Whole genome phylogenies reflect long-tailed distributions of recombination rates in many bacterial species

Thomas Sakoparnig[1], Chris Field[1,2], and Erik van Nimwegen[1,*]
1 Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland
2 Current address: Institute of microbiology, ETH-Zurich, Zurich, Switzerland
∗ E-mail: erik.vannimwegen@unibas.ch

Abstract

Although homologous recombination is accepted to be common in bacteria, so far it has been challenging to accurately quantify its impact on genome evolution within bacterial species. We here introduce methods that use the statistics of single-nucleotide polymorphism (SNP) splits in the core genome alignment of a set of strains to show that, for many bacterial species, recombination dominates genome evolution. Each genomic locus has been overwritten so many times by recombination that it is impossible to reconstruct the clonal phylogeny and, instead of a consensus phylogeny, the phylogeny typically changes many thousands of times along the core genome alignment.

We also show how SNP splits can be used to quantify the relative rates with which different subsets of strains have recombined in the past. We find that virtually every strain has a unique pattern of recombination frequencies with other strains and that the relative rates with which different subsets of strains share SNPs follow long-tailed distributions. Our findings show that bacterial populations are neither clonal nor freely recombining, but structured such that recombination rates between different lineages vary along a continuum spanning several orders of magnitude, with a unique pattern of rates for each lineage. Thus, rather than reflecting clonal ancestry, whole genome phylogenies reflect these long-tailed distributions of recombination rates.

Introduction

The only illustration that appears in Darwin's Origin of species [1] is of a phylogenetic tree. Indeed, the tree has become the archetypical concept representing biological evolution. Since every biological cell that has ever lived was the result of a cell division, all cells are connected through cell divisions in a giant tree that stretches all the way back to the earliest cells that existed on earth. Thus, the study of biological evolution in some sense corresponds to the study of the structure of this giant cell-division tree. Indeed, virtually all models of evolutionary dynamics formulate the dynamics as occurring along the branches of a tree, and many mathematical and computational methods have been developed for inference and modeling of evolutionary dynamics along the branches of a tree, e.g. [2, 3].

It is thus natural that the first step in the analysis of a set of related biological sequences is to reconstruct the phylogenetic tree that reflects the cell division history of the sequences, i.e their 'ancestral phylogeny'. Once the ancestral relationships between the sequences are known, the evolution of the sequences can then be modeled along the branches of this tree. This strategy has been employed from the earliest days of sequence analysis [4] and is almost invariable applied in the analysis of microbial genome sequences, which is the main topic of this work.

A second key concept in models of evolutionary dynamics is the idea of a 'population' of organisms that are mutually competing for resources and that, for purposes of mathematical modeling, can be considered exchangeable in the sense that they are subjected to the same environment. Indeed, populations of exchangeable individuals form the basis of almost all mathematical population genetics models (see e.g. [5]), including coalescent models for phylogenies [6]. Although it is of course well recognized that, in the real world, populations are structured into sub-populations with varying degrees of interaction between them, population genetics models almost by definition assume that at some level there are sub-populations of exchangeable individuals sharing a common environment.

In this paper we present evidence that we believe challenges the usefulness of applying these two concepts for describing genome evolution in prokaryotes. First, we find that for most bacterial species recombination is so frequent that, within an alignment of strains, each genomic locus has been overwritten by recombination many times and the phylogeny typically changes tens of thousands of times along the genome. Moreover, for most pairs of strains, none of the loci in their pairwise alignment derives from their ancestor in the ancestral phylogeny, and the vast majority of genomic differences result from recombination events, even for very close pairs. Consequently, the ancestral phylogeny cannot be reconstructed from the genome sequences using currently available methods and, more generally, the strategy of modeling microbial genome evolution as occurring along the branches of an ancestral phylogeny breaks down.

Second, we show that the structure represented in whole genome phylogenies of microbial strains does not reflect ancestry, but instead the relative rates with which different lineages have recombined in the past. Whereas almost every short genomic segment follows a different phylogeny, these phylogenies are not uniformly randomly sampled from all possible phylogenies, but sampled from highly biased distributions. In particular, the relative frequencies with which particular sub-clades of strains occur in the phylogenies at different loci follow roughly power-law distributions and each strain has a distinct distribution of co-occurrence frequency with the other stains. Since each strain has a unique 'finger print' of recombination rates with the lineages of other strains, the assumption that at some level strains can be considered as exchangeable members of a population, also fundamentally breaks down.

