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 close related bacterial strains in complex environments

Hugo R Barajas de la Torre [1], **Miguel Romero** [1], **Shamayim Martínez-Sánchez** [1], **Luis David Alcaraz** Corresp. [1]

[1] Facultad de Ciencias. Departamento de Biología Celular, Universidad Nacional Autónoma de México

Corresponding Author: Luis David Alcaraz
Email address: lalcaraz@ciencias.unam.mx

Background. Comparative genomics between closely related bacterial strains aids to distinguish important features like pathogenesis, antibiotic resistance, and phylogenetic structure. *Streptococcus* is relevant because public health and food safety and it are well-represented (>100 genomes) in databases of publicly available databases. Streptococci are cosmopolitan, and there are multiple sources of isolation, from humans to dairy products. The *Streptococcus* have been classified by morphology, serum types, 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 *Streptococcus* and using their core genome as a simplified tool to assess strain specific abundances in metagenomic sequences.

Methods. A 16S rRNA gene phylogeny has been calculated for 108 strains, belonging to 16 *Streptococcus* species and compared the results to a dendrogram using the GSS with all homologous shared information available in the genomes. Additionally, genus core and pan-genome were calculated. The core genome sequences identity was analyzed and the core genome was used as a seed to discriminate abundances between close related strains in metagenomic samples.

Results. A total of 404 proteins are shared by all 108 *Streptococcus* genomes, which are the core genome. The core identity values ranges across all the compared strains and outgroups are reported. Lower sequence identity variation (90-100%) within the core belongs to ribosomal and translation-related proteins. It was found out that 48 proteins (11.8%) of the core genome are considered a hypothetical protein and those proteins host the larger sequence identity variations within the core. The sequence identity of the core genome identity diminishes as GSS score between species increases. The GSS dendrogram recovers most of the clades in the 16S rRNA gene phylogeny with the advantage to distinguish between 16S polytomies (unresolved nodes). Finally, our proposed core genome was used to distinguish the abundances of close related strains within human oral metagenomes being able to get strain relative abundances between healthy and caries infected (with *S. mutans*) individuals.

Discussion. The clinical and food safety importance of *Streptococcus* genus gives a playground to test multiple comparative genomic scenarios due to its excellent genomic coverage. 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. Combination of core genome and rapid alignment tools allows to estimate abundance and discriminate in a strain-specific manner in metagenomic samples. Here it is shared with the community both GSS genomic dendrogram and core genome to explore possibilities within streptococci.

Global genomic similarity and core genome sequence diversity of the *Streptococcus* genus as a toolkit to identify close related bacterial strains 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.26665v1 | CC BY 4.0 Open Access | rec: 13 Mar 2018, publ: 13 Mar 2018

Abstract

Background. Comparative genomics between closely related bacterial strains aids to distinguish important features like pathogenesis, antibiotic resistance, and phylogenetic structure. *Streptococcus* is relevant because public health and food safety and it are well-represented (>100 genomes) in databases of publicly available databases. Streptococci are cosmopolitan, and there are multiple sources of isolation, from humans to dairy products. The *Streptococcus* have been classified by morphology, serum types, 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 *Streptococcus* and using their core genome as a simplified tool to assess strain specific abundances in metagenomic sequences.

Methods. A 16S rRNA gene phylogeny has been calculated for 108 strains, belonging to 16 *Streptococcus* species and compared the results to a dendrogram using the GSS with all homologous shared information available in the genomes. Additionally, genus core and pan-genome were calculated. The core genome sequences identity was analyzed and the core genome was used as a seed to discriminate abundances between close related strains in metagenomic samples.

Results. A total of 404 proteins are shared by all 108 *Streptococcus* genomes, which are the core genome. The core identity values ranges across all the compared strains and outgroups are reported. Lower sequence identity variation (90-100%) within the core belongs to ribosomal and translation-related proteins. It was found out that 48 proteins (11.8%) of the core genome are considered a hypothetical protein and those proteins host the larger sequence identity variations within the core. The sequence identity of the core genome identity diminishes as GSS score between species increases. The GSS dendrogram recovers most of the clades in the 16S rRNA gene phylogeny with the advantage to distinguish between 16S polytomies (unresolved nodes). Finally, our proposed core genome was used to distinguish the abundances of close related strains within human oral metagenomes being able to get strain relative abundances between healthy and caries infected (with *S. mutans*) individuals.

Discussion. The clinical and food safety importance of *Streptococcus* genus gives a playground to test multiple comparative genomic scenarios due to its excellent genomic coverage. 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. Combination of core genome and rapid alignment tools allows to estimate abundance and discriminate in a strain-specific manner in metagenomic samples. Here it is shared with the community both GSS genomic dendrogram and core genome to explore possibilities within streptococci.

Background

Streptococcus sp. is a bacteria genus that englobes more than 40 different species, hosting a diverse range of human and animal pathogens like the etiological agents from caries to meningitis, but they can be commensal species inhabiting animal guts and respiratory tract (Killian, 2007). It is a well-known genus which classification and taxonomy have been done by multiple criteria since morphologic, biochemical profiles, serum types, and recently it has been done using the comparison of 16S ribosomal RNA (rRNA) gene phylogenies (Kawamura et al., 1995), and there are Multilocus Sequence Types (MLST) for 8 streptococci species (Jolley & Maiden, 2010). The Streptococci are divided in 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 cluster (Kilian et al., 2008). There are multiple genome sequences available for the streptococci, most of them are for species isolated from ill humans, bovine, swine, and dairy product samples (Supplemental Information 1).

