Published in final edited form as:

Infect Genet Evol. 2012 March ; 12(2): 274–277. doi:10.1016/j.meegid.2011.11.004.

Functional bias of positively selected genes in *Streptococcus* genomes

Haruo Suzuki¹ and Michael J. Stanhope^{*,1}

¹Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA

Abstract

Rates of nonsynonymous substitution (dN) significantly higher than rates of synonymous substitution (dS) have been used as evidence of positive selection for the fixation of advantageous point mutations. It has been suggested that positive selection contributes to the evolution of virulence factors and certain functional categories in bacterial pathogens. The genus *Streptococcus* contains a number of important human and agricultural pathogens. Here we assessed positive selection across 13 *Streptococcus* species, and their relationship with virulence factors and functional categories. We found that known virulence genes were subject to positive selection pressure as much as other genes. After false discovery rate correction for multiple comparisons, no functional categories were significantly over- or underrepresented in positively selected genes relative to other genes. Our results suggest that within the genus *Streptococcus* positive selection based on dN/dS ratios is not distributed with bias across biological functions.

Keywords

Streptococcus; positive selection; gene functional categories; virulence factor

1. Introduction

Positive selection for the fixation of advantageous point mutations is an important force in the adaptation of microorganisms to different environmental niches, to optimize infection processes and to escape host immune response (Anisimova et al., 2007). There are many reports suggesting that positive selection contributes to the evolution of virulence genes or certain functional categories in bacterial pathogens such as *Escherichia coli* (Chen et al., 2006), *Streptococcus* (Anisimova et al., 2007; Lefebure and Stanhope, 2007), *Listeria monocytogenes* (Orsi et al., 2008), *Salmonella* (Soyer et al., 2009), and *Campylobacter* (Lefebure and Stanhope, 2009). A previous analysis of gene families and protein domains from a large and diverse set of species suggested that ‘evidence for positive selection was found principally for gene products with membrane or extracellular localizations, mostly involved in pathogenesis, belonging to different aspects of cell communication or to viral

© 2012 Published by Elsevier B.V.

***Corresponding author.** Michael J. Stanhope, Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA, Phone: 607-253-3859, Fax: 607-253-3083, mjs297@cornell.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

life cycles' (Aris-Brosou, 2005). It follows that proteins which have important roles in pathogenicity might be more likely to be subject to positive selection.

Streptococcus species cause a wide variety of human diseases, and possesses many different virulence genes (Chen et al., 2005; Davies et al., 2007; Yang et al., 2008). Several studies have reported that positive selection contributes to the evolution of virulence genes in *Streptococcus* species such as the gene encoding streptokinase (*ska*) in group A streptococcus (GAS) (Kalia and Bessen, 2004), the virulence gene *pauA* in *Streptococcus uberis* (Zadoks et al., 2005), several virulence genes in group B *Streptococcus* (GBS) (Springman et al., 2009), and the capsule virulence determinant of *Streptococcus pneumoniae* (Lysenko et al., 2010). It was suggested that positive selection in *Streptococcus pyogenes* was more frequent in recently transferred genes than in core genes (Marri et al., 2006). A sites test of positive selection found no evidence to support the hypothesis that pathogen-specific accessory genes are more likely to be subject to positive selection than core genes in *Streptococcus* species (Anisimova et al., 2007) although a large fraction (29%) of positively selected genes can be connected to virulence. A branch-site test of positive selection on core genes of *Streptococcus* species found positive selection pressure was unevenly distributed across lineages and functional categories, and some of the positively selected genes are known virulence factors (Lefebure and Stanhope, 2007). These previous studies led us to the hypothesis that *Streptococcus* species have a greater frequency of positive selection in virulence factors and certain functional categories compared to other genes. To test this hypothesis, we assessed positive selection across homologous sequences from the complete genomes of 13 *Streptococcus* species.

