isolated in 2013 from flatland farm soil; strain ML3W^T was isolated in 2012 from *Myotis lucifugus* wing; strain B29^T was isolated in 2015 from rhizospheres). KEGG mapping showed that CZ-2^T has 26 genes coding for transporters, mainly related to phosphate-specific transport (PstS protein is a phosphate-binding protein, PstA and PstC proteins are considered to form the transmembrane channel of the Pst system, ATP binding PstB protein probably interacts with PstA-PstC at the cytoplasmic side of the membrane), lipopolysaccharide (LptF and LptG, together with LptB, functions to extract lipopolysaccharide from the inner membrane en route to the outer membrane), lipoprotein (lipoprotein-releasing system permease protein LolC and LolE), iron complex (a periplasmic iron-binding protein (FhuD) and cytoplasmic membrane proteins (FhuB and FhuC)), heme (heme trafficking system

membrane protein CcmB and CcmC), molyblate (molybdenum transport ATP-binding protein ModF) and the ABCB subfamily (ATP-binding lipopolysaccharide transport protein MsbA). Importantly, the *PstS*, *PstC*, *PstA*, *LolC*, *LolE*, *CcmC*, *CcmB*, *LptF*, *LptB*, *FtsX*, *FhuD*, *FhuB*, *FhuC* and *MsbA* genes encode transporters that are present in all five *Sphingobacterium* strains (Table 4). We consider these transporters essential for the survival of *Sphingobacterium* strains. The *MalK*, *MsmX*, *MsmK*, *SmoK*, *AglK*, *MsiK*, *RbsB*, *RbsA*, *MglA*, *latA*, *BtuF* and *ZnuC* genes encode transporters that are present in HMA12^T and B29^T but absent in CZ-2^T, 21 and ML3W^T. We hypothesize that these transporters may contribute to the adaptation of strain HMA12^T and B29^T to its soil environment. Alternatively, these transporters may help strains HMA12^T and B29^T to import a wider range of nutrients from soil organic matter.

Betaine is widely distributed in nature and has been found in many microorganisms, such as bacteria, archaea and fungi. The main function of betaine is to protect microorganisms from drought, osmotic stress and temperature stress. Meanwhile, betaine plays an important role in methyl group metabolism. Strains 21, ML3WT^T, HMA12^T and B29T^T feature several genes whose protein products are predicted to play a role in transport of the osmoprotectants glycine and betaine and may thus contribute resistance to environmental stresses.

Conclusions

ANI, rDNA phylogeny and genome sequencing were used to identify bacterial strain CZ-2^T, isolated from tobacco leaves infected with tobacco wildfire diseases in Guiyang County, Hunan Province, China, as a novel species within the genus *Sphingobacterium*. Therefore, we propose CZ-2 as the type strain of *Sphingobacterium tobaci* sp. CZ-2^T nov. Whole-genome sequencing (NGS and TGS) was used to derive the high quality, finished genome sequence assembly for this gram-negative bacterial strain within the genus *Sphingobacterium*. Genome-wide comparisons with other sequenced *Sphingobacterium spp. provided an extensive list of core genes, which are highly conserved to contribute to the adaptation of Sphingobacterium to its native environment; we also found a list of species-unique genes, some of which are proposed to contribute to adaptation to the soil environment.*

Declarations

Acknowledgements

The authors thank other members of the laboratory at Department of Bioinformatics, College of Plant Protection, Hunan Agricultural University, Changsha, China for their help during the manuscript preparation

Funding information

This research was supported by China Postdoctoral Science Foundation (No.2015T80870 and No.2014M562109), China Scholarship Council (No.201708430002) and Scientific Research Fund of

Hunan Provincial Education Department (NO.17C0770).

Availability of data and materials

All the Sphingobacterium sp. CZ-2^T sequencing data obtained in this work are available in the National Center for Biotechnology Information (NCBI) under the GenBank ID CP038159. Furthermore, all genome annotated in this study are provided in additional files.

Author's contributions

CXH, WRY, HSN, LY and CT compiled the data, performed analysis, and wrote first draft of the MS, TSQ and HJY performed comparative genomics analysis, ZW planned the study, drew the conclusions and contributed to the writing of the manuscript. All authors have read and approved the manuscript.