Bacteria phylogenetics has been done using multiple criteria to define bacteria species. The current standard is based 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 (Fox, Wisotzkey & Jurtshuk, 1992; Stackebrandt & Goebel, 1994). A recent controversy about the use of multiple coding genes alignments known as multi locus sequence typing is standard practice for bacteria pathogenic strains, and even recent discussion has arisen as the definition for a standard in bacteria molecular phylogenetics species concept is fuzzy (Fraser et al., 2009).

With the astounding amount of bacteria genomes been sequenced in the last years (77,107 with available data in GenBank, February 2018; (Liolios et al., 2010)) it is

possible to perform further detailed phylogenetic reconstructions like the use of core genomes, and understanding the biological diversity of a strain-specific set of genes known as the pan-genome (Tettelin et al., 2005). The core genome is a concept that involves the identification of a shared set 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 biological cohesion. For example, when describing the *Bacillus* genus core genome it was determined that 814 genes were orthologous and common to 20 strains compared, when describing a defining genus features like the ability to form endospores; the study put into the spotlight genes that were part of the core genome and were master regulators for endospore formation (Alcaraz et al., 2010).

The core genome is now accessible through software pipelines that identify shared ortholog genes (Contreras-Moreira & Vinuesa, 2013). Nonetheless, the pan-genomic variability of a group shows that traditional phylogenetic reconstructions only take into account vertical inherited genes and discard strain-specific genes out of the analysis. It is our concern that traditional shared by all requisite of phylogenetics to draw the relationships of bacteria discard relevant elements of the biology of these organisms like horizontal gene transfer (HGT), gene families expansions, and their pan-genomic variability, which is enough to have innocuous and pathogenic strains that are indistinguishable using traditional phylogenetic methods. We think that a metric representing actual genomic distances from pairwise shared homologous genes between a set of bacteria strains will allow to answer the most common question when sequencing the genome of a new strain: How related is the strain to their known relatives?

The Genomic Similarity Score (GSS) has been used before successfully, and it has been used to get 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 metric that depends on the normalized bit-scores of reciprocal best BLAST hits between a shared set of predicted proteomes. GSS takes values from 0 to 1; when a compared pair of proteomes are identical, it has a maximum value of 1, two unrelated

proteomes will have 0 value (Moreno-Hagelsieb & Janga, 2007). Best reciprocal BLAST hits have been used to identify orthologs when comparing complete genomes (Moreno-Hagelsieb & Janga, 2007). The paired GSS values can be parsed into a distance matrix between a group of organisms which can be turned into a distance dendrogram. If outgroups are included in the comparison, it will allow to guide and polarize the dendrogram.

In this work the GSS score was used for the *Streptococcus* spp., comparing 108 strains belonging to 16 different species, compared the resulting dendrogram against a 16S rRNA gene phylogenetic reconstruction. Secondly, a core genome was built with the 108 strains and assess their conservancy regarding sequence identity, and measure how much sequence diversity is residing in the core genome of *Streptococcus* spp. Additionally, the core genome was used to discriminate between closely related strains in metagenomic sequences of highly *Streptococcus* dominated environments like the human mouth, where strains of the very same genus are differential for causing caries or health status (Belda-Ferre et al., 2012; Alcaraz et al., 2012; López-López et al., 2017).

Methods

Analyzed genomes and ortholog mapping.

Predicted proteomes for 108 selected *Streptococcus* spp., 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.

Genomic Similarity Score (GSS)

The GSS was conducted 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, then normalized the bit-score maximum values to a self-comparison of each proteome. Values of GSS have a range from 0-1, and GSS formula is calculated in the following form:

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

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 tree of life 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 one of the 108 streptococci genomes. The multiple alignments for 16S rRNA gene were done using structural RNA information using the software ssu-align (v0.1) (Nawrocki, 2009). The resulting 16S rRNA phylogeny was plotted by Neighbor-Joining method using MEGA 5.2 (Tamura et al., 2013). GSS calculations 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 were used: *S. agalactiae* 2-22 (FO393392; 1548 proteins). From the RBH calculations, results were compared, and the union set of proteins for all the 108 streptococci are defined as the core genome. From the local alignments from RBH comparisons global alignments were performed using Needleman-Wunsch method implemented in needleall of EMBOSS suite (Rice, Longden

& Bleasby, 2000), global alignments were used to calculate global sequence identity from each core genome predicted protein.

Pan-genome

A non-redundant pan-genome of the *Streptococcus* genus was calculated using concatenating all the predicted proteins of each analyzed strain (Supplemental Information 2) and then parsing the result to cd-hit (Huang et al., 2010) clustering using an identity cut-off value of 70% to build protein families.

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). A minimum length of 15 amino acids and a minimum identity of 60% were required. Sequences 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 in 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 against reference core genome for each streptococci species using Nucmer and Promer from the Mummer suite (Marçais et al., 2018). A cut-off value of 90% identity (amino acid) was the choice for identifying each metagenomic read and then assign it to individual species.

Results

Phylogenetic and genome similarity of the *Streptococcus* genus.

A reference phylogenetic reconstruction was done as a reference for our study 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*.

