Comparative Genomic Analysis of the *Streptococcus dysgalactiae* Species Group: Gene Content, Molecular Adaptation, and Promoter Evolution

Haruo Suzuki[1], Tristan Lefébure[†,1], Melissa Jane Hubisz[2], Paulina Pavinski Bitar[1], Ping Lang[§,1], Adam Siepel*[,2], and Michael J. Stanhope*[,1]

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

[2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York

[†]Present addresses: Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, Université de Lyon, France; and CNRS, Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, Université Lyon 1, Villeurbanne, France.

[§]Present address: Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, New York.

*Corresponding authors: E-mail: mjs297@cornell.edu; acs4@cornell.edu.

Accepted: 24 January 2011

Abstract

Comparative genomics of closely related bacterial species with different pathogenesis and host preference can provide a means of identifying the specifics of adaptive differences. *Streptococcus dysgalactiae* (SD) is comprised of two subspecies: *S. dysgalactiae* subsp. *equisimilis* is both a human commensal organism and a human pathogen, and *S. dysgalactiae* subsp. *dysgalactiae* is strictly an animal pathogen. Here, we present complete genome sequences for both taxa, with analyses involving other species of *Streptococcus* but focusing on adaptation in the SD species group. We found little evidence for enrichment in biochemical categories of genes carried by each SD strain, however, differences in the virulence gene repertoire were apparent. Some of the differences could be ascribed to prophage and integrative conjugative elements. We identified approximately 9% of the nonrecombinant core genome to be under positive selection, some of which involved known virulence factors in other bacteria. Analyses of proteomes by pooling data across genes, by biochemical category, clade, or branch, provided evidence for increased rates of evolution in several gene categories, as well as external branches of the tree. Promoters were primarily evolving under purifying selection but with certain categories of genes evolving faster. Many of these fast-evolving categories were the same as those associated with rapid evolution in proteins. Overall, these results suggest that adaptation to changing environments and new hosts in the SD species group has involved the acquisition of key virulence genes along with selection of orthologous protein-coding loci and operon promoters.

Key words: *Streptococcus dysgalactiae* subsp. *equisimilis*, *S. dysgalactiae* subsp. *dysgalactiae*, gene content, molecular adaptation, promoter evolution.

Introduction

Streptococcus dysgalactiae (SD) is one of several Lancefield group C, G, and L streptococci falling within the pyogenic group of *Streptococcus*. In 1996, Vandamme et al., based on observed differences in physiological and biochemical properties, proposed dividing the species into two subspecies: 1) *S. dysgalactiae* subsp. *equisimilis* (SDE), including human strains belonging to lancefield groups C and G and 2) *S. dysgalactiae* subsp. *dysgalactiae* (SDD) for strains of animal origin belonging to groups C and L. Other studies, including pulse field gel electrophoresis, DNA–DNA reassociation experiments, multilocus enzyme electrophoresis, phenotypic studies, and phylogenetic analysis of various gene sequences, have since confirmed the necessity to divide SD into two subspecies which are correlated with source host (Bert et al. 1997; Vieira et al. 1998; Glazunova et al. 2010). SDE was primarily regarded as a human commensal organism (Rolston 1986) but is now recognized as an increasingly important human pathogen, which can cause a spectrum of human diseases, including cellulitis, peritonitis, septic arthritis, pneumonia, endocarditis, acute

pharyngitis, bacteremia, and toxic shock syndrome (Brandt and Spellerberg 2009). Many of these infections are similar to those caused by the important human pathogen *Streptococcus pyogenes*, and the two organisms have been shown to share many of the same virulence genes (Davies et al. 2007). SDD on the other hand is strictly an animal pathogen and a major cause of bovine mastitis. At present, there are no published accounts of genome sequence data for either of these subspecies, although GenBank (Benson et al. 2009) contains a genome sequence for a strain of SDE isolated from a patient with streptococcal toxic shock syndrome (STSS), which has been briefly mentioned in the literature (Sunaoshi et al. 2009; Takahashi et al. 2010). Genomic comparisons of SDD and SDE are likely to provide useful information regarding host adaptation and pathogenesis, given the relatively strict host demarcation between the two subspecies and their distinct disease phenotypes.

Arguably, the three most significant molecular characteristics responsible for biological differences between lineages of bacteria are: 1) presence and absence of particular loci, 2) differences in orthologous proteins driven by selection, and 3) differential gene regulation. Analysis of genomic gene content differences may provide some insight into the adaptive differences between these two subspecies, however, it is also important to understand the role of natural selection in explaining any observed sequence differences at shared protein-coding loci, as well as the potential role of noncoding functional elements. This latter aspect is a particularly underexplored area of bacterial comparative genomics. Here, we provide an examination of the evolution of promoter regions in the different lineages pertaining to the two subspecies SD, as well as *S. pyogenes*, and assess whether different rates of evolution are associated with different metabolic functions or biochemical categories of loci. Although we regard this as an initial, preliminary step toward understanding adaptive evolution of noncoding DNA in streptococci, it is nonetheless our hope that this three-faceted approach to studying molecular adaptation will ultimately lead to identifying the relative importance of each of these adaptive mechanisms in bacteria of different taxonomic groups and environmental situations.

The principal aims of this paper then are to describe genome sequences for SDD and SDE and to present comparative evolutionary analyses leading to hypotheses regarding genes and putative noncoding regions that could be linked to the specific features of host adaptation and pathogenesis of these two subspecies. In addition, we chose as our SDE isolate a strain of putative commensal origin, isolated from a skin infection, in order to provide a comparison to the STSS-linked strain of SDE, for which there is currently a genome sequence on National Center for Biotechnology Information (NCBI).

Materials and Methods

Genome Sequencing

SDD strain NADC Z-8 is a Lancefield group C streptococci isolated from a bovine udder infection in the early 1970s (McDonald TJ and McDonald JS 1976). Strain D166B is a Lancefield group G *Streptococcus dysgalactiae* subsp. *equisimilis* (SDE1), collected in 1939 from a blister of a child with epidermolysis bullosa, an inherited skin disorder that causes blistering in response to minor injury. The strains can be found at the American Type Culture Collection under accession numbers ATCC 27957 and ATCC 12394, respectively. Both these strains were sequenced as part of this work. We also included in our analysis a genome sequence for SDE GGS_124 (stG480.0), which was isolated from patients with STSS (GenBank accession number AP010935). We refer to this sequence as SDE2.

Roche/454 pyrosequencing was used to determine the sequence of both genomes. A total of 1015102 and 1028577 single-end reads and 656086 and 635598 paired-end reads resulted from the FLX sequencer, for SDD and SDE1, respectively. De novo assembly with Newbler yielded 354 contigs arranged in 14 scaffolds for SDD (NADC Z-8) and 9 scaffolds comprising 195 contigs for SDE1 (D166B) for an average coverage of 56.1 and 56.5, respectively. Physical maps of both genomes were determined by OpGen Technologies, Inc. using restriction enzyme BglII and the optical mapping technique. The order and orientation of the scaffolds was determined by aligning the scaffold on the optical map using Opgen Mapviewer. Small inter and intrascaffold gaps were closed by polymerase chain reaction (PCR) and sequenced using the Sanger approach, while 17 large gaps were amplified with long range PCR and sequenced on the Illumina GA2 sequencer. The Illumina reads were assembled with Velvet (Zerbino and Birney 2008) using a large range of parameters, and the best assembly was selected using the N50 statistic.

Genome Characterization and Gene Content

Genome annotation for SDD and SDE1 was done by NCBI Prokaryotic Genomes Automatic Annotation. Basic genome features were determined using the G-language Genome Analysis Environment version 1.8.11 (Arakawa et al. 2003; Arakawa and Tomita 2006). Statistical tests and graphics were implemented using R, version 2.11.1 (R_Development_Core_Team 2010). Circular genome maps were generated using the Circular Genome Viewer (Stothard and Wishart 2005; Grant and Stothard 2008). Pairwise comparisons of nucleotide sequences were performed using bl2seq (BlastN; *E* value cutoff of 1×10^{-2}) (Altschul et al. 1997) and displayed using GenomeMatcher (Ohtsubo et al. 2008). Transcription units (TUs) were predicted using PathoLogic (Karp et al. 2009), based on features, such as intergenic distances, direction of

transcription, known functional relationships between genes, and comembership in pathways or protein complexes (Karp et al. 2009). Genome sequences for SDD and SDE1 have been deposited in GenBank under accession numbers AEGO00000000 and CP002215, respectively.

The complete listings of the chromosomes of *Streptococcus* used in this study (a total of 45, including SDD, SDE1, and SDE2) are shown in supplementary table S1 (Supplementary Material online). To characterize the gene content of SDD and SDE1 and compare it with that of other *Streptococcus* taxa, orthologous genes were identified using OrthoMCL (Li et al. 2003). OrthoMCL clusters proteins based on sequence similarity, using an all-against-all Blast comparison of proteomes, followed by normalization of interproteomic differences, and Markov graph clustering (MCL) to define ortholog groups using amino acid sequences from all 45 strains. OrthoMCL was implemented using a BlastP *E* value cutoff of 1×10^{-5} and the default MCL inflation parameter of 1.5 (for a complete listing of the predicted orthologous groups, see supplementary table S2, Supplementary Material online). The dissimilarity in gene content patterns (binary data for the presence or absence of each protein) between two genomes was measured by the Jaccard distance (one minus the Jaccard coefficient), and the distance matrix was subject to hierarchical cluster analysis (unweighted pair group method with arithmetic mean [UPGMA]). To identify variations in virulence gene sets among the 45 *Streptococcus* strains, we used the Virulence Factors Database (VFDB) (Chen et al. 2005), which contains 2,294 proteins, experimentally verified as virulence loci, from various pathogenic bacteria, including 88 proteins from *Streptococcus*. Using TBlastN, we searched for high-similarity matches between each of the 88 proteins and each of the 45 *Streptococcus* proteomes. We also used OrthoMCL to determine *Streptococcus* virulence gene orthologs in the complete VFDB, and finally, we performed similar searches against a previously assembled list of 211 putative *Streptococcus* virulence loci (Davies et al. 2007).

Prophage regions in SD chromosomes were predicted using Prophinder (Lima-Mendez et al. 2008). This tool performs BlastP to detect homologs of phage proteins stored in the ACLAME database (Leplae et al. 2010), and then regions with a high density of phage-like proteins are identified as putative prophages. To detect integrative conjugative elements (ICEs) and related elements, we used TBlastN to find matches between an ICE from SDE NS3396 (ICESde3396; GenBank accession, EU142041) (Davies et al. 2009) and each of the 45 *Streptococcus* genomes.

Both to support our own analyses and to facilitate use of our data by the broader community, we created a publicly available *Streptococcus* Genome Browser based on the UCSC Genome Browser platform (Kent et al. 2002; Rhead et al. 2010). Our comparative analyses using this browser were performed with SDE1 as a reference genome. Protein-coding annotations were augmented with predictions of RNA genes based on the Rfam database and INFERNAL software (Gardner et al. 2009) and predictions of transcription factor binding sites based on motifs from RegTransBase (Kazakov et al. 2007). Pairwise alignments between SDE1 and each of four other genomes—SDE2, SDD, *S. pyogenes* strain MGAS315 (SPY1), *S. pyogenes* strain MGAS10750 (SPY2), and *S. equi subsp. equi* strain 4047 (SEE)—were produced using lastz (http://www.bx.psu.edu/miller_lab). These alignments were then processed using the UCSC alignment "chains and nets" pipeline (Kent et al. 2002) to eliminate paralogous alignments and identify regions of conserved synteny. A genome-wide multiple alignment was then obtained using multiz (Blanchette et al. 2004).

