Genome Biology and Evolution
SMBE
GBE

Phylogenomics and the Dynamic Genome Evolution of the Genus *Streptococcus*

Vincent P. Richards[1], Sara R. Palmer[2], Paulina D. Pavinski Bitar[1], Xiang Qin[3], George M. Weinstock[4], Sarah K. Highlander[3,5], Christopher D. Town[6], Robert A. Burne[2], and Michael J. Stanhope[1,*]

[1]Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University

[2]Department of Oral Biology, University of Florida

[3]Human Genome Sequencing Center, Baylor College of Medicine

[4]The Genome Institute, Washington University in St. Louis

[5]Department of Molecular Virology and Microbiology, Baylor College of Medicine

[6]J. Craig Venter Institute, Rockville, MD

*Corresponding author: E-mail: mjs297@cornell.edu.

Accepted: March 4, 2014

Data deposition: Whole-genome shotgun projects have been deposited at DDBJ/EMBL/GenBank under the accessions AEUV00000000, AEUX00000000, AEUW00000000, AEUU00000000, AEUY00000000, AEUZ00000000, AWEX00000000, and AMOO00000000. The versions described in this paper are AEUV02000000, AEUX02000000, AEUW02000000, AEUU01000000, AEUY02000000, AEUZ02000000, AWEX01000000, and AMOO01000000.

Abstract

The genus *Streptococcus* comprises important pathogens that have a severe impact on human health and are responsible for substantial economic losses to agriculture. Here, we utilize 46 *Streptococcus* genome sequences (44 species), including eight species sequenced here, to provide the first genomic level insight into the evolutionary history and genetic basis underlying the functional diversity of all major groups of this genus. Gene gain/loss analysis revealed a dynamic pattern of genome evolution characterized by an initial period of gene gain followed by a period of loss, as the major groups within the genus diversified. This was followed by a period of genome expansion associated with the origins of the present extant species. The pattern is concordant with an emerging view that genomes evolve through a dynamic process of expansion and streamlining. A large proportion of the pan-genome has experienced lateral gene transfer (LGT) with causative factors, such as relatedness and shared environment, operating over different evolutionary scales. Multiple gene ontology terms were significantly enriched for each group, and mapping terms onto the phylogeny showed that those corresponding to genes born on branches leading to the major groups represented approximately one-fifth of those enriched. Furthermore, despite the extensive LGT, several biochemical characteristics have been retained since group formation, suggesting genomic cohesiveness through time, and that these characteristics may be fundamental to each group. For example, proteolysis: mitis group; urea metabolism: salivarius group; carbohydrate metabolism: pyogenic group; and transcription regulation: bovis group.

Key words: comparative genomics, phylogenetics, gene gain and loss, enrichment, lateral gene transfer.

Introduction

The genus *Streptococcus* comprises approximately 72 species of Gram-positive bacteria including numerous species that have a severe impact on human health, inflicting significant morbidity and mortality (Köhler 2007). In addition, several species are responsible for substantial economic losses to agriculture. For example, *Streptococcus pyogenes* (Group A *Streptococcus*; GAS) is among the top ten causes of human mortality due to infectious disease, inflicting a wide range of diseases that include necrotizing fasciitis, toxic shock syndrome, pharyngitis, impetigo, puerperal sepsis, scarlet fever, glomerulonephritis, and rheumatic fever (Carapetis et al. 2005; Ralph and Carapetis 2013). *Streptococcus pneumoniae*, despite being a common member of the normal microbial flora of the nose and throat, is the leading cause of bacterial disease worldwide, and in addition to common diseases such as otitis media, the pathogen is responsible for life-threatening sepsis, meningitis, and pneumonia (O'Brien and Nohynek

© The Author(s) 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

2003). Similarly, *S. agalactiae* (Group B *Streptococcus*) is a commensal of the genital and gastrointestinal tract yet can cause severe invasive disease in adults and neonates (e.g., pneumonia, meningitis, and septicemia) (Baker 2000; Balter et al. 2000; Dermer et al. 2004). *Streptococcus mutans,* another commensal, is implicated as a leading cause of tooth decay, and although not life threatening, the economic burden of treatment is substantial (Loesche 1986). Several *Streptococcus* species can cause bovine mastitis (e.g., *S. uberis*, *S. agalactiae*, *S. dysgalactiae* subsp. *dysgalactiae*, and *S. canis*), with *S. uberis* and *S. agalactiae* responsible for major economic loss to the dairy industry (Zadoks et al. 2011).

The species within the genus *Streptococcus* display a wide range of epidemiological and ecological characteristics. For example, many species are restricted to humans or a single animal host, for example, *Streptococcus equi* subsp. *equi* is restricted to horses, whereas *S. agalactiae* infects multiple hosts ranging from humans to teleosts. Some species are regarded as contagious (transmitted directly between hosts), whereas others are regarded as environmental (transmitted between the environment and host; for example, *S. uberis* can be transmitted from soil to cows). One species within the group (*S. thermophilus*) is nonpathogenic and is used extensively in the dairy industry (Price et al. 2011).

Early classification for *Streptococcus* was based primarily on hemolytic reaction and Lancefield group antigens, which divided the genus into two groups: the pyogenic and viridans (Sherman 1937). The pyogenic group was beta-hemolytic and isolated from a range of human and animal sources, whereas the viridans group was mostly alpha-hemolytic and isolated predominantly from the human oral cavity. Subsequent to this, Bentley et al. (1991) and Kawamura et al. (1995) utilized 16S rRNA sequences to divide the genus into six major groups. The pyogenic group remained, but viridans was split into five subgroups whose names reflected one of the species within each of the groups: anginosus, mitis, salivarius, bovis, and mutans. It should be noted, however, that neither of these studies attempted to provide phylogenetic support for these groupings. Subsequently, Facklam (2002) devised an identification scheme based on phenotypic characteristics that could delineate species into these same groups plus an additional one that he named sanguinis. Combining sequence data from 16S rRNA and the RNase P RNA gene (*rnpB*), Täpp et al. (2003) explored phylogenetic relationships among 50 *Streptococcus* species. Although they recovered all the major groups except sanguinis (this group was paraphyletic), relationships among the groups were poorly resolved. Here, we make use of 46 *Streptococcus* genome sequences (44 distinct species), including eight species for which genome sequences were previously unavailable, to provide the first genomic level insight into the evolutionary history of all the major groups of this genus, as well as the genetic basis underlying their functional diversity. Toward a better understanding of the evolution of this functional diversity, we also examine gene gain/loss events that occurred on all lineages through the course of their evolution.

Materials and Methods

Sequence Data

Details regarding the 46 genome sequences (47 including the outgroup, see below) used in our analyses are presented in table 1. Of these sequences, 39 were obtained directly from National Center for Biotechnology Information (NCBI). The remaining eight were sequenced to near completion as part of this study. The following six were sequenced using a combination of Roche 454 and Illumina sequencing technologies and assembled using Celera Assembler v6.1 (Myers et al. 2000): *S. criceti* (HS-6), *S. ictaluri* (707-05), *S. macacae* (NCTC 11558), *S. porcinus* (Jelinkova 176), *S. pseudoporcinus* (LQ 940-04), and *S. urinalis* (2285-97). *Streptococcus equi* subsp. *ruminatorum* (CECT 5772) was sequenced using Roche 454 technology and assembled using Newbler v2.3. All of these genome sequences were annotated using the Prokaryotic Genome Automated Annotation Pipeline at NCBI. *Streptococcus iniae* (9117) was sequenced using Roche 454 technology and assembled using Newbler v1.1. In the case of *S. iniae*, gene prediction and manually curated annotation were performed as described previously (Highlander et al. 2007). For two of these species, strains from different hosts were included: *S. agalactiae,* human and bovine and *S. parauberis*, bovine and flounder. The origin of *S. gordonii* (Challis substr CH1) is probably human blood or an endocarditis valve; however, this is unconfirmed (Vickerman M, personal communication). *Lactobacillus crispatus*, from the closely related nonpathogenic family *Lactobacillaceae*, was used as an outgroup (Price et al. 2011).