Genomic similarity score (GSS) dendrogram shows the same clades using 16S rRNA (Fig. 1B). However, it rearranges the Pyogenic group, where *S. agalactiae* which is included in the Pyogenic in the 16S phylogeny, and GSS shows it as the basal group for the Pyogenic clade. Another rearrangement of GSS when comparing is the Suis group 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 GSS dendrogram distances are longer enough to distinguish discrete groups among close related strains like is visible for inner clades of Suis, Pyogenic, Mutans, and Mitis groups. Remarkably, resolved clades are formed in GSS dendrogram for stains of *S. pneumoniae* and *S. pseudopneumoniae* whereas 16S does not allow to distinguish inner relationships, showing polytomies. Also, Suis GSS group shows clear resolved branches when comparing to 16S rRNA phylogeny.

Core genome sequence diversity

Our streptococci core genome has 404 proteins shared by all the 108 analyzed strains. It is a relatively small number when comparing to the genus average protein content that is 1,929 in each strain, the core then represents one-fifth of the average predicted proteome for each strain, and 33,039 protein families compose the total pan-genome of the streptococci at 70% identity (Supplemental Information 3). Paired global alignments were performed to understand the pairwise identity of each compared protein and how the identity varies within the core genome (Fig. 2). The identity conservation is probably showing evidence for selective constraints even within the core genome (Supplemental

Information 4). The individual proteins composing the core genome were plotted showing the pairwise identity of the alignments between a reference sequence where *S. pyogenes* was chosen as the reference because of its top phylogenetic position both in 16S and in GSS dendrogram (Fig. 2). Identity of the predicted proteins, of the core genome, diminish along species increase their genomic distances (GSS), sorting the proteins by their identity level allowed us to find out that the identity ranges are enormous with distances spanning from 100% to less than 25% identity for the global alignment. Note that the similarity percentage for amino acid substitutions were not included (like changing a polar amino acid for another one), global alignments are used as a refinement for calculating sequence identity in a precise and exhaustive way to refine the initial blast local alignment strategy. Based on the level of sequence diversity of the pairwise alignments, alignment of a core protein sequences with high identity (>90%) is proposed and could be used to discriminate between streptococci of close related strains in environmental shotgun sequencing samples. Core genomes for each of the streptococci species described here are available for the community (Supplemental Information 5).

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 families are related to the translational machinery, including ribosomal proteins and translation-related proteins (Z=3.08 core; Z=0.88 pan-genome). Cell division related proteins are better represented 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 core genome predicted proteins with high group identity (>90%) are mostly related to the translation process, and the top 10 are exclusively ribosomal proteins (Supplemental Information 4). As identity decrease, 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 4). Hypothetical proteins in the core proteome are abundant (48 out of 404; 11.81%).

Using the core genome to scan oral metagenomes

Using the core proteome relative abundance estimations for each *Streptococcus* species in the oral microbiome were performed. Oral metagenomes were chosen because of the many streptococci with high abundance (4 to >20%) on them (Supplemental Information 5). Two oral metagenomes were chosen: a patient with active caries and a healthy adult that have 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 clearly depleted (17 metagenomic fragments) in the healthy patient (NOCA_01) and highly represented (127 metagenomic fragments) in the patient with caries. Sorting the number of fragmented metagenomic sequences aligned against each reference metagenome and filtering them with high identity levels (≥ 90%) shows that is possible to generate strain-specific profiles (Table 1).

Discussion

Hosting multiple pathogenic strains clinical criteria like their hemolysis capabilities have historically classified *Streptococcu*s, and through their cell wall antigenic properties (Kayser, Bienz & Eckert, 2011). Molecular phylogenetics has aided streptococci classification (Kawamura et al., 1995; Kilian et al., 2008). The streptococci have been the beginning for interesting comparative genomics studies, genomic variability within the same species in detail started with the definition of relevant concepts like pan-genome and core genome when sequencing and comparing the genomes of strains further than the accepted reference in *S. agalactiae (Tettelin et al., 2005)*.

The core genome is an ever-changing concept; if more genomes are added into the comparison, the union set will be lower each time. In this work, information about 404 coding genes of the core genome, done with 108 strains compared is presented. To support our statements, the first core genome for the group was 611 genes comparing 26 genomes (Lefébure & Stanhope, 2007); a second effort is about 547 genes using 64 genomes (Van den Bogert et al., 2013); a third reconstruction gave 369 core predicted

proteins in their 138 selected strains (Gao et al., 2014). Additionally, core genome allows us to have a shared set of genes between multiple species, and it is possible to detail about the metabolic profile they are coding for. Interestingly, 11.81% of the core genes of streptococci are of unknown function (Supplemental Information 4), with sequence diversity, and represent an opportunity to test them as therapeutic targets.

Here, a catalog of predicted proteins which were evaluated for their degree of similarity is provided, and then used as a seed for searching particular strains into metagenomic samples. Additionally, we think that traditional phylogenetic methodology is necessary to understand a vertical group evolution. However, the bacteria have amazing capabilities of natural moving of genes through conjugation, transformation, and competence, with high rates of recombination, pose a challenge for traditional phylogenetics (Frost et al., 2005; Francino & Pilar Francino, 2012). Pan-genomic variability gives the chance to adapt to particular environments through slight additions or deletions into the genomic repertoire (Tettelin et al., 2008; Mira et al., 2010; Vernikos et al., 2015). The GSS is trying to get insights into bacteria strains similarity considering all the possible amount of homologous genetic information shared by pairs of bacteria, no matter if it is vertical or horizontal transmitted and it translates into overall similarity and this approach has been used previously (Janga & Moreno-Hagelsieb, 2004; Moreno-Hagelsieb & Janga, 2007; Alcaraz et al., 2010; Moreno-Hagelsieb et al., 2013). The main advantage of GSS is that uses both core and pan-genomic information to estimate relatedness between strains.

