MICROBIAL GENOMICS
Bases to Biology

Short Paper

Recombination produces coherent bacterial species clusters in both core and accessory genomes

Pekka Marttinen,[1,4] Nicholas J. Croucher,[2] Michael U. Gutmann,[3] Jukka Corander[3] and William P. Hanage[4]

[1]Aalto University, Espoo, Finland

[2]Imperial College, London, UK

[3]University of Helsinki, Helsinki, Finland

[4]Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA, USA

Correspondence: Pekka Marttinen (pekka.marttinen@aalto.fi)

DOI: 10.1099/mgen.0.000038

Background: Population samples show bacterial genomes can be divided into a core of ubiquitous genes and accessory genes that are present in a fraction of isolates. The ecological significance of this variation in gene content remains unclear. However, microbiologists agree that a bacterial species should be 'genomically coherent', even though there is no consensus on how this should be determined.

Results: We use a parsimonious model combining diversification in both the core and accessory genome, including mutation, homologous recombination (HR) and horizontal gene transfer (HGT) introducing new loci, to produce a population of interacting clusters of strains with varying genome content. New loci introduced by HGT may then be transferred on by HR. The model fits well to a systematic population sample of 616 pneumococcal genomes, capturing the major features of the population structure with parameter values that agree well with empirical estimates.

Conclusions: The model does not include explicit selection on individual genes, suggesting that crude comparisons of gene content may be a poor predictor of ecological function. We identify a clearly divergent subpopulation of pneumococci that are inconsistent with the model and may be considered genomically incoherent with the rest of the population. These strains have a distinct disease tropism and may be rationally defined as a separate species. We also find deviations from the model that may be explained by recent population bottlenecks or spatial structure.

Keywords: computational modeling; core/accessory genome; evolution; recombination; speciation.

Abbreviations: HGT, horizontal gene transfer; HR, homologous recombination.

Data statement: So supplementary table animations has been deposited in Figshare. All supporting data, code and protocols have been provided within the article or through supplementary data files.

Data Summary

1. Supplementary Animations have been deposited in Figshare: http://figshare.com/s/6471c982669011e58c4806ec4b8d1f61
2. R code to run the model has been deposited in Figshare: http://figshare.com/s/c70dd5e0669011e59ff906ec4bbcf141

Introduction

Bacterial diversity can be described in terms of the alleles of core genes common to all strains and the additional acces-

Received 14 August 2015; Accepted 6 October 2015

sory genes present in a subset of strains. For example, as little as 11 % of all *Escherichia coli* genes described are present in all strains of the species (Perna *et al.*, 2001; Touchon *et al.*, 2009), and the concepts of the 'core' and 'pan' genomes are now commonplace. Variation in gene content is often assumed to be selective, reflecting different ecological specialization, but this has rarely been formally tested (Baltrus, 2013) and evidence exists that the selective consequences of horizontal gene transfer (HGT) may be surprisingly small (Knöppel *et al.*, 2014). The profusion of large population-based studies of individual pathogens presents us with an opportunity to test different models of diversification, explicitly examining the expected core and accessory genome distribution.

Models of diversification in the core genome point to the vital role of homologous recombination (HR) in forming clusters of related strains and maintaining population structure (Fraser *et al.*, 2007, 2009; Doroghazi & Buckley, 2011). However, these models do not account for recombination events affecting the gene content. On the other hand, several models have successfully provided insight into how gene content evolves to produce the characteristic U-shaped histogram of gene frequencies observed at multiple levels of taxonomy (an example is shown in Fig. 1a) (Baumdicker *et al.*, 2012; Collins & Higgs, 2012; Haegeman & Weitz, 2012; Lobkovsky *et al.*, 2013). Many extensions also exist: expanding population (Baumdicker *et al.*, 2012), genes with different fitnesses (Lobkovsky *et al.*, 2013), and multiple gene categories with different deletion/acquisition rates (Collins & Higgs, 2012; Haegeman & Weitz, 2012). These models have included rates of acquisition and loss of genes, but have not modelled the divergence of the core simultaneously with that of the accessory genome nor investigated the potential for gene exchange by HR. The recent emergence of population genomics has produced datasets of hundreds or thousands of genomes from the same species, sampled in a systematic fashion (Croucher *et al.*, 2013; Chewapreecha *et al.*, 2014). Here, we present a model that includes both core genome and gene content variation, and use it to examine a well-characterized collection of 616 *Streptococcus pneumoniae* genome sequences (Croucher *et al.*, 2013).

The joint distribution of core genome and gene content divergence in the data shows that gene content, measured here in terms of clusters of orthologous groups (Tatusov *et al.*, 1997), diverges approximately linearly with core genome sequence (Croucher *et al.*, 2014) (Fig. 1b). The dominant feature in the distribution is the concentration of the majority (~ 86 %) of the distances within a small, clearly delineated region. This mode results from the fact that all but one of the 15 major sequence clusters detected in the population are approximately equally distant from each other by both metrics (Croucher *et al.*, 2013). Another small mode near the origin corresponds to distances between very closely related strains and the small mode in the top-right corner represents strains in the single more divergent cluster; these strains have previously been characterized as 'atypical pneumococci' (Croucher *et al.*, 2014).

Impact Statement

Bacterial species should be 'genomically coherent', but what this means is unclear due to the horizontal gene transfer that they exhibit. We fit a simulation of diversification in the core and accessory genome, including horizontal transfer, to a sample of >600 pneumococcal genomes, capturing the major features of the data and providing estimates of key parameters highly consistent with independent empirical measurements. The model predicts the surprising observation that all but one of the major strain clusters in the data are equidistant from each other as measured in terms of either core or accessory genome divergence – a feature that we show can be produced by biologically plausible recombination rates. Notably, the model is neutral with regard to the fitness of the different gene combinations that make up each genome. Deviations from model prediction indicate a departure from neutral expectations worthy of further investigation: strains that are more divergent than expected may be defined as a distinct species, suggesting a rational basis for the definition of a genomically coherent species. Strains that are more closely related may reflect short-term selective and epidemiological processes.

Methods

Model. Previous models have considered the observed diversity in the core genome of loci present in all strains (Fraser *et al.*, 2007, 2009). We extend this to include the accessory genome, with parameters governing the gain and loss of genes. Here, we present an overview of our approach; a detailed description of the model and the model fitting algorithm are provided in the supplementary text, Figs S1–S5, and Tables S1 and S2. Briefly, we simulate a population of sequences according to the Wright–Fisher model by sampling with replacement from the previous generation, with the following events possible at each generation: gene introduction, gene deletion, HR (replacing the recipient allele with the donor allele), HGT between two strains (altering the genome content in the recipient) and mutation. Our model is parsimonious, with just five free parameters representing rates of the different events, and is neutral with respect to the success of individual genes or lineages, and the resulting association between the core and accessory loci. A small multiplicative fitness penalty (using a factor of 0.99; values in the range from 0.95 to 0.999 produced similar results; see Figs S14–S16 for sensitivity analyses) is imposed for each gene exceeding a prespecified genome size, to prevent the genome growing without limit. Recombination

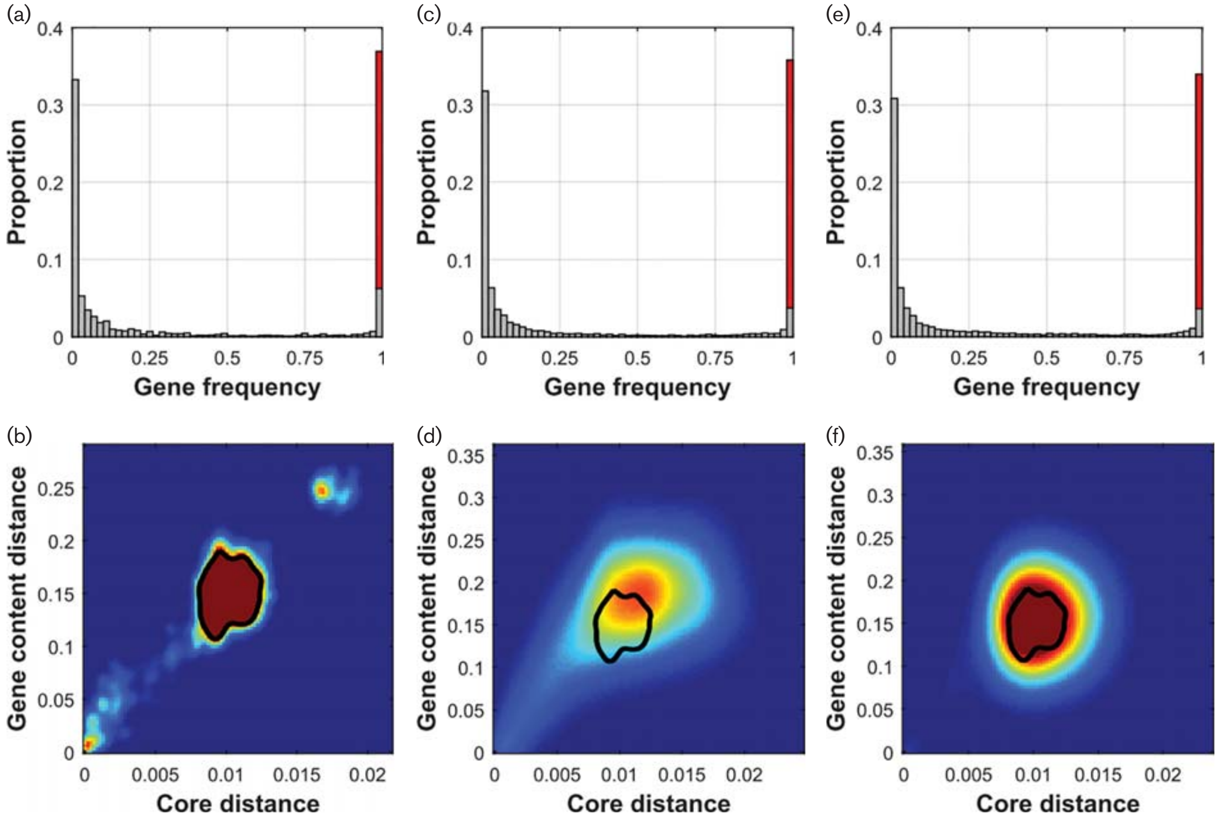

Fig. 1. Gene frequency histograms (a, c, e) and strain distance distributions (b, d, f). The frequency histograms (a, c, e) show the number of very rare or common genes is much larger than the number of genes at intermediate frequencies; the red column represents the core genome (the overlapping grey bar represents frequencies f with $0.98<f<1$). The distance distributions (b, d, f), obtained by averaging over the whole simulation after discarding initial samples, are based on pairwise comparisons of strains, showing the core genome (Hamming) distance on the x-axis and the gene content (Jaccard) distance on the y-axis (see Methods). A contour line encompassing the mode in the real data is shown in the simulated distributions for easier comparison. The columns show results in the real data (a, b), in the model with learned parameter values (c, d) and in the model with between-strain recombination increased by a factor of 10 (e, f).