Ethics approval and consent to participate

Not application.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature. 517(7536):608.

Chaisson MJP, John H, Dennis MY, Sudmant PH, Maika M, Fereydoun H, Francesca A, Urvashi S, Richard S, Matthew B (2014) Resolving the complexity of the human genome using single-molecule sequencing. Nature. 517(7536):608-611.

Delcher A, Bratke K, Powers E, Salzberg S (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 23(6):673-679.

Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE (2012) Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nature biotechnology. 30(12):1232-1239.

Galperin MY, Makarova KS, Wolf YI, Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Research. 43(Database issue):261-269.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, Dunn C, Baker C, Armstrong J, Diekhans M, Paten B, Shendure J, Wilson RK, Haussler D, Chin CS, Eichler EE (2016) Long-read sequence assembly of the gorilla genome. Science. 352(6281):aae0344.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal Systematic Evolutionary Microbiology. 57(Pt 1):81-91.

Greer EL, Blanco MA, Gu L, Sendinc E, Liu J, Aristizabal-Corrales D, Hsu CH, Aravind L, He C, Shi Y (2015) DNA Methylation on N6-Adenine in C. elegans. Cell. 161(4):868-878.

Hibi A and Kumano Y (2017) Sphingobacterium spiritivorum bacteremia due to cellulitis in an elderly man with chronic obstructive pulmonary disease and congestive heart failure: a case report. Journal of Medical Case Reports. 11(1):277.

Holmes B, Owen RJ, Hollis DG (1982) Flavobacterium spiritivorum, a New Species Isolated from Human Clinical Specimens. International Journal of Systematic Bacteriology. 32(32):157-165.

Huertacepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research. 44(D1):D286-D293.

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL (2017) ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Research. 27(5):768-777.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature biotechnology. 36(4):338-345.

John E, Adrian F, Jeremy G, Khai L, John L, Geoff O, Paul P, David R, Primo B, Brad B (2009) Real-time DNA sequencing from single polymerase molecules. Science. 323(5910):133-138.

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopaedia of Genes and Genomes. Nucleic Acids Research. volume 28(1):27-30(24).

Kim M, Oh HS, Park SC, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International Journal Systematic Evolutionary Microbiology. 64(Pt 2):346-351.

Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, Na H, Park SC, Jeon YS, Lee JH, Yi H (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. International Journal Systematic Evolutionary Microbiology. 62(Pt 3):716-721.

Konings WN, Albers SV, Koning S, Driessen AJM (2002) The cell membrane plays a crucial role in survival of bacteria and archaea in extreme environments. Antonie van Leeuwenhoek. 81(1-4):61-72.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. 27(5):722.

Lagesen K, Hallin P, Rodland E, Staerfeldt H, Rognes T, Ussery D (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research. 35(9):3100.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 10(3):R25.

Lee DH, Hur JS, Kahng HY (2013) Sphingobacterium cladoniae sp. nov., isolated from lichen, Cladonia sp., and emended description of Sphingobacterium siyangense. International Journal of Systematic & Evolutionary Microbiology. 63(2):755-760.

Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S, Goodwin S, McCombie WR, Schatz M (2016) Third-generation sequencing and the future of genomics. BioRxiv.048603.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754-1760.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research. 25(5):955-964.

Minoru K, Susumu G, Yoko S, Miho F, Mao T (2012) KEGG for integration and interpretation of large-scale molecular data sets.

Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. Bmc Bioinformatics. 11(1):1-11.

Richter M, Rossello-Mora R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences of the United States of America. 106(45):19126-19131.

Roberts RJ, Carneiro MO, Schatz MC (2013) The advantages of SMRT sequencing. Genome biology. 14(6):405.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution. 4(4):406-425.

Schatz MC, Delcher ALSalzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Research. 20(9):1165-1173.

Sean P, Damian S, Kalliopi T, Alexander R, Michael K, Jean M, Roland A, Thomas R, Ivica L, Tobias D (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Research. 40(D1):D284-D289.

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research. 28(1):33-36.