The structure of the paper is as follows. To present our analyses, we will focus on a collection of 91 wild *E. coli* strains that were isolated over a short period from a common habitat [7]. After introducing these strains, we introduce the main puzzle of bacterial whole genome phylogeny: although the phylogenies of individual genomic loci are all distinct, the phylogeny inferred from any large collection of genomic loci converges to a common structure, e.g. [8–10]. We first study recombination by studying pairs of strains, extending a recent approach by Dixit et al. [11] to model each pairwise alignment as a mixture of ancestrally inherited and recombined regions. We show that, as the distance to the pair's common ancestor increases, the fraction of the genome covered by recombined segments increases, and at some pairwise distance all clonally inherited DNA disappears. Importantly, this distance is far below the typical divergence of pairs of strains such that for the vast majority of pairs, none of the DNA in their genome alignment stems from their common ancestor.

Much of the new analysis methodology that we introduce is based on bi-allelic SNPs (which constitute almost all SNPs in the core alignment). Although bi-allelic SNPs have been studied to estimate the number of recombinations along alignments of sexually reproducing species [12], as far as we are aware they have received very little attention in the study of prokaryotic genomes. We show that virtually all bi-allelic SNPs correspond to single mutational events in the history of its genomic locus, so that each SNP provides a bi-partition that occurs at the phylogeny at that locus. We show various ways in which these bi-allelic SNPs can be used to investigate which SNPs are consistent with given phylogenies, and each other, and use them to quantify the amount of phylogenetic variation along the alignment. We use these SNPs to show there is no consensus phylogeny, that the phylogeny changes every few SNPs along the core phylogeny, and to estimate a lower bound on the ratio of recombination to mutation events in a genome alignment.

We then show how these SNPs can also be used to quantify the relative rates with which different lineages share mutations, and show that these rates follow approximately scale free distributions, indicating that there is 'population structure' on every scale. Finally, we define entropy profiles of phylogenetic variability of each strain and show that these entropy profiles provide a unique phylogenetic fingerprint of each strain.

We close by showing how all the statistics that we developed for *E. coli* apply to a set of other bacterial species including: *B. subtilis*, *H. pylori*, *M. tuberculosis*, *S. enterica*, and *S. aureus*. We show that, with the exception of *M. tuberculosis* where all strains are very closely related and no pair has yet been fully recombined, all other species follow the same general behavior as *E. coli.* Thus, for almost all bacterial species that we studied, there is no common or consensus phylogeny, but many thousands of different phylogenies along the core genome. These phylogenies are drawn from a distribution with scale-free properties, and each strain has a unique fingerprint of recombination with the others. We feel that these observations necessitate a new way of thinking about how to model genome evolution in prokaryotes.

Results

To illustrate our methods we focus on the SC1 collection of wild *E. coli* isolates that were collected in 2003 – 2004 near the shore of St. Louis river in Duluth, Minnesota [7]. We sequenced 91 strains from this collection together with the K12 MG1655 lab strain as a reference. In a companion paper [13] we discuss this collection in more detail and extensively analyze the evolution of gene content and phenotypes of this collection. Here we focus on sequence evolution in the core genome of these strains. Although the SC1 strains were collected from a common habitat over a short period of time, they show a remarkable diversity, with no two identical strains, all known major groups of *E. coli* represented, as well as an 'out group' of 9 strains that are more than 8% diverged at the nucleotide level from other *E. coli* strains (see Suppl. Fig. S1 for a phylogenetic tree constructed using maximum likelihood on the joint core genome of the SC1 strains and 189 reference strains, [13]).

Phylogenies of individual loci disagree with the phylogeny of the core genome

To construct a core genome alignment of the SC1 strains and K12 MG1655 we used the REALPHY software [14] (see Methods), resulting in a multiple alignment across all 92 strains of $2'756'541$ base pairs long. REALPHY used PhyML [15] to reconstruct a phylogeny from the core genome alignment and will refer to this tree as the *core tree* from here on (Fig. 1).

We first checked to what extent the alignments of individual genomic loci are statistically consistent with the core tree. For each 3 kilobase block of the core alignment we used PhyML to reconstruct a phylogeny and then compared its log-likelihood with the log-likelihood that can be obtained when the phylogeny is constrained to have the topology of the core tree. We find that essentially all 3 Kb alignment blocks reject the core tree (Suppl. Fig. S2, left panel). Moreover, each alignment block rejects the topologies that were constructed from all other alignment blocks (Suppl. Fig. S2, right panel).