events are accepted with a probability that decreases exponentially with increasing sequence divergence, reflecting a log-linear decline in the frequency of recombination with the divergence of donor and recipient sequences, as observed in empirical studies (Vulić *et al.*, 1997; Majewski *et al.*, 2000; Zawadzki *et al.*, 1995). To reduce computational complexity further, we use a low-dimensional representation for the gene sequences and approximate the real distances with Monte Carlo simulation.

Model fitting. A normal maximum-likelihood approach to model fitting is computationally infeasible, so we use simulation-based inference, and match summary statistics between the simulated and real data; this resembles the simulated method of moments (McFadden, 1989; Pakes & Pollard, 1989; Gouriéroux & Monfort, 1997; Wood, 2010). To determine the parameter value maximizing the similarity between simulated and real data, we model the overall similarity score over a range of plausible parameter values by non-parametric regression (Rasmussen, 2006; Gutmann & Corander, 2015). The model fitting procedure involves a subjective decision on selecting data summaries to use when comparing between real and simulated data. We used one multivariate and three scalar summaries, all of which varied systematically in the simulation, allowing unambiguous identification of the model parameters. The multivariate summary was the U-shaped gene frequency histogram (Fig. 1a), which was highly informative about gene deletion and introduction rates (Fig. S6). For determining the HGT and HR rates, we defined two additional data summaries, termed here as the 'clonality score' and the 'linkage score', respectively. These measure the randomness of the distribution of the accessory genes in the population and the correlation between core loci, with high rates resulting in low scores (Figs S4 and S5). The slope of the distance distribution (Fig. 1b) was used as the last statistic informative about mutation rate. Namely, high mutation rate stretched the distribution along the x-axis, resulting in a more gradual slope.

Table 1. Estimates for two parameters: r/m (the number of substitutions introduced by recombinations versus mutations) and the ratio of gene introduction/deletion rates

The second column reports the estimate from the model and the third column an estimate from a detailed genomic analysis (see Methods).

Parameter	Model estimate	Genomic analysis
r/m	8.0	11.3
Gene introduction/deletion	1.3	1.4

Distance metrics. The Hamming distance between two strains, used to measure the core genome divergence, measures the proportion of differing sites in the core genome alignment. The Jaccard distance, used to measure the gene content divergence, equals the number of genes present in one and absent in the other strain, divided by the total number of genes present in either one of the strains.

Data. For simplicity, we use a term ‘gene’ to refer to a cluster of orthologous groups throughout this paper. Core gene alignments, cluster annotation of the strains, the gene presence–absence matrix and a phylogenetic tree have been described previously (Croucher *et al.*, 2013). As an additional data cleaning step, we removed all genes whose alignment length was <265 bp, which corresponded to the 0.05th quantile of the lengths of the alignments of the core genes. This step was added to increase confidence in the genes detected. This left us with 2692 accessory genes and 1191 core genes in the 616 pneumococcal isolates. The detailed genomic analysis estimates for gene introduction and deletion rates in the real data, provided in Table 1, were obtained by estimating maximum-likelihood reconstructions of the genes along the fixed phylogeny, using an R function ace from package APE (Paradis *et al.*, 2004). The number of substitutions introduced by recombinations versus mutations, r/m, was computed as the mean over estimates reported for the sequence clusters (Croucher *et al.*, 2013).

Results and Discussion

Our fitted model predicts a stationary mode in the distance distribution, in the same location as in the real data, and increasing the recombination rate does not alter its location (Fig. 1). Thus, the mode appears to represent a limit for divergence in the population similar to what has previously been reported from gene sequence models (Fraser *et al.*, 2007), but, strikingly, we see a similar limit in the divergence of gene content. Note that the model was fit without assuming the mode, using metrics in the model fitting process that were independent of the mode. Altering the recombination rate has a major impact on dynamics. Whilst the position of the mode is consistent when averaged over time, it can move markedly over short timescales and separate into multiple clusters (see Animations S1–S3). With extremely low recombination rates, the observed mode does not emerge and the model output is merely distinct groups of closely related strains drifting rapidly apart from each other. After the mode emerges, increasing recombination within the population (i.e. the HR and HGT rates, see Model), whilst maintaining other parameters in their fitted values, does not change its location but rather stabilizes it. This indicates the impact of recombination on the population structure as measured here saturates when the distribution of alleles/genes between strain clusters is close to random, which is the required condition for the mode to emerge. The saturation can be seen in the levelling of the scores used in model fitting (Figs S4 and S5). For example, when two loci have become relatively uncorrelated due to recombination, further recombination has little impact.

Fitted values of the five parameters are shown in Table S1. In addition to the raw values, we recorded information of all events during the simulation, from which we computed the total number of substitutions introduced by HR and mutation, and the total number of gene introductions (caused either by an introduction of a new gene into the population or a within population gene transfer) and the overall number of gene losses (caused either by deletions or within population gene transfers). The resulting estimates of the ratio of recombination to mutation and the acquisition and loss of novel loci reported in Table 1 broadly recapture estimates from previous work analysing sets of whole-genome alignments (Croucher *et al.*, 2013).

There are important ways in which our model does not capture the observed data; one such is the small peak in the distance distribution close to the origin. To determine what might produce this, we extended the model in two simple ways. (1) We created a geographically biased sample, reflecting the way the real data were collected, similar to previous work on relating genetic divergence to short transmission chains (Fraser *et al.*, 2005). (2) We examined the impact of a population bottleneck, acting as a collective proxy for processes whereby some strains leave more progeny than others, including recent selection (Fraser *et al.*, 2009). For example, a recent vaccine introduction has led to rapid changes in the prevalences of certain serotype groups in the population (Croucher *et al.*, 2013). Outputs from these extensions demonstrate that both mechanisms can contribute to the peak, whilst leaving the main mode in the distribution intact, and further work will focus on estimating their relative significance (Fig. 2). Another major feature contradicting the expectation is the separate mode in the upper right corner of the distance distribution corresponding to a sequence cluster (SC12) divergent from the main group. Animation S2 shows how such additional modes emerge with decreased recombination, suggesting limited exchange between SC12 and the rest of the population. Notably, the previously reported recombination rate for SC12, detecting recombination as anomalous tracts of SNPs in the alignment, is relatively high (Fig. S9). This suggests that SC12 may be recombining with strains unrepresented in the population or conceivably that the SNPs in question are the consequence of some

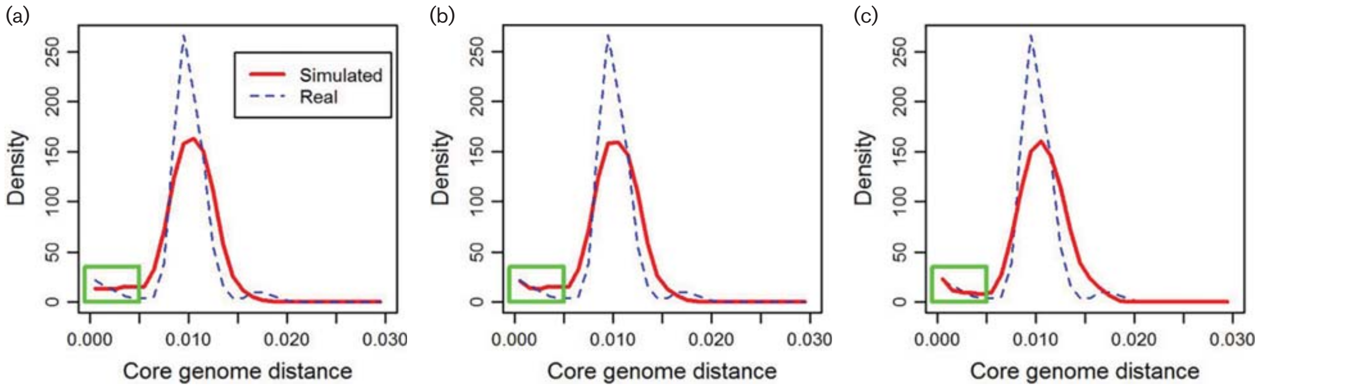

Fig. 2. Effects of geographical sampling bias and a recent bottleneck on the core genome Hamming distance distribution. Strains from a simulated generation, representative of the average shape, were selected as the initial population (a). The green rectangle highlights the region of interest, showing the increase in the number of closely related strain pairs in the real data. (b) The distance distribution after taking a geographically structured sample, averaged over 20 independent replicates (red curve). (c) The effect of a population bottleneck, obtained by selecting a specified number of strains (here 100 out of 2000 strains in total) as possible ancestors from which the next generation was sampled with replacement. Bottlenecks of other sizes are shown in Fig. S10. The distribution for the real data is shown in each panel for comparison.

selective process that means SC12 does not fit our model (additional results are presented in the supplementary text and Figs S6–S16).