In this work, it was possible to infer a GSS dendrogram that resembles the primary literature accepted groups of streptococci. GSS shows it strength in resolving strain relatedness if comparing clade structure and distances when compared to 16S rRNA gene phylogeny (Fig. 1). In the 16S rRNA phylogeny, *S. mutans* and *S. equi* have noticeable long branches when comparing to the rest of the species, the 16S resolution does not allow us to distinguish differences between *S. mutans* nor *S. equi.* When observing the same groups in GSS dendrogram, it is possible to distinguish clusters and noticeable distances for species like *S. mutans* and *S. equi* (Fig. 1B). Within group resolution is greatly improved 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 and with clear branching and long enough distances to identify each strain within a species.

The expansive growing of metagenomic and metatranscriptomic data needs to have a framework to distinguish between closely related strains. Some environments host intra-genus diversity with implications for health like in the case for human vaginal microbiomes extensively 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 relative simple (404 genes) dataset were it is possible to align all the metagenomic information (reads, contigs) to the references and estimate species abundances based in the coverage and identity of each aligned fragment (Fig. 3). Despite the biological relevance, or connecting it to essential genes (Goodall et al., 2017), the core genome of a group provides a working tool to discriminate between closely related strains. Nonetheless, it is important to understand sequence identity variation within core genome, providing a basis for differential selective level for each predicted protein even within genus shared genes (Supplemental information 5). Understanding core genome variations, could be fully 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 phylogenetic tools in bacteria are invaluable, and the community will keep using them. However, they do not get the dynamism occurring in bacteria genomes and other tools like the GSS allow us to distinguish genome level relatedness between strains, even between closely related ones. Incorporating all pair of pan-genomic homologous proteins pairs into the comparisons no matter their evolutionary origin is a strength of comparisons like GSS. A practical use for the core genome of the streptococci is shown to classify abundances of different species and strains into metagenomic samples. Finally, we provide the community the range of sequence diversity for the *Streptococcus* core proteins, which is impressive and will need further analysis to define if the range of sequence identity correlates with selective pressures for core genes.

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 Prof. Víctor Valdés-López, Prof. Luisa A. Alba-Lois, Prof. Claudia Segal, and Viviana Escobar for their kindly support that made this work possible.

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.

363 Alcaraz LD., Moreno-Hagelsieb G., Eguiarte LE., Souza V., Herrera-Estrella L., Olmedo G. 2010.
364 Understanding the evolutionary relationships and major traits of Bacillus through
365 comparative genomics. *BMC genomics* 11:332. DOI: 10.1186/1471-2164-11-332.
366 Belda-Ferre P., Alcaraz LD., Cabrera-Rubio R., Romero H., Simón-Soro A., Pignatelli M., Mira A.
367 2012. The oral metagenome in health and disease. *The ISME journal* 6:46–56. DOI:
368 10.1038/ismej.2011.85.
369 Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009.
370 BLAST : architecture and applications. *BMC bioinformatics* 10:421. DOI: 10.1186/1471-
371 2105-10-421.
372 Cardenas E., Cole JR., Tiedje JM., Park J-H. 2009. Microbial Community Analysis using RDP II
373 (Ribosomal Database Project II):Methods, Tools and New Advances. *Environmental*
374 *Engineering Research* 14:3–9. DOI: 10.4491/eer.2009.14.1.003.
375 Ciccarelli FD. 2006. Toward Automatic Reconstruction of a Highly Resolved Tree of Life.
376 *Science* 311:1283–1287. DOI: 10.1126/science.1123061.
377 Contreras-Moreira B., Vinuesa P. 2013. GET_HOMOLOGUES, a Versatile Software Package for
378 Scalable and Robust Microbial Pangenome Analysis. *Applied and environmental*
379 *microbiology* 79:7696–7701. DOI: 10.1128/aem.02411-13.
380 Fox GE., Wisotzkey JD., Jurtshuk P Jr. 1992. How close is close: 16S rRNA sequence identity
381 may not be sufficient to guarantee species identity. *International journal of systematic*
382 *bacteriology* 42:166–170. DOI: 10.1099/00207713-42-1-166.
383 Francino MP., Pilar Francino M. 2012. The Ecology of Bacterial Genes and the Survival of the
384 New. *International journal of evolutionary biology* 2012:1–14. DOI:
385 10.1155/2012/394026.
386 Fraser C., Alm EJ., Polz MF., Spratt BG., Hanage WP. 2009. The bacterial species challenge:
387 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. *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.

Jolley KA., Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics* 11:595. DOI: 10.1186/1471-2105-11-595.

Kawamura Y., Hou XG., Sultana F., Miura H., Ezaki T. 1995. Determination of 16S rRNA sequences of Streptococcus mitis and Streptococcus gordonii and phylogenetic