2. Materials and methods

2.1. Data collation

Genome sequence data were manipulated using the Bioperl version 1.6.1 (Stajich et al., 2002) and the G-language Genome Analysis Environment version 1.8.12 (Arakawa et al., 2003). Statistical tests and graphics were implemented using R, version 2.11.1 (R_Development_Core_Team, 2010). Complete genome sequences of 13 *Streptococcus* strains in GenBank format (Benson et al., 2011) were retrieved from the National Center for Biotechnology Information site at <ftp://ftp.ncbi.nih.gov/genomes/Bacteria>. One strain was randomly selected from species for which multiple strains have been sequenced (Table 1). All protein-coding sequences longer than 300 nucleotides were retrieved from the 13 *Streptococcus* strains (Table 1). A group of homologous proteins or protein family was built by all-against-all protein sequence comparison using BLASTP (Altschul et al., 1997) followed by Markov clustering with an inflation factor of 2.5 (van Dongen, 2000). Homologous proteins were identified by BLASTP on the criteria of an E-value cutoff of $1e-5$, and minimum aligned sequence length coverage of 50% of a query sequence. This approach yielded 3874 protein families containing 22709 individual proteins from the 13 *Streptococcus* strains. Alignments required at least four sequences for phylogenetic reconstruction. Of these 3874 protein families, 1485 were retained after excluding (i) protein families shared by less than 4 strains, (ii) those whose ratio of minimum to maximum sequence length were less than 0.5, and (iii) those related to mobile elements such as transposase and integrase.

2.2. Positive selection

The homologous sequences were first aligned at the amino acid level using Probalign (Roshan and Livesay, 2006), then backtranslated to nucleotides. Protein families including alignments with >50% of the sites removed, based on a posterior probability cut-off <0.6, were discarded from the analysis. Among the resulting 1264 protein families, the numbers of

sequences varied from 4 to 38, and sequence alignment length ranged from 303 to 5310 nucleotides, the median being 876 nucleotides (Supplementary Table S1). The alignments were tested for intragenic recombination based on the single breakpoint (SBP) analysis and KH test in the HyPhy package (Pond et al., 2005; Kosakovsky Pond et al., 2006). A phylogenetic tree was determined for each of the genes using PhyML (Phylogenetic estimation using Maximum Likelihood) (Guindon and Gascuel, 2003; Guindon et al., 2010) with the GTR + Gamma substitution model of nucleotide evolution, and the Subtree Pruning-Regrafting (SPR) branch-swapping method. The median of total branch length of the 1264 gene trees was 4.195 nucleotide substitutions per site (Supplementary Table S1). Each gene tree was used to determine its optimal nucleotide substitution model. To provide evidence of non-neutral evolution, with rates of nonsynonymous substitution (dN) significantly different from rates of synonymous substitution (dS), we used three different codon-based maximum likelihood methods in the HyPhy package (Pond et al., 2005): single-likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), and random effects likelihood (REL). To determine sites evolving non-neutrally, a significance level of 0.05 was used for the SLAC and FEL methods, and a minimum Bayes factor of 95 was used for the REL method. Functional categories from multiple databases were used to assess categories in which positively selected genes were over- or under-represented: the Clusters of Orthologous Groups (COG) (Tatusov et al., 2000; Tatusov et al., 2001), the JCVI role categories (Davidsen et al., 2010), KEGG (Kanehisa and Goto, 2000), SEED (Overbeek et al., 2005), Virulence Factors Database (VFDB) (Chen et al., 2005), and the Gene Ontology (GO) (Ashburner et al., 2000) database. We searched protein sequences against the Pfam (Finn et al., 2010) library of hidden Markov models (HMMs) using HMMER (<http://hmmer.janelia.org/>), and converted Pfam accession numbers to GO terms using the 'pfam2go' mapping (<http://www.geneontology.org/external2go/pfam2go>). We identified putative virulence factors based on a combination of SEED, VFDB, and MvirDB (Zhou et al., 2007). MvirDB integrates sequence information from multiple microbial databases of protein toxins, virulence factors, and antibiotic resistance genes.