Molecular Adaptation of Protein-Coding Loci

Genome-wide positive selection (PS) was assessed using our previously developed pipeline, described in detail elsewhere (Lefébure and Stanhope 2007, 2009). The sequences considered included SDE1, SDE2, SDD, SPY1, SPY2, and SEE. Preliminary phylogenetic analysis involving the single-copy orthologous genes from the genomes of the *Streptococcus* taxa included in this study, confirmed the monophyly of this group, and suggested SD as the sister group to *S. pyogenes* (supplementary fig. S1, Supplementary Material online). Orthologous gene content information, determined using OrthoMCL, was used to delimit the core genome of this set of taxa. The sequences were first aligned at the amino acid level using Probalign (v1.1) (Roshan and Livesay 2006), then backtranslated to DNA, and alignment columns with a posterior probability <0.6 were removed. Alignments with $>50\%$ of the sites removed, based on this 0.6 posterior probability cutoff, were discarded from further analysis. To minimize the influence of recombination in the PS scan, the alignments were tested for intragenic recombination using GARD (Kosakovsky Pond et al. 2006). When a recombination breakpoint was found to be significant, the alignment was broken into two or more gene fragments. For each of the resulting alignments, a gene tree was reconstructed using PhyML (Guindon and Gascuel 2003; Guindon et al. 2010) employing a general time revisable + gamma model of evolution, the maximum likelihood criteria, and the subtree pruning-regrafting branch-swapping method. There was a clear consensus of the gene trees toward a single species tree topology. This species tree topology (SDE1/SDE2 joined by SDD, followed by the two *S. pyogenes* strains and finally *S. equi* subsp. *equi*) was used to detect putative lateral gene transfers (LGTs) based on phylogenetic signal. Each gene tree search was bootstrapped (500 pseudoreplicates) with PhyML using the nearest-neighbor interchange branch-swapping method, and genes supporting strongly conflicting bipartitions were considered LGTs and removed from the analysis. Using each of these non-LGT

alignments and the species tree topology, PS was assessed on each of the SDE1, SDE2, SDE, SDD, and SD lineages using the branch-site test implemented in CodeML program of PAML version 4b (Yang 2007). The likelihoods of model "A" and model "1a" were compared, and *P* values were calculated under the assumption that the likelihood ratio has a chi-square distribution with one degree of freedom (Zhang et al. 2005). Multiple testing adjustments were performed by using a false discovery rate (FDR) approach at a 5% significance level (Benjamini and Yekutieli 2001).

Evolutionary Rates

Protein-Coding Sequences. We also examined the evolutionary rates of protein-coding sequences in a category-specific manner. For this analysis, we made use of the ratio of nonsynonymous to synonymous substitution rates, $\omega = d_N/d_S$. First, we considered overall differences in evolutionary rate between categories, ignoring differences among branches of the phylogeny. Each category *C* partitioned the genes and corresponding alignments into two sets: those assigned to *C* and those not assigned to *C*. We concatenated all the alignments assigned to *C* into one alignment, denoted X_C, and all the remaining alignments into another, denoted $X_{\underline{C}}$, and performed a likelihood ratio test (LRT) of the null hypothesis that the values of ω for X_C and $X_{\underline{C}}$ are equal ($\omega_C = \omega_{\underline{C}}$) against the alternative hypothesis that they are unequal ($\omega_C \neq \omega_{\underline{C}}$). Under both the null and alternative hypotheses, the transition–transversion rate ratio (κ), the equilibrium codon frequencies (under the F3 × 4 parameterization), and the branch lengths of the tree were shared for X_C and $X_{\underline{C}}$ and were estimated by maximum likelihood. This led to a comparison of nested models differing by one parameter, and *P* values were computed by assuming twice that the difference in log likelihoods obeyed the asymptotic chi-squared distribution with one degree of freedom under the null hypothesis. These tests were performed using custom software developed by one of the authors (M.J.H.), which supports LRTs like those implemented in the PAML package (Yang 2007), but with a somewhat more flexible parameterization (Kosiol et al. 2008). In practice, these tests were almost always highly significant because even small differences in ω were supported by large amounts of data. Therefore, we have focused our interpretation on the ω estimates themselves rather than on the likelihood ratios or *P* values, which tend to reflect the sizes of the gene sets more than the magnitude of the effect.

In addition, we performed a series of clade-specific LRTs, in which the null model has a single ω parameter, but the alternative model has two such parameters: ω_f, which applies to a designated set of "foreground" branches (say, those in the SDE/SDD clade; supplementary fig. S2, Supplementary Material online), and ω_b, which applies to the remaining "background" branches. These are therefore tests of the null hypothesis that $\omega_f = \omega_b$ against the alternative hypothesis that $\omega_f \neq \omega_b$ for each foreground set of interest. We applied these LRTs to all category-specific alignments, X_C, for foreground sets corresponding to each clade of interest (supplementary fig. S2, Supplementary Material online). Notice that these tests contrast sets of branches rather than sets of sites and therefore do not depend on the complementary alignments, $X_{\underline{C}}$. These LRTs are instances of the "branch test" implemented in PAML (the null model is known as M0 and the alternative model as "Model B") and were performed using the CodeML program. *P* values were computed by assuming an asymptotic chi-square distribution with one degree of freedom, as above. These tests, like the category tests, were usually significant, so we have focused on the ω estimates (specifically, on the ratio ω_f / ω_b) in interpreting them.

Gene Ontology categories (Ashburner et al. 2000) were assigned to clusters of orthologous genes by homology. The *Streptococcus* genes were compared with all bacterial proteins from the Uniref90 database using BlastP, and the GO categories of matches with *E* values $< 10^{-5}$ were obtained using the uniProt GOA database. Each gene was also explicitly mapped to all parents of its assigned categories in the GO hierarchy. For the category-by-category analyses, only the 127 GO categories containing ten genes or more were considered.

Promoter Sequences. We identified sequences upstream of predicted protein encoding transcriptional units (ending at the next unit) as putative promoters and assigned each such promoter sequence to the union of GO categories of its constituent genes. We then tested for significant differences in evolutionary rates between promoters of different categories, using methods similar to those above. In this case, we made use of a neutral phylogenetic model estimated from 4-fold degenerate (4D) sites in our protein-coding gene set and estimated global scale factors for the phylogeny rather than the ω parameter. As above, each category *C* induced two alignments, a concatenated alignment of the promoter sequences in the category, X_C, and a concatenated alignment of all other promoter sequences, $X_{\underline{C}}$. Once again, we considered both tests of all branches of the phylogeny and tests of particular foreground branches of interest. In the all-branch test, the null hypothesis was that the average evolutionary rate for X_C, denoted r_C, and the average rate for $X_{\underline{C}}$, denoted $r_{\underline{C}}$, were equal ($r_C = r_{\underline{C}}$), where *r* acts as a global scaling constant for the branches of the phylogeny. The alternative hypothesis is that these two rates are unequal ($r_C \neq r_{\underline{C}}$). Parameter estimation and likelihood computation was performed using the phyloFit program (Siepel and Haussler 2004), holding fixed all other parameters of the neutral model (the branch-length proportions, equilibrium frequencies, and substitution rate matrix). The clade-specific test is analogous

to the one for protein-coding sequences, with one *r* parameters for the null model and two for the alternative model (one for the foreground and one for the background branches). It is equivalent to the "subtree test" implemented in the phyloP program (Pollard et al. 2010), and this program was used for parameter estimation and *P* value computation. In this case, we focus on the values of the *r* parameters for interpretation.

Results

Genome Characterization and Gene Content

The single circular chromosome of SDD strain ATCC_27957 contains 2141837 bp and 2,107 protein-coding sequences (CDS) with a G + C content of 39.3% and that of SDE1 strain ATCC_12394 contains 2159491 bp and 2,070 CDS with a G + C content of 39.5% (supplementary fig. S3, Supplementary Material online). All three genomes have a clear shift in GC skew, likely correlated with the origin of replication (Lobry 1996; Touchon and Rocha 2008). The number of predicted operons for each of SDD, SDE1, and SDE2 were 457, 456, and 442, respectively. The number of tRNAs was 57 and the number of rRNA operons was 5, for all three genome sequences.

Homology comparisons of the three SD (inclusive of the two subspecies SDE, SDD) genomes against one another indicated that the chromosomes were highly syntenous along almost their entire lengths (supplementary fig. S4, Supplementary Material online). The all-against-all BlastP comparison followed by OrthoMCL yielded 9,053 protein families containing individual proteins from the 45 strains (see supplementary table S2, Supplementary Material online). Of the 9,053 protein families, 7,442 were absent in the three SD strains. An analysis of gene content for the three SD genomes reveals that a set of 1,471 proteins were common to SDD and SDE, 305 proteins were unique to SDD, 194 proteins were unique to SDE, and 254 and 241 genes were unique to each of the two SDE strains (supplementary fig. S5, Supplementary Material online). Differences in gene content (binary data for presence or absence of different protein families) among these 45 *Streptococcus* genomes, represented by a UPGMA dendrogram (fig. 1) indicates the overall gene content of SD is most similar to *S. pyogenes*, and the two strains of SDE, although showing more variation in gene content than is typical of different strains of *S. pyogenes*, are nonetheless more similar to one another than to anything else (fig. 1). Enrichment tests across functional categories, based on pairwise comparisons of these strains and *S. pyogenes*, identified only a few significantly underrepresented or overrepresented (depending on the pairwise comparison) JCVI mainrole categories (Davidsen et al. 2010) and, in particular, included "mobile and extrachromosomal element functions" and "cellular processes" (supplementary fig. S6, Supplementary Material online). The JCVI mainrole

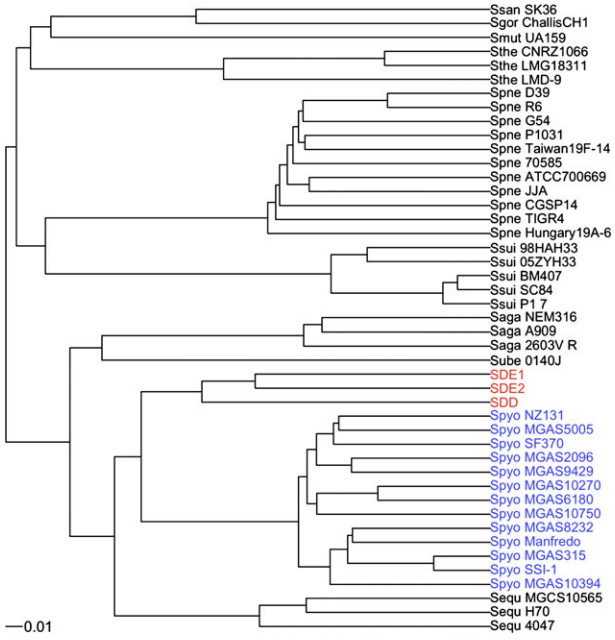

Fig. 1.—A dendrogram constructed by hierarchical clustering (UPGMA) based on dissimilarities in gene content (binary data for presence or absence of protein families) among the 45 *Streptococcus* strains. The dissimilarities were measured using Jaccard distance (one minus the Jaccard coefficient), ranging from 0 to 1, represented by the horizontal bar at the base of the figure. Species that comprise the primary focus of this paper appear in color. Species abbreviations are as follows: Ssan, *Streptococcus sanguinis*; Sgor, *Streptococcus gordonii*; Smut, *Streptococcus mutans*; Sthe, *Streptococcus thermophilus*; Spne, *Streptococcus pneumoniae*; Ssui, *Streptococcus suis*; Saga, *Streptococcus agalactiae*; Sube, *Streptococcus uberis*; Spyo, *Streptococcus pyogenes*; Sequ_MGCS10565, *Streptococcus equi* subsp. *zooepidemicus*; Sequ_H70, *Streptococcus equi* subsp. *zooepidemicus*; Sequ_4047, *Streptococcus equi* subsp. *equi*.

category mobile and extrachromosomal element functions and subrole categories "prophage functions" and "pathogenesis" were underrepresented in SDE1 relative to SDE2 and SDD, consistent with the presence of prophages in these latter two strains.