Conclusion

We imposed a soft limit on genome size by assuming in our model a small fitness penalty for increasing genome size beyond a given threshold (see Methods). An analogous assumption has also been used by others (Vogan & Higgs, 2011), and whilst some selective pressure against larger genomes likely exists, the approach seems overly simplified. The limit is needed for computational reasons, but it also accounts for the empirical observation that genome sizes are not constantly increasing. Importantly, the limit does not produce any heritable fitness differences between different combinations of genes and the results are robust over a wide range of possible parameter values. Previous models have approached the same issue by either letting genomes grow (Baumdicker *et al.*, 2012) or by coupling gene introductions and deletions (Haegeman & Weitz, 2012; Lobkovsky *et al.*, 2013), both of which also seem arbitrary. In reality, several explanations may underlie the observation. In our model, the assumption facilitates the fitting of the gene frequency histogram as a stationary condition, from which the dominant mode in the distance distribution follows, given sufficient shuffling of genes between strains by recombination. Surprisingly, no additional assumptions, such as niche adaptation or selection on individual genes, are needed to explain the mode. The equidistant sequence clusters predicted by the model are consistent with previous findings showing the majority of differences in gene content between strain clusters to be related to combinations of loci, rather than unique cluster-defining genes (Croucher *et al.*, 2014).

We have developed a parsimonious model of genome evolution and shown that it can capture important features of a bacterial population, including the distance distribution between the strains and the gene frequency histogram. In addition, we have used it to detect characteristics of data that are not concordant with neutral expectations. We have demonstrated the importance of recombination in producing the population structure, as represented by either the gene content or the core genome divergence. Despite several ways in which the model is idealized, it broadly estimates the population genetic parameters well. A remarkable fact is that the model predicts the population of equidistant strain clusters observed in the real data without recourse to selection or niche adaptation; however, we emphasize that our purpose here is not to reject selection, but merely to point out its redundancy in explaining this striking feature of the population structure. We used our model as a null hypothesis to detect features not expected by neutral processes. For example, closely related strains required an additional explanation, such as a bottleneck. Furthermore, strains that were more divergent than expected, forming a distinct mode in the distance distribution, may be rationally defined as distinct species. Thus, our model might serve as a definition for a '(preferably) genomically coherent' species, which is an aspiration of systematicists in response to the growth of genomic data. Improved annotation of accessory genomes, coupled with extensions of our model, will enable us to ask whether the observed gene combinations are more or less frequent than we would expect to see by chance.

The model we have developed offers insights into the processes that generate genotypic clusters associated with species in recombinogenic bacteria (see also Shapiro & Polz, 2014). There are obvious similarities to the biological species concept in eukaryotes, in which sexual reproduction operates as a cohesive force preventing divergence of lineages (e.g. Higgs & Derrida, 1992). However, the differences between eukaryotic and prokaryotic reproduction make this a more general version of the concept that is also capable of considering recombination between things we might term species, without being sufficient to prevent those species clusters becoming distinct. The divergent cluster of 'atypical pneumococci' may be considered a separate species by our criteria, i.e. that it forms a distinct mode that cannot be explained by the mean recombination rate within the population. There are multiple mechanisms by which this could have occurred, but common to them all is insufficient recombination between the two clusters, allowing them to diverge. The cause of that barrier is impossible to determine from the present analysis, but could be intrinsic (the two clusters do not recombine efficiently) or ecological (isolates in the two clusters do not encounter each other often enough for recombination to efficiently shuffle their genomes). Further work is necessary to distinguish between these possibilities.

Acknowledgements

This work was funded by the Academy of Finland (grant no. 251170 to the Finnish Centre of Excellence in Computational Inference Research COIN; grant no. 259272 to PM).

References

Baltrus, D. A. (2013). Exploring the costs of horizontal gene transfer. *Trends Ecol Evol* **28**, 489–495.

Baumdicker, F., Hess, W. R. & Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. *Genome Biol Evol* **4**, 443–456.

Chewapreecha, C., Harris, S. R., Croucher, N. J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D. M., Mather, A. E. & other authors (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet* **46**, 305–309.

Collins, R. E. & Higgs, P. G. (2012). Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol Biol Evol* **29**, 3413–3425.

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

Croucher, N. J., Coupland, P. G., Stevenson, A. E., Callendrello, A., Bentley, S. D. & Hanage, W. P. (2014). Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* **5**, 5471.

Doroghazi, J. R. & Buckley, D. H. (2011). A model for the effect of homologous recombination on microbial diversification. *Genome Biol Evol* **3**, 1349–1356.

Fraser, C., Hanage, W. P. & Spratt, B. G. (2005). Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci U S A* **102**, 1968–1973.

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

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

Gouriéroux, C. & Monfort, A. (1997). *Simulation-based Econometric Methods*. Oxford: Oxford University Press.

Gutmann, M. U. & Corander, J. (2015). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, in press, arXiv:1501.03291.

Haegeman, B. & Weitz, J. S. (2012). A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* **13**, 196.

Higgs, P. G. & Derrida, B. (1992). Genetic distance and species formation in evolving populations. *J Mol Evol* **35**, 454–465.

Knöppel, A., Lind, P. A., Lustig, U., Näsvall, J. & Andersson, D. I. (2014). Minor fitness costs in an experimental model of horizontal gene transfer in bacteria. *Mol Biol Evol* **31**, 1220–1227.

Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. (2013). Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol Evol* **5**, 233–242.

Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. (2000). Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol* **182**, 1016–1023.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* **57**, 995–1026.

Pakes, A. & Pollard, D. (1989). Simulation and the asymptotics of optimization 262 estimators. *Econometrica* **57**, 1027–1057.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in r language. *Bioinformatics* **20**, 289–290.

Perna, N. T., Plunkett, G. III, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J. & other authors (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Shapiro, B. J. & Polz, M. F. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology* **22**, 235–247.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631–637.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C. & other authors (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5**, e1000344.

Vogan, A. A. & Higgs, P. G. (2011). The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. *Biol Direct* **6**, 1.

Vulić, M., Dionisio, F., Taddei, F. & Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci U S A* **94**, 9763–9767.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**, 1102–1104.

Zawadzki, P., Roberts, M. S. & Cohan, F. M. (1995). The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. *Genetics* **140**, 917–932.

Data Bibliography

1. **Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J. & Hanage, W. P.** (2015). Figshare. http://figshare.com/s/6471c982669011e58c4806ec4b8d1f61

2. **Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J. & Hanage, W. P.** (2015). Figshare. http://figshare.com/s/c70dd5e0669011e59ff906ec4bbcf141

Recombination produces coherent bacterial species clusters in both core and accessory genomes - Supplementary material

Pekka Marttinen, Nicholas J. Croucher, Michael U. Gutmann,
Jukka Corander, and William P. Hanage

Contents

1 Detailed Methods

1.1 Model

As the basis of our simulation, we use the Wright-Fisher model, in which the next generation of strains is sampled with replacement from the current generation. The model may be modified to incorporate fitness by sampling strains with unequal probabilities. Motivated by the biological observation that genome sizes are not constantly increasing, we include a multiplicative fitness penalty (using a factor 0.99) to an otherwise neutral model for each gene exceeding a pre-specified number of genes to prevent unrestricted growth.

In our model, we use two separate data structures to represent the genome of each strain: 1) gene content component, represented by a binary indicator vector showing the presence and absence of genes, and 2) gene sequence component, represented by a fixed number of gene sequences for which detailed evolution is simulated (Supplementary Figures 1A and 1B). The two components are present in the same strains, but otherwise their evolution is independent. In the gene content component we assume no fitness differences between genes, and, in particular, make no distinction between core and accessory genes. Results concerning gene content, for example the proportion of core and the gene frequency histogram, are derived from this component. Because one of our goals is to investigate the relationship between gene content and core sequence divergences (Figure 1), we include the gene sequence component into our model and use it to compute the core genome distances. For simplicity, we assume the same genes to be present in all strains in the gene sequence component.

The evolution of the gene content is driven by the following events, taking place between the sampling of strain generations with frequencies specified by the parameters of the model:

1. Introduction of a novel gene into the population

2. Deletion of a randomly selected gene from a randomly selected strain

3. Horizontal gene transfer between two strains in the population, resulting in the gene presence/absence status of the donor to be copied to the recipient (Supplementary Figure 1C).

For modeling the detailed evolution in the gene sequence component, we use the following events:

1. Mutation

2. Homologous recombination, in which an allele of the recipient is replaced by an allele from the donor (Supplementary Figure 1C).

A detailed parameterization of the model is shown in Tables S1 and S2. Our motivation for separating gene content and core sequence evolution stems from computational savings resulting from the fact that it is not necessary to simulate the detailed sequence evolution for all genes. As a preliminary experiment, we implemented also a model with the two components combined and similar results were obtained; however, fitting the model took considerably longer (results not shown).

Our model assumes that a gene can enter the population only once, after which its evolution is driven by drift and recombination. Furthermore, as observed with several real data sets, the potential of bacteria to recombine decreases with decreasing sequence similarity. Motivated by this, we make the same assumption as Fraser et al. (2007), and accept recombination proposals with probability that decreases exponentially with increasing sequence divergence using parameter values observed in real data sets (see Table S2 for details).

To further reduce computational complexity, gene sequences, whose detailed evolution is simulated, are represented using a low-dimensional feature space analogous to Fraser et al. (2007). In detail, each gene is represented by a vector of 10 integers, serving as abstract sequence features. Every time a mutation occurs in the gene, a randomly selected feature is incremented by one. The distance between two strains computed using the feature representation underestimates the real sequence distance due to an increased probability of two mutations occuring at the same location and we correct for the bias by mapping the estimated distances to the expected true distances, using a mapping obtained with Monte Carlo simulation (Supplementary Figure 2), which has an improved accuracy compared to an analytical correction used by Fraser et al. (2007) in the range of distances relevant for this study.