Although it thus appears that the phylogeny at each genomic locus is statistically significantly distinct, it is still possible that all these phylogenies are highly similar. In order to quantify the differences between the core tree and the phylogenies of 3Kb blocks we calculated, for each split in the core tree, the fraction of 3Kb blocks for which the same split occurred in the phylogeny reconstructed from that alignment block. As shown in the top row of Fig. 1, the phylogenies of individual blocks differ substantially from the core tree: roughly two-thirds of the splits in the core tree occur in less than half of 3Kb block phylogenies and half of the core tree splits occur in less than a quarter of all 3Kb block phylogenies. Especially the splits higher up in the core tree do not occur in the large majority of block phylogenies.

These observations are not particularly novel. There is by now a vast and sometimes contentious literature on the role of recombination in prokaryotic genome evolution which is beyond the scope of this article to review. We thus focus on a few key points that are central to the questions and methods we study here. First, systematic studies of complete microbial genomes have shown that horizontal gene transfer is relatively common and can significantly affect phylogenies of individual loci, e.g. [16, 17]. Such observations caused some researchers to question whether trees can be meaningfully used to describe genome evolution [18]. However, many in the researchers field feel that, given careful study, a major phylogenetic backbone can be extracted from genomic data. For example, it has been observed that, whenever a phylogeny is reconstructed from the alignments of a large number of genomic loci, one

