PLOS GENETICS

RESEARCH ARTICLE

Long-term evolution of *Streptococcus mitis* and *Streptococcus pneumoniae* leads to higher genetic diversity within rather than between human populations

Charlotte Davison[1], **Sam Tallman**[1¤], **Megan de Ste-Croix**[1], **Martin Antonio**[2,3,4], **Marco R. Oggioni**[1,5], **Brenda Kwambana-Adams**[2,6,7,8], **Fabian Freund**[1], **Sandra Beleza**[1]*

1 Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom, **2** Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine, Fajara, The Gambia, **3** Centre for Epidemic Preparedness and Response, London School of Hygiene & Tropical Medicine, London, United Kingdom, **4** Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom, **5** Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, **6** Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom, **7** Malawi Liverpool Welcome Programme, Blantyre, Malawi, **8** Division of Infection and Immunity, University College London, London, United Kingdom

¤ Current address: Genomics England, London, United Kingdom
* sdsb1@leicester.le.ac.uk

Abstract

Evaluation of the apportionment of genetic diversity of human bacterial commensals within and between human populations is an important step in the characterization of their evolutionary potential. Recent studies showed a correlation between the genomic diversity of human commensal strains and that of their host, but the strength of this correlation and of the geographic structure among human populations is a matter of debate. Here, we studied the genomic diversity and evolution of the phylogenetically related oro-nasopharyngeal healthy-carriage *Streptococcus mitis* and *Streptococcus pneumoniae*, whose lifestyles range from stricter commensalism to high pathogenic potential. A total of 119 *S. mitis* genomes showed higher within- and among-host variation than 810 *S. pneumoniae* genomes in European, East Asian and African populations. Summary statistics of the site-frequency spectrum for synonymous and non-synonymous variation and ABC modelling showed this difference to be due to higher ancestral bacterial population effective size (N_e) in *S. mitis*, whose genomic variation has been maintained close to mutation-drift equilibrium across (at least many) generations, whereas *S. pneumoniae* has been expanding from a smaller ancestral bacterial population. Strikingly, both species show limited differentiation among human populations. As genetic differentiation is inversely proportional to the product of effective population size and migration rate (N_em), we argue that large N_e have led to similar differentiation patterns, even if *m* is very low for *S. mitis*. We conclude that more diversity within than among human populations and limited population differentiation must be common features of the human microbiome due to large N_e.

OPEN ACCESS

Citation: Davison C, Tallman S, de Ste-Croix M, Antonio M, Oggioni MR, Kwambana-Adams B, et al. (2024) Long-term evolution of *Streptococcus mitis* and *Streptococcus pneumoniae* leads to higher genetic diversity within rather than between human populations. PLoS Genet 20(6): e1011317. https://doi.org/10.1371/journal.pgen.1011317

Editor: Xavier Didelot, University of Warwick, UNITED KINGDOM

Received: November 22, 2023

Accepted: May 23, 2024

Published: June 6, 2024

Data Availability Statement: The S. mitis genomic dataset was deposited in NCBI-SRA (https://www.ncbi.nlm.nih.gov/sra) and are available under accession numbers provided in Supplementary Table 6. S. pneumoniae genomes were collected from the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/browser/home), available under the study accession numbers: PRJEB2357 (Asian dataset), PRJEB2417 (European dataset), and PRJEB3084 (African dataset). The reference genomes sequence used for the read mapping are available from GenBank (S. mitis NCTC 12261, NCBI accession: NZ_CP028414.1, GenBank accession: GCA_000960005.1; and S. pneumoniae R6, NCBI accession: NC_003098.1, GenBank accession: AE007317.1). Scripts for running the ABC computation and the Ne calculations are available at https://github.com/fabfreund/strepto_demography.git.

Funding: This work was supported by MRC-UK MR/M01987X/1 awarded to SB; by BBSRC-UK and University of Leicester funded Midlands Integrative Biosciences Training Partnership (MIBTP), grant number BB/M01116X/1, to CD; University of Leicester Doctoral Program Training Fellowship to ST. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

The genetic variation of human-associated bacteria and the evolutionary mechanisms leading to that variation are crucial for the establishment of highly contextual interactions with their host, giving rise to phenotypes such as virulence. Here, we studied *Streptococcus mitis* and *Streptococcus pneumoniae*, who share a common ancestor but evolved lifestyles on different ends of the pathogenic potential spectrum: *S. mitis* is mainly commensal, whereas *S. pneumoniae* has high pathogenic potential. Genomic variation of worldwide healthy-carriage strains is considerably higher for *S. mitis*. We show this to be due to a larger ancestral population in *S. mitis*, whose population effective size (N_e) and mutation gain and loss (by genetic drift) have been kept stable over the generations. In contrast, *S. pneumoniae* population has expanded from an ancestral population with smaller N_e. We also observe low genetic differentiation among populations in both species. We deduce that this is due to large N_e which, even with limited dispersal rate as in *S. mitis*, leads to significant effective dispersal among populations. As both species' properties of N_e and dispersal are common among human-associated bacteria, more diversity within rather than among human populations and limited population differentiation must be common features of the human microbiome.

Introduction

There is evidence that healthy carriage of human-associated bacteria have coevolved with humans over thousands of years [1–3], and that this has generated significant bacterial genetic diversity among hosts [4]. Reconstructions of within-species single nucleotide variants (SNVs) from metagenomic samples from the gut showed that bacterial strain diversity in human populations is globally consistent with theoretical expectations of long-term evolution by stochastic fluctuation of allele frequencies over the generations (genetic drift) and purifying selection [4,5]. However, these studies have also uncovered a wide range of genetic variation among species. This is reflective of independent evolutionary histories enabling bacteria with different ecological trajectories. Importantly, these evolutionary histories have resulted from interactions with humans that range from bacteria with a stricter commensal lifestyle to others exhibiting high pathogenic potential. Characterization of the genetic variation of human-associated bacterial species on different ends of the pathogenic potential spectrum and the evolutionary mechanisms leading to that variation is, therefore, crucial for understanding bacterial adaptation to their host environments, as well as the evolution of phenotypes such as virulence.

An interesting case of phylogenetic related species that show contrasting interactions with humans are the two upper respiratory tract inhabitants *Streptococcus mitis* and *Streptococcus pneumoniae*. *S. mitis* is one of the most prevalent species of the oropharyngeal microbiome [6,7], where it resides as a commensal. Because of its abundance at birth and throughout life, *S. mitis* is an excellent model for comprehensive analysis of variation and diversification of the oral microbiome. Compositional similarity between mother and child indicates vertical (familial) transmission of the oral microbiome [8]. However, early investigations into genetic diversity of *S. mitis*, although impeded by difficulties in distinguishing related species within the Mitis phylogenetic group [3,9], found multiple genotypes in new-born infants both matching and not matching parental genotypes, suggesting some degree of horizontal transmission of the Mitis group including *S. mitis* [10,11]. This is corroborated by recent oral metagenome studies that discovered a median sharing of 32% of oral strains between cohabitating individuals, and 3% between non-cohabiting individuals in the same population [12]. High *S*. *mitis* diversity within and between hosts is supported by analyses of the highly differentiating *gdh* gene, of a combination of multi-locus sequence alignment (MLSA) genes, and of a small number of genomes [13–15]. These patterns were associated with their lower rate of transmission between unrelated and non-cohabitating individuals leading to lineage isolation and, consequently, to preclusion of potential homogenization of the mitis gene pool due to homologous recombination [14,15]. According with this view, discrete clusters of genetic variation tracking host genomic variation across populations should be observed, for instance in a similar manner to the stomach-colonizer and vertically transmitted *Helicobacter pylori* [16]. Recent studies have shown this to be the case for gut microbiome species [17], although the correlation between host and commensal genetic diversities seems to be weak [18–20]. The apportionment of genetic diversity of human commensals within and between human populations is still a subject of debate, and the factors that contribute to bacterial population differentiation need further consideration. This is important because the amount of population variation within and among populations contribute to the evolutionary potential of a species.

On the other hand, *S*. *pneumoniae* is a frequent colonizer of the nasopharynx that is associated with high human mortality and morbidity worldwide [21]. Although nasopharyngeal colonization is usually asymptomatic, it is considered an essential step preceding invasive and non-invasive pneumococcal disease [22,23]. In addition, carriage serves as a source of pneumococci that can be transmitted within and between households. The rate of horizontal transmission between households is known to be high in *S*. *pneumoniae* [24,25], which will influence the level and distribution of genetic variation in the overall population of *S*. *pneumoniae* across host populations. Therefore, pneumococcal carriage is a key stage in the evolution of this organism. However, *S*. *pneumoniae* population genetic variation has been mostly characterised in relation to invasive disease [26–34] or in the context of changes after vaccine introduction [35–43], and few recent studies address the genetic variation of carriage *S*. *pneumoniae* within a population [44–49].

Differences in transmission rates between *S*. *mitis* and carriage *S*. *pneumoniae* (considerably higher for *S*. *pneumoniae*) are expected to lead to more geographic isolation of the former in comparison to the latter, and this is expected to lead to the observation of higher geographic structure in *S*. *mitis* than in *S*. *pneumoniae*. However, under the standard neutral model, genetic differentiation among populations is inversely proportional to the product of effective population size and migration rate ($N_e m$; [50]). This means that in populations with large N_e, even in cases where m is minimal, effective migration will overcome genetic drift (proportional to $1/N_e$) leading to low long-term population differentiation.

Here, we sought to gain a better understanding about the evolution and population differentiation of healthy carriage of *S*. *mitis* and *S*. *pneumoniae* across human populations. We collected bacterial samples from European, East Asian and African hosts, which compose the major divisions of human genetic diversity [51], in this way, sampling across deeper evolutionary time scales. We observe that the apportionment of genetic diversity is higher within populations than between populations, and that this is related with the large N_e for both species. However, genomic variation is considerable higher for *S*. *mitis* due to a bigger ancestral bacterial population that has been maintained close to mutation-drift equilibrium across (at least many) generations; in *S*. *pneumoniae*, contemporary diversity is inferred to be due to a population expansion from an ancestral population with smaller N_e. Both species have been evolving under purifying selection, but there are signatures of diversifying selection which have not led to high intermediate-frequency alleles.

Results

S. mitis has considerably higher within-host diversity than carriage *S. pneumoniae*

We collected a representation of within- and between-host genomic variation of healthy colonisation of *S. mitis* and *S. pneumoniae* across major human population groups. The *S. mitis* dataset consists of a total of 101 newly collected and 18 publicly available isolates from 46 independent hosts from Africa (n = 17 hosts/32 isolates), East Asia (n = 6 hosts/37 isolates), and Europe (n = 24 hosts/50 isolates). More than one isolate was collected from 18 hosts, ranging from 2 to 10 isolates per host. The African dataset included 20 isolates from five pairs of related individuals (two mother-child pairs, two sibling pairs and one sibling trio), whereas the European and East Asian isolates were obtained from unrelated individuals only.

To obtain a sample of the true general diversity of carriage *S. pneumoniae* populations that is not biased by the action of strong recent selection, the *S. pneumoniae* dataset was obtained from published pre-vaccination studies [42,44,46] and comprises a total of 810 isolates from asymptomatic hosts (carriage isolates) from Africa (n = 90 hosts /230 isolates), East Asia (n = 480 hosts/isolates) and Europe (n = 100 hosts/isolates). The African dataset included within-host genomic variation, with the number of isolates collected from 71 hosts ranging from two to five. In this dataset, the total number of serotypes, sequence types and Global Pneumococcal Sequence Clusters (GPSCs) of shared evolutionary history as previously defined [52,53] is 63, 253 and 135.

Genetic diversity and presence of clonal relationships within the host, within families and between hosts were assessed for both species by calculating the number of SNV differences between every pair of isolates in the total dataset and within each of those three levels of host relatedness. For *S. mitis*, the mean number of pairwise SNV differences in the total sample is 30,190 (Interquartile range, IQR: 29,050–32,580). We observed clonal relationships in ~one percent of total pairwise comparisons (<1000 SNVs), mostly from within-the host, but also between related individuals (two sibling pairs) and between unrelated individuals' isolates collected in the same geographic region (five pairs; S1A Fig). Pairwise SNV comparisons within the host (mean = 24,366; IQR: 25,298–30,358) and between related hosts (mean = 25,411; IQR: 24,822–32,769) are slightly smaller than pairwise SNV comparisons between unrelated hosts (mean = 30,531; IQR: 29,189–38,379) (Kruskal-Wallis test, p-value$\leq$ 7.08x10^{-5}; S1C Fig).

