bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Analysis of recent and ancestral recombination reveals high-resolution population structure in *Streptococcus pneumoniae*

Rafal Mostowy[1], Nicholas J. Croucher[1], Cheryl P. Andam[2], Jukka Corander[3], William P. Hanage[2], Pekka Marttinen[4*]

[1]Department of Infectious Disease Epidemiology, St. Mary's Campus, Imperial College London, UK; [2]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard TH Chan School of Public Health; [3]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, Finland; [4]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland.

[*]pekka.marttinen@aalto.fi

Abstract

Bacterial population genomic analyses rely on identification of genetic recombinations, but growing databases represent a challenge for computational methods to detect these recombinations and interpret sequence ancestry. We introduce a novel algorithm called `fastGEAR` which identifies major lineages in diverse microbial alignments and recombinations between them. The algorithm can detect not only recent recombinations, but also the ancestral ones affecting entire lineages. Using simulated data, `fastGEAR` demonstrates outstanding power to detect ancestral recombination events compared to other state-of-the-art methods. The utility is further demonstrated by analysing 616 whole genomes of *Streptococcus pneumoniae*, providing novel insights into the evolution of recombinogenic bacteria.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Background

Microbial genomes are constantly subjected to a number of evolutionary processes, including mutation, gene gain and loss, genetic rearrangement and recombination, the latter here broadly defined as any form of horizontal transfer of DNA. The importance of recombination in prokaryotic evolution has been recognized for some time (Feil *et al.*, 2001) and genomic studies have become an important source of data to measure its contribution (Polz *et al.*, 2013). Comparative studies of prokaryotic genomes have found that the vast majority of their genes have been laterally transferred at least once in the past (Dagan and Martin, 2007; Dagan *et al.*, 2008), with around 20% of genes being acquired recently (Popa *et al.*, 2011). Furthermore, when measured over shorter time scales, many bacterial species were found to recombine so frequently that the impact of recombination on their genetic diversification was shown to be greater than that of mutation alone (Vos, 2009).

The prevalence of recombination is suggestive of its importance for microbial evolution, with potential adaptive benefits. Genetic exchange between different strains has been argued to play an important role in shaping of bacterial communities (Polz *et al.*, 2013; Marttinen *et al.*, 2015; Shapiro, 2016) and the emergence of new bacterial species (Fraser *et al.*, 2007, 2009; Shapiro *et al.*, 2012). Bacterial recombination has also proved a powerful adaptive weapon against major forms of clinical interventions: antibiotics and vaccines (Hanage *et al.*, 2009; Croucher *et al.*, 2011; Perron *et al.*, 2012). As most evolutionary models (e.g., phylogenetic analysis) assume no recombination, a good understanding of the impact of recombination on bacterial genomes is crucial for the correct interpretation of any genomic analysis.

Currently, popular methods used for detecting recombination include ClonalFrame (Didelot and Falush, 2007) and ClonalFrameML (Didelot and Wilson, 2015), Gubbins (Croucher *et al.*, 2014b), and BratNextGen (Marttinen *et al.*, 2012). The three former approaches follow the line of methods based on phylogenetic trees (Husmeier, 2005; Minin *et al.*, 2005; Webb *et al.*, 2009), and look for clusters of polymorphisms on each branch of a phylogenetic tree. On the other hand, BratNextGen uses Hidden Markov Models (HMMs) to model the origin of changes in the alignment, where clonality (lack of recombinations) represents one origin and other origins represent foreign recombinations. All these methods specialise in identifying imports originating in external sources, and are therefore appropriately applied to a single bacterial lineage at a time. Thus, they rely on another method to identify the underlying population structure, which limits their ability to provide insight into species-wide patterns of exchange. With the recent development of high-throughput sequencing methods which can process tens of thousands of bacterial whole-genomes, such analyses have become increasingly interesting and necessary.

Here we present an approach to fulfill such a demand, which identifies both the population structure of a sequence alignment and detects recombinations between the inferred lineages as well as from external origins. The method locates both recent as well as ancestral recombination events affecting entire lineage, as

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

shown in Figure 1. Our approach is similar to the popular STRUCTURE software with the linkage model (Falush *et al.*, 2003) but with the following crucial differences: (a) it is computationally scalable to thousands of bacterial genomes, (b) it provides insight into the mosaicism of entire bacterial populations by inferring also the ancestral recombination events. As our method can quickly infer the *ge*nomic *ar*rangement of large bacterial datasets, we called it `fastGEAR`.

We assess the accuracy of `fastGEAR` using extensive simulations, and compare it to three other state-of-the-art methods. We then use it to analyse a dataset of 616 whole-genomes of a recombinogenic pathogen *Streptococcus pneumoniae* sampled in Massachusetts, USA, in a paediatric carriage study (Croucher *et al.*, 2013). The results not only efficiently reproduce many previous findings, but also provide some novel insights into the evolution of *S. pneumoniae* as a species.

Results

Our approach is based on a Hidden Markov Model (HMM) of nucleotide frequencies at polymorphic sites of the alignment analysed (see Figure 2). In brief, we first identify lineages in the alignment, and in these lineages we infer 'recent recombinations' as foreign genetic fragments present in a subset of strains in the lineage. We then identify 'ancestral recombinations' as foreign genetic fragments present in all strains in the lineage. Finally, we test for significance of both recent and ancestral recombinations. The comprehensive description of the approach is given in Methods and details in text S1.

Performance on simulated data

To give an example of `fastGEAR` performance, we performed coalescent simulations involving $P = 3$ lineages with a given effective population size for each lineage N_e and the most recent common ancestor (MRCA) time T. We then simulated recombination events between them in three different modes: recent, intermediate and ancestral. Finally, we compared the resulting, true population structure with the one inferred by `fastGEAR` in Figure 3. We see that `fastGEAR` not only correctly identifies the lineages, but also finds all recombinations. The inference of recent recombinations is generally better than the inference of ancestral recombinations. Of particular importance is that `fastGEAR` does much better at predicting the direction of recombination events for recent recombinations. Such direction is difficult, if not impossible, to determine for older, ancestral recombinations. However, we see that the population genetic structure is correctly inferred in all three examples, even in the difficult case of multiple, overlapping recombinations occurring at different time scales.

To systematically assess performance of `fastGEAR` we performed two different sets of *in silico* experiments. First, we examined how well `fastGEAR` detects recombinations for different population parameters. To this end, we varied the within-population distance (achieved by changing N_e) and the between-

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

population distance (achieved by changing T). The results are shown in Figure S1. We see that `fastGEAR` generally detects recombinations well, particularly the recent ones as they share higher resemblance to the origin and are thus by definition easier to detect. The false-positive rate was low for all types of recombinations detected and did not vary with the between- or the within-lineage distance. By contrast, we observed that the proportion of detected recombinations was highly dependent on the between-lineage distance. This is because in the absence of clear population genetic structure, populations are relatively closely related and there are too few polymorphisms to signal the presence of a recombination. Furthermore, a higher within-lineage distance often affected the inference of ancestral recombinations as it generated the intra-lineage population genetic structure. Thus, as expected, performance of `fastGEAR` depends on the underlying population genetic structure.