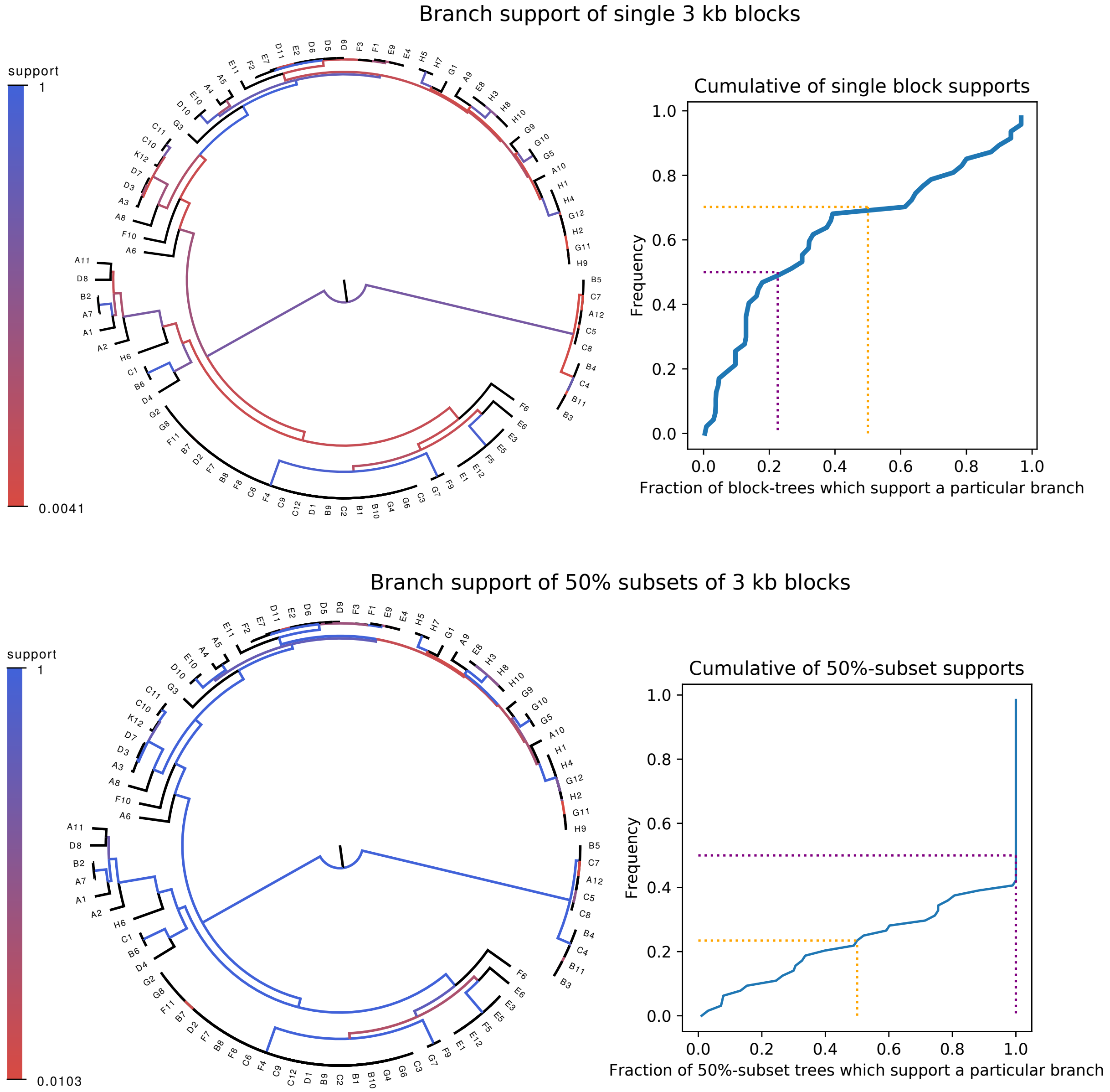

Figure 1. Whereas phylogenies of individual alignment blocks differ substantially from the core tree, phylogenies reconstructed from a large number of blocks are highly similar to the core tree. **Top left:** For each split (i.e. branch) in the core tree, the color indicates what fraction of the phylogenies of 3 Kb blocks support that bi-partition of the strains. **Top right:** Cumulative distribution of branch support, i.e. fraction of 3 Kb blocks supporting each branch. The dotted lines indicated show the fraction of branches that have less than 50% support (yellow) and the median support per branch (purple). **Bottom left and bottom right :** As in the top row, but now based on phylogenies reconstructed from random subsets of 50% of all 3 Kb blocks as opposed to individual blocks.

obtains the same or highly similar phylogenies, e.g. [8–10]. We also observe this behavior for our strains. Phylogenies reconstructed from a random sample of 50% of all 3Kb blocks look highly similar to the core tree, i.e. with two thirds of the core tree's splits occurring in *all* phylogenies (Fig. 1, bottom row).

How should we interpret this convergence of phylogenies to the core phylogeny as increasing numbers of genomic loci are included? One interpretation, proposed by some researchers, is that once a large number of genomic segments is considered, effects of horizontal transfer are effectively averaged out, and the phylogeny that emerges corresponds to the clonal ancestry of the strains, e.g. [8, 19]. Based on this idea, several tools have been developed that detect recombination events by comparing local phylogenies with the overall reference phylogeny constructed from the entire genome [20, 21]. In contrast, some recent studies have argued that recombination is so common in some bacterial species that it is impossible to meaningfully reconstruct the clonal ancestry from the genome sequences, and that these species should be considered freely recombining, e.g. [22]. However, if members of the species are freely recombining, one would expect the core tree to take on a star-like structure as opposed to the clear and consistent phylogenetic structure that phylogenies converge to as more genomic regions are included in the analysis. Addressing this puzzle is one of the topics of this work.

Quantifying recombination through analysis of pairs of strains

As a first analysis of the impact of recombination, we follow an approach recently proposed by Dixit et al. based on the pairwise comparison of strains [11]. The simplest measure of the distance between a pair of strains is their nucleotide divergence, i.e. the fraction of mismatching nucleotides in the core genome alignment of the two strains. For pairs of strains with very low divergence, e.g. D6 and F2 at 4×10^{-4} divergence (Fig. 2A), the effects of recombination are almost directly visible in the pattern of SNP density along the genome. While the SNP density is very low along most of the genome, i.e. $0 - 2$ SNPs per kilobase, there are a few segments, typically tens of kilobases long, where the SNP density is much higher and similar to the typical SNP density between random pairs of *E. coli* strains, i.e. $10 - 30$ SNPs per kilobase. These high SNP density regions almost certainly result from horizontal transfer events in which a segment of DNA from another *E. coli* strain, for example carried by a phage, made it into one of the ancestor cells of this pair, and was incorporated into the genome through homologous recombination. For pairs of increasing divergence, e.g. the pair C10-D7 with divergence 0.002 in Fig. 2B, the frequency of these recombined regions increases, until eventually the majority of the genome is covered by such regions (pair D6-H10 in Fig. 2C).

For close pairs, the histograms of SNP densities also clearly separate into two components: a majority of clonally inherited regions with up to at most 3 SNPs per kilobase, and a long tail of recombined regions with up to 50 or 60 SNPs per kilobase (Fig. 2D-E). As explained in the methods, we can accurately model the distributions of SNP densities as a mixture of a Poisson distribution for the clonally inherited regions plus a negative binomial for the recombined regions (solid line fits in Fig. 2D-F). In this way we can estimate, for each pair of strains, the fraction of the genome that is clonally inherited, and the number of SNPs that fall in clonally inherited versus recombined regions. Using a Hidden Markov model on close pairs, we also estimated the distribution of lengths of recombined regions (see Methods), finding that recombined blocks are typically in the range of $10 - 70$ kilobases long (Fig. 2J).