Gene Clustering and Phylogenetic Analyses

Amino acid sequences were delineated into clusters with putative shared homology using the Markov clustering (MCL) algorithm (van Dongen 2000) as implemented in the MCLBlastLINE pipeline (available at http://micans.org/mcl, last accessed March 20, 2014). Throughout the article, we refer to a set of gene sequences delineated by this method as an MCL gene cluster. The pipeline uses MCL to assign gene sequences to clusters with putative shared homology based on a Blastp search between all pairs of protein sequences using an *E*-value cut off of 1e-5. The MCL algorithm was implemented using an inflation parameter of 1.8. Simulations have shown this value to be generally robust to false positives and negatives (Brohee and van Helden 2006). Nucleotide sequences corresponding to each MCL gene cluster were aligned using Probalign v1.1 (Roshan and Livesay 2006).

For the phylogenetic analysis, we selected those MCL gene clusters that were shared among all taxa and contained only single gene copies for each taxon (the core set) ($n = 159$)

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

Table 1
Genome Sequence Details

Species	Source	Tissue/Presentation	Accession No.
Streptococcus agalactiae (A909)	Human	Neonate	NC_007432
S. agalactiae (FSL S3-026)	Bovine	Mastitis	AEXT01
S. anginosus (1 2 62CV)	Human	Rectal biopsy	ADME01
S. australis (ATCC 700641)	Human	Saliva	AEQR01
S. bovis (ATCC 700338)	Human	Synovial fluid	AEEL01
S. canis (FSL Z3-227)	Bovine	Mastitis	AIDX01
S. constellatus subsp. *pharyngis* (SK1060)	Human	Throat	AFUP01
S. criceti (HS-6)	Hamster	Caries lesion	AEUV02
S. cristatus ATCC 51100	Human	Periodontal abscess	AEVC01
S. downei (F0415)	Human	Oral cavity	AEKN01
S. dysgalactiae subsp. *equisimilis* (GGS 124)	Human	Toxic shock syndrome	NC_012891
S. dysgalactiae subsp. *dysgalactiae* (ATCC 27957)	Bovine	Mastitis	AEGO01
S. equi subsp. *equi* (4047)	Horse	Strangles	NC_012471
S. equi subsp. *ruminatorum* (CECT 5772)	Sheep	Mastitis	AWEX01
S. equi subsp. *zooepidemicus* (MGCS10565)	Human	Nephritis	NC_011134
S. equinus (ATCC 9812)	Human	Gastrointestinal tract	AEVB01
S. gallolyticus (UCN34)	Human	Blood	NC_013798
S. gordonii (Challis substr CH1 ATCC 35105)	Human?	Blood/endocarditis valve?	NC_009785
S. ictaluri (707-05)	Catfish		AEUX02
S. infantarius subsp. *infantarius* (ATCC BAA-102)	Human	Feces	ABJK01
S. infantis (ATCC 700779)	Human	Tooth surface and pharynx	AEVD01
S. iniae (9117)	Human	Blood	AMOO01
S. intermedius (F0413)	Human	Dental plaque	AFXO01
S. macacae (NCTC 11558)	Macacae	Dental plaque	AEUW02
S. macedonicus (ACA-DC 198)	Cheese		NC_016749
S. mitis (B6)	Human	Blood	NC_013853
S. mutans (UA159)	Human	Oral cavity	NC_004350
S. oralis (Uo5)	Human	Oral cavity	NC_015291
S. parasanguinis (ATCC 15912)	Human	Oral cavity	NC_015678
S. parauberis (KCTC 11537)	Flounder		NC_015558
S. parauberis (NCFD 2020)	Bovine	Mastitis	AEUT01
S. pasteurianus (ATCC 43144)	Human	Blood	NC_015600
S. peroris (ATCC 700780)	Human	Tooth surface and pharynx	AEVF01
S. pneumoniae (670 6B)	Human	Nasopharyngeal	NC_014498
S. porcinus (Jelinkova 176)	Swine	Hemorrhagic lymph nodes	AEUU01
S. pseudopneumoniae (IS7493)	Human	Sputum; HIV	NC_015875
S. pseudoporcinus (LQ 940-04)	Human	Female genitourinary tract	AEUY02
S. pyogenes (MGAS10394)	Human	Throat	NC_006086
S. ratti (FA-1)	Rat	Oral cavity	AJTZ01
S. salivarius (JIM8780)	Human	Blood	NC_015760
S. sanguinis (SK36)	Human	Dental plaque	NC_009009
S. suis (05ZYH33)	Human	Toxic shock syndrome	NC_009442
S. thermophilus (CNRZ1066)	Yogurt		NC_006449
S. uberis (0140J)	Bovine	Mastitis	NC_012004
S. urinalis (2285-97)	Human	Urine	AEUZ02
S. vestibularis (ATCC 49124)	Human	Oral cavity	AEVI01
Lactobacillus crispatus (ST1)	Chicken	Crop	NC_014106

(supplementary table S1, Supplementary Material online). Recombination for genes in these clusters was assessed using a combination of four methods: 1) Genetic Algorithm Recombination Detection (GARD), 2) the Pairwise Homoplasy Index (PHI), 3) the Neighbor Similarity Score (NSS), and 4) Maximum χ^2. GARD is a phylogenetic method that searches for gene segments with incongruent phylogenetic topologies (Kosakovsky Pond et al. 2006a). PHI and NSS are compatibility methods that examine pairs of sites for homoplasy (Jakobsen and Easteal 1996; Bruen et al. 2006). Maximum χ^2 is a

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

substitution distribution method that searches for significant clustering of substitutions at putative recombination break points (Maynard Smith 1992). The tests were implemented using GARD (Kosakovsky Pond et al. 2006b) and PhiPack (Bruen et al. 2006). We compared two data sets for subsequent analyses. In the first, MCL gene clusters showing evidence for recombination for all four methods were removed. In the second, gene clusters showing evidence for recombination for at least three of the four methods were removed. Using the first approach, 23 clusters (14.5%) were removed leaving 136 clusters. Using the second approach, 84 clusters (52.8%) were removed leaving 75 clusters. For each data set, rooted maximum likelihood (ML) phylogenies (gene trees) were constructed from each MCL gene cluster using PhyML v3.0 (Guindon et al. 2010). Then, a species tree based on the consensus of the gene trees was constructed using the Triple Construction Method as implemented in the program Triplec (Ewing et al. 2008). This procedure is based on the observation that the most probable three-taxon tree consistently matches the species tree (Degnan and Rosenberg 2006). The method searches all input trees for the most frequent of the three possible rooted triples for each set of three taxa. Once found, the set of rooted triples are joined to form the consensus tree using the quartet puzzling heuristic (Strimmer and vonHaeseler 1996). The method has been shown to outperform majority rule and greedy consensus methods (Degnan et al. 2009). The species tree built using 136 MCL gene clusters is shown in figure 1, and the species tree built using 75 clusters is shown in supplementary figure S1, Supplementary Material online. In general, the two trees shared the same topology. However, a notable discrepancy was that *S. agalactiae* (a pyogenic species) clustered with the bovis species group in the 75-cluster species tree. This grouping was supported by 52.9% of the gene trees. The 75-cluster species tree also showed more conflict among the gene trees. Consequently, given the questionable placement of *S. agalactiae* in this tree coupled with its higher gene tree conflict, we elected to use the 136-cluster data set for subsequent analyses. The performance of gene tree consensus approaches, in particular the triple consensus approach, improves with the addition of more gene trees (Ewing et al. 2008). Consequently, it appears that for our data set, the benefit of adding additional genes, likely outweighed the possible confounding effect of an accompanying increase in recombination.

We utilized a second phylogenetic approach, gene alignment concatenation, to construct a species tree. This approach has the benefit of providing branch lengths. Specifically, the 136 MCL gene cluster alignments were concatenated and all invariant sites removed. The resulting alignment (68,803 bp) was used to build an ML phylogeny using GARLI v2.0 (Zwickl 2006). The search was performed using the GTR + G substitution model, which was determined to be the best fit for the data using the Akaike Information Criterion in MODELTEST (Posada and Crandall 1998). GARLI's auto termination option was set, allowing it to run until no significant improvement in topology was attained. Three search replicates were performed. Branch support was provided by generating 200 bootstrap replicates.

