Signatures of non-neutral processes within the population structure of *Streptococcus pneumoniae*

José Lourenço[a,1], Eleanor R. Watkins[a], Uri Obolski[a], Samuel J. Peacock[a], Callum Morris[b], Martin C. J. Maiden[a], and Sunetra Gupta[a]

[a]Department of Zoology, University of Oxford, Oxford, OX1 3PS, United Kingdom; [b]University of Durham

This manuscript was compiled on October 24, 2016

Populations of *Streptococcus pneumoniae* are typically structured into groups of closely related organisms or lineages. Here, we employ a machine learning technique to try and tease out whether these lineages are maintained by selection or by neutral processes. Our results indicate that lineages of *S. pneumoniae* evolved through selection on the *groESL* operon, an essential component of its survival machinery. This operon contains genes which encode chaperone proteins that enable a very large range of proteins to fold correctly within the physical environment of the nasopharynx and therefore will be in strong epistasis with several other genes. These features of *groESL* would explain why lineage structure is so stable within *S. pneumoniae* despite high levels of horizontal genetic transfer. *S. pneumoniae* is also antigenically diverse, exhibiting a variety of distinct capsular serotypes. We show that associations may arise between lineage and capsular serotype due to immune selection and direct resource competition but these can be more easily perturbed in the presence of external pressures such as vaccination. Overall, our analyses indicate that the evolution of *S. pneumoniae* can be conceptualized as the rearrangement of modular functional units occurring on several different timescales under different selection pressures: some patterns have locked in early (such as the epistatic interactions between *groESL* and a constellation of other genes) and preserve the differentiation of lineages, while others (such as the associations between capsular serotype and lineage) remain in continuous flux.

Pneumococcus | Selection | Competition | Metabolic

Many bacterial pathogen populations contain a number of co-circulating lineages bearing unique signatures of alleles at selected housekeeping loci [1] and also at a whole genome level [2–4]. The maintenance of these discrete lineages is hard to ascribe to purely neutral processes, given the high rate of genetic exchange in these pathogen populations [5]. We have previously proposed that extensive co-adaptation between loci may give rise to these patterns, as even small fitness differences among different combinations of alleles can lead to the loss of less 'fit' lineages under intense competition for resources [6]. Bacterial populations may also segregate into a set of successful 'metabolic types' which are able to co-circulate by virtue of exploiting separate metabolic niches and thereby avoiding direct resource competition and immune pressures [7]. As an example, specific differences in the ability to absorb particular carbohydrate resources have been observed in functional genomics studies of *Streptococcus pneumoniae* [8], and these may reflect specialization upon different resources within the same environment as a means of avoiding competition.

Several bacterial pathogens are also antigenically diverse: *S. pneumoniae*, for example, can exist in over 90 different serologically distinguishable states or 'serotypes' [9]. Many bacterial populations – including *S. pneumoniae* - exhibit strong associations between antigenic type and lineage, at least at the level of MLST [7, 10]. Such associations may have arisen through neutral processes; alternatively, as we have previously demonstrated, they may represent the outcome of a combination of immune selection acting upon antigen genes and direct resource competition acting upon metabolic genes and virulence factors [6, 10]. Distinguishing between these two hypotheses is complicated by the high levels of linkage disequilibrium observed across the whole genome [10, 11]. However, the alternative hypotheses make very different predictions about how the system would respond to perturbation by vaccination, particularly when only a subset of antigenic types are included in the vaccine, as is the case for *S. pneumoniae*. Under these circumstances, antigenic types that are not included in the vaccine may be expected to increase in frequency but, under the neutral model, it would be highly unlikely that they would do so in association with the genotypes previously associated with vaccine serotypes. By contrast, if the associations were primarily generated by selection, one would expect non-vaccine serotypes to become associated with genotypes that were previously commonly associated with vaccine serotypes [10]. Evidence for this phenomenon of Vaccine

DRAFT

Significance Statement

Populations of *Streptococcus pneumoniae* (the pneumococcus) appear to form stable clusters of closely related organisms despite the fact that they frequently exchange genetic material. In this paper we show that, rather than emerging by chance, these clusters have evolved to maintain their differences so that they may avoid competing with each other. Our work suggests that these clusters are fundamentally determined by variation within a set of genes which encode "chaperones" that help other pneumococcal proteins fold correctly under changes in the physical environment they would encounter while trying to infect their vertebrate host. These chaperone proteins are also targets of immunity and therefore may have originally diverged to minimise immunological interference between pneumococci, thereby necessitating changes across the whole genome.

S.G., J.L. and E.W. designed the study; S.G. and J.L. conducted the study. All authors were involved in analysis and interpretation of results and available data and in writing the paper.

We declare we have no competing interests.

[1]To whom correspondence should be addressed. E-mail: jose.lourenco@zoo.ox.ac.uk

Induced Metabolic Shift (VIMS) has been found among pre- and post-vaccine isolates collected in the USA [2, 4, 12–14] and South Korea [15].

Establishing the contribution of co-adaptation and metabolic competition in the maintenance of lineage structure is thus important as the outcome of certain interventions, such as vaccination, depends crucially on these underlying determinants of population structure. Here, we assess the potential to achieve this by machine learning techniques, which work by attempting to identify relevant features based on information supplied on a set of potential predictor variables for each individual genetic sample. Random forests (RF) are one of such methods currently witnessing a surge of attention, owing to its unique advantages in dealing with large datasets of both numerical and categorical data, as well as having low computational overhead, a nonparametric nature and a well defined probabilistic output [16]. A random forest algorithm (RFA) is an ensemble method that combines the information of multiple, regression or classification trees built around predictor variables towards a response variable. The output of an RFA is composed both of the classification success rates of the response variable and a ranking of the predictor variables (scores) quantifying their relative role in the classification process. RFA-based methods are widely applied in genome-wide association studies of cancer and chronic disease risk [17], drug resistance [18], species classification [19], and in the analysis of microarray data [20]. In the context of host-pathogen systems, machine learning techniques have been shown to be able to successfully ascertain host tropism, for instance by identifying the key sites that determine host specificity of zoonotic viruses [21], by analyzing the probability of *Escherichia coli* cattle strains more likely to be virulent to humans [22], and by selecting the clear genetic distinctions in both avian and human proteins of Influenza viruses [23, 24].

In this paper, we undertake a feature selection analysis of a dataset containing 616 whole genomes of *S. pneumoniae* collected in Massachusetts (USA), including 133 samples from 2001, 203 from 2004 and 280 from 2007 [3, 25], thus representing the bacterial population at the point of PCV7 introduction in year 2000, and any changes that may have followed. These data have been used in numerous studies, including analysis of post-vaccine epidemiological and genetic changes [3, 10, 26], maintenance of population structure [2], *beta*-lactam resistance [27], determinants of colonization [28] and constraints on serotype switching [29]. Each isolate in this dataset contains information on its capsular serotype (determined by serological means), and had also been assigned to one of a number of monophyletic Sequence Clusters (SC) using a phylogenetic and clustering analysis on a core genome built from all putative protein-coding sequences that were present in a single copy in all genomes [3]. Using a machine learning technique and a previous allelic annotation of 2135 genes among these isolates (using ATCC 700669 serotype 23F as reference [10], table S1), we attempt to identify the relative contribution of each gene in maintaining the observed population structure in terms of (i) capsular serotype and (ii) Sequence Cluster (SC). We find a clear distinction between the sets and functions of genes highly informative for serotype versus SC, suggesting that different selective processes have led to the emergence and maintenance of *S. pneumoniae*'s population structure.

Results

Genes which predict serotype do not perform well in predicting Sequence Cluster.

We first assessed the success of the combined variation in 2135 genes of known and unknown function in identifying the Sequence Cluster (SC) to which isolates belonged, this being a measure of shared ancestry (as per [3]). Classification of SC by RFA was accurate (Fig. S1B) with all SC types being predicted with success close to 100%. By contrast, the success rate in identifying the capsular serotypes of the 616 whole genomes, although also very high, was not perfect. None of housekeeping genes included in MLST classification performed better than average in predicting serotype or SC (Fig. 1).

As might be expected, genes within the capsular locus (defined as being within but not including the genes *dexB* and *aliA*) achieved high scores in predicting serotype but these did not score above average in predicting SC (Fig. 1). We noted that many of these genes contained what appeared to be a high proportion of deletions but, in fact, had simply eluded allelic notation on account of their high diversity at the level of the population. For certain genes, such as those encoding the polysaccharide polymerase Wzy and the flippase Wzx, the allelic notation process failed at least 50% of the time for over 90% of the isolates, essentially working only for 23F (the reference genome) and the closely related 23A and 23B serotypes. In general, the degree of success in allelic notation of each gene was closely linked to the potential for alignment with its counterpart in the 23F reference genome (Fig. S4). Nonetheless, the same shift towards lower RFA scores of capsule associated genes in predicting SC rather than serotype was observed upon performing these classification exercises after excluding all genes which contain $> 50\%$ (Fig. S2) or $> 10\%$ (Fig. S3) of gene mismatches/deletions. When imposing an exclusion criterion of $> 10\%$ we retained only the genes *wze*, *wzg* and *wzh* (in addition to two pseudogenes), and these could also clearly be seen to shift from above the upper 97.5% limit into the neutral expectation of RFA scores when predicting SC (Fig. S3).