In the second set of experiments, we compared the performance of `fastGEAR` to other recombination-detection methods: STRUCTURE, Gubbins and ClonalFrameML. To compare with the former two phylogeny-based approaches that are designed to be used in a lineage-by-lineage manner, we ran `fastGEAR` both using entire alignment and on each lineage individually. The results are shown in Figure 4. First, we saw a comparable performance of `fastGEAR` to STRUCTURE in detecting recent and intermediate recombinations. The two methods are generally quite similar in their approach, however STRUCTURE explores the entire parameter space using MCMC whereas `fastGEAR` uses the most probable clustering and point estimates for hyperparameters. It is therefore not surprising that in all the examined cases STRUCTURE slightly outperformed `fastGEAR` in the proportion of detected recent and intermediate recombinations (typically by around 10%-20%) but at a much higher computational cost (between 1.5-2 orders of magnitude; see Figure S2). However, in reality STRUCTURE may not always perform better as it was here run knowing the true value of the number of populations K and the true lineage memberships of different strains as priors, while `fastGEAR` had no such knowledge. We thus conclude that the performance of `fastGEAR` to detect recent recombinations, when run with entire alignment, was comparable to that of STRUCTURE but in a fraction of CPU time.

We found the performance of `fastGEAR` run in a lineage-by-lineage manner to be lower than when using the full alignment. This is because `fastGEAR`, similarly to STRUCTURE, gains its statistical power from having actual origins of recombinations. In spite of the decreased power, the lineage-by-lineage results were comparable to Gubbins, which in addition uses the information about the ancestry of sequences in the alignment and exploits it when detecting recombinations. Conversely, the performance of the lineage-by-lineage runs of `fastGEAR` was clearly lower than the one of ClonalFrameML, but comparable when run on the entire alignment. However, the ClonalFrameML analysis was conditioned on the true phylogeny. The difference in detection power between Gubbins and ClonalFrameML could stem from the former using a more conservative multiple-testing correction in recombination-detection.

Importantly, throughout simulations `fastGEAR` detected ancestral recombinations equally well to recent and intermediate recombinations. This is par-

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ticularly encouraging as none of the other methods could detect those events. The false detection rate of ancestral recombinations was generally zero with the exception of very distant lineages, and even then it was only 0.3 recombinations per alignment on average. In spite of such an encouraging result, caution is warranted when interpreting the results, in which we recommend both visual inspection of Bayes factors and the good understanding of the used dataset (see also Discussion).

Analysis of *Streptococcus pneumoniae* data

We next applied `fastGEAR` to a whole-genome collection of 616 isolates of *S. pneumoniae* to infer the bacterial population structure and analyse the distribution of recombinations, both recent and ancestral, across the genome. To this end, we analysed the COGs independently (see Methods). The goal of the analysis was to gain a detailed view of the relationships between the strains in the data set, and ultimately to better understand the impact of recombination on pneumococcal evolution.

High-resolution view of population structure

To investigate the population structure of the entire collection of isolates, we calculated the proportion of shared ancestry (PSA) matrix, which summarizes the `fastGEAR` results for all 2,113 COG alignments. Specifically, for each pair of isolates we analysed the population structure at each COG with putative lineages and the list of recent and ancestral recombinations detected in the COG. If the COG was present in both isolates, we computed the proportion of the length of the COG sequence in which the two isolates were assigned to the same lineage. If the COG was absent in either or both of the isolates, they were not compared at this COG; if multiple copies of the gene were found, then all possible comparisons between the two isolates were included, and correspondingly taken into account in the total length of sequence compared.

The resulting PSA matrix together with a previously published core-gene-based phylogeny and 15 monophyletic sequence clusters (SCs), which can be taken as lineages, is shown in Figure 5. Overall, the PSA results are highly concordant with the tree and the SCs. First, strains within SCs share almost all of their ancestry, such that the average PSA within different SCs ranges from 85% up to 98%, which is visible as blocks of high PSA on the diagonal. Second, these blocks correspond well to the clades of the phylogeny. Third, the sequence cluster SC12, which has previously been identified as 'atypical', non-encapsulated pneumococci (Croucher *et al.*, 2014a) and appears distant from the rest of the population in the phylogeny, shares considerably less of its ancestry (approximately 60%) than other SCs share with each other. We also note that the polyphyletic SC16, which includes all strains in the phylogeny which are not part of SCs 1-15 (and is this not shown), consists of multiple blocks of high PSA. These individual groups are similar to other SCs, with the difference that they just are too small to be identified as separate SCs. Thus, we see

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

that `fastGEAR` can produce a high-resolution view of the bacterial population genomic structure. Even though the PSA matrix and the phylogeny are in good accordance, our results highlight some details of the population structure not apparent in the phylogeny; for example, a pair of isolates between SC5 and SC8 in Fig. 5 that seem to share a large proportion of their ancestry with SC8.

A conspicuous feature of the PSA matrix is the lack of hierarchy between different SCs. Indeed, the different lineages (except for SC12) are approximately equidistant from each other, sharing from 71% to 81% of their ancestry, a pattern that can be explained by frequent recombination between the SCs (Fraser *et al.*, 2007; Marttinen *et al.*, 2015). To better demonstrate this, we computed the amount of private ancestry for each strain, defined as the proportion of the strain where the origin was not found in any other SC than the one to which the strain belonged (Figure S3). The results show that all SCs have very little private ancestry; even the divergent SC12 has only about 15% of its ancestry private, i.e., not found in any other SC. These findings are consistent with the analysis of accessory genome content, which hypothesised that SC12 pneumococci may constitute a different streptococcal species altogether (Croucher *et al.*, 2014a).

To investigate the impact of recombination on the core genome further, we analysed the population structure of 96 housekeeping genes from an extended MLST set (Crisafulli *et al.*, 2013). The results for all the 96 genes are shown in Figure S4, for a subset of 25 genes in Figure 5, and for all core genes (i.e., present in at least 95% of the isolates) in Figure S5. Two observations are particularly striking. First, we see that for the vast majority of genes the inferred number of lineages is much smaller than 15 (median: 3, 95% quantile: 2-6). Second, the population structure is highly variable across the genome including at the 96 most essential genes, significantly deviating from a clonal model of diversification (Figures S6 and S7). These findings lead to two important conclusions: (i) the pneumococcal-wide population structure, as represented by the SCs, emerges as the average of highly variable population structures of individual genes, and (ii) variable population structures of individual genes reflects their different evolutionary histories, and thus imply high rates of recombination at almost all bacterial genes, even the most conserved ones.

Population structure of clinically relevant genes

We next investigated the population structure of clinically relevant genes, including three penicillin-binding proteins, which determine resistance to beta-lactam antibiotics (*pbp1a*, *pbp2b* and *pbp2x*), and common genes at the vaccine-targeted capsule biosynthesis locus (*dexB*, *wzg*, *wzh*, *wzd*, *wze*, located upstream the *cps* locus, and *rmlA*, *rmlC*, *rmlB*, *rmlD*, located downstream the *cps* locus). Interestingly, as *pbp1a* and *pbp2x* are located in the vicinity of the capsular locus, we investigated the relationship between the population structure of these two protein groups.

The results are shown in Figure 6. As expected, we found a very strong association between the population structure of *wz*-genes (present in all serotypes), the rhamnose synthesis operon genes (present in about half of serotypes) and the

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

actual serotypes. Conversely, we saw no association between these three groups and the *dexB* gene. This likely reflects the fact that recombination events driving serotype switching often span the *wz*- and *rml*-genes but not *dexB*, and that the former two groups are likely to be horizontally transferred together. A probable explanation for this is the selective pressure to maintain the serogroup as recombination breakpoints in the middle of the locus would likely disrupt the genetic content of the locus (Croucher *et al.*, 2015). Interestingly, our analysis pointed to *rmlA* gene as one of recombination hotspots (see also next subsection) but no such hotspots were found in within the *wz*-genes. One possible explanation is that there may be a possible stronger epistatic interaction between *wz*-genes and the capsule-determining genes compared to between *rml*-genes and the capsule-determining genes.