To identify core virulence genes and variations in gene sets among the 45 *Streptococcus* strains, we examined presence and absence, as well as divergence, of 88 *Streptococcus* virulence genes retrieved from the VFDB (fig. 2 and supplementary fig. S7-*A*, Supplementary Material online). *Streptococcus pyogenes*, *S. equi*, and SD clustered together based on VFDB gene content as did SDE1 and SDE2. A number of virulence loci were unique to the *S. pyogenes*, *S. equi*, and SD group and a number were uniquely absent. Several VFDB loci were present in SDE and *S. pyogenes* but absent from SDD. These included streptolysin O (*slo*) and streptokinase A (*ska*). Similarly, there were a few VFDB loci common to SDE2 and many of the *S. pyogenes* isolates, but absent from SDE1, suggesting that they might be relevant to the

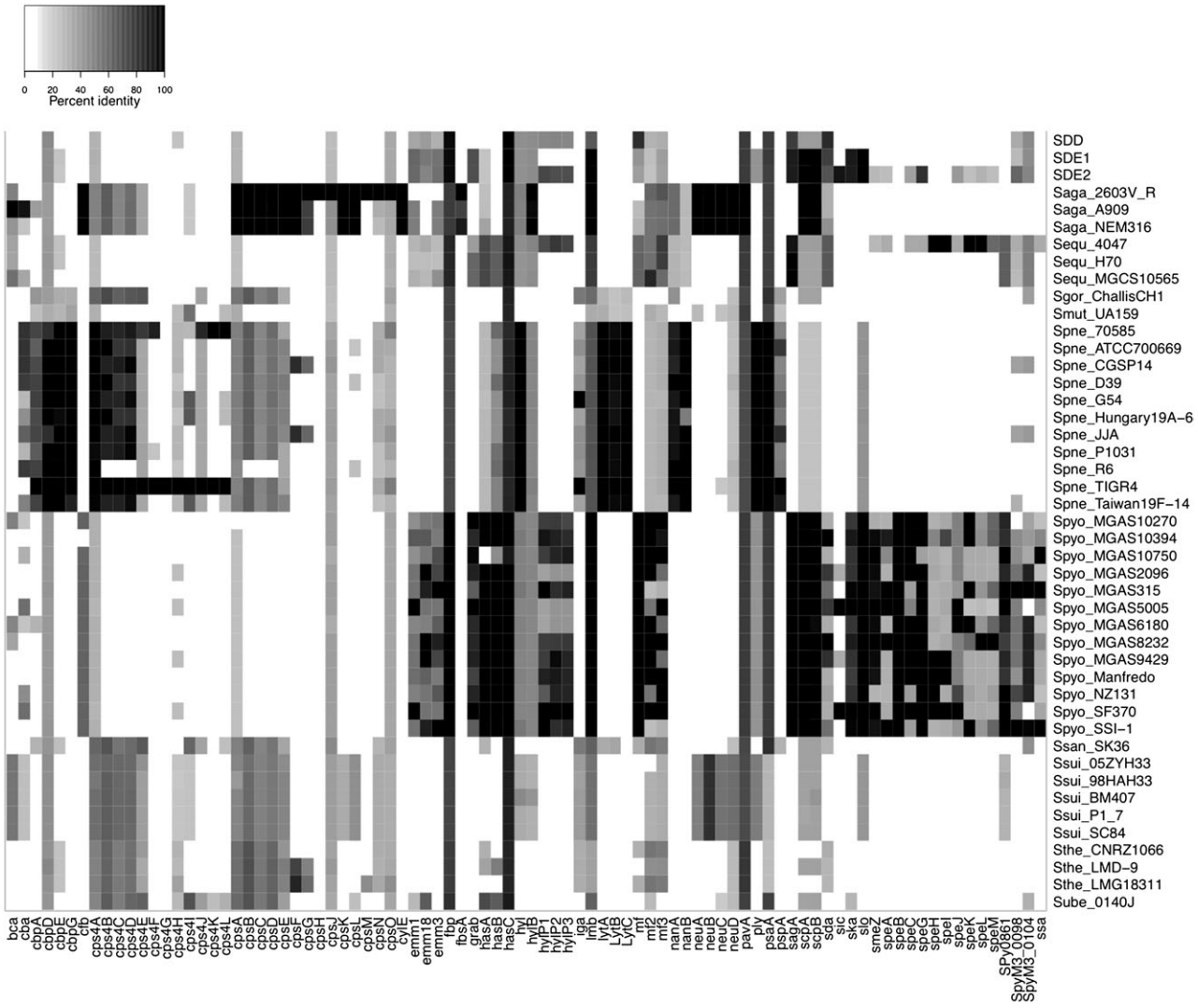

Fig. 2.—Heatmap showing % identity of Blast best hit of the 45 *Streptococcus* proteomes, against the 88 *Streptococcus* virulence genes from VFDB.

disease phenotype of SDE2. These included hyaluronidase (*hylP1*, *hylP2*, and *hylP3*) and exotoxin G (*speG*). A broader comparison that includes not only the *Streptococcus* VFDB genes but instead all the loci from this database indicates that the three strains carry between 84 and 92 VFDB orthologous genes. Comparative analysis of the set of putative *Streptococcus* virulence genes from Davies et al. (2007) (211 orthologous genes) resulted in a similar clustering to that based on the VFDB (supplementary fig. S7-*B*, Supplementary Material online). A complete list of virulence loci from each of VFDB and the Davies et al. set of 211, for this set of 45 *Streptococcus* species, appears in supplementary table S2 (Supplementary Material online) (column VFDB and Davies, respectively).

A well-known virulence factor in *S. pyogenes* is the M protein. Recently, it has been suggested that the M protein is part of a 47-kb pathogenicity island, of assumed ancient origin, due to its presence in all currently genome sequenced strains of *S. pyogenes* (Panchaud et al. 2009). We used TBlastN to look for the presence of this pathogenicity island in each of SDE1, SDE2, and SDD. Although the majority of the genes are present in all three of these strains (fig. 3), they are not present as a contiguous island, as is the case in all available genome sequences of *S. pyogenes* strains. Subsets of this island, comprising several contiguous genes, are present and this is more the case in SDE than in SDD. Furthermore, the level of similarity of the individual loci to a reference *S. pyogenes* strain (SF370) tends to be significantly less for SDD than for SDE. An inhibitor of complement-mediated lysis (Spy_2016) was present in SDE2 but absent in SDE1. All three strains possessed the M protein but again it was more similar between SDE and *S. pyogenes* than between SDD and

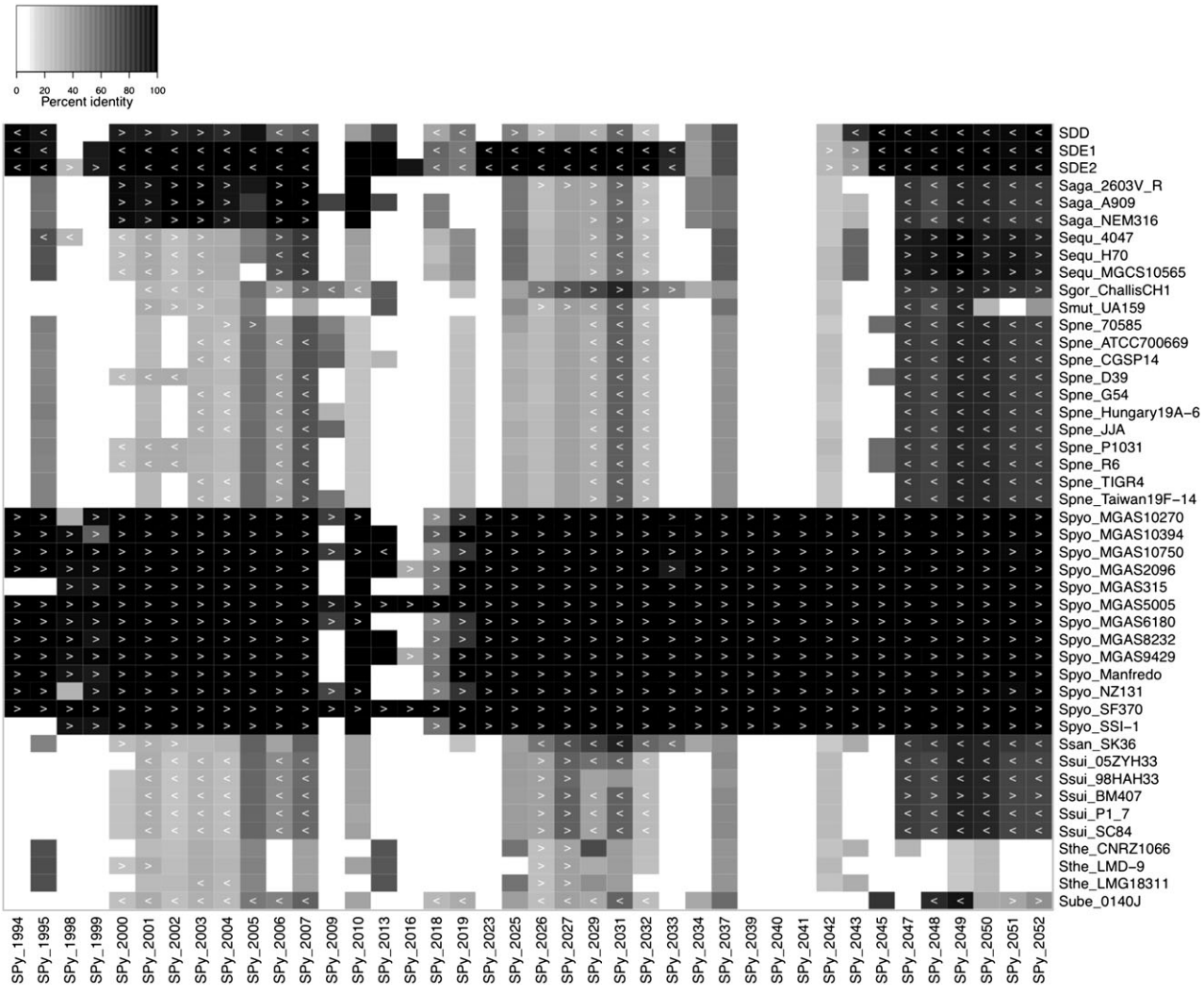

Fig. 3.—Heatmap showing % identity of Blast best hit of the 45 *Streptococcus* proteomes, against the 47 kb pathogenicity island in *Streptococcus pyogenes* SF370; arrows refer to gene orientation.