Gene Gain and Loss

We assessed gene gain/loss on the species tree using the parsimony based gene tree species tree reconciliation approach implemented in the program AnGST (David and Alm 2011). The reconciliation is obtained by inferring a minimum set of the following evolutionary events: gene loss, gene duplication, speciation, lateral gene transfer (LGT), and gene birth or genesis. We constrained gene transfer events to only occur between contemporaneous lineages on the phylogeny. To enforce this option in AnGST, the species tree required branch lengths scaled in units of time. We therefore rescaled the branch lengths on the concatenation species tree to units of time for this analysis. This rescaling was performed using a semiparametric method based on the penalized likelihood method of Sanderson (2002) as implemented in the chronopl function within the ape R package (Paradis et al. 2004). To obtain more accurate estimates for the ingroups, the outgroup and its long connecting branch was removed (Magallon and Sanderson 2005). The optimum value for the likelihood smoothing parameter (lambda) was determined using cross validation of lambda values from 10^{-4} to 10^{6}. Although Bayesian phylogenetic approaches might produce more accurate estimates of time scaled branch lengths, the penalized likelihood method was deemed suitable given the high computational demands a Bayesian approach would likely require to attain convergence on this amount of data and that precise dating of nodes in the phylogeny was not our objective here.

Gene trees for all MCL gene clusters containing three or more genes were constructed using PhyML v3.0 with the GTR + I + G substitution model. AnGST can account for gene tree phylogenetic uncertainty by creating a "chimeric" gene tree via the amalgamation of bootstrap replicates. We utilized this option by providing AnGST with 500 bootstrap replicates for every gene tree that contained three or more taxa. One MCL gene cluster for sequences annotated as an ATP-binding cassette (ABC) transporter protein contained a particularly large number of sequences (1,087). Phylogenetic analysis of this cluster was mostly unresolved, resulting in a large polytomy. This cluster was excluded from the AnGST analysis. For MCL gene clusters containing two genes (doublets), a gene could be seen either once in two separate taxa or twice within one taxa. For these genes, a gene tree is uninformative. Therefore, we followed Kamneva et al. (2012) and attempted to explain the evolutionary history of these genes without a gene tree by overlaying the position of the two genes onto the species tree and using the same set of evolutionary events and event penalties (see later) used in

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

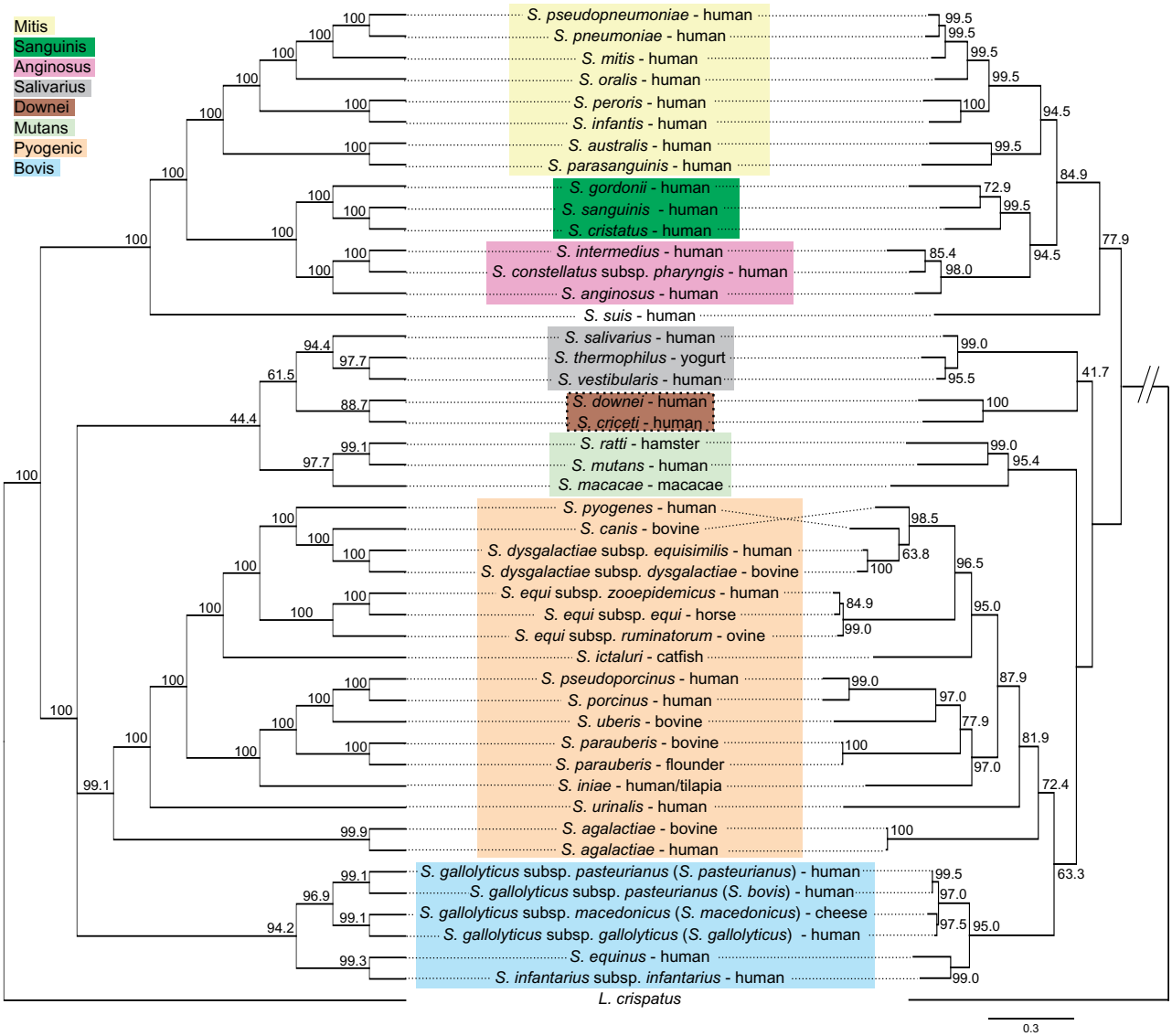

Fig. 1.—Phylogenies derived from a core set of 136 genes. Left: Consensus of the phylogenetic signal from each gene (numbers on branches show the proportion of genes that support a particular grouping). Right: ML phylogeny derived from a concatenation of the genes (numbers on branches show bootstrap support). Each of the major eight groups is color shaded. The putative downei group is shown with a dashed line. Previous nomenclature for species within the bovis group is shown within parentheses.

the reconciliation procedure to calculate the most parsimonious evolutionary scenario of gene birth followed by vertical transmission. LGT was inferred if the number of losses resulted in a higher penalty for loss than LGT. In this scenario, we counted one birth and one LGT. However, the direction of this transfer was unknown. If the two genes were from the same genome, we counted one birth and one duplication. It should be noted, however, that gene gain/loss analyses are limited to the genomes included and that some births may be the result of a LGT from a genome not included in the analysis.

AnGST event penalties were first determined by David and Alm (2011) using an approach that selected the combination of penalties that minimized the average change in genome size between ancestor and descendant within the species tree (genome flux). Specifically, their approach fixed the loss penalty to 1.0 and then ran multiple reconciliations adjusting LGT and duplication penalties. For a wide range of eukaryote, archaeal, and bacterial taxa, they found that a LGT penalty of 3.0 and a duplication penalty of 2.0 minimized genome flux. More recently, Kamneva et al. (2012) determined event penalties using the same approach for a data set containing only

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

bacteria from the Planctomycetes, Verrucomicrobia, and Chlamydiae (PVC) phyla. Their penalties were LGT = 5.0 and duplication = 3.0. Both studies showed that the LGT penalty had the strongest effect on genome flux. We compared reconciliations using both sets of penalties described earlier and also a reconciliation using equal penalties (LGT = 1.0, duplication = 1.0).