413 relationships among members of the genus Streptococcus. *International journal of*
414 *systematic bacteriology* 45:406–408. DOI: 10.1099/00207713-45-2-406.
415 Kayser FH., Bienz KA., Eckert J. 2011. *Medical Microbiology*. Thieme.
416 Kerepesi C., Bánky D., Grolmusz V. 2014. AmphoraNet: the webserver implementation of the
417 AMPHORA2 metagenomic workflow suite. *Gene* 533:538–540. DOI:
418 10.1016/j.gene.2013.10.015.
419 Kilian M., Poulsen K., Blomqvist T., Håvarstein LS., Bek-Thomsen M., Tettelin H., Sørensen
420 UBS. 2008. Evolution of Streptococcus pneumoniae and its close commensal relatives.
421 *PloS one* 3:e2683. DOI: 10.1371/journal.pone.0002683.
422 Lefébure T., Stanhope MJ. 2007. Evolution of the core and pan-genome of Streptococcus:
423 positive selection, recombination, and genome composition. *Genome biology* 8:R71. DOI:
424 10.1186/gb-2007-8-5-r71.
425 Liolios K., Chen I-MA., Mavromatis K., Tavernarakis N., Hugenholtz P., Markowitz VM., Kyrpides
426 NC. 2010. The Genomes On Line Database (GOLD) in 2009: status of genomic and
427 metagenomic projects and their associated metadata. *Nucleic acids research* 38:D346–54.
428 DOI: 10.1093/nar/gkp848.
429 López-López A., Camelo-Castillo A., Ferrer MD., Simon-Soro Á., Mira A. 2017. Health-
430 Associated Niche Inhabitants as Oral Probiotics: The Case of Streptococcus dentisani.
431 *Frontiers in microbiology* 8. DOI: 10.3389/fmicb.2017.00379.
432 Marçais G., Delcher AL., Phillippy AM., Coston R., Salzberg SL., Zimin A. 2018. MUMmer4: A
433 fast and versatile genome alignment system. *PLoS computational biology* 14:e1005944.
434 DOI: 10.1371/journal.pcbi.1005944.
435 Meyer F., Bagchi S., Chaterji S., Gerlach W., Grama A., Harrison T., Paczian T., Trimble WL.,
436 Wilke A. 2017. MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-
437 throughput metagenome analysis. *Briefings in bioinformatics*. DOI: 10.1093/bib/bbx105.

438 Mira A., Martín-Cuadrado AB., D'Auria G., Rodríguez-Valera F. 2010. The bacterial pan-
439 genome:a new paradigm in microbiology. *International microbiology: the official journal of*
440 *the Spanish Society for Microbiology* 13:45–57. DOI: 10.2436/20.1501.01.110.
441 Moreno-Hagelsieb G., Janga SC. 2007. Operons and the effect of genome redundancy in
442 deciphering functional relationships using phylogenetic profiles. *Proteins: Structure,*
443 *Function, and Bioinformatics* 70:344–352. DOI: 10.1002/prot.21564.
444 Moreno-Hagelsieb G., Latimer K. 2008. Choosing BLAST options for better detection of
445 orthologs as reciprocal best hits. *Bioinformatics* 24:319–324. DOI:
446 10.1093/bioinformatics/btm585.
447 Moreno-Hagelsieb G., Wang Z., Walsh S., ElSherbiny A. 2013. Phylogenomic clustering for
448 selecting non-redundant genomes for comparative genomics. *Bioinformatics* 29:947–949.
449 DOI: 10.1093/bioinformatics/btt064.
450 Nawrocki EP. 2009. Structural RNA Homology Search and Alignment Using Covariance Models.
451 PhD Thesis. Washington University.
452 Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
453 language. *Bioinformatics* 20:289–290. DOI: 10.1093/bioinformatics/btg412.
454 Pruesse E., Quast C., Knittel K., Fuchs BM., Ludwig W., Peplies J., Glockner FO. 2007. SILVA: a
455 comprehensive online resource for quality checked and aligned ribosomal RNA sequence
456 data compatible with ARB. *Nucleic acids research* 35:7188–7196. DOI:
457 10.1093/nar/gkm864.
458 R Development Core Team. 2003. *The R Reference Manual: Base Package*. Network Theory.
459 Rice P., Longden I., Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software
460 Suite. *Trends in genetics: TIG* 16:276–277.
461 Rusch DB., Halpern AL., Sutton G., Heidelberg KB., Williamson S., Yooseph S., Wu D., Eisen
462 JA., Hoffman JM., Remington K., Beeson K., Tran B., Smith H., Baden-Tillson H., Stewart
463 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 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.