Protein families were classified into two categories: those showing evidence of positive selection (PS^+) and those showing no evidence of positive selection (PS^*). A 2×2 contingency table was constructed for each functional category from the COG, JCVI, KEGG, SEED, VFDB, and GO databases: (a) the number of PS^+ in this category; (b) the number of PS^+ that are not in this category; (c) the number of PS^* in this category; (d) the number of PS^* that are not in this category. The odds ratio was calculated as ad/bc . Fisher's exact test was used to test significance of over- or under-representation of each functional category in PS^+ genes relative to PS^* genes, and the p-value was adjusted for the multiple comparisons by controlling the false discovery rate (Benjamini and Yekutieli 2001).

3. Results and discussion

We tested for positive selection across homologous genes from the complete genomes of 13 *Streptococcus* species using three different methods (SLAC, FEL and REL). Of the 1264 protein families tested, 5, 196, and 220 had at least one positively selected site, identified by the SLAC, FEL, and REL methods, respectively (Figure 1). Of these, two protein families, DNA-directed RNA polymerase subunit beta (protein family identification number PFID 315 in Supplementary Table S1) and nucleoside-triphosphatase (PFID 613), were judged as being under positive selection by all the three methods. Two protein families, DNA polymerase III subunit alpha (Gram-positive type) and ADP-ribose pyrophosphatase, were judged as being under positive selection by both the SLAC and FEL methods. A total of 39 protein families were judged as being under positive selection by both the FEL and REL methods. These 39 protein families included a number of putative virulence factors deposited in VFDB (trans-acting positive regulator, *atxA*, from *Bacillus anthracis* Sterne)

and MvirDB (argininosuccinate lyase, molecular chaperone DnaK, selenocysteine lyase, ascorbate-specific PTS system enzyme IIC from *Streptococcus agalactiae*, 4-alpha-glucanotransferase from *Bacteroides thetaiotaomicron* VPI-5482, and carboxylase from *Streptomyces coelicolor* A3 2). The SLAC, FEL, and REL methods identified 1, 153, and 179 positively selected genes that were unique to each method. A total of 888 protein families showed no evidence of positive selection by any of the three methods. The maximum numbers of positively selected sites for any particular gene detected by the SLAC, FEL, and REL methods were 1, 5, and 30, respectively (Supplementary Table S1). The FEL method identified 5 positively selected sites within the gene encoding alanyl-tRNA synthetase (PFID 489). The REL method found 30 positively selected sites in the feoB gene encoding Fe²⁺ transport system protein B (PFID 1498), which has been reported as a virulence factor (Iron uptake; Ferrous iron uptake) in other bacteria such as *Legionella pneumophila* Philadelphia 1 and *Staphylococcus aureus* subsp. *aureus* Mu50. REL was relatively liberal (more likely to identify false positive results) in its detection of positively selected sites, SLAC was conservative, and FEL was intermediate, as reported previously (Kosakovsky Pond and Frost, 2005).

There is no tendency for protein families with more sequences to be more likely to show positive selection. The number of sequences showed no effect for SLAC (Wilcoxon rank sum test; p-value = 0.079) and FEL (Wilcoxon rank sum test; p-value = 0.193), although the median number of sequences for positively selected genes (13) was higher than that for other genes (12). The number of sequences showed an opposite effect for REL (Wilcoxon rank sum test; p-value = 0.004), positive selection being more likely to be detected in small data sets. The median number of sequences for positively selected genes (10) was smaller than that for other genes (12).