1.2 Model fitting

Our model has in total five free parameters: three governing the evolution of gene content (deletion rate, novel gene introduction rate, horizontal gene transfer rate) and two governing the detailed sequence evolution (mutation rate, homologous recombination rate). Given the complexity of the model, maximizing its likelihood is computationally infeasible. We employ a simulation-based inference method instead, which resembles the simulated method of moments (McFadden, 1989; Pakes and Pollard, 1989; Gourieroux and Monfort, 1997; Wood, 2010). The basic idea is to fit the model by matching summary statistics of the real data. A two-step algorithm is used for fitting the model:

- Fit the parameters for the gene content evolution using 1) the gene frequency distribution, and 2) the median clonality score (see below) over genes present in approximately half (40-60%) of the strains.
- Fit the parameters for the detailed sequence evolution using 1) the slope of the gene content vs. core genome distance relationship, and 2) the median linkage score (see below) over all core gene pairs.

Each optimization step consists of simulating multiple artificial data replicates over a set of values for parameters to be optimized, and measuring the similarity between the simulated and real data statistics using a similarity measure (see below). Due to simulation variability, the similarity score between the simulated and real data sets fluctuates even if exactly the same parameter values are used in different simulation runs. For this reason, we do not have a closed form formula for the relation between similarity score and parameters. We learn the relation for the range of plausible parameter values by non-parametric regression (Rasmussen, 2006). Our estimate is obtained as the parameter value that maximizes the learned regression function, which represents the expected similarity between the simulated and real data. Supplementary Figure 3 illustrates this procedure when learning the parameters for the detailed sequence evolution.

The model fitting procedure incorporates a subjective decision of selecting data summaries to use when matching the real and simulated data sets. Ideally, the summaries would identify the parameters unambiguously. When fitting the gene content component, for example, the gene frequency distribution statistic was found to be highly informative about deletion and novel gene introduction rates; however, it did not contain sufficient information for identifying the horizontal gene transfer rate. For learning the horizontal gene transfer and homologous recombination rates, we defined two additional data summaries, the clonality score and the linkage score, respectively. Each of these two scores was found to vary monotonically with the recombination rates, such that high rates indicated low clonality or linkage scores (Supplementary Figures 4 and 5). Details of the two scores are provided below.

The *clonality score* for a gene is defined on the basis of the fact that the gene divides the strains into two groups, those with the gene, and those without. In the absence of horizontal gene transfer, the two groups would correspond to different branches of a phylogenetic tree, and, consequently, the within-group strain distances would be expected to be smaller than the between-group distances. We define the clonality score as the quantile of within-group distances that corresponds to the 0.01st quantile of the between group distances. Thus, it measures the excess of closely related strains sharing the gene (or its absence) relative to the proportion of closely related strains with differing gene presence/absence status. The median score over all genes present in approximately half of the strains was used as the final summary as these are the most informative about recombination events (if a gene is very rare or common, the chance of seeing it donated is low). The *linkage score* for a pair of genes is defined as follows: the distances between the strains are calculated using sequences for each gene independently. The Spearman correlation of the distances between the genes is taken as the linkage score for the gene pair. In the absence of homologous recombination, the distances are expected to be highly similar, resulting in a high linkage score.

When fitting the gene content component, the similarity between real and simulated data was measured using

$$d_1 = -\log KL - \frac{1}{2s_{real}^2}\left(c_{simu} - c_{real}\right)^2,$$

where c_{real} is the median clonality score over genes having frequency 0.4-0.6 in the real data, s_{real}^2 is the variance of the median clonality score obtained by bootstrapping, c_{simu} is the corresponding median clonality score in the simulated data, and KL is the Kullback-Leibler divergence between the real and simulated frequency histograms. To account for sampling, the simulated histogram was computed by averaging over histograms for 30 bootstrap samples of 616 strains (the number of strains in the real data), sampled from the 2000 strains simulated. Before computing the KL-divergence, the gene frequency distributions were discretized into 7 bins using boundaries: (0, 0.02, 0.05, 0.2, 0.95, 0.98,1) and a single bin for genes with frequency exactly 1. Thus, the bins simultaneously captured all main characteristics of the frequency distribution: the proportion of the core genome, the slopes at each end of the histogram, and a bin to combine intermediate frequencies. When fitting the gene sequence component, the similarity between real and simulated data was measured using

$$d_2 = -\log((s_{simu} - s_{real})^2) - \log((l_{simu} - l_{real})^2),$$

where s_{simu} and s_{real} are the slopes of the distance distribution in the simulated and real data sets and l_{simu} and l_{real} are the median linkage scores between all core gene pairs in the simulated and real data sets.

2 Detailed Results

2.1 Gene frequency histogram

Results from a model fitted by matching the gene frequency histograms and the clonality scores between the real and simulated data sets are shown in Supplementary Figure 6. The histograms were obtained by running the model for 40,000 generations, and combining the results at a 1,000 generation interval after discarding the first 10,000 generations, which yielded approximately the same number of genes as observed in the real data. The figure shows a simulated histogram with the optimized parameter values and illustrates the impact of each parameter on the results.

The results show that the overall gene content distribution can be fitted accurately by modifying only three parameters: novel gene introduction rate, deletion rate and horizontal gene transfer (recombination) rate. Intuitively, increasing the rate at which novel genes are introduced in the population has a major impact on the proportion of genes present in a small proportion of strains. Furthermore, the deletion rate influences the ratio of the number of genes present in all strains (the core) and the number of genes present in almost all strains, as the former become the latter through deletions. Recombination rate has a minor impact on the gene frequency histogram. On the other hand, the clonality score increases from 0.046 with a high recombination rate to 0.52 with a low recombination rate, with the fitted rate yielding a clonality score equal to 0.11 (the value in the real data is 0.12).

The main visually detectable quantitative difference between the real data and the optimized model is that the proportion of genes with frequency between 98 and 100%, i.e., genes that are almost core, corresponding to the rightmost grey column, is slightly higher in the real data (6%) than in the optimized model (4%). The fit of this aspect could be improved by increasing the deletion rate; however, this would lead to an excess of other high-frequency genes (see the panel with high deletion rate in Supplementary Figure 6). One possible reason for the larger proportion of the 'almost core' genes in the real data is that some of them are actually core, but have not been annotated as such due to inconsistencies in the gene prediction algorithm's output.

As an example of a real data feature related to the gene content, not compatible with the model assumpations, rare genes (present in 2-4 strains) were typically found in closely related strains in the simulation, as a result of inheritance from a common ancestor, but not in the real data (Supplementary Figure 7). A detailed inspection revealed that many rare genes had originated through frameshift mutations (not included in the model) and the proportion of frameshifts among genes found in distant strains was significantly higher (58%) than among genes found in closely related strains (34%, p=3.7e-6). Note that our data were treated to remove likely false positive gene predictions (see Methods in the main text).

An important difference between our model and previous models is the inclusion of within-population HGT events, which we have shown to play a central role in generating the observed distribution of accessory loci. Another recent model has assumed genes can be donated from one strain to another (Baumdicker and Pfaffelhuber, 2013). The key difference in our new model is that HGT may lead not only to an acquisition, but also to a deletion of a gene, which is biologically motivated. This avoids the problem related to an excess of genes at intermediate to high frequencies not seen in real data

(Baumdicker and Pfaffelhuber, 2013), which follows when each HGT event increases the number of genes in the population. With our formulation, within-population HGT is not expected to change the gene frequency histogram, because the number of donor-recipient pairs resulting in a gene deletion is equal to the number of pairs leading to a gene acquisition. This explains the success of the previous models to explain the gene frequency distribution without HGT (Baumdicker et al., 2012; Collins and Higgs, 2012; Haegeman and Weitz, 2012; Lobkovsky et al., 2013). One consequence of our formulation is that rare genes are deleted with a higher frequency than commons genes, because the number of strains that can 'donate the absence' of the gene is higher. This is connected with recent results indicating that low frequency genes have different acquisition and deletion rates than other genes (Collins and Higgs, 2012); however, as discussed above, many rare genes seemingly transfer faster than expected even by our model.

2.2 Population structure

The fitted recombination rate resulted in a population structure with multiple strain clusters approximately equally distant from each other (Supplementary Figure 8), yielding the best match with the key features of the overall population structure observed in the real data, in which 16 sequence clusters were detected (Croucher et al., 2013). However, some aspects of the population structure were not accurately captured by the simple model. For example, the separate small mode in the top-right corner and the peak close to the origin in the heat map (Figure 1) seem not well accommodated by the fitted model, although the model assigned some probability mass to these regions also. The separate mode suggests that the corresponding sequence cluster 12 may have a lowered ability to recombine with the rest of the population. The peak close to the origin, on the other hand, means that strains in some sequence clusters are more closely related to each other than expected by neutral variation. A more detailed inspection of the real data revealed different sequence clusters to have different distance distributions, highlighting the fact that the fitted model was obtained by averaging over many independent evolutionary processes (Supplementary Figure 9).

In an attempt to gain a better understanding of mechanisms that could underlie the peak near the origin in the distance distribution, we experimented with two simple extensions of our model, compatible with the available background information (Croucher et al., 2013). In the first extension, a geographically structured sample was taken from the whole population, to account for the observed relatedness of strains from the same location and sequence cluster. We included each sampled strain multiple times following the joint distribution of sampling sites and sequence clusters in the real data, thus representing an upper bound on the effect achievable by the geographic structuring. In the second extension, a bottleneck was simulated on the population, acting as a simple proxy for other processes by which some strains produce more offspring than others, such as periodic selection and selective sweeps (Fraser et al., 2009). Example outputs from the extensions demonstrate the capability of both of them to explain some of the peak while leaving the main mode intact (Figure 2 and Supplementary Figure 10). However, we emphasize that the real data likely represent an outcome of many of the processes acting continuously and in conjunction, and with varying relative importances and timescales within different sequence clusters. Actually fitting the models, and selecting between them and more complicated alternatives would require explicit quantitative characterization of the differences between the sequence clusters, which is beyond the scope of this work.