Furthermore, we saw a strong association between the population structure of capsular genes and the bacterial lineages (sequence clusters or SCs), and this is expected due to a known strong association between pneumococcal lineages and serotypes. Within-lineage serotype variation was well reflected in changes in the capsular population structure, thereby well reflecting the historical serotype-switching events. These changes were consistent with a previous lineage-by-lineage phylogenetic analysis of such events (Croucher *et al.*, 2015). Additionally, our analysis revealed other serotype-switching recombinations outside the 15 monophyletic clusters.

The population structure of the three β-lactam genes was very strongly associated, with mosaicism in one gene being a good predictor of the mosaicism in the other two genes. This confirms previous findings which suggested a strong epistatic interaction of these genes (Croucher *et al.*, 2013). To further investigate the association between mosaicism and resistance at those genes, we applied a simple linear model and examined how well increased levels of MIC can be explained by the number of recent and ancestral recombinations in β-lactam genes. We found a strong positive correlation between ancestral and recent recombinations in both *pbp2b* ($p = 4.05 \times 10^{-12}$ for recent recombinations; $p < 2.2 \times 10^{-16}$ for ancestral recombinations) and *pbp1a* ($p < 2.2 \times 10^{-16}$ for recent recombinations; $p < 2.2 \times 10^{-16}$ for ancestral recombinations) with elevated MICs (see Figure S8). However, such associations were not seen for *pbp2x* ($p = 0.59$ for recent recombinations; ancestral recombinations associated with a significant tendency to lower MICs), which `fastGEAR` finds to be the most extensively modified of these genes. This relationship is likely due to earlier findings that mosaicism of *pbp2x* gene does not result in decreased susceptibility to penicillin of strains carrying the mosaic allele (Dowson *et al.*, 1994).

Comparison of recombination levels across different proteins

We next compared the levels of recent and ancestral recombination between different proteins. Consistent with a constant rate of recombination over the history of this population, both measures were significantly correlated ($R^2 = 0.46$, $p < 2.2 \times 10^{-16}$) with the mean number of recent recombinations almost twice the number of ancestral recombinations (1.4 vs. 2.7). However, in 20% of genes

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

we found a greater number of ancestral recombinations which could reflect a complex relationship between recombination rate and selection (see Figure S9). Among the genes with the highest number of ancestral and recent recombinations (Tables 1 and 2 respectively) we found many loci previously identified as recombination hotspots. These proteins can be classified into several groups.

The first group are mobile genetic elements, which include integrative and conjugative elements, prophages, phage-related chromosomal islands and insertion sequences. This is not surprising as frequent between- and within-lineage recombination of mobile genetic elements has reported previously (Croucher *et al.*, 2014a). The second group are proteins which are engaged in the interactions with the host. Pneumococcal surface protein C (*pspC*), which plays a central role in pathogenesis of the pneumococcus, was a top hit for both recent and ancestral recombinations (Kadioglu *et al.*, 2008). Another top hit, as discussed above, was the first of the rhamnose genes (*rmlA*; cf. Fig. 6), which often serves as a breakpoint in serotype switching events. We also found a high number of recent recombinations in the zinc metalloprotease *zmpA*, which cleaves human immunoglobulin A1 (Weiser *et al.*, 2003). The third group are genes involved in determining resistance to antibiotics, including sulphamethoxazole resistance (*folC*), as well as β-lactams (*pbp1a*, *pbp2b* and *pbp2x*) discussed above (see also Figure S10). With the exception of the mosaic *zmpA* sequences, these proteins were previously identified as recombination hotspots in globally disseminated lineages (Croucher *et al.*, 2011, 2014c).

We also found highly recombinogenic proteins which have not been previously identified as recombination hotspots. One example is the chromosome partitioning SMC protein, which functions in chromosomal segregation during cell division (Britton *et al.*, 1998) and is one of the top hits in both recent and ancestral recombinations. Other genes that were also inferred to undergo high levels of recombination are phenylalanyl- and valyl-tRNA synthetases, enzymes that attach the amino acids phenylalanine and valine to their cognate tRNA molecules during the translation process. Previous reports show that recombination and horizontal gene transfer frequently occur in aminoacyl-tRNA synthetases (*aaRS*) (Woese *et al.*, 2000). The horizontal acquisition of *aaRS* variants may be implicated in resistance to antibiotics (Woese *et al.*, 2000), and at least one atypical additional *aaRS* has been found on Pneumococcal Pathogenicity Island 1 (Croucher *et al.*, 2009). Although within-species recombination of *aaRS* has not been widely investigated, our results suggest that this process plays an important role in the evolution of pneumococci.

Discussion

In this article we introduced a novel tool called `fastGEAR` to analyse the population genetic structure in bacteria. Specifically, `fastGEAR` identifies major lineages and infers recombination events between them as well as those originating from outside the sample population. Simultaneous inference of the population structure and between-population recombinations using Hidden Markov Models

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

is analogous to an earlier approach called STRUCTURE (Falush *et al.*, 2003) but is novel in terms of both the ability to infer ancestral exchanges between those populations and computational scalability. When tested on simulated data, `fastGEAR` demonstrated a comparable accuracy to STRUCTURE but in a fraction of CPU time (cf. Figure S2). Unlike ChromoPainter/fineStructure (Lawson *et al.*, 2012) – another related method that investigates the similarity of human haplotypes using other haplotypes as possible origins for the target haplotype – `fastGEAR` is designed for detecting recombination between groups of sequences. Thus, our method is a notable addition to the currently available approaches for recombination detection in bacterial genomes, particularly so due to its ability to cope with increasingly large collections of whole-genome data.

Event though `fastGEAR` detected ancestral recombinations exceptionally well in simulated data, a few points should be kept in mind when interpreting the results. First, the term 'ancestral' is relative and does not have to reflect the time of recombination; it merely reflects the fact that the recombination happened before the strains in the affected lineage diverged. In fact, ancestral recombinations which occurred recently will be easier to detect than ancestral recombinations which occurred a long time ago. This is because our method, broadly speaking, is opposite to the clonal-frame-like approaches (Didelot and Falush, 2007; Croucher *et al.*, 2014b; Didelot and Wilson, 2015), where recombinations are divergent segments among highly similar sequences; here the ancestral recombinations are highly similar segments between diverse lineages. Second, it is important to emphasise that `fastGEAR` cannot reliably infer the direction of ancestral recombinations because this would require additional assumptions about relationships between the ancestral sequences. In the results presented we have resolved this issue by always marking the lineage with fewer strains as recombinant, assuming that lineage sizes are indicative of their ages, but the potential of sampling bias should be considered when interpreting results.

