A peer-reviewed version of this preprint was published in PeerJ on 14 January 2019.

Barajas HR, Romero MF, Martínez-Sánchez S, Alcaraz LD. 2019. Global genomic similarity and core genome sequence diversity of the *Streptococcus* genus as a toolkit to identify closely related bacterial species in complex environments. PeerJ 6:e6233 https://doi.org/10.7717/peerj.6233

Global genomic similarity and core genome sequence diversity of the *Streptococcus* genus as a toolkit to identify closely related bacterial species in complex environments.

Hugo R. Barajas[1], Miguel Romero[1], Shamayim Martínez-Sánchez[1], and Luis D. Alcaraz[1]

[1] Departamento de Biología Celular, Facultad de Ciencias, Universidad Nacional Autónoma de México (UNAM). Mexico city, Mexico.

Corresponding Author:

Luis D. Alcaraz[1]

Laboratorio de Biología Molecular y Genómica. Departamento de Biología Celular, Facultad de Ciencias e Instituto de Ecología. Universidad Nacional Autónoma de México (UNAM), 04510 Mexico City, Mexico.

Email address: lalcaraz@ciencias.unam.mx

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

Abstract

Background. Comparative genomics between closely related bacterial strains can distinguish important features determining pathogenesis, antibiotic resistance, and phylogenetic structure. The *Streptococcus* genus is relevant to public health and food safety and it is well-represented (>100 genomes) in databases of publicly available databases. Streptococci are cosmopolitan, with multiple sources of isolation, from humans to dairy products. The *Streptococcus* genus has been classified by morphology, serotypes, 16S rRNA gene, and Multi Locus Sequence Types (MLST). The Genomic Similarity Score (GSS) is proposed as a tool to quantify genome level relatedness between species of *Streptococcus*. The *Streptococcus* core genome can be used to assess strain specific abundances in metagenomic sequences.

Methods. A 16S rRNA gene phylogeny was calculated for 108 strains, belonging to 16 *Streptococcus* species and compared to a dendrogram using GSS pairwise distances for the same genomes. The core and pan-genome were calculated for these 108 genomes. The core genome sequences were analyzed and used as a resource to discriminate homologous fragment reads from closely related strains in metagenomic samples.

Results. A total of 404 proteins are shared by all 108 *Streptococcus* genomes, which is the core genome. The pairwise amino acid identity values of the core proteins for all the compared strains and outgroups are reported. Lower sequence identity variation (90-100%) is predominantly found in core clusters containing ribosomal and translation-related proteins. For 48 core proteins (11.8%) no functional assignment could be made and those proteins have larger sequence identity variations than other core proteins. The sequence identity of the core genome diminishes as GSS score between species decreases. The GSS dendrogram recovers most of the clades in the 16S rRNA gene phylogeny while distinguishing between 16S polytomies (unresolved nodes). Finally, the core genome was used to distinguish between closely related species within human oral

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

metagenomes.

Discussion. The *Streptococcus* genus provides a benchmark dataset for comparative genomic studies due to the breath depth of genomic coverage. Comparing metagenomic shotgun fragment reads to the core genome using rapid alignment tools allows species-specific abundance estimates in metagenomic samples. Understanding of genomic variability and strains relatedness is the goal of tools like GSS, which make use of both pairwise shared core and pan-genomic homologous shared sequences for its calculation.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

Background

Streptococcus is a bacterial genus that encompasses more than 40 different species, from a diverse range of human and animal pathogens like the etiological agents for caries and meningitis, to commensal species inhabiting animal guts and respiratory tracts (Kilian, 2012). Classification within the genus has been done by morphology, biochemical profiles, serum types, and more recently using the comparison of 16S ribosomal RNA (rRNA) gene phylogenies (Kawamura et al., 1995). There are also Multi Locus Sequence Types (MLST) for eight streptococci species (Kawamura et al., 1995). The Streptococci are divided into six main paraphyletic groups, because of clinical or practical ease, named: *pyogenes, mitis, anginosus, salivarius, bovis,* and *mutans* according to the representative species for each clade (Kilian et al., 2008). There are multiple genome sequences available for the Streptococci; most of the species used in this work were isolated from humans, bovine, swine, and dairy product samples (Supplemental Information 1).

Bacterial phylogenetics has been done using multiple criteria to define bacteria species. The current standards are based either on genome wide Average Nucleotide Identity (ANI) above 95% for estimating an overall genome related index (OGRI) (Konstantinidis & Tiedje, 2005; Konstantinidis, Ramette & Tiedje, 2006; Chun et al., 2018) or on 16S rRNA gene sequence comparison with a 97% identity or above the threshold to identify a bacterium species (Stackebrandt & Goebel, 1994). Protein translation is universal to cellular life, and thus the conservation of the molecular-associated machinery has been used as a molecular taxonomic marker due to its high conservation across the tree of life, including the 16S rRNA gene. However, 16S rRNA has a slow evolutionary rate which does not allow enough resolution to distinguish between closely related species (Stackebrandt & Goebel, 1994). The use of multiple coding genes alignments known as multi locus sequence typing (MLST) is standard practice for distinguishing between strains of pathogenic bacteria. Even what should define a bacterial species based on molecular phylogenetics is fuzzy (Fraser *et a*l., 2009).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

The astounding amount of available bacterial genomes (77,107 in GenBank, February 2018); (Liolios *et al*., 2010) allows genomic phylogenetic reconstructions based on the pan-genome (Tettelin *et al*., 2005). The core genome for a set of related genomes is a concept that involves the identification of orthologous genes common to a species (Goodall *et al*., 2017), and even genus (Alcaraz *et al*., 2010). The biological relevance of the core genome is to be discussed and analyzed yet because it tends to decrease if more genomes are added to the comparison. However, it provides a set of genes that are probably responsible for a genus evolutionary cohesion. For example, 20 strains encompassing 13 species of the *Bacillus* genus were determined to share 814 core genes which defined genus features like the ability to form endospores (Alcaraz *et al*., 2010).

The core genome is automatically computable by software pipelines that identify shared orthologous genes (Contreras-Moreira & Vinuesa, 2013). Traditional phylogenetic reconstructions only use vertically inherited core genes ignoring clade-specific genes. Ignoring these genes discards relevant elements of the biology of these organisms like horizontal gene transfer (HGT), gene family expansions, and gene content variability. Innocuous and pathogenic strains can be indistinguishable using traditional phylogenetic methods. We think that a metric representing actual genomic distances from pairwise shared homologous genes within a set of bacterial genomes will answer the most common question when sequencing the genome of a new strain: How related is the strain to known relatives?