The effect of recombination can be understood as follows: on one hand, it acts as a diversifying force for closely related strains, as the strains acquire recombinations from other distant strains. On the other hand, recombination prevents a strain from diverging further than the average strain distance by mixing genes between the strains. As a result of the two forces acting in opposite directions, the distance distribution ends up consisting of a single mode, clearly separate from the origin. Moreover, once recombination rate is sufficiently high for the mode to emerge, further increase does not change its location. This can be understood by noticing that recombination does not remove variation at any single locus, it only shuffles it across the strains. Because the strain distances are obtained as an average over distances at individual loci, it follows that after shuffling the different loci, the pairwise strain distances become concentrated around the mean locus-wise distance. The essentially same principles apply in the directions of both the axes, although details differ: in the direction of the x-axis (core genome distance), homologous recombinations are shuffling the variation in gene sequences caused by mutations. In the direction of the y-axis (gene content distance), the variation in gene content caused by deletions and introductions of novel genes is mixed between the strains by horizontal gene transfers.

By comparing the distance distributions between the three years in which the strains were sampled

(2001, 2004, 2007), one can test the conclusion that the mode in the distance distribution represents a stationary property of the population. Indeed, the distributions seems highly similar (Supplementary Figure 11), as expected on the basis of the results from the model; however, we note that mapping the time-scales between simulation and real data is not straightforward.

2.3 Sensitivity analyses

Our model does not assume separate core and accessory genomes, but the core emerges stochastically when genes become fixed. For comparison, we investigated assuming part of the core 'stable', i.e., deletion of these loci recuded fitness to zero, leaving no descendants. The results show that models with less than 30% of the core stable (of the whole core) could be fitted approximately equally well to the real data (Supplementary Figure 12). The fit decreased when the proportion of stable core was increased beyond 30%, when no parameter combination was able reproduce the frequency distribution adequately. In detail, the core ended up too large, and the proportion of common accessory genes with frequency between 50% to 100% too small when many stable core genes were assumed (Supplementary Figure 13). This result is in contrast with a recent estimate that genes in the stable category would account for approximately 80% of the core genome, which was obtained by fitting a model assuming one stable (essential) and two accessory gene categories to the frequency histogram (Collins and Higgs, 2012); however, the histogram was based on 14 *S. pneumoniae* genomes only.

We included in our model a multiplicative fitness penalty, equal to 0.99, for each gene beyond a pre-specified limit for the number of genes. Changing this parameter in the range from 0.95 to 0.999 does not affect the gene frequency histogram (Supplementary Figure 14), and, consequently, also the location of the mode in the gene content Jaccard distance is unaffected. The only measurable difference was that the average genome size increased from 0.95 to 1.02 relative to the limit for the genome size, and development of further summaries is required to formally fit the parameter. In the analyses the number of strains in the population was equal to 2000. Neither decreasing this to 1000 nor increasing to 4000 affects the main conclusions (Supplementary Figures 15 and 16). The only notable difference is a minor decrease in the overall variation in the population over time with respect to increasing population size, as expected, resulting in smaller variance in the summaries.

References

Baumdicker, F., W. R. Hess, and P. Pfaffelhuber. 2012. The infinitely many genes model for the distributed genome of bacteria. Genome Biology and Evolution 4:443–456.

Baumdicker, F., and P. Pfaffelhuber. 2013. The infinitely many genes model with horizontal gene transfer. arXiv preprint arXiv:1301.6547 .

Collins, R. E., and P. G. Higgs. 2012. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Molecular Biology and Evolution 29:3413–3425.

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

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

Fraser, C., W. P. Hanage, and B. G. Spratt. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480.

Gourieroux, C., and A. Monfort. 1997. Simulation-based econometric methods. Oxford University Press.

Haegeman, B., and J. S. Weitz. 2012. A neutral theory of genome evolution and the frequency distribution of genes. BMC Genomics 13:196.

Lobkovsky, A. E., Y. I. Wolf, and E. V. Koonin. 2013. Gene frequency distributions reject a neutral model of genome evolution. Genome Biology and Evolution 5:233–242.

McFadden, D. 1989. A method of simulated moments for estimation of discrete response models without numerical integration. Econometrica 57:995–1026.

Pakes, A., and D. Pollard. 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57:1027–1057.

Rasmussen, C. E. 2006. Gaussian processes for machine learning. Citeseer.

Wood, S. N. 2010. Statistical inference for noisy nonlinear ecological dynamic systems. Nature 466:1102–1104.

3 Supplementary Figures

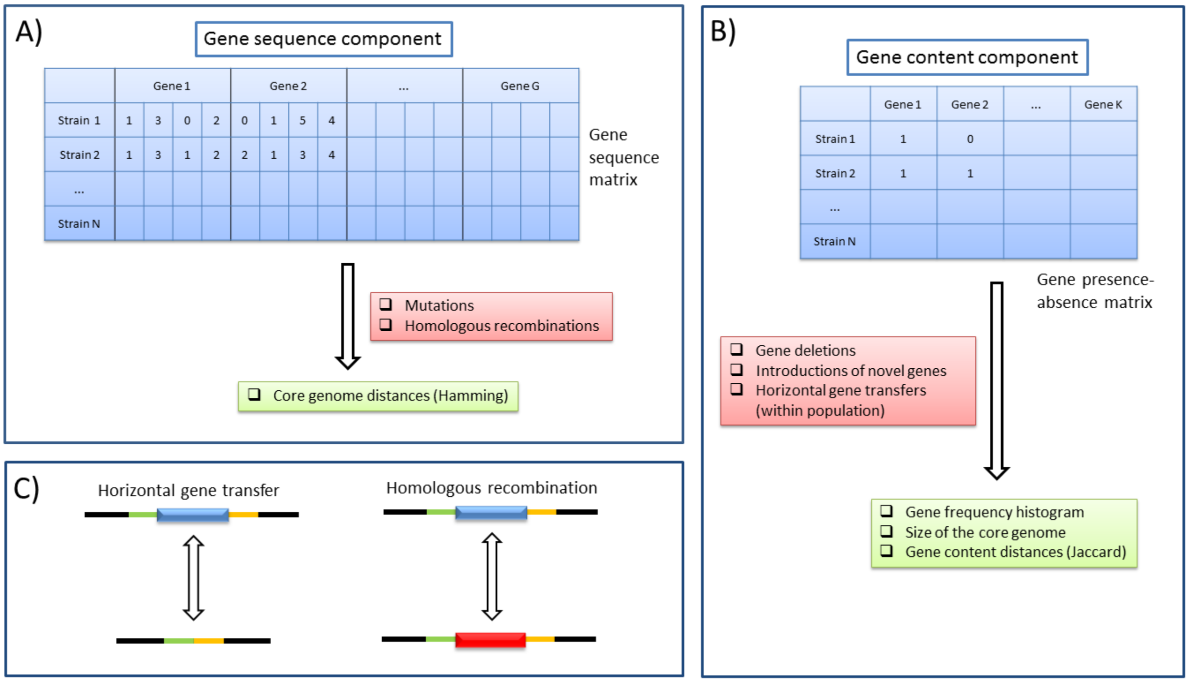

Figure 1: Schematic illustration of the model. A/B) Gene sequence/gene content components of the model. The related data structures are colored blue, the evolutionary forces acting on the components red, and the outputs derived from the components green. C) Recombination events implemented in the model. In a horizontal gene transfer, a sequence encompassing a gene (blue rectangle) may replace a sequence in another strain without the gene, or *vice versa*. In a homologous recombination an allele (blue rectangle) is replaced by another allele (red rectangle) of the same gene, or *vice versa*.

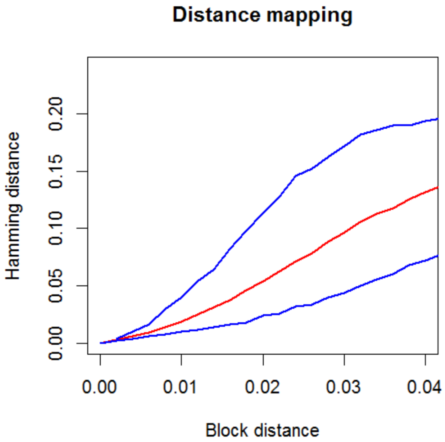

Figure 2: Mapping of sequence distances from low-dimensional sequence representation to Hamming distances resulting from the same number of mutations following the Jukes-Cantor model. The mapping was derived from results obtained by simulating the low-dimensional and the full model in parallel multiple times. The red line shows the mean of the distribution, blue lines the 5th and 95th percentiles.

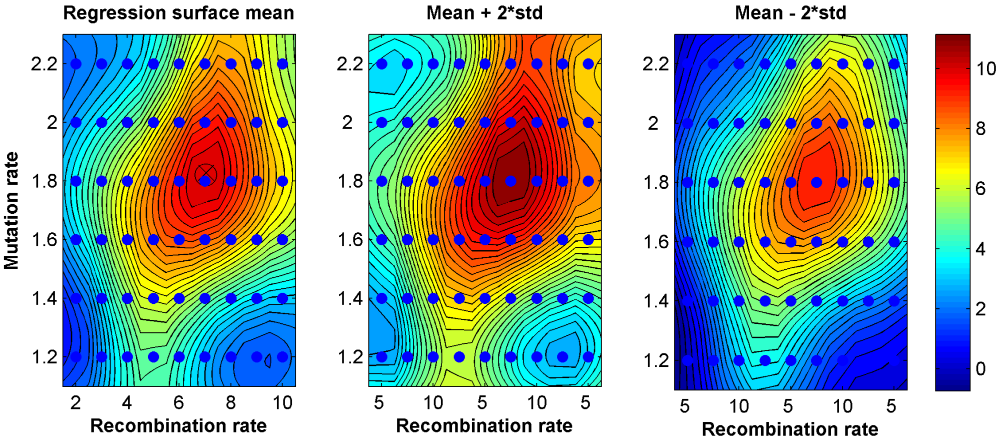

Figure 3: A distribution for the similarity scores between the simulated and real data. The similarity is computed using formula $-\log((s_{simu} - s_{real})^2) - \log((l_{simu} - l_{real})^2)$, where s_x refers to the slope of the gene content vs. core genome distance distribution, l_x to the median linkage score over all core gene pairs, and the subscript x specifies whether real or simulated data is in question. The blue dots show the parameter combinations at which the simulations were run and the cross denotes the location of the optimal parameter value.