Hierarchal Clustering and Enrichment

Hierarchal clustering among genomes using presence/absence of MCL gene clusters was performed using the complete linkage method and binary distances as implemented in the R package pvclust (Suzuki and Shimodaira 2006). Support for groupings was obtained by calculating approximately unbiased *P* values using 500 bootstrap replicates.

Gene Ontology (GO) terms were assigned to all *Streptococcus* genomes using Blast2GO v.2.5.0 (Gotz et al. 2008). Relative enrichment (overrepresentation) of GO terms among lineages was assessed using Fisher exact tests. The test was performed using the Gossip statistical package (Blüthgen et al. 2005) implemented within Blast2GO. The false-discovery rate procedure of Benjamini and Hochberg (1995) was used to correct for multiple hypothesis testing (FDR = 0.05).

Results and Discussion

Genome Sequencing

The number of contigs, genome length, number of CDS, rRNAs, tRNAs, and %GC for each of the eight genomes sequenced as part of this study are shown in table 2. With the exception of *S. equi* subsp. *ruminatorum* (CECT 5772) (133 contigs), these genomes were assembled to a high level of contiguity (average contig number = 3, range = 1–8), generally consistent with bacterial genomes categorized as "noncontiguous finished" (Chain et al. 2009).

Phylogenetic Relationships

In general, the consensus and concatenation species trees were well supported and concordant (fig. 1). Both trees recovered the pyogenic, bovis, salivarius, and anginosus groups described in previous phylogenetic studies (Bentley et al. 1991; Kawamura et al. 1995; Täpp et al. 2003) and also showed the sanguinis group to be monophyletic. All these groups were well supported. Recognizing the sanguinis group as a distinct monophyletic entity subsequently renders the eight mitis species a monophyletic grouping that is also well supported. *Streptococcus parasanguinis* fell within the mitis group and not within the sanguinis group with *S. sanguinis*, confirming previous studies that showed these two species not to be sister taxa (Bentley et al. 1991; Kawamura et al. 1999; Täpp et al. 2003). Monophyly for the five species previously included in the mutans group, however, was not well supported. Although *S. mutans*, *S. ratti*, and *S. macacae* all formed a well-supported clade, *S. downei* and *S. criceti*, which also grouped tightly, clustered instead with the three salivarius species. However, inclusion of these two species in the salivarius group does not appear justified as the grouping was weakly supported in both trees, and these two species were distantly related to the three salivarius species, which were connected by relatively short branch lengths and have been shown to be phenotypically distinct. We suggest that *S. downei* and *S. criceti* might best be considered as comprising a separate taxonomic assemblage, which we tentatively propose as the downei group. At the tip of the trees, there were two minor discrepancies. Within the sanguinis group, the position of *S. gordonii* and *S. cristatus* was switched, and within the pyogenic group, the position of *S. pyogenes* and *S. canis* was switched. At the base of the trees, relationships among the salivarius, downei, mutans, pyogenic, and bovis groups were poorly resolved, possibly reflecting the effect of frequent LGT during the early diversification of these groups. Relationships among the mitis, sanguinis, and anginosus groups, however, were well supported.

Clustering and Gene Gain/Loss Analyses

We identified 19,000 MCL clusters containing a total of 96,242 gene copies. Of these clusters, 4,322 (22.7%) contained three or more gene copies, 1,753 (9.2%) contained two gene copies (doublets), and 12,925 (68.0%) contained one gene copy (singletons). Kamneva et al. (2012) reported

Table 2
Genome Characteristics for the Eight *Streptococcus* Species Sequenced as Part of This Study

Species	Contigs	Base Pairs	CDS	rRNA	tRNA	%GC
Streptococcus porcinus (str. Jelinkova 176)	1	2,025,881	1,956	15	59	36.8
S. macacae (NCTC 11558)	1	1,916,985	1,871	15	65	41
S. urinalis (2285-97)	1	2,130,431	2,207	14	60	37.8
S. criceti (HS-6)	2	2,417,851	2,282	15	61	42.2
S. iniae (9117)	3	1,246,519	2,025	10	55	36.8
S. pseudoporcinus (LQ 940-04)	5	1,816,306	2,027	15	57	37.1
S. ictaluri (707-05)	8	2,234,402	2,473	9	45	38.2
S. equi subsp. *ruminatorum* (CECT 5772)	133	2,140,742	2,127	4	43	41.4

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

similar proportions for PVC bacteria: three or more gene copies = 19.4%, doublets = 10.4%, singletons = 70.2%. Genome flux was minimized using a LGT penalty of 3.0 and a duplication penalty of 2.0 (the default penalties established by David and Alm [2011]). Using these penalties, we detected a dynamic pattern of gene gain/loss through the phylogeny. The early evolution of *Streptococcus* (the branches prior to the eight major groups) was characterized by more gene gain than loss (balance of gains and losses = 444) (fig. 2), whereas evolution for each of the major groups including the branches leading to each group and excluding the terminal branches was characterized in general by more gene loss than gain (overall balance = −1,210) (fig. 2 and table 3). Finally, the terminal branches were characterized by much higher gain than loss (overall balance = 12,265).

These results are similar in part to several previous studies. For example, the weighted parsimony gene gain/loss study of Makarova et al. (2006) for 11 Lactobacillaceae/Leuconostocaceae species that are closely related to *Streptococcus* (Price et al. 2011) showed an overall pattern of gene loss. Similarly, using AnGST, Kamneva et al. (2012) showed a general pattern of genome shrinkage for PVC bacteria. Using a probabilistic gene birth-and-death model to investigate the evolution of Archaea, Csuros and Miklos (2009) also showed a pattern of overall gene loss. However, this pattern was additionally punctuated by periods of gene gain. They suggested a scenario where populations successful at occupying new environments due to a major physiological or metabolic invention, then underwent genomic streamlining (gene loss) as they diversified. More recently, using a model of