An earlier study reported that 29% of positively selected genes across species of *Streptococcus* were known or hypothetical virulence factors in streptococci (Anisimova et al., 2007). Of the 1264 protein families here tested, 406 are known virulence factors of some bacteria, and 103 are known virulence factors of *Streptococcus* species (Supplementary Table S1). Of the 5 positively selected genes identified by the SLAC method, 2 are virulence factors based on MvirDB: rpoB encoding DNA-directed RNA polymerase subunit beta in *Streptococcus agalactiae* NEM316 (PFID 315), and the mutT-like protein in *Streptomyces coelicolor* A3 2 (PFID 674). Of the 196 positively selected genes identified by the FEL method, 61 are virulence factors of some bacteria and 17 are virulence factors of *Streptococcus* species, including for example *Streptococcus pyogenes* virulence regulators (SpeB-SpeF extended regulon) (PFID 202), the gene covS encoding two-component sensor histidine kinase in *Streptococcus agalactiae* 2603V/R (PFID 1158), and M protein trans-acting positive regulator (PFID 1642). Of the 220 positively selected genes identified by the REL method, 71 are virulence factors of some bacteria and 19 are virulence factors of *Streptococcus* species, including glyceraldehyde 3-phosphate dehydrogenase (PFID 599), and the gene yesM encoding two-component sensor histidine kinase (PFID 1678). Virulence genes were not significantly over- or underrepresented in the set of positively selected genes relative to other genes (Fisher's exact test p-values were 0.63, 0.80, and 0.87 for SLAC, FEL, and REL, respectively). The results suggest that known virulence genes are subject to positive selection pressure as much as other genes.

We performed enrichment tests to examine over- or underrepresented functional categories in positively selected genes relative to other genes. In positively selected genes identified by the SLAC method, the KEGG pathway map "Purine metabolism" (odds ratio = Infinite) was overrepresented (Supplementary Table S2). In positively selected genes identified by the FEL method, the JCVI subcategory "Protein folding and stabilization" (odds ratio = 8.56) was overrepresented (Supplementary Table S3). In positively selected genes identified by

the REL method, the JCVI subcategory “Serine family” (odds ratio = 7.8) was overrepresented, while the JCVI main category “DNA metabolism” (odds ratio = 0.27) was underrepresented (Supplementary Table S4). After false discovery rate correction for multiple comparisons, no functional categories were significantly over- or underrepresented in positively selected genes relative to other genes ($P < 0.05$).

An earlier study analyzed a database of gene families and protein domains from a large and diverse set of species, and concluded that genes involved in complex functions such as transcription and translation are significantly less likely to undergo positive selection (Aris-Brosou, 2005). In the genus *Campylobacter*, no GO term was significantly overrepresented in positively selected genes, and ribosome-related genes were never found to be under positive selection (Lefebure and Stanhope, 2009). However, this is not the case in the genus *Streptococcus* tested in the present study. Of the 37 ribosomal proteins, 0, 6, and 8 were identified as being under positive selection by SLAC, FEL, and REL, respectively. The corresponding odds ratio values for the COG category “Translation, ribosomal structure and biogenesis” were 0, 1.47, and 0.86. Note that SLAC found only five positively selected genes and it is this poor sampling that results in the misleading odds ratio of 0. There are however, potential artifacts involving translational genes related to the confounding effects of codon bias and the detection of positive selection (Aris-Brosou, 2005). It has been shown that highly expressed genes such as ribosomal proteins in fast-growing bacteria exhibit strong codon usage biases, putatively reflecting the action of translational selection to optimize the speed and accuracy of translation (Ikemura, 1985; Eyre-Walker, 1996).

Positive selection can be falsely detected due to the presence of recombination (Anisimova et al., 2003; Shriner et al., 2003). Of the 5, 196, and 220 protein families identified as being under positive selection by the SLAC, FEL, and REL methods, 4, 106, and 106 were also judged as being recombinant based on the single breakpoint (SBP) analysis (Supplementary Table S1). Excluding these recombinant genes from the enrichment tests (data not shown) did not change our main conclusion, i.e., that certain functional categories or virulence factors were not over- or underrepresented in positively selected genes relative to other genes.