S. pyogenes. The VFDB has three genes encoding M protein, emm1, emm3, and emm18, which are found in *S. pyogenes* serotype M1 (SF370), M3 (MGAS315), and M18 (MGAS8232), respectively. Based on OrthoMCL, emm1 and emm18 are regarded as orthologs and present in SDE; emm3 is regarded as a distinct ortholog and present in SDD (supplementary table S2, Supplementary Material online). The M protein of SDD has low % identity to all three M proteins (fig. 2; emm3, 31.53; emm18, 39.25; emm1, 34.82), as reported previously (Brandt and Spellerberg 2009).

Prophinder (Lima-Mendez et al. 2008) detected two putative prophage regions in the chromosomes of SDD and SDE2. Prophage regions were not identified in the genome sequence for SDE1. These SDD/SDE2 prophages showed varying degrees of sequence similarity and were homologous to prophage from *S. pyogenes*. The two prophages in SDE2 were similar to the M3 GAS phages 315.3 and 315.5, with mean percent nucleotide sequence identity of 90.55% and 92.95%, respectively for the regions homologous to the *S. pyogenes* versions of these elements (fig. 4); six putative virulence genes (based on the Davies et al. compilation and the VFDB) were associated with these elements. Our sequence for SDD contains two putative prophages, both of which have homology to the M3 GAS phage 315.3 (fig. 4). The SDD prophages had 90.17% and 90.59% sequence identity for the regions homologous to 315.3; eight putative virulence genes (Davies et al.) were associated with these elements. The 315.3-like prophages in SDD and SDE2 are integrated in different regions (supplementary fig. S3, Supplementary Material online).

An abundance of duplicated regions were apparent from homology comparisons of the SDD and SDE genomes against themselves. These duplications primarily reflect

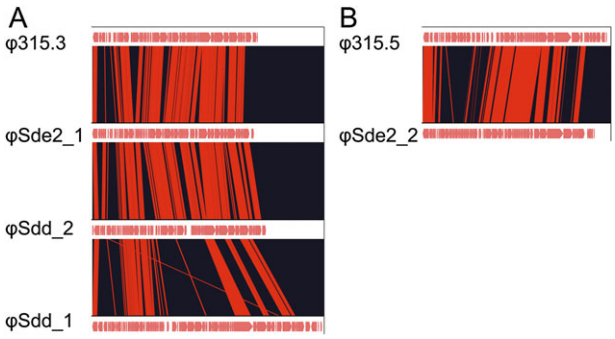

Fig. 4—Pairwise comparisons of prophages from SDD (φSdd_1 and φSdd_2) and SDE2 (φSde2_1 and φSde2_2) and *Streptococcus pyogenes* MGAS315 (φ315.3 and φ315.5). The colored bars separating sequences (red and green) represent similarity matches identified by Blast analysis. Red lines link matches in the same orientation; green lines link matches in the reverse orientation.

a proliferation of transposons, particularly in comparison to the genomes of *S. pyogenes*. The number of orthologous genes annotated as transposase in each of SDD, SDE1, and SDE2 are 45, 27, and 69 (93, 60, and 105 copies), whereas the numbers present in genomes of *S. pyogenes* range from 9–31 (11–47 copies) (supplementary table S1, Supplementary Material online). This ranks SDE and SDD among the top of *Streptococcus* genomes in terms of transposase abundance, along with strains of *S. pneumoniae*. Recent work has identified a type of conjugative transposon, referred to as ICE, as moderately abundant in *S. agalactiae* genomes (Brochet et al. 2008), with a variant (ICESde3396) present within the genome of SDE (strain NS3396; Davies et al. 2009). Comparisons of the present strains of SDE and SDD reveal the presence of an ICE in SDE1 that is similar to ICESde3396 (fig. 5; 53 hits ranging from 29% to 100% identity). Differences in gene content between the ICE from SDE1 and ICESde3396 are primarily associated with an internal region encompassing 13 loci unique to ICESde3396 including an arsenic resistance operon, permease, recombinase, lysin, cell wall hydrolase, and holin. Blast searches of *Streptococcus* genomes, with ICE identified from other species, indicate partial ICEs are widely distributed throughout the genus, however, more complete elements are relatively uncommon (supplementary fig. S8, Supplementary Material online). We find no evidence for ICESde3396 in SDE2, whereas SDD does possess truncated versions (fig. 5). At least one gene of ICESde3396 (ICESde3396_52) can encode a putative virulence factor: a surface associated agglutinin receptor described in other streptococci (Brady et al. 1992). This gene was detected in SDE1 (fig. 5).

The current version of our *Streptococcus* Genome Browser (fig. 6) is publicly available at: http://strep-genome.bscb.cornell.edu/cgi-bin/hgGateway. It makes use of the SDE1 genome as reference and shows the SDE2 and SDD genomes in alignment with it. Alignments with three outgroup species are also shown: *S. pyogenes* strains MGAS315 (SPY1) and MGAS10750 (SPY2) and *S. equi subsp. equi* strain 4047 (SEE). At present, this browser has several mapping and sequencing tracks (G + C content, GC, and AT skew), gene-related tracks (predicted genes and transcriptional units), and comparative genomic tracks (pairwise and multiple alignments, conservation scores). A track showing the genes predicted to be under PS, as described below, is also available. In addition to standard browsing capabilities (zooming and scrolling, selecting tracks of interest, searching by gene name), the browser supports a flexible query and download interface (the Table Browser), rapid sequence searches via BLAT (a description of the differences between Blast and BLAT can be found at: http://genome.ucsc.edu/FAQ/FAQblat.html), and other standard UCSC Genome Browser features.

Evolution of Protein Sequences

Molecular Selection of Protein-Coding Loci. A total of 1,066 single-copy core orthologous loci were identified by OrthoMCL from the genomes of SDE1, SDE2, SDD, two strains of *S. pyogenes*, and *S. equi* subsp. *equi*. The number of genes retained after Probalign was 1,042, which in turn resulted in 1,253 GARD fragments. Of these fragments, 464 were judged to be recombinant, leaving 789 nonrecombinant fragments (including 673 genes) for PS analysis. The total number of recombinant CDS (those with at least one recombinant fragment) was 367, representing 34% of the core genome. From the set of 789 nonrecombinant fragments, a total of 68 genes or gene fragments were judged to be under PS (P-value < 0.05); five genes were under selection on more than one lineage (cvpA, colicin V production protein; purD, phosphoribosylamine-glycine ligase; purN, phosphoribosylglycinamide formyltransferase; recX, recombination regulator; citG, triphosphoribosyl-dephospho-CoA synthase; supplementary table S3, Supplementary Material online). Selection was most evident on the branch leading to the SD ancestor with 40 significant cases (P-value < 0.05; supplementary table S3, Supplementary Material online). There were 10 and 11 instances of PS on the branches leading to SDD and SDE, respectively, with 8 on SDE1 and 4 on SDE2 (supplementary table S3, Supplementary Material online). The genes under PS were distributed among the broad COG categories as follows: 19 cases involving information storage and processing; 8 involving cellular processes, 23 involving metabolism, and the remainder poorly characterized. However, after a strict correction for multiple comparisons, only 6, 3, and 1 genes were judged to be significant (FDR adjusted P-value < 0.05) in the SD, SDE1, and SDE2 branches, respectively. The genes on the SD branch that passed the statistical correction included citrate lyase ligase, enolase, translation elongation factor G (EF-G), ribosomal protein L16/L10E,

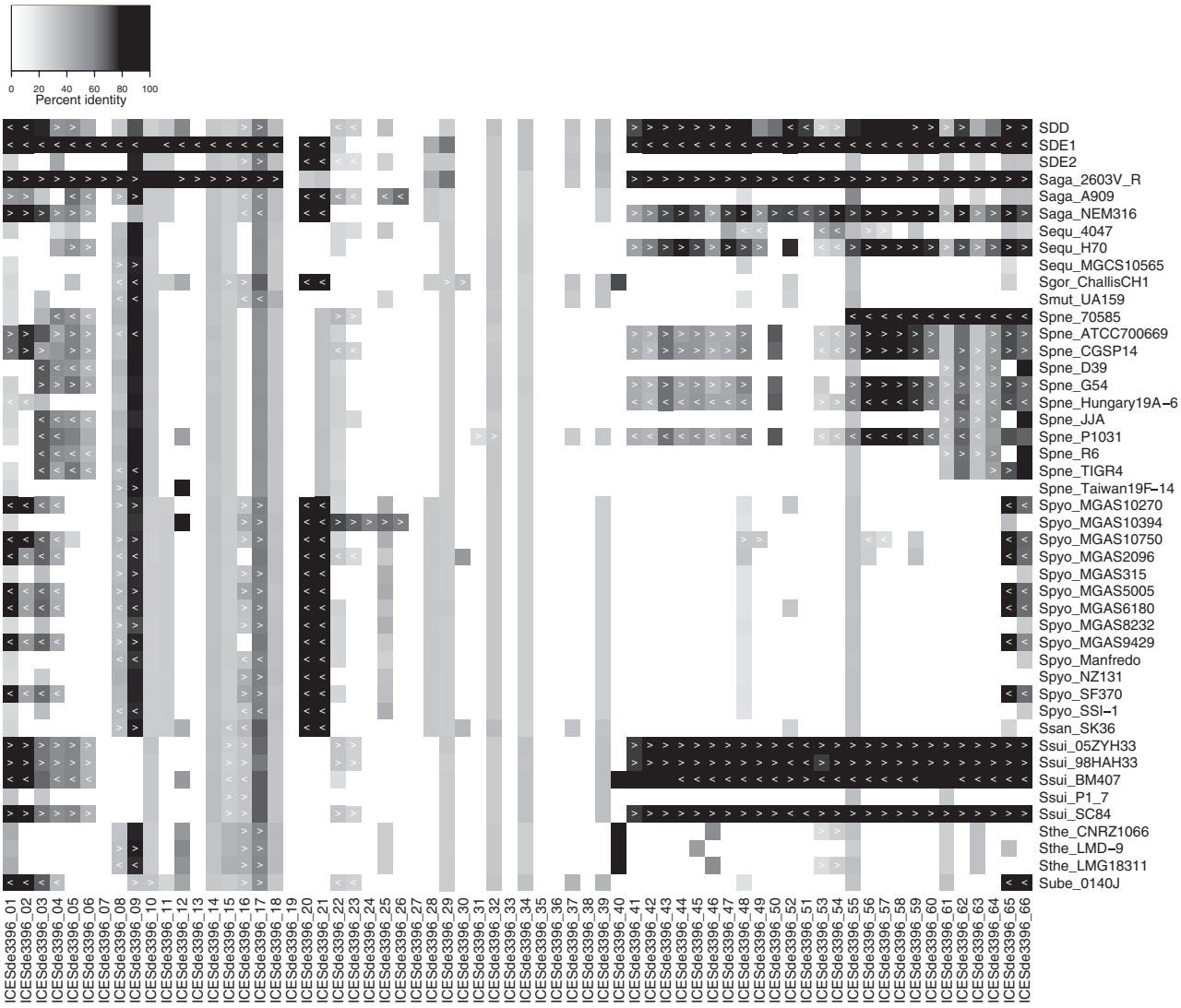

Fig. 5.—Heatmap showing % identity of Blast best hit of the 45 *Streptococcus* proteomes, against the ICEs from SDE NS3396 (ICESde3396); arrows refer to gene orientation.

metal-dependent hydrolase, and a predicted esterase. The genes on the SDE1 branch, included triphosphoribosyl-dephospho-CoA synthetase and ribosomal proteins L10 and L21. The single gene on the SDE2 branch was an arginine–ornithine antiporter.