Although our method does not assume a phylogeny, it nevertheless relies on some lineages between which recombinations are detected, and inferring the lineages is an important first step on which reliable downstream analysis can be based. Assuming consistent lineages over the length of one gene is more justified than over larger genomic regions, as demonstrated also by our results. For this reason we chose to analyse the *S. pneumoniae* data gene-by-gene, rather than concatenating multiple genes for joint analysis. The gene-by-gene analysis has additional benefits of being straightforward to parallelize and possible to apply to whole-genome core alignments. The latter is particularly appealing as it permits insight into the population structure and evolution of diverse microbial datasets. For these reasons, this is the way we currently recommend to use the method in practice. One downside of the gene-by-gene analysis is that there is no straightforward way for making inferences about long recombinations spanning multiple genes.

Our statistical approach combines HMMs to identify putative recombinations with a post-processing step to compute the significances of the recombinations. These steps use information in the sequence data differently: the HMMs are based on allele frequencies at polymorphic sites, whereas the significances are

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

computed using variations in SNP frequency along the sequence. The need for a separate post-processing step follows from the limitation of the HMMs that they can only tell whether two lineages are the same or different, but not how different they are. Consequently, very close or distant lineages are easily handled by the HMMs, but there always seems to be some intermediate distance for which HMMs may produce short segments of false positive recombinations, regardless of the exact way the HMM is formulated (for example, we experimented with various ways to handle the hyperparameters). The post-processing will produce a bias towards removing short, diverged segments as longer ancestral recombinations often reach higher significance. This is useful from a biological point of view because such sort segments may also emerge as a combined result of mutation and positive selection. By assigning higher significance to longer fragments the chance of those fragments representing horizontal and not vertical evolution is increased. Nevertheless, a visual check of significance and a good understanding of the data analysed is highly recommended.

The usefulness of `fastGEAR` became evident when we applied it on a large collection of whole-genomes of *Streptococcus pneumoniae* from Massachusetts, by analysing the population structure of all individual genes present in at least fifty isolates (COGs). To give an example of computational efficiency of the algorithm, we recorded the runtime for the three penicillin-binding proteins (*pbp1a, pbp2x, pbp2b*), which had particularly complex population structures. On a 2.3 Ghz laptop, these were: 329, 308 and 234 seconds, respectively, reflecting the approximate linear scaling of the runtime with respect to the number of polymorphic sites (823, 720, 558 SNPs in the three datasets, respectively). For the majority of genes the run time was around two minutes or less. The analysis of all COGs produced for the first time a high-resolution view of the species-wide population structure. The population structure was consistent with previous studies of the fifteen major monophyletic groups but it also permitted insight into the ancestral composition of smaller clusters as well as the relationships between the clusters. Analysis of recombinations within individual genes not only correctly identified many known major recombination hotspots in the pneumococcus but also pointed to potentially novel ones (SMC protein, *valS*, *aaRS*).

Furthermore, our method also provided insight into clinically-driven evolution of *S. pneumoniae*. First, we found a strong relationship between individual recombination events (both recent and ancestral) and adaptation to clinical forms of intervention, including resistance to β-lactam antibiotics and serotype switching at vaccine-targeted capsular locus. Second, and perhaps surprisingly, `fastGEAR` did not find evidence for strong linkage between the capsular and neighboring genes, including those that confer penicillin resistance. There are reports of simultaneous transfer of these clinically important regions both from natural populations (Brueggemann *et al.*, 2007; Coffey *et al.*, 1999) and *in vitro* studies (Trzciński *et al.*, 2004). Our observations might indicate divergent selective pressures exerted by antibiotics in the community, and host immunity. However it is known that the prevalence of penicillin is higher in some serotypes than others, so our results suggest that in the majority of cases this is con-

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

vergent evolution rather than a single event. All of these approaches demonstrate the potential of `fastGEAR` to efficiently provide insight into the important, epidemiologically-relevant changes in the population structure.

Interestingly, our analysis also provided some fundamental understanding of the evolution of *S. pneumoniae*. The analysis of the population structure of the core genome itself demonstrated a highly variable and mosaic structure of individual core genes, including 96 essential housekeeping genes. This shows the scale of genome-wide recombination of the pneumococcus and highlights its major role in the evolution of bacterial lineages. These results also indicate an important limitation of phylogenetic approaches to studying the evolution of the pneumococcus and likely other highly recombinogenic bacteria: while core genome based trees are likely to efficiently reproduce within-lineage ancestry patterns, their deeper branches may not well represent the between-lineage relationships at many individual genes, but rather emerge as an average over various population structures at different loci.

While developed and tested with bacterial genomes in mind, there is nothing in the method *per se* to exclude it from the analysis of other pathogens, including viruses. Nevertheless, `fastGEAR` assumes the isolates to be haploid, for which reason we expect `fastGEAR` to be particularly useful in questions related to microbial evolution.

Conclusions

`fastGEAR` offers a novel approach to simultaneously infer the population structure and recombinations (both recent and ancestral) between lineages of diverse microbial populations. We expect the method will bring novel insight into the evolution of recombinogenic microbial species, particularly so when recombination rates are high enough for the species concept to be challenging to define.

Methods

Overview of the algorithm

Here we give a general high-level description of the method, and the details are presented in supplementary text S1. The algorithm takes as input an alignment of bacterial DNA sequences and performs the following four tasks:

1. Identify lineages in the alignment.

2. Identify recent recombinations in the lineages, where the 'recent recombinations' are defined as those that are present in a subset of strains in a lineage.

3. Identify ancestral recombinations in the lineages, where the 'ancestral recombinations' are defined as those that are present in all strains that belong to the lineage.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4. Test of significance of the putative recombinations.

The distinction between recent and ancestral recombinations is whether the recombination event happened before or after the most recent common ancestor of the lineage in which it was detected (Fig. 1).

(1) Identifying lineages To identify lineages in a data set, we start by running a previously published clustering algorithm (Corander and Marttinen, 2006) included in the Bayesian Analysis of Population Structure (BAPS) software (Corander *et al.*, 2003). This produces C strain clusters which represent population structure among the strains. However, the clusters as such are not optimal for recombination analysis for two reasons. First, the algorithm may assign two otherwise identical sets of sequences into distinct clusters due to one of the sets experiencing a recombination event, resulting in a poor representation of the overall population structure. Second, the algorithm may detect clusters that have diverged very recently, and the closeness of such clusters may result in added noise in recombination detection. Due to these reasons, given the clustering pattern, we infer lineages using a Hidden Markov Model (HMM) approach (Fig. 2A). In more detail, we compare allele frequencies for each cluster pair using a HMM, where the hidden states of the HMM represent equality of the frequencies at the polymorphic sites. All pairwise comparisons are summarised as a distance matrix, which tells the proportion of the length of the sequences where two clusters are considered different. We apply the standard complete linkage clustering with cutoff 0.5 to this distance matrix, resulting in a grouping of clusters into L groups, which are taken as lineages in the data. This means that two clusters will usually be considered as part of the same lineage if their sequences are considered similar for at least 50% of the sequence length, although this is not strictly enforced by the complete linkage algorithm.

(2) Detecting recent recombinations To identify recent recombinations, we analyse each lineage by applying a HMM approach for strains assigned to the lineage, this time with hidden states representing the origins of the different polymorphic sites in the strain (Fig. 2B). Possible origins are the other lineages detected in the data, as well as an unknown origin, not represented by any strain in the data. The positions which are assigned to a different origin than the identified lineage of the strain are considered recombinations. After analysing all strains in the lineage, the hyperparameters of the HMM are updated. Further iterations of detecting recombinations and updating hyperparameters are carried out until approximate convergence. The final reported recent recombinations are those sequence positions where the probability of the assigned lineage is less than some threshold, where we have used a conservative threshold value equal to 0.05. If a sequence position is considered recombinant, then the origin is set to be the lineage with the highest probability at this position. We note that also the full probability distributions are available from our implementation.