Finally, we performed the same analysis excluding all genes which showed mismatches or deletions above a threshold of 1%. This eliminated all of the genes considered above as belonging within the capsular locus, although many flanking genes were retained and a number of these achieved the top 2.5% of RFA scores in predicting serotype (Fig. 2A, Table 1): 38% of the top genes occurred within 10 genes downstream and upstream of the capsular locus, and 66% were situated within 60 genes (a distance amounting to 2.8% of the genome). None of the genes achieving the top 2.5% of RFA scores in predicting serotype (shown in red in Fig. 2) remained in the top 2.5% category when asked to predict SC. Similarly, all genes which achieved top scores in predicting SC (Table 2) were only of average importance in elucidating serotype (shown in green in Fig. 2). Interestingly, the MLST gene *spi* gained a place among the top-scoring genes for SC (Table 2) under this stringent cutoff.

Top-scoring genes for serotype classification mediate competitive interactions.

We found that a large proportion of non-capsular genes which

DRAFT

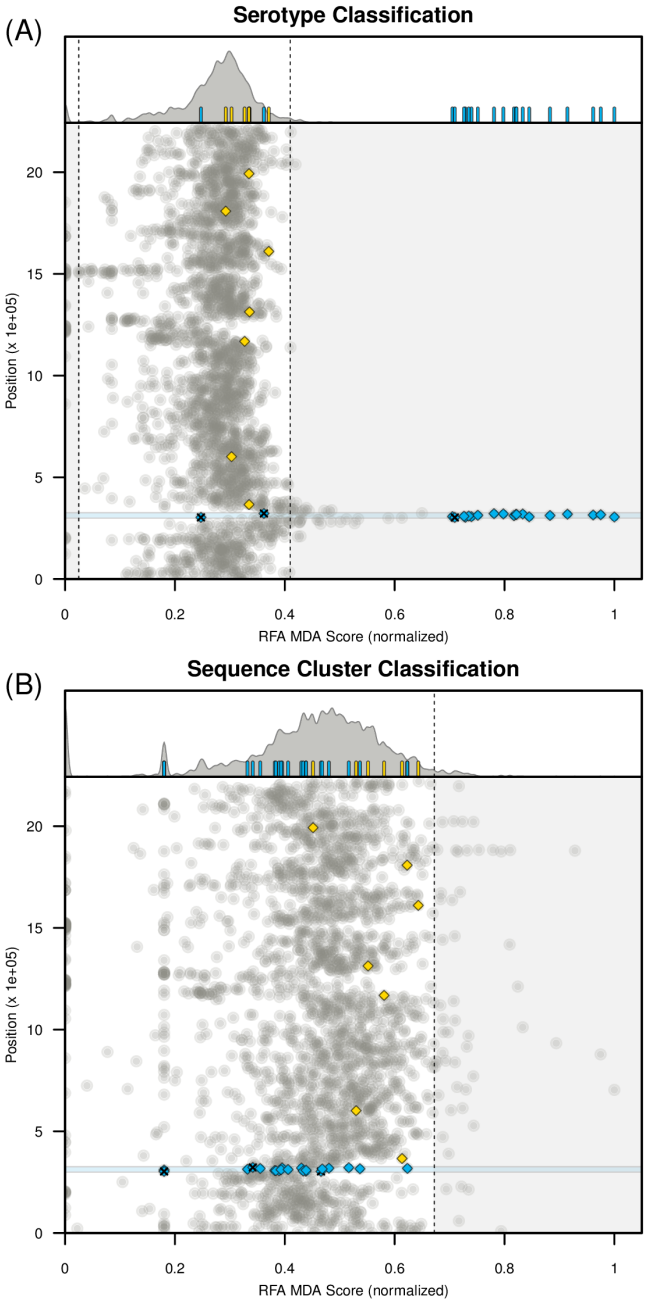


Fig. 1. Random forest classification. *(A)* Random forest analysis (RFA) for serotype classification. *(A, top)* Density function of normalised RFA scores with 95% boundaries marked by the dashed lines. Small bars highlight the position and types of particular genes. *(A, bottom)* Genomic position for each gene in the dataset against their normalised RFA score. The circular genome is presented in a linear form, with the first gene being *dnaA* and the last gene *parB*. MLST genes are marked in yellow diamonds (*spi, xpt, glkA, aroE, ddlA, tkt*) and genes within the capsular locus with blue diamonds (pseudogenes tagged with 'x'). *(B)* RFA for sequence cluster classification; figure details the same as in *A*. Blue shared areas mark the capsular locus (genes within *aliA* and *dexB*).

were associated with serotype to be involved in metabolic functions linked to resource competition, at least in related streptococcal species. The top-scoring gene *trpF*, for example, is known to be essential for the biosynthesis of tryptophan for *S. pneumoniae* [30], and more generally of the biosynthesis of aromatic amino acids in at least 9 species of bacteria [31]. Another top-scoring gene, *fabG*, encodes the β-ketoacyl-ACP reductase, the only known keto-acid reductase in bacterial fatty acid biosynthesis [32]. The gene *lysC* codes for an aspartokinase involved in lysine production and aminoethyl cysteine resistance in *Corynebacterium glutamicum* [33]. We also found two genes, *mvaD* and *mvaK2*, of the mevalonate pathway scoring highly for serotype prediction. This pathway, also known as the HMG-CoA reductase pathway, can be found in bacteria, eukaryotes and archaea [34]. One of its main products, the isopentenyl pyrophosphate (IPP), is used to make isoprenoids, a diverse class of over 30,000 biomolecules. In bacteria, the principal products of IPP include the lipid carrier undecaprenol (involved in wall biosynthesis), plus a range of menaquinones and ubiquinones both involved in electron transport, and the latter also in aerobic cellular respiration [34–36]. In *S. pneumoniae*, these two genes are essential for growth and are proposed to be part of a single operon [35].

ATP-binding cassette (ABC) transporter genes, which are critical for intake and metabolism, were found 5 times more frequently in the top genes classifying serotype compared to those determining SC (Table 1). As part of one such transporter, the SpuA protein is involved in α-glucan metabolism, whose main substrate is glycogen (polysaccharide of glucose), an abundant resource in human lung epithelial cells [37, 38]. The gene *patB* also encodes part of an ABC efflux pump in *S. pneumoniae*, responsible for resistance to fluoroquinolenes [39–41]. Among other ABC transporters, two other genes were located within the *pit* operon which is involved in iron uptake. In line with our findings, the *pit* operon has previously been shown to exhibit strain-specific variation [42]. In contrast, our approach did not select 2 other operons involved in iron uptake (*piu* and *pia*), which are conserved between *S. pneumoniae* strains [42] and therefore unlikely to be predictors of serotype. Another important regulator of iron transport among the top-scoring genes is *gnd* [43]. The latter is transcriptionally linked to another top gene, *ritR*, which is orthologous to the streptococcal global regulator *covR*, for which there is conclusive evidence from *S. pyogenes*, *S. suis* and *S. agalactiae* of regulatory functions on capsular biosynthesis [44–46]. Another ABC transporter known as Ecs is represented in the top list by one of its two genes, *ecsA*. The substrate of Ecs is so far unknown, but obligatory anaerobes or microaerophilic bacteria do not carry the Ecs transporter, and its function is therefore argued to be related to respiration [47]. Finally, transport can be achieved by a multitude of systems alternative to ABC transporters, such as 'passive' channels like the top-scoring sodium symporter GlyP [48]. Sodium is one of the main electrolytes in human saliva, existing there at a higher concentration than in blood plasma, and differentiation in sodium transport, similarly to iron or glucose transport, could potentially be under selection for niche specialization.

High RFA scores for serotype were also found among a number of genes flanking the capsular locus which are involved in the cell wall peptidoglycan biosynthesis pathway [9].