Evolutionary Rates of Protein-Coding Sequences. We examined average rates of protein-coding evolution in these genomes, using $\omega = d_N/d_S$, the ratio of nonsynonymous to synonymous substitution rates, as a general-purpose indicator for the long-term impact of natural selection on protein-coding sequences. The average ω for all genes and all branches of this six species phylogeny is strikingly low ($\omega = 0.064$), implying that strong purifying selection has dominated during their evolution. However, there are statistically significant differences between gene categories in these average rates, with category-specific estimates of ω ranging from about $\omega = 0.02$ to $\omega = 0.12$ (fig. 7). Several of the categories that exhibit increased rates of evolution are associated with metabolism and biosynthesis. Other fast-evolving categories relate to antibiotic responses or the extracellular region. Nevertheless, purifying selection appears to dominate, with $\omega < 0.12$ for all categories.

We also estimated a separate value of ω for each branch of the phylogeny, (supplementary fig. S2-*A*, Supplementary Material online). Not surprisingly, the long internal branches of the tree (on which most substitutions have occurred) have ω estimates similar to the genome-wide average. However, the external branches of the phylogeny—leading to the three SD taxa, SDD, SDE1, and SDE2, and to the two *S. pyogenes* strains SPY1 and SPY2—have ω estimates 2–4 times as large as the average. Among the external branches, the

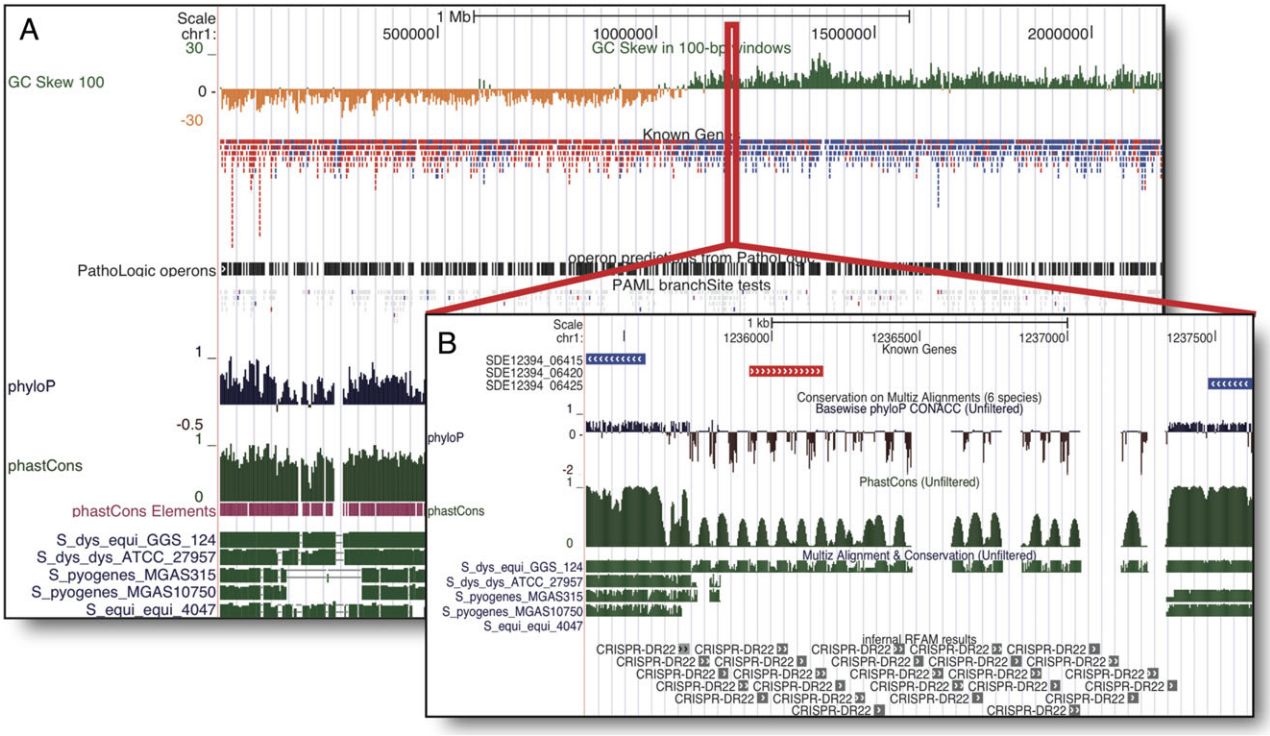

Fig. 6.—(*A*) *Streptococcus* Genome Browser (http://strep-genome.bscb.cornell.edu/cgi-bin/hgGateway). The reference genome for the browser is SDE ATCC 12394 (denoted SDE1 in this article). The two other *dysgalactiae* strains, SDE GGS 124 (SDE2) and SDD are shown via alignments with SDE1, as are two *Streptococcus pyogenes* strains (MGAS315 and MGAS10750) and an *S. equi equi* strain (4047). Selected tracks from the browser are shown, including a measure of G + C skew in 100-bp windows [computed as (C − G)/(C + G)], the gene annotations for SDE1, predicted operons from PathoLogic, genes predicted to be under PS using PAML, conservation scores produced by the phyloP (Pollard et al. 2010) and phastCons (Siepel et al. 2005) programs, predicted conserved elements from phastCons, and genome-wide multiple alignments produced with the multiz program (Blanchette et al. 2004). Notice the pronounced correlation between the direction of replication and both the G + C skew and the direction in which genes are transcribed (the origin of replication is at position 0; genes in red are transcribed on the positive strand and genes in blue on the negative strand), as has been observed with other *Streptococcus* genomes (Ferretti et al. 2001). Other tracks not shown here include predicted transcription factor binding sites and RNA genes, and alignment chains and nets revealing regions of conserved synteny (Kent et al. 2003). (*B*) Illustration of lineage-specific evolutionary patterns evident from aligned SD genomes. Shown is a browser display of a cluster of CRISPR-DR22 noncoding RNAs, which were annotated using the Rfam database and INFERNAL software (Gardner et al. 2009). Notice the relatively high levels of conservation inside the CRISPR elements (green peaks in phastCons track), contrasting with high levels of divergence in the spacer regions between them (red downward spikes in phyloP track), which is typical for CRISPR elements (Marraffini and Sontheimer 2010). This array of noncoding RNAs appears to be present in both sequenced SDE genomes but does not align with the other genomes because of extensive rearrangements or gains and losses of elements or because high levels of sequence divergence prohibit an alignment from being obtained.

short branches leading to the two *equisimilis* subspecies, SDE1 and SDE2, show particularly elevated rates of amino acid evolution, with ω estimates near 0.2.

Next, we looked at differences among GO categories in clade-specific evolutionary rates, by contrasting ω estimates for particular clades of the phylogeny with background estimates for the remainder of the tree (see Materials and Methods; supplementary fig. S2-*B*, Supplementary Material online). Categories of genes showing the most rapid evolution in SDD/SDE included "transcriptional regulator activity," "two-component response regulator activity," "phosphorylation," "biosynthetic process" (fig. 8). Similar categories were enriched with respect to the SDD and SDE lineages, "cell wall" appeared among the fast-evolving categories for the SDD lineage, possibly because the cell wall is often a trigger for immune responses to bacterial pathogens. By contrast, many of the categories showing the least rapid evolution in the *dysgalactiae* lineages reflected housekeeping activities expected to be conserved, such as "structural constituent of ribosome," "translation," and "protein transport."

Evolution of Promoter Sequences

To examine rates of evolution in promoter regions in our genomes, we grouped predicted genes into TUs and then identified upstream regions between TUs as putative promoter sequences. We associated each promoter with the union of GO categories of its component genes and then performed a series of category-specific tests of evolutionary rates in promoter regions. These tests were similar to our tests of coding regions, but instead of contrasting nonsynonymous

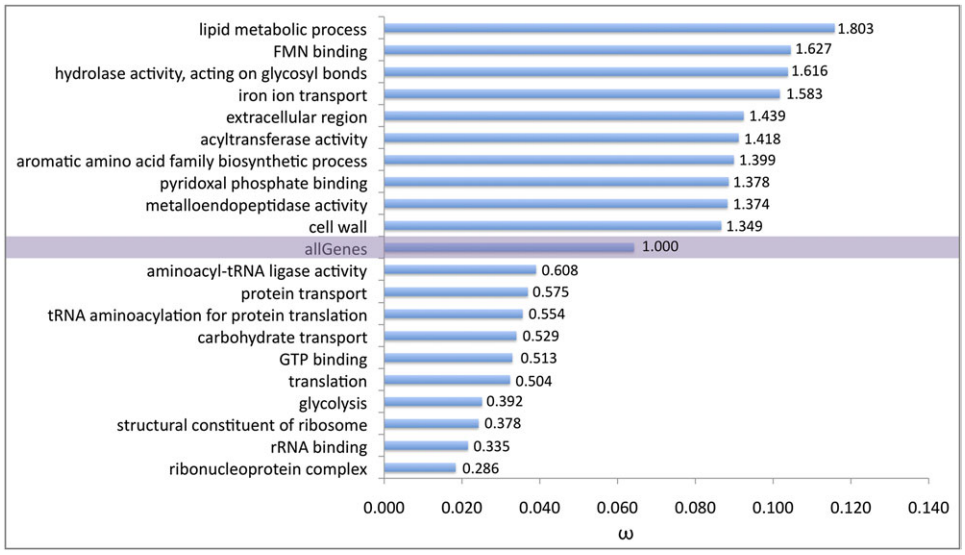

Fig. 7.—Rates of protein evolution. Estimates of ω (d_N/d_S) across all branches of the six-species phylogeny for all 673 genes (highlighted in purple) and for genes assigned to each of several Gene Ontology categories. Shown are the estimates for all genes (middle), the ten fastest-evolving categories (top), and the ten slowest-evolving categories (bottom). The numbers at right indicate ratios with respect to the estimate for all genes. For all categories shown, the differences from the average are highly statistically significant ($p0$), according to a LRT (see Materials and Methods).

and synonymous substitution rates, they contrasted the overall rate of evolution in each promoter region with a "neutral rate" estimated from 4-fold degenerate (4D) sites in coding regions (see Materials and Methods). As with the coding regions, we considered both the overall rate of evolution, across all branches of the phylogeny, and the rates in particular clades relative to the remainder of the tree.

We found that, on average, promoter sequences have experienced fairly strong purifying selection, with evolutionary rates only 37% as high as those in 4D sites. However, there is substantial variation among functional categories in the evolutionary rates of promoter sequences (fig. 9). Many of the categories showing increased rates of promoter evolution are similar to those associated with rapid evolution in protein-coding regions, such as cell wall, "extracellular region," pathogenesis, and "DNA metabolic process." Transcription factors are heavily represented among the genes whose promoters have significantly reduced rates of evolution (as indicated by categories such as "transcription factor activity," "regulation of transcription," and "binding").

