White Paper Application

Project Title: Evolutionary genomics and population genetics of pathogenic streptococci

White Paper Submission Date (MM/DD/YY): 11/02/09

Investigator Contact:

Name	Michael J. Stanhope
Position	Professor of Evolutionary Genomics
Institution	Cornell University
Address	Department of Population Medicine and Diagnostic Sciences
State	NY
ZIP Code	14853
Telephone	607 253 3859
Fax	607 253 3440
E-Mail	mjs297@cornell.edu

All white papers will be evaluated based on the following sections.

1. Executive Summary *(Please limit to 500 words.)*

Provide an executive summary of the proposal.

The bacteria genus *Streptococcus* includes some of the most important human pathogens, causing a wide range of different disease, and inflicting significant morbidity and mortality throughout the world. Our long term goal is to reach a thorough understanding of the molecular specifics correlated with adaptive differences within and between the pathogenic taxa of the genus *Streptococcus*. Towards this goal we will implement modern phylogenetic and population genetic principles and procedures in the analysis of comparative genome sequence data gathered from multiple strains of the same pathogenic species and from representative isolates of different species. We are particularly interested in the role of positive selection in the diversification of *Streptococcus* species and the different components of their genomes. Positive selection is the fixation of advantageous mutations driven by natural selection, and is the fundamental process behind adaptive changes in genes and genomes, leading to evolutionary innovations and species differences. Our central hypothesis is that many loci, sites within loci, and noncoding functional elements, identified as being under lineage specific, or positive selection pressure, in our evolutionary analyses, will be key loci involved in the colonization, persistence, survival, and propensity to cause disease in these pathogenic streptococci. More specifically, the aims of this proposal are to: (1) Generate and annotate genome sequence data for multiple species of *Streptococcus* within the Pyogenes, and Mutans groups, chosen such that they will yield various sister group comparisons involving important human pathogens; (2) Assess the role of positive selection and lateral gene transfer in the diversification of the genomes of the species within the three key pathogenic group of streptococci: Pyogenes, Mutans and Mitis; (3) Collect comparative genome sequence data across multiple strains of the important human pathogens *Streptococcus agalactiae*, and *Streptococcus pneumoniae* for the purpose of analyzing selection pressure and demographic history, employing new population genetic based methods developed by us and others; (4) Implement several new methods developed by us in a comprehensive survey of noncoding functional elements in the *Streptococcus* Pyogenes, Mutans and Mitis groups; (5) Use the results

arising from the selection analyses, and identification of noncoding functional elements, to create gene knockout and site specific mutants for the human oral pathogen *Streptococcus mutans*, and assay the phenotypic effects.

The fundamental data on comparative genomics and molecular adaptation, arising from this project, will serve as a framework for the development of novel therapeutic and preventative strategies for pathogenic species of *Streptococcus*. The results from these evolutionary analyses will serve as a guide for future follow-up, cause effect experimentation.

2. Justification

Provide a succinct justification for the sequencing or genotyping study by describing the significance of the problem and providing other relevant background information.

2.0 Introduction

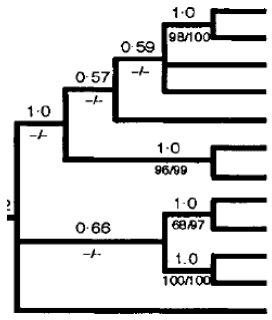
Streptococci are a diverse group of bacteria, comprising over 60 species, and include taxa which are members of human and animal commensal microflora, as well a variety of important human, zoonotic, and agricultural pathogens. Recent molecular phylogenies support several taxonomic divisions within the genus, including Pyogenes, Bovis, Mutans, Salivarius, Anginosus, and Mitis (e.g. Tapp et al. 2003). The major human pathogens occur in the Pyogenes, Mitis and Mutans groups.

Comparative sequence data on a genomic scale provides the opportunity to explain major biological differences between organisms at the molecular level. Arguably the three most significant molecular characteristics responsible for biological differences between organisms are: (1) presence and absence of particular loci, (2) molecular selection differences and (3) gene regulation. It is our purpose here to examine all three of these, in an attempt to explain adaptive differences between *Streptococcus* taxa and putative ecotypes of the Pyogenes, Mitis and Mutans groups.