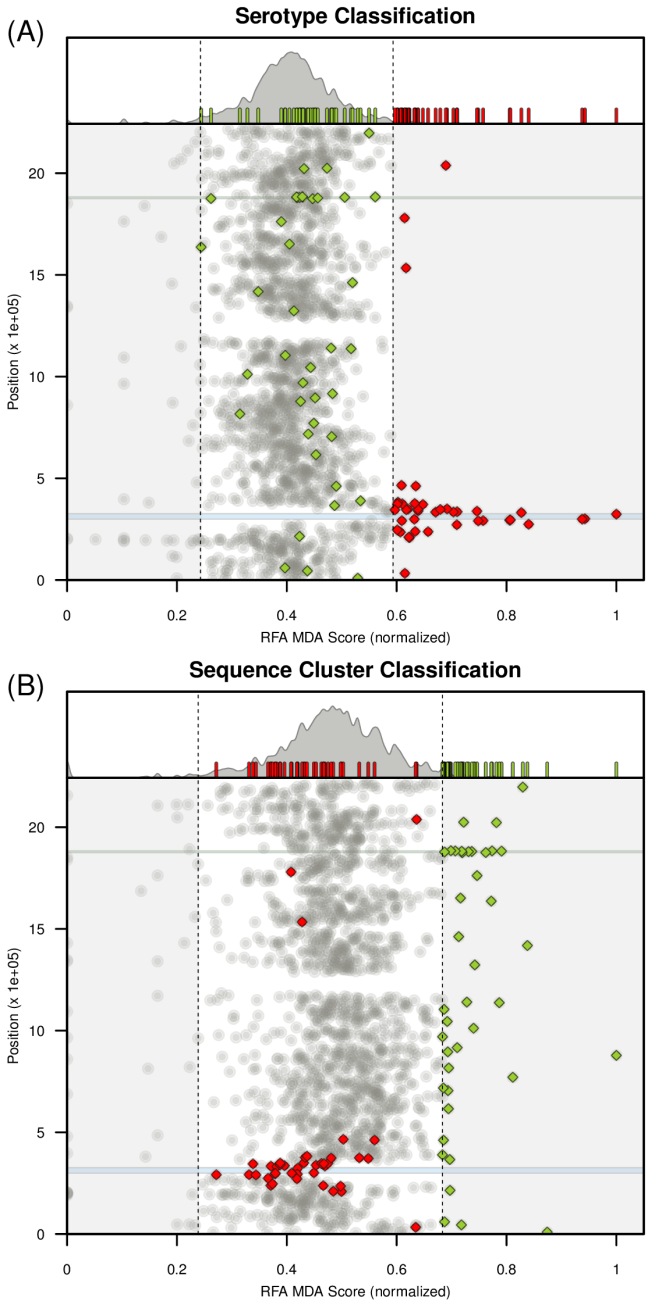


Fig. 2. Random forest classification excluding data with gene mismatches. *(A)* Random forest analysis (RFA) for serotype classification when excluding genes for which the allelic notation process had $< 99\%$ positive matches with the reference genome. *(A, top)* Density function of normalised RFA scores with the 95% boundaries marked by the dashed lines. Small bars highlight the position and types of particular genes. *(A, bottom)* Genomic position for each gene in the dataset against their normalised RFA score. The circular genome is presented in a linear form, with the first gene being *dnaA* and the last gene *parB*. Red and green damonds mark the top 2.5% ranking genes for serotype and Sequence Cluster classification, respectively. *(B)* RFA for Sequence Cluster classification; figure details the same as in *A*. Blue shared areas mark the capsular locus (genes within *aliA* and *dexB*). Green shaded areas mark the genes contiguous and including the *groESL* operon (Table 2).

These include the penicillin-binding protein genes *pbpX* and *pbp1A*, the 16S rRNA cytosine-methyltransferase gene *mraW* and the phospho-N-acetylmuramoyl-pentapeptide-transferase gene *mraY*. Mutations in these genes can lead to penicillin resistance, and single-nucleotide positions in all three genes have been shown to associate strongly with *S. pneumoniae* β-lactam resistance in genome-wide association studies (GWAS) performed on the dataset used in this study [3], in a Thai study containing 3,085 isolates [49], and in a Canadian study on 11,083 isolates [50]. It is of relevance to note that in *S. Pneumoniae*, *pbp1A* is also involved in the formation of the septum during cell division [51] and is associated in a two-gene operon with the top-scoring gene *recU*, coding for the Holliday junction resolvase, required for homologous DNA recombination, repair and chromosome segregation [52, 53]. Finally, resistance to various classes of cell wall-inhibitory antibiotics (ex. methicillin, vancomycin, daptomycin) in *S. Aureus* is regulated via the vra operon, by up or downregulation of a set of genes commonly designated as the cell wall stimulon [54]. We find this operon represented by two entries, the *vraT* and *vraT* genes.

In addition to genes clearly related to critical resource functions, transport and antibiotic resistance, we also found some of the top-scoring entries to be involved in functions associated with direct inter- and intra-species competition, either through factors related to immune escape or warfare. For instance, *blpH* is part of the BlpABCSRH pathway [55], which regulates production of class II bacteriocins and related immunity proteins [56, 57]. In related species, the aminotransferase GlmS is also known to upregulate the production of ammonia thereby increasing acid tolerance and survival [58]. The capsular flanking gene *luxS* is also a good example, as it is part of a *Staphylococcus epidermidis* quorum-sensing system in biofilm formation, and linked to pneumolysin expression, a key player in interference with the host immune response [59, 60]. Finally, the top-scoring *lytC* gene encodes a lysozyme (or glycoside hydrolase) which can be found in a number of secretions, such as tears, saliva and mucus, with the potential to damage (inter-species) bacterial cell walls by catalyzing hydrolysis of linkages and residues in peptidoglycans and chitodextrins [61, 62].

Several top-scoring genes for SC classification are also key determinants of phenotype.

A number of top scoring genes (ex. *sodA*, *groEL*, *groES*, *lmb*) in predicting SC have previously been demonstrated to be powerful discriminators of genealogy in a range of bacterial species. For instance, *sodA*, encoding for the manganese superoxide dismutase, critical against oxidative stress and linked to both survival and virulence, has been highlighted in numerous studies for its relevance in identification of rare clones of pneumococci [63, 64] and Streptococci at the species level [65, 66]. Also, the *lmb* gene encodes for an extracellular protein with a key role in physiology and pathogenicity [67, 68], and homologs of this protein have been documented to be present and discriminatory of at least 25 groups of the *Streptococcus* genus with possible similar functions [69, 70].

Certain top-scoring genes were strongly associated with phenotype such as cell-shape, virulence or invasiveness. For instance, glycolytic enzymes (GE) such as the one encoded by the top-scoring gene *pdhB* have long been regarded as virulence factors [71] and are involved in cytosol-located metabolic

processes. When transported to the surface, the PdhB protein-complex is known to interact with host factors such as the extracellular matrix and fibrinolysis system [72]. Critically, *Mycoplasma pneumoniae's pdhB* is involved in the degradation of human fibrinogen and is also able to bind human fibronectin [72, 73]. Fibronectin is commonly found in human saliva, presenting a vast set of functions, from prevention to colonization of the oral cavity and pharynx, to involvement in adhesion and wound healing [74]. Another top gene, *pclA*, encodes for the pneumococcal collagen-like protein A, a top candidate for human collagen mimicry [75], involved in host-cell adherence and invasion [76]. Binding to fibronectin and collagen are common strategies employed by various invading bacterial pathogens to colonize or disseminate within the host [77, 78]. In ovococcus bacteria such as *S. pneumoniae* the function of the top-scoring protein MreD (the Rod shape-determining protein) is unknown. It is therefore down to speculation on why this protein is a good predictor of SC, but since the depletion of MreD protein can cause cells to stop growing, become spherical, form chains and lyse, its selection hints on the possibility that variation in this gene may dictate specific lineage differences in cell-shape phenotype [79]. We also find the genes designated as SPN23F11320 and SPN23F09460 to be relevant for SC classification, which in our dataset represent about 13% of all non-putative GCN5-related, N-acetyltransferases of the (GNAT) family. These are key proteins involved in acetylation, and there is growing evidence in the literature of their role in regulation of central carbon metabolism and phenotype through epigenetics [80, 81].

Overall, the characteristics of these top-scoring genes differed significantly from those which were successful in predicting serotype and, contrary to expectations from a population structured mainly by neutral evolution, we found the top-scoring genes for SC (ancestry) not to be uniformly distributed across the genome. Most strikingly, 25% of the top scoring genes for SC were contiguous and contained the *groESL* operon, which includes the GroEL and GroES chaperon proteins (Table 2). Other studies have reported the power of the *groESL* operon and its proteins to ascertain phylogeny and classification within the *Streptococcus* genus [82] and between species of the *Viridans* and *Mutans* Streptococci groups [83, 84]. We also noted the top-scoring gene *recX* is in close proximity to the *groESL* operon, which encodes a regulatory protein that inhibits the RecA recombinase in multiple species of bacteria [85–88].

Discussion

We have presented a novel technique for attempting to distinguish the effects of selection from neutral processes giving rise to population structure by applying a machine learning algorithm to genomic data. Our strategy involves applying a Random Forest Algorithm (RFA) to predict particular features (serotype or Sequence Cluster) of each isolate from information on the allelic composition of all isolates. By comparing the contribution of different genes as reflected in their RFA scores in predicting serotype or Sequence Cluster, inferences can be made concerning the evolutionary processes underlying their formation, relationship and maintenance at the population level. We performed this analysis on a dataset containing 616 whole genomes of *S. pneumoniae* collected in Massachusetts (USA) [3] , for each of which we had obtained allelic profiles of 2135 genes [10].