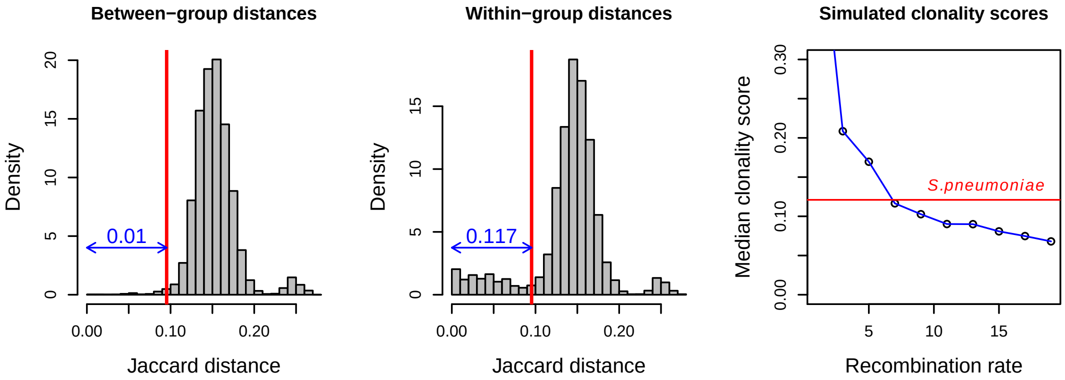

Figure 4: Illustration of the clonality scores. A gene selected randomly from the real data divides the strains into two groups, those with and without the gene. The first and second panels show the between-group and within-group gene content (Jaccard) strain distances. The quantile of the within-group distances corresponding to the 0.01st quantile of the between-group distances is defined as the clonality score of the gene (here 0.117). The last panel shows how the median clonality score (computed over genes with frequency 40-60%) varies in a simulation as a function of horizontal gene transfer rate.

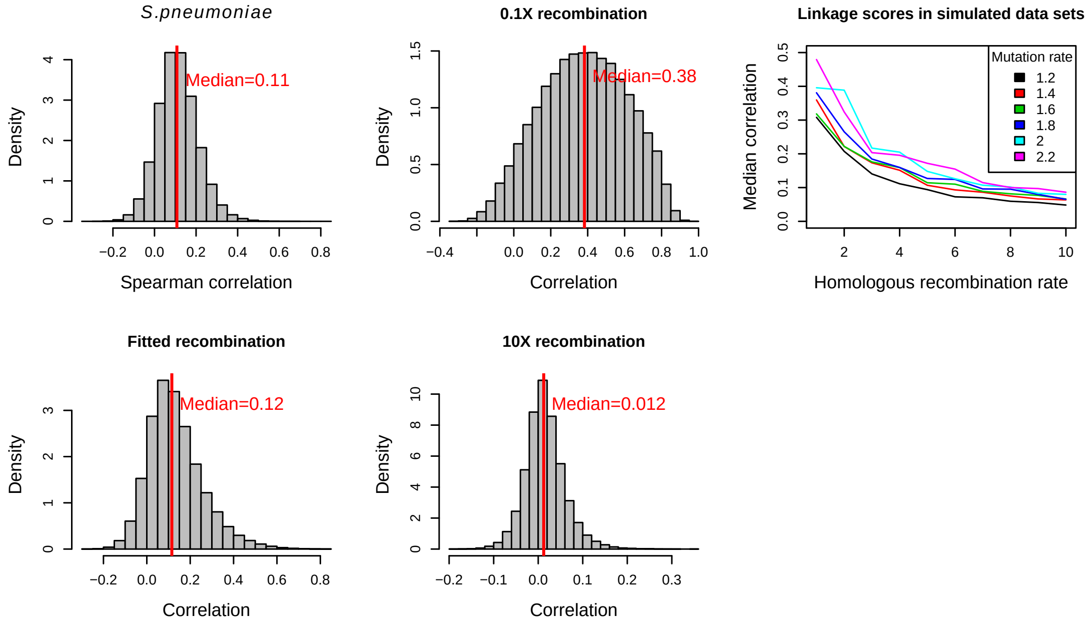

Figure 5: Illustration of the linkage scores. A linkage score for a pair of core genes is defined as the Spearman correlation of strain Hamming distances computed separately at the two genes. The panels show the distribution of linkage scores in the real data, and in simulated data sets with fitted/decreased/increased homologous recombination rates. The panel on the right shows how the median linkage score over all core gene pairs varies in the simulations as a function of the homologous recombination and mutation rates. To investigate the sensitivity of the median score to different levels of variation observed in different genes in the real data, we re-computed the median score after removing all genes with less than 16 or 27 SNPs (10th and 20th percentiles of the SNP count distribution). The median score changed from 0.108 to 0.112 to 0.116, indicating a negligible impact on the recombination rate estimate.

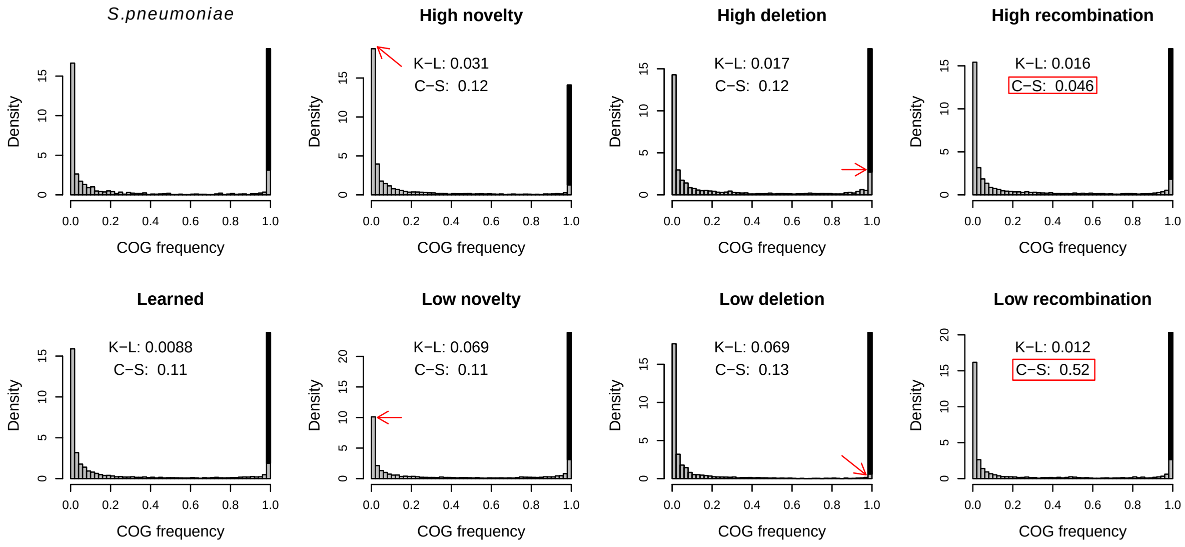

Figure 6: Gene frequency distributions. The x-axis shows the proportion of strains in which a gene is present, such that rare genes appear on the left, and common genes on the right. The black column represents genes present in all strains, *i.e.*, the core genome. Individual panels, in columnwise order, show results for the *S. pneumoniae* data, simulated data with fitted parameter values, simulated data with increased/decreased novel gene introduction rates, simulated data with increased/decreased deletion rates, and simulated data with increased/decreased homologous recombination rates. Kullback-Leibler (K-L) divergence between the real and simulated histograms, and the clonality score (C-S) are shown for the simulated results (the clonality score for the real data was approximately 0.12). The read arrows and boxes highlight features most affected by the respective parameters.

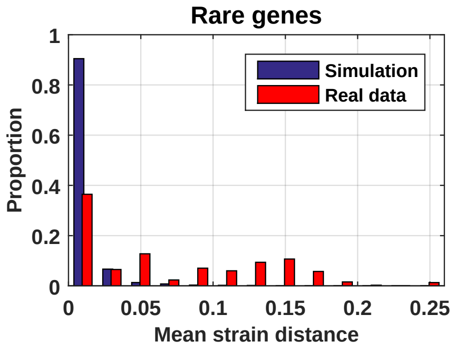

Figure 7: Mean gene content (Jaccard) distances between strains sharing a rare gene. The results show that strains sharing a rare gene are usually closely related to each other in a simulation (blue). In the *S. pneumoniae* data, a relatively large proportion of rare genes are found in distant strains (red).