The Genomic Similarity Score (GSS) has been used to obtain a non-redundant set of genomes (Janga & Moreno-Hagelsieb, 2004; Moreno-Hagelsieb & Janga, 2007; Alcaraz *et al*., 2010; Moreno-Hagelsieb *et al*., 2013). The GSS is a pairwise metric that depends on the normalized bit-scores of reciprocal best BLAST hits between orthologous proteins. GSS takes values from 0 to 1: when all orthologous proteins between two proteomes are identical it has a maximum value of 1, two genomes with no orthologous proteins have a value of 0 (Moreno-Hagelsieb & Janga, 2007). Best reciprocal BLAST hits have been used to identify orthologs when comparing complete genomes (Moreno-Hagelsieb & Janga, 2007). The pairwise GSS values can define a distance matrix between a set of genomes which can be turned into a distance dendrogram. Outgroups can be included in the comparison to root the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

dendrogram.

The GSS score was used to generate a dendrogram for the 108 strains comprising 16 species of *Streptococcus* for comparison to a 16S rRNA gene phylogenetic reconstruction. A core genome was built from the 108 strains to measure the sequence diversity of *Streptococcus*. Additionally, the core genome was used to discriminate between closely related strains in metagenomic sequences of *Streptococcus* dominated environments like the human mouth, where strains of the same genus are differential for causing caries (Belda-Ferre *et al*., 2012; Alcaraz *et al*., 2012; López-López *et al*., 2017).

Methods

Analyzed genomes and ortholog mapping.

Predicted proteomes for 108 strains of *Streptococcus*, representing 16 different species were downloaded from NCBI Genbank (Supplemental Information 1). Orthologs were defined as Reciprocal Best Hits (RBH) of pairwise comparisons using the BLASTp program (Camacho *et al*., 2009), the following parameters were used as previously suggested (Moreno-Hagelsieb & Latimer, 2008): e-value cutoff set to 1e-6 '-evalue 1e-6', mask low complexity regions of the query sequence only during the search phase '-soft_masking "true"', and perform an alignment with the Smith-Waterman algorithm to compute the bitscore '-use_sw_tback'. Then, hits with an alignment length shorter than 60% of the length of the query sequence were discarded. Detailed scripting procedure of RBH is available (Supplemental Information 2).

Genomic Similarity Score (GSS)

The GSS was calculated as previously reported (Janga & Moreno-Hagelsieb, 2004; Moreno-Hagelsieb & Janga, 2007; Alcaraz *et al*., 2010; Moreno-Hagelsieb *et al*., 2013). Briefly, from the RBH of pairwise comparisons of predicted proteomes, the raw bit-score was parsed for each pair of aligned sequences of the proteomes and summed, then the self-scores of proteome a were summed and used to normalize

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

the summed raw scores. Values of GSS have a range from 0-1, and GSS formula is calculated in the following form:

$$GSS_a = \frac{\sum_{i=1}^{n} compScore_i}{\sum_{i=1}^{n} selfScore_i}$$

Where *compScore* is the bitscore of protein *i* against its reciprocal best hit and *selfScore* is the bitscore of the alignment of protein *i* against itself in proteome *a*. Since *selfScore* might differ in proteome *a* and *b*, the final GSS for the proteome pair *ab* is the arithmetic mean of GSS_a and GSS_b. We used two bacilli species (*Bacillus subtilis* 168*,* and *B. licheniformis*) as outgroups for the comparisons of GSS values, as *Bacillus* is the external group to *Streptococcus* according to a whole genome phylogeny (Ciccarelli, 2006). An inverse (1-GSS) distance matrix was built and used to compute a Neighbor-Joining tree using the ape library v. 3.5 (Paradis, Claude & Strimmer, 2004) for R v. 3.3.1 (R Development Core Team, 2003). A control phylogeny was built using 16S rRNA full-length sequence from each of the 108 streptococci genomes. The multiple alignment for the 16S rRNA genes was done using structural RNA information using the software ssu-align (v0.1) (Nawrocki, 2009). The resulting 16S rRNA phylogeny was plotted using the Neighbor-Joining method from MEGA 5.2 (Tamura *et al*., 2013). GSS calculation protocols are available as Supplemental Information 2.

Core genome calculations

As a reference for all the core genome comparisons the smallest predicted proteome of all the streptococci analyzed strains was used: *S. agalactiae* 2-22 (FO393392; 1548 proteins). From the RBH calculations, results were compared, and the intersection set of orthologous proteins for all the 108 streptococci was defined as the core genome. From the local alignments from RBH comparisons, global alignments were performed using the Needleman-Wunsch method implemented in needleall of the EMBOSS suite (Rice, Longden & Bleasby, 2000), global alignments were used to calculate global sequence identity for each core protein. Additionally,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

the core genome was defined using the software package GET_HOMOLOGUES (Contreras-Moreira & Vinuesa, 2013) with the blastp program to perform comparisons and the BDBH algorithm to define orthologous clusters. The minimum alignment coverage was set to 60% and and the maximum E-value to 1e-06. Only clusters that included at least one sequence from all the analyzed genomes were considered for further analysis. Only protein coding genes were considered.

Pan-genome calculation

The *Streptococcus* genus pan-genome was calculated by clustering all the predicted proteomes using cd-hit (Huang *et al*., 2010) with an identity cut-off value of 70%. This clustering method allows to generate protein family without constraints of inparalog groupings that collapses large gene family (*i.e.,* ABC transporters). Additionally, GET_HOMOLOGUES was used as a second method to obtain the genus pan-genome. BLASTp (Camacho *et al*., 2009) hits with at least 70% sequence identity, a minimum of 75% alignment length coverage, and an E-value of 1e-6 were considered. The OrthoMCL algorithm (Li, Stoeckert & Roos, 2003) was used to group sequences. Only protein coding genes were considered.

ANI calculation

Average Nucleotide Identity (ANI) was calculated using pyani (Marçais *et al*., 2018) for the 108 genomes used in this study (Supplemental Information 1) with two methods: Mummer (Marçais *et al*., 2018) using minimum lengths of exact match (20 nt), maximum gaps (90 nt); and BLASTN+ (Camacho *et al*., 2009) with 1020 nucleotide windows.

Core genome and pan-genome annotation

The core and pan-genomes were annotated using MG-RAST (Huang *et al*., 2010; Meyer *et al*., 2017) and their M5NR database (Wilke *et al*., 2012). Annotation required a minimum alignment length of 15 amino acids and 60% identity.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

Streptococci coding genes were uploaded to MG-RAST because it is possible to compare them with multiple metagenomes, in particular, human oral metagenomes where *Streptococcus* species composition has repercussions for health or disease status (Belda-Ferre *et al*., 2012; Alcaraz *et al*., 2012; López-López *et al*., 2017).