Table 1. Top genes for serotype prediction

SPN23F		(name) Function
00400		Hypothetical protein
02300		(pitA) Ferric iron ABC transporter, permease protein
02320		(pitB) Ferric iron ABC transporter, ATP-binding protein
02540	a	(glmS) Glucosamine-frutose-6-phosphate aminotransferase
02550	a	Luciferase-like monooxygenase / Oxidoreductase
02560	a	(spuA) Surface-anchored pullulanase
02600		(polC) DNA polymerase III PolC-type
02870	b	Maltodextrin glucosidase
02880	b	(basA) Glutathione peroxidase family protein
03060	c	**(mraW) 16S rRNA cytosine-methyltransferase**
03070	c	**(ftsL) Cell division protein**
03080	c	**(pbpX) Penicillin binding protein / cell division protein**
03090	c	**(mraY) Phospho-N-acetylmuramoyl-pentapeptide-transferase**
03110		**(clpL) ATP-dependent Clp proteinase**
03130	d	**(luxS) S-ribosylhomocysteinase lyase**
03140	d	**ATP-dependent Zinc protease**
03150	d	**(dexB) Glucan-1 6-alpha-glucosidase**
03390		**(aliA) Oligopeptide ABC transporter**
03410	e	**(pbp1A) Transpeptidase / Penicillin-binding protein**
03420	e	**(recU) Holliday junction resolvase**
03430	e	**Hypothetical protein**
03450	c	**23S rRNA / guanine-methyltransferase**
03470	f	**(gnd) 6-phosphogluconate dehydrogenase**
03480	f	**(ritR) Response regulator**
03540	g	(mvaD) Mevalonate diphosphate decarboxylase
03550	g	(mvaK2) Mevalonate kinase
03560	g	(fni) Isopentenyl-diphosphate delta-isomerase
03570	g	(vraT) Cell wall-active antibiotics response protein
03580	g	(vraS) Sensor histidine kinase
03840		(glyP) Sodium glycine symporter
03860	h	(shetA) Exfoliative toxin A
03870	h	(serS) Seryl-tRNA synthetase
03890		(lysC) Aspartokinase
03960		(fabG) 3-oxoacyl-acyl-carrier protein reductase
04740		(ecsA) ABC transporter ATP-binding protein
04790		(blpH) Histidine kinase of the competence regulon ComD
15900		(lytC) Glucan-binding domain / Lysozyme M1
18330		(trpF) Phosphoribosylanthranilate isomerase
20980		(patB) Multidrug resistance ABC transporter, ATP-binding protein

Genes in **bold** flank the capsular locus up to 10 genes in distance. Letters *a* to *h* in the second column denote groups of contiguous genes.

Classification success of Sequence Cluster (SC) to which each isolate belonged was achieved almost perfectly by the RFA. This is a reflection of the strong correspondence between taxonomy and classification trees based on genetic information, as explored in recent studies [19], and demonstrated by Austerlitz and colleagues when comparing the success of RFA, neighbour-joining and maximum-likelihood (PhyML) methodologies on simulated and empirical genetic data [89]. Classification of serotype by the RFA was more variable and, most importantly, there was no overlap between the genes which appeared to be most important in determining serotype and those which scored highly in identifying SC. As might be expected, genes of the capsular locus (cps) and many of those flanking it achieved high RFA scores in predicting serotype but did not perform better than average in predicting SC. Interestingly, none of the genes among the MLST loci showed a consistently strong association with SC across sensitivity

DRAFT

Table 2. Top genes for Sequence Cluster prediction

SPN23F		(name) Function
00090		Phospholycenate mutase
00540		(recO) DNA recombination and repair protein
00660		(vanZ) Teicoplanin resistance protein
02370		Transcriptional regulator
03790		(spi) Signal peptidase I
04050		Hypothetical protein
04730		Histidine triad nucleotide-binding protein
06210		ABC transporter, ATP-binding protein
06880		(sodA) Manganese superoxide dismutase
07240		Hypothetical protein
07340		Hydrolase / Haloacid dehalogenase-like family
07930		(iscU) Putative iron-sulfur cluster assembly scaffold protein
08320		Putative membrane protein
09040		O-methyltransferase family protein C1
09280		(lmb) Laminin-binding protein
09460		N-acetyltransferase (GNAT) family protein
10040		Cytosolic protein containing multiple CBS domains
10480		Hypothetical protein
10670		(pdhB) Acetoin dehydrogenase E1 component β-subunit
11320	a	Acetyltransferase (GNAT) family protein
11630	a	(licA) Choline kinase
11660		(carB) Membrane protein / O-antigen and teichoic acid
13490		Hypothetical protein
14640		(lta) Bacterocin transport accessory protein
15100		(pclA) Putative NADPH-dependent FMN reductase
16930		Hypothetical protein
17080		Hypothetical protein
18130		Hypothetical protein
19240	b	**(recX) Regulatory protein**
19250	b	**Cysteinyl-tRNA synthase related protein**
19300	c	**(groEL) Heat shock protein 60 family chaperone**
19310	c	**(groES) Heat shock protein 60 family co-chaperone**
19330	d	**Short-chain dehydrogenase**
19340	d	**(ytpR) Phenylalanyl-tRNA synthetase domain protein**
19360	e	**Hypothetical protein**
19370	e	**Hypothetical protein**
19380	f	**Membrane protein**
19390	f	**Response regulator of LytR/AlgR family**
20880		Hydrolase, haloacid dehalogenase-like family
20900		(thrC) Threonine synthase
22500		(mreD) Rod shape-determining protein

Genes in **bold** include and flank the *groESL* operon. Letters *a* to *f* in the second column denote groups of contiguous genes.

experiments and all performed no better at predicting SC than serotype.

We encountered difficulties in using the entire dataset due to the large number of putative deletions recorded. Some of these proved to be a result of the extreme diversity of genes (such as *wzx* and *wzy* within the capsular locus) which interfered with their alignment to the reference serotype 23F genome (ATCC 700669). In the entire dataset, around 7% of the genes had over $> 80\%$ deletions/mismatches recorded, 10% had $> 50\%$ deletions/mismatches recorded, and just over 25% of the data had to be discarded if we rejected all genes with an excess of $> 1\%$ of deletions/mismatches. Given these limitations, we repeated the RFA analysis under various cutoffs for percentage of gene deletions/mismatches in a series of sensitivity exercises. While this did not affect the trend of genes within and flanking the cps locus to shift to lower RFA scores when comparing prediction of serotype against prediction of SC, it thwarted

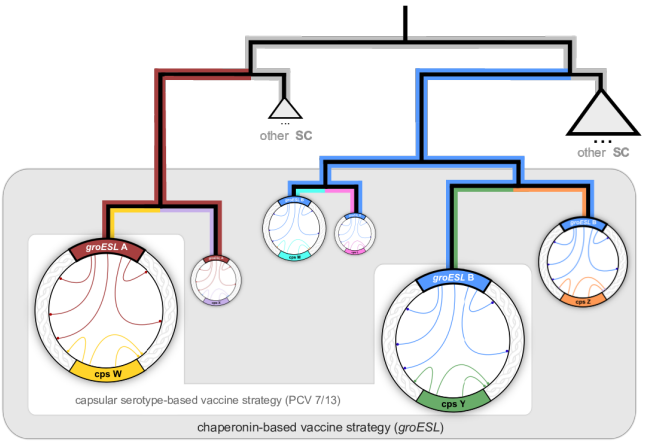


Fig. 3. Population structure and vaccination. Conceptual representation of phylogenetic relationships between serotypes and Sequence Clusters (SC), where the former are defined by variation at the cps locus (arbitrarily designated X, W, Y, Z, M, and L, respectively coloured yellow, purple, green, orange, cyan and pink) and the latter are linked to variation in the *groESL* operon (arbitrarily designated A and B and respectively coloured red and blue). Circles symbolize genotypes, with size relative to their prevalence. Inner genome arcs represent epistatic links: those with the *groESL* operon extend across the genome, while links with the cps locus are more local. Within our framework and according to observed patterns [3], most SCs will be dominantly associated with a single serotype. Current vaccine strategies (white area) that target a selection of capsular serotypes can lead to the expansion of non-vaccine serotypes (VISR, [26, 29]), potentially within the same sequence cluster (VIMS [10]). Vaccine strategies based on *groESL* variants (grey area) would target entire SCs instead, including all uncommon serotypes within and thereby preventing their expansion.

efforts to ascertain whether any specific associations with serotype existed among other highly variable surface proteins of interest: PspA, choline binding protein CpbA/PspC, the IgA proteases or the histidine triad proteins. Future work of this methodology will rely on the development of more robust methods of allele classification for this category of genes, an area still lacking adequate approaches.

By eliminating all genes with $> 1\%$ of deletions/mismatches, we were left with 1581 genes which likely corresponds to the 1500 'core' cluster of orthologous genes (COGs) identified by Croucher et al [2] in their recent analysis of the same dataset. Within this more restricted set, we also observed a clear disjunction between genes that score highly in predicting serotype and the top-scoring genes for predicting SC. Not surprisingly, a significant proportion of genes that were good markers for serotype were found to flank the capsular locus (shown in bold in Table 1), although there were a number which were distal to it. A high proportion of genes scoring highly for serotype prediction were associated with key functions in metabolism and very likely defined unique 'metabolic types', but since most were in proximity to the cps locus, it was not possible to determine whether these had become segregated through resource competition [10] or by physical and/or functional associations with this locus. The presence of several co-functional, co-transcribed or co-localizing sets of genes (eg. the *gnd* and *ritR* genes, the *pit*, *mva* and *vra* operons, and the penicillin-binding genes) on this list (Table 1) argues, however, that the evolution of these serotype-associated traits may best be understood within a modular framework in which different serotypes are characterized by particular combinations of these units.