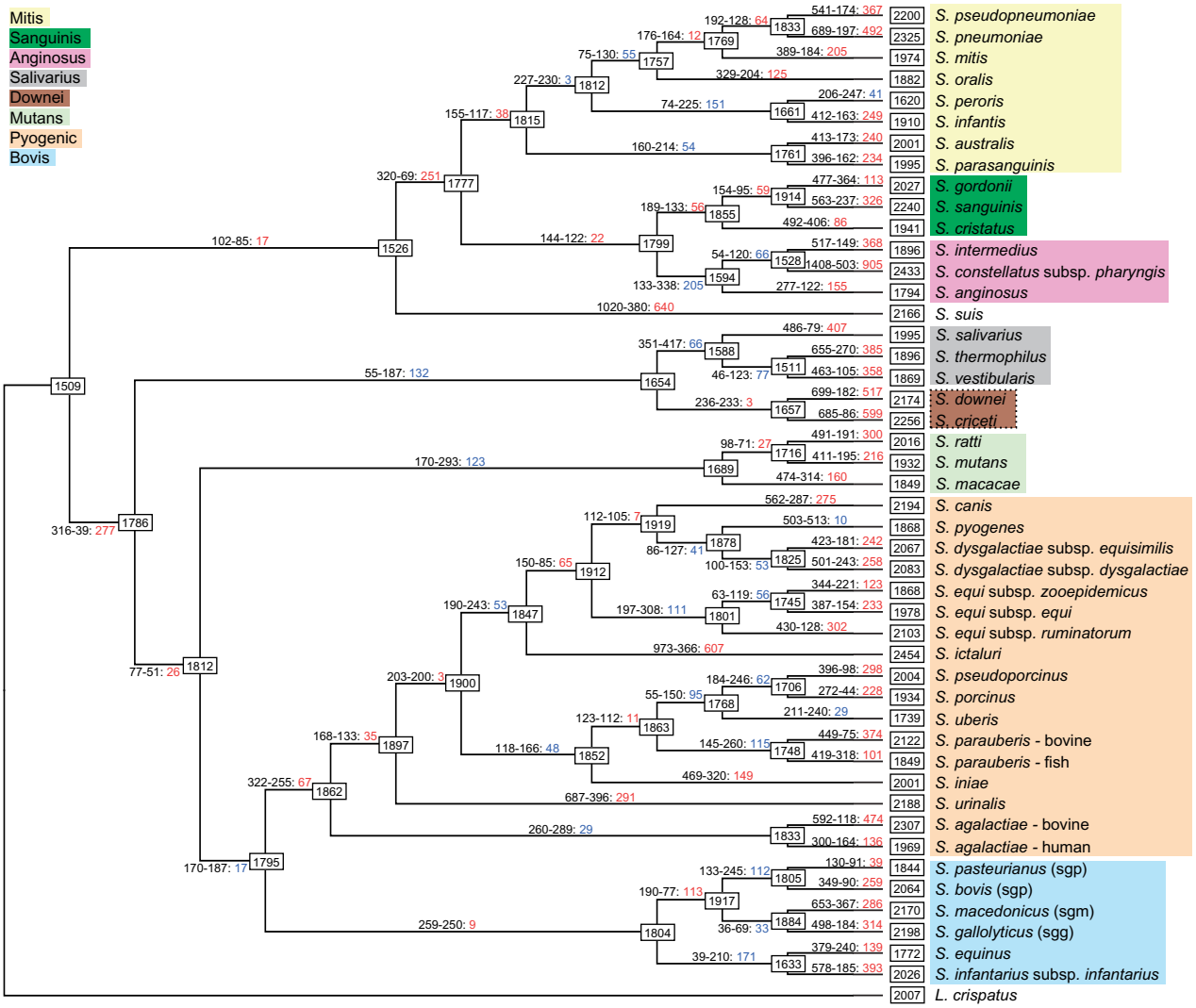

Fig. 2.—Result of gene gain/loss analysis. Boxes on nodes and tips of the phylogeny show genome size (genome size calculations exclude one MCL gene cluster that was excluded from the analysis—see Materials and Methods). Numbers on branches show the number of gains, losses, and the overall balance (gain–loss: balance), with red indicating an overall gain and blue an overall loss. Color shading for each major *Streptococcus* group follows figure 1.

Genome Biology and Evolution SMBE

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution

SMBE

Table 3
Gene Gain/Loss Summary

Group	Gains	Losses	Balance
Mitis[a]	1,059	1,208	−149
Sanguinis[a]	343	228	115
Anginosus[a]	187	458	−271
Salivarius[a]	397	540	−143
Downei[a]	236	233	3
Mutans[a]	268	364	−96
Pyogenic[a]	2,476	2,951	−475
Bovis[a]	657	851	−194
Mitis[b]	3,375	1,504	1,871
Sanguinis[b]	1,532	1,007	525
Anginosus[b]	2,202	774	1,428
Salivarius[b]	1,604	454	1,150
Downei[b]	1,384	268	1,116
Mutans[b]	1,376	700	676
Pyogenic[b]	7,935	3,866	4,069
Bovis[b]	2,587	1,157	1,430

[a]Gain/loss for each group including the branches leading to each group and excluding the terminal branches.
[b]Gain/loss for terminal branches for each group.

virtual cells evolving to maintain homeostasis, Cuypers and Hogeweg (2012) showed a similar general pattern of genome evolution where genomes undergo an initial period of expansion as they adapt to new environments followed by a longer period of streamlining where genes redundant in the new environment are lost. Over the broad timescale of our data, our results are similar to the latter studies, showing an initial period of genome expansion as the major groups began to diversify. Then, after the groups formed, genomes in general experienced a period of reductive evolution (streamlining). Finally, there was a more recent period of genome expansion as the majority of present day species evolved. Also using AnGST, David and Alm (2011) showed a similar pattern over an even larger time scale for the three domains of life, with an initial period of expansion during the Archaean followed by a period of streamlining. However, they did not show a more recent genomic expansion for prokaryotes. As highlighted by the authors, this discrepancy is likely explained by the fact that their analysis did not include singletons. In our analysis, singletons were responsible for the majority of gains on the terminal branches. For example, the overall balance of gains and losses on the terminal branches would be reduced from 12,265 to 704 if the singletons were removed. However, this high number of singletons should be considered in the context that, for the most part, our phylogeny only contained a single strain for each species. If it had been possible to include multiple strains for each species, we would have been able to better assess their dispensable genomes and therefore detect more gene losses; the overall proportion of singletons would have decreased, and the proportion of losses toward the terminal branches of the phylogeny would have increased (e.g., Lefebure and Stanhope 2007). Nevertheless, there is still an overwhelming pattern of high gain on the terminal branches, which has been reported numerous times for a wide range of bacterial species including those within the genus *Streptococcus* (Hao and Golding 2004; Hao and Golding 2006; Marri et al. 2006, 2007; Lefebure and Stanhope 2007; Kamneva et al. 2012). Additionally, several of these studies (including two that focused on *Streptococcus*) provided evidence that laterally acquired genes on terminal branches had a role in adaptation (Hao and Golding 2004; Hao and Golding 2006; Marri et al. 2006, 2007; Lefebure and Stanhope 2007; Kamneva et al. 2012).

Of the proportion of genes gained on the terminal branches, 6.8% were identified as LGT (see supplementary fig. S2, Supplementary Material online, for a breakdown of all evolutionary events). The majority of the remainder were gains classified as genes born on these branches (singletons). However, given that these genes were not present in any of the remaining taxa, it is possible that these genes were acquired via LGT from bacteria not included in our analysis. This possibility was explored by Lefebure et al. (2012) who performed a similar AnGST analysis on 15 *Streptococcus* species and showed that approximately two-thirds of the genes born on the *S. pyogenes* branch were likely LGTs from species not included in their analysis (they had significant Blast hits with the NCBI nr database). Although the remainder may have been born de novo on this branch, it is also possible that homologous genes have yet to be sequenced. Overall, our findings suggest that a large proportion of the *Streptococcus* pan-genome (all MCL gene clusters) has been involved in LGT. For example, 18.0% of all MCL clusters were directly identified as being involved in LGT and 68.0% of all clusters were singletons, many of which are likely to be LGTs. Assuming that two-thirds of the genes born on terminal branches are actually LGTs, this suggests that over 60% of the pan-genome has been subject to LGT. If we assume that all the genes born on the terminal branches are LGTs, it suggests that over 80% of the pan-genome has been subject to LGT.

From the perspective of gene gain/loss (turnover) for individual species, *S. constellatus* subsp. *pharyngis* is notable, as this taxon showed considerably more turnover than any other (gain = 1,408, loss = 503) (fig. 2), suggesting that perhaps this species was shifting or expanding its niche (Hao and Golding 2006; Marri et al. 2007). This high turnover was also reflected in the hierarchal clustering analysis (presence/absence of MCL gene clusters) (fig. 3) where this species was placed as an outlier to all remaining *Streptococcus* species. Similarly, the important human pathogens *S. pyogenes* and *S. pneumoniae* also showed high gene turnover, ranking 6th and 9th, respectively. In contrast, the two strains of *S. agalactiae* showed considerably less turnover ranking 21st (bovine isolate) and 40th (human isolate). The higher turnover for the bovine isolate compared with the human isolate might reflect a more recent adaptation to this environment. Similarly, *S. parauberis*, which was traditionally associated with the bovine