Metagenomic comparisons

Fragment recruitment analysis (Rusch *et al*., 2007) was done to compare oral metagenomes from a healthy and diseased individuals against the *Streptococcus* reference core genome for each streptococci species using Nucmer from the Mummer suite (Marçais *et al*., 2018). A cut-off value of 90% identity (nucleotide) was chosen for classifying each metagenomic read to an individual species. Using minimum lengths of exact match (20 nt) and maximum gaps (90 nt).

Results

Phylogenetic and genome similarity of the *Streptococcus* genus.

A 16S rRNA phylogenetic reconstruction was done as a reference and confirms previously proposed clades (Fig. 1A) (Kawamura *et al*., 1995). There is a Pyogenic clade containing multiple species: *S. pyogenes, S. dysgalactiae, S. equi, S. uberis, S. parauberis, S. agalactiae,* and *S. pneumoniae.* A second clade is the Salivarius group formed just by *S. thermophilus* and *S. salivarius.* The Mutans clade groups the following species: *S. mutans, S. infantarius, S. lutetiensis, S. macedonicus*, and *S. gallolyticus.* The species *S. suis* has its clade with multiple strains of the same species. A fifth clade known as Mitis group is the basal group: *S. pneumoniae, S. pseudopneumoniae, S. mitis, S. pasteurianus, S. parasanguinis, S. sanguinis, S.gordonii, S. oligofermentans,* and *S. intermedius*. The external groups are *Bacillus subtilis* 168 and *B. licheniformis*.

Besides, ANI was calculated for all the Streptococci genomes. ANI was able to discriminate main Pyogenic and Suis clades (Fig. 1 B); but it does break Mutans, Salivarius, which are supported both by 16S phylogeny and GSS dendrogram (Fig. 1 A,C). Interestingly, there is an ANI clade formed by a mix of Pyogenic, Mitis, and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

Salivarius groups, not supported neither by GSS or 16S phylogeny. ANI correlogram is available (Supplemental Information 3).

The GSS dendrogram has the same clades as the 16S rRNA (Fig. 1C); however, GSS rearranges the Pyogenic group, where *S. agalactiae* is included interior to the Pyogenic clade in the 16S phylogeny, GSS shows it as the basal group for the Pyogenic clade. Another rearrangement of GSS has the Suis group as a sister clade to the Mitis group, but in the 16S rRNA phylogeny, Suis is placed as a sister clade to the Pyogenic group. It is noticeable that the GSS dendrogram distances are large enough to distinguish discrete groups among closely related strains like such as the inner clades of Suis, Pyogenic, Mutans, and Mitis groups. Resolved clades are formed in the GSS dendrogram for strains of *S. pneumoniae* and *S. pseudopneumoniae*; whereas, 16S rRNA does not distinguish inner relationships, but rather allows polytomies. Also, the Suis GSS clade shows clearly resolved branches when comparing to the 16S rRNA phylogeny.

Core genome sequence diversity

According to the RBH method, the 108 streptococci core genome has 404 proteinsis a small number compared to the average protein content of 1,929 for the 108 strains., The core proteins represent one-fifth of the average predicted proteome. The total pan-genome comprises 33,039 protein clusters (families) at 70% identity (Supplemental Information 4). According to the GET_HOMOLOGUES method, the core genome is composed of 306 proteins and the pan-genome of 36,387. Comparisons of the core proteins obtained by both methods find 255 proteins, the RBH method finds 149 unique proteins, while GET_HOMOLOGUES finds only 51 unique proteins (Supplementary Information 5).

Paired global alignments were performed to analyze variation across species and strains over the core genome (Fig. 2). For each core cluster, the individual proteins were plotted showing the pairwise identity of the protein compared to a reference sequence from *S. pyogenes* which was chosen as the reference because of its top phylogenetic position both in 16S and in GSS dendrogram (Fig. 2). The high sequence identity (mean=77.6±11.5) for the core proteome is suggesting evidence

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

for selective constraints (Supplemental Information 5). Identity to *S. pyogenes* over the core genome, diminishes as the genomic distances to other species increase. The range of protein sequence diversity in the core proteome goes from 25 - 100% identity. Based on the core proteome sequence diversity, we were able to describe a set of phylogenetic markers that can be used as DNA references to identify and discriminate between closely related species in metagenomes using high nucleotide identity cut-offs (>90%). Core genes for each of the streptococci species described here are available for the community in FASTA format (Supplemental Information 6).

Core genome functional analysis

Normalized abundances (Z-scores) of the pan-genome against the core were compared to stress out the over-represented protein categories in the core (Fig. 3). The most abundant genes in the 404 protein core clusters are related to the translational machinery, including ribosomal proteins and translation-related proteins (Z=3.08 core; Z=0.88 pan-genome). There are more cell division related proteins in the core genome (Z=-0.87), than in the pan-genome (Z=-1.06). Membrane and cell envelope coding genes (M) are better represented in the core genome (Z=0.22; Z=0.10 pan-genome). The most conserved core proteins (average pairwise identity >90%) are mostly related to the translation process, and the top 10 are exclusively ribosomal proteins (Supplemental Information 5). As average pairwise identity decreases for the core proteins, several transport proteins appear along with multiple transport-related proteins, transcriptional regulators, phosphatases, recombinases, peptidases, multidrug and efflux transporters (MATE), and hypothetical proteins (Fig. 2; Supplemental Information 5). There is also a high proportion of core proteins with unknown function (48 out of 404; 11.81%).

Using the core genome to scan oral metagenomes

Metagenomic shotgun reads from oral microbiome samples were mapped to the individual sequences from the core genome to estimate relative abundance for each *Streptococcus* species. Oral metagenomes were chosen because of the many streptococci with high abundance (4 to >20%) (Supplemental Information 7). Two oral metagenomes were chosen: a patient with active caries and a healthy adult that

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

never suffered from caries (Belda-Ferre *et al*., 2012). In both metagenomes, the species with the most recruited number of fragments was *S. pneumoniae* (Fig. 4 and Table 1), but the caries etiological agent *S. mutans* is depleted (17 metagenomic fragments) in the caries-free individual (NOCA_01) and abundant (127 metagenomic fragments) in the patient with caries. Recruiting metagenomic sequences against each reference core genome and filtering alignments with high identity levels (≥ 90%) shows that is possible to generate species-specific profiles (Fig. 4, Table 1).