4. Conclusions

Our results suggest that known virulence genes from *Streptococcus* species are subject to positive selection pressure as much as other genes. Previous studies of positive selection in various pathogenic bacteria have revealed many virulence factors that can be under positive selection pressure (Anisimova et al., 2007; Lefebure and Stanhope, 2007; Suzuki et al., 2011). In such cases, researchers inclined to attribute the positive selection to the virulence potential of the locus. However, our findings suggest that virulence factors, at least in *Streptococcus*, may not be correlated with a history of positive selection. We cannot rule out the possibility that current virulence factor databases such as VFDB and MvirDB (Chen et al., 2005; Zhou et al., 2007; Yang et al., 2008) are far from comprehensive and thus a set of known virulence factors were underestimated. After false discovery rate correction for multiple comparisons, no functional categories were significantly over- or underrepresented in positively selected genes relative to other genes in *Streptococcus* species. A previous study analyzed a database of gene families and protein domains from a large and diverse set of species, which may mask the uniqueness of particular species groups, and concluded that gene function is strongly correlated with its tendency to undergo adaptive evolution (Aris-Brosou, 2005). Our findings suggest that a correlation of gene function and selection may not be applicable to *Streptococcus* and by extension possibly other specific genera.

Highlights

We test functional bias of positively selected genes in *Streptococcus* genomes.
 Known virulence genes were subject to positive selection as much as other genes.
 No functional categories were enriched in positively selected genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Tristan Lefébure for helpful discussion, Sergei L Kosakovsky Pond for his advice on HyPhy, and Robert Bukowski, Andrew Dolgert and Linda Woodard for their help with the parallelization of the analyses on a Linux cluster at the Computational Biology Service Unit (CBSU) and Center for Advanced Computing (CAC) of Cornell University. 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 M.J.S.

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25:3389–3402. [PubMed: 9254694]
- Anisimova M, Bielawski J, Dunn K, Yang Z. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evol Biol.* 2007; 7:154. [PubMed: 17760998]
- Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics.* 2003; 164:1229–1236. [PubMed: 12871927]
- Arakawa K, Mori K, Ikeda K, Matsuzaki T, Kobayashi Y, Tomita M. G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics.* 2003; 19:305–306. [PubMed: 12538262]
- Aris-Brosou S. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol Biol Evol.* 2005; 22:200–209. [PubMed: 15483330]
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25:25–29. [PubMed: 10802651]
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2011; 39:D32–D37. [PubMed: 21071399]
- Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. Q. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* 2005; 33:D325–D328. [PubMed: 15608208]
- Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P, Armstrong JR, Fulton RS, Latreille JP, Spieth J, Hooton TM, Mardis ER, Hultgren SJ, Gordon JI. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc Natl Acad Sci U S A.* 2006; 103:5977–5982. [PubMed: 16585510]
- Davidson T, Beck E, Ganapathy A, Montgomery R, Zafar N, Yang Q, Madupu R, Goetz P, Galinsky K, White O, Sutton G. The comprehensive microbial resource. *Nucleic Acids Res.* 2010; 38:D340–D345. [PubMed: 19892825]
- Davies MR, McMillan DJ, Beiko RG, Barroso V, Geffers R, Sriprakash KS, Chhatwal GS. 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.* 2007; 44:1442–1454. [PubMed: 17479940]