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

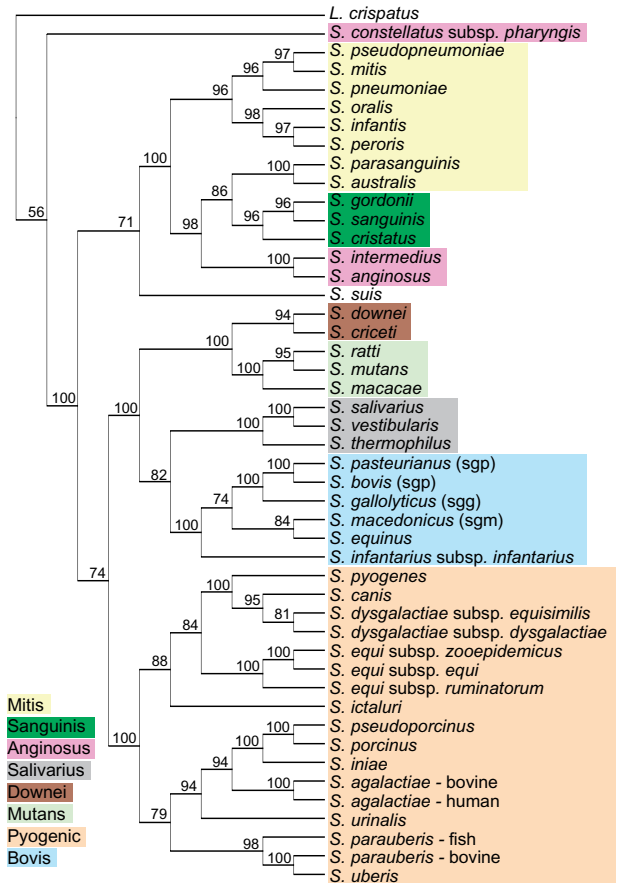

Fig. 3.—Hierarchal clustering among genomes using presence/absence of MCL gene clusters. Approximately unbiased *P* values are shown on branches. Color shading for each major *Streptococcus* group follows figure 1.

environment (Williams and Collins 1990), has recently been identified as an emerging fish pathogen (Nho et al. 2011), and the fish isolate (19th) showed considerably more turnover than the bovine isolate (38th). *Streptococcus suis*, which remained an outlier to the mitis, sanguinis, and anginosus groups, also showed particularly high turnover (gain = 1,020, loss = 380, ranking 2nd). This species is a major porcine pathogen; however, the species has recently been identified as a particularly virulent emerging zoonotic pathogen and as an etiological agent for streptococcal toxic shock syndrome (STSS) (Lun et al. 2007). The *S. suis* strain included in our analysis (98HAH33) was isolated from a fatal case of STSS from an outbreak in China (Chen et al. 2007). Again, the high gene turnover for this species may reflect its recent adaptation to the human environment.

A pattern of rapid adaptive radiation is perhaps best illustrated by the pyogenic group where there is little correlation between evolutionary relationship and host species (fig. 1). For example, the *S. equi* subspecies are all separated by relatively short branch lengths yet have distinct niches: *S. equi* subsp. *equi* is typically restricted to horses, *S. equi* subsp. *ruminatorum* has been isolated from mastitic sheep and goats, and *S. equi* subsp. *zooepidemicus* can infect a wide range of hosts, which suggests a recent diversification of these varied niches. Similarly, *S. dysgalactiae* subsp. *equisimilis* is typically isolated from the human environment, whereas *S. dysgalactiae* subsp. *dysgalactiae* is typically isolated from the bovine environment. There are also human and bovine ecotypes within *S. agalactiae* and bovine and fish ecotypes within *S. parauberis*, suggesting more recent adaptation. It is likely that LGT is a major evolutionary mechanism responsible for this rapid adaptation (Alm et al. 2006; Marri et al. 2006, 2007). Both genome relatedness and physical proximity are probable major factors affecting the frequency of this LGT.

With the exception of *S. parasanguinis* and *S. australis*, which clustered with the sanguinis group, and *S. constellatus* subsp. *pharyngis,* which was an outlier to all *Streptococcus* species, the hierarchal clustering analysis recovered all the major groups (fig. 3). However, relationships within the groups were not concordant with the species tree, suggesting that LGT is more likely to occur within the groups than between them. Lack of concordance with the species tree within the groups also suggests that relatedness has less of an influence on LGT and that other factors such as shared environment may be playing a role. Indeed, previous studies have provided good evidence for LGT among numerous *Streptococcus* species within a shared bovine environment (Richards et al. 2011, 2012). Furthermore, our analysis included five distantly related pyogenic species isolated from the bovine environment, and we detected LGT between ten different species-pair combinations involving those taxa. From the perspective of the genus as a whole, a factor contributing to the strong support for the major groups in the hierarchal clustering analysis might be the effect of group-specific gene loss (streamlining). Consequently, the combined effect of the reduced likelihood of intergroup LGT and group-specific streamlining may have contributed to genomic cohesiveness of the major taxonomic groups throughout the evolution of *Streptococcus*.

GO Term Enrichment and Genes Born on Branches Leading to Major Streptococcus *Groups*

Each of the eight phylogenetic groups was tested for enrichment of GO terms relative to the other groups. All groups showed enrichment for at least one of the three GO term domains of biological process (P), molecular function (F), and cellular component (C) (supplementary table S2, Supplementary Material online). The number of terms enriched for each group was as follows: mitis = 98; sanguinis = 17; anginosus = 6; salivarius = 34; downei = 6; mutans = 25; pyogenic = 241; and bovis = 198. For each GO term attached to a particular gene, there are typically

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

more general parent terms connected to it. The GO can be represented by a directed acyclic graph (DAG) and the parent terms become increasingly more general moving up through DAG levels. For the enrichment test, the term count for a particular gene includes all parent terms in the DAG (Blüthgen et al. 2005). Reducing the terms to their most specific, reduced the term count as follows: mitis = 47; sanguinis = 7; anginosus = 1; salivarius = 18; downei = 1; mutans = 5; pyogenic = 120; and bovis = 64 (supplementary table S3, Supplementary Material online).

To gain perspective on those genes that have likely been important for the evolution of each of the eight phylogenetic groups, we delineated the GO terms for the genes born on the branches leading to each group (supplementary table S3, Supplementary Material online). We then compared the terms on each group's branch to those that were enriched (most specific terms) for the same group. We omitted cellular component terms, as there were relatively few or no terms from this domain enriched for each group. In the discussion that follows, we primarily focused on terms that were both born on a group's branch and also enriched for the group. We acknowledge the obvious caveat that our analysis does not contain genome sequences for all *Streptococcus* species. Nevertheless, we provide the most complete assessment of the evolution of *Streptococcus* biochemical characteristics to date.

The Mitis Group

The mitis group showed enrichment for proteolysis, with genes annotated with this term occurring in the highest frequency. Genes with this term were distributed fairly evenly among species within the group (the number of genes assigned the term for each species ranged from 39 to 67 [average ~ 55]). The mitis group is primarily composed of commensal organisms of the upper respiratory tract and pioneer colonizers of dental plaque. Proteases have important roles in both these environments. For example, in plaque biofilms, bacteria utilize proteases to exploit salivary proteins as a nutrient source (Bradshaw et al. 1994; Wickstrom et al. 2009), and many species of pathogenic bacteria secrete proteases that interfere with host defenses and/or damage host cells or tissues (Harrington 1996; Miyoshi and Shinoda 2000; Potempa et al. 2000). Proteases can also aid bacteria spread and dissemination through tissue. For the mitis group, proteolysis appears to be a defining feature, as in addition to the enrichment, there were 17 genes born on the mitis branch that were also annotated with this term, suggesting that this characteristic has been retained throughout the group's evolutionary history. The characteristic is likely important to the group as a whole, since once it was gained; no species within the group lost it (not subject to genomic streamlining).

Terms for response to antibiotic and antibiotic transport were enriched for the mitis group. Studies focusing on antibiotic resistance for *Streptococcus* have indicated that in general the viridans group shows substantial resistance to antibiotics (in particular beta-lactam antimicrobials) (Facklam 2002). However, none of the genes born on the mitis branch were annotated with GO terms for response to antibiotic and antibiotic transport, suggesting that resistance to antimicrobials was acquired more recently via LGT. In contrast, a notable feature of the mitis group was the enrichment for *N*-acetyltransferase activity combined with the occurrence of ten genes born on the group's branch with this term. Further examination of the genes responsible for the enrichment showed approximately half (50.7%) to be annotated as GCN5-related *N*-acetyltransferases (GNAT). More specifically, the number of GNAT genes for each of the eight mitis species ranged from 5 to 22 (average ~ 14). Some members of the GNAT family of acetyltransferases confer resistance to aminoglycoside antibiotics (Vetting et al. 2005), and this resistance has been reported for several viridans species (Collatz et al. 1984; Horaud and Delbos 1984). Our findings suggest that for the mitis group, this resistance might be due in part to an intrinsic resistance of the group as a whole to aminoglycosides.