The clade-specific tests indicated that, as with protein-coding sequences, promoter sequences have, on average, evolved somewhat faster in the *dysgalactiae* species than in the remainder of the phylogeny. Promoter sequences have evolved at roughly 1.3 times background rate in the SDD/SDE clade and at 1.5 times the background rate on the SDD lineage (fig. 10), suggesting some mixture of increased PS and relaxation of constraint. By contrast, promoters have evolved at slightly reduced rates in the SDE lineages. Many of the categories of promoters that are evolving most rapidly in the *dysgalactiae* lineages relative to the background rate are similar to the fast-evolving categories overall, in both protein-coding and promoter sequences. These include several categories related to metabolism, biosynthesis, and pathogenesis. Two new categories that emerge are "barrier septum formation" and "cell cycle," suggesting adaptation in the fundamental processes of cell division and cytokinesis, perhaps relating to changes in growth rate. These signals appear to come primarily from the SDD lineage because they are not evident in the SDE tests.

Discussion

This study provided an evolutionary genomic assessment of the roles of gene content, molecular adaptation of protein-coding loci, and the evolution of control region sequences in the diversification and adaptation of the SD species group. In several instances, this also led to the identification of genes or classes of genes that could be linked to the different pathogenic properties of the two strains of SDE and the taxon SDD.

The gene content of these closely related strains was very similar, with 12–16% of their genomes unique, and no obvious biochemical functional category differentially associated with these unique portions. Arguably, the most significant gene content difference between all three of these strains lies in the assortment of virulence loci that each carries, and this appears to be at least partially linked to the presence of mobile elements. The prophages of SDD and

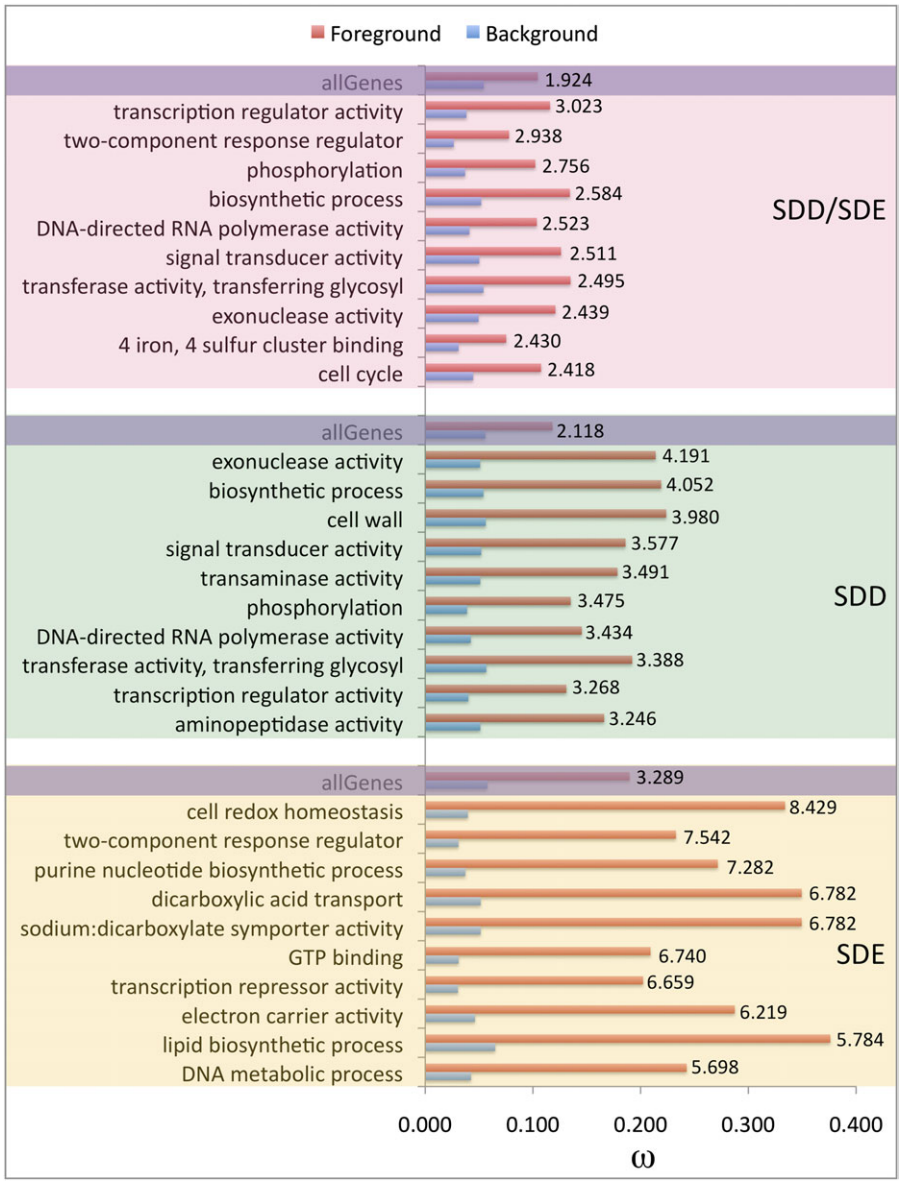

Fig. 8.—Rates of clade-specific protein evolution. Estimates of ω (d_N/d_S) for three clades of interest (foreground; see supplementary fig. S2-*B*, Supplementary Material online) versus estimates for the remaining branches of the tree (background). Shown are estimates for all genes and for the ten GO categories showing the greatest increase in ω per clade. The categories are ranked by the ratio of foreground:background ω estimates (see labels at right). All these differences are highly statistically significant ($p0$) by a LRT (see Materials and Methods).

SDE2 carry a number of prophage proteins that are considered putative virulence genes in *S. pyogenes*, including hyaluronidase (in the case of SDD and SDE2) and streptodornase type D (in the case of SDD), both of which are documented as *Streptococcus* virulence genes (Davies et al. 2007). If one broadens the consideration of virulence loci to include genes identified as having virulence phenotypes in other genera of bacteria (based on VFDB), there are a number of further putative virulence loci associated with the prophage in both SDE2 and SDD, including, for example, an amidase protein in SDD involved in peptidoglycan biosynthesis (N-acetylmuramoyl-L-alanine amidase). These SD phage elements display considerable similarity with M3 GAS phages 315.3 and 315.5, suggesting a history of phage LGT involving these taxa. An earlier report by Davies et al. (2007) had proposed LGT involving M3 GAS phage 315.1 and SDE. Our genome sequence data indicates that this history has also included SDD and involved elements other than 315.1. SDD and SDE2 share 315.3 homologous elements, but they are integrated in different regions suggesting their LGT independence. SDE and *S. pyogenes* share the same host and thus their participation in interspecies LGT

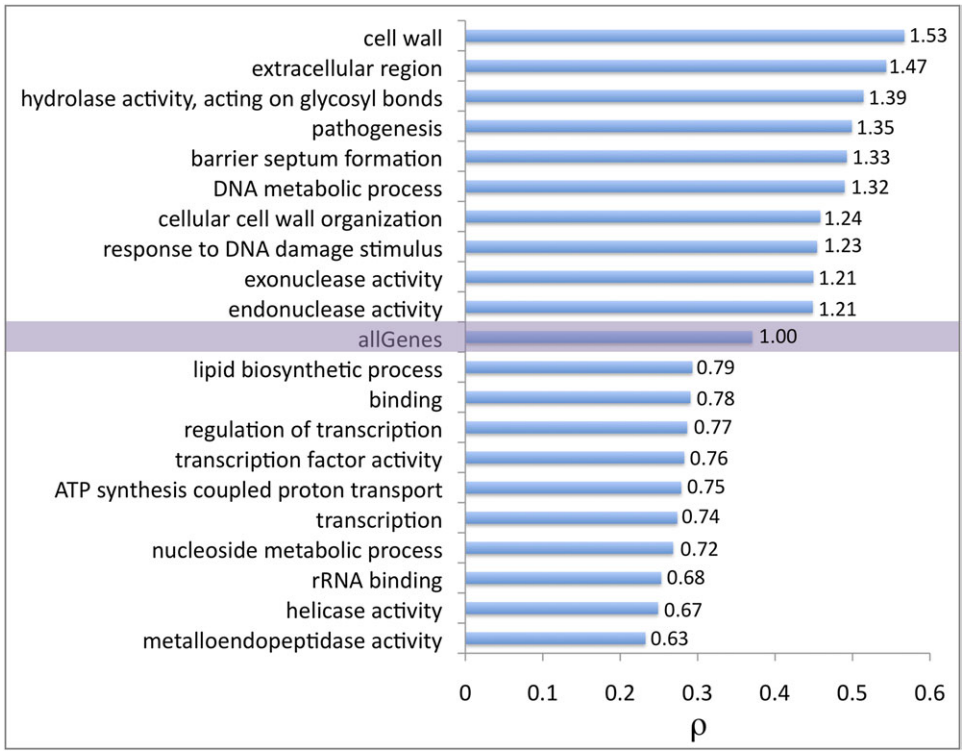

Fig. 9.—Rates of promoter evolution. Estimated rates of evolution for promoter sequences as a fraction of the neutral rate (r). The neutral rate is estimated from 4-fold degenerate (4D) sites in coding regions, and the parameter r is estimated as a scaling factor by maximum likelihood (see Materials and Methods). Estimates are shown for all genes (middle), the ten fastest-evolving GO categories (top), and the ten slowest-evolving GO categories (bottom). Promoters were defined as upstream sequences of predicted transcriptional units and were assigned the GO categories of all constituent genes. The numbers at right indicate ratios with respect to the rate for all genes. Notice that, on average, promoter regions evolve at about 37% the rate of 4D sites, suggesting that they have generally experienced fairly strong purifying selection. However, there is considerable variation across GO categories, with category-specific rates ranging from $r = 0.23$ (0.63 times the average) to $r = 0.57$ (1.53 times the average). The rates for the ten fastest and ten slowest categories are all significantly different from the rates at other promoters by a likelihood ratio test ($P < 10^{-11}$).

involving mobile elements is perhaps not unexpected; however, SDD is not associated with the human host. It is tempting to assume that this was an ancient event, prior to the divergence of the two subspecies, however, the different genomic positions of the elements in the two taxa, combined with their otherwise high level of genomic synteny, argues against a LGT event in the SD ancestor. Although SDE is primarily thought of as an agent of human infection, there are also reports that link it to animal disease (Laus et al. 2007). It seems likely that on occasion SDE and SDD have shared the same environment providing the opportunity of exchange of mobile elements. Whatever the actual scenario that facilitated the genetic exchange, the presence of these GAS-like phages in both SDE and SDD highlights the important role these elements can play in interspecies disease transmission.