(3) Detecting ancestral recombinations To identify ancestral recombinations, we analyse all lineage pairs using the same approach as in step (1), such that the latent variables for the different sequence positions have two possible states, either the lineages are the same or different, with the recent recombi-

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

nation sequences treated as missing data. Putative ancestral recombinations between lineages correspond to regions of the alignment where the inferred lineages are the same - hence a portion of the genome in isolates that are overall assigned to different lineages, may be considered to be part of the same lineage. However, it is important to note that the direction of a recombination can not be identified using this approach. To resolve this issue, we always mark the lineage with fewer strains as the recombination recipient in our results. The convention may be justified by the principle of maximum parsimony, as it results in fewer strains in the data set carrying a recombinant segment (but see Discussion).

(4) Test of significance The HMMs produce probabilities for sequence positions of having their origins in the different lineages, which can be used as a measure of statistical strength of the findings. However, in our experiments we encountered two kinds of false positive findings: first, recent recombinations in strains that were outliers in the data set; second, ancestral recombinations between lineages that were diverged to the verge of not being considered the same by the HMM, but not completely different either (see Discussion on the limitations of the HMMs). To prune these false positive findings, we monitor the locations of SNPs between the target strain and its ancestral lineage (for recent recombinations) or between the two lineages (for ancestral recombinations) within and between the claimed recombinant segments. We apply a simple binomial test to compute a Bayes factor (BF; (Bernardo and Smith, 2001, see, e.g.)), that measures how strongly the changes in SNP density support a recombination, and we use a threshold BF=1 for recent recombinations and BF=10 for ancestral recombinations for additional pruning of recombinations proposed by the HMM analyses. These thresholds represent a compromise between false positive rate and power to detect recombinations. Recombinations with the BF less than the threshold are not reported at all, and the estimated BFs for the remaining recombinations are included in the output.

Simulations

Details of simulations are given in supplementary text S1. In brief, to generate *in silico* data we first created a phylogeny using a coalescent simulation framework (Excoffier *et al.*, 2013) assuming $P = 3$ demes which diverged T generations ago, each with a clonal population of effective size N_e and with mutation rate μ. A sample of n isolates was drawn from each population. An alignment of length L was created conditional on the phylogeny and recombinations were simulated by donating a homologous DNA fragment from a prespecified donor population to the target population, after which the fragment evolved according to the phylogeny of the target population. Recent recombinations were assumed to occur on average several generations before the present; intermediate recombinations were assumed to occur sometime between present and the youngest of all P most recent common ancestors for each population; ancestral recombinations were assumed to occur before the oldest of all P most recent common ancestors for each population. The recombination size was modelled as a geometrically distributed variable with Γ_r being the mean size of recent and intermediate re-

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

combinations and Γ_a the mean size of ancestral recombinations. We assumed on average R_r recombination events per population for recent recombinations (with targets chosen randomly), R_i recombination events per population for intermediate recombinations and R_a recombination events in total for ancestral recombinations.

The accuracy of `fastGEAR` was assessed by quantifying the number of wrong recombinations (false-positives) and missed recombinations (false-negatives). To account for non-independence of recent and ancestral recombinations affecting multiple isolates, we clustered similar recombinations together with 95% identity threshold and counted each cluster as a single event. Inferred recombinations were then compared to true recombinations by comparing the isolates in which they occurred and position at which they occurred (assuming any overlap), which determined the number of false-positives and the proportion of all recombinations detected. Due to the difficulties in identifying direction of recombination, a detected recombination was considered a true-positive if the resulting population structure was correct, even if the recipient was not identified correctly.

Data from *Streptococcus pneumoniae*

We analysed a collection of 616 *Streptococcus pneumoniae* genome strains sampled in Massachusetts, for which whole-genome sequences were described in the original publication (Croucher *et al.*, 2013). The assembled data were scanned for putative protein-coding sequences, which were grouped according to their similarity, resulting in 5,994 clusters of orthologous genes (COGs). From these, we selected into our analysis those with at least 50 sequences, and we only included proteins that were within 75% and 125% of the median length of the COG. After this filtering, we kept COGs with at least five distinct protein sequences included in the data set, resulting in a total of 2,113 COGs included into our analysis. Unique sequences were aligned with Muscle (Edgar, 2004). All DNA sequences associated with each protein sequence were then back-translated into a full codon alignment. A core alignment was constructed using COGs present once in each genome assembly. This alignment was previously used to produce a maximum likelihood phylogeny of the data, and analysed by BAPS to produce 16 sequence clusters (SCs), of which 15 were monophyletic (Croucher *et al.*, 2013).

References

Bernardo, J. M. and Smith, A. F. 2001. *Bayesian theory*. IOP Publishing.

Britton, R. A., Lin, D. C.-H., and Grossman, A. D. 1998. Characterization of a prokaryotic smc protein involved in chromosome partitioning. *Genes & development*, 12(9): 1254–1259.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Brueggemann, A. B., Pai, R., Crook, D. W., and Beall, B. 2007. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathogens*, 3(11): e168.

Coffey, T. J., Daniels, M., Enright, M. C., and Spratt, B. G. 1999. Serotype 14 variants of the Spanish penicillin-resistant serotype 9V clone of Streptococcus pneumoniae arose by large recombinational replacements of the cpsA-pbp1a region. *Microbiology*, 145(8): 2023–2031.

Corander, J. and Marttinen, P. 2006. Bayesian identification of admixture events using multilocus molecular markers. *Molecular ecology*, 15(10): 2833–2843.

Corander, J., Waldmann, P., and Sillanpaa, M. J. 2003. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1): 367–374.

Crisafulli, G., Guidotti, S., Muzzi, A., Torricelli, G., Moschioni, M., Masignani, V., Censini, S., and Donati, C. 2013. An extended multi-locus molecular typing schema for Streptococcus pneumoniae demonstrates that a limited number of capsular switch events is responsible for serotype heterogeneity of closely related strains from different countries. *Infect. Genet. Evol.*, 13: 151–161.

Croucher, N. J., Walker, D., Romero, P., Lennard, N., Paterson, G. K., Bason, N. C., Mitchell, A. M., Quail, M. A., Andrew, P. W., Parkhill, J., *et al.* 2009. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniae Spain23F ST81. *Journal of bacteriology*, 191(5): 1480–1489.

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J. H., Ko, K. S., *et al.* 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science*, 331(6016): 430–434.

Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., Bentley, S. D., Hanage, W. P., and Lipsitch, M. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics*, 45(6): 656–663.

Croucher, N. J., Coupland, P. G., Stevenson, A. E., Callendrello, A., Bentley, S. D., and Hanage, W. P. 2014a. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nature Communications*, 5.

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., and Harris, S. R. 2014b. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic acids research*, page gku1196.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Croucher, N. J., Hanage, W. P., Harris, S. R., McGee, L., van der Linden, M., de Lencastre, H., Sá-Leão, R., Song, J.-H., Ko, K. S., Beall, B., *et al.* 2014c. Variable recombination dynamics during the emergence, transmission and disarmingof a multidrug-resistant pneumococcal clone. *BMC biology*, 12(1): 49.