Discussion

*Streptococcu*s species have historically been classified by their cell wall antigenic properties (Kayser, Bienz & Eckert, 2011) and clinical criteria for pathogenic strains like hemolysis capabilities. More recently molecular phylogenetics has aided streptococci classification (Kawamura *et al*., 1995; Kilian *et al*., 2008). Analysis of genomic variability within the same species expanded with the definition of relevant concepts like the pan-genome and the core genome for *S. agalactiae* (Tettelin *et al*., 2005).

The core genome is dependent on the set of genomes being analyzed, for each genome added the size of the core would decrease if any genes are not present for that genome. Besides this, different methods can estimate different core and pan-genome sizes, this have been shown in previous works (Fouts *et al*., 2012). In this work, 404 core proteins comprise the core genome according to the RBH method and given the 108 strains compared, while GET_HOMOLOGUES gets 306 proteins. Historically, the first core genome for streptococci was 611 genes for 26 genomes (Lefébure & Stanhope, 2007); a second effort wast 547 genes for 64 genomes (Van den Bogert *et al*., 2013); a third reconstruction gave 369 core genes for 138 strains (Gao *et al*., 2014). Interestingly, 11.81% of the core genes of streptococci are of unknown function (Supplemental Information 5), representing an opportunity as possible therapeutic targets.

The core genome for streptococci provides a platform for investigating what is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

essential to the lifestyle of these organisms and also can be used to analyze their presence in metagenomic samples. Additionally, we think that traditional phylogenetic methodology is necessary to understand vertical group evolution and GSS or similar measures of whole genome relatedness are an improvement over marker gene-based methods. However, bacteria have amazing capabilities to transfer genes through conjugation, transformation, and competence, with high rates of recombination, which pose a challenge for traditional phylogenetics (Frost *et al*., 2005; Francino, 2012). Pan-genomic analysis shows the variability within a species which may indicate adaptation to particular environments through additions or deletions to the genomic repertoire (Tettelin *et al*., 2008; Mira *et al*., 2010; Vernikos *et al*., 2015). The GSS measures bacterial strain similarity over all homologous genetic elements shared by a pair of bacteria, no matter if it is vertically or horizontally transmitted (Janga & Moreno-Hagelsieb, 2004; Moreno-Hagelsieb & Janga, 2007; Alcaraz *et al*., 2010; Moreno-Hagelsieb *et al*., 2013). Recently, new standards are establishing in the bacterial taxonomic rules trying to make use of whole genome information, and ANI is the preferred choice to discriminate between species (Chun *et al*., 2018). Working at with genus level, involves methods able identify homologous sequences, here we found protein sequence diversity with distances spanning from 100% to less than 25% identity for the global alignments. The main advantage of GSS is that it uses both core and pan-genomic information to estimate relatedness between strains. Proteins are the choice to find homologs with large evolutionary distances (Rost, 1999). ANI will be the choice when comparing inside strains of the same species (Chun *et al*., 2018), but it discards homologous information due to the shortcome of comparing nucleotides when comparing long time diverging lineages, in Streptococci there are estimates about 0.5 billion years of the last common ancestor (Battistuzzi, Feijao & Hedges, 2004). Multiple sequenced strains redundancy complicates comparative genome analysis, information beyond nucleotide clustering is needed; genome redundancy elimination by using information like distance matrix or phylogenetic information like GGRaSP (Clarke *et al*., 2018), GSS could easily integrate to tools like GGRaSP.

The GSS dendrogram is consistent with the accepted clades of streptococci. GSS provides better resolution of clade structure and distances than the 16S rRNA gene based phylogeny (Fig. 1). Within group resolution is greatly improved in the GSS

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

dendrogram for several streptococci species like *S. pyogenes, S. suis, S. mutans,* and *S. pneumoniae* which are practically indistinguishable using 16S but GSS shows monophyletic clades for each species with clear branching and long enough distances to identify each strain within a species (Fig 1C).

The growth of metagenomic data needs a framework to distinguish between closely related strains. Some environments host intra-genus diversity with implications for health like human vaginal microbiomes dominated by *Lactobacillus* species (Gajer *et al*., 2012), and the human oral microbiome (Belda-Ferre *et al*., 2012; Simón-Soro *et al*., 2013). There are multiple ways to bin metagenomic diversity from nucleotide k-mer frequencies (Ulyantsev *et al*., 2016), using phylogenomic markers (Segata *et al*., 2012), AMPHORA (Segata *et al*., 2012; Kerepesi, Bánky & Grolmusz, 2014), through annotation of ribosomal genes (Pruesse *et al*., 2007; Cardenas *et al*., 2009), and lowest common ancestor binning (Huson *et al*., 2007; Meyer *et al*., 2017). In this work, the use of the core genome of a genus provides a relatively simple (404 genes) dataset to align metagenomic information (reads, contigs) against and estimate species abundances based on the coverage and identity of each aligned fragment (Fig. 4). Despite the biological relevance, or connecting it to essential genes (Goodall *et al*., 2017), the core genome of a clade provides a resource to discriminate between closely related strains. Sequence identity variation within the core genome provides a basis for understanding the differential selective pressure for each core cluster (Supplemental information 5). Core genome variation could be exploited in practical and biological meaningful ways like probe and diagnosis design or understanding conserved but highly variable proteins.

Conclusions

The core genome of bacteria, no matter if species, genus or whatever preferred level should be an open repository and recalculated each time a new strain is sequenced, and shared with the scientific community, maybe through a “living” paper that self-updates with new genomes. Here is presented a working version of streptococci core genome with 404 predicted proteins. Additionally, core genome and pan-genome are not just mathematical concepts only, the functional metabolic roles of the known genes are relevant and also its natural variations.Traditional marker gene based

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

phylogenetic tools in bacteria are invaluable; however, they do not capture the dynamism occurring in bacterial genomes and other tools like the GSS better distinguish genome level relatedness between species. A practical use for the core genome of the streptococci is to classify abundances of different species and strains in metagenomic samples. Finally, the range of sequence diversity within each *Streptococcus* core cluster will need further analysis to determine if the level of sequence identity correlates with selective pressures.

Acknowledgements

We are grateful to the Facultad de Ciencias UNAM community for their warm and sincere welcoming and giving us the opportunity to do our research with them, particularly to Víctor Valdés-López, Luisa A. Alba-Lois, Claudia Segal, and Viviana Escobar for their helpful support that made this work possible. The authors sincerely appreciate Granger Sutton reviewer work whose suggestions significantly improve our manuscript, along with the comments of other two anonymous reviewers.

Funding

HBT, MR, and SMS had fellowships from CONACyT. LDA got funding from DGAPA-PAPIIT-UNAM TA2001171 and SEP-CONACyT Ciencia Básica 237387.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

References

Alcaraz LD., Belda-Ferre P., Cabrera-Rubio R., Romero H., Simón-Soro A., Pignatelli M., Mira A. 2012. Identifying a healthy oral microbiome through metagenomics. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 18 Suppl 4:54–57. DOI: 10.1111/j.1469-0691.2012.03857.x.