2.1 The Pyogenes Group

The Pyogenes group includes a number of human, animal and zoonotic pathogens, and is named after the important human pathogen, *Streptococcus pyogenes* (Group A *Streptococcus*; GAS), which is responsible for a wide range of human diseases, including pharyngitis, impetigo, puerperal sepsis, necrotizing fasciitis ("flesh-eating disease"), scarlet fever, the postinfection sequelae glomerulonephritis and rheumatic fever (Cunningham 2000; Musser and Krause 1998). *Streptococcus agalactiae* (Group B *Streptococcus*, GBS) is the leading cause of bacterial sepsis, pneumonia, and meningitis, in U.S and European neonates (Schuchat and Wenger 1994). Although *S. agalactiae* normally behaves as a commensal organism that colonizes the genital or gastrointestinal tract of healthy adults, it can cause life threatening invasive infection in susceptible hosts, such as newborns, pregnant women, and nonpregnant adults with chronic illnesses (Schuchat and Wenger 1994). The newly described species *S. pseudoporcinus*, previously classified as *S. porcinus*, appears to be associated with the genitourinary tract of women (Bekal et al. 2006; Duarte et al. et al. 2005) and as a cause of pre-term stillbirth (Martin et al. 2004). Importantly, *S. pseudoporcinus* occupies the same niche as does the human pathogen *S. agalactiae* and has serological cross-reactivity with the group B streptococcal reagent used to identify *S. agalactiae*. Thus, there is a distinct possibility that human infections and/or colonization by S. *pseudoporcinus* may be greatly underestimated. *Streptococcus dysgalactiae* subsp. *equisimilis*, (human group C or G *Streptococcus*, GCS/GGS), is generally regarded as a human commensal organism (Rolston 1986) but can cause a spectrum of human diseases very similar to that caused by *S. pyogenes* Fujita etal. 1982; Gaunt and Seal 1987; Stryker et al. 1982; Zaoutis et al. 2004). Zoonotic pathogens in the Pyogenes group include *S. iniae*, which can cause cellulitis, discitis and endocarditis (Koh et al. 2004; Weinstein et al. 1997); *S. equi* subsp. *zooepidemicus*, which can cause meningitis (Latorre et al. 1993), pneumonia (Rose et al. 1980), septic arthritis (Collazos et al. 1992), and endocarditis (Martinez-Luengas et al. 1982); *S. equi* subsp. *equi*, a rare cause of childhood meningitis (Ferretti et al. 2001); and *S. canis*,

which is found in polymicrobial skin and soft tissue infections (Galperine et al. 2007) and, rarely, as monomicrobial cause of bacteremia (Bert and Lambert-Zechovsky 1997; Takeda et al. 2001; Whatmore et al. 2001) or urinary tract infections (Galperine et al. 2007).



	Customary niche	Genome sequence
S. dysgalactiae subsp. *dysgalactiae*	Bovine	MJS – final gap closure
S. dysgalactiae subsp. *equisimilis*	Human	MJS – final gap closure + 1
S. agalactiae	Human, Bovine	Available (n=8)
S. phocae	Seal	This proposal
S. porcinus	Porcine	This proposal
S. iniae	Fish, Human	In progress
S. parauberis	Fish	This proposal
S. canis	Animal	MJS – final gap closure
S. pyogenes	Human	Available (n=14)
S. equi subsp. *equi*	Equine	Available
S. equi subsp. *zooepidemicus*	Equine, Human	Available (n=2)
S. uberis	Bovine	Available

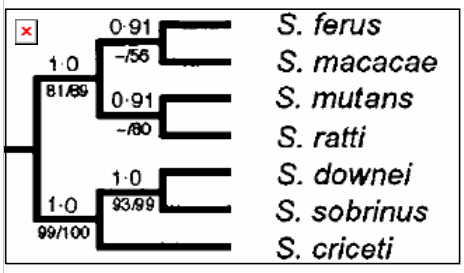
Fig. 1. The Pyogenes group of Streptococci (from (Tapp et al. 2003), with niche adaptation and availability of genome sequence data. *Streptococcus pseudoporcinus* and *S. urinalis* are two additional species, which we propose to sequence, not included in this figure.

Genome sequence data are available, nearly complete, or in process for a number of taxa within the Pyogenes group (Fig. 1). Available data include 8 genome sequences for *S. agalactiae* (e.g. Tettelin et al. 2005), 14 for *S. pyogenes* (e.g. Beres et al. 2004; Nakagawa et al. 2003; Smoot et al. 2002), and 85 draft sequences of *S. pyogenes* in the form of Solexa reads on the NCBI sequence read archive database. Genome sequences have recently been completed for *S. uberis* (Ward et al. 2009), *S. equi* subsp. *equi*, *S. equi* subsp. *zooepidemicus* (Holden et al. 2009) and *S. dysgalactiae* subsp *equisimilis*. *S. iniae* is in final assembly (CRIS accession number 0205306). We are in final gap closure for *S. dysgalactiae* subsp *equisimilis*, *S. dysgalactiae* subsp *dysgalactiae*, and S. canis. In addition to the species indicated in Fig. 1, we propose to also include *S. urinalis* and *S. pseudoporcinus*; both are typical of humans – *S. urinalis* generally falls in as the outgroup to the Pyogenic group and pseudocporcinus is the sister group to *porcinus*.

2.2 The Mutans Group

Current phylogenetic classifications include the taxa depicted in Fig. 2, for the Mutans group of streptococci, with the recent addition of *Streptococcus dentirousetti, Streptococcus devriesei*, and *Streptococcus orisuis*. *S. mutans* is implicated as the principal causative agent of human dental caries (tooth decay) (Loesche 1986). Dental caries is one of the most common infectious diseases afflicting humans, and tends to be left untreated in many underdeveloped areas, leading to considerable human suffering. The economic burden associated with treating dental infections, reaches staggering amounts, estimated, in 1984 at about 24 billion dollars per year in the USA alone (54). Although many species of bacteria have been associated with dental plaque, *S. mutans* is the key species consistently linked with the formation of human dental caries (Loesche 1986). *S. mutans* is also an occasional cause of non-oral infections, principally subacute bacterial endocarditis (Ajdic et al. 2002). The main virulence factors associated with *S. mutans* are adhesion, acidogenicity, acid tolerance, potent biofilm forming abilities, and enhanced systems for the assimilation of diverse carbohydrate sources (Burne 1998; Napimoga et al. 2005). *S. sobrinus* is the next most prevalently linked species to human dental caries (Loesche 1986). Although *S. mutans* has been shown to be more prevalent in dental plaque samples, *S. sobrinus* may be more closely associated with high caries activity (Fujiwara et al. 1991; Hirose et al. 1993). Okada et al. (2005) presented evidence that children harboring both these species have a higher incidence of dental caries than with *S. mutans* alone. The oral cavity of a variety of animal species, as well as occasionally

humans, is the niche for the remaining *Streptococcus* species of this group (Fig. 2).