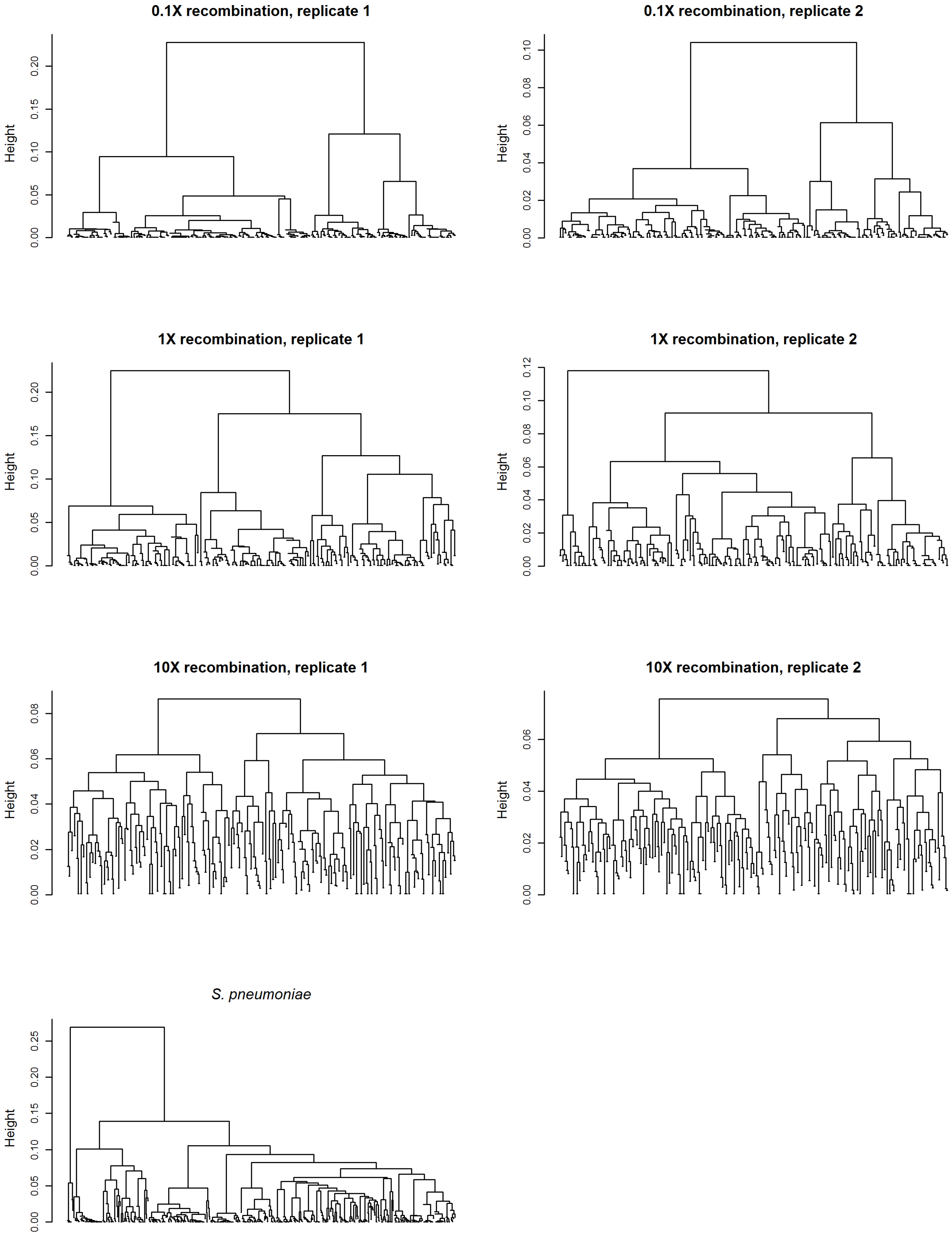

Figure 8: Simulated phylogenetic trees. The trees are based on core Hamming distances, and estimated using the simple complete linkage hierarchical clustering. Results for the fitted (1X), decreased (0.1X), and increased (10X) recombination rates are shown, along with the *S. pneumoniae* data for reference. The long branch in the tree for the real data separates strains in the divergent sequence cluster 12 from other strains. The characteristics of the tree with the decreased rate include dense clusters in the ends of long branches. On the other hand, the increased rate corresponds to a tree with star-tree like characteristics, with none of the strains very close or distant from the other strains. The fitted rate results in a compromise between the two extremes.

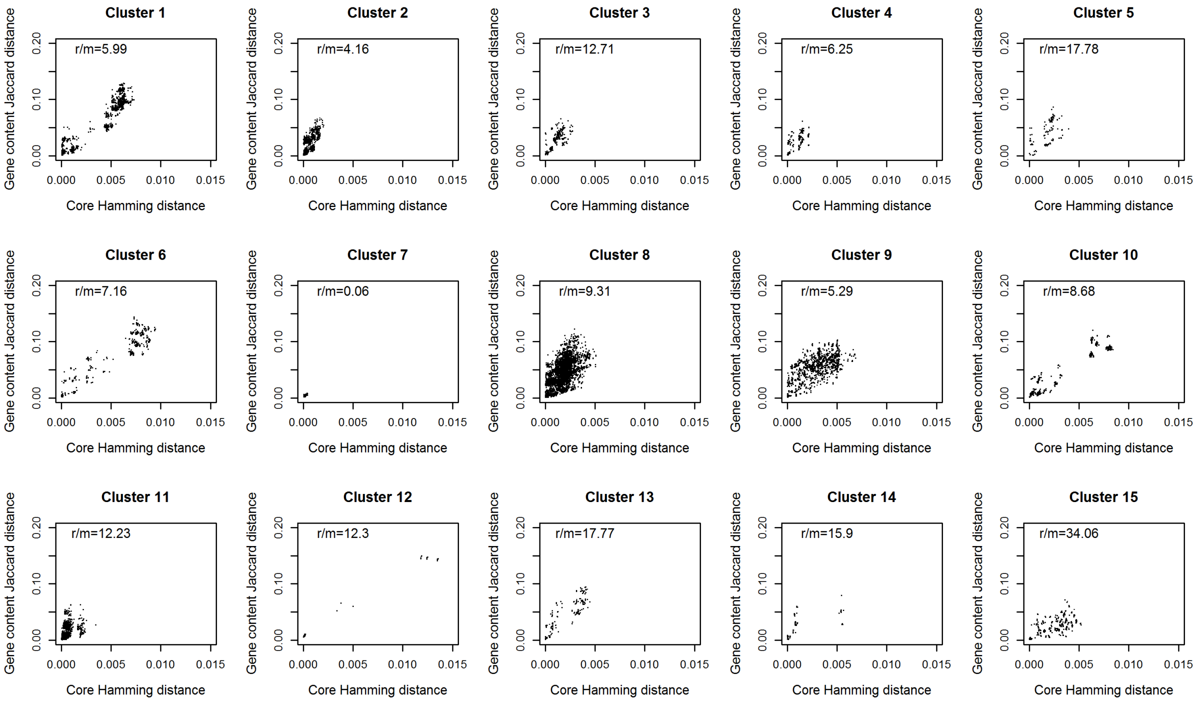

Figure 9: Gene content Jaccard distances vs. core genome Hamming distances for strains within different sequence clusters in the *S. pneumoniae* data set. The r/m values in the panels show estimates for the numbers of substitutions introduced by recombinations vs. mutations in the sequence clusters, and are taken from Croucher *et al.* (2013). No apparent relation between r/m and the shape of the distribution seems to exist.

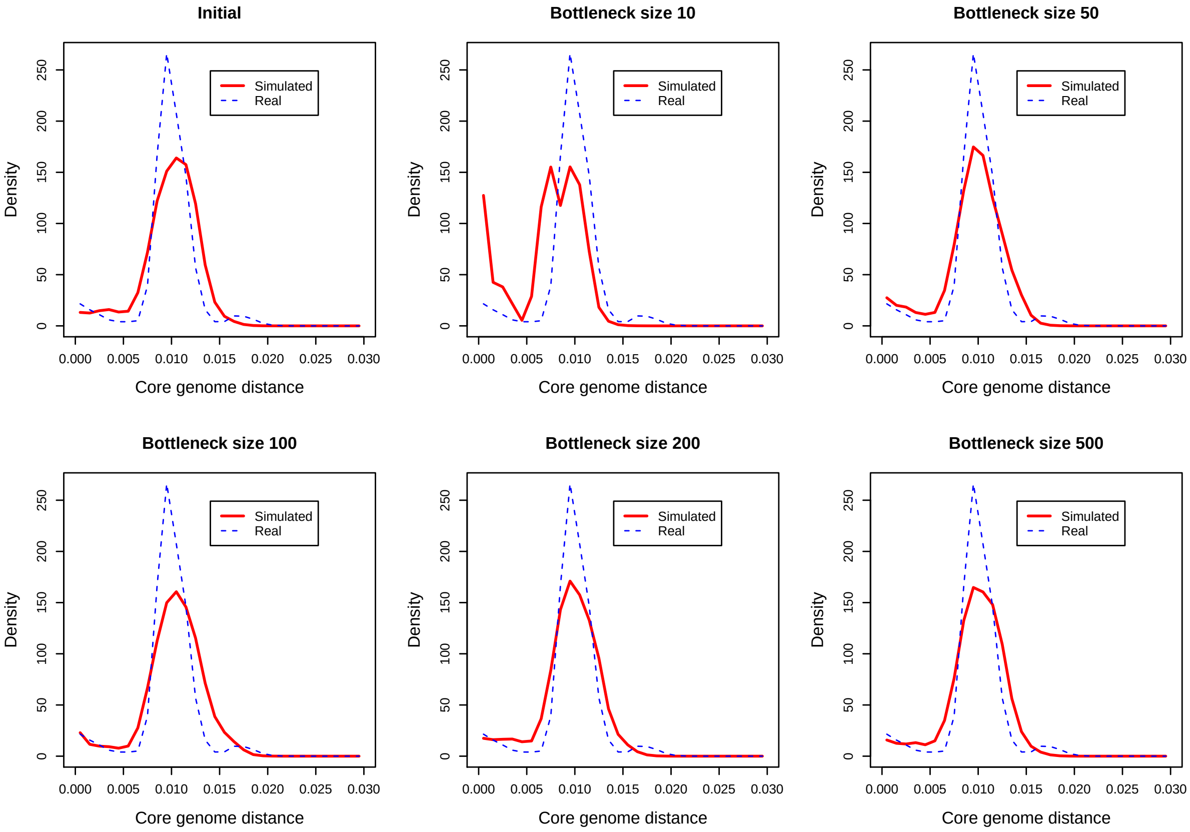

Figure 10: Effect of a population bottleneck on the core genome Hamming distance distribution. Strains from a simulated generation, representative of the average shape, were selected as the initial population. A bottleneck was simulated by selecting a specified number of strains (out of 2,000 strains in total) as possible ancestors from which the next generation was sampled with replacement. The bottleneck with size 100 seems to produce the most similar peak near the origin to the one observed in the real data.