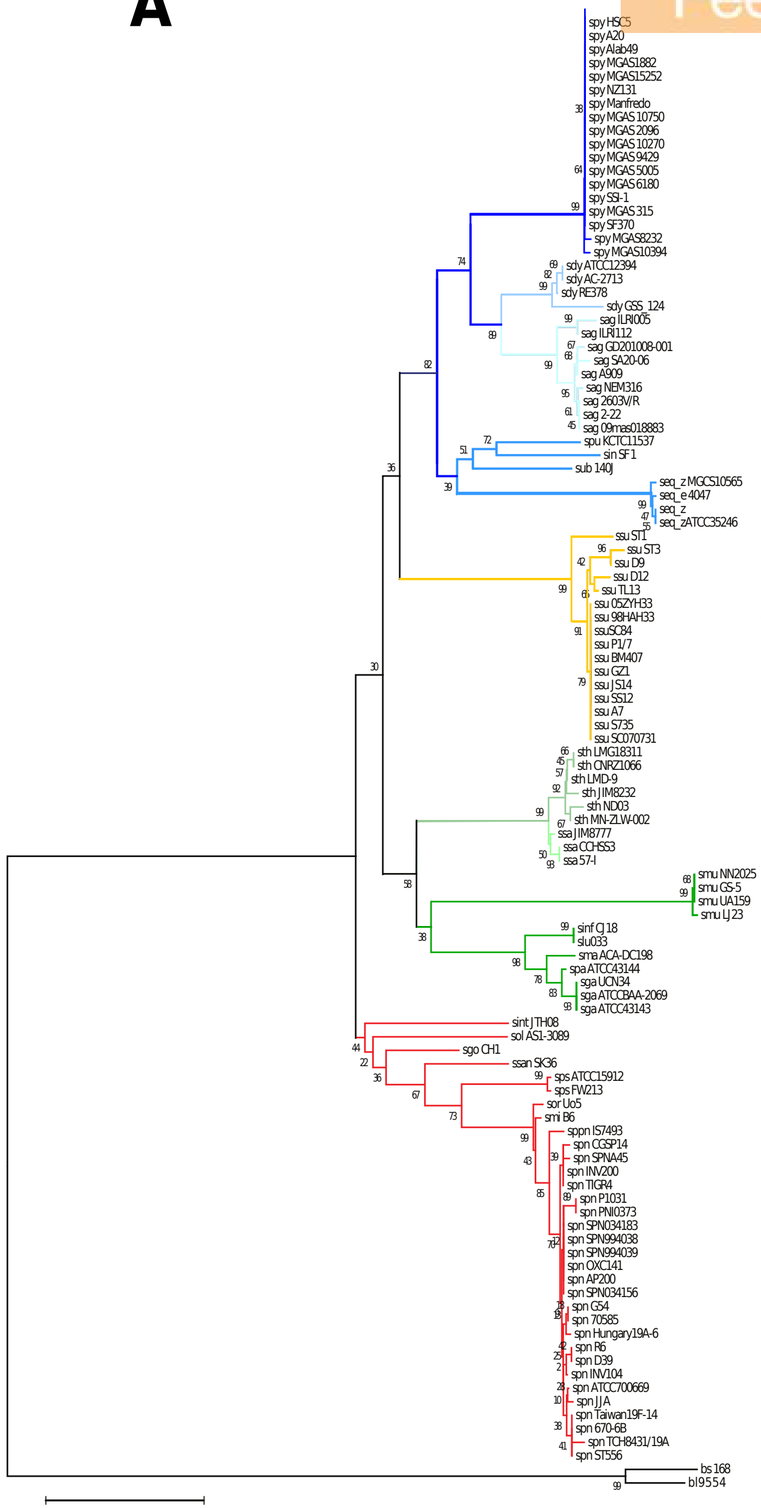
Wilke A., Harrison T., Wilkening J., Field D., Glass EM., Kyrpides N., Mavrommatis K., Meyer F. 2012. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC bioinformatics* 13:141. DOI: 10.1186/1471-2105-13-141.

Figure 1(on next page)

Streptococcus genus phylogenetic reconstruction and Genomic Similarity Score (GSS) dendrograms.

(A) Neighbor-Joining 16S rRNA reconstruction, with 1,000 bootstraps. (B) Genomic similarity score (GSS) dendrogram. Some of the major paraphyletic groups of streptococci due to clinical or practical uses (Killian, 2007) are Pyogenic, Suis, Salivarius, Mutans, and Mitis. Abbreviations of the tree are indicated: *spy=S. pyogenes, sdy=S.dysgalactiae, sag=S. agalactiae, spu=S. parauberis, sin=S. iniae, sub=S.uberis, seq_z=S.equi subsp. zooepidemicus, seq_z=S.equi subsp. equi, ssu=S. suis, sth=S.thermophilus, ssa=S.salivarius, smu=S. mutans, sint=S. intermedius, sol=S. oligofermentans, ssan=S. sanguinis, sgo=S. gordonii, sps=S. parasanguinis, spas=S. pasteurianus, sor=S. oralis, spn=S. pneumoniae, sppn=S. pseudopneumoniae, smi=S. mitis, sga=S. gallolyticus, sma=S. macedonicus, slu=S. lutetiensis, sinf=S. infantarius, bs=B. subtilis, and bl=B. licheniformis.*