Teng C, Zhou Z, Molnar I, Li X, Tang R, Chen M, Wang L, Su S, Zhang W, Lin M (2015) Whole-genome optical mapping and finished genome sequence of Sphingobacterium deserti sp. nov., a new species isolated from the Western Desert of China. PLoS One. 10(4):e0122254.

Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, Liu Y, Byrum SD, Mackintosh SG, Zhong M, Tackett A, Wang G, Hon LS, Fang G, Swenberg JA, Xiao AZ (2016) DNA methylation on N(6)-adenine in mammalian embryonic stem cells. Nature. 532(7599):329-333.

Yabuuchi E, Kaneko T, Yano I, Moss CW, Miyoshi N (1983) Sphingobacterium gen. nov., Sphingobacterium spiritivorum comb. nov., Sphingobacterium multivorum comb. nov., Sphingobacterium mizutae sp. nov., and Flavobacterium indologenes sp. nov.: Glucose-Nonfermenting Gram-Negative Rods in CDC Groups IIK-2 and IIb. International Journal of Systematic Bacteriology. 33(3):580–598.

Yoo SH, Weon HY, Jang HB, Kim BY, Kwon SW, Go SJ, Stackebrandt E (2007) Sphingobacterium composti sp. nov., isolated from cotton-waste composts. Int J Syst Evol Microbiol. 57(7):1590-1593.

Yoon SH, Ha SM, Lim J, Kwon S, Chun J (2017) A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie van Leeuwenhoek. 110(10):1281-1286.

Yuki M, Masumi I, Shujiro O, Yoshizawa AC, Minoru K (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research. 35(suppl_2):W182-W185.

Zhang G, Huang H, Liu D, Cheng Y, Liu X, Zhang W, Yin R, Zhang D, Zhang P, Liu J, Li C, Liu B, Luo Y, Zhu Y, Zhang N, He S, He C, Wang H, Chen D (2015) N6-methyladenine DNA modification in Drosophila. Cell. 161(4):893-906.

Zhang J, Zheng JW, Cho BC, Hwang CY, Fang C, He J, Li SP (2012) Sphingobacterium wenxiniae sp. nov., a cypermethrin-degrading species from activated sludge. International Journal of Systematic & Evolutionary Microbiology. 62(3):683-687.

Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA, Wyatt PW, Bharadwaj R, Makarewicz AJ, Li Y, Belgrader P, Price AD, Lowe AJ, Marks P, Vurens GM, Hardenbol P, Montesclaros L, Luo M, Greenfield L, Wong A, Birch DE, Short SW, Bjornson KP, Patel P, Hopmans ES, Wood C, Kaur S, Lockwood GK, Stafford D, Delaney JP,

Wu I, Ordonez HS, Grimes SM, Greer S, Lee JY, Belhocine K, Giorda KM, Heaton WH, McDermott GP, Bent ZW, Meschi F, Kondov NO, Wilson R, Bernate JA, Gauby S, Kindwall A, Bermejo C, Fehr AN, Chan A, Saxonov S, Ness KD, Hindson BJ, Ji HP (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature biotechnology. 34(3):303-311.

Zhou Z, Tang R, Chen M, Lin M and Zhang W (2012) Corynebacterium deserti sp. nov., isolated from desert sand. International Journal of Systematic & Evolutionary Microbiology. 62(Pt 4):791.

Tables

Table 1. Statistics of sequencing data

Illumina		*PacBio*	
Features	*Statistics of raw data*	*Features*	*Statistics of raw data*
Total reads num	*23,648,762*	*Total reads num*	*26,473*
Total bases (bp)	*3,547,314,300*	*Total bases (bp)*	*176,693,851*
Q20 (%)	*95.95*		
Largest (bp)			*41,645*
Average coverage	*903*		*45*
Average length (bp)	*150*		*6,674*