The Sanguinis Group

Of the terms enriched for the sanguinis group, the most frequent was *N*-acetyltransferase, and there were nine genes with this term born on the group's branch. Examination of the genes responsible for this enrichment showed that 78.3% were annotated as GNAT, and the number of GNAT genes for each of the three sanguinis species ranged from 20 to 30 (average ~ 24). These findings are similar to those for the mitis group and suggest the same possibility of an intrinsic ability or potential for resistance to aminoglycoside antibiotics.

The Salivarius Group

Notable enrichment for the salivarius group was for urease activity, urea metabolic process, nickel cation binding, and both cobalt and calcium transport. The genes responsible for the urease and urea enrichment belong to an inducible urease operon, which allows for the metabolism of urea found in saliva (Chen et al. 1998). Specifically, the genes included those encoding all three of the structural proteins (alpha, beta, and gamma) and one of the four accessory genes. These genes were also born on the group's branch suggesting that this characteristic has been retained throughout the group's history. However, not all *S. salivarius* strains possess the operon. Geng et al. (2011) showed it to be absent in strain SK126, suggesting a recent loss of the operon from some strains within this species (perhaps reflecting the early stages of streamlining). The urease enzyme requires nickel to function (Chen et al. 1998), explaining the enrichment for nickel cation binding. In addition, Chen and Burne (2003) showed that for *S. salivarius*, a three gene cobalt ATP-dependent

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

binding cassette (ABC) transporter immediately 3′ to the urease operon (*ureMQO*) was functioning as a nickel transporter, likely explaining the enrichment for cobalt transport. A search of the chromosome, wgs, and refseq_genomic databases at NCBI for the complete urease operon (*ureABCEFGDMQO*) showed it to be only present in the three salivarius species. Interestingly, *S. thermophilus* has been reported as nonureolytic in the literature (Facklam 2002). One possible explanation for this might be that similar to *S. salivarius*, certain strains of *S. thermophilus* lack the operon.

The salivarius group also showed enrichment for transposase activity (molecular function) and DNA-mediated transposition (biological process), with these terms born on the group's branch. The terms occurred relatively frequently, suggesting potential for high levels of recombination. However, the gene gain/loss analysis showed the group to rank last regarding the number of genes exchanged among all 46 taxa (LGT counts for each group were normalized by dividing the count by the number of taxa in the group) (supplementary table S4, Supplementary Material online), suggesting that recombination was more likely between strains within a species than among species. The pyogenic group showed a very similar pattern for these two terms. Both were enriched, occurred very frequently, and were born on the group's branch. This pattern suggests independent origins for the transposases in the salivarius and pyogenic groups and that these transposases have tended to remain within their respective groups through evolutionary time, again suggesting that recombination mediated by these genes was more likely to have occurred among more closely related sequences.

The Pyogenic Group

For the pyogenic group, the majority of enriched terms involved transport and metabolism; more specifically, the phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS) and the metabolism of carbohydrates. Terms for the PTS were among the most frequent and the genes with this term (and also several associated with carbohydrate metabolism) were born on the group's branch, suggesting that this characteristic has been retained throughout the group's history. The MCL clustering analysis delineated the genes with PTS GO terms into eight clusters. Six contained PTS permeases (IIA, B, and C), and two contained membrane regulatory proteins. Most hexose sugars and disaccharides used by streptococci are taken up by the PTS (Price et al. 2011), and their enrichment for the pyogenic group could reflect the wide range of hosts and environments inhabited by the species of this group. The pyogenic group also showed enrichment for the pathogenesis GO term. Although the genes responsible were distributed among all species in the group, they were not distributed evenly (supplementary table S5, Supplementary Material online). Not surprisingly, *S. pyogenes* had the highest number of these genes (19). The next two highest species were *S. dysgalactiae* subsp. *equisimilis* (12) and *S. equi* subsp. *equi* (12), with the former an emerging human pathogen (Brandt and Spellerberg 2009). In total, the MCL clustering analysis showed *S. pyogenes* to share 11 pathogenesis genes with other pyogenic species (supplementary table S5, Supplementary Material online). *Streptococcus dysgalactiae* subsp. *equisimilis* shared the most genes (eight), likely contributing to its emergence as a human pathogen. *Streptococcus equi* subsp. *equi* was the next highest with five. None of the genes with pathogenesis GO terms were born on the pyogenic group's branch, indicating that these genes were acquired via LGT.

The Bovis Group

The bovis group showed enrichment for regulation of transcription (DNA dependent) and sequence-specific DNA-binding transcription factor activity, with genes annotated with these two terms occurring in the highest frequency. These results suggest an enhanced regulatory ability for the bovis group, where strains can activate/deactivate multiple metabolic pathways to survive in resource poor environments such as the colon. Genes for both terms were also born on the group's branch, suggesting that this regulatory ability has been retained throughout the group's history. Examination of the genes responsible for this enrichment, that were also born on the group's branch, revealed six MCL clusters that in general exclusively contained transcriptional regulator genes from each of the bovis species (supplementary table S6, Supplementary Material online). Notably, one of the clusters contained genes from the multiple antibiotic resistance regulator (MarR) family of regulators. For *Escherichia coli*, this family of regulators has been shown to regulate genes responsible for resistance to antibiotics and other toxic chemicals (Alekshun and Levy 1999). Annotations for genes in the other clusters suggested that they belonged to the LysR, Crp-Fnr, and GntR families of regulators and were involved in regulation of anaerobic metabolic pathways and pyridoxine (vitamin B6) metabolism. The transcriptional regulator genes identified here represent useful targets for further study. For example, ascertaining virulence for the genes regulated, which could be particularly valuable given the possible involvement in antibiotic resistance for several of the regulators identified.

Conclusion

Revealing a pattern of genomic expansion and streamlining for *Streptococcus*, our study adds to the emerging view that genomes evolve in this manner (Cuypers and Hogeweg 2012) and lends support to the proposal that this process may be a generic pattern of evolving systems (Cuypers and Hogeweg 2012). Furthermore, our results demonstrate that despite LGT affecting a substantial proportion of the *Streptococcus*

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

pan-genome, many of the major groups have retained distinct core characteristics since their formation.

Supplementary Material

Supplementary tables S1–S7 and figures S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Olga Kamneva for assistance with the gene/gain loss analysis. This work was supported in whole or part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract number HHSN272200900007C and grant number AI073368 and by United States Department of Agriculture grant 2006-35600-16569 to S.K.H.

Literature Cited

Alekshun MN, Levy SB. 1999. The mar regulon: multiple resistance to antibiotics and other toxic chemicals. Trends Microbiol. 7:410–413.

Alm E, Huang K, Arkin A. 2006. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol. 2:e143.

Baker C. 2000. Group B streptococcal infections. In: Stevens D, Kaplan E, editors. Streptococcal infections. Clinical aspects, microbiology, and molecular pathogenesis. New York: Oxford Univeristy Press. p. 222–237.

Balter S, Whitney C, Schuchat A. 2000. Epidemiology of group B streptococcal infections. In: Fischetti V, Novick R, Ferreti J, Portnoy D, Rood J, editors. Gram-positive pathogens. Washington (DC): ASM Press.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 57:289–300.

Bentley RW, Leigh JA, Collins MD. 1991. Intrageneric structure of *Streptococcus* based on comparative-analysis of small-subunit ribosomal-RNA sequences. Int J Syst Bacteriol. 41:487–494.

Blüthgen N, et al. 2005. Biological profiling of gene groups utilizing gene ontology. Genome Inform. 16:106–115.

Bradshaw DJ, Homer KA, Marsh PD, Beighton D. 1994. Metabolic cooperation in oral microbial communities during growth on mucin. Microbiology 140 (Pt 12):3407–3412.

Brandt CM, Spellerberg B. 2009. Human infections due to *Streptococcus dysgalactiae* subspecies *equisimilis*. Clin Infect Dis. 49:766–772.