- Eyre-Walker A. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol Biol Evol.* 1996; 13:864–872. [PubMed: 8754221]
- Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. The Pfam protein families database. *Nucleic Acids Res.* 2010; 38:D211–D222. [PubMed: 19920124]
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010; 59:307–321. [PubMed: 20525638]
- Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 2003; 52:696–704. [PubMed: 14530136]
- Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 1985; 2:13–34. [PubMed: 3916708]
- Kalia A, Bessen DE. Natural selection and evolution of streptococcal virulence genes involved in tissue-specific adaptations. *J Bacteriol.* 2004; 186:110–121. [PubMed: 14679231]
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000; 28:27–30. [PubMed: 10592173]
- Kosakovsky P, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 2005; 22:1208–1222. [PubMed: 15703242]
- Kosakovsky P, Posada D, Gravenor MB, Woelk CH, Frost SD. GARD: a genetic algorithm for recombination detection. *Bioinformatics.* 2006; 22:3096–3098. [PubMed: 17110367]
- Lefebvre T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* 2007; 8:R71. [PubMed: 17475002]
- Lefebvre T, Stanhope MJ. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*. *Genome Res.* 2009; 19:1224–1232. [PubMed: 19304960]
- Lysenko ES, Lijek RS, Brown SP, Weiser JN. Within-host competition drives selection for the capsule virulence determinant of *Streptococcus pneumoniae*. *Curr Biol.* 2010; 20:1222–1226. [PubMed: 20619820]
- Marri PR, Hao W, Golding GB. Gene gain and gene loss in *Streptococcus*: is it driven by habitat? *Mol Biol Evol.* 2006; 23:2379–2391. [PubMed: 16966682]
- Orsi RH, Sun Q, Wiedmann M. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. *BMC Evol Biol.* 2008; 8:233. [PubMed: 18700032]
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Newweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005; 33:5691–5702. [PubMed: 16214803]
- Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics.* 2005; 21:676–679. [PubMed: 15509596]
- R_Development_Core_Team R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.
- Roshan U, Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics.* 2006; 22:2715–2721. [PubMed: 16954142]
- Shriner D, Nickle DC, Jensen MA, Mullins JI. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res.* 2003; 81:115–121. [PubMed: 12872913]
- Soyer Y, Orsi RH, Rodriguez-Rivera LD, Sun Q, Wiedmann M. Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected *Salmonella* serotypes. *BMC Evol Biol.* 2009; 9:264. [PubMed: 19912661]
- Springman AC, Lacher DW, Wu G, Milton N, Whittam TS, Davies HD, Manning SD. Selection, recombination, and virulence gene diversity among group B streptococcal genotypes. *J Bacteriol.* 2009; 191:5419–5427. [PubMed: 19581371]

- Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 2002; 12:1611–1618. [PubMed: 12368254]
- Suzuki H, Lefebure T, Hubisz MJ, Pavinski Bitar P, Lang P, Siepel A, Stanhope MJ. Comparative genomic analysis of the *Streptococcus dysgalactiae* species group: gene content, molecular adaptation, and promoter evolution. *Genome Biol Evol.* 2011; 3:168–185. [PubMed: 21282711]
- Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 2000; 28:33–36. [PubMed: 10592175]
- Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 2001; 29:22–28. [PubMed: 11125040]
- van Dongen, S. Graph Clustering by Flow Simulation. University of Utrecht; 2000.
- Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet.* 2002; 32:569–577. [PubMed: 12457190]
- Yang J, Chen L, Sun L, Yu J, Jin Q. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res.* 2008; 36:D539–D542. [PubMed: 17984080]
- Zadoks RN, Schukken YH, Wiedmann M. Multilocus sequence typing of *Streptococcus uberis* provides sensitive and epidemiologically relevant subtype information and reveals positive selection in the virulence gene *pauA*. *J Clin Microbiol.* 2005; 43:2407–2417. [PubMed: 15872274]
- Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* 2007; 35:D391–D394. [PubMed: 17090593]

Figure 1.

A venn diagram showing the number of protein families judged as being under positive selection by the three methods (SLAC, FEL, and REL).

Table 1

The 13 *Streptococcus* strains analyzed in this study.

NCBI Accession Number	Organism names
NC_004116	<i>Streptococcus agalactiae</i> 2603V/R
NC_012891	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> GGS_124
NC_012471	<i>Streptococcus equi</i> subsp. <i>equi</i> 4047
NC_013798	<i>Streptococcus gallolyticus</i> UCN34
NC_009785	<i>Streptococcus gordonii</i> str. Challis substr. CH1
NC_013853	<i>Streptococcus mitis</i> B6
NC_004350	<i>Streptococcus mutans</i> UA159
NC_003028	<i>Streptococcus pneumoniae</i> TIGR4
NC_002737	<i>Streptococcus pyogenes</i> M1 GAS
NC_009009	<i>Streptococcus sanguinis</i> SK36
NC_009443	<i>Streptococcus suis</i> 98HAH33
NC_006449	<i>Streptococcus thermophilus</i> CNRZ1066
NC_012004	<i>Streptococcus uberis</i> 0140J