	Customary niche	Genome sequence
S. ferus	Rodent and pig	N/A
S. macacae	Macaque	This proposal
S. mutans	Human	Available
S. ratti	Human and rodent	MJS – final gap
S. downei	Human and macaque	This proposal
S. sobrinus	Human	In progress
S. criceti	Human and rodent	This proposal

Fig. 2. The Mutans group of Streptococci (from Tapp et al. 2003), with host adaptation and availability of genome sequence data.

2.3 The Mitis Group

The most important human pathogen in the Mitis group is *S. pneumoniae*, which is the leading cause of community acquired pneumonia, as well as a major cause of meningitis, otitis media, and sepsis. The putative sister group to *S. pneumoniae* is *S. mitis*, a human commensal species found in nose, throat and mouth, and implicated as a donor in LGT events of antibiotic resistance loci involving *S. pneumoniae* (e.g. Dowson et al. 1993; Reichmann et al. 1997). *S. sanguinis* is relatively closely related to *mitis* and *pneumoniae*, found in the human oral cavity and is implicated in the development of caries and periodontal disease, as well as the taxon most commonly associated with native-valve endocarditis (Mylonakis and Calderwood 2001). Including both complete and draft sequences, a total of 11 genome sequences are available for *S. pneumoniae* (Hiller et al. 2007). There is a recent complete sequence for *S. sanguinis* (Xu et al. 2007) and there are 2 strains of *S. mitis* in progress. A sequence for *S. gordonii*, classified within the closely related Anginosus group, has also recently been released (Vickerman et al. 2007). Although we do not plan to obtain genome sequence data of any additional taxa from this group, we will conduct comparative genomic analyses of this accumulating body of data, and we do propose to take advantage of our extensive collection of *S. pneumoniae* isolates (see e.g. Stanhope et al. 2005; Stanhope et al. 2007) to conduct population genomic analyses of this species (described below in section 2.5).

2.4 Phylogenetic Assessment of Molecular Adaptation

The species from the Pyogenic group that we propose to sequence include the following: *S. phocae, S. porcinus, S. parauberis, S. urinalis*, and *S. pseudoporcinus*. Concomitant with already existing data, the resulting set of comparative genomic sequences would be unprecedented in terms of their representation of overall phylogenetic diversity, for an interspecific taxonomic group of bacteria (Fig. 1). These taxa were also chosen in order to provide multiple sister group comparisons involving a range of host groups and disease, with the human host always the key component. Sister group comparisons are particularly valuable because they afford the possibility of more directly evaluating the changes on the branch leading to the development of a specific human pathogen. For example, current phylogenies support a sister group relationship between *S. canis* and *S. pyogenes* (e.g. Facklam 2002; Tapp et al. 2003). Despite high prevalence of *S. canis* in pets and frequent and intimate contact between pets and humans *S. canis* infections in humans are extremely rare (Bert and Lambert-Zechovsky 1997; Takeda et al. 2001; Wahtmore et al. 2001). What makes *S. pyogenes* the most virulent pathogen in the group, while its sister species is rare in humans? The species *S. pseudoporcinus*, appears to share important overlap in habitat with *S. agalactiae*. How does the genomic composition of this newly described taxon compare, not only to *S. porcinus*, its animal host sister species, but also to *S. agalactiae*, and what genes have a LGT history

involving S. *agalactiae* and *S. pseudoporcinus*? In addition to providing a thorough representation of diversity for the Pyogenes group, our choice of these 5 taxa has comparative benefits that are specifically associated with their similarities and differences in pathogenesis and host or niche adaptation. Comparisons can be made between divergent species of the same host, sister taxa with the same or distinct host, and subspecies with distinct hosts or disease manifestations, with the human host an integral part of all these comparisons (Fig. 1). The phylogenetic density of this dataset and the sister group comparisons, will provide an ability to accurately assess genes gained and lost on each of the branches of the phylogeny and thus reconstruct the evolutionary acquisition of genes associated with particular pathogenic properties.

Analysis of genomic gene content differences, and more specifically genes gained and lost, will provide insight into adaptive differences, however, it is also essential to understand the role of natural selection in explaining the observed sequence differences at shared protein coding loci, as well as the possible role of noncoding functional elements. Positive selection is the fixation of advantageous mutations driven by natural selection, and is the fundamental process behind adaptive changes in genes and genomes, leading to evolutionary innovations and species differences. Analysis of the genome sequence data captured across strains, for signatures of positive selection, would proceed using both phylogenetic and population genetic based approaches, because both have their specific advantages. Ziheng Yang, Rasmus Nielsen and colleagues (Zang et al. 1997; Nielsen and Yang 1998; Yang et al. 1998; Yang et al. 2000; Yang et al. 2005; Zang et al. 2005) have developed powerful phylogenetic methods to detect molecular selection in protein coding genes. Their methods compare synonymous and nonsynonymous substitution rates in protein coding genes and regard a nonsynonymous rate elevated above the synonymous rate as evidence for positive or Darwinian selection. A significant advancement of many earlier methods, which averaged over sites and time, their methods are designed to detect positive selection at individual sites and lineages. The branch sites test would be performed on all branches, for all orthologous loci of the genomes of the species within this group, employing our bioinformatics pipeline specifically constructed for this purpose (Lefebure and Stanhope 2007; Lefebure and Stanhope 2009).