A



Pyogenic

Suis

Salivarius

Mutans

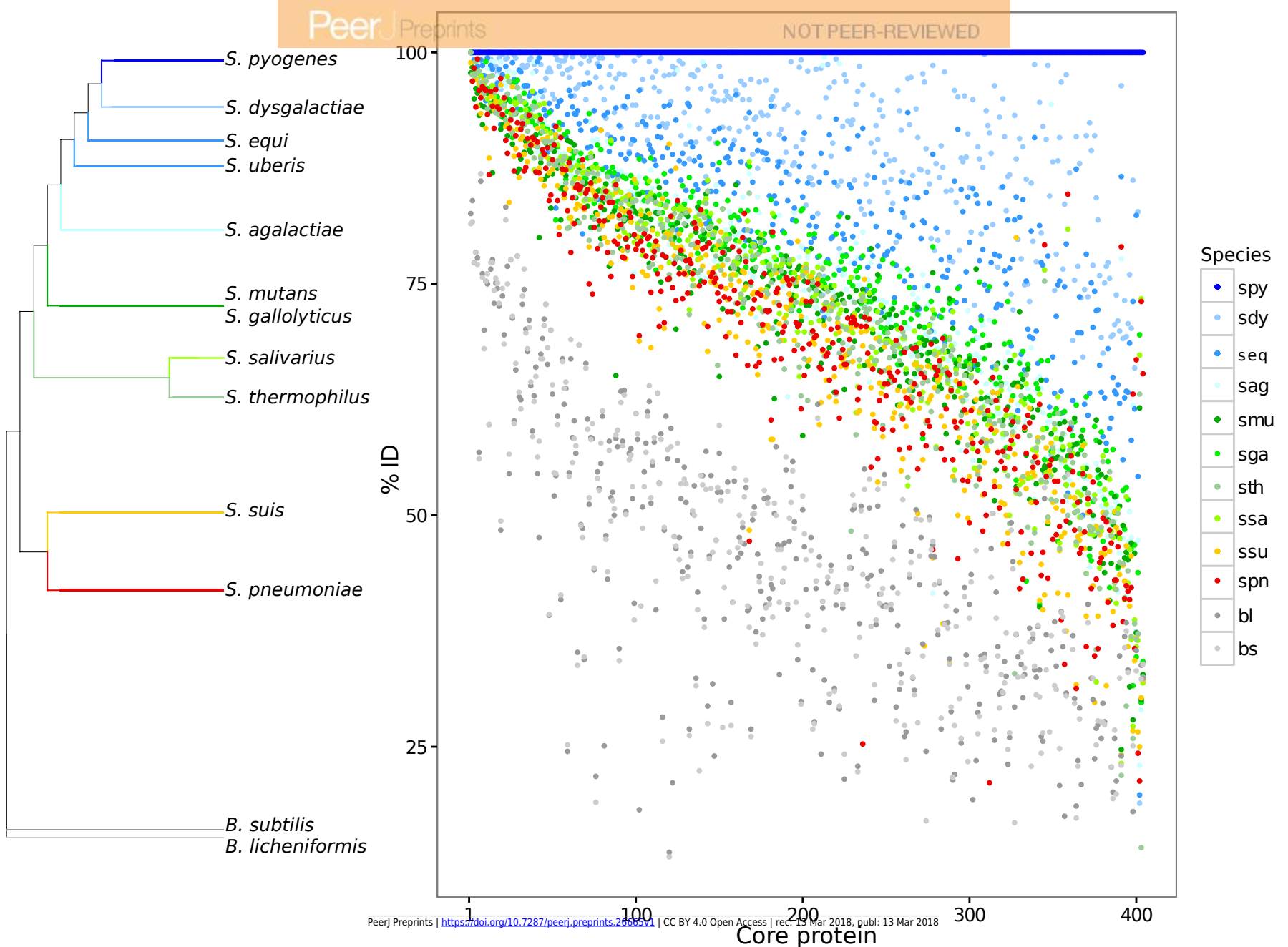
Mitis

B

Figure 2(on next page)

Core genome variability amongst different streptococci clades.

Each streptococci core gene is plotted against the S. pyogenes core genome, the pairwise global protein sequence alignment identity is plotted and ordered from the higher identity to the lowest. Outgroups of Bacillus are used as lower boundary identity limits. Identity values increases parallel to GSS distances (left pane). Abbreviations: *spy=S. pyogenes, sdy=S.dysgalactiae, sag=S. agalactiae, ssu=S. suis, sth=S.thermophilus, ssa=S.salivarius, smu=S. mutans, spn=S. pneumoniae, sga=S. gallolyticus, seq=S. equi, bs=B. subtilis, and bl=B. licheniformis.*