Table 2. Genome statistics results

Attribute		*Values*	*% of Total*
Genome size (bp)		*3,925,977*	*100%*
	DNA coding region (bp)	*3,463,143*	*88.21%*
	DNA GC content (bp)	*1,597,100*	*40.68%*
	GC content in gene region (bp)	*1,441,125*	*41.61%*
	Gene average length	*1,000*	
	Gene density	*0.881*	
Total genes		*3,462*	*100%*
	tRNA	*50*	*1.44%*
	rRNA	*9*	*0.26%*
	Genes assigned to NR	*3310*	*95.61%*
	Genes assigned to SWSS	*1952*	*56.38%*
	Genes assigned to KEGG	*1380*	*39.86%*
	Genes assigned to COG	*2,021*	*58.38%*
	Genes assigned to GO	*1,952*	*56.38%*

Table 3. Number of genes associated with each of the general GO functional categories

Term type	*term*	*number*	*percent*	*GO*
Molecular function	*electron carrier activity*	*1*	*0.0039*	*GO:0009055*
	protein binding transcription factor activity	*4*	*0.0156*	*GO:0000988*
	binding	*130*	*0.5058*	*GO:0005488*
	transporter activity	*23*	*0.0895*	*GO:0005215*
	catalytic activity	*121*	*0.4708*	*GO:0003824*
	molecular transducer activity	*19*	*0.0739*	*GO:0060089*
	structural molecule activity	*22*	*0.0856*	*GO:0005198*
	nucleic acid binding transcription factor activity	*18*	*0.0700*	*GO:0001071*
	antioxidant activity	*6*	*0.0233*	*GO:0016209*
Cellular component	*organelle*	*23*	*0.0895*	*GO:0043226*
	cell part	*62*	*0.2412*	*GO:0044464*
	membrane part	*24*	*0.0934*	*GO:0044425*
	nucleoid	*1*	*0.0039*	*GO:0009295*
	membrane	*53*	*0.2062*	*GO:0016020*
	macromolecular complex	*27*	*0.1051*	*GO:0032991*
	cell	*62*	*0.2412*	*GO:0005623*
	organelle part	*7*	*0.0272*	*GO:0044422*
Biological process	*regulation of biological process*	*34*	*0.1323*	*GO:0050789*
	cellular component organization or biogenesis	*9*	*0.0350*	*GO:0071840*
	developmental process	*3*	*0.0117*	*GO:0032502*
	cellular process	*144*	*0.5603*	*GO:0009987*
	response to stimulus	*15*	*0.0584*	*GO:0050896*
	single-organism process	*99*	*0.3852*	*GO:0044699*
	signaling	*5*	*0.0195*	*GO:0023052*
	localization	*47*	*0.1829*	*GO:0051179*
	metabolic process	*181*	*0.7043*	*GO:0008152*
	biological regulation	*36*	*0.1401*	*GO:0065007*

Molecular function: the elemental activities of a gene product at the molecular level, such as binding or catalysis; Cellular component: the parts of a cell or its extracellular environment; Biological process: operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs and organisms.

Table 4. Transport system genes of strain CZ-2^T with other four Sphingobacterium strains

	gene	CZ-2^T	21	HMA12^T	ML3WT^T	B29T^T
Phosphate	**PstS**	+	+	+	+	+
	PstC PstA	+	+	+	+	+
		+	+	+	+	+
	PstB	+	+	+	+	-
Lipoprotein	**LolC** LolE	+	+	+	+	+
		+	+	+	+	+
	LilD	+	+	-	-	-
Heme	CcmD	-	-	-	-	-
	CcmC CcmB	+	+	+	+	+
		+	+	+	+	+
	CcmA	-	-	+	-	-
Lipopolysaccharide	**LptF** LptG	+	+	+	+	+
		+	+	-	+	+
	LptB	+	+	+	+	+
Cellar division involvment	**FtsX**	+	+	+	+	+
	FtsE	+	+	-	+	-
Iron complex	**FhuD**	+	+	+	+	+
	FhuB	+	+	+	+	+
	FhuC	+	+	+	+	+
Molyblate	ModA	-	-	+	+	+
	ModB	-	-	-	-	-
	ModC ModF	-	-	-	-	-
		+	+	+	+	-
ABCB Subfamily	**MsbA**	+	+	+	+	+

1. S. spiritivorum HMA12^T: flatland farm soil; S. spiritivorum ML3W^T: wing (Host: Myotis lucifugus); S. spiritivorum B29^T: Rhizosphere (Belgium: Beverlo); +: predicted; -: missed.