In *S. pneumoniae*, the presence of clonal relationships was detected both within and between hosts at a larger degree than in *S. mitis* (S1B and S1D Fig). The mean number of pairwise SNV differences between serotypes is 7,309 (IQR:6,310–7,748), similar to the mean number of pairwise SNV differences within serotypes (mean = 7,265, IQR: 5,348–10,481). Considering genomic variation in carriage *S. pneumoniae* organised in GPSCs, the mean number of pairwise SNV differences is greater between (7,393; IQR: 6,319–7,866) than within the defined clusters (1,880, IQR: 128–3,113). Analysing the within-the-host pairwise SNV differences distribution in the African sample, we observe that ~97% of the pairwise comparisons within GPSCs and within serotypes are smaller than 100; the remaining pairwise SNV comparisons are above the first quartile of between-host pairwise SNV differences. We then set a threshold of 100 pairwise SNV differences to define clonal relationships in *S. pneumoniae*, which maintains the serotype (n = 62), sequence type (n = 247) and GSPC clusters (n = 134) diversity across all geographic regions (Fig 1B). However, we also performed pangenome and core-genome diversity and population structure analyses with the more conservative thresholds of 1000 (as in *S. mitis*; 0.8% percentile of the pairwise SNV distribution), 2000 (1% percentile of pairwise SNV distribution) and 5600 (5% percentile of pairwise SNV distribution)

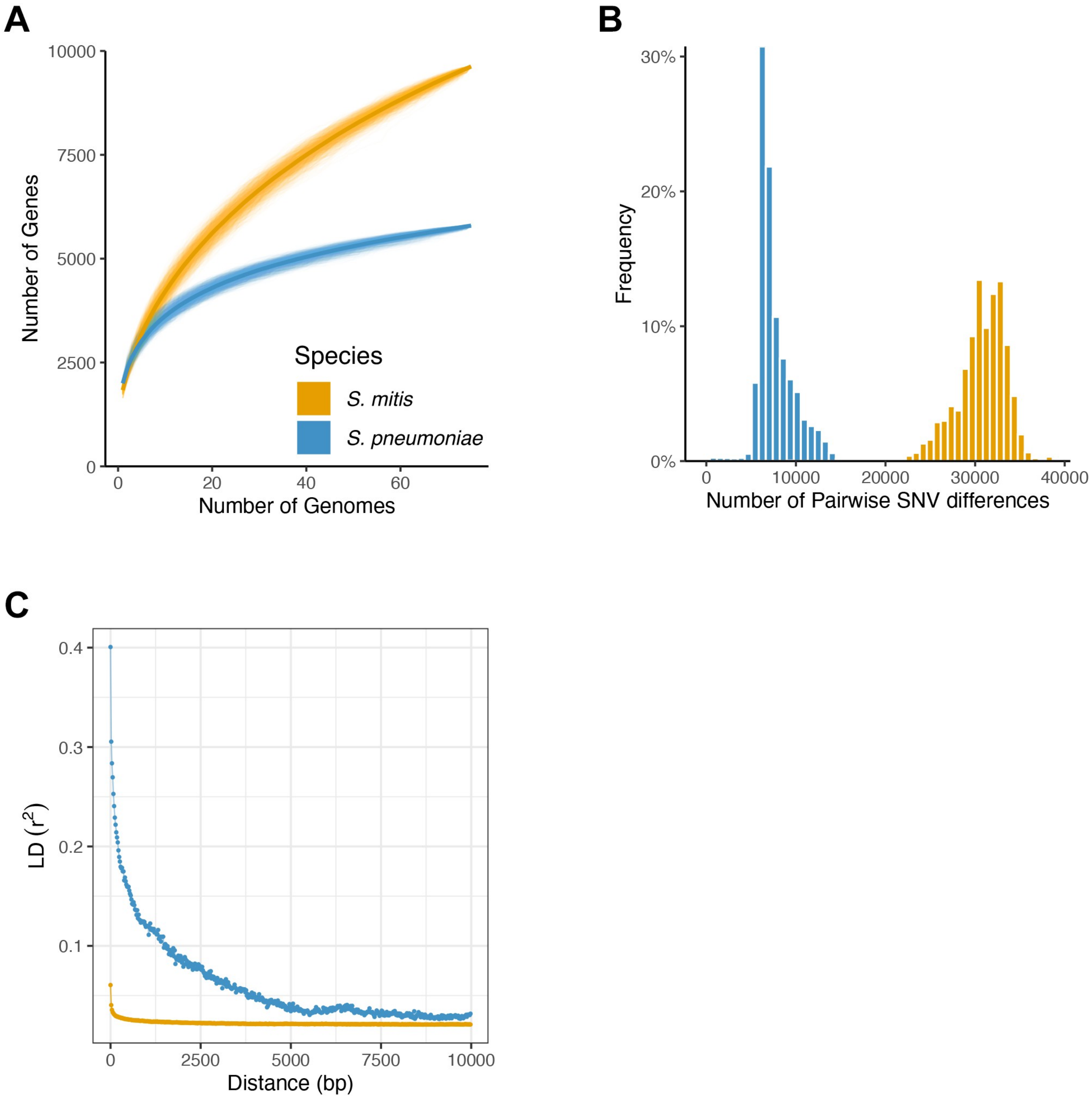

Fig 1. *S. mitis* and *S. pneumoniae* genetic variation. A. *S. mitis* and *S. pneumoniae* pangenomes according to power law fit. The number of genes is plotted as a function of the number of genomes. In *S. mitis*, the order of the 75 unrelated genomes was permuted 1000 times. In *S. pneumoniae*, the same procedure was applied to each of 1000 random samples of 75 unrelated genomes. Power law regression was fitted to the mean number of genes obtained across all permutations. Fitting the same model to the median gave similar results, albeit with lower goodness-of-fit to the *S. pneumoniae* data. The parameterisation that best fitted the data was: $Y = aX^b + c$. For *S. mitis*, $b = 0.42$; for *S. pneumoniae*, $b = 0.12–0.32$; mean = 0.2. **B.** Pairwise SNV differences distribution for *S. mitis* and *S. pneumoniae* after controlling for clonality (pairwise SNV differences threshold of 1000 for *S. mitis* and of 100 for *S. pneumoniae*). **C.** LD (r^2) decay with distance. *S. mitis*, orange; *S. pneumoniae*, blue.

https://doi.org/10.1371/journal.pgen.1011317.g001

pairwise SNV differences, to confirm that results and main conclusions are not affected by closer relatedness of few pairs of isolates (S2 and S3 Figs).

Considering the established cut-off for clonality for both species (1000 pairwise SNV differences for *S. mitis*; 100 pairwise SNV differences for *S. pneumoniae*), we observe higher within-host diversity for *S*. *mitis*: 64 out of 242 (26%) and 155 out of 236 (66%) within-host pairwise comparisons are defined as clonal relationships for *S*. *mitis* and *S*. *pneumoniae*, respectively. We can conclude that human hosts are generally colonised by multiple and divergent strains of *S*. *mitis*, whereas *S*. *pneumoniae* populations within the host are most often dominated by a single strain.

As we are interested in investigating global population genetic structure and long-term evolutionary dynamics of both species, we extracted the maximal datasets composed solely by unrelated strains (that is, datasets composed by isolates whose pairwise SNV differences are bigger than the cut-offs for clonality defined above). In total, 75 *S*. *mitis* isolates (25 African, 32 European, 18 Asian) from 45 hosts, and 353 *S*. *pneumoniae* isolates (78 African, 68 European, 207 Asian) from 339 hosts from across the three geographic regions were used in the analyses.

S. mitis has considerably higher between-host diversity than carriage *S. pneumoniae*

Phylogenetic analyses of geographically restricted (mainly European) *S*. *mitis* and mainly invasive *S*. *pneumoniae* genomes have shown higher diversification of *S*. *mitis* than *S*. *pneumoniae* [15]. Here, we analyse if these patterns of genome variation for both species are observed in our extended dataset of exclusively carriage isolates from across geographically distant regions, which corresponds to analysing the scale of diversity in these species associated with host ancestry and with deeper evolutionary timescales.

We first evaluated the sets of genes available to both species for their evolutionary success. The pangenome–defined as the total set of genes observed across all sampled isolates—and mean individual genome sizes are similar across the three geographic regions for *S*. *mitis* and *S*. *pneumoniae* (S1 Table). To compare the pangenome and genome sizes between the two species, we have analysed 1000 random samples of unrelated *S*. *pneumoniae* with the same size as *S*. *mitis* (n = 75) and present the mean estimates. Individual genome size is smaller for *S*. *mitis*–on average 1832 genes (SD = 84) for *S*. *mitis* and 2005 genes (SD = 71) for *S*. *pneumoniae* (Mann-Whitney test, *P*-value$< 2.20 \text{ x}10^{-16}$). However, *S*. *mitis* gene repertoire is significantly larger than that of *S*. *pneumoniae*. In a total of 75 strains, the *S*. *mitis* pangenome was estimated to harbour 9626 genes, of which ~10% were found in all strains (core genes, 951 genes). At the same level of resolution, *S*. *pneumoniae* pangenome size is 6064 genes (mean across 1000 random samples; SD = 250), of which 18% comprise core genes (mean ± SD = 1092 ± 29). Furthermore, the impact of each additional genome on the size of the pan genome is greater for *S*. *mitis* than *S*. *pneumoniae* (Fig 1A). This is independent of any pairwise SNV cut-off considered to define clonality for *S*. *pneumoniae* (S2 Fig). Analysing permuted data with power law regression [54], we estimated that each *S*. *mitis* genome adds on average 2.8 times more genes to the pangenome than any *S*. *pneumoniae* genome (Fig 1A).

Based on the core genomes extracted for each species, nucleotide diversity (π) within species, estimated accounting for core genome length and sample size, was similar across the three geographic locations: for *S*. *mitis*, π values for the core genome are 0.034, 0.032 and 0.033 for the African, Asian and European bacterial populations of isolates, respectively; for *S*. *pneumoniae*, π values for the core genome are 0.008, 0.010 and 0.008 for the African, Asian and European bacterial populations of isolates, respectively. π values obtained for *S*. *pneumoniae* when considering higher thresholds of pairwise SNV differences were very similar (S2 Fig). Considering the species as a whole, π is higher in *S*. *mitis* (π = 0.034) than in *S*. *pneumoniae* (π = 0.010) (Fig 1B compares the distribution of pairwise SNV differences between species). *S*. *mitis* is therefore considered to have substantially higher genetic diversity than its close relative pneumococcus in terms of both pangenome content and sequence variation.