Alcaraz LD., Moreno-Hagelsieb G., Eguiarte LE., Souza V., Herrera-Estrella L., Olmedo G. 2010. Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC genomics* 11:332. DOI: 10.1186/1471-2164-11-332.

Battistuzzi FU., Feijao A., Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC evolutionary biology* 4:44. DOI: 10.1186/1471-2148-4-44.

Belda-Ferre P., Alcaraz LD., Cabrera-Rubio R., Romero H., Simón-Soro A., Pignatelli M., Mira A. 2012. The oral metagenome in health and disease. *The ISME journal* 6:46–56. DOI: 10.1038/ismej.2011.85.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST : architecture and applications. *BMC bioinformatics* 10:421. DOI: 10.1186/1471-2105-10-421.

Cardenas E., Cole JR., Tiedje JM., Park J-H. 2009. Microbial Community Analysis using RDP II (Ribosomal Database Project II):Methods, Tools and New Advances. *Environmental Engineering Research* 14:3–9. DOI: 10.4491/eer.2009.14.1.003.

Chun J., Oren A., Ventosa A., Christensen H., Arahal DR., da Costa MS., Rooney AP., Yi H., Xu X-W., De Meyer S., Trujillo ME. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International journal of systematic and evolutionary microbiology* 68:461–466. DOI: 10.1099/ijsem.0.002516.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

Ciccarelli FD. 2006. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science* 311:1283–1287. DOI: 10.1126/science.1123061.

Clarke TH., Brinkac LM., Sutton G., Fouts DE. 2018. GGRaSP: A R-package for selecting representative genomes using Gaussian mixture models. *Bioinformatics* . DOI: 10.1093/bioinformatics/bty300.

Contreras-Moreira B., Vinuesa P. 2013. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and environmental microbiology* 79:7696–7701. DOI: 10.1128/aem.02411-13.

Fouts DE., Brinkac L., Beck E., Inman J., Sutton G. 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research* 40:e172. DOI: 10.1093/nar/gks757.

Francino MP. 2012. The Ecology of Bacterial Genes and the Survival of the New. *International journal of evolutionary biology* 2012:1–14. DOI: 10.1155/2012/394026.

Fraser C., Alm EJ., Polz MF., Spratt BG., Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323:741–746. DOI: 10.1126/science.1159388.

Frost LS., Leplae R., Summers AO., Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. *Nature reviews. Microbiology* 3:722–732. DOI: 10.1038/nrmicro1235.

Gajer P., Brotman RM., Bai G., Sakamoto J., Schütte UME., Zhong X., Koenig SSK., Fu L., Ma ZS., Zhou X., Abdo Z., Forney LJ., Ravel J. 2012. Temporal dynamics of the human vaginal microbiota. *Science translational medicine* 4:132ra52. DOI: 10.1126/scitranslmed.3003605.

Gao X-Y., Zhi X-Y., Li H-W., Klenk H-P., Li W-J. 2014. Comparative genomics of the bacterial genus *Streptococcus* illuminates evolutionary implications of species groups.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

PloS one 9:e101229. DOI: 10.1371/journal.pone.0101229.

Goodall E., Robinson A., Johnston I., Jabbari S., Turner K., Cunningham A., Lund P., Cole J., Henderson I. 2017. The essential genome of *Escherichia coli* K-12. DOI: 10.1101/237842.

Huang Y., Niu B., Gao Y., Fu L., Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26:680–682. DOI: 10.1093/bioinformatics/btq003.

Huson DH., Auch AF., Qi J., Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome research* 17:377–386. DOI: 10.1101/gr.5969107.

Janga SC., Moreno-Hagelsieb G. 2004. Conservation of adjacency as evidence of paralogous operons. *Nucleic acids research* 32:5392–5397. DOI: 10.1093/nar/gkh882.

Kawamura Y., Hou XG., Sultana F., Miura H., Ezaki T. 1995a. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. *International journal of systematic bacteriology* 45:406–408. DOI: 10.1099/00207713-45-2-406.

Kayser FH., Bienz KA., Eckert J. 2011. *Medical Microbiology*. Thieme.

Kerepesi C., Bánky D., Grolmusz V. 2014. AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene* 533:538–540. DOI: 10.1016/j.gene.2013.10.015.

Kilian M. 2012. *Streptococcus* and *enterococcus*. In: *Medical Microbiology*. 183–198. DOI: 10.1016/b978-0-7020-4089-4.00031-7.

Kilian M., Poulsen K., Blomqvist T., Håvarstein LS., Bek-Thomsen M., Tettelin H., Sørensen UBS. 2008. Evolution of *Streptococcus pneumoniae* and its close commensal relatives. *PloS one* 3:e2683. DOI: 10.1371/journal.pone.0002683.

Konstantinidis KT., Ramette A., Tiedje JM. 2006. The bacterial species definition in

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

the genomic era. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 361:1929–1940. DOI: 10.1098/rstb.2006.1920.

Lefébure T., Stanhope MJ. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome biology* 8:R71. DOI: 10.1186/gb-2007-8-5-r71.

Liolios K., Chen I-MA., Mavromatis K., Tavernarakis N., Hugenholtz P., Markowitz VM., Kyrpides NC. 2010. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research* 38:D346–54. DOI: 10.1093/nar/gkp848.

Li L., Stoeckert CJ Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13:2178–2189. DOI: 10.1101/gr.1224503.

López-López A., Camelo-Castillo A., Ferrer MD., Simon-Soro Á., Mira A. 2017. Health-Associated Niche Inhabitants as Oral Probiotics: The Case of *Streptococcus dentisani*. *Frontiers in microbiology* 8. DOI: 10.3389/fmicb.2017.00379.

Marçais G., Delcher AL., Phillippy AM., Coston R., Salzberg SL., Zimin A. 2018a. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* 14:e1005944. DOI: 10.1371/journal.pcbi.1005944.

Meyer F., Bagchi S., Chaterji S., Gerlach W., Grama A., Harrison T., Paczian T., Trimble WL., Wilke A. 2017. MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Briefings in bioinformatics*. DOI: 10.1093/bib/bbx105.

Mira A., Martín-Cuadrado AB., D'Auria G., Rodríguez-Valera F. 2010. The bacterial pan-genome:a new paradigm in microbiology. *International microbiology: the official journal of the Spanish Society for Microbiology* 13:45–57. DOI: 10.2436/20.1501.01.110.