DRAFT

Genes that were highly informative for SC classification were also not uniformly distributed across the genome, with around a quarter of them co-localizing within and around the *groESL* operon (shown in bold in Table 2), encoding a chaperonin system which remains a paradigm of macromolecular machinery for protein folding [90]. Apart from assisting protein folding by preventing inappropriate interactions between non-native polypeptides [90], this system may also buffer deleterious effects of mutations on protein foldability and stability [91], with important consequences for protein evolution. The protein GroEL is also highly immunogenic for different bacterial species and has been shown to provide strain-specific protection in vaccine studies [92–94]. This raises the radically alternative possibility that sequence clustering may have arisen from immune selection operating on these genes in conjunction with epistatic interactions between the relevant heat shock proteins and the loci encoding the proteins they are chaperoning. The associations between serotype and SC may thus be primarily driven by immune selection operating on multiple immunogenic loci (in this case, cps and *groESL*) causing them to be organized into non-overlapping combinations, as predicted by strain theory of host-pathogen systems [95, 96]. It has previously been proposed that immune selection acting jointly on capsular and sub-capsular antigens could account for the maintenance of these associations [29]. Immunological selection of unique combinations of cps and *groESL*, however, has the additional advantage of consolidating the link with a range of other genes across the genome through essential epistatic and highly specific (chaperoning) interactions with GroEL and GroES [90].

Our results are in broad agreement with the framework proposed by Croucher and colleagues [2], based on their analysis of the same dataset, in which lineage structure is maintained by infrequent transfer of modular elements ("macroevolution") and provides a stable backdrop for more frequent, and often transient, "microevolutionary" changes (see Figure 3). The differentiation of the *groESL* operon is potentially a striking example of "macroevolution", being specific not only to *S. pneumoniae* sequence clusters but also serving to genealogically distinguish closely related bacterial species [82–84]. We propose that this is the evolutionary outcome of a combination of immune selection and epistasis operating on specific modules, such as *groESL*, rather than neutral processes. Selection would also operate at a "microevolutionary" level in creating (more transient) associations between SC and serotype as means of avoiding immunological and direct resource competition [6, 10, 29]. We note that genes belonging to the Rec family are positioned in close proximity to both the contiguous clusters of top-scoring genes for SC and serotype (Tables 1 and 2) and would argue that these endorse the role of restriction-modification systems (RMS) in protecting the modularity of the genome [2], and that population structure arises through selection favouring particular combinations of variants of these modules. Our analyses support the hypothesis that lineage structure in maintained by co-adaptation and competition [6, 10] and show, unexpectedly, that these selection pressures converge upon the same locus, namely the *groESL* operon, and strongly endorse the development of vaccines targeting the associated chaperone protein GroEL to avoid vaccine induced changes in the population structure such as Vaccine Induced Serotype Replacement (VISR, [26, 29]) or Vaccine Induced Metabolic Shift (VIMS, [10]) which have the potential of greatly reducing the benefits of capsular serotype targeted interventions.

Materials and Methods

Sequence Data and Allelic Annotation. We used a dataset sequenced by Croucher et al, comprising 616 carriage *S. pneumonaie* genomes isolated in 2001, 2004 and 2007 from Massachusetts (USA). The data includes 133, 203, 280 samples from 2001, 2004, 2007, respectively; and is stratified into 16 samples of serotype 10A, 50 of 11A, 7 of 14, 24 of 15A, 60 of 15BC, 8 of 16F, 5 of 17F, 6 of 18C, 73 of 19A, 33 of 19F, 1 of 21, 21 of 22F, 33 of 23A, 23 of 23B, 17 of 23F, 11 of 3, 4 of 31, 5 of 33F, 6 of 34, 49 of 35B, 18 of 35F, 2 of 37, 9 of 38, 47 of 6A, 17 of 6B, 33 of 6C, 3 of 7C, 11 of 7F, 4 of 9N, 6 of 9V and 14 of NT (see [3] for collection details). In summary, allelic notation was carried out using the BIGSdb software with an automated BLAST process [97], and the genomes were analysed using the Genome Comparator tool (with ATCC 700669, serotype 23F, accession number FM211187, as the reference). Alleles identical to the reference were classified as '1', with subsequent sequences, differing at least by one base, labelled in increasing order. Genes were further classified as allele 'X' when genetic data present had no match to the genome of interest, or were found to be truncated or non-coding (see S1 Dataset of [10] for a visual representation of allele annotation and diversity). The allelic matrix as obtained by this approach and used in the RFA analysis is herein made available in supplementary Table S1, which also includes the Accession Numbers, gene name, gene product, gene position in reference genome, and year of collection, Sequence Cluster and serotype of each sample.

Random Forest Approach. We implement a machine learning approach based on a Random Forest Algorithm (RFA) to predict particular features (serotype or Sequence Cluster) of each pneumococci isolate from information on the allelic composition of 2135 genes [16]. In summary, the RFA process takes the following pseudo-steps: (I) the response variable and predictor variables are chosen by the user; (II) a predefined number of independent bootstrap samples are drawn from the dataset with replacement, and a classification tree is fit to each sample containing roughly 2/3 of the data, for which predictor variable selection on each node split in the tree is conducted using only a small random subset of predictor variables; (III) the complete set of trees, one for each bootstrap sample, composes the random forest, from which the status (classification) of the response variable is predicted as an average (majority vote) of the predictions of all trees. Compared to single classification tress, RFs increase prediction accuracy, since the ensemble of slight different classification results adjusts for the instability of the individual trees and avoids data overfitting [98].

Here we use randomForest: Breiman and Cutler's Random Forests for Classification and Regression, a software package for the R-statistical environment [99]. Predictor variables are set to be each gene in our genome samples and the response variable is set to the serotype or Sequence Cluster classification of each genome (as per [3]). We use the Mean Decrease Accuracy (MDA), or Breiman-Cutler importance, as a measure of predictor variable importance, for which classification accuracy after data permutation of a predictor variable is subtracted from the accuracy without permutation, and averaged over all trees in the RF to give an importance value [98]. For the results presented in the main text, we assume the predictor variables to be numerical (as opposed to categorical). This assumption is known to introduce RF biases, as classification is effectively made by regression and artificial correlations between allele numbering and the features being selected (serotype and Sequence Cluster) may be present. The assumption is herein necessary since the RFA R-based implementation (version 3.6.12) has an upper limit of 53 categories per predictor variable and we find some genes to present up to 6 times this limit in allele diversity. The categorical constraint is a common feature of RFA implementations, as predictor variables with N categories imply 2^N possible (binary) combinations for an internal node split, making

DRAFT

the RFA method computationally impractical. Given this inherent RFA limitation, we implemented an input shuffling strategy to minimize potential bias. For this, M random permutations of each gene's allelic numbering in the original dataset is performed, effectively creating M independent input matrices. The RFA is run over the input matrices and in the main results we present each gene's average MDA score. A sensitivity analysis was performed by comparing RFA results between two independent sets of $M = 50$ input matrices (effectively comparing 100 independent runs) (Fig. S5). Results suggest that the existing biases in independent runs of the RFA due to the assumption of numerical predictors are virtually mitigated with our shuffling approach, specially for experiments classifying serotype.

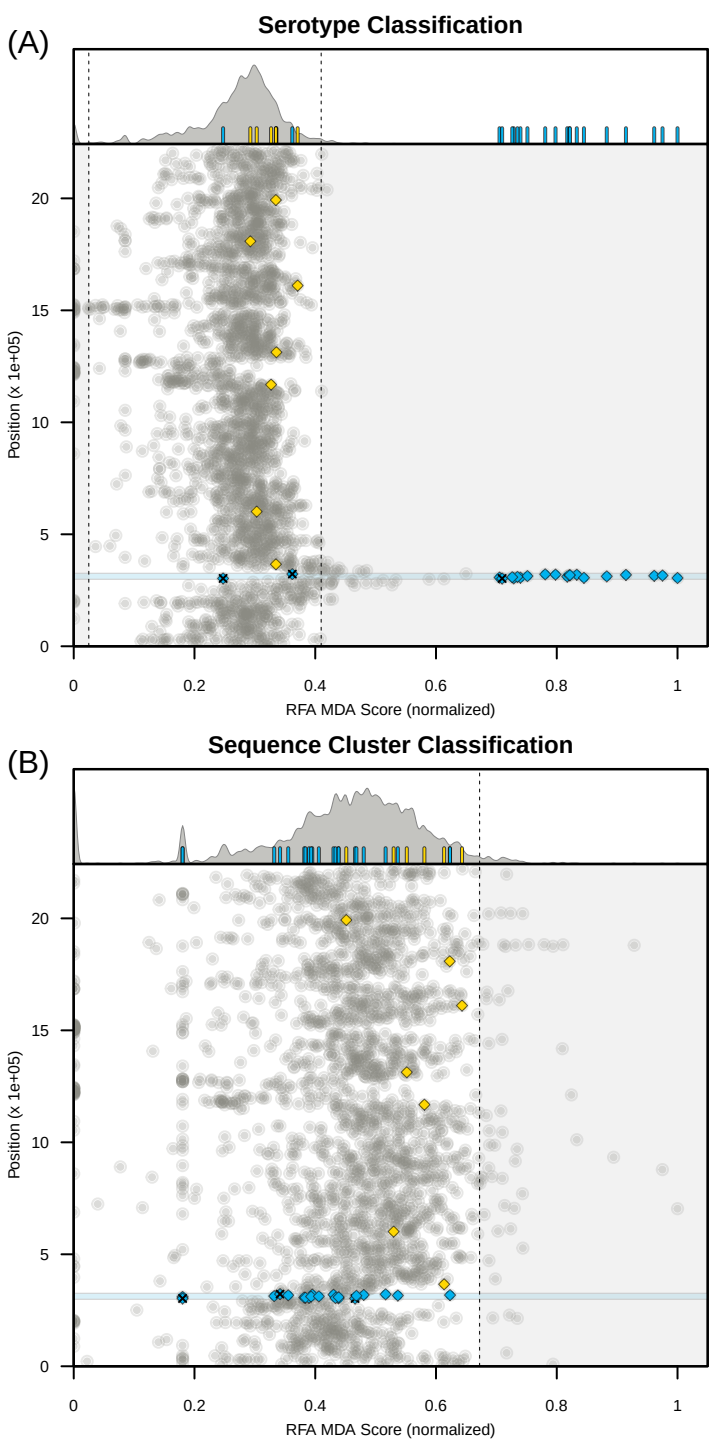
ACKNOWLEDGMENTS. The authors acknowledge the sequence data and constructive comments by Angela Brueggemann and Andries van Tonder.