Limited population differentiation in both *S. mitis* and carriage *S. pneumoniae*

Phylogenetic analysis and PCA show both species to have little genetic differentiation among geographic locations (Figs 2, S3 and S4). Concordant with these analyses, between bacterial population divergence values in *S. mitis* and *S. pneumoniae* among geographic regions are >3-fold lower (average Hudson's F_{ST} = 0.043 in *S. mitis*; 0.039 in *S. pneumoniae*; S2 Table)

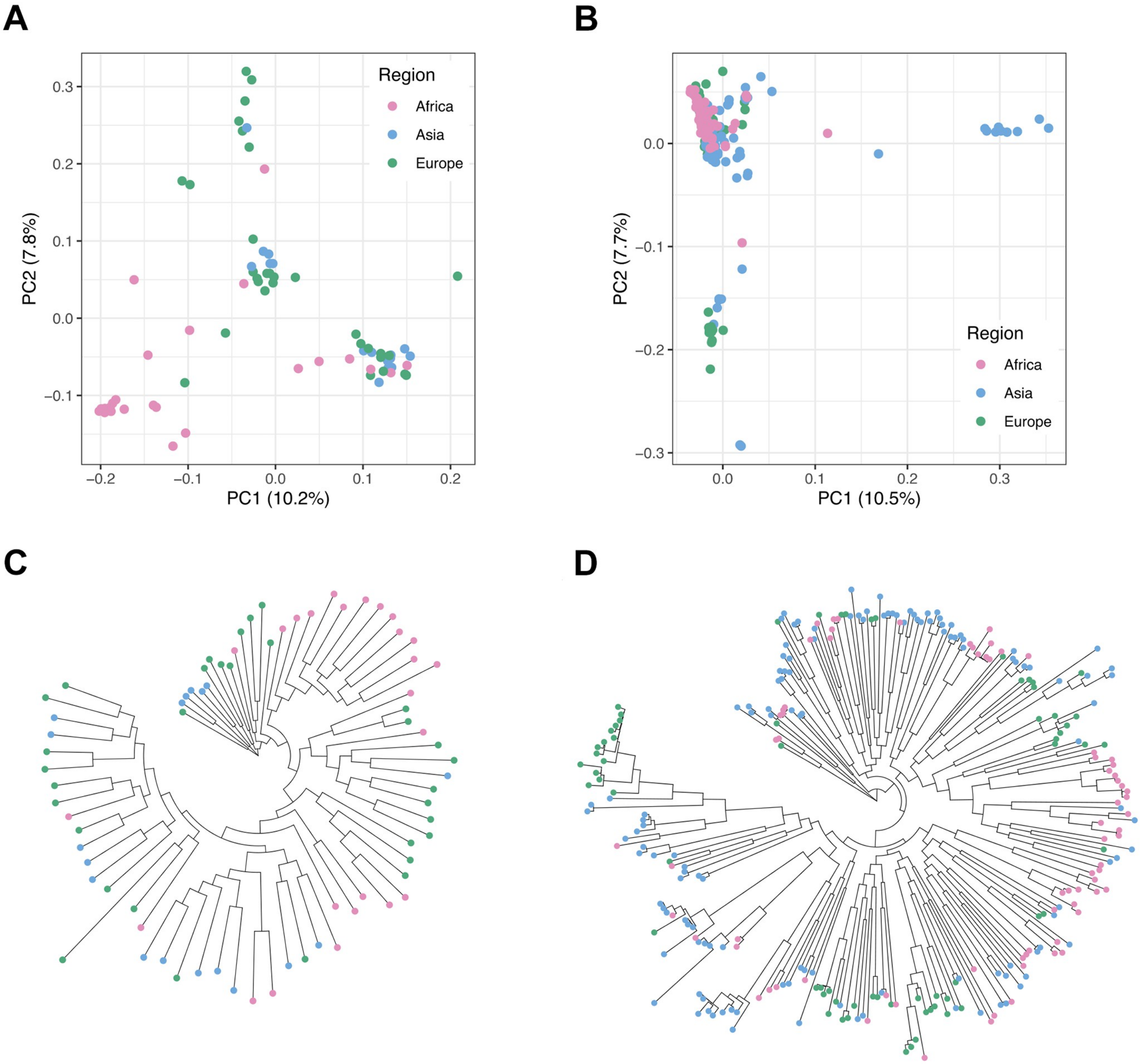

Fig 2. Population genetic structure in *S. mitis* and *S. pneumoniae*. A. plots of PC1 versus PC2 for *S. mitis*. **B.** Plots of PC1 versus PC2 for *S. pneumoniae*. **C.** Maximum likelihood unrooted phylogenetic tree for *S. mitis*. **D.** Maximum likelihood unrooted phylogenetic tree for *S. pneumoniae*. Analyses presented here for *S. pneumoniae* do not include outlying serotype NT isolates, whose higher genetic variation has been attributed to higher recombination rates for this clade [44]. Graphical representation of the analyses for the full *S. pneumoniae* sample are in S4 Fig. The two phylogenetic trees are in the same scale. Colour code corresponds to geographic region: Africa, pink Asia, blue; Europe, green.

https://doi.org/10.1371/journal.pgen.1011317.g002

than that in humans from the same geographic regions (average Hudson's F_{ST} = 0.135; [55]). Analysis of molecular variance confirms that most of the genetic variation in *S. mitis* and *S. pneumoniae* is segregating within rather than between geographically distinct bacterial populations: the within-population component explains 95.3% and 96.3%, whereas the among-geographic-regions component explains 4.7% and 3.7% of the genetic variation, respectively for *S. mitis* and *S. pneumoniae*. One caveat in our analysis is that our *S. mitis* Asian sample was mainly collected in the UK and, therefore, might include one or two genomes that are closer to European genomes via horizontal transmission [12,56], lowering F_{ST} values between these two bacterial populations. However, these will not affect our overall observation of more within-population rather than between-population diversity.

We can conclude that, although the two species conform to contrasting population models of transmission and dispersal, both *S. mitis* and *S. pneumoniae* genomic variation is mostly randomly assorted according to geography and host ancestry. The lack of strong population differentiation in both species means that the evolutionary history of the isolates can be modelled by assuming that isolates form a single, non-structured population.

Genetic diversity difference between *S. mitis* and *S. pneumoniae* is not due to mutation or recombination

The considerably higher level of *S. mitis* genetic diversity in comparison to *S. pneumoniae* can be attributed to deterministic (mutation and recombination) and population level processes (genetic drift and selection) which dictate the frequency of the allelic variation in the population. To evaluate the roles of mutation and recombination in driving bacterial variation in *S. mitis* and *S. pneumoniae*, we used the coalescence analytical framework implemented in mcorr [57], which is appropriate for species composed of highly divergent strains and for which the underlying phylogeny is difficult to estimate accurately. In addition, the estimation of evolutionary parameters in mcorr is based on synonymous variation only, which minimises the effect of selection.

Rates of recombination to mutation (δ/μ) were >1 for both *S. mitis* and *S. pneumoniae* (Table 1), showing recombination to have had a greater role in the species diversification than mutation in both species and replicating previous results for *S. pneumoniae* [39,58]. Mutational divergence (also known as effective mutation, $\theta_{\text{pool}} = \theta_{\text{pool}} = 2\bar{T}\mu$, where μ is the mutation rate, and $\bar{T}$ is the mean pairwise coalescence time across all loci in the overall population and equals $N_e/2$ in the mcorr coalescence model) is significantly higher than the sample's diversity for both species, indicating that their population gene pool is highly diverse and that recombination within species can happen between highly divergent sequences (with divergences as high as 18% for *S. mitis* and 10% in *S. pneumoniae*; Table 1).

Although mutational divergence is similar between species (see also S3 Text), recombination divergence (or effective recombination, $\phi_{\text{pool}} = 2\bar{T}\delta$, where δ is recombination rate) is considerably higher in *S. mitis* (Table 1). Recombination frequency in both species was also evaluated by analysing the decay of LD with physical distance (Fig 1C). Neutral (based on synonymous variation alone) LD r^2 values for *S. pneumoniae* decay from a maximum of ~0.3 to

Table 1. mcorr estimates of recombination and mutation parameters for *S. mitis* and *S. pneumoniae*. n, sample size; n_{genes}, number of genes analysed; d_{sample}, diversity of the sample; θ_{pool}, mutational divergence; and ϕ_{pool}, recombinational divergence of the of the species' gene pool; δ/μ, the relative rate of recombination to mutation; $\bar{f}$, the mean recombination fragment length; c, recombination coverage. A definition of these parameters is in Methods.

Species	n	n_{genes}	d_{sample}	θ_{pool}	ϕ_{pool}	δ/μ	$\bar{f}$ (bp)	c
S. mitis	75	950	0.046	0.18	0.82	4.61	2511	0.32
S. pneumoniae	353	972	0.012	0.10	0.14	1.41	1004	0.13

https://doi.org/10.1371/journal.pgen.1011317.t001

low, considered non-significant, values—less than 0.1—within 1500bps; LD r^2 maximum value is ~ 0.06 for *S. mitis* and halves within 150 bps. These values together with mcorr recombination parameters, indicate a history of frequent genetic exchange that is preventing the presence of strong population structure in both species. The low levels of neutral LD r^2 for *S. mitis* are consistent with a quasi-sexual mode of evolution for this species [5].

Importantly, neutral LD r^2 values in *S. mitis* are considerably lower and decay to minimal values faster than LD r^2 in *S. pneumoniae* (Fig 1C). In non-structured populations such as the ones under study, neutral LD between alleles is maintained by a balance between genetic drift and recombination [59]. One explanation for the observed differences in long-range LD for both species is that recombination rates are higher in *S. mitis*. However, under this hypothesis, recombination to mutation rates in serotype lineages in *S. pneumoniae* would be expected to be consistently lower than in *S. mitis*, and this pattern is not observed [58]. Therefore, the most likely explanation for lower neutral LD in *S. mitis* is long-term evolution based on lower levels of genetic drift. This hypothesis is also consistent with the genomic diversity and population structure patterns observed in *S. mitis* (see Discussion) and is compatible with its quasi-sexual mode of evolution. In summary, the difference in genetic diversity between the two species is not due to mutation or recombination.

Long-term evolutionary dynamics for *S. mitis* and carriage *S. pneumoniae*

We computed the minor-allele frequency distribution of variant sites, called the folded site-frequency spectrum ($_f$SFS; Fig 3A and 3B), independently for synonymous and nonsynonymous variation, each to evaluate the role of genetic drift (derived from demographic history of the species) and selection occurring during each species' evolutionary history, respectively [60]. Notably, because we inferred high effective recombination for both species (Table 1), we assume that selection only affects a reduced number of closely linked neutral sites around the selected variant.

S. pneumoniae shows an excess of rare synonymous (<2%) variants compared to *S. mitis* (chi-square (1df), p-value < 2.20 x10^{-16}) (Fig 3A). In contrast, a higher number of intermediate frequency synonymous variants (>5%) is observed in *S. mitis*. Conversely, the overall *S. mitis* nonsynonymous $_f$SFS is skewed towards rare variation in comparison to *S. pneumoniae* (chi-square (1 df), p-value < 2.20 x10^{-16}; Fig 3B). The distributions of synonymous variation suggest contrasting demographic models of non-growth versus population growth in *S. mitis* versus *S. pneumoniae* and the distribution of non-synonymous variation, potentially, different types and magnitudes of selection.

To test for these hypotheses, both species' $_f$SFS were summarised by calculating Tajima's D (TD) per core genome and per core gene, for synonymous (TDsyn) and nonsynonymous (TDnon) variants (Fig 3C and 3D). Overall core genome TDs were compared with simulated TDs obtained under a standard population equilibrium and selective neutral model (Methods). In *S. mitis*, TDnon is consistently smaller than TDsyn (Fig 3C). The overall TDsyn is -0.36 (p-value = 0.380), whereas TDnon is -1.91 (p-value = 0.003). This confirms that the evolutionary model most consistent with *S. mitis* genomic variation is one of a stationary population in equilibrium (at least over many generations) and natural selection acting at nonsynonymous sites. In contrast, there is a linear tendency for larger values of TDnon to be associated with larger values of TDsyn in *S. pneumoniae*, and both overall metrics are significantly smaller than zero (overall TDnon = -1.90, p-value = 0.001; overall TDsyn = -1.40, p-value = 0.039). This confirms that the $_f$SFS of *S. pneumoniae* is more consistent with a model of population growth influencing the scale of genetic diversity observed in *S. pneumoniae* and natural selection acting at nonsynonymous sites.