Moreno-Hagelsieb G., Janga SC. 2007. Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. Proteins: Structure,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

Function, and Bioinformatics 70:344–352. DOI: 10.1002/prot.21564.

Moreno-Hagelsieb G., Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24:319–324. DOI: 10.1093/bioinformatics/btm585.

Moreno-Hagelsieb G., Wang Z., Walsh S., ElSherbiny A. 2013c. Phylogenomic clustering for selecting non-redundant genomes for comparative genomics. *Bioinformatics* 29:947–949. DOI: 10.1093/bioinformatics/btt064.

Nawrocki EP. 2009. Structural RNA Homology Search and Alignment Using Covariance Models. PhD Thesis. Washington University.

Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290. DOI: 10.1093/bioinformatics/btg412.

Pruesse E., Quast C., Knittel K., Fuchs BM., Ludwig W., Peplies J., Glockner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research* 35:7188–7196. DOI: 10.1093/nar/gkm864.

R Development Core Team. 2003. *The R Reference Manual: Base Package*. Network Theory.

Rice P., Longden I., Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics: TIG* 16:276–277.

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein engineering, design & selection: PEDS* 12:85–94. DOI: 10.1093/protein/12.2.85.

Rusch DB., Halpern AL., Sutton G., Heidelberg KB., Williamson S., Yooseph S., Wu D., Eisen JA., Hoffman JM., Remington K., Beeson K., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews-Pfannkoch C., Venter JE., Li K., Kravitz S., Heidelberg JF., Utterback T., Rogers Y-H., Falcón LI., Souza V., Bonilla-Rosso G., Eguiarte LE., Karl DM., Sathyendranath S., Platt

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018

T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari MR., Strausberg RL., Nealson K., Friedman R., Frazier M., Venter JC. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology* 5:e77. DOI: 10.1371/journal.pbio.0050077.

Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O., Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* 9:811–814. DOI: 10.1038/nmeth.2066.

Simón-Soro A., Belda-Ferre P., Cabrera-Rubio R., Alcaraz LD., Mira A. 2013. A tissue-dependent hypothesis of dental caries. *Caries research* 47:591–600. DOI: 10.1159/000351663.

Stackebrandt E., Goebel BM. 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International journal of systematic and evolutionary microbiology* 44:846–849. DOI: 10.1099/00207713-44-4-846.

Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* 30:2725–2729. DOI: 10.1093/molbev/mst197.

Tettelin H., Masignani V., Cieslewicz MJ., Donati C., Medini D., Ward NL., Angiuoli SV., Crabtree J., Jones AL., Durkin AS., Deboy RT., Davidsen TM., Mora M., Scarselli M., Margarit y Ros I., Peterson JD., Hauser CR., Sundaram JP., Nelson WC., Madupu R., Brinkac LM., Dodson RJ., Rosovitz MJ., Sullivan SA., Daugherty SC., Haft DH., Selengut J., Gwinn ML., Zhou L., Zafar N., Khouri H., Radune D., Dimitrov G., Watkins K., O'Connor KJB., Smith S., Utterback TR., White O., Rubens CE., Grandi G., Madoff LC., Kasper DL., Telford JL., Wessels MR., Rappuoli R., Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences of the*

United States of America 102:13950–13955. DOI: 10.1073/pnas.0506758102.

Tettelin H., Riley D., Cattuto C., Medini D. 2008. Comparative genomics: the bacterial pan-genome. *Current opinion in microbiology* 11:472–477.

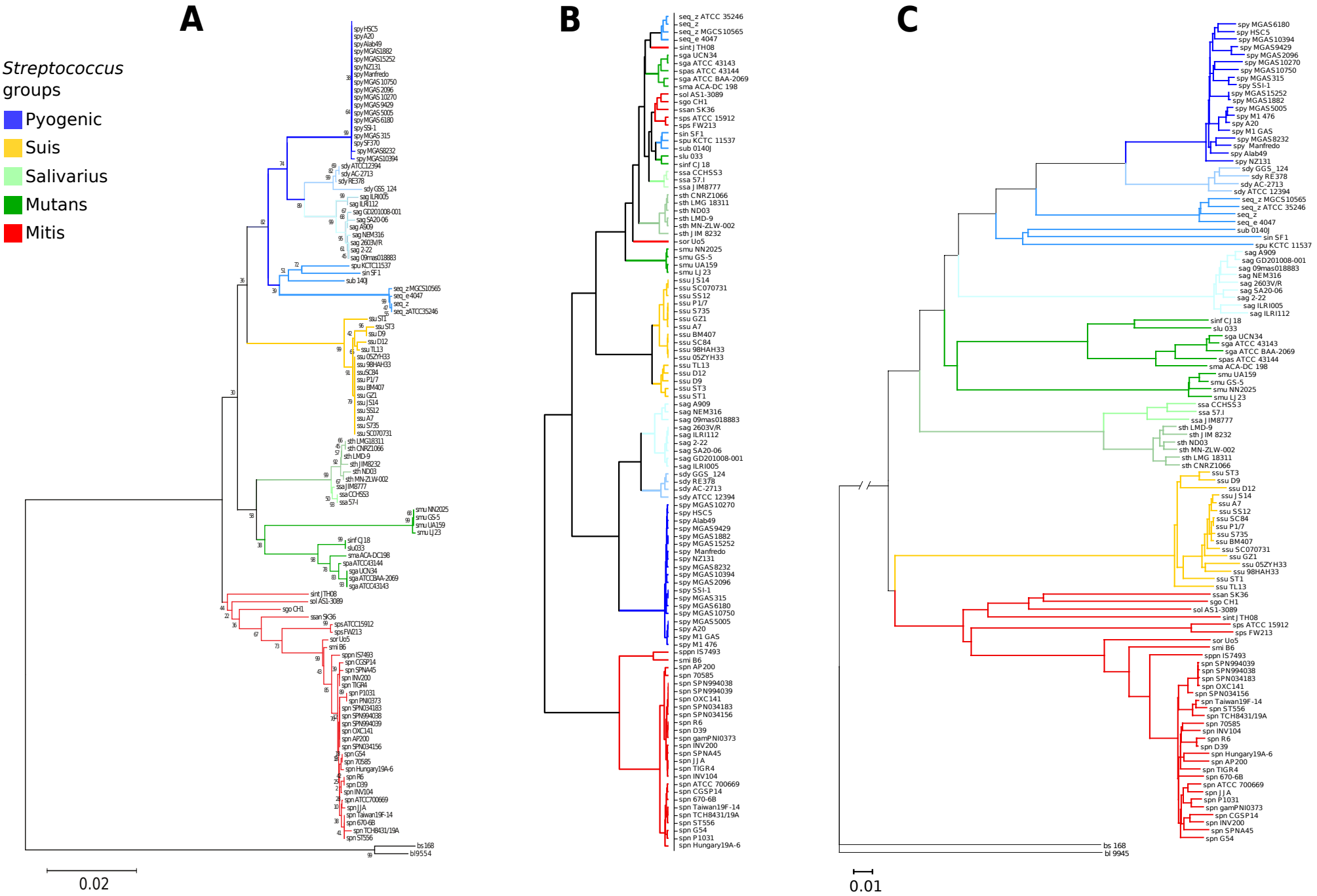
Ulyantsev VI., Kazakov SV., Dubinkina VB., Tyakht AV., Alexeev DG. 2016. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics* 32:2760–2767. DOI: 10.1093/bioinformatics/btw312.