Figures

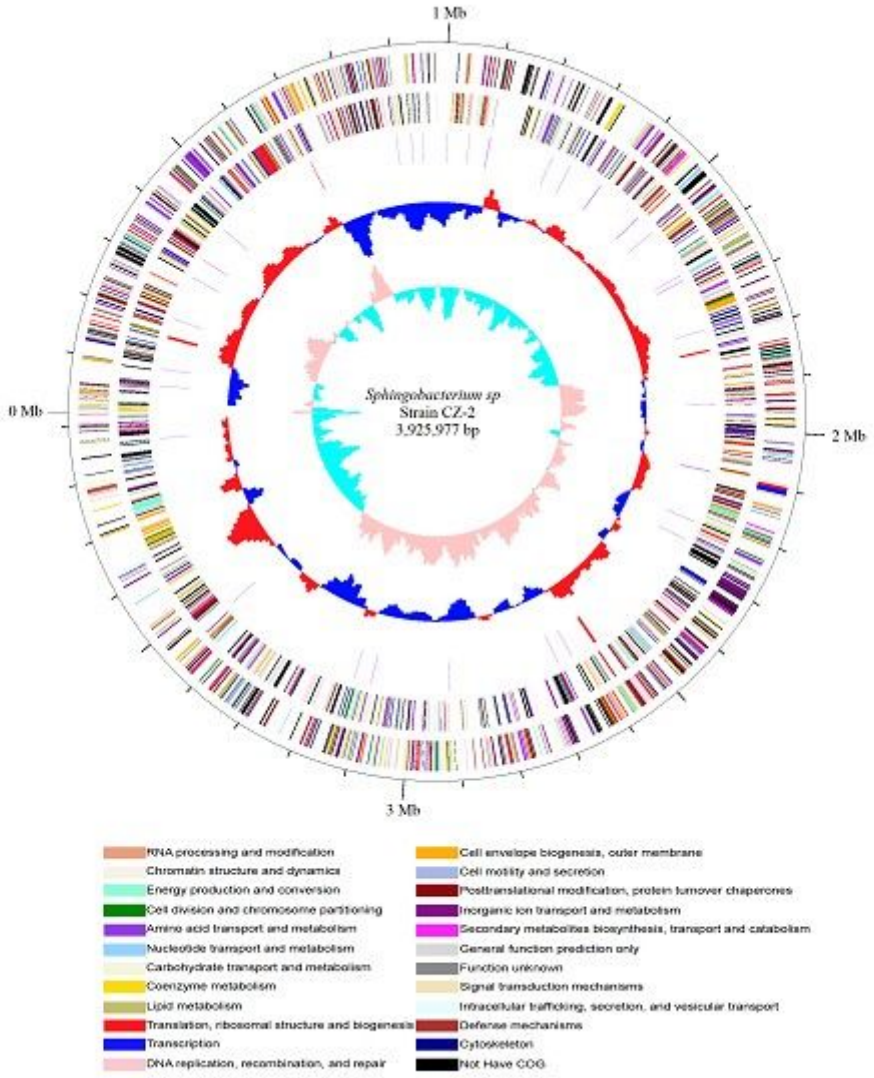

Figure 1

Genome map of Sphingobacterium sp. CZ-2T. Rings from the outermost to the centre: (1) scale marks; (2) protein-coding genes on the forward strand; (3) protein-coding genes on the reverse strand (color-coded by functional category); (4) rRNA (red) and tRNA genes (purple); (5) GC content; and (6) GC skew. Protein-coding genes are colour coded according to their COG categories.