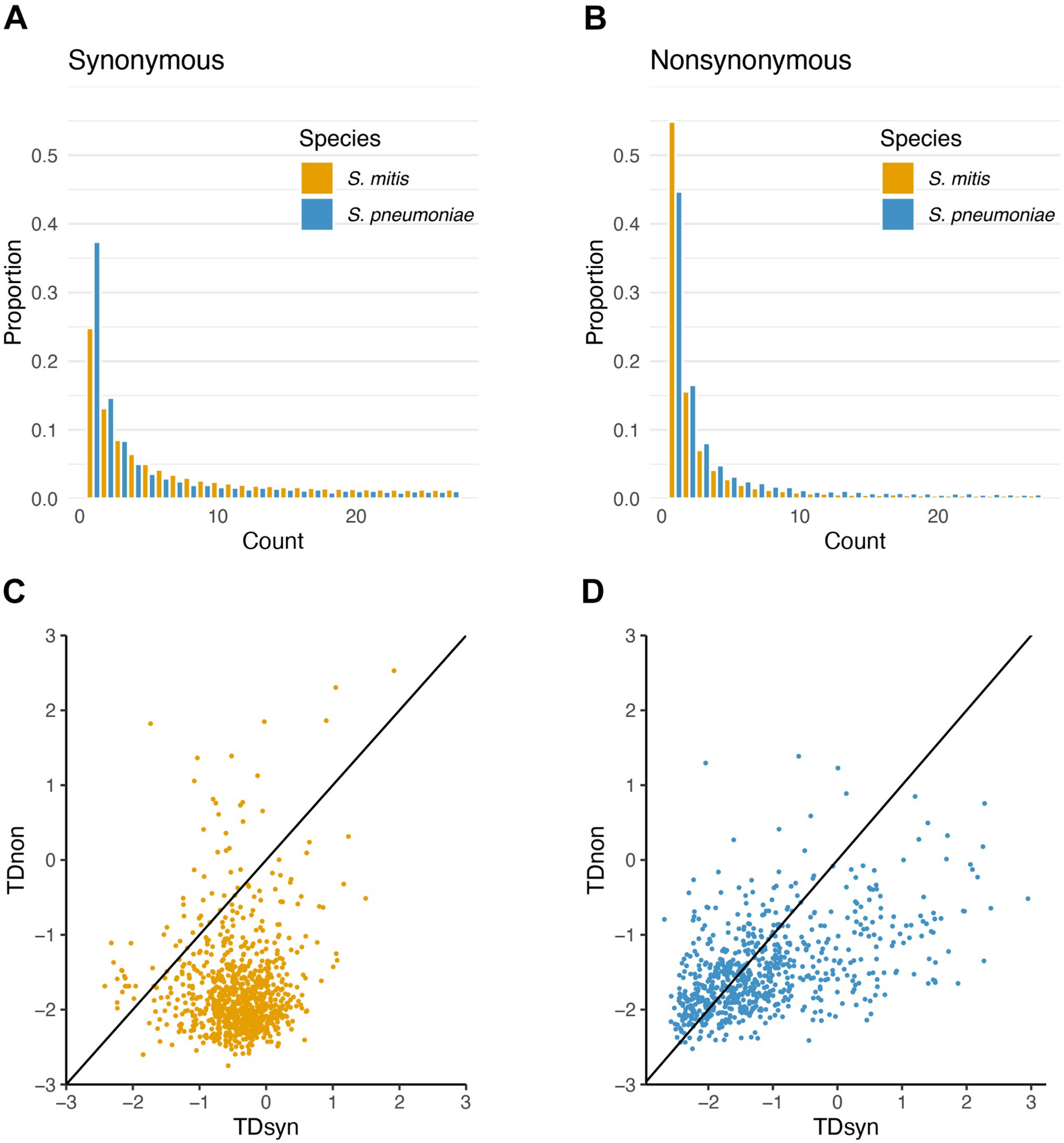

Fig 3. Evolutionary dynamics of *S. mitis* and *S. pneumoniae*. A. Folded site frequency spectrum for synonymous variation. **B.** Folded site frequency spectrum for nonsynonymous variation. To get a comparable result, the SFS was calculated based on the same sample size for both species. **C.** Tajima's D for synonymous variation (TDsyn) vs Tajima's D for nonsynonymous variation (TDnon) in *S. mitis*. Line corresponds to the diagonal (where x = y). **D.** Tajima's D for synonymous variation (TDsyn) vs Tajima's D for nonsynonymous variation (TDnon) in *S. pneumoniae*. Line corresponds to the diagonal (where x = y). *S. mitis*, orange; *S. pneumoniae*, blue.

https://doi.org/10.1371/journal.pgen.1011317.g003

We further investigated the demographic history of *S*. *pneumoniae* and *S*. *mitis* using an approximated Bayesian computation (ABC) framework. For each species, we simulated 30,000 250-Kb windows for *S*. *mitis* and 75,000 250-Kb windows for *S*. *pneumoniae* with sample size and number of synonymous (neutral) SNVs matching the ones estimated for overlapping sliding windows of the same size across the sampled genomes (overlapping by 200 Kb) and whose SFS is derived from an analytical approach [61,62] that incorporates the Kingman coalescence model with exponential growth characterized by a specified growth rate *g* (including no growth ($g = 0$), that is constant population sized). We then compared the genetic diversity obtained from the simulations with the observed genetic diversity via a Random Forest ABC classifier to get the posterior growth rate parameter distributions (see Methods and S1 Text). This approach does not model recombination (we assumed each SNV to be independent due the absence of significant LD in both species; Fig 1C); however, running ABC on genetic diversity of genomic windows simulated under an exponential growth model and incorporating the recombination rate estimated for *S*. *pneumoniae* gave similar results (S2 Text).

Our approach shows support for significant population growth in *S*. *pneumoniae* (median growth rate = 2.7; 95% confidence interval, 95% CI = [1.4–9.6], in coalescent time units of N_e; Table A in S1 Text). In contrast, the model that best fits *S*. *mitis* genetic diversity includes a very small, non-significant, (median) growth rate of 0.25 (95% CI = [0.2–0.4], Table B in S1 Text), compatible with an evolutionary model with very low growth or with one of stationary population for at least many generations.

A sliding window analysis (with sizes of 250Kb, overlapping by 50Kb, or 100Kb, overlapping by 10Kb as for our ABC analysis; S1 and S2 Texts) of TDnon and TDsyn across *S*. *mitis* and *S*. *pneumoniae* genomes showed that TDnon is consistently more negative than TDsyn (e.g. for 100Kb windows, TDnon-TDsyn ranges between [-0.84, -0.16] in *S*. *pneumoniae* and [-1.84, -1.26] in *S*. *mitis*). This overall trend is evidence for widespread purifying selection in both species leading to rarer non-synonymous alleles (in contrast with synonymous alleles), that reduce the non-synonymous nucleotide diversity in comparison to the number of non-synonymous segregating sites, in this way, leading to more negative TDnon. Of note, we do not expect directional selection to contribute meaningfully to this signature in these highly recombining species due to the absence of genetic hitchhiking.

To further investigate the role of natural selection in *S*. *mitis* and *S*. *pneumoniae*, we analysed two summary statistics of genetic diversity for all genes and compared them with null distributions obtained from sampling randomly non-synonymous and synonymous variants from across the two species' genomes (see Methods). We first analysed the ratio of the number of non-synonymous variants to the number of synonymous variants [46] and observed a significant fraction of genes with an excess (p-value<0.05; corresponding to the action of diversifying selection) and deficiency (p-value>0.95; corresponding to the action of purifying selection) of non-synonymous variation in both species (binomial test p-value<2.2e-16; S5 Fig). We also evaluated the ratio of nucleotide diversity for non-synonymous to nucleotide diversity for synonymous variation (π_N/π_S) and, in this case, both species' gene distributions show smaller non-synonymous genetic diversity in comparison to the null distribution (S6 Fig). Altogether, we can conclude that the signature of diversifying selection influencing the genetic diversity in both species is not leading to an excess of intermediate-frequency non-synonymous variants.

Long-term evolution of effective population sizes

Neutral genetic diversity within populations is determined by the population effective size and the mutation rate, a notion that is captured in the population parameter θ. As we confirmed experimentally that the mutation rate and growth cycle are comparable between the two species (S3 Text), differences in θ between *S. mitis* and *S. pneumoniae* will be due to their N_e. We have then inferred the coalescent N_e using the observed number of segregating sites for each species (which $\approx \theta E(L_n)$, where $E(L_n)$ is the expected length [in coalescent time units of N_e] of the sample's genealogy), and incorporating estimated growth rates to control for the demographic history in both species (see Methods). Considering the mutation rate from [35], this indicated that *S. mitis*'s contemporary N_e to 2.8x higher than *S. pneumoniae*'s N_e: *S. mitis* N_e = 274,288 (IQR = [268,555–279,736]); *S. pneumoniae* N_e = 96,597 ([93,851–99,167] calculated based on the median, and minimum and maximum growth rates—S2 Text). We conclude that the ancestral population of *S. pneumoniae* prior to expansion was significantly smaller than the concurrent *S. mitis* population and that *S. pneumoniae*'s population expansion was not sufficient to recover similar contemporary N_e or genetic diversity levels to *S. mitis*.

Discussion

Here, we have studied the genomic diversity, population differentiation and evolution of two phylogenetically related human-associated bacterial species, whose modes of transmission, dispersal, and interaction with host lead to different expectations in the magnitude of genomic variation (higher for *S. mitis*) and of population differentiation/structure (more significant for *S. mitis*). Indeed, we have determined higher historical N_e in *S. mitis* that has led to more substantial levels of genomic diversity in this species compared to *S. pneumoniae*, even though contemporary N_e between species is more comparable. However, both species present limited geographic differentiation across host populations contrary to our expectations based on modes of transmission. We now discuss how different patterns of natural colonization, transmission, dispersal, interactions with the host, and demographic histories have led to these populational patterns.

Contrary to *S. pneumoniae* (our data; [47]) and common species in the gut microbiome [5,63], where a dominant strain is retained over prolonged periods of the host life, our results show that *S. mi*tis exhibits multiple colonization. According to [13] diversifying lineages may coexist for long stretches of time, generating within-host variation that will be able to respond to strong selection pressures such as changes in diet or antibiotic intake. The presence of multiple highly divergent lineages within the same host makes *S. mitis* less amenable to be studied using metagenomic samples; here we developed a culture system that allows to obtain isolates for within-the host studies more effectively. Our data confirms that additional *S. mitis* lineages are acquired within households but can also be acquired horizontally within contained social networks, as recently described for the oral microbiome more generally [12,56]. Multiple colonization and some degree of horizontal transmission leads to effective recombination between strains (S1 Text), contrary to what was proposed [14].

We replicate the previously observed higher levels of genome variation in *S. mitis* than in *S. pneumoniae* [14] and extend these observations to the pangenome level. We determined that *S. mitis* variation is consistent with the most common evolutionary model for human commensals of population in mutation-drift equilibrium, with some degree of purifying selection [4,5]. The higher N_e calculated for this species leads to a highly diverse populational gene pool (S1 Text) within which horizontal gene transfer (HGT) can occur. In addition, our results go in agreement with recent evidence indicating that *S. mitis* pangenomes are shaped by cross-species HGT [14,64]. Although the *S. mitis* pangenome diversity can be explained by neutral evolution [65], future work should indicate if common accessory genes confer adaptive advantage within and across *S. mitis* populations [66].

Modelling of *S. pneumoniae* genomic diversity agrees with a population expansion as a demographic model for *S. pneumoniae*. As *S. mitis* and *S. pneumoniae* have the same host, the distinct demographic history between both species must be linked with behavioural differences. As [15] proposes, frequent horizontal transmission makes *S. pneumoniae* more dependent on having enough hosts for successful dispersal and, consequently, population growth in S. *pneumoniae* is most likely intimately linked with the increase in human population density [67]. In addition, we argue that the horizontal transmission features of continuous founding bottlenecks and recovery are analogous to range expansion processes, and these properties can lead to the same genomic patterns as an (instantaneous) population expansion if the number of 'migrants' or bacterial cells between hosts is large ($Nm>>1$, where N is the within-host bacterial population effective size; [68]). Within-host deep-sequencing showed transmission bottlenecks sizes between donor and recipient to be higher than one [47], the within-host diversity of one colonization event (defined from cultured isolates) is consistent with N_e ranging from 1 to 72 bacterial cells [46], and mice colonization experiments allowed to estimate within-host N_es of ~100 [69]. In fact, a multiple merger coalescent genealogy was previously inferred for *S. pneumoniae* [62], which is concordant with range expansion [70]. It would be interesting to understand how between-host transmission affects N_e over a great number of generations. Another aspect of *S. pneumoniae*'s genetic diversity is that it is also due to purifying selection albeit to a lower degree than *S. mitis* (S5 Fig).

An important question that has been debated recently is the amount of commensal strain variation within and among different geographic regions or ethnic groups [17,18]. Here we show that two important components of the oro-nasopharyngeal microbiome in humans show little differentiation across human populations. In an island model under neutral equilibrium, F_{ST} is equal to $1/(1+2N_em)$ in haploids where m is the migration rate [50]. Studies on the transmissibility of the oral genome showed that strain dispersal across human populations is minimal but not null [12]. Therefore, for *S. mitis* the large N_e leads to values of $N_em >>1$, even if m is very small, reducing the differentiation among human populations. In contrast, there are reports of pathogenic outbreaks extending for long geographic distance within Europe and across the world for *S. pneumoniae* [35,71], and we detected pairs of highly related strains between continents (within our dataset 350 pairs of strains involving 125 strains from different geographic regions had <2000 SNV pairwise differences [<1%percentile]), demonstrating meaningful migration rates. Therefore, although N_e is smaller in *S. pneumoniae*, it may have been and currently is sufficiently high enough that together with higher m make $N_em >>1$. Estimates of N_e, HGT, between-host transmission and of dispersal are known to vary among commensal species, but there is evidence for a great fraction of the human microbiome that they are high enough to lead to the same patterns as in *S. mitis* and *S. pneumoniae* [4,12,56,72,73]. In conclusion, our results support the view that more diversity within than among human populations and little population differentiation must be common features of the human microbiome due to large N_e.