Van den Bogert B., Boekhorst J., Herrmann R., Smid EJ., Zoetendal EG., Kleerebezem M. 2013. Comparative genomics analysis of *Streptococcus* isolates from the human small intestine reveals their adaptation to a highly dynamic ecosystem. *PloS one* 8:e83418. DOI: 10.1371/journal.pone.0083418.

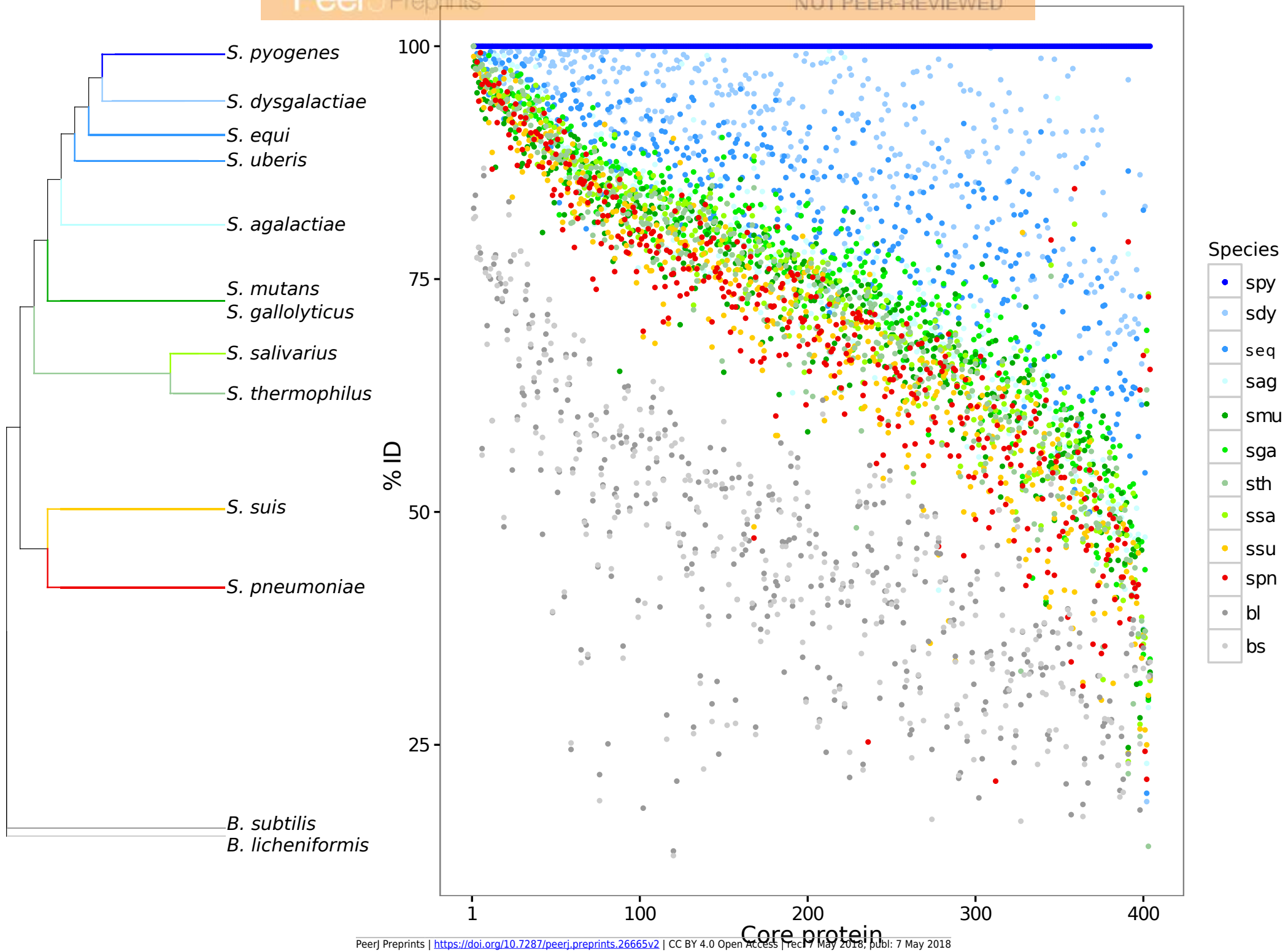
Vernikos G., Medini D., Riley DR., Tettelin H. 2015. Ten years of pan-genome analyses. *Current opinion in microbiology* 23:148–154. DOI: 10.1016/j.mib.2014.11.016.

Wilke A., Harrison T., Wilkening J., Field D., Glass EM., Kyrpides N., Mavrommatis K., Meyer F. 2012. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC bioinformatics* 13:141. DOI: 10.1186/1471-2105-13-141.