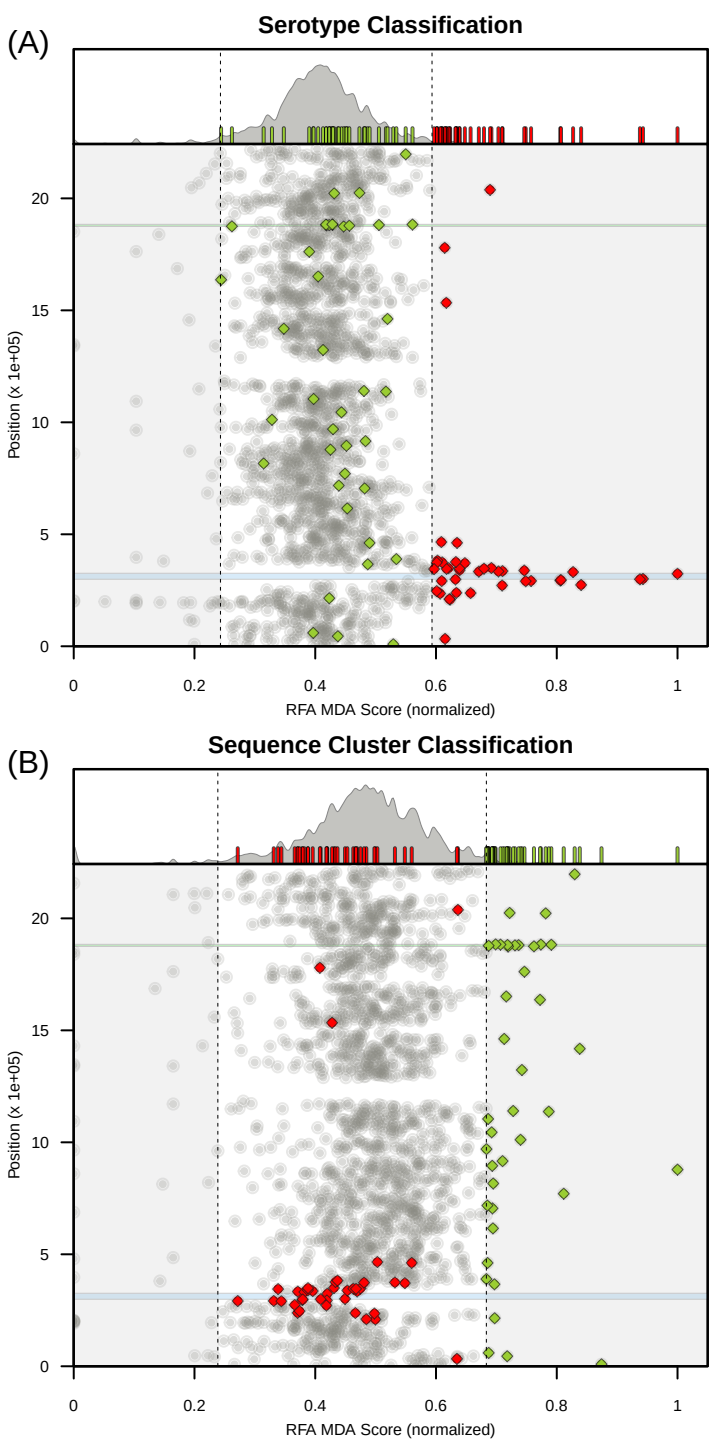
1. Maiden MC et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* 95(6):3140–5.
2. Croucher NJ et al. (2014) Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nature Communications* 5(March 2016):1–12.
3. Croucher NJ et al. (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics* 45(6):656–63.
4. Cremers AJH et al. (2015) The post-vaccine microevolution of invasive Streptococcus pneumoniae. *Scientific reports* 5:14952.
5. Henriques-Normark B, Blomberg C, Dagerhamn J, Bättig P, Normark S (2008) The rise and fall of bacterial clones: Streptococcus pneumoniae. *Nature reviews. Microbiology* 6(11):827–37.
6. Buckee CO et al. (2008) Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. *Proceedings of the National Academy of Sciences of the United States of America* 105(39):15082–7.
7. Watkins ER, Maiden MC, Gupta S (2016) Metabolic competition as a driver of bacterial population structure. *Future Microbiology* pp. fmb–2016–0079.
8. Bidossi A et al. (2012) A functional genomics approach to establish the complement of carbohydrate transporters in Streptococcus pneumoniae. *PloS one* 7(3):e33320.
9. Wen Z, Liu Y, Qu F, Zhang JR (2016) Allelic Variation of the Capsule Promoter Diversifies Encapsulation and Virulence In Streptococcus pneumoniae. *Scientific Reports* 6:30176.
10. Watkins ER et al. (2015) Vaccination Drives Changes in Metabolic and Virulence Profiles of Streptococcus pneumoniae. *PLoS pathogens* 11(7):e1005034.
11. Müller-Graf CDM et al. (1999) Population biology of Streptococcus pneumoniae isolated from oropharyngeal carriage and invasive disease. *Microbiology* 145:3283–3293.
12. Beall BW et al. (2011) Shifting genetic structure of invasive serotype 19A pneumococci in the United States. *The Journal of infectious diseases* 203(10):1360–8.
13. Metcalf BJ et al. (2016) Strain features and distributions in pneumococci from children with invasive disease before and after 13-valent conjugate vaccine implementation in the USA. *Clinical Microbiology and Infection* 22(1):60.e9–60.e29.
14. Miernyk KM et al. (2016) Population structure of invasive Streptococcus pneumoniae isolates among Alaskan children in the conjugate vaccine era, 2001 to 2013. *Diagnostic microbiology and infectious disease* 86(2):224–230.
15. Choe YJ et al. (2016) Emergence of antibiotic-resistant non-vaccine serotype pneumococci in nasopharyngeal carriage in children after the use of extended-valency pneumococcal conjugate vaccines in Korea. *Vaccine* 34:4771–4776.
16. Breiman L (2001) Random forests. *Machine Learning* 45:5–32.
17. Meng Ya, Yu Y, Cupples LA, Farrer La, Lunetta KL (2009) Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* 10(1):78.
18. Alam MT et al. (2014) Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome Biology and Evolution* 6:1174–1185.
19. Slabbinck B et al. (2010) From learning taxonomies to phylogenetic learning: Integration of 16S rRNA gene data into FAME-based bacterial classification. *BMC Bioinformatics* 11(1):69.
20. Kursa MB (2014) Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics* 15(1):8.
21. Aguas R, Ferguson NM (2013) Feature Selection Methods for Identifying Genetic Determinants of Host Species in RNA Viruses. *PLoS Computational Biology* 9(10).
22. Lupolova N, Dallman TJ, Matthews L, Bono JL, Gally DL (2016) Support vector machine applied to predict the zoonotic potential of E. coli O157 cattle isolates. *Proceedings of the National Academy of Sciences* p. 201606567.
23. Eng CLP, Tong JC, Tan TW (2014) Predicting host tropism of influenza A virus proteins using random forest. *BMC Medical Genomics* 7:S1–S1.
24. Eng CLP, Tong JC, Tan TW (2016) Distinct host tropism protein signatures to identify possible zoonotic influenza a viruses. *PLoS ONE* 11:1–12.
25. Croucher NJ et al. (2015) Population genomic datasets describing the post-vaccine evolutionary epidemiology of Streptococcus pneumoniae. *Scientific data* 2:150058.
26. Chang Q et al. (2015) Stability of the pneumococcal population structure in Massachusetts as PCV13 was introduced. *BMC infectious diseases* 15(1):68.
27. Chewapreecha C et al. (2014) Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genetics* 10(8).
28. Li Y et al. (2015) Identification of pneumococcal colonization determinants in the stringent response pathway facilitated by genomic diversity. *BMC genomics* 16:369.
29. Croucher NJ et al. (2015) Selective and Genetic Constraints on Pneumococcal Serotype Switching. *PLoS Genetics* 11(3):1–21.
30. Jürgens C et al. (2000) Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proceedings of the National Academy of Sciences of the United States of America* 97(18):9925–30.
31. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS (2003) Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiology Letters* 222(2):211–220.
32. Patel MP et al. (2005) Kinetic and chemical mechanisms of the fabG-encoded Streptococcus pneumoniae β-ketoacyl-ACP reductase. *Biochemistry* 44(50):16753–16765.
33. Kalinowski J, Bachmann B, Thierbach G, Pühler A (1990) Aspartokinase genes lysC alpha and lysC beta overlap and are adjacent to the aspartate beta-semialdehyde dehydrogenase gene asd in Corynebacterium glutamicum. *Molecular & general genetics : MGG* 224(3):317–24.