Superantigen and streptolysin S genes are regarded as the most important virulence factors contributing to invasive streptococcal infection (Abdelsalam et al. 2010); the 12 superantigens (*smeZ*, *speA*, *speB*, *speC*, *speG*, *speH*, *speI*, *speJ*, *speK*, *speL*, *speM*, and *ssa*) are available in VFDB and were therefore included in our comparative analysis. Based on OrthoMCL analysis, the superantigen *speG* was found in SDE2 but not identified in SDD and SDE1. *speG* has been detected previously in SDE (Tanaka et al. 2008) and in some SDD isolates taken from moribund fish (Abdelsalam et al. 2010). In *S. pyogenes*, many streptococcal superantigen genes (e.g., *speA*, *speC*, *speH*, *speI*, *speK*, *speL*, *speM*, and *ssa*) are located on prophages, whereas *smeZ*, *speG*, and *speJ* are located on chromosomes (Proft et al. 2003). It has been suggested that the superantigen *speA* is key to STSS in *S. pyogenes* (Talkington et al. 1993), whereas other studies have implicated *speC* (Holm et al. 1992; Demers et al. 1993) and still others have suggested a lack of association between STSS and *speA* or *speC* (Hsueh et al. 1998). In SDD/SDE, *speG* has been identified as the important superantigen locus (Hashikawa et al. 2004; Zhao et al. 2007; Brandt and Spellerberg 2009; Abdelsalam et al. 2010). The presence of *speG* in SDE2 and its absence in SDE1 is correlated with the presence and absence, respectively, of STSS in these strains. Although *speG* is chromosomal, in the SDD fish isolates, it has been linked to two

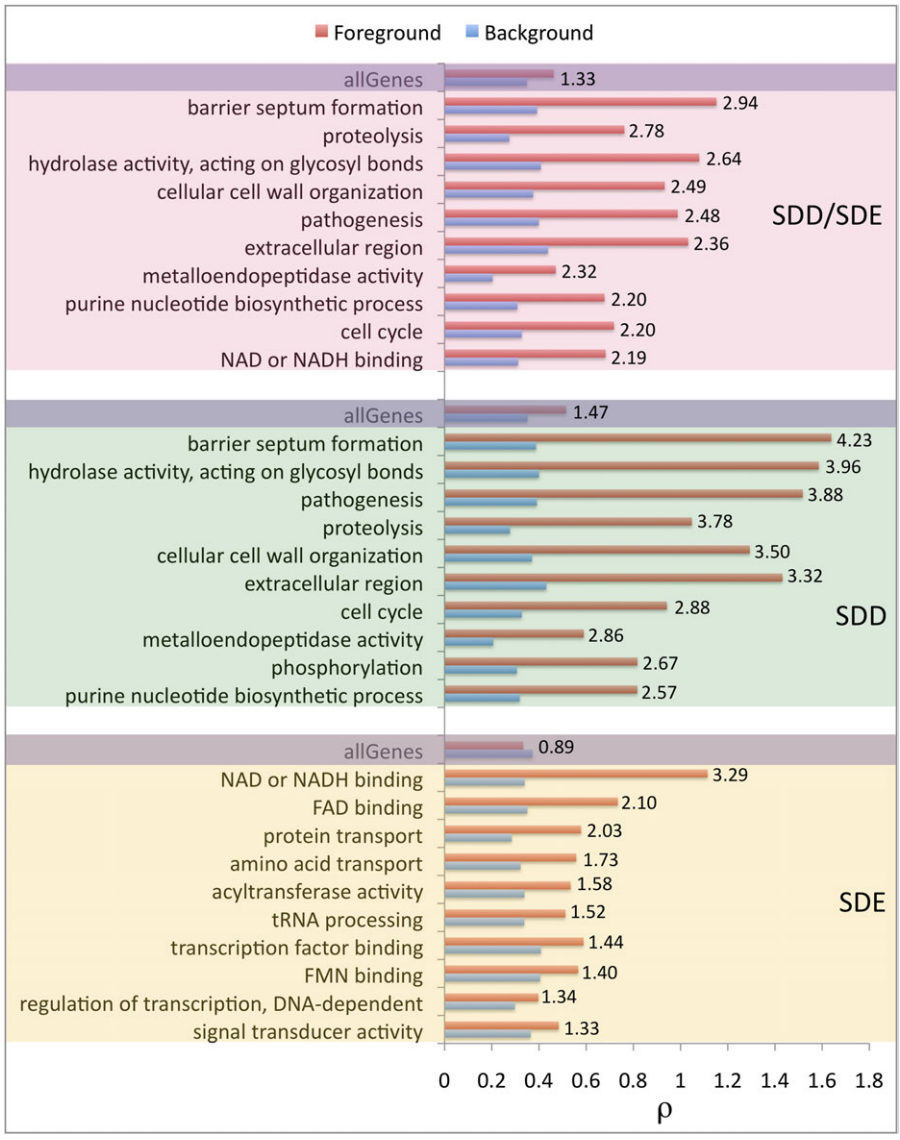

Fig. 10.—Rates of clade-specific promoter evolution. Estimated rates of promoter evolution for three clades of interest (foreground; see supplementary fig. S2-*B*, Supplementary Material online) versus estimates for the remaining branches of the tree (background). All estimates are obtained by maximum likelihood and are relative to a neutral rate estimated from 4D sites. The categories are ranked by the ratio of foreground:background rates (see labels at right). The ten categories with the highest ratios for each clade are shown. Notice that rates are somewhat elevated in SDD/SDE and SDD across all genes, but they are much more elevated for certain categories than for others. The foreground rate is significantly different from the background rate in all cases by a LRT, except for "FMN binding" in the SDE clade ($P = 0.06$).

IS elements (IS981SC and IS1161) (Abdelsalam et al. 2010) which are prevalent throughout *Streptococcus* (Bourgoin et al. 1999; Lowe et al. 2007; McShan et al. 2008); however, we find no evidence that *speG* in SDE2 is associated with an IS element.

Another mobile element found in the genome of SDD, and in this case SDE1, is known as the ICE, recently reported to be abundant in *S. agalactiae* (Brochet et al. 2008). ICE and related elements (*cis*-mobilizable element; CIME) have been found in various *Streptococcus* species, including ICE*St1*, ICE*St3*, CIME302, DeltaCIME308, and CIME19258 from *S. thermophilus* (Pavlovic et al. 2004), ICE*SE2* from *S. equi* subsp. *equi* strain 4047 (Heather et al. 2008), and ICE-*Sa*2603 from *S. agalactiae* 2603V/R, which is closely related to ICE*Sde*3396 from SDE NS3396 (Davies et al. 2009). Strain SDE1 possesses an ICE related to ICESde3396, differing primarily in the absence of an internal region including about 13 loci. SDD carried multiple, but smaller, ICEs. We find evidence for only remnant ICEs in *S. pyogenes* genomes but did detect somewhat more complete elements in *S.*

pneumoniae and *Streptococcus suis*. ICEs have been reported previously in human strains of *S. suis* and noted to carry antibiotic resistance loci (Holden et al. 2009). ICEs are common in *S. agalactiae*, and *S. pyogenes* can readily accommodate prophage. SDE and SDD appear to harbor both types of mobile elements, suggesting that they can act as donor/recipient of DNA via both elements and involving both species. Haenni et al. (2010) have shown that a high proportion of bovine strains of *S. agalactiae* have functional ICEs (Haenni et al. 2010). *Streptococcus agalactiae* and SDD are both agents of bovine mastitis, and thus, it is likely they often coexist in the same environment, providing a means for genetic exchange. Recent reports have linked SDD to cases of STSS-like syndrome in bovine (Chenier et al. 2008). Because SDD can accommodate both prophage and ICEs, combined with a frequently sympatric distribution with *S. agalactiae*, we suggest this taxon might be expected to develop a variety of virulence attributes.

The branch-site test indicated a total of 73 genes were under PS (P value < 0.05), however, after a correction for multiple comparisons, only ten of these were significant (FDR adjusted P value < 0.05). A number of the genes from this latter set are reported to be linked to virulence in other bacteria. For example, enolase is an important virulence component in a number of Gram-positive bacterial pathogens, including *S. pyogenes*, *S. pneumoniae*, *S. suis*, and *Staphylococcus aureus*. In both *S. pneumoniae* and *S. pyogenes*, enolase binds plasminogen (blood clot–dissolving protein) and enhances the formation of proteolytic plasmin activity, an important feature of pathogenesis in these organisms (Bergmann et al. 2003; Sun et al. 2004). PS on the SD enolase may reflect conformation changes in plasminogen between *S. pyogenes* and SD or possibly an alternative function for SD enolase. Esterases are widespread in bacterial pathogens but their roles in virulence and pathogenesis have been somewhat unclear. Recently, however, a CovRS-regulated secreted esterase (Sse) has been shown to have a role in the virulence of GAS in subcutaneous infection, severe soft tissue infections and in systemic dissemination of GAS from the skin (Zhu et al. 2009). PS of an esterase along the SD lineage indicates the development of a specific functional significance to this locus since separation of SD from the putative SD/*S. pyogenes* ancestor. The SD esterase under selection is not an orthologue to Sse, and PS, although suggestive of functional significance, is not necessarily linked to virulence. Nonetheless, the pathogenicity of GAS Sse together with the PS results suggest this SD esterase should be given some consideration as a putative virulence factor in SD.

Several of the genes under PS are involved in the process of translation, including three different ribosomal proteins and translation EF-G. Translation proteins are known to be expressed constitutively at very high levels and exhibit a strong codon bias toward a subset of synonymous codons (so-called translationally optimal codons), which are those most efficiently and accurately recognized by the most abundant tRNA species in the cell (Karlin and Mrazek 2000; Henry and Sharp 2007). The efficiency and accuracy of translation is especially important for cells' rapid growth. Species exposed to selection for rapid growth have more rRNA operons, more tRNA genes, and a greater propensity for translationally selected codon usage bias (Sharp et al. 2005). Such translational selection was previously detected in *S. agalactiae* 2603V/R, *S. pyogenes* M1 GAS SF370, and *S. pneumoniae* R6 (Sharp et al. 2005). We performed a within-group correspondence analysis (Suzuki et al. 2008) and detected translational selection in all the 45 *Streptococcus* strains including the three SD strains (data not shown). It is possible, therefore, that the PS we detect for these particular proteins may be linked to this translational codon usage bias and increased growth efficiency. Coincident with this hypothesis regarding growth, our results regarding promoter evolution, suggest that the gene categories barrier septum formation and cell cycle are evolving rapidly in the SD lineage, suggesting adaptation in the processes of cell division and cytokinesis, perhaps relating to changes in growth rate.

Because our power to detect PS in individual genes was limited, we attempted to gain further insight into the evolution of the SDD/SDE proteomes by pooling data across genes, either by GO category or by branch or clade in the phylogeny. These analyses indicated that purifying selection has dominated in the evolution of the proteome of these species, but we did observe somewhat elevated rates of evolution in several gene categories that could indicate adaptation to changing environments, such as metabolism, antibiotic response, biosynthesis, and extracellular region. We also observed a fairly pronounced increase in the rate of protein evolution on external branches of the tree, particularly the branches leading to the two SDE subspecies. These elevated average rates may reflect recent PS as these species have adapted to their different niches, although they could also be influenced by differences in effective population sizes or weakly deleterious mutations not yet eliminated from populations. One interesting finding that emerged from our clade-specific analysis of protein-coding rates was that several GO categories associated with two-component signal transduction systems (TCSs)—such as two-component response regulator activity, "phyosphorylation," and "signal transducer activity"—were among those showing the most rapid evolution in SDD and SDE relative to other portions of the phylogeny (fig. 8). TCSs are known to play an important role in adaptive gene regulation in group A streptococci (Kreikemeyer et al. 2003). The CovRS TCS is particularly important as a global regulator critical in the transition to invasive infection (Churchward 2007). Therefore, these elevated rates of protein evolution

in TCS-related genes could reflect infection-related adaptive changes in SD lineages.