From this analysis we see that, whenever the pairwise divergence is less than 0.001, the large majority of blocks is clonally inherited, which is indicated as the light-green segment in Fig. 2G. However, over a narrow range of divergence between 0.001 and 0.01 the fraction of clonally inherited DNA drops dramatically (yellow segment in Fig. 2G) and at a divergence of about 0.014 essentially the entire alignment has been overwritten by recombination and all clonally inherited DNA is lost (blue segment in Fig. 2G). Notably, 80% of all pairs of strains lie in this fully recombined regime (Fig. 2I). Thus, for the large majority of pairs of strains, none of the DNA in their alignment derives from their clonal ancestor, making it impossible to estimate the distance to their clonal ancestor from comparing their DNA. Moreover, as shown in Fig. 2H, even for pairs that are so close that most of their genomes are

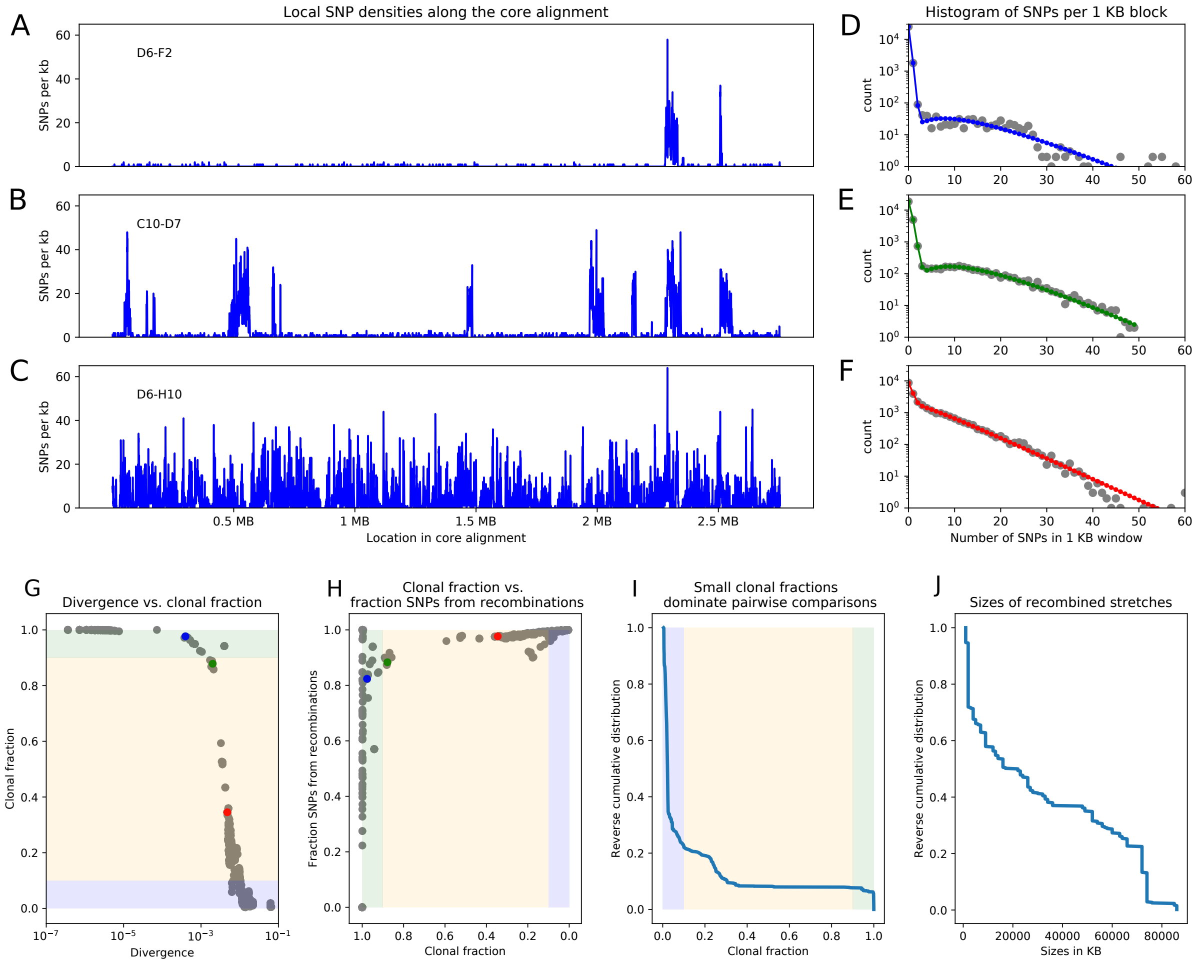

Figure 2. Summary statistics of pairwise analysis for the SC1 strains. A-C: SNPs densities (per kilobase) along the core genome for three pairs of strains at overall nucleotide divergences of 4×10^{-4} (D6-F2), 0.002 (C10-D7), and 0.0048 (D6-H10). **D-F**: Corresponding histograms for the number of SNPs per kilobase (dots) together with fits of the mixture model for D6-F2 (blue), C10-D7 (green), and D6-H10 (red). Note the vertical axis is on a logarithmic scale. **G**: For each pair of strains (dots), the fraction of the genome that was inherited clonally is shown as a function of the nucleotide divergence of the pair, shown on a logarithmic scale. The three pairs that were shown in panels A-F are shown as the blue, green, and red dots. The light green, yellow, and blue segments show strains that are mostly clonal, a mixture of clonal and recombined, and fully recombined, respectively. **H**: Fraction of all SNPs that lie in recombined regions as a function of the clonally inherited fraction of the genome. **I**: Reverse cumulative distribution of the clonal fractions of the pairs. **J**: Reverse cumulative distribution of the lengths of recombined segments for pairs that are in the mostly clonal regime. The mean length of recombined regions is $31'197$, with first quartile 2000, median $19'500$, and third quartile $66'000$.

clonally inherited, the large majority of the SNPs derives from the recombined regions.

For later comparison with the data on other species, we summarize our observations from the pairwise analysis by a few key statistics. First, half of the genome is recombined at a *critical divergence* of 0.0032. Second, at this critical divergence, the fraction of all SNPs that is in recombined regions is 0.95. Third, the fraction of mostly clonal pairs is 0.077, and finally, the fraction of fully recombined pairs is 0.78 (see Methods). All these statistics suggest that pairwise divergences between strains are almost entirely driven by recombination and do not reflect distances to their clonal ancestors. To understand how a consistent phylogenetic structure can still emerge when the full core genomes of all strains are compared, we need to go beyond studying pairs.