34. Buhaescu I, Izzedine H (2007) Mevalonate pathway: A review of clinical and therapeutical implications. *Clinical Biochemistry* 40(9-10):575–584.
35. Wilding EI et al. (2000) Identification , Evolution , and Essentiality of the Mevalonate Pathway for Isopentenyl Diphosphate Biosynthesis in Gram-Positive Cocci. 182(15):4319–4327.
36. Holstein SA, Hohl RJ (2004) Isoprenoids: remarkable diversity of form and function. *Lipids* 39(4):293–309.
37. van Bueren AL, Higgins M, Wang D, Burke RD, Boraston AB (2007) Identification and structural basis of binding to host lung glycogen by streptococcal virulence factors. *Nature structural & molecular biology* 14(1):76–84.
38. Abbott DW et al. (2010) The molecular basis of glycogen breakdown and transport in Streptococcus pneumoniae. *Molecular Microbiology* 77(1):183–199.
39. Garvey MI, Baylay AJ, Wong RL, Piddock LJV (2011) Overexpression of patA and patB, which encode ABC transporters, is associated with fluoroquinolone resistance in clinical isolates of Streptococcus pneumoniae. *Antimicrobial Agents and Chemotherapy* 55(1):190–196.
40. El Garch F et al. (2010) Fluoroquinolones induce the expression of patA and patB, which encode ABC efflux pumps in Streptococcus pneumoniae. *Journal of Antimicrobial Chemotherapy* 65(10):2076–2082.
41. Boncoeur E et al. (2012) PatA and PatB form a functional heterodimeric ABC multidrug efflux transporter responsible for the resistance of streptococcus pneumoniae to fluoroquinolones. *Biochemistry* 51(39):7755–7765.
42. Jomaa M et al. (2006) Immunization with the iron uptake ABC transporter proteins PiaA and PiuA prevents respiratory infection with Streptococcus pneumoniae. *Vaccine* 24(24):5133–5139.
43. Ulijasz AT, Andes DR, Glasner JD, Weisblum B (2004) Regulation of Iron Transport in Streptococcus pneumoniae by RitR , an Orphan Response Regulator Regulation of Iron Transport in Streptococcus pneumoniae by RitR , an Orphan Response Regulator. *Journal of Bacteriology* 186(23):8123–8136.
44. Graham MR et al. (2002) Virulence control in group A Streptococcus by a two-component gene regulatory system: global expression profiling and in vivo infection modeling. *Proceedings of the National Academy of Sciences of the United States of America* 99(21):13855–60.
45. Lamy MC et al. (2004) CovS/CovR of group B streptococcus: a two-component global regulatory system involved in virulence. *Molecular microbiology* 54(5):1250–68.
46. Pan X et al. (2009) The orphan response regulator CovR: a globally negative modulator of virulence in Streptococcus suis serotype 2. *Journal of bacteriology* 191(8):2601–12.
47. Jonsson IM et al. (2010) Inactivation of the Ecs ABC transporter of Staphylococcus aureus attenuates virulence by altering composition and function of bacterial wall. *PLoS ONE* 5(12).
48. Reizer J, Reizer A, Saier MH (1994) A functional superfamily of sodium/solute symporters. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes* 1197(2):133–166.
49. Chewapreecha C et al. (2014) Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics* 46(3):305–309.
50. Pillai DR et al. (2009) Genome-wide dissection of globally emergent multi-drug resistant serotype 19A Streptococcus pneumoniae. *BMC Genomics* 10:642.
51. Paik J, Kern I, Lurz R, Hakenbeck R (1999) Mutational analysis of the Streptococcus pneumoniae bimodular class A penicillin-binding proteins. *Journal of Bacteriology* 181(12):3852–3856.
52. Morlot C, Zapun A, Dideberg O, Vernet T (2003) Growth and division of Streptococcus pneumoniae: localization of the high molecular weight penicillin-binding proteins during the cell cycle. *Molecular microbiology* 50(3):845–55.
53. Pedersen LB, Setlow P (2000) Penicillin-binding protein-related factor A is required for proper chromosome segregation in Bacillus subtilis. *Journal of bacteriology* 182(6):1650–8.
54. Boyle-Vavra S, Yin S, Jo DS, Montgomery CP, Daum RS (2013) VraT/YvqF is required for methicillin resistance and activation of the VraSR regulon in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy* 57(1):83–95.
55. Knutsen E, Ween O, Håvarstein LS (2004) Two Separate Quorum-Sensing Systems Upregulate Transcription of the Same ABC Transporter in Streptococcus pneumoniae. *Journal of Bacteriology* 186:3078–3085.
56. De Saizieu A et al. (2000) Microarray-based identification of a novel Streptococcus pneumoniae regulon controlled by an autoinduced peptide. *Journal of Bacteriology* 182(17):4696–4703.
57. Reichmann P, Hakenbeck R (2000) Allelic variation in a peptide-inducible two-component system of Streptococcus pneumoniae. *FEMS microbiology letters* 190:231–236.
58. Moye ZD, Burne RA, Zeng L (2014) Uptake and metabolism of N-acetylglucosamine and glucosamine by Streptococcus mutans. *Applied and Environmental Microbiology* 80(16):5053–5067.
59. Joyce EA et al. (2004) LuxS Is Required for Persistent Pneumococcal Carriage and Expression of Virulence and Biosynthesis Genes. *Infection and Immunity* 72(5):2964–2975.
60. Xu L et al. (2006) Role of the luxS Quorum-Sensing System in Biofilm Formation and Virulence of Staphylococcus epidermidis Role of the luxS Quorum-Sensing System in Biofilm