The species from the Mutans group that we propose to sequence include the following: *S. macacae, S. downei*, and *S. criceti*. In the case of this Mutans group our choice of taxa for genome sequencing centers on sister group comparisons of the two important human pathogens *S. mutans* and *S. sobrinus*. The taxa indicated in Fig. 2 will provide the ability to evaluate which genes are under positive selection and which were gained and lost in the diversification of *S. mutans* from its common ancestry with *S. ratti*, and similarly for *S. sobrinus* after its separation from common ancestry with *S. downei*.

2.5 Population Genomic Analysis of Molecular Adaptation

Our population genetic assessment of molecular adaptation and the population genomic sequencing proposed herein, involves the choice of two taxa: *S. agalactiae*, and *S. penumoniae*. We have recently acquired the necessary data to conduct *S. mutans* population genomic analyses and Illumina sequence reads are available on NCBI (Sequence Read Archive: SRP000775) for a population genomic dataset of 85 *S. pyogenes* isolates. Our choice of S. *agalactiae* and *S. pneumoniae*, will therefore, provide a set of four important human pathogens, within the same genus, of very different niches and pathogenic properties to compare across population genomic datasets. In addition to the important pathogenic characteristics of these species, with regard to human disease, S. *agalactiae* also infects bovine, and there is evidence for a bovine ecotype (Bisharat et al. 2004; Sukhnanand et al. 2005). Thus, this taxon provides a within species molecular adaptation comparison to be made across human and bovine, which could prove particularly informative with regard to identifying key human pathogenic loci. Powerful population genetic methods are available for inferring selection (e.g. Tajima 1989; McDonald and Kreitman 1991; Bustamante et al. 2002) and have a number of advantages in comparison to the phylogenetic based approaches, including: (1) requiring a lower threshold for amino acid divergence relative to synonymous divergence in order to detect positive selection, and thus greater power, (2) ability to detect recent selection that may not have left a phylogenetic signature (such as lineage-specific

fixation of a single novel amino acid mutation), and (3) ability to model recombination in the data which is a known source of false positive signal in phylogenetics (Anisimova et al. 2003).

One of the important recent discoveries in comparative bacterial genomics is that a bacterial species' genome is comprised of a core genome, including genes common to all isolates of that species, and a dispensable genome, consisting of genes not present in all the isolates, which together constitute the pan-genome for that species (Tettelin et al. 2005; Lefebure and Stanhope 2007). Recent work of ours involving two species of gastrointestinal pathogens, with genomes only slightly smaller than typical *Streptococcus* genomes suggests that with high coverage Solexa genome sequence data it is possible to de novo assemble both dispensable and core components of the genome, and with many multiple isolates eventually reach an estimate of the pan-genome of the species (Fig. 3).

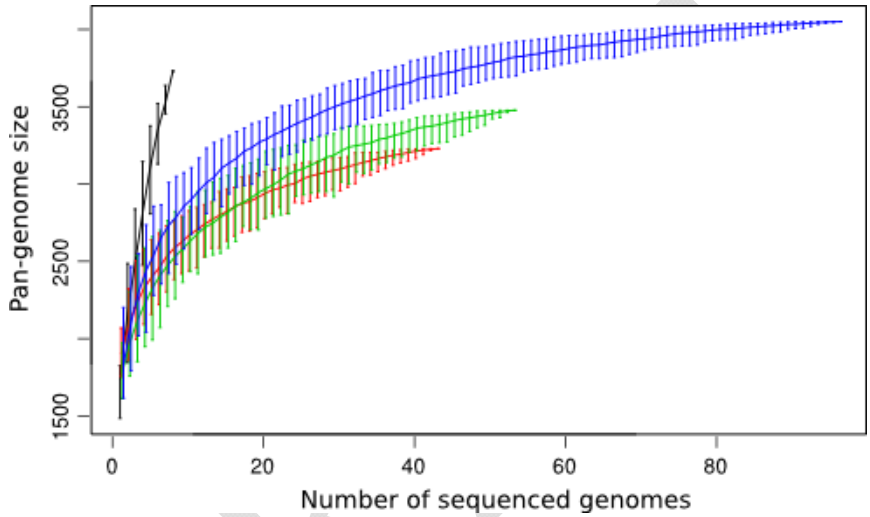


Fig. 3. Gene accumulation curves estimating pan genome size for *Campylobacter*. The vertical bars correspond to standard deviations after repeating one hundred random input orders of the genomes. Red= *C. coli*; Green = *C. jejuni*; blue = *C.jejuni+C. coli*; Black = *Campylobacter* genus