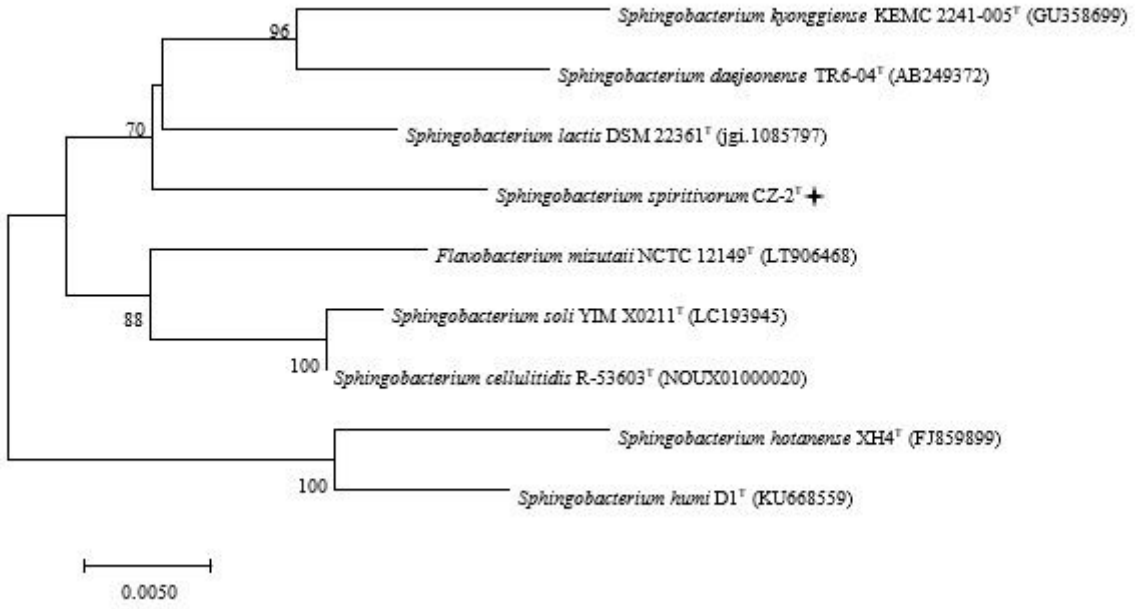

Figure 2

Neighbour-joining phylogenetic analysis based on 16S rRNA gene sequences from the EZ BioCloud database (accession numbers are given in parentheses) and depicting the position of strain CZ-2 among members of the genus Sphingobacterium. Multiple alignment, distance calculations (according to the Kimura 2-parameter model) and clustering were performed using the MEGA-7 software package. Bootstrap values (≥50) based on 1000 resamplings are shown at the nodes. Filled circles indicate that the corresponding nodes were also recovered in trees reconstructed with maximum parsimony and minimum evolution algorithms. Bar, 0.005 nucleotide substitutions per nucleotide position.