Material and methods

Ethics statement

African *S. mitis* samples were collected by MRC The Gambia Unit, under ethical approval granted by The Gambia Government/Medical Research Council Joint Ethics Committee.

European and East Asian samples were collected under ethics approval granted by Ethics Committee for Medicine and Biological Sciences at the University of Leicester (protocol 14610). All samples were collected from Live Participants in Leicester, UK, and all participants gave their written consent.

Bacterial isolate collection

Buccal swabs are part of an extended dataset collected from across The Gambia from family trios consisting of mother, child (3–10 years old) and a baby (less than 2 years old). Suspected streptococcal species (based on plate morphology and presence of alpha-haemolysis) were analysed via MALDI-TOF mass spectrometry, and putative Mitis group isolates had their genomes sequenced (see below). Mitis group species identification was done analysing MLSA variation as in [74]. The final dataset consisted of 32 isolates from one baby/two siblings' trio, two sibling/baby duos, two mother/child duos, and six unrelated individuals.

European samples were collected from the cheeks and tongue individuals natural and living in the UK for the past 2–5 years. East Asian samples were collected from five Chinese individuals that had lived in the UK for less than six months, maintained a Chinese diet and had either no romantic partner or a Chinese partner who met the same criteria. We complemented these datasets with publicly available European and East Asian whole genomes from NCBI-SRA (https://www.ncbi.nlm.nih.gov/sra; S3 Table), which we confirmed via phylogenetic and F_{ST} analyses that they did not differ significantly from our two sample sets.

S. pneumoniae carriage samples matching the broad geographic origin of the *S. mitis* dataset (Europe: UK, East Asia: Thailand and Africa: The Gambia) were collated from the literature [42,44,46]. Serotypes and sequence types are disclosed in respective publications and genomic sequences are available on the European Nucleotide Archive under study accessions: PRJEB2357 (Asian dataset), PRJEB2417 (European dataset), and PRJEB3084 (African dataset). We only analysed isolates from non-vaccinated individuals. For the East Asian dataset, we considered the subset of isolates collected in the first year of a 3-year sampling period, to obtain a random sample of a size comparable to the other studied geographic regions. This subsample might include more than one isolate per individual, but because we did not have access to that information, we did not consider it for within-host analyses. For the African and European datasets, we considered only the within-host carriage isolates that were collected pre-vaccination.

S. mitis isolation and growth

The European and East Asian samples were cultured on Mitis Salivarius Agar (5% Sucrose (Fisher Scientific), 1.5% Peptone (Oxoid), 1.5% agar (BioGene), 0.5% Tryptone (Oxoid), 0.4% di-potassium hydrogen orthophosphate (Fisons), 0.1% Glucose (Fisons), 7.5E-3% Trypan Blue (Sigma), 8E-5% Crystal Violet (Sigma)) [75]. Putative *S. mitis* colonies were identified based on flat, "rough" morphology and a blue colouring in the centre of the colony following 18 hours growth at 37˚C, 5% CO2, and on the sequencing of the house keeping gene *glucose-6-dehydrogenase* (*gdh*) for unambiguous differentiation between Mitis group species [13].

This isolation technique had a significantly better yield than that used for African samples: 95% of isolates predicted as *S. mitis* based on *gdh* phylogeny clustering were confirmed as *S. mitis* based on whole genome data (European and East Asian isolates) compared to 37% of isolates where species was predicted with MALDI-TOF mass spectrometry (African isolates).

DNA extraction, sequencing and *de novo* sequence assembly

DNA extraction was performed using Wizard Genomic DNA purification kit (Promega) and manufacturer's instructions for Gram-positive bacteria. Samples were sequenced on an Illumina HiSeq 4000 at the Oxford Genomics Centre (UK), to a mean coverage of 83.4x for African isolates and 300x for East Asian and European isolates.

Raw FASTQ reads were quality normalised with Trimmomatic 2.0 [76]. Quality was confirmed via FastQC 0.11.5 [77]. Trimmed, paired reads were assembled to contig level using

SPAdes 3.12 [78]. Assembly was improved to scaffold level using the “assembly improvement” pipeline from Sanger Pathogens [79]. Assembly quality was assessed via QUAST 4.3 [80]. The mean N50 across isolates was 239.6Kb (42.3Kb-931Kb) and the GC% content (39.2%-40.6%, mean = 40.1%) was mirrored across isolates. Accession numbers for this dataset are in S3 Table.

Serotyping and determination of Global Pneumococcal Sequence Clusters

Capsular serotyping and determination of Global Pneumococcal Sequence Clusters (GPSCs) of shared descent for *S. pneumoniae* [52,53] were performed by uploading newly assembled genomes in the web tool PathogenWatch (https://pathogen.watch/). Two genomes from the Asian *S. pneumoniae* dataset were classified as *S. pseudopneumoniae* and were therefore removed from the dataset.

Core and accessory genome extraction and pangenome analysis

Scaffolds were annotated using Prokka 1.14.6 with default settings [81] and the full set of non-redundant genes was extracted with Roary 3.13 [82]. Alignment was made invoking MAFFT [83] and considering a sequence identity of 85%. The percentage sequence identity was empirically determined, by considering the percentage that minimised the change in the number of genes per change in the parameter. Core genomes were formed as a concatenation of genes identified to be present in 100% of input sequences. The total core genome sizes obtained were 900,605 bp for *S. mitis* and 816,449 bp for *S. pneumoniae*.

To evaluate differences of pangenome size between *S. mitis* and *S. pneumoniae*, we performed an iterative procedure in which we permuted the input order of the bacterial genomes and assessed the rate of new genes identified per addition of each genome. For *S. mitis*, the input order of the (unrelated) genomes was permuted 1000 times. For *S. pneumoniae*, because final sample size of (unrelated) isolates was significantly larger than for *S. mitis*, we employed the same procedure to each of 1000 random samples of unrelated genomes of the same size as the one for *S. mitis*. Obtained curves were fitted applying power law regression to the mean number of number of genes obtained across permutations [54].

Phylogenetic analysis

Phylogenetic trees were built from FASTA alignments with FastTree 2.1 [84], using the generalised time-reversible model of nucleotide evolution and re-scaled based on likelihoods reported under the discrete gamma model with 20 rate categories. This is a standard approximation for accounting for variable evolutionary rates across sites and uncertainty in these rates [85]. Trees were visualised using iTOL [86].

Variant and gene mapping and annotation

Trimmed fastq files were mapped to type strains *S. mitis* NCTC 12261 (NCBI accession: NZ_CP028414.1) and *S. pneumoniae* R6 (NCBI accession: NC_003098.1) using BWA mem, and variants called using SAMtools 1.3.2 mpileup [87,88]. In order to be called a core variant, a site read depth of $>10\%$ of the mean genome-wide read depth, as well as minimum sequencing and mapping quality scores of 30, were implemented.

Synonymous and nonsynonymous variants were called in comparison to the type strains using snpEff [89]. To generate genetic diversity estimates per gene (see below), we considered the gene coordinates recorded in the genome assemblies of type strains.

Genetic diversity, population structure, and folded site frequency spectrum analyses

These analyses used biallelic variant sites only. Pairwise SNV differences were calculated with the software SNP-dists 0.6 [90]. Pairwise F_{ST} between populations and nucleotide diversity (π) of core genomes were calculated with the Python package scikit-allel [91]. We calculated Hudson's F_{ST} estimator which is less influenced by differences in sample size and SNV ascertainment scheme [55]. Pairwise F_{ST}'s are based on the set of SNVS segregating in both populations; however, considering the set of SNVs segregating in either population gave similar F_{ST} values. Analysis of Molecular Variance (AMOVA) based on the pairwise differences between sequences was performed in ARLEQUIN ver3.5.2.2 [92].

Principal component analyses (PCA) of population structure were conducted in PLINK v1.9 [93]. PLINK V1.9 was also used to determine the Minor allele frequency (MAF) of variant sites, which was used to generate the $_{f}$SFS for synonymous and nonsynonymous sites.

Estimation of mutation and recombination parameters

Recombination and mutation populational parameters were estimated using mcorr [57]. Aligned-sequenced gene fasta files resulting from our Prokka/Roary/MAFFT pipeline were converted to XMFA files [94], which were used as input files in mcorr. Only sequences with <2% gaps of the total alignment length were used in the analysis. Mcorr computes the correlation profile, P(l), in each species by averaging over the correlation profiles of synonymous substitutions for all gene sequence pairs in the sample, and then averaging over all genes. When homologous recombination is present, the probability of observing a correlated substitution (P(l)) decreases with distance between any two loci (l), at a rate that is proportional to the recombination rate. Otherwise, the function is constant. By fitting P(l) to an analytical form inferred based on a coalescence model with recombination, mcorr estimates parameters that characterize the more diverse species' gene pool with which the sampled genomes have recombined: the mutational divergence ($\theta_{\text{pool}} = 2\bar{T}\mu$, where μ is the mutation rate), the recombinational divergence ($\phi_{\text{pool}} = 2\bar{T}\delta$, where δ is recombination rate), the relative rate of recombination to mutation ($\theta_{\text{pool}}/\phi_{\text{pool}} = \delta/\mu$), and the mean recombination fragment length ($\bar{f}$), where $\bar{T}$ is the mean pairwise coalescence time across all loci in the overall population ($\bar{T} = Ne/2$) in the considered coalescence model). It also calculates the average fraction of the sampled genomes that were brought in by recombination (recombination coverage, *c*). Mcorr was implemented with default settings and 1000 bootstraps.

We also computed the decay of Linkage disequilibrium (LD) with physical distance. The LD statistic r^2 between all pairs of synonymous SNVs with minor allele frequency bigger than 0.01 was calculated up to a distance of 10 Kb (*—ld-window-k 10*) with PLINK 1.9, using the command flags*—ld-window 5000* to allow computation of r^2 between all SNVs, and*—ld-window-r2 0* to obtain the tabling of all r2 values (>0).

Per gene analyses

Nucleotide diversity (π), Watterson's θ, and Tajima's D (TD) per gene and for synonymous and non-synonymous substitutions independently using scikit-allel. We tested if the TD statistic deviated from a population equilibrium and selective neutral model by generating 10,000 random samples under this model and with the same diversity as the one observed for both species in ARLEQUIN ver3.5.2.2 [92]; p-values of the D statistic correspond to the proportion of random D statistics less or equal to the observation. If this fraction is <0.05, we concluded that the observed TD is significantly different from zero and not evolving according to the standard neutral model. We then used observed TDsyn statistics to determine the action of specific demographic processes (TDsyn<0 is expected under population growth, and TDsyn>0 under population contraction), and observed TDnon statistics to determine the action of natural selection (TDnon< 0 is expected under purifying and positive selection and TDnon>0 is expected under balancing selection) on the species' genetic variation.

To build null (selectively neutral) distributions of genetic variation in *S. mitis* and *S. pneumoniae* populations and test for the presence of natural selection, we took advantage of the fact that bacterial genomes are gene heavy and extracted the synonymous and non-synonymous variants from random sites across *S. mitis* and *S. pneumoniae* genomes to build randomly sampled windows. We only retained windows that have a minimum of one non-synonymous variant and one synonymous variant. We have then calculated two summary statistics of genetic diversity: the ratio of the number of non-synonymous variants to the number of synonymous variants and the ratio of the nucleotide diversity for non-synonymous variation to the nucleotide diversity for synonymous variation (π_N/π_S). These were then compared with observed ratios at genes with more than one non-synonymous variant and one synonymous variant as well.

Specifically in the case of the ratio of the number of non-synonymous variants to the number of synonymous variants, for a gene of length *n*, we randomly picked *n* sites from across the genome and counted the number of non-synonymous and synonymous variants. This process was repeated 1,000 times for each gene in each species. We then calculated the fraction of simulated ratios that were bigger than the observed ratio, giving us a p-value for an excess of non-synonymous variants for each gene. We then tested if the fraction of genes with p-value<0.05 was more significant than expected under the by performing a binomial test, assuming a null distribution of uniform p-values.