Croucher, N. J., Kagedan, L., Thompson, C. M., Parkhill, J., Bentley, S. D., Finkelstein, J. A., Lipsitch, M., and Hanage, W. P. 2015. Selective and genetic constraints on pneumococcal serotype switching. *PLoS Genet*, 11(3): e1005095.

Dagan, T. and Martin, W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 104(3): 870–875.

Dagan, T., Artzy-Randrup, Y., and Martin, W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 105(29): 10039–10044.

Didelot, X. and Falush, D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics*, 175(3): 1251–1266.

Didelot, X. and Wilson, D. J. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.*, 11(2): e1004041.

Dowson, C., Johnson, A., Cercenado, E., and George, R. 1994. Genetics of oxacillin resistance in clinical isolates of Streptococcus pneumoniae that are oxacillin resistant and penicillin susceptible. *Antimicrobial agents and chemotherapy*, 38(1): 49–53.

Edgar, R. C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5): 1792–1797.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., and Foll, M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet.*, 9(10): e1003905.

Falush, D., Stephens, M., and Pritchard, J. K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4): 1567–1587.

Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M.-S., Day, N. P., Enright, M. C., Goldstein, R., Hood, D. W., Kalia, A., Moore, C. E., *et al.* 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences*, 98(1): 182–187.

Fraser, C., Hanage, W., and Spratt, B. 2007. Recombination and the nature of bacterial speciation. *Science*, 315(5811): 476–480.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G., and Hanage, W. P. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science*, 323(5915): 741–746.

Hanage, W., Fraser, C., Tang, J., Connor, T., and Corander, J. 2009. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science*, 324(5933): 1454–1457.

Husmeier, D. 2005. Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics*, 21(suppl 2): ii166–ii172.

Kadioglu, A., Weiser, J. N., Paton, J. C., and Andrew, P. W. 2008. The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. *Nat. Rev. Microbiol.*, 6(4): 288–301.

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. 2012. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1): e1002453–e1002453.

Marttinen, P., Hanage, W. P., Croucher, N. J., Connor, T. R., Harris, S. R., Bentley, S. D., and Corander, J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research*, 40(1): e6–e6.

Marttinen, P., Croucher, N. J., Gutmann, M., Corander, J., and Hanage, W. P. 2015. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*. Accepted for publication.

Minin, V., Dorman, K., Fang, F., and Suchard, M. 2005. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, 21(13): 3034–3042.

Perron, G. G., Lee, A. E., Wang, Y., Huang, W. E., and Barraclough, T. G. 2012. Bacterial recombination promotes the evolution of multi-drug-resistance in functionally diverse populations. *Proc. Biol. Sci.*, 279(1733): 1477–1484.

Polz, M. F., Alm, E. J., and Hanage, W. P. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics*, 29(3): 170–175.

Popa, O., Hazkani-Covo, E., Landan, G., Martin, W., and Dagan, T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.*, 21(4): 599–609.

Shapiro, B. J. 2016. How clonal are bacteria over time? *Curr. Opin. Microbiol.*, 31: 116–123.

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., and Alm, E. J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336(6077): 48–51.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Trzciński, K., Thompson, C. M., and Lipsitch, M. 2004. Single-step capsular transformation and acquisition of penicillin resistance in Streptococcus pneumoniae. *Journal of Bacteriology*, 186(11): 3447–3452.

Vos, M. 2009. Why do bacteria engage in homologous recombination? *Trends in microbiology*, 17(6): 226–232.

Webb, A., Hancock, J. M., and Holmes, C. C. 2009. Phylogenetic inference under recombination using bayesian stochastic topology selection. *Bioinformatics*, 25(2): 197–203.

Weiser, J. N., Bae, D., Fasching, C., Scamurra, R. W., Ratner, A. J., and Janoff, E. N. 2003. Antibody-enhanced pneumococcal adherence requires iga1 protease. *Proceedings of the National Academy of Sciences*, 100(7): 4215–4220.

Woese, C. R., Olsen, G. J., Ibba, M., and Söll, D. 2000. Aminoacyl-trna synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews*, 64(1): 202–236.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The source code for Matlab and a compiled executable for Unix/Linux operating system are available at `https://users.ics.aalto.fi/~pemartti/fastGEAR/`.

Author's contributions

PM designed and implemented the method; RM and JC participated in designing the method; PM and RM designed the analysis of the pneumococcus data and PM carried out the analysis; RM and PM designed the simulations while NJC made substantial contributions to the design of simulations; RM carried out the simulations; NJC processed and prepared the data while CPA and WPH made substantial contributions; NJC, CPA and WPH helped analyse and interpret the results; PM and RM wrote the manuscript; All authors carefully read, edited, and approved the final manuscript.

Acknowledgements

This work was funded by the Academy of Finland (grants no. 286607 and 294015 to PM) and Junior Research Fellowship from Imperial College London (RM). The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project. Authors would like to thank Xavier Didelot for helpful comments on the manuscript and David Aanensen and Yonatan Grad for insightful discussions.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

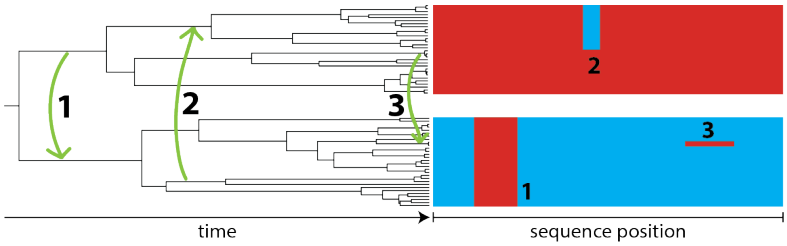

Figure 1: **Simulations of bacterial recombinations.** The diagram shows the underlying simulation method, and here the case of $P = 2$ populations is considered: blue and red. Populations were simulated under a clonal model of evolution for a given set of parameters (see Methods). Three types of recombinations were then simulated using the clonal alignment. Ancestral recombinations (event 1) occurred before the most recent common ancestor of both populations, and thus were present in all isolates of the recipient lineage. Intermediate recombinations (case 2) occurred sometime between the time when populations emerged and present time ($t = 0$), and thus were typically present in multiple isolates. Recent recombinations (case 3) occurred in the last few generations, and thus were typically present in few isolates.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