SNPs in the core genome alignment correspond to splits in the local phylogeny

Whereas there may not be a single phylogeny that captures the evolution of our genomes, we will assume that each *single position* in the core genome alignment, i.e. each alignment column, has evolved according to some phylogenetic tree. A key insight is that our set of strains is sufficiently closely related that, for almost all of these alignment columns, the number of substitutions that have occurred in their evolutionary history is either zero or one. In particular, of the $2'457'464$ columns in the core genome alignment, only 10.85% are polymorphic. Moreover, almost all of these SNP columns are bi-allelic, i.e. for 93.6% of the SNPs only 2 nucleotides appear, 6.3% have 3 nucleotides, and in 0.2% all 4 nucleotides occur, suggesting that most positions have not undergone any substitutions, and that columns with multiple substitutions are rare. Notably, these statistics are still inflated due to the occurence of an outgroup of 9 strains that is far removed from the other strains (the clade from B5 to B3 visible on the right in Fig. 1). We observe that almost 36% of all SNPs correspond to SNPs in which all 9 strains of this outgroup have one nucleotide, and all other 83 strains have another nucleotide. If we remove the outgroup from our alignment, the fraction of SNP positions in the alignment drops from 10.85% to 6.7%, and the fraction of SNPs that are bi-allelic increases to 95.5%.

We analyzed the frequencies of columns with 1, 2, 3 and 4 different nucleotides that are expected under a simple substitution model, separately analyzing positions that are under least selection (third positions of 4-fold degenerate codons) and positions under most selection (second positions in codons), and either including or excluding the outgroup (see Methods). These analyses indicate that around 98% of all bi-allelic SNP columns correspond to columns in which only a single substitution took place.

Since almost all bi-allelic SNPs correspond to a single substitution, each such SNP provides an important piece of information about the phylogeny at that position in the alignment: whatever this phylogeny is, it must contain a split, i.e. a branch bipartitioning the set of strains, such that all strains with one letter occur on one side of the split, and all strains with the other letter on the other side (Fig. 3).

As illustrated in Fig. 3, pairs of SNPs can either be consistent with a common phylogeny, i.e. columns X and Y or columns Y and Z, or they can be inconsistent with a common phylogeny, i.e. columns X and Z. The pairwise comparison of SNP columns for consistency with a common phylogeny is known as the four-gamete test in the literature on sexual species [12] but has so far rarely been used for quantifying recombination in bacteria (see [23] for the only exception we are aware of). In the rest of this paper we show how analysis of bi-allelic SNPs (which from now on we will just call SNPs) can be systematically used to quantify recombination in bacterial species.