In the case of π_N/π_S, we calculated π_N and π_S independently in 120,000 as follows. We first randomly sampled 1Kb windows (~mean size of *S. mitis* and *S. pneumoniae* genes), recorded the observed counts of non-synonymous and synonymous variants. We then extracted as many as non-synonymous and synonymous variants as in each sampled window randomly from across the genome and calculated π_N/π_S with for these random SNV sets. We then compared this distribution with the observed distribution of π_N/π_S for all genes.

ABC computation

As high (*S. pneumoniae*) to very high recombination rates (*S. mitis*) lead to noticeably strong LD decay (Fig 1C), we assumed a simplified model where all synonymous SNVs are in linkage equilibrium and, therefore, appear on independent coalescence trees (this is, essentially, the Poisson Random Field approach from [95]). On a specific coalescent tree, a SNV has frequency *i* (that is, it is present in *i* sequences in the sample) with probability $B_i/\sum_{j=2}^{n} B_j$, where B_i is the summed length of all genealogical branches that are connected to exactly *i* sampled lineages. Thus, if we observe many SNVs (as in a large window), the proportion of SNVs that show a specific allele frequency is approximately given by $\mathrm{E}(B_i/\sum_{j=2}^{n} B_j) \approx \mathrm{E}(B_i)/\sum_{j=2}^{n} \mathrm{E}(B_j)$ (law of large numbers). Here, we used the analytical approach from [61] as implemented in [62] that uses a Kingman coalescence model with exponential growth of specified growth rate *g* to compute $E(B_i)$ and, in this way, generate a theoretical SFS under this model. Growth rates were drawn from a uniform prior distribution on the discrete set of $g \in [0,25]$ with step 0.1 (in coalescence units of N_e generations). We then obtained 30,000 simulated 250Kb windows for *S. mitis* and 75,000 simulated 250Kb windows for *S. pneumoniae* with the number of SNVs drawn randomly from the observed number of (synonymous) SNVs in sliding windows of the same size (with an offset of 50 Kb between adjacent windows) across the *S. mitis* and *S.*

pneumoniae core genomes, and with a SFS derived from the described coalescent model with a specified growth rate *g*. We used a windowed approach to assess whether the genomic signal is stable across the genome. For *S*. *pneumoniae*, we restricted the analysis to the windows mapping to the first 950 Kb of the reference genome, where read coverage is consistently high ($>$10% of mean depth of coverage) among all isolates (encompass a continuous genomic region of the core genome), and where the number of observed segregating sites in consistent across windows (S1 Text).

For every simulated and observed window, a set of summary statistics of genetic diversity was recorded: the 5% percentiles of the minor allele frequencies; the 10% percentiles of pairwise Hamming distances between sequences, scaled by their maximum within the sample (simulated or observed); and Tajima's D. Calculations on observed data were based on synonymous (neutral) variation.

To estimate posterior distributions and median point estimates of the exponential growth rate, we used the Random Forest ABC model selection procedure (ABC-RF) from the *abcrf* R-package [96]. To evaluate the accuracy of the model, we report the absolute error for simulations with growth rate $g = 0$ (averaged over all true simulations with $g = 0$) and the normalized mean absolute error NMAE, which averages |estimated value–true value|/true value across all simulations, for all other true growth rates combined (S1 Text).

To assess whether the model fits the observed genetic diversity, we performed graphical posterior predictive checks by calculating the Tajima's D of simulated 1,500 250Kb-windows obtained using as parameters the 25% percentile, the median, and the 75% percentile of the estimated growth rates, and comparing these values with the observed Tajima's D of the corresponding windows from which the growth rates were extracted from (S1 Text).

Scripts for running this analysis are available on GitHub at https://github.com/fabfreund/strepto_demography.git.

N_e calculation

N_e and magnitude difference in N_e between species was calculated using coalescence theory and the observed number of segregating sites, S_n. Ignoring the effect of selection, a sample of size n from a haploid species is expected to have $\mathrm{E}(S_n) \approx S_n = \mathrm{N_e}\mathit{mu}\mathrm{E}(\mathrm{L_n})$, where $\mathrm{E}(\mathrm{L_n})$ is the expected length (in coalescent time units) of the genealogy of the sample [97], and *mu* is the per-genome, per-generation mutation rate ($mu = \mu gl$, where μ is the pneumococcal mutation rate of 1.57×10^{-6} $\mathrm{site}^{-1}\mathrm{year}^{-1}$ [35], *g* is the generation time of 14/cell divisions /year [46], and *l is* the core genome size for *S*. *mitis* and *S*. *pneumoniae*). We have then inferred N_e considering a population expansion model for both species, where $\mathrm{E}(\mathrm{L_n})$ was calculated using the recursion from [98], as implemented in [99]. Scripts for running this analysis are available on GitHub at https://github.com/fabfreund/strepto_demography.git.

Supporting information

S1 Text. Accuracy and model fit of the ABC approach used to evaluate *S. mitis* and *S. pneumoniae* demographic histories. Fig (i). Posterior predictive checks of the ABC approach implemented to investigate *S. mitis* demographic history. The density plots show, from left to right, the distribution of Tajima's D obtained from simulating 1,500 250Kb-windows (blue) using as parameters the 25% percentile, the median, and the 75% percentile of estimated growth rates from across windows, and the corresponding to the observed Tajima's D estimated across the 34 windows considered (Table A in S1 Text). **Fig (ii). Posterior predictive checks of the ABC approach implemented to investigate *S. pneumoniae* demographic history**. The density plots show, from left to right, the distribution of Tajima's D obtained from

simulating 1,500 250Kb-windows (blue) using as parameters the 25% percentile, the median, and the 75% percentile of estimated growth rates from across windows, and the corresponding to the observed Tajima's D estimated across the 15 windows considered (Table B in S1 Text). **Table A. Observed genetic diversity indices (S and Tajima's D) and posterior estimates of growth rate obtained in the ABC-RF approach implemented to investigate *S. mitis* demographic history.** Presented are the median, 2.5% and 97.5% percentiles of growth rates obtained from 30,000 simulations. S, number of segregating sites. **Table B**. **Observed genetic diversity indices (S and Tajima's D) and posterior estimates of growth rate obtained in the ABC-RF approach implemented to investigate *S. pneumoniae* demographic history.** Presented are the median, 2.5% and 97.5% percentiles of growth rates obtained from 30,000 simulations. S, # of segregating sites.
(PDF)

S2 Text. Assessing the influence of recombination in the estimation of growth rates for *S. pneumoniae* via the ABC approach in FastSimBac. Fig (i). Comparison between the distribution of observed number of segregating sites (pink; across 83 *S. pneumoniae* core genome windows) and the distribution of simulated number of segregating sites used in the ABC approach implemented to investigate *S. pneumoniae* demographic history. The distribution of simulated S values follows the observed distribution closely. S, number of segregating sites. **Table A**. **Observed genetic diversity indices (S and Tajima's D) and posterior estimates of growth rate obtained in the ABC-RF approach that considers all independence of SNVs (the first 3 columns with posterior estimates) and in the ABC-RF approach that considers the estimated recombination rate (designated as 'FastSimBac'; the last 3 columns with posterior estimates) used to investigate the *S. pneumoniae* demographic history.** Presented are the median, 2.5% and 97.5% percentiles (%iles) of growth rates. S, number of segregating sites.
(PDF)

S3 Text. Higher neutral genetic diversity in *S. mitis* than in *S. pneumoniae* is not due to differences in their mutation rate. Fig (i). Experimental confirmation of comparable mutation rates between *S. mitis* and *S. pneumoniae.* Three isolates per species per experiment were used, which were performed in biological triplicate. Error bars show SD from mean. **A** Spontaneous mutation rate for inhibitory concentration of Streptomycin (black) and Rifampicin (grey). No significant difference between species was identified (two-way ANOVA) for either Streptomycin (*P-value* = 0.36) or Rifampicin (*P-value* = 0.41). **B** Cell viability with and without UV exposure. Number of viable cells was not statistically significant at the 0.05 level between cells exposed to UV and those that were not. Significance between states tested by unpaired t-test (*P-value*: B2C2 = 0.38, C5T6 = 0.36, S1092G24C4 = 0.09, G54 = 0.08, D39 = 0.69, TIGR4 = 0.43). **C** Twenty-four-hour growth curves from starting OD600 of 0.002 (reaching 0.157–0.642) demonstrating comparable growth rate between species.
(PDF)

S1 Fig. Pairwise SNV differences distribution for *S. mitis* and *S. pneumoniae*. A. Pairwise SNV differences distribution for *S. mitis* total sample (n = 119). **B.** Pairwise SNV differences distribution for *S. pneumoniae* total sample (n = 810). **C.** Pairwise SNV differences distribution between related hosts, between unrelated hosts and within host for *S. mitis*. **D.** Pairwise SNV differences distribution between unrelated hosts and within hosts for *S. pneumoniae* (for the African dataset, n = 230).
(PDF)

S2 Fig. *S. pneumoniae* genetic diversity considering bigger thresholds of pairwise SNV differences to control for clonal relationships. The number of genes is plotted as a function of the number of genomes. Power law regression was fitted to the mean number of genes obtained across all permutations and for 1000 random samples of size 75 as in *S. mitis*. The parameterisation that best fitted the data was: $Y = aX^b + c$; **A.** Analysis of *S. pneumoniae* pangenomes for the isolates with a minimum of 1000 pairwise SNV differences ($b = 0.16–0.32$; mean = 0.23). Core genomes nucleotide diversity (π) estimates for Africa (Af), Asia (As) and European (Eu) samples: $\pi_{Af} = 0.009$; $\pi_{As} = 0.010$; $\pi_{Eu} = 0.008$. **B.** Analysis of *S. pneumoniae* pangenomes for the isolates with a minimum of 2000 pairwise SNV differences ($b = 0.16–0.34$; mean = 0.24. Core genomes nucleotide diversity estimates: $\pi_{Af} = 0.009$; $\pi_{As} = 0.010$; $\pi_{Eu} =$ 0.008. **C.** Analysis of *S. pneumoniae* pangenomes for the isolates with a minimum of 2000 pairwise SNV differences ($b = 0.19–0.31$; mean = 0.28). Core genomes nucleotide diversity estimates: $\pi_{Af} = 0.010$; $\pi_{As} = 0.011$; $\pi_{Eu} = 0.008$. *S. mitis*, orange; *S. pneumoniae*, blue.
(PDF)

S3 Fig. Population structure analyses for *S. pneumoniae* considering bigger thresholds of pairwise SNV differences to control for clonal relationships. A&B. PCA analysis of *S. pneumoniae* genetic variation for the isolates with a minimum of 1000 pairwise SNV differences (A, including serotype NT; B, excluding serotype NT). F_{ST}s for Africa (Af), Asia (As) and European (Eu) population pairs: F_{ST}(Af-As) = 0.0177 +/- 0.0027; F_{ST}(Af-Eu) = 0.0146 +/- 0.0016; F_{ST}(As-Eu) = 0.0310 +/- 0.0031. **C&D.** PCA analysis of *S. pneumoniae* genetic variation for the isolates with a minimum of 2000 pairwise SNV differences (C, including serotype NT; D, excluding serotype NT). F_{ST}s: F_{ST}(Af-As) = 0.0235 +/- 0.0027; F_{ST}(Af-Eu) = 0.0069 +/- 0.0018; F_{ST}(As-Eu) = 0.0253 +/- 0.0025. **E&F.** PCA analysis of *S. pneumoniae* genetic variation for the isolates with a minimum of 5600 pairwise SNV differences (E, including serotype NT; F, excluding serotype NT). F_{ST}s: F_{ST}(Af-As) = 0.0032 +/- 0.0018; F_{ST}(Af-Eu) = -0.0088 +/- 0.0034; F_{ST}(As-Eu) = 0.0042 +/- 0.0017. The NT cluster comprise nonencapsulated isolates which were previously reported to have higher recombination rates generating significantly more diversity within this cluster [44]. Colour code corresponds to geographic region: Africa, pink; Asia, blue; Europe, green. Abbreviations: pair diff, pairwise SNV differences; ST-NT, serotype NT.
(PDF)