Although most studies of adaptive evolution have focused on protein-coding sequences, there is growing interest in rapid evolution of noncoding regions as well, particularly in *cis*-regulatory regions near core promoters (Andolfatto 2005; Haygood et al. 2007; Torgerson et al. 2009). Using simple methods, we examined evolutionary rates in the core promoters of six *Streptococcus* genomes (including SDD, SDE1, and SDE2), taking advantage of the fact that these species are sufficiently closely related that their genomes can easily be aligned in noncoding as well as in coding regions. We found clear evidence of purifying selection in these promoter regions but also found that certain categories of genes have significantly faster-evolving promoters than others. Furthermore, many of the fast-evolving categories are similar to those associated with rapid evolution in protein-coding regions (e.g., cell wall, extracellular region, pathogenesis, and DNA metabolic process), suggesting that many of the same types of genes are evolving adaptively for both protein function and gene expression. Interestingly, transcription factors were prominent among genes having reduced rates of promoter evolution, possibly because their pronounced influence on the expression patterns of multiple downstream genes decreases their flexibility for regulatory adaptation (Duret and Mouchiroud 2000; Wray et al. 2003).

This work has provided a view of bacterial pathogen evolution, involving the diversification of closely related subspecies, that suggests adaptation to changing environments and new hosts involves changes in gene content, as well as selection of orthologous protein-coding loci and operon promoters. Differences in gene content can have major influences on virulence attributes and rapidly alter the characteristics of recipient isolates through the process of LGT, particularly facilitated through mobile elements. Adaptive evolution of protein-coding genes, although undoubtedly a much slower process, is nonetheless a component of the overall differentiation of even such closely related species and may at least partly proceed through the coordinated evolution of biochemical categories of genes. Although we regard this as a very preliminary step toward an understanding of the evolution of noncoding functional elements in bacteria, it is clear that promoters in the SD species group are evolving differently in different lineages and different categories of genes and may in some instances be evolving in concert with protein-coding genes of the same biochemical category.

Supplementary Material

Supplementary figure S1–S8 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Kazuharu Arakawa for his advice on the G-language Genome Analysis Environment, to Yoshiyuki Ohtsubo for his advice on GenomeMatcher, to Paul Stothard for his advice on Circular Genome Viewer, to Gipsi Lima Mendez and Ariane Toussaint for their advice on Prophinder, and to Renee Setter for help in figure preparation. This work was supported by the National Institute of Allergy and Infectious Disease, US National Institutes of Health, under grant number AI073368-01A2 awarded to A.S. and M.J.S.

Literature Cited

Abdelsalam M, Chen SC, Yoshida T. 2010. Dissemination of streptococcal pyrogenic exotoxin G (*spegg*) with an IS-like element in fish isolates of *Streptococcus dysgalactiae*. FEMS Microbiol Lett. 309:105–113.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149–1152.

Arakawa K, et al. 2003. G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. Bioinformatics 19:305–306.

Arakawa K, Tomita M. 2006. G-language System as a platform for large-scale analysis of high-throughput omics data. J Pesticide Sci. 31:282–288.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 29:1165–1188.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. Nucleic Acids Res. 37:D26–D31.

Bergmann S, et al. 2003. Identification of a novel plasmin(ogen)-binding motif in surface displayed alpha-enolase of *Streptococcus pneumoniae*. Mol Microbiol. 49:411–423.

Bert F, Branger C, Poutrel B, Lambert-Zechovsky N. 1997. Differentiation of human and animal strains of *Streptococcus dysgalactiae* by pulsed-field gel electrophoresis. FEMS Microbiol Lett. 150:107–112.

Blanchette M, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Bourgoin F, Pluvinet A, Gintz B, Decaris B, Guedon G. 1999. Are horizontal transfers involved in the evolution of the *Streptococcus thermophilus* exopolysaccharide synthesis loci? Gene 233:151–161.

Brady LJ, Piacentini DA, Crowley PJ, Oyston PC, Bleiweis AS. 1992. Differentiation of salivary agglutinin-mediated adherence and aggregation of mutans streptococci by use of monoclonal antibodies against the major surface adhesin P1. Infect Immun. 60:1008–1017.

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

Brochet M, Couve E, Glaser P, Guedon G, Payot S. 2008. Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. J Bacteriol. 190:6913–6917.

Chen L, et al. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 33:D325–D328.

Chenier S, Leclere M, Messier S, Fecteau G. 2008. *Streptococcus dysgalactiae* cellulitis and toxic shock like syndrome in a Brown Swiss cow. J Vet Diagn Invest. 20:99–103.
Churchward G. 2007. The two faces of Janus: virulence gene regulation by CovR/S in group A streptococci. Mol Microbiol. 64:34–41.
Davidsen T, et al. 2010. The comprehensive microbial resource. Nucleic Acids Res. 38:D340–D345.
Davies MR, et al. 2007. Virulence profiling of *Streptococcus dysgalactiae* subspecies *equisimilis* isolated from infected humans reveals 2 distinct genetic lineages that do not segregate with their phenotypes or propensity to cause diseases. Clin Infect Dis. 44:1442–1454.
Davies MR, Shera J, Van Domselaar GH, Sriprakash KS, McMillan DJ. 2009. A novel integrative conjugative element mediates genetic transfer from group G Streptococcus to other {beta}-hemolytic Streptococci. J Bacteriol. 191:2257–2265.
Demers B, et al. 1993. Severe invasive group A streptococcal infections in Ontario, Canada: 1987–1991. Clin Infect Dis. 16:792–800 discussion 801–792.
Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.
Ferretti JJ, et al. 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci U S A. 98:4658–4663.
Gardner PP, et al. 2009. Rfam: updates to the RNA families database. Nucleic Acids Res. 37:D136–D140.
Glazunova OO, Raoult D, Roux V. 2010. Partial *recN* gene sequencing: a new tool for identification and phylogeny within the genus *Streptococcus*. Int J Syst Evol Microbiol. 60:2140–2148.
Grant JR, Stothard P. 2008. The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. 36:W181–W184.
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.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.
Haenni M, et al. 2010. Diversity and mobility of integrative and conjugative elements in bovine isolates of *Streptococcus agalactiae, S. dysgalactiae* subsp. *dysgalactiae*, and *S. uberis*. Appl Environ Microbiol. 76:7957–7965.
Hashikawa S, et al. 2004. Characterization of group C and G streptococcal strains that cause streptococcal toxic shock syndrome. J Clin Microbiol. 42:186–192.
Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet. 39:1140–1144.
Heather Z, et al. 2008. A novel streptococcal integrative conjugative element involved in iron acquisition. Mol Microbiol. 70:1274–1292.
Henry I, Sharp PM. 2007. Predicting gene expression level from codon usage bias. Mol Biol Evol. 24:10–12.
Holden MT, et al. 2009. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen *Streptococcus suis*. PLoS One 4:e6072.
Holm SE, Norrby A, Bergholm AM, Norgren M. 1992. Aspects of pathogenesis of serious group A streptococcal infections in Sweden, 1988-1989. J Infect Dis. 166:31–37.
Hsueh PR, et al. 1998. Invasive group A streptococcal disease in Taiwan is not associated with the presence of streptococcal pyrogenic exotoxin genes. Clin Infect Dis. 26:584–589.
Karlin S, Mrazek J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol. 182:5238–5250.
Karp PD, et al. 2009. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 11:40–79.
Kazakov AE, et al. 2007. RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. Nucleic Acids Res. 35:D407–D412.
Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 100:11484–11489.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.
Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098.
Kosiol C, et al. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144.
Kreikemeyer B, McIver KS, Podbielski A. 2003. Virulence factor regulation and regulatory networks in *Streptococcus pyogenes* and their impact on pathogen-host interactions. Trends Microbiol. 11:224–232.
Laus F, et al. 2007. Clinical and epidemiological investigation of chronic upper respiratory diseases caused by beta-haemolytic Streptococci in horses. Comp Immunol Microbiol Infect Dis. 30:247–260.
Lefébure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8:R71.
Lefébure T, Stanhope MJ. 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus. *Campylobacter* Genome Res. 19:1224–1232.
Leplae R, Lima-Mendez G, Toussaint A. 2010. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res. 38:D57–D61.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 24:863–865.
Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 13:660–665.
Lowe BA, Miller JD, Neely MN. 2007. Analysis of the polysaccharide capsule of the systemic pathogen *Streptococcus iniae* and its implications in virulence. Infect Immun. 75:1255–1264.
Marraffini LA, Sontheimer EJ. 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet. 11:181–190.
McDonald TJ, McDonald JS. 1976. Streptococci isolated from bovine intramammary infections. Am J Vet Res. 37:377–381.
McShan WM, et al. 2008. Genome sequence of a nephritogenic and highly transformable M49 strain of *Streptococcus pyogenes*. J Bacteriol. 190:7773–7785.
Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M. 2008. GenomeMatcher: a graphical user interface for DNA sequence comparison. BMC Bioinformatics 9:376.
Panchaud A, et al. 2009. M-protein and other intrinsic virulence factors of *Streptococcus pyogenes* are encoded on an ancient pathogenicity island. BMC Genomics 10:198.
Pavlovic G, Burrus V, Gintz B, Decaris B, Guedon G. 2004. Evolution of genomic islands by deletion and tandem accretion by site-specific

recombination: ICE *St1*-related elements from *Streptococcus thermophilus*. Microbiology 150:759–774.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.
Proft T, Sriskandan S, Yang L, Fraser JD. 2003. Superantigens and streptococcal toxic shock syndrome. Emerg Infect Dis. 9:1211–1218.
R_Development_Core_Team. 2010. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rhead B, et al. 2010. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 38:D613–D619.
Rolston KV. 1986. Group G streptococcal infections. Arch Intern Med. 146:857–858.
Roshan U, Livesay DR. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics 22:2715–2721.
Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33:1141–1153.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol. 21:468–488.
Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. Bioinformatics 21:537–539.
Sun H, et al. 2004. Plasminogen is a critical host pathogenicity factor for group A streptococcal infection. Science 305:1283–1286.
Sunaoshi K, et al. 2009. Molecular *emm* genotyping and antibiotic susceptibility of *Streptococcus dysgalactiae* subsp. *equisimilis* isolated from invasive and non-invasive infections. J Med Microbiol. 59:82–88.
Suzuki H, Brown CJ, Forney LJ, Top EM. 2008. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res. 15:357–365.
Takahashi T, et al. 2010. Clinical aspects of invasive infection with *Streptococcus dysgalactiae* subsp. *equisimilis* in elderly patients. J Infect Chemother. 16:68–71.
Talkington DF, et al. 1993. Association of phenotypic and genotypic characteristics of invasive *Streptococcus pyogenes* isolates with clinical components of streptococcal toxic shock syndrome. Infect Immun. 61:3369–3374.
Tanaka D, et al. 2008. Genetic features of clinical isolates of *Streptococcus dysgalactiae* subsp. *equisimilis* possessing Lancefield's group A antigen. J Clin Microbiol. 46:1526–1529.
Torgerson DG, et al. 2009. Evolutionary processes acting on candidate *cis*-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet 5:e1000592.
Touchon M, Rocha EP. 2008. From GC skews to wavelets: a gentle guide to the analysis of compositional asymmetries in genomic data. Biochimie 90:648–659.
Vieira VV, et al. 1998. Genetic relationships among the different phenotypes of *Streptococcus dysgalactiae* strains. Int J Syst Bacteriol. 48(Pt 4):1231–1243.
Wray GA, et al. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 20:1377–1419.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.
Zhao J, et al. 2007. Cloning, expression, and characterization of the superantigen streptococcal pyrogenic exotoxin G from *Streptococcus dysgalactiae*. Infect Immun. 75:1721–1729.
Zhu H, Liu M, Sumby P, Lei B. 2009. The secreted esterase of group a *streptococcus* is important for invasive skin infection and dissemination in mice. Infect Immun. 77:5225–5232.

Associate editor: Takashi Gojobori