We propose generating similar such data and analyses for S. *agalactiae* and S. *pneumoniae* and feel such information is significant for a number of reasons: (1) Many species of bacteria, including *S. agalactiae*, appear to evolve host preferred groups, and since many of the genes associated with specific features of pathogenesis are contained within the dispensable portion of the genome, knowledge of the complete repertoire of loci will greatly facilitate the identification of genes correlated with these different groups; (2) Knowledge of the dispensable genome components and their distribution across different niches is key to understanding the pathobiology of the species; (3) With the knowledge of the complete repertoire of genes for *S. agalactiae* and *S. pneumoniae* also comes the ability to perform large scale surveys of gene presence and absence, using microarrays, across isolates of multiple hosts and niches from around the world, that would, for the first time (for any species of bacteria), include the complete pan-genome in the comparisons; (4) A growing body of comparative information on bacteria pan-genomes would ultimately lead to an understanding of the role of different niches in the evolution of dispensable genomes in pathogenic bacteria.

2.6 Analysis of Noncoding Functional Elements

Comparative studies of bacterial genomes have so far tended to focus on differences in gene content

(e.g., Makarova et al. 2006; McClelland et al. 2000; Tettelin et al. 2002) and in the coding sequences of orthologous genes (Charlesworth and Eyre-Walker 2006). The Pyogenes and Mutans groups—with their diverse species at modest evolutionary distances—provide an ideal opportunity to look also at differences between species in noncoding functional elements such as *cis*-regulatory elements, structural RNA genes, and riboswitches. Recent findings in eukaryotes suggest that noncoding sequences may be far more important than previously thought in determining differences between species (Andolfatto 2005; Pollard et al. 2006; Ponting and Lunter 2006; Prabhakar et al. 2006). Despite the greater importance of LGT in bacteria, their simpler mechanisms for regulation, and their greatly reduced noncoding DNA content, bacterial species and strains, too, are undoubtedly defined not only by their repertoire of protein-coding genes but also their regulatory machinery and structural RNAs.

We propose to conduct a comprehensive survey of noncoding functional elements in the Pyogenes, Mutans, and Mitis groups, focusing on genomic differences likely to help explain differences in phenotype. Our new sequence data, together with existing sequences from each of these groups, will provide an unprecedented opportunity to study the evolution and possible function of bacterial noncoding DNA. Recent work by Halpern et al. (2007) suggests there are conserved noncoding motifs in different groups of bacteria, including some streptococci, that could be involved in genome stability. There has also been work on computational identification of noncoding RNAs in *E. coli* (Argaman et al. 2001; Rivas et al. 2001; Wassarman et al. 2001), and on the use of phylogenetic footprinting to identify regulatory elements in proteobacteria (McCue et al. 2001; McGuire et al. 2000; Rajewsky et al. 2002), however, most currently available bacterial species are too distant to permit a detailed analysis of the evolution of noncoding sequences. In contrast, within the Pyogenes group, many intergenic regions can be aligned and analyzed in detail (Fig. 4).

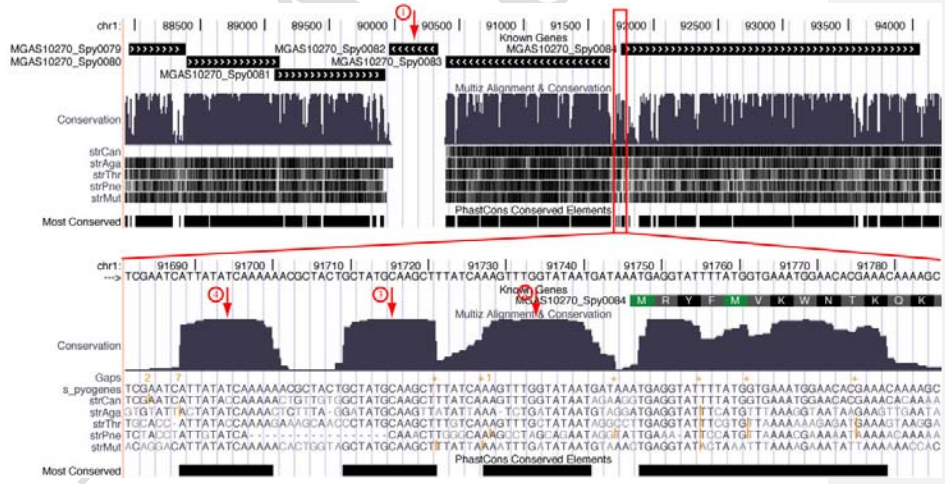


Fig. 4. Prototype *Streptococcus* Genome Browser, showing known genes, cross-species alignments, and phastCons scores. One of the six genes shown is specific to *S. pyogenes* (arrow 1). At bottom a detailed view of the promoter region of the *pbp1b* gene shows conserved -10 (arrow 2) and -35 (arrow 3) elements, and a possible conserved upstream element (arrow 4). This browser is available at http://genome-mirror.bscb.cornell.edu/cgi-bin/hgTracks?db=strPyo1.