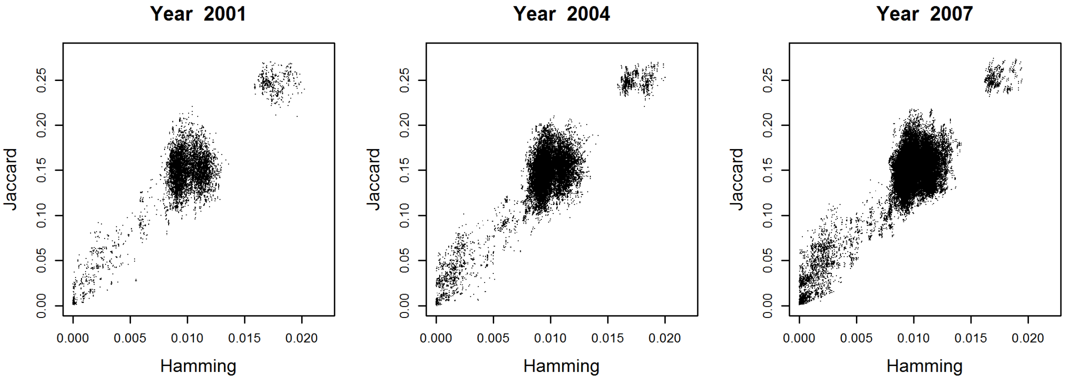

Figure 11: Gene content Jaccard distances vs. core genome Hamming distances for strains sampled in different years in the *S. pneumoniae* data set. The numbers of isolates in the different years were: 133 (2001), 203 (2004), and 280 (2007).

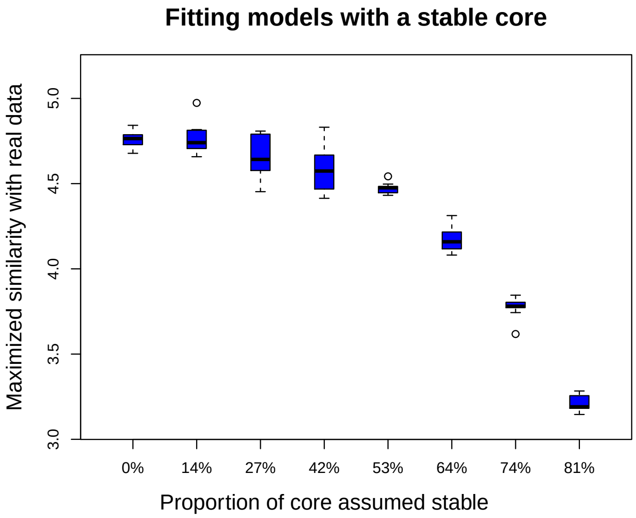

Figure 12: Maximized similarity scores between simulated and real data for different proportions of stable core genome out of the whole core genome. Different numbers of stable core genes were assumed, and the model was optimized 10 times independently. The boxplots show the similarity scores between the real data and the optimized models in the 10 optimization rounds.

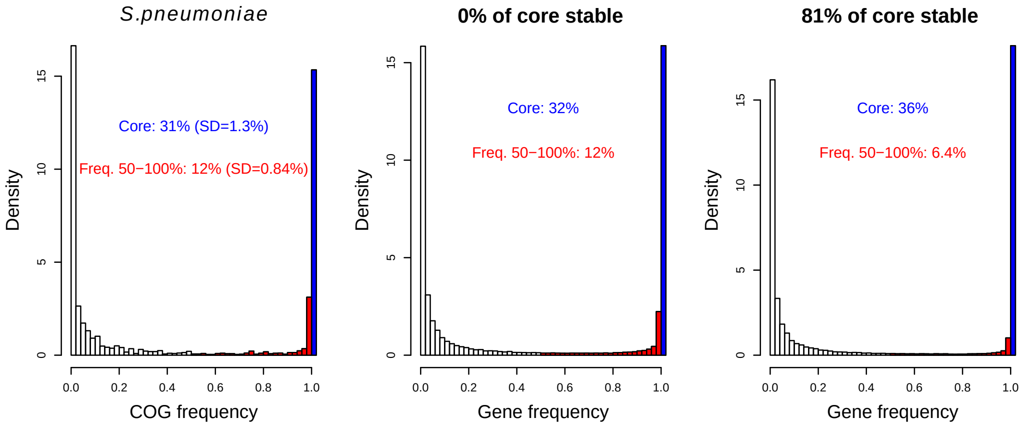

Figure 13: Effect of assuming a stable core on the gene frequency histogram. The left-most histogram shows for reference the frequencies in the real data. The other two histograms show fitted histograms, averaged over 10 optimization rounds, from two different models, one assuming no stable core, the other assuming that on average 81 per cent of core (out of the whole core) is stable. Additional annotation in each panel specifies: the proportion of genes that are present in all strains, i.e. the core (blue), proportion of non-core genes that are present in 50-100 per cent of the strains (red).

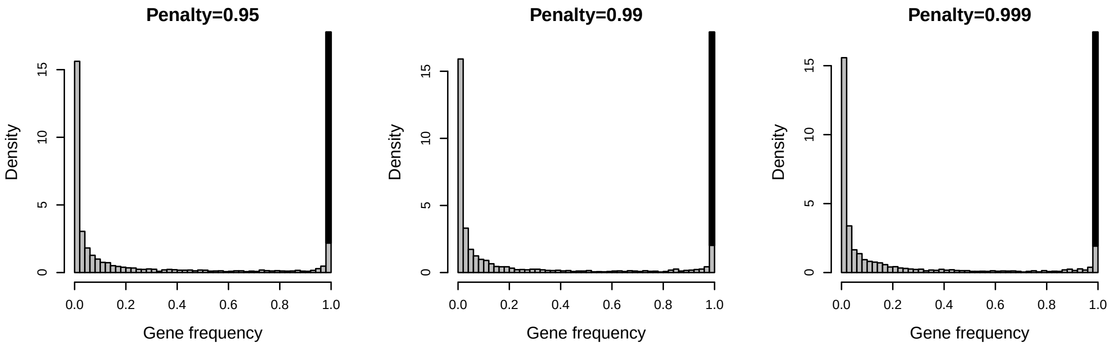

Figure 14: Gene frequency histograms when the fitness penalty for increasing the number of genes is changed in the range from 0.95 to 0.999.

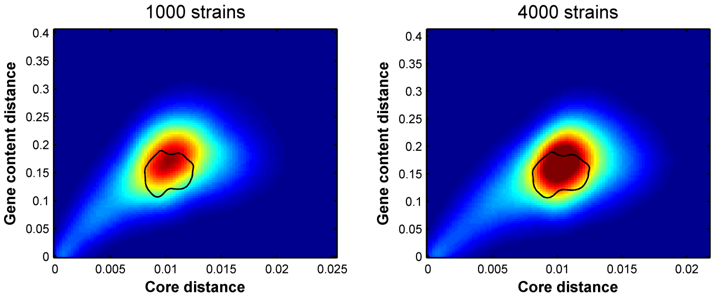

Figure 15: Impact of the number of simulated strains on the distance distributions.

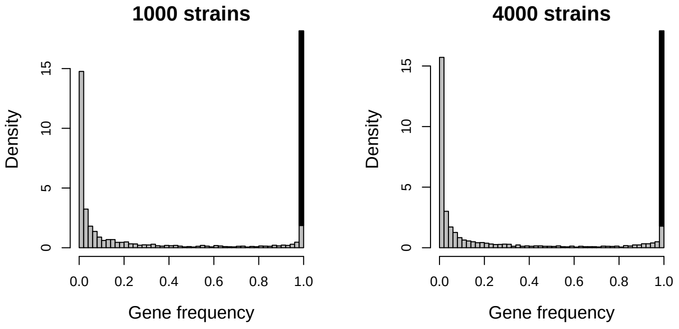

Figure 16: Impact of the number of simulated strains on the gene frequency histograms.

4 Supplementary Tables

Parameter	**Description**	**Fitted value**
deletion.rate	Mean number of gene deletions per generation (relative to the size of the core genome).	0.066
novel.gene.introduction.rate	Mean number of introductions of novel genes per generation (relative to the size of the core genome).	0.18
horizontal.gene.transfer.rate	Mean number of horizontal gene transfer attempts per generation per gene.	7.4
mutation.rate	Mean number of mutations per generation per gene (gene sequence component).	1.8
homologous.recombination.rate	Mean number of homologous recombination event attempts per generation per gene (gene sequence component).	7.0

Table 1: Evolutionary parameters in the model.

Parameter	**Description**	**Value**
num.strains	Number of sequences simulated.	2000
sequence.component.size	Number of genes for which detailed evolution is simulated in the gene sequence component.	40
genome.size	Number of genes that can be present in the gene content component in a strain without fitness cost.	60
fitness.cost.per.gene	Fitness cost per excess gene applied to strains which have the number of genes larger than *genome.size*	0.99
rec.acceptance.par	A recombination attempt is accepted with probability 10^{-Ax}, where A is *rec.acceptance.par*, and x is the local sequence divergence calculated over the gene affected by the recombination (homologous recombination) or, when the local divergence is not available in the horizontal gene transfers (full sequences not simulated), the overall Jaccard distance between the donor and the recipient (mapped to the corresponding Hamming distance).	18
gene.length	The length in base pairs of a gene for which detailed evolution is simulated	500

Table 2: Simulation meta-parameters

5 Supplementary Animation Captions

Animation 1: Evolution of the strain distance distribution with fitted parameter values. The simulation was run for 25,000 generations and the animation was created by plotting the distance distribution at 100 generation interval. Fig. 1e of the main text was created by averaging over the generations, after discarding the first 10,000 generations. (*animation_fitted_recombination.avi*)

Animation 2: Evolution of the strain distance distribution with between strain recombination rate multiplied by a factor of 1/10. (*animation_0.1X_recombination.avi*)

Animation 3: Evolution of the strain distance distribution with between strain recombination rate multiplied by a factor of 10. (*animation_10X_recombination.avi*)