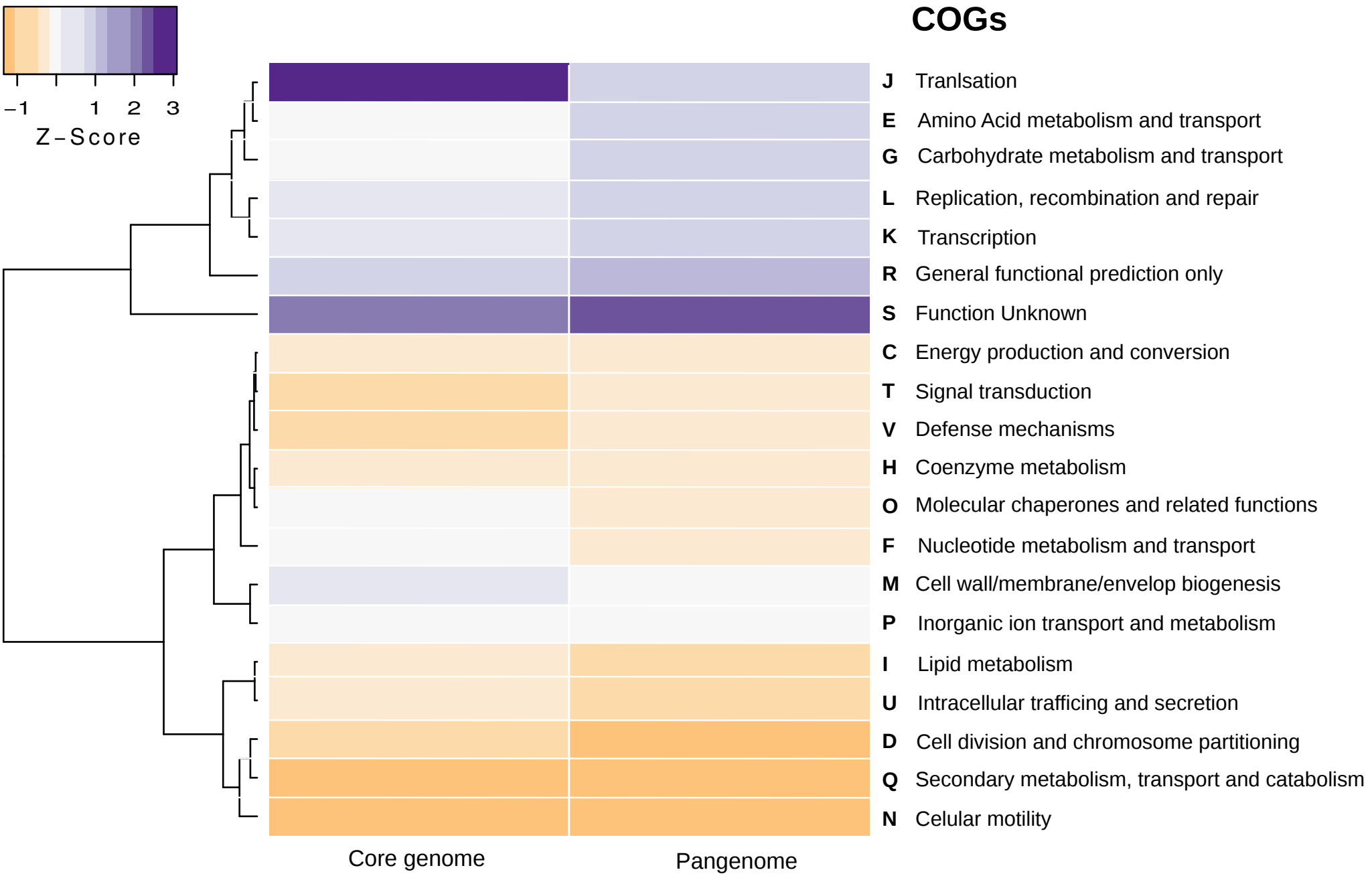
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018



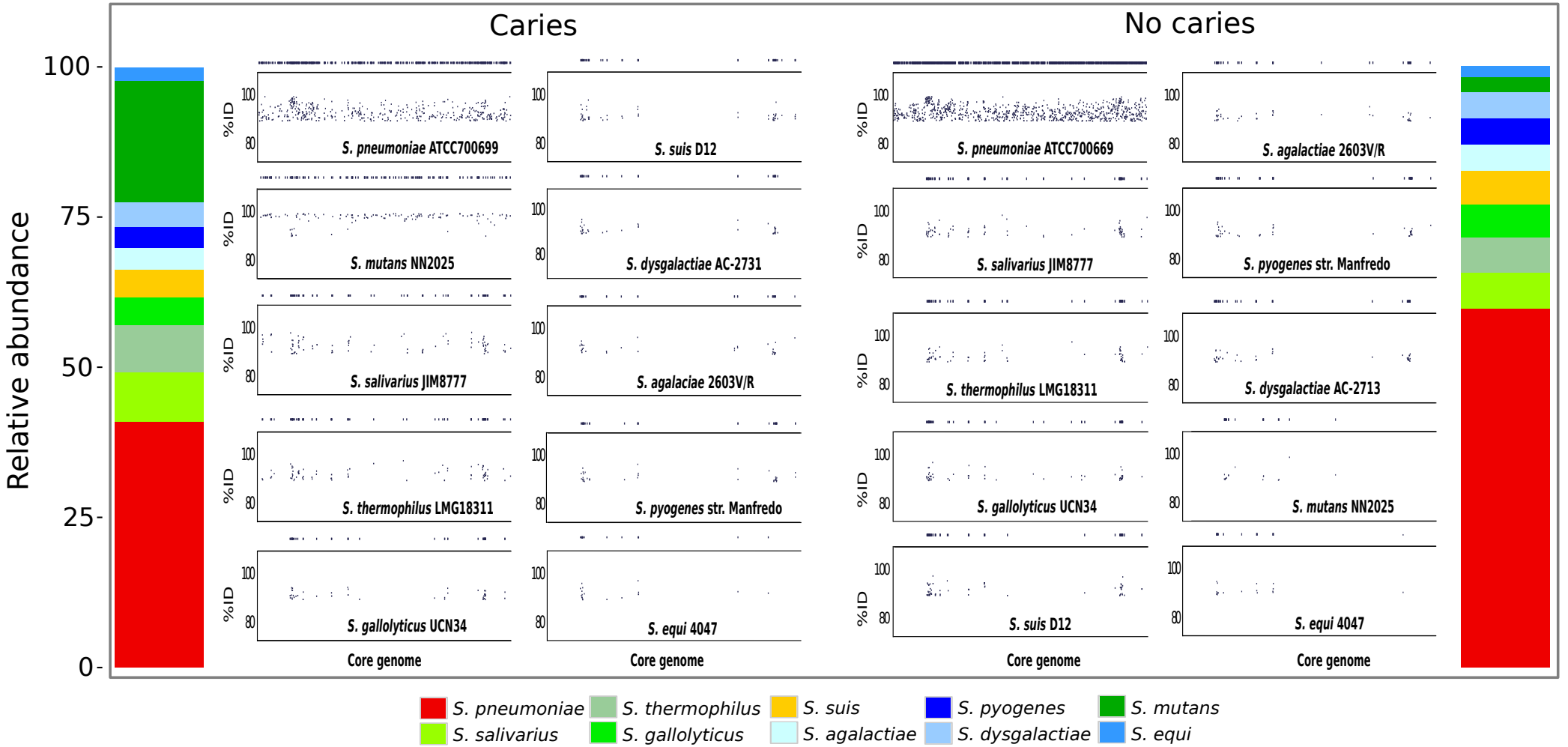
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v2 | CC BY 4.0 Open Access | rec: 7 May 2018, publ: 7 May 2018