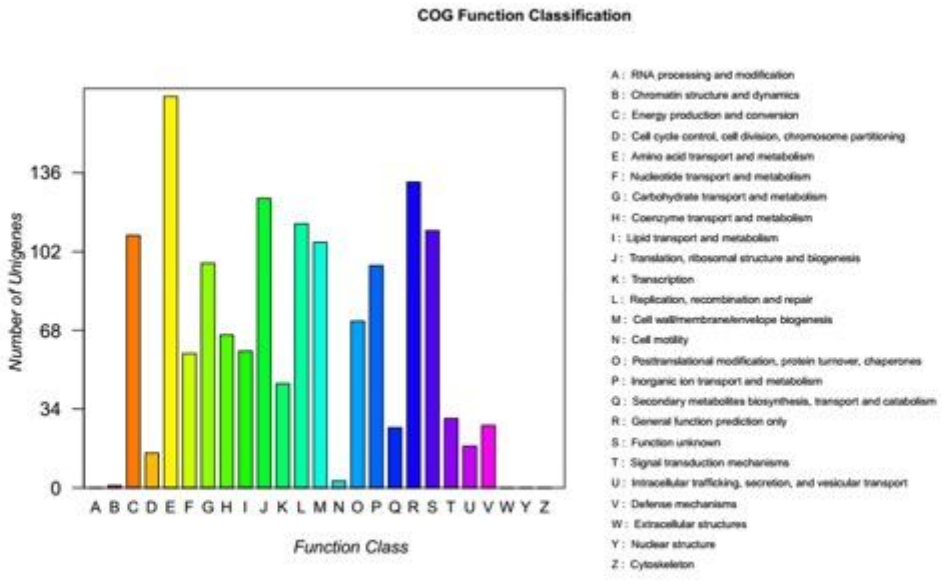

Fig. 3 COG functional categories

Figure 3

COG functional categories

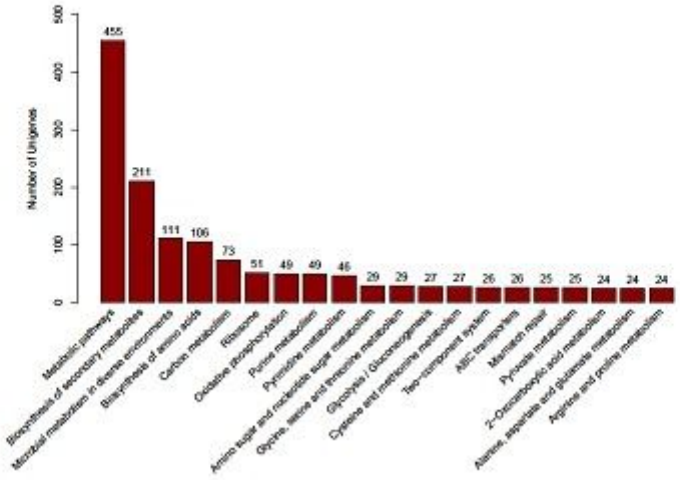

Figure 4

Number of genes associated with the KEGG pathways (top 20). KEGG pathway is a collection of annually drawn pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development

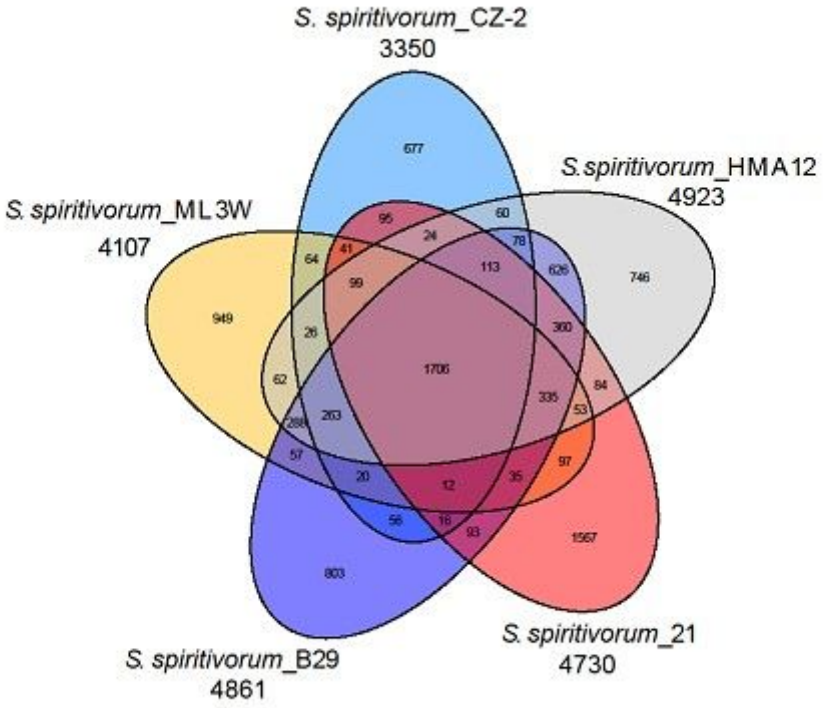

Figure 5

Venn diagram depicting orthologous groups of predicted proteins encoded in five Sphingobacterium genomes

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- *TableS5.KEGGpathway.xlsx*
- *FigureS1.ReadslengthdistributionofPacBiodatainCZ2T.pdf*
- *TableS2.rRNAannotatedinCZ2Tstraingenome.xlsx*
- *ChecklistforsubmissionGENG.docx*
- *renamed47305.xlsx*
- *TableS1.tRNAannotatedinCZ2Tstraingenome.xlsx*
- *FigureS3.NRGOCOGKEGGSWSS.venn.pdf*
- *TableS4.COGannotationandsummary.xlsx*
- *FigureS2.GOfunctionalcategories.pdf*