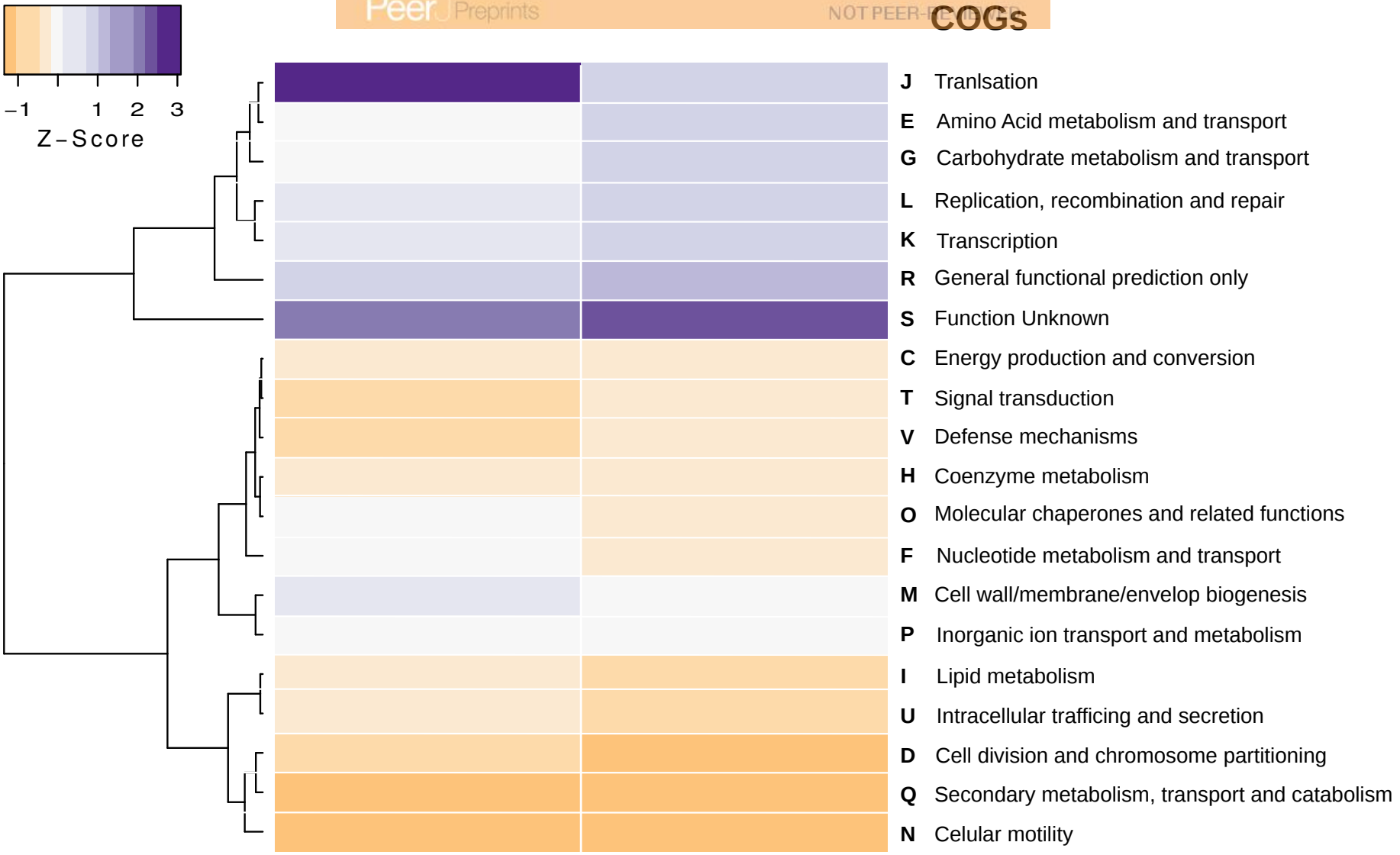


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2668v1 | CC BY 4.0 Open Access | rec: 13 Mar 2018, publ: 13 Mar 2018

Figure 3(on next page)

Streptococcus core and pan-genome summary of general functions profiles according to the Cluster of Orthologous Groups.

Complete annotation is available in Supplementary Information 2.

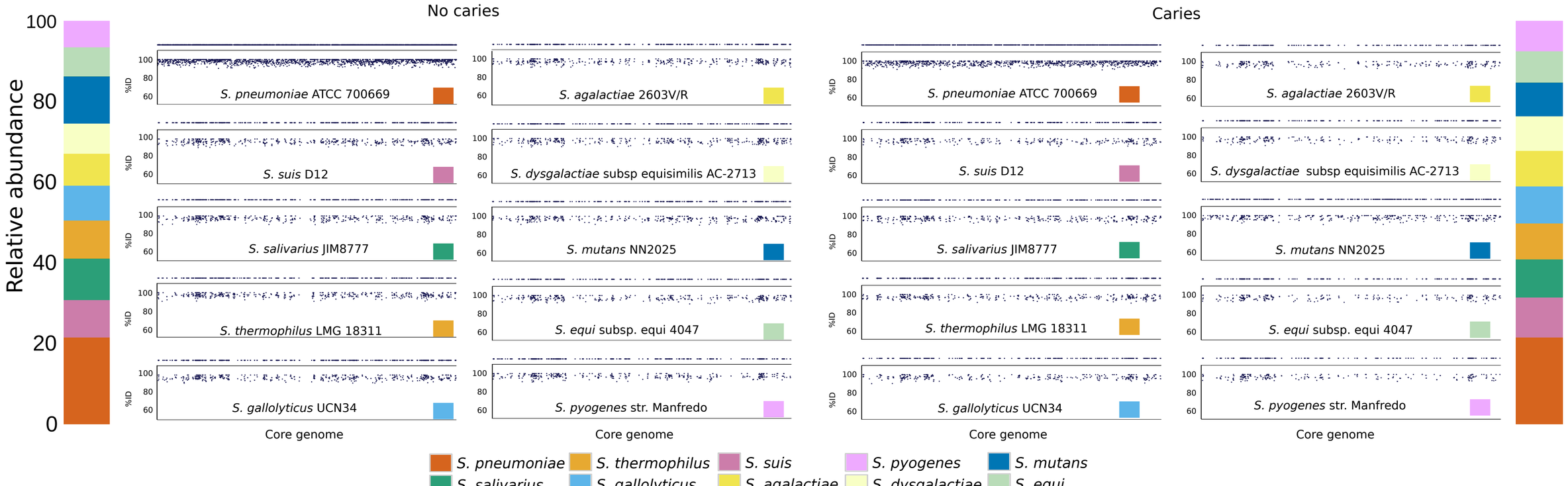


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v1 CC BY 4.0 Open Access | rec: 13 Mar 2018, publ: 13 Mar 2018

Figure 4(on next page)

Metagenomic fragment recruitment against Streptococci core genomes.

Metagenomic reads from a healthy (right pane) and diseased individual (dental caries; left pane) were aligned against the core genomes of 10 different species of Streptococci. Left and right bar plots indicate the species gene relative abundance in each metagenome.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26665v1 | CC BY 4.0 Open Access | rec: 13 Mar 2018, publ: 13 Mar 2018

Table 1(on next page)

Promer metagenomic recruitments against core genomes.

The number of metagenomic contigs recruited and in parenthesis the number of core genes aligned. NOCA is no caries patient, CA is a patient with active caries (Belda-Ferre et al. 2012).

Tables

Table 1. Promer metagenomic recruitments against core genomes. The number of metagenomic contigs recruited and in parenthesis the number of core genes aligned. NOCA is no caries patient, CA is a patient with active caries (Belda-Ferre et al. 2012).

	Metagenomic recruitments	
Species	NOCA_01	CA_04P
S. agalactiae	42 (24)	31 (20)
S. thermophilus	67 (32)	75 (42)
S. pyogenes	40 (24)	34 (19)
S. pneumoniae	867 (329)	418 (221)
S. equii	18 (10)	18 (12)
S. gallolyticus	54 (31)	43 (25)
S. mutans	17 (13)	127 (109)
S. salivarius	77 (33)	87 (45)
S. suis	60 (30)	49 (25)
S. dysgalactiae	40 (24)	36 (22)