Brohee S, van Helden J. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7:488.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172: 2665–2681.

Carapetis JR, Steer AC, Mulholland EK, Weber M. 2005. The global burden of group A streptococcal diseases. Lancet Infect Dis. 5:685–694.

Chain PS, et al. 2009. Genomics. Genome project standards in a new era of sequencing. Science 326:236–237.

Chen C, et al. 2007. A glimpse of streptococcal toxic shock syndrome from comparative genomics of *S. suis* 2 Chinese isolates. PLoS One 2:e315.

Chen YY, Burne RA. 2003. Identification and characterization of the nickel uptake system for urease biogenesis in *Streptococcus salivarius* 57.I. J Bacteriol. 185:6773–6779.

Chen YY, Weaver CA, Mendelsohn DR, Burne RA. 1998. Transcriptional regulation of the *Streptococcus salivarius* 57.I urease operon. J Bacteriol. 180:5769–5775.

Collatz E, Carlier C, Courvalin P. 1984. Characterization of high-level aminoglycoside resistance in a strain of *Streptococcus pneumoniae*. J Gen Microbiol. 130:1665–1671.

Csuros M, Miklos I. 2009. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol. 26:2087–2095.

Cuypers TD, Hogeweg P. 2012. Virtual genomes in flux: an interplay of neutrality and adaptability explains genome expansion and streamlining. Genome Biol Evol. 4:212–229.

David LA, Alm EJ. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469:93–96.

Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. 2009. Properties of consensus methods for inferring species trees from gene trees. Syst Biol. 58:35–54.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Dermer P, Lee C, Eggert J, Few B. 2004. A history of neonatal group B *Streptococcus* with its related morbidity and mortality rates in the United States. J Pediatr Nurs. 19:357–363.

Ewing GB, Ebersberger I, Schmidt HA, von Haeseler A. 2008. Rooted triple consensus and anomalous gene trees. BMC Evol Biol. 8:118.

Facklam R. 2002. What happened to the streptococci: overview of taxonomic and nomenclature changes. Clin Microbiol Rev. 15: 613–630.

Geng J, Huang SC, Li S, Hu S, Chen YY. 2011. Complete genome sequence of the ureolytic *Streptococcus salivarius* strain 57.I. J Bacteriol. 193:5596–5597.

Gotz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36:3420–3435.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hao W, Golding GB. 2004. Patterns of bacterial gene movement. Mol Biol Evol. 21:1294–1307.

Hao W, Golding GB. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 16:636–643.

Harrington DJ. 1996. Bacterial collagenases and collagen-degrading enzymes and their potential role in human disease. Infect Immun. 64: 1885–1891.

Highlander SK, et al. 2007. Subtle genetic changes enhance virulence of methicillin resistant and sensitive *Staphylococcus aureus*. BMC Microbiol. 7:99.

Horaud T, Delbos F. 1984. Viridans streptococci in infective endocarditis: species distribution and susceptibility to antibiotics. Eur Heart J. 5(Suppl C), 39–44.

Jakobsen IB, Easteal S. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput Appl Biosci. 12:291–295.

Kamneva OK, Knight SJ, Liberles DA, Ward NL. 2012. Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. Genome Biol Evol. 4:1375–1390.

Kawamura Y, Hou XG, Sultana F, Miura H, Ezaki T. 1995. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. Int J Syst Bacteriol. 45:406–408.

Kawamura Y, Whiley RA, Shu SE, Ezaki T, Hardie JM. 1999. Genetic approaches to the identification of the mitis group within the genus *Streptococcus*. Microbiology 145(Pt 9):2605–2613.

Köhler W. 2007. The present state of species within the genera *Streptococcus* and *Enterococcus*. Int J Med Microbiol. 297:133–150.

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014

Genome Biology and Evolution SMBE

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006a. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol. 23:1891–1901.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006b. GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098.

Lefebure T, Richards VP, Lang P, Pavinski-Bitar P, Stanhope MJ. 2012. Gene repertoire evolution of *Streptococcus pyogenes* inferred from phylogenomic analysis with *Streptococcus canis* and *Streptococcus dysgalactiae*. PLoS One 7:e37607.

Lefebure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8:R71.

Loesche WJ. 1986. Role of *Streptococcus mutans* in human dental decay. Microbiol Rev. 50:353–380.

Lun ZR, Wang QP, Chen XG, Li AX, Zhu XQ. 2007. *Streptococcus suis*: an emerging zoonotic pathogen. Lancet Infect Dis. 7:201–209.

Magallon SA, Sanderson MJ. 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. Evolution 59: 1653–1670.

Makarova K, et al. 2006. Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A. 103:15611–15616.

Marri PR, Hao W, Golding GB. 2006. Gene gain and gene loss in *Streptococcus*: is it driven by habitat? Mol Biol Evol. 23: 2379–2391.

Marri PR, Hao W, Golding GB. 2007. The role of laterally transferred genes in adaptive evolution. BMC Evol Biol. 7(Suppl 1), S8.

Maynard Smith J. 1992. Analyzing the mosaic structure of genes. J Mol Evol. 34:126–129.

Miyoshi S, Shinoda S. 2000. Microbial metalloproteases and pathogenesis. Microbes Infect. 2:91–98.

Myers EW, et al. 2000. A whole-genome assembly of *Drosophila*. Science 287:2196–2204.

Nho SW, et al. 2011. Complete genome sequence and immunoproteomic analyses of the bacterial fish pathogen *Streptococcus parauberis*. J Bacteriol. 193:3356–3366.

O'Brien KL, Nohynek H. 2003. Report from a WHO Working Group: standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*. Pediatr Infect Dis J. 22:e1–e11.

Paradis E, Claude J, Strimmer K. 2004. ape: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.

Potempa J, Banbula A, Travis J. 2000. Role of bacterial proteinases in matrix destruction and modulation of host responses. Periodontol 2000 24:153–192.

Price CE, Zeyniyev A, Kuipers OP, Kok J. 2011. From meadows to milk to mucosa—adaptation of *Streptococcus* and *Lactococcus* species to their nutritional environments. FEMS Microbiol Rev. 36:949–971.

Ralph AP, Carapetis JR. 2013. Group A *Streptococcal* diseases and their global burden. Curr Top Microbiol Immunol. 368:1–27.

Richards VP, et al. 2011. Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. Infect Genet Evol. 11:1263–1275.

Richards VP, et al. 2012. Genome characterization and population genetic structure of the zoonotic pathogen, *Streptococcus canis*. BMC Microbiol. 12:293.

Roshan U, Livesay DR. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics 22: 2715–2721.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 19: 101–109.

Sherman J. 1937. The *Streptococci*. Bacteriol Rev. 1:3–97.

Strimmer K, vonHaeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol. 13: 964–969.

Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22:1540–1542.

Täpp J, Thollesson M, Herrmann B. 2003. Phylogenetic relationships and genotyping of the genus *Streptococcus* by sequence determination of the RNase P RNA gene, rnpB. Int J Syst Evol Microbiol. 53:1861–1871.

van Dongen S. 2000. Graph clustering by flow simulation [Ph.D. dissertation]. [Utrecht (The Netherlands)]: University of Utrecht.

Vetting MW, et al. 2005. Structure and functions of the GNAT superfamily of acetyltransferases. Arch Biochem Biophys. 433:212–226.

Wickstrom C, Herzberg MC, Beighton D, Svensater G. 2009. Proteolytic degradation of human salivary MUC5B by dental biofilms. Microbiology 155:2866–2872.

Williams AM, Collins MD. 1990. Molecular taxonomic studies on *Streptococcus uberis* types I and II. Description of *Streptococcus parauberis* sp. nov. J Appl Bacteriol. 68:485–490.

Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH. 2011. Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J Mammary Gland Biol Neoplasia. 16: 357–372.

Zwickl D. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. [Ph.D. dissertation]. [Austin (TX)]: The University of Texas at Austin.

Associate editor: Takashi Gojobori

Downloaded from http://gbe.oxfordjournals.org/ at Clemson University on November 30, 2014