We will adapt and extend computational methods we have pioneered and applied extensively in eukaryotes (Pedersen et al. 2006; Siepel et al. 2005; Siepel and Haussler 2004a; Siepel and Haussler 2004b; Siepel et al. 2006) to identify three types of candidate noncoding functional elements in

Streptococcus genomes: conserved regulatory elements, conserved RNA structures, and conserved elements of indeterminate function. We will identify elements showing significant evidence of lineage-specific selection, in addition to those conserved across species—e.g., regulatory elements conserved in most species but apparently under relaxed constraint in, say, *S. pyogenes*, or RNA structures specific to, say, the sister species *S. agalactiae* and *S. dysgalactiae*. In addition, we will identify noncoding elements that are evolving significantly faster than synonymous sites in flanking genes, and therefore are likely to be under (possibly lineage-specific) positive selection.

We expect that this work will allow for a comprehensive annotation of the noncoding DNA of Pyogenes, Mutans, and Mitis group species, and will identify important noncoding differences between species. These genomic differences may help to explain the different biological niches of the organisms, and could potentially suggest new strategies for vaccines or therapies.

3a. **Rationale for Strain Selection**

Phylogenetic Dataset for Pyogenes and Mutans Groups

Our phylogenetic based study of adaptation involves both an evaluation of genes gained and lost in the diversification of human pathogens and an assessment of positive selection of shared loci along the branch leading to the human pathogen. Both these approaches to studying adaptation center around a very basic tenant of comparative evolutionary biology which suggests sister group comparisons provide the ability to evaluate which characteristics evolved on each lineage after separation from their common ancestry. In our case, because the sister group comparisons involve a human pathogen and an animal pathogen, this should provide significant answers to the question of what makes the human pathogen, a human pathogen. Thus, we propose sequencing several species of *Streptococcus* which are not human pathogens, but which provide the necessary comparative framework to address key components in the evolution of the human pathogen. In order to evaluate this, at a minimum, one requires genome sequence from the human pathogen, its putative sister group and a closely related outgroup. This minimum requirement is what we have planned in proposing the species for sequencing from the Mutans group: *S. macacae, S. downei*, and *S. criceti*. In the case of the Pyogenes group, there is greater known phylogenetic diversity, greater diversity in pathogenic features, and greater host diversity. There are also more genome sequences already available for a number of these species. Therefore, in this case we plan a choice of isolates that both maximizes sister group comparisons involving the human pathogens, while also taking advantage of existing data, to provide an unparalleled dataset of phylogenetic density, for any group of bacteria, that will afford an accurate assessment of genes gained and lost and under positive selection in the diversification of this entire group. For the Pyogenes group, in order to accomplish this, we propose sequencing the following taxa: *S. phocae, S. porcinus, S. parauberis, S. urinalis*, and *S. pseudoporcinus*.

Population Genomic Dataset for *S. agalactiae* and *S. penumoniae*

In the case of our population genomic analyses we propose sequencing multiple isolates of S. *agalactiae* and *S. penumoniae*. Both are key human pathogens, with different pathogenic features and *S. agalactiae* has the added comparative benefit of being a bovine pathogen, with a putative bovine ecotype. For *S. agalactiae* there are currently 8 genome sequences available, including both complete and draft sequences, with no documented indication on the Gold Genome database or NCBI of other strains in progress. We propose sequencing 60 strains from human sources and 30 from bovine sources. In the case of the human strains they come from a wide diversity of human sources of infection, geographic locations, serotype, and sequence types. In the case of the bovine strains they involve a diversity of locations and sequence types taken from within the putative bovine ecotype. This total of 90 isolates

will provide sufficient numbers to conduct the population genetic analysis on both the human and bovine sets, and should also provide sufficient numbers to assemble the entire pan-genome of the species. Recent experience of ours involving two species with only slightly smaller genome sizes, suggests we can assemble the entire pan-genome of an organism with a 1.8 mb genome from approximately 50-60 isolates. On average, S. *agalactiae* is about 300-400kb larger than this, and we estimate that a total of 80-100 isolates should be sufficient to estimate the entire pan-genome of the species. This will also provide a dataset, in terms of isolate numbers, comparable in size to that of *S. pyogenes* (Sequence Read Archive: SRP000775), as well as our dataset of *S. mutans*, and therefore provide the possibility of a comparative analysis between these species based on a balanced dataset.

In the case of *S. penumoniae* we propose including 60 isolates for sequencing and these arise from a diverse collection of MJS, originating from the Alexander Network collection, established by GSK. All 60 of these isolates were part of a larger study of 216 isolates published recently (Stanhope et al. 2007; Stanhope et al. 2008). These 216 isolates comprise a diverse international representation, collected over an eight year period, all of them have been genotyped, the vast majority also been serotyped, and their susceptibility to a range of at least 10 different antibiotics has been determined. We have chosen 60 isolates for sequencing that maximize all these features of diversity. These 60 genome sequences would be added to the set of 11 sequences already published. The principal reason we propose fewer strains for *S. pneumoniae*, than for *agalactiae*, is that there are more sequences for this species currently available and more listed as in progress (as many as an additional 50). Although it is impossible to predict how many of these in progress sequences would be completed to coincide with the analysis of our dataset, it is likely that some of them would be available and thus ultimately we envision a dataset of similar size in terms of numbers to *S. agalactiae, S. pyogenes*, and *S. mutans*.

3b. Strain Information:

Included as attachments to this proposal are lists of strains with the requested information, for 454 genome sequencing of new species, to accompany the phylogenetic based analyses, and lists of strains for Solexa sequencing and the associated population genomic analyses.

3c. Nature, Availability & Source of Reagents/Samples:

All isolates proposed for sequencing have been retrospectively collected, are in our possession, the vast majority have DNA extracted, and could be shipped to the GSC at anytime.