S4 Fig. Phylogenetic and population structure for *S. pneumoniae*'s total sample (including serotype NT). A. PCA computed in Plink v1.9 [93]. **B.** Maximum likelihood unrooted phylogenetic tree obtained using FastTree [84]. There are subclades of Asian lineages which belong exclusively to the PCA outlying serotypes NT (three subclades with isolates belonging mainly to GPSCs 28, 42, 60, 66, 118) and 19F (one subclade classified as GPSC-1), although both NT and 19F clades also include African and European lineages. The NT cluster comprise nonencapsulated isolates which were previously reported to have higher recombination rates generating significantly more diversity within this cluster [44]. Colour code corresponds to geographic region: Africa, pink; Asia, blue; Europe, green.
(PDF)

S5 Fig. Distribution of p-values for whether genes have a higher ratio of number of non-synonymous variants to number of synonymous variants than a random set of simulated ratios for *S. mitis* (A) and *S. pneumoniae* (B). The horizontal line corresponds to the expected number of genes with p-values<0.05, under a null uniform distribution of p-values.
(PDF)

S6 Fig. QQplots of the null π_N/π_S distribution obtained from simulated 1Kb windows (wins) versus the observed π_N/π_S distributions in genes from *S. mitis* (A) and *S. pneumoniae* (B).
(PDF)

S1 Table. Pangenome statistics for *S. pneumoniae* and *S. mitis*. Pangenome and core genome sizes units are number of genes. N, sample size. Regional (continental) random sample of unrelated isolates show means of 1000 random samples with size equal to that of the smaller regional sample size within each species. 'All random sample of unrelated isolates' in *S. pneumoniae* shows the mean of 1000 random samples with size equal to the observed *S. mitis* unrelated sample size.
(PDF)

S2 Table. Between population divergence estimates (Hudson's F_{ST} +/- SE). *S. mitis*, above diagonal; *S. pneumoniae*, below diagonal.
(PDF)

S3 Table. Genome accession and geographical origin of *Streptococcus mitis* genomes.
(PDF)

Acknowledgments

We thank Ebony Cave for bioinformatic support during the writing of this paper.

Author Contributions

Conceptualization: Marco R. Oggioni, Sandra Beleza.

Data curation: Charlotte Davison, Megan de Ste-Croix.

Formal analysis: Charlotte Davison, Sam Tallman, Fabian Freund, Sandra Beleza.

Funding acquisition: Sandra Beleza.

Investigation: Charlotte Davison, Sam Tallman, Megan de Ste-Croix, Fabian Freund, Sandra Beleza.

Methodology: Fabian Freund.

Resources: Martin Antonio, Brenda Kwambana-Adams.

Supervision: Sandra Beleza.

Visualization: Charlotte Davison, Fabian Freund, Sandra Beleza.

Writing – original draft: Charlotte Davison, Fabian Freund, Sandra Beleza.

Writing – review & editing: Charlotte Davison, Sam Tallman, Megan de Ste-Croix, Martin Antonio, Marco R. Oggioni, Brenda Kwambana-Adams, Fabian Freund, Sandra Beleza.

References

1. Denapaite D, Rieger M, Kondgen S, Bruckner R, Ochigava I, Kappeler P, et al. Highly Variable *Streptococcus oralis* Strains Are Common among Viridans Streptococci Isolated from Primates. Msphere. 2016; 1(2). ARTN e00041-15 https://doi.org/10.1128/mSphere.00041-15 PMID: 27303717

2. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, et al. Cospeciation of gut microbiota with hominids. Science. 2016; 353(6297):380–2. https://doi.org/10.1126/science.aaf3951 PMID: 27463672

3. Jensen A, Scholz CFP, Kilian M. Re-evaluation of the taxonomy of the Mitis group of the genus Streptococcus based on whole genome phylogenetic analyses, and proposed reclassification of *Streptococcus dentisani* as *Streptococcus oralis* subsp. *dentisani* comb. nov., *Streptococcus tigurinus* as *Streptococcus oralis* subsp. *tigurinus* comb. nov., and *Streptococcus oligofermentans* as a later synonym of *Streptococcus cristatus*. Int J Syst Evol Microbiol. 2016; 66(11):4803–20. Epub 20160817. https://doi.org/10.1099/ijsem.0.001433 PMID: 27534397.

4. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013; 493(7430):45–50. https://doi.org/10.1038/nature11711 PMID: 23222524

5. Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. Plos Biol. 2019; 17(1). ARTN e3000102 https://doi.org/10.1371/journal.pbio.3000102 PMID: 30673701

6. Li K, Bihan M, Methe BA. Analyses of the Stability and Core Taxonomic Memberships of the Human Microbiome. Plos One. 2013; 8(5). ARTN e63139 https://doi.org/10.1371/journal.pone.0063139 PMID: 23671663

7. Zaura E, Nicu EA, Krom BP, Keijser BJF. Acquiring and maintaining a normal oral microbiome: current perspective. Front Cell Infect Mi. 2014; 4. ARTN 85 https://doi.org/10.3389/fcimb.2014.00085 PMID: 25019064

8. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe. 2018; 24 (1):133–45 e5. https://doi.org/10.1016/j.chom.2018.06.005 PMID: 30001516.

9. Hoshino T, Fujiwara T, Kilian M. Use of phylogenetic and phenotypic analyses to identify nonhemolytic streptococci isolated from bacteremic patients. J Clin Microbiol. 2005; 43(12):6073–85. https://doi.org/10.1128/JCM.43.12.6073-6085.2005 PMID: 16333101.

10. Hohwy J, Reinholdt J, Kilian M. Population dynamics of *Streptococcus mitis* in its natural habitat. Infect Immun. 2001; 69(10):6055–63. https://doi.org/10.1128/Iai.69.10.6055-6063.2001

11. Kirchherr JL, Bowden GH, Richmond DA, Sheridan MJ, Wirth KA, Cole MF. Clonal diversity and turnover of *Streptococcus mitis* bv. 1 on shedding and nonshedding oral surfaces of human infants during the first year of life. Clin Diagn Lab Immunol. 2005; 12(10):1184–90. https://doi.org/10.1128/CDLI.12.10.1184-1190.2005 PMID: 16210481.

12. Valles-Colomer M, Blanco-Miguez A, Manghi P, Asnicar F, Dubois L, Golzato D, et al. The person-to-person transmission landscape of the gut and oral microbiomes. Nature. 2023; 614(7946):125–+. https://doi.org/10.1038/s41586-022-05620-1 PMID: 36653448

13. Bek-Thomsen M, Tettelin H, Hance I, Nelson KE, Kilian M. Population diversity and dynamics of *Streptococcus mitis*, *Streptococcus oralis*, and *Streptococcus infantis* in the upper respiratory tracts of adults, determined by a nonculture strategy. Infect Immun. 2008; 76(5):1889–96. Epub 20080303. https://doi.org/10.1128/IAI.01511-07 PMID: 18316382

14. Kilian M, Poulsen K, Blomqvist T, Havarstein LS, Bek-Thomsen M, Tettelin H, et al. Evolution of *Streptococcus pneumoniae* and its close commensal relatives. Plos One. 2008; 3(7):e2683. Epub 20080716. https://doi.org/10.1371/journal.pone.0002683 PMID: 18628950.

15. Kilian M, Riley DR, Jensen A, Bruggemann H, Tettelin H. Parallel Evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to Pathogenic and Mutualistic Lifestyles. Mbio. 2014; 5(4). ARTN e01490-14 https://doi.org/10.1128/mBio.01490-14 PMID: 25053789

16. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of human migrations in *Helicobacter pylori* populations. Science. 2003; 299(5612):1582–5. https://doi.org/10.1126/science.1080857 PMID: 12624269

17. Suzuki TA, Fitzstevens JL, Schmidt VT, Enav H, Huus KE, Mbong Ngwese M, et al. Codiversification of gut microbiota with humans. Science. 2022; 377(6612):1328–32. https://doi.org/10.1126/science.abm7759 PMID: 36108023

18. Good BH. Limited codiversification of the gut microbiota with humans. bioRxiv. 2023:2022.10.27.514143. https://doi.org/10.1101/2022.10.27.514143

19. Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, et al. The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host & Microbe. 2019; 26(5):666–+. https://doi.org/10.1016/j.chom.2019.08.018 PMID: 31607556

20. Karcher N, Pasolli E, Asnicar F, Huang KD, Tett A, Manara S, et al. Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol. 2020; 21(1):138. Epub 20200608. https://doi.org/10.1186/s13059-020-02042-y PMID: 32513234.

21. Wahl B, O'Brien KL, Greenbaum A, Majumder A, Liu L, Chu Y, et al. Burden of *Streptococcus pneumoniae* and *Haemophilus influenzae* type b disease in children in the era of conjugate vaccines: global,

regional, and national estimates for 2000–15. Lancet Glob Health. 2018; 6(7):e744–e57. https://doi.org/10.1016/S2214-109X(18)30247-X PMID: 29903376.

22. Bogaert D, De Groot R, Hermans PW. *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. Lancet Infect Dis. 2004; 4(3):144–54. https://doi.org/10.1016/S1473-3099(04)00938-7 PMID: 14998500.

23. Simell B, Auranen K, Käyhty H, Goldblatt D, Dagan R, O'Brien KL. The fundamental link between pneumococcal carriage and disease. Expert Review of Vaccines. 2012; 11(7):841–55. https://doi.org/10.1586/erv.12.53 PMID: 22913260

24. Senghore M, Chaguza C, Bojang E, Tientcheu PE, Bancroft RE, Lo SW, et al. Widespread sharing of pneumococcal strains in a rural African setting: proximate villages are more likely to share similar strains that are carried at multiple timepoints. Microb Genom. 2022; 8(2). https://doi.org/10.1099/mgen.0.000732 PMID: 35119356.

25. Tonkin-Hill G, Ling C, Chaguza C, Salter SJ, Hinfonthong P, Nikolaou E, et al. Pneumococcal within-host diversity during colonization, transmission and treatment. Nat Microbiol. 2022; 7(11):1791–804. Epub 20221010. https://doi.org/10.1038/s41564-022-01238-1 PMID: 36216891

26. Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, et al. Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone *Streptococcus pneumoniae* (Spain23F) ST81. J Bacteriol. 2009; 191(5):1480–9. https://doi.org/10.1128/Jb.01343-08 PMID: 19114491

27. Blomberg C, Dagerhamn J, Dahlberg S, Browall S, Fernebro J, Albiger B, et al. Pattern of Accessory Regions and Invasive Disease Potential in *Streptococcus pneumoniae*. J Infect Dis. 2009; 199 (7):1032–42. https://doi.org/10.1086/597205 PMID: 19203261

28. Thomas JC, Figueira M, Fennie KP, Laufer AS, Kong Y, Pichichero ME, et al. *Streptococcus pneumoniae* clonal complex 199: genetic diversity and tissue-specific virulence. Plos One. 2011; 6(4):e18649. Epub 20110414. https://doi.org/10.1371/journal.pone.0018649 PMID: 21533186.

29. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Linares J, et al. The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. Genome Biology. 2012; 13(11). ARTN R103 https://doi.org/10.1186/gb-2012-13-11-r103 PMID: 23158461

30. Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, Feltwell T, et al. Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. PLoS Genet. 2013; 9(10):e1003868. Epub 20131010. https://doi.org/10.1371/journal.pgen.1003868 PMID: 24130509.

31. Chaguza C, Ebruke C, Senghore M, Lo SW, Tientcheu PE, Gladstone RA, et al. Comparative Genomics of Disease and Carriage Serotype 1 Pneumococci. Genome Biol Evol. 2022; 14(4). ARTN evac052 https://doi.org/10.1093/gbe/evac052 PMID: 35439297

32. Cremers AJH, Mobegi FM, van der Gaast-de Jongh C, van Weert M, van Opzeeland FJ, Vehkala M, et al. The Contribution of Genetic Variation of *Streptococcus pneumoniae* to the Clinical Manifestation of Invasive Pneumococcal Disease. Clin Infect Dis. 2019; 68(1):61–9. https://doi.org/10.1093/cid/ciy417 PMID: 29788414

33. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Seron MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nature Communications. 2019; 10. ARTN 2176 https://doi.org/10.1038/s41467-019-09976-3 PMID: 31092817

34. Gladstone RA, Siira L, Brynildsrud OB, Vestrheim DF, Turner P, Clarke SC, et al. International links between *Streptococcus pneumoniae* vaccine serotype 4 sequence type (ST) 801 in Northern European shipyard outbreaks of invasive pneumococcal disease. Vaccine. 2022; 40(7):1054–60. https://doi.org/10.1016/j.vaccine.2021.10.046 PMID: 34996643

35. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011; 331(6016):430–4. https://doi.org/10.1126/science.1198545 PMID: 21273480.

36. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nature Communications. 2014; 5(1):5471. https://doi.org/10.1038/ncomms6471 PMID: 25407023

37. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013; 45(6):656–+. https://doi.org/10.1038/ng.2625 PMID: 23644493

38. Cremers AJH, Mobegi FM, de Jonge MI, van Hijum SAFT, Meis JF, Hermans PWM, et al. The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. Sci Rep-Uk. 2015; 5. ARTN 14952 https://doi.org/10.1038/srep14952 PMID: 26492862

39. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, et al. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol. 2017; 1(12):1950–60. https://doi.org/10.1038/s41559-017-0337-x PMID: 29038424

40. Lochen A, Croucher NJ, Anderson RM. Divergent serotype replacement trends and increasing diversity in pneumococcal disease in high income settings reduce the benefit of expanding vaccine valency. Sci Rep-Uk. 2020; 10(1). ARTN 18977 https://doi.org/10.1038/s41598-020-75691-5 PMID: 33149149

41. Azarian T, Martinez PP, Arnold BJ, Qiu XT, Grant LR, Corander J, et al. Frequency-dependent selection can forecast evolution in *Streptococcus pneumoniae*. Plos Biol. 2020; 18(10). https://doi.org/10.1371/journal.pbio.3000878 PMID: 33091022

42. Gladstone RA, Devine V, Jones J, Cleary D, Jefferies JM, Bentley SD, et al. Pre-vaccine serotype composition within a lineage signposts its serotype replacement—a carriage study over 7 years following pneumococcal conjugate vaccine use in the UK. Microb Genom. 2017; 3(6):e000119. Epub 20170609. https://doi.org/10.1099/mgen.0.000119 PMID: 29026652.

43. Croucher NJ, Chewapreecha C, Hanage WP, Harris SR, McGee L, van der Linden M, et al. Evidence for Soft Selective Sweeps in the Evolution of Pneumococcal Multidrug Resistance and Vaccine Escape. Genome Biol Evol. 2014; 6(7):1589–602. https://doi.org/10.1093/gbe/evu120 PMID: 24916661

44. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014; 46(3):305–+. https://doi.org/10.1038/ng.2895 PMID: 24509479

45. Everett DB, Cornick J, Denis B, Chewapreecha C, Croucher N, Harris S, et al. Genetic Characterisation of Malawian Pneumococci Prior to the Roll-Out of the PCV13 Vaccine Using a High-Throughput Whole Genome Sequencing Approach. Plos One. 2012; 7(9). ARTN e44250 https://doi.org/10.1371/journal.pone.0044250 PMID: 22970189

46. Chaguza C, Senghore M, Bojang E, Gladstone RA, Lo SW, Tientcheu PE, et al. Within-host microevolution of *Streptococcus pneumoniae* is rapid and adaptive during natural colonisation. Nature Communications. 2020; 11(1). https://doi.org/10.1038/s41467-020-17327-w PMID: 32651390

47. Tonkin-Hill G, Ling C, Chaguza C, Salter SJ, Hinfonthong P, Nikolaou E, et al. Pneumococcal within-host diversity during colonization, transmission and treatment. Nature Microbiology. 2022; 7(11):1791–+. https://doi.org/10.1038/s41564-022-01238-1 PMID: 36216891

48. Kremer PH, Ferwerda B, Bootsma HJ, Rots NY, Wijmenga-Monsuur AJ, Sanders EA, et al. Pneumococcal genetic variability in age-dependent bacterial carriage. Elife. 2022; 11. ARTN e69244 https://doi.org/10.7554/eLife.69244 PMID: 35881438

49. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife. 2017; 6. ARTN e26255 https://doi.org/10.7554/eLife.26255 PMID: 28742023

50. Wright S. The genetical structure of populations. Ann Eugen. 1951; 15(4):323–54. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x PMID: 24540312.

51. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245.

52. Gladstone RA, Lo SW, Goater R, Yeats C, Taylor B, Hadfield J, et al. Visualizing variation within Global Pneumococcal Sequence Clusters (GPSCs) and country population snapshots to contextualize pneumococcal isolates. Microb Genomics. 2020; 6(5). ARTN 000357 https://doi.org/10.1099/mgen.0.000357 PMID: 32375991

53. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine. 2019; 43:338–46. Epub 20190416. https://doi.org/10.1016/j.ebiom.2019.04.021 PMID: 31003929

54. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008; 11(5):472–7. https://doi.org/10.1016/j.mib.2008.09.006 PMID: 19086349

55. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting F-ST: The impact of rare variants. Genome Research. 2013; 23(9):1514–21. https://doi.org/10.1101/gr.154831.113 PMID: 23861382

56. Brito IL, Gurry T, Zhao S, Huang K, Young SK, Shea TP, et al. Transmission of human-associated microbiota along family and social networks. Nat Microbiol. 2019; 4(6):964–71. Epub 20190325. https://doi.org/10.1038/s41564-019-0409-6 PMID: 30911128

57. Lin MZ, Kussell E. Inferring bacterial recombination rates from large-scale sequencing datasets. Nat Methods. 2019; 16(2):199–+. https://doi.org/10.1038/s41592-018-0293-7 PMID: 30664775

58. Lin MZ, Kussell E. Correlated Mutations and Homologous Recombination Within Bacterial Populations. Genetics. 2017; 205(2):891–917. https://doi.org/10.1534/genetics.116.189621 PMID: 28007887

59. Sved JA, Hill WG. One Hundred Years of Linkage Disequilibrium. Genetics. 2018; 209(3):629–36. https://doi.org/10.1534/genetics.118.300642 PMID: 29967057

60. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 2005; 102(22):7882–7. Epub 20050519. https://doi.org/10.1073/pnas.0502300102 PMID: 15905331

61. Spence JP, Kamm JA, Song YS. The Site Frequency Spectrum for General Coalescents. Genetics. 2016; 202(4):1549–61. Epub 20160216. https://doi.org/10.1534/genetics.115.184101 PMID: 26883445

62. Freund F, Kerdoncuff E, Matuszewski S, Lapierre M, Hildebrandt M, Jensen JD, et al. Interpreting the pervasive observation of U-shaped Site Frequency Spectra. PLoS Genet. 2023; 19(3):e1010677. Epub 20230323. doi: 10.1371/journal.pgen.1010677. PMID: 36952570

63. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017; 27(4):626–38. Epub 20170206. https://doi.org/10.1101/gr.216242.116 PMID: 28167665

64. Kalizang'oma A, Chaguza C, Gori A, Davison C, Beleza S, Antonio M, et al. *Streptococcus pneumoniae* serotypes that frequently colonise the human nasopharynx are common recipients of penicillin-binding protein gene fragments from Streptococcus mitis. Microb Genomics. 2021; 7(9). ARTN 000622 https://doi.org/10.1099/mgen.0.000622 PMID: 34550067

65. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. Isme J. 2017; 11(7):1719–21. https://doi.org/10.1038/ismej.2017.36 PMID: 28362722

66. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nature Microbiology. 2017; 2(4). ARTN 17040 https://doi.org/10.1038/nmicrobiol.2017.40 PMID: 28350002

67. Barrett R, Kuzawa CW, McDade T, Armelagos GJ. Emerging and re-emerging infectious diseases: The third epidemiologic transition. Annu Rev Anthropol. 1998; 27:247–71. https://doi.org/10.1146/annurev.anthro.27.1.247

68. Excoffier L. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Mol Ecol. 2004; 13(4):853–64. https://doi.org/10.1046/j.1365-294x.2003.02004.x PMID: 15012760

69. Li Y, Thompson CM, Trzcinski K, Lipsitch M. Within-host selection is limited by an effective population of *Streptococcus pneumoniae* during nasopharyngeal colonization. Infect Immun. 2013; 81(12):4534–43. Epub 20130930. https://doi.org/10.1128/IAI.00527-13 PMID: 24082074

70. Birzu G, Hallatschek O, Korolev KS. Genealogical structure changes as range expansions transition from pushed to pulled. P Natl Acad Sci USA. 2021; 118(34). ARTN e2026746118 https://doi.org/10.1073/pnas.2026746118 PMID: 34413189

71. Gladstone RA, Siira L, Brynildsrud OB, Vestrheim DF, Turner P, Clarke SC, et al. International links between Streptococcus pneumoniae vaccine serotype 4 sequence type (ST) 801 in Northern European shipyard outbreaks of invasive pneumococcal disease. Vaccine. 2022; 40(7):1054–60. Epub 20220105. https://doi.org/10.1016/j.vaccine.2021.10.046 PMID: 34996643

72. Bobay LM, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 2018; 18(1):153. Epub 20181012. https://doi.org/10.1186/s12862-018-1272-4 PMID: 30314447.

73. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. Trends Microbiol. 2010; 18 (7):315–22. Epub 20100506. https://doi.org/10.1016/j.tim.2010.04.002 PMID: 20452218

74. Bishop CJ, Aanensen DM, Jordan GE, Kilian M, Hanage WP, Spratt BG. Assigning strains to bacterial species via the internet. BMC Biol. 2009; 7:3. Epub 20090126. https://doi.org/10.1186/1741-7007-7-3 PMID: 19171050.

75. Dworkin M, Falkow S. The prokaryotes: a handbook on the biology of bacteria. 3rd ed. New York; London: Springer; 2006. v. < 1–6 > p.

76. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. Epub 20140401. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

77. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. (Online). 2010.

78. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19(5):455–77. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

79. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. Microb Genomics. 2016; 2(8). ARTN 000083 https://doi.org/10.1099/mgen.0.000083 PMID: 28348874

80. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29(8):1072–5. https://doi.org/10.1093/bioinformatics/btt086 PMID: 23422339

81. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30(14):2068–9. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063

82. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015; 31(22):3691–3. https://doi.org/10.1093/bioinformatics/btv421 PMID: 26198102

83. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30(14):3059–66. https://doi.org/10.1093/nar/gkf436 PMID: 12136088

84. Price MN, Dehal PS, Arkin AP. FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. Plos One. 2010; 5(3). ARTN e9490 https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

85. Yang ZH. Maximum-Likelihood Phylogenetic Estimation from DNA-Sequences with Variable Rates over Sites—Approximate Methods. J Mol Evol. 1994; 39(3):306–14. https://doi.org/10.1007/BF00160154 PMID: 7932792

86. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021; 49(W1):W293–W6. https://doi.org/10.1093/nar/gkab301 PMID: 33885785

87. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

88. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;303.3997v2.

89. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. Fly. 2012; 6(2):80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672

90. Seeman T, Klötzl F, Page AJ. https://github.com/tseemann/snp-dists. 2018.

91. Alistair M, Harding N. cggh/scikit-allel: v1.3.3. Zenodo. http://doi.org/10.5281/zenodo.822784. 2017.

92. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010; 10(3):564–7. https://doi.org/10.1111/j.1755-0998.2010.02847.x PMID: 21565059

93. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81 (3):559–75. https://doi.org/10.1086/519795 PMID: 17701901

94. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. Plos One. 2010; 5(6). ARTN e11147 https://doi.org/10.1371/journal.pone.0011147 PMID: 20593022

95. Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992; 132 (4):1161–76. https://doi.org/10.1093/genetics/132.4.1161 PMID: 1459433.

96. Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A. ABC random forests for Bayesian parameter inference. Bioinformatics. 2019; 35(10):1720–8. https://doi.org/10.1093/bioinformatics/bty867 PMID: 30321307

97. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975; 7(2):256–76. https://doi.org/10.1016/0040-5809(75)90020-9 PMID: 1145509.

98. Polanski A, Kimmel M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 2003; 165(1):427–36. https://doi.org/10.1093/genetics/165.1.427 PMID: 14504247

99. Eldon B, Birkner M, Blath J, Freund F. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics. 2015; 199(3):841–56. Epub 20150109. https://doi.org/10.1534/genetics.114.173807 PMID: 25575536