SNP statistics are inconsistent with a single consensus phylogeny

As a first test, we investigated to what extent the SNPs support the branches in the core tree. Since each branch in the core tree corresponds to a split, we calculated what fraction of SNPs correspond to a branch in the core tree, and what fraction are inconsistent with the core tree. Overall, 58% of the SNPs that are shared by at least 2 strains correspond to a branch of the core tree, whereas 42% clash with it

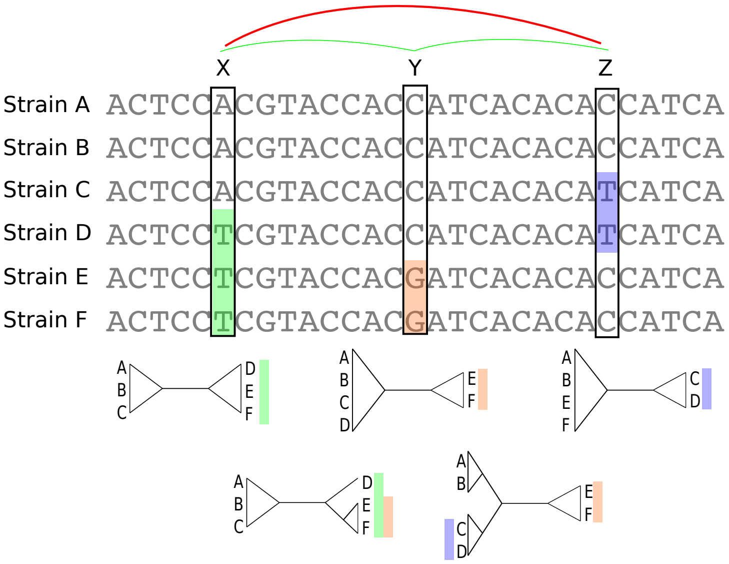
X
Y
Z
Strain A ACTCCACGTACCACCATCACACACCATCA
Strain B ACTCCACGTACCACCATCACACACCATCA
Strain C ACTCCACGTACCACCATCACACATCATCA
Strain D ACTCCTCGTACCACCATCACACATCATCA
Strain E ACTCCTCGTACCACGATCACACACCATCA
Strain F ACTCCTCGTACCACGATCACACACCATCA
A B C
D E F
A B C D
E F
A B E F
C D
A B C
D E F
A B
C D
E F

Figure 3. SNP columns correspond to phylogeny splits. A segment of a multiple alignment of 6 strains containing 3 bi-allelic SNPs, X, Y, and Z. The 3 diagrams below the alignment show that each SNP constrains the local phylogeny to contain a particular split, i.e. bi-partition of the strains. In this example, the neighboring pairs of SNPs (X, Y) and (Y, Z) are both consistent with a common phylogeny and can be used to further resolve the phylogeny in the local segment of the alignment as shown in the bottom two diagrams. However, SNPs X and Z are mutually inconsistent with a common phylogeny indicating that somewhere between X and Z a recombination event must have occurred.

(SNPs that occur in only a single strain are consistent with any phylogeny). However, this relatively high fraction results almost entirely from SNPs on the single branch connecting the outgroup to the other strains, which is responsible for almost 36% of all SNPs. When the outgroup is removed, only 27.4% of all SNPs are consistent with the core tree. Since the core tree was constructed using a maximum likelihood approach that assumes the entire alignment follows one common tree, we investigated to what extent the number of tree supporting SNPs can be improved by specifically constructing a tree to maximize the number of supporting SNPs (see Methods). However, such trees only marginally improve the number of supporting SNPs by 0.1%.

To assess the extent to which SNPs are consistent with individual branches of the core tree we counted, for each branch, the number of supporting SNPs S that match the split, and the number of clashing SNPs C that are inconsistent with the split, to calculate the fraction $f = S/(S + C)$ of SNPs supporting the branch. Figure 4 shows that, for two-thirds of the branches, there are more clashing than supporting SNPs. Moreover, for as many as half of the branches in the core tree, the fraction of supporting SNPs is less than 5%, i.e. there are 20-fold more clashing than supporting SNP columns.