3d. Proposed Methods and Protocols to Prepare Reagents:

The vast majority of strains proposed for sequencing already have DNA extracted. We have had good success with other facilities performing the same sort of sequencing as what we propose here for other similar species of bacteria (*S. mutans* for example) and therefore we expect that there should not be any quality control issues. If there are, then we can repeat as required.

3e. Proposed Duration for Sample/Reagent Preparation *(If sample collection / preparation is prospective)***:** *(MM/DD/YYYY) – MM/DD/YYYY)*

3f. Proposed Date of Sample/Reagent Shipment: *(MM/DD/YYYY):*
At the GSC's request.

4. Approach to Data Production:

4a. Phylogenetic Dataset for Pyogenes and Mutans Groups
The phylogenetic dataset proposed here involves de novo sequencing of new species of *Streptococcus*, which have a genome size averaging about 2 Mb. We have found this to be best accomplished by 454 sequencing involving a 50:50 mix of shotgun and 3K Long-Tag Paired End sequencing with approximately 20X coverage on the genome sequencer FLX system (half of a full run). The resulting gaps in the assembled scaffold sequences can be easily closed using conventional PCR. Inter-scaffold gaps are more difficult, because the orientation of the scaffolds is not always clear and the gaps are usually longer. Generally, however, we find that with the range of complete genomes available for different species of *Streptococcus*, we can determine the correct orientation of the scaffolds through interspecific genomic comparisons and perform long range PCR. If this does not work then we resort to high resolution, ordered, whole genome restriction maps, performed by the company OpGen, and incorporate these data along with the 454 sequence reads. We envision obtaining the raw sequence reads from the GSC, assembling the sequence data with Newbler, and proceeding to close the genomes of the new species that are generated as part of this effort; closed genomes would be deposited on NCBI and also available on our "Streptococcus Genome Browser" (currently in development).

4b. Population Genomic Dataset for *S. agalactiae* and *S. penumoniae*
The population genomic datasets for *S. agalactiae* and *S. pneumoniae* would not require complete genome sequences but would require genome sequencing of significant numbers of isolates. Currently the best approach for this type of data production is Solexa/Illumina sequencing of individual strains. The level of sequence coverage depends on what you wish to do with the resulting data. Although the generation of SNP data across the core genome arguably only requires coverage of around 10-25X, we propose generating sequence coverage closer to 100X, because in our experience, for organisms with genomes approaching this 2 Mb size, something between 75-100X Solexa coverage allows the de novo assembly of both the core and non-core (dispensable) portions of the genome and thus, with the numbers of isolates we propose here, ultimately the entire pan-genome of the species. Therefore, with the new 86 bp potential of the Solexa apparatus, and the possibility of about 10 million reads per lane, we suggest 3 or 4 indexed strains per lane would provide this level of coverage and afford the assembly of the dispensable portion of the genome. Again, we envision receiving the raw sequence reads from the GSC, from which we would assemble the sequence using Velvet software.

5. Community Support and Collaborator Roles:

Identifying which genes and sites within genes are of key functional significance, across a pathogen's genome is fundamental to our understanding of disease biology. Comparative genomic sequence analysis plays a pivotal role in this endeavor by providing robust and

powerful analytical tools for leveraging evolutionary divergence and convergence to study function across a range of different organisms. The bacteria genus *Streptococcus* includes some of the most important human pathogens, causing a wide range of different disease, and inflicting significant morbidity and mortality throughout the world, as well as resulting in significant economic burden. These pathogenic species can be phylogenetically divided into three key systematic groupings within the genus, known as Pyogenes, Mitis, and Mutans. The Pyogenes group includes the important human pathogens *S. agalactiae* and *S. pyogenes*, as well as a number of additional human, and zoonotic pathogens. The Mitis group includes *S. pneumoniae* and the Mutans group includes several species linked to dental caries, including *S. mutans* and *S. sobrinus*. The fundamental data on comparative genomics and molecular adaptation, arising from this project, will serve as a framework for the development of novel therapeutic and preventative strategies for these pathogenic species of *Streptococcus*. The results from these evolutionary analyses will serve as a guide for future follow-up, cause effect experimentation. As part of a "proof of concept" component of the RO1 work associated with this sequencing proposal, based on the results arising from the selection analyses, and identification of noncoding functional elements, we will create gene knockout and site specific mutants for strains of *S. mutans*, and assay their phenotypic effects. These experiments are designed to test our hypothesis that many loci, sites within loci, and noncoding functional elements, identified as being under positive selection pressure in our evolutionary analyses, will be key loci involved in the colonization, persistence, survival, and propensity to cause disease in these organisms. All of our genome sequence data and the analyses associated with it will be displayed on our *Streptococcus* Genome Browser where we intend to display information correlating gene gain/loss and molecular adaptation within genes, to disease characteristics of the various pathogens. Concomitant with the associated publications, we anticipate this will allow us to reach the widest possible public audience and the work can be of greatest benefit to an extensive range of scientists interested in streptococcal biology.

Funds from the existing award (NIH/NIAID; 1 R01 A1 073368-01A1) related to this work are available to cover postdoc salaries, technicians, programmers, laboratory consumables, and OpGen expenses as necessary.