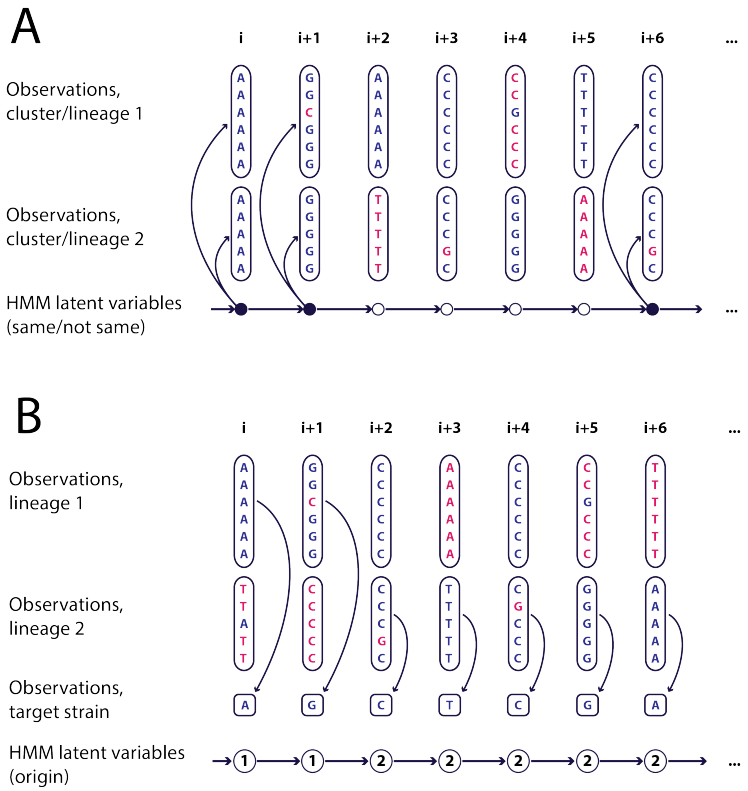

Figure 2: **Hidden Markov models to detect recombination.** (A) Hidden Markov model used for identifying lineages and inferring ancestral recombinations. Each column represents a polymorphic site in the alignment and rows represent strains. The observed states of the chain are nucleotides within each cluster (in the case of identifying lineages) or lineage (in the case of identifying ancestral recombinations). The latent states of the chain represent identity of allele frequencies in the two lineages at the polymorphic sites. (B) Hidden Markov model used for identifying recent recombinations. The observed states are frequencies within each lineage and the latent states are the possible origins of the target strain. The possible origins include all observed lineages plus an unknown state.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

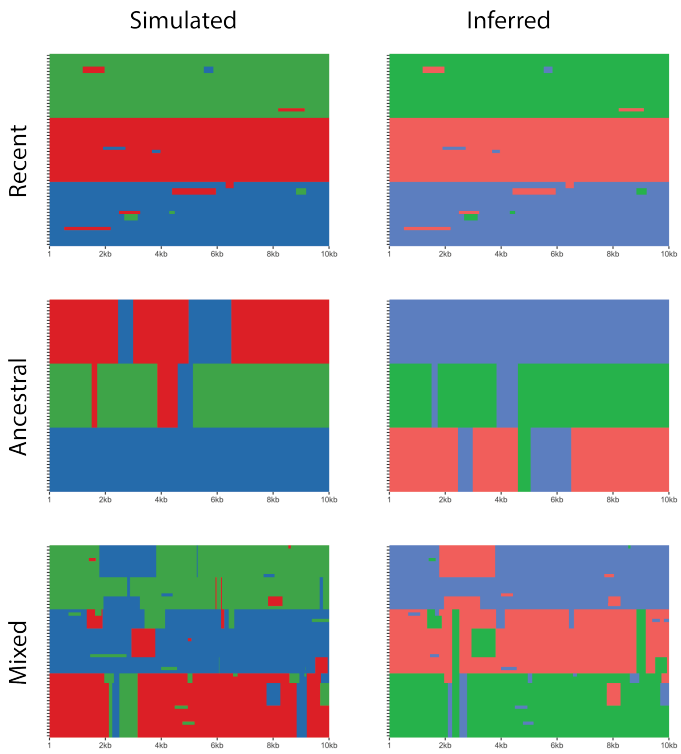

Figure 3: **Visual assessment of the inferred population genetic structure.** The figure shows the population genetic structure of the simulated data. In each panel, the rows correspond to sequences, columns correspond to positions in the alignment and colours show different populations. The left column shows the simulated, true structure while the right column shows the population genetic structure inferred by `fastGEAR`. The order of the sequences in both columns is identical, and the colours are assigned randomly, thus populations are in the same order (1,2,3) but can be of different colour on the left and on the right. Three figure rows correspond to three different simulation scenarios: only recent recombinations (top), only ancestral recombinations (middle), and all three types of recombinations (bottom). The following parameters were used in the simulations: $P = 3$, $n = 20$, $N_e = 50$, $T = 2 \times 10^4$, $\mu = 2 \times 10^{-6}$, $L = 10$kb (all rows); $\Gamma_r = 800$ and $R_r = 5$, $R_i = R_a = 0$ (top panel); $\Gamma_a = 800$ and $R_a = 3$, $R_r = R_i = 0$ (middle panel); $\Gamma_a = 500$, $\Gamma_r = 500$ and $R_a = 3$, $R_i = 4$ and $R_r = 6$ (bottom panel).

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

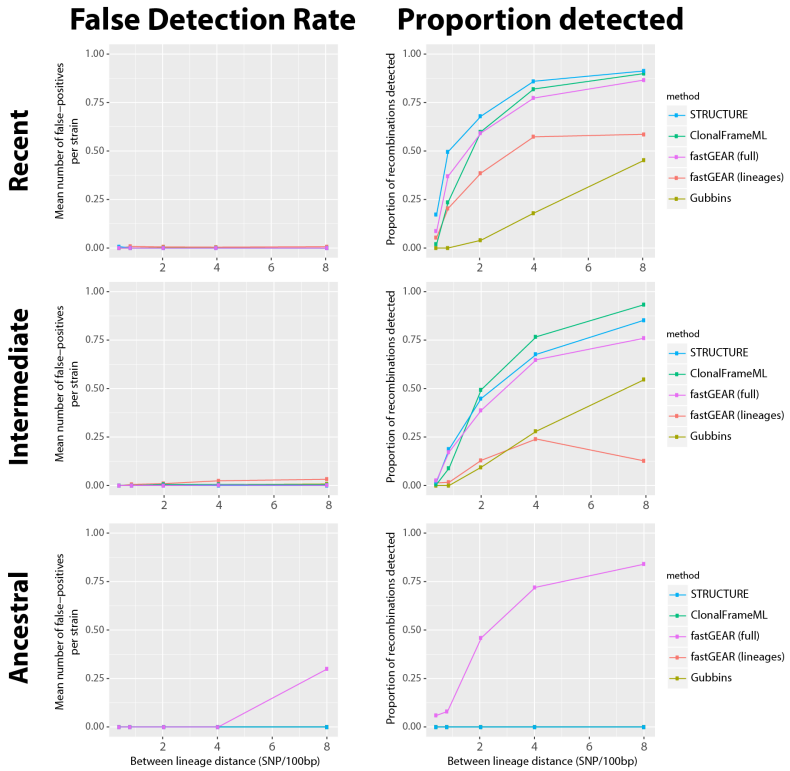

Figure 4: **Comparison of `fastGEAR` to other bacterial recombination detection methods.** The figure shows performance of `fastGEAR` in detecting recombinations when compared to other recombination-detection methods for bacterial genomes: STRUCTURE, Gubbins and ClonalFrameML. Top row shows results for recent, middle row for intermediate, and bottom row for ancestral simulated recombinations. Both recent and intermediate simulated recombinations were detected by `fastGEAR` in the same way as 'recent' recombinations. The left column shows the false detection rate, namely, the mean number of false-positive recent recombinations per strain (top/middle) and ancestral recombinations per alignment (bottom). The right column shows the proportion of detected true recombinations. Horizontal axis shows the between-population distance per 100bp (simulated by varying T between 10^3 and 2×10^4). Different lines show performance of different approaches. Blue line shows results of STRUCTURE run for 400,000 generations (200,000 burn-in), with true populations set as prior and with three independent chains to test for convergence of the MCMC. Green line shows results of ClonalFrameML conditioned on the true phylogeny and run on lineage-by-lineage basis. Magenta line shows results of `fastGEAR` run on the full alignment. Red line shows results of `fastGEAR` run lineage-by-lineage. Green line shows results of Gubbins run lineage-by-lineage. Each point represents the mean of ten independent simulations. The following parameters were used in the simulations: $P = 3$, $n = 30$, $\mu = 2 \times 10^{-6}$, $L = 20$kb, $\Sigma_r = 300$, $\Sigma_a = 600$ and $R_r = 5$, $R_i = R_a = 5$.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