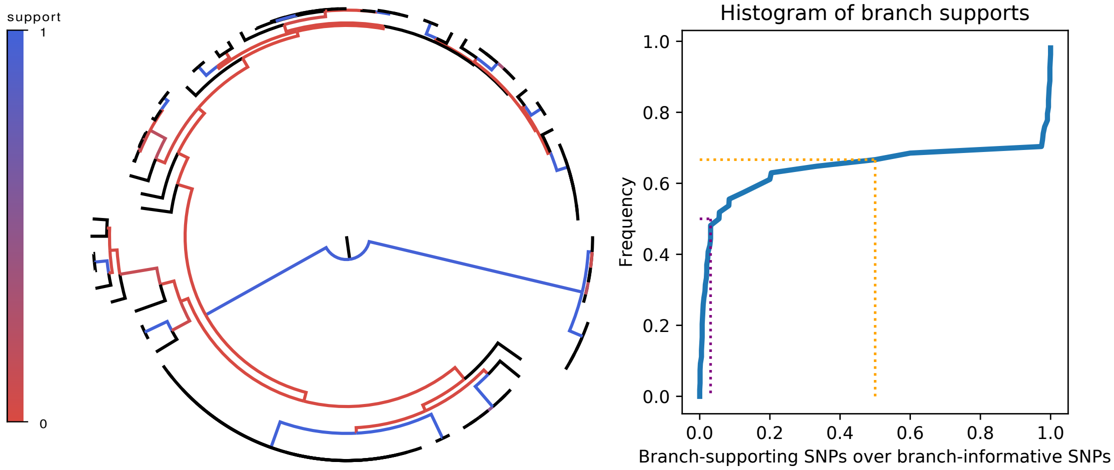

Figure 4. SNP support of the branches of the core genome tree. **Left Panel**: Fraction of supporting versus clashing SNPs for each branch of the core tree. **Right panel**: Cumulative distribution of the fraction of supporting SNPs across all branches. The purple and orange dotted lines show the median and the frequency of branches with 50% or less support, respectively.

Besides the brach to the outgroup, the only branches for which supporting SNPs outnumber clashing SNPs are branches toward groups of highly similar strains near the bottom of the tree. We thus wondered if it would be possible to construct well supported subtrees for clades of closely-related strains near the bottom of the tree. We devised a method that builds subtrees bottom-up by iteratively fusing clades so as to minimize the number of clashing SNPs at each step (see Methods and Suppl. Fig S3, left panel). As shown in Suppl. Fig. S3, while the fraction of clashing SNPs is initially low, it rises quickly as soon as the average divergence within the reconstructed subtrees exceeds 10^{-4}, which is more than 100-fold below the typical pairwise distance between *E. coli* strains. Thus, while some groups of very closely-related strains can be unambiguously identified, only a minute fraction of sequence divergence falls within these groups, and the bulk of the sequence variation between the strains is not consistent with a single phylogeny.

It is also conceivable that there is a single dominant phylogeny for most strains, but that this is concealed from view when analyzing the full alignment because of a subset of strains with aberrant behavior. To investigate this, we focused on the smallest subsets of strains that have meaningfully different phylogenetic tree topologies. For a quartet of strains (I, J, M, N), there are 3 possibly binary trees, i.e with (I, J) and (M, N) nearest neighbors, with (I, M) and (J, N), or with (I, N) and (J, M) (See Suppl. Fig. S4). We selected quartets of roughly equidistant strains and checked, for each quartet, whether the SNPs clearly supported one of the tree possible topologies. However, we find that alternative topologies are always supported by a substantial fraction of the SNPs, and that for most quartets the most supported topology is supported by less than half of the SNPs (Suppl. Fig S4).

Thus, consistent with our analysis of pairs of strains, all these results show that the core tree does not capture the sequence relationships between the strains. In fact, rather than a single phylogeny representing the evolutionary relationships between the strains, the SNP data suggest a large number of different phylogenies across the core genome alignment. It may thus seem all the more puzzling that, when trees are constructed from sufficiently many genomic loci, the core tree reliably emerges (Fig. 1, bottom). To underscore this puzzle, we observed that if we remove all SNP columns from the core genome alignment that correspond to branches of the core tree, and then reconstruct a phylogeny from this edited alignment, the resulting tree is still highly similar to the core tree (Suppl. Fig. S5). However, almost *all* SNPs of this edited alignment clash with the tree that is reconstructed from it. Thus, the core tree reconstructed from an alignment does not need to match the phylogeny of any of the genomic loci. Rather, the core tree represents some sort of *average* of the distribution of phylogenies across the genome. Note that, whenever a quantity x has a multi-modal distribution, it can easily occcur that there is almost no probability for any sample of x to occur near its average $\langle x \rangle$. Similarly, the actual phylogenies occurring across the core genome alignment may all be very different from the global 'average' phylogeny that the core tree represents.

Phylogeny changes every few dozens of base pairs along the core alignment

So far we have analyzed SNP consistency without regard to their relative positions. We now analyze to what extent mutually consistent SNPs are clustered along the alignment. In particular, we calculate the lengths of segments along the alignment that are consistent with a single phylogeny.