The various individuals on this project and a brief description of their roles on the project are listed below:

Michael J. Stanhope (Principal Investigator), Professor of Evolutionary Genomics, supervises two of the projects postdocs, and the laboratory technician, conducts evolutionary analyses, prepares manuscripts, and oversees overall project management. Dr. Stanhope has over 20 years of experience in molecular evolutionary biology, including both computational and laboratory components, covering a wide range of specific disciplines, including a six year period leading a group of evolutionary biologists in the Bioinformatics department of GlaxoSmithKline, where the focus was microbial. Dr. Stanhope's current line of research is predominately devoted to microbial evolutionary genomics, involving various microbial pathogens.

Carlos Bustamante (Co-Investigator), Professor in the Department of Biological Statistics and Computational Biology, jointly supervises (along with Dr. Siepel) one of the postdocs, conducts population genetic analyses, prepares manuscripts, and develops and refines population genetic analysis methods, specifically for our *Streptococcus* data. Dr. Bustamante's research focuses on developing statistical methods for parameter

estimation and hypothesis testing in population genetics and molecular phylogenetics. He has been actively engaged in developing methods for estimating the relative contributions of demographic forces (e.g., population structure, population size expansion / contraction) and selective forces on the history of natural populations using data from standing genetic variation as well as fixed differences between populations, and has applied such approaches to genome scan assessments of molecular adaptation in humans.

Adam Siepel (Co-Investigator), Assistant Professor in the Department of Biological Statistics and Computational Biology, co-supervises one of the postdocs (along with Dr. Bustamante), oversees the construction of the *Streptococcus* Genome Browser, develops and applies computational methods to identify noncoding functional elements, and prepares manuscripts. Dr. Siepel's expertise is in comparative genomics, software development, and evolutionary bioinformatics. He has been engaged in several eukaryotic genome sequencing projects, has significant experience developing bioinformatics software (National Center for Genome Resources), has developed widely used methods for identifying coding and noncoding functional elements, and was until recently a member of the University of California at Santa Cruz Genome Browser team. He has, therefore, a detailed understanding of the current versions of the browser, an inside perspective on how it can be adapted to our *Streptococcus* data, and a long term working relationship with the UCSC team.

Robert Burne (Co-Investigator), Professor and Chair of Oral Biology at the University of Florida, supervises a postdoc in his laboratory, at the University of Florida, responsible for performing the genetic engineering experiments with *S. mutans*, which target loci identified in the evolutionary analyses. This portion of the work begins in July/2010. Dr. Burne also provides *S. mutans* isolates for the population genomic analyses of this species, as well as consulting expertise on the biology of this organism. Dr. Burne's laboratory studies the molecular mechanisms governing the ability of bacteria that are capable of causing diseases in humans, to modulate their virulence in response to environmental influences.

Melissa Hubisz is the project's programmer, works under Dr. Siepel's supervision and is responsible for all software and database development for the *Streptococcus* Genome Browser, as well as other miscellaneous programming duties as required on the project.

Paulina Pavinski Bitar is the project laboratory technician. She works in Dr. Stanhope's laboratory and performs basic microbiology and molecular biology laboratory procedures, such as isolating DNA from the experimental strains, primer design, PCR, genome closing, and collating the resulting sequence data received from the sequencing facility.

Tristan Lefebure and Haruo Suzuki are Research Associates, trained in evolutionary bioinformatics, that work in Dr. Stanhope's laboratory, and are involved in comparative genomic and evolutionary genetic analyses of the resulting genome sequence data.

Omar Cornejo is a postodoc working under Dr. Siepel and Dr. Bustamante's supervision. He is involved in the development, implementation, and genome-wide application of new computational methods for identifying noncoding functional elements, and for detecting positive and/or lineage-specific selection in noncoding elements. In addition, he also contributes to the analysis of the population genetic data for *S. mutans*, *S. agalactiae*, and *S. pneumoniae,* as well as develops new models for population demographic history and selection.

6. Compliance Requirements:

6a. Review NIAID's Reagent, Data & Software Release Policy:

http://www3.niaid.nih.gov/research/resources/mscs/data.htm

http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-013.html

Accept ☒ Decline ☐

6b. Public Access to Data and Materials:

Regarding the de novo sequencing of species of *Streptococcus* for the phylogenetic based analyses, genome sequence data, in the form of draft sequences, will be released to NCBI within 45 days of its generation. Closed genomes will then be released to NCBI as soon as the closed circle is verified. In addition, all our genome sequence data will be available on our Genome Browser for internet download. Isolates will be made available to NIAID, and in the case of the new species, are already available on ATCC or some other public repository. All of our genome sequence data and the analyses associated with it will be displayed on our *Streptococcus* Genome Browser where we intend to display information correlating gene gain/loss and molecular adaptation within genes, to disease characteristics of the various pathogens.

For the population genomic datasets, sequence data will be deposited on NCBI under their sequence read archive database, within 45 day of its generation. The majority of these isolates are already available on public repositories, and the remainder will be made available by us. In the case of *S. pneumoniae*, we have information on serotype and antibiotic susceptibility that will accompany these isolates for public repository and appearing on our Genome Browser.

6c. Internal Review Board (IRB) / IACUE Yes ☐ No ☐

Investigator Signature:

Investigator Name: **Michael J. Stanhope** **Date:** **Nov. 2/2009**