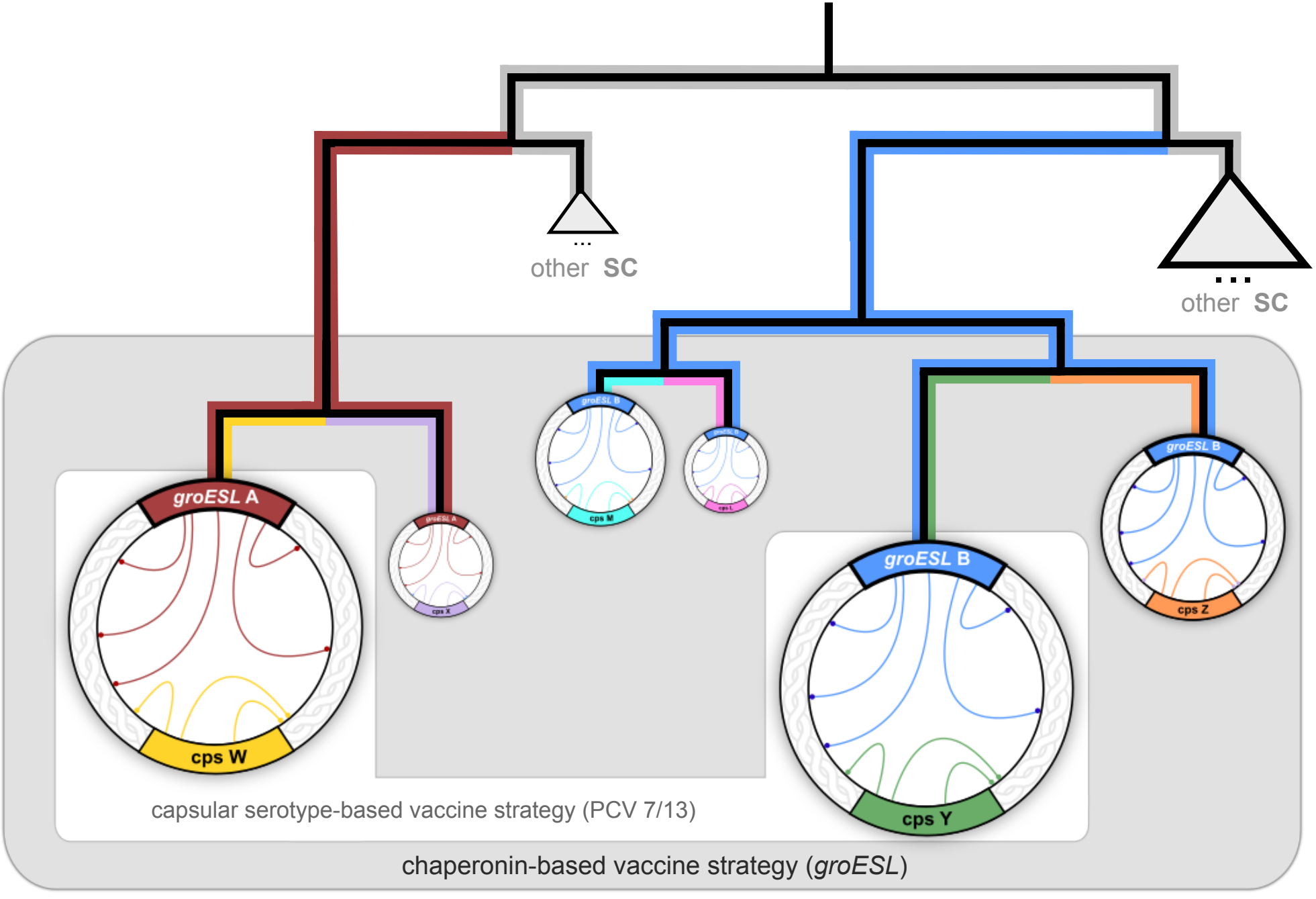
DRAFT

Formation and Virulence of Staphylococcus epidermidis. *Infection and immunity* 74(1):488–496.
61. García P, González MP, García E, García JL, López R (1999) The molecular characterization of the first autolytic lysozyme of Streptococcus pneumoniae reveals evolutionary mobile domains. *Molecular Microbiology* 33(1):128–138.
62. Eldholm V, Johnsborg O, Haugen K, Ohnstad HS, Havastein LS (2009) Fratricide in Streptococcus pneumoniae: Contributions and role of the cell wall hydrolases CbpD, LytA and LytC. *Microbiology* 155(7):2223–2234.
63. Obregón V et al. (2002) Molecular peculiarities of the lytA gene isolated from clinical pneumococcal strains that are bile insoluble. *Journal of Clinical Microbiology* 40(7):2545–2554.
64. Arbique JC et al. (2004) Accuracy of phenotypic and genotypic testing for identification of Streptococcus pneumoniae and description of Streptococcus pseudopneumoniae sp. nov. *Journal of Clinical Microbiology* 42(10):4686–4696.
65. Poyart C, Quesne G, Coulon S, Berche P, Trieu-Cuot P (1998) Identification of streptococci to species level by sequencing the gene encoding the manganese-dependent superoxide dismutase. *Journal of Clinical Microbiology* 36:41–47.
66. Martín-Galiano AJ, Balsalobre L, Fenoll A, De la Campa AG (2003) Genetic characterization of optochin-susceptible viridans group streptococci. *Antimicrobial Agents and Chemotherapy* 47:3187–3194.
67. Spellerberg B et al. (1999) Lmb, a protein with similarities to the LraI adhesin family, mediates attachment of Streptococcus agalactiae to human laminin. *Infection and immunity* 67(2):871–8.
68. Terao Y, Kawabata S, Kunitomo E, Nakagawa I, Hamada S (2002) Novel laminin-binding protein of Streptococcus pyogenes, Lbp, is involved in adhesion to epithelial cells. *Infection and Immunity* 70(2):993–997.
69. Zhang YM et al. (2014) Prevalent distribution and conservation of streptococcus suis lmb protein and its protective capacity against the chinese highly virulent strain infection. *Microbiological Research* 169:395–401.
70. Wahid RM et al. (2005) Immune response to a laminin-binding protein (Lmb) in group a streptococcal infection. *Pediatrics International* 47(2):196–202.
71. Pancholi V, Chhatwal GS (2003) Housekeeping enzymes as virulence factors for pathogens. *International journal of medical microbiology : IJMM* 293(6):391–401.
72. Gründel A, Pfeiffer M, Jacobs E, Dumke R (2016) Network of surface-displayed glycolytic enzymes in Mycoplasma pneumoniae and their interactions with human plasminogen. *Infection and Immunity* 84(3):666–676.
73. Dallo SF, Kannan TR, Blaylock MW, Baseman JB (2002) Elongation factor Tu and E1 beta subunit of pyruvate dehydrogenase complex act as fibronectin binding proteins in Mycoplasma pneumoniae. *Molecular microbiology* 46(4):1041–51.
74. Pankov R, Yamada KM (2002) Fibronectin at a glance. *Journal of cell science* 115(Pt 20):3861–3.
75. Doxey AC, McConkey BJ (2013) Prediction of molecular mimicry candidates in human pathogenic bacteria. *Virulence* 4:453–466.
76. Paterson GK, Nieminen L, Jefferies JMC, Mitchell TJ (2008) PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiology Letters* 285(2):170–176.
77. Eberhard T, Virkola R, Korhonen T, Kronvall G, Ullberg M (1998) Binding to Human Extracellular Matrix by Neisseria meningitidis. *Infection and immunity* 66(4):1791–1794.
78. Agarwal V et al. (2013) Streptococcus pneumoniae Endopeptidase O (PepO) is a multifunctional plasminogen-and fibronectin-binding protein, facilitating evasion of innate immunity and invasion of host cells. *Journal of Biological Chemistry* 288(10):6849–6863.
79. Land AD, Winkler ME (2011) The requirement for pneumococcal MreC and MreD is relieved by inactivation of the gene encoding PBP1a. *Journal of Bacteriology* 193:4166–4179.
80. Li J et al. (2016) Epigenetic Switch Driven by DNA Inversions Dictates Phase Variation in Streptococcus pneumoniae. *PLoS Pathogens* 12(7):1–36.
81. Favrot L, Blanchard JS, Vergnolle O (2016) Bacterial GCN5-Related N -Acetyltransferases: From Resistance to Regulation. *Biochemistry* 55(7):989–1002.
82. Glazunova OO, Raoult D, Roux V (2009) Partial sequence comparison of the rpoB, sodA, groEL and gyrB genes within the genus Streptococcus. *International Journal of Systematic and Evolutionary Microbiology* 59:2317–2322.
83. Teng Lj et al. (2002) groESL Sequence Determination , Phylogenetic Analysis , and Species Differentiation for Viridans Group Streptococci groESL Sequence Determination , Phylogenetic Analysis , and Species Differentiation for Viridans Group Streptococci. *Journal of Clinical Microbiology* 40:3172–3178.
84. Hung WC, Tsai JC, Hsueh PR, Chia JS, Teng LJ (2005) Species identification of mutans streptococci by groESL gene sequence. *Journal of Medical Microbiology* 54:857–862.
85. Bergé M, Mortier-Barrière I, Martin B, Claverys JP (2003) Transformation of Streptococcus pneumoniae relies on DprA- and RecA-dependent protection of incoming DNA single strands. *Molecular Microbiology* 50(2):527–536.
86. Venkatesh R et al. (2002) RecX protein abrogates ATP hydrolysis and strand exchange promoted by RecA: insights into negative regulation of homologous recombination. *Proceedings of the National Academy of Sciences of the United States of America* 99(19):12091–12096.
87. Stohl EA et al. (2003) Escherichia coli RecX inhibits RecA recombinase and coprotease activities in vitro and in vivo. *Journal of Biological Chemistry* 278(4):2278–2285.
88. Galvão CW et al. (2012) The RecX protein interacts with the RecA protein and modulates its activity in herbaspirillum seropedicae. *Brazilian Journal of Medical and Biological Research* 45(12):1127–1134.
89. Austerlitz F et al. (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC bioinformatics* 10 Suppl 1(Suppl 14):S10.
90. Hayer-Hartl M, Bracher A, Hartl FU (2016) The GroEL-GroES Chaperonin Machine: A Nano-Cage for Protein Folding.
91. Williams TA, Fares MA (2010) The effect of chaperonin buffering on protein evolution. *Genome Biology and Evolution* 2(1):609–619.
92. Kim SN, Kim SW, Pyo SN, Rhee DK (2001) Molecular cloning and characterization of groESL operon in Streptococcus pneumoniae. *Mol Cells* 11(3):360–368.
93. Cao J et al. (2013) Protection against pneumococcal infection elicited by immunization with multiple pneumococcal heat shock proteins. *Vaccine* 31(35):3564–3571.
94. Péchiné S, Hennequin C, Boursier C, Hoys S, Collignon A (2013) Immunization using GroEL decreases Clostridium difficile intestinal colonization. *PLoS ONE* 8(11).
95. Gupta S, Ferguson N, Anderson R (1998) Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science (New York, N.Y.)* 280(5365):912–915.
96. Lourenço J, Wikramaratna PS, Gupta S (2015) MANTIS: an R package that simulates multilocus models of pathogen evolution. *BMC bioinformatics* 16(1):176.
97. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403–10.
98. Friedman J, Hastie T, Tibshirani R (2001) *No Title*. (Berlin: Springer series in statistics), First edit edition.
99. Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* 2(3):18—-22.

DRAFT


(A)
Serotype Classification
Position (x 1e+05)
0
5
10
15
20
0
0.2
0.4
0.6
0.8
1
RFA MDA Score (normalized)
(B)
Sequence Cluster Classification
Position (x 1e+05)
0
5
10
15
20
0
0.2
0.4
0.6
0.8
1
RFA MDA Score (normalized)


(A)
Serotype Classification
Position (x 1e+05)
0
5
10
15
20
0
0.2
0.4
0.6
0.8
1
RFA MDA Score (normalized)
(B)
Sequence Cluster Classification
Position (x 1e+05)
0
5
10
15
20
0
0.2
0.4
0.6
0.8
1
RFA MDA Score (normalized)


...
other SC
...
other SC
groESL A
cps W
groESL A
cps X
groESL B
cps M
groESL B
cps L
groESL B
cps Y
groESL B
cps Z
capsular serotype-based vaccine strategy (PCV 7/13)
chaperonin-based vaccine strategy (groESL)