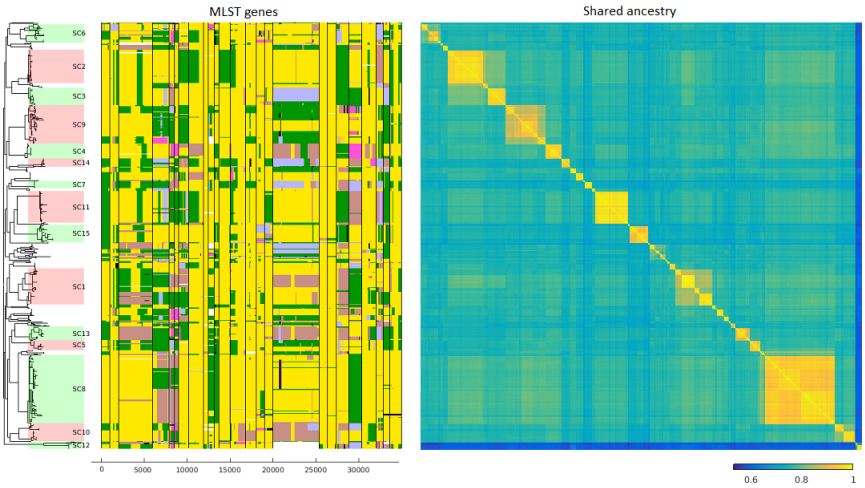

Figure 5: **Population structure in the pneumococcal data** The phylogeny and the sequence clusters (SCs) on the left show the core-genome-based tree with 15 major monophyletic clusters. Middle panel shows `fastGEAR` output for 25 out of 96 housekeeping genes, as discussed in the text; results for all 96 genes are qualitatively the same and shown in Figure S4. The colors represent different lineages identified in the analysis. The results for the different genes were obtained by running `fastGEAR` independently, but the lineage colors at different genes were permuted to approximately minimize the average entropy of the colour distributions of the strains over all the genes. White colour denotes missing data. The PSA matrix on the right shows the proportion of shared ancestry between the isolates in the data set, ranging from blue (distant) to yellow (closely related).

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

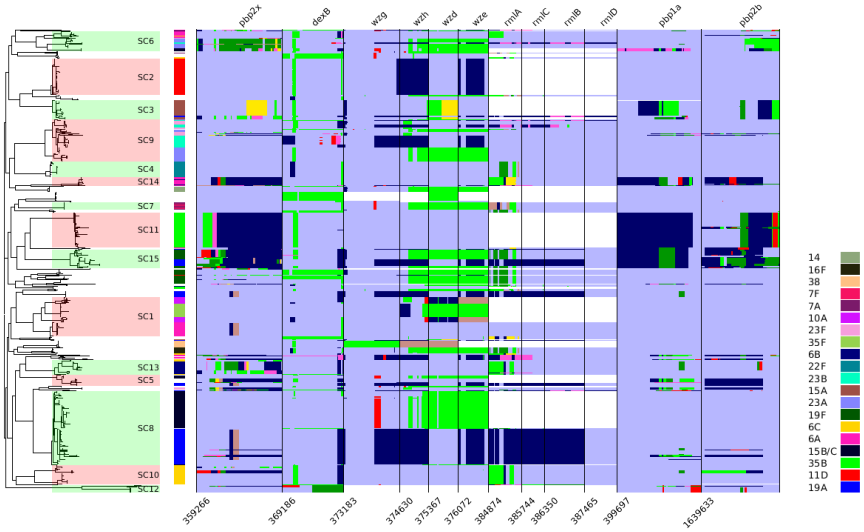

Figure 6: **Detailed results for the clinically relevant genes in the pneumococcal data.** The phylogeny on the left shows the population structure and the SCs in the data. The annotation on the right-hand side of the phylogeny shows the serotype of isolates for the 20 most common serotypes. The main panel shows the `fastGEAR` outputs for penicillin-binding proteins and capsule-flanking genes, whose names are shown above and locations in 670-6B reference (Genbank: CP002176) are shown below. Note that the colours of the lineages in the results are selected independently of the colours of the serotypes.

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

COG	Number SNPs	Number sequences	Clusters	Lineages	Number ancestral	Gene name
CLS00019	788	288	17	10	58	Pneumococcal surface protein C
CLS00082	818	2149	24	17	51	Insertion sequence IS1167
CLS01094	543	616	21	8	47	Chromosome partition protein Smc
CLS00114	618	2320	25	19	39	Insertion sequence IS1381
CLS00092	297	325	14	11	34	Bacteriophage DNA binding protein
CLS00949	1177	378	10	7	32	Tn5252 relaxase
CLS00355	720	616	12	7	31	Penicillin-binding protein 2x
CLS01539	191	616	20	10	30	Ribosomal protein L11 methyltranferase
CLS01729	332	612	14	7	23	Two component system histidine kinase
CLS02405	258	336	11	7	23	Capsule locus glucose-1-phosphate thymidyl transferase RmlA
CLS00543	320	616	22	9	22	Phenylalanyl-tRNA synthetase (PheS)
CLS02258	310	229	11	7	22	Protein found in phage-related chromosomal islands

Table 1: COGs with the highest number of ancestral recombination events (>20) identified by fastGEAR

bioRxiv preprint doi: https://doi.org/10.1101/059642; this version posted June 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

COG	Number SNPs	Number sequences	Clusters	Lineages	Number recent	Gene name
CLS00019	788	288	17	10	69	Pneumococcal surface protein C
CLS01094	543	616	21	8	69	Chromosome partition protein Smc
CLS00355	720	616	12	7	57	Penicillin-binding protein 2x
CLS00543	320	616	22	9	40	Phenylalanyl-tRNA synthetase (PheS)
CLS00534	536	616	16	5	37	Valyl-tRNA synthetase (ValS)
CLS01539	191	616	20	10	37	Ribosomal protein L11 methyl-transferase
CLS00380	583	606	23	4	35	Large cell wall surface anchored protein
CLS01435	558	615	12	4	34	Penicillin-binding protein 2b
CLS02424	2021	248	10	4	34	Zinc metalloprotease A
CLS00326	490	616	11	7	31	Folylpolyglutamate synthase (FolC)
CLS00824	263	614	17	7	31	Transporter
CLS01432	393	616	15	8	29	Cell wall synthesis protein MurF
CLS00114	618	2320	25	19	28	Insertion sequence IS1381
CLS00389	326	597	15	5	28	Choline binding protein C/J
CLS00185	284	616	18	6	27	Glucose-inhibited division protein GidA
CLS01729	332	612	14	7	27	Two component system histidine kinase
CLS02317	741	97	7	5	27	Phage protein
CLS00357	574	614	9	5	26	ATP-dependent protease ATP-binding protein ClpL
CLS01858	242	613	19	5	26	Mannosidase
CLS01987	138	565	20	6	26	Membrane protein
CLS02463	421	64	9	6	26	Part of large cell wall surface anchored protein PsrP

Table 2: Genes with the highest number of recent recombination events (>25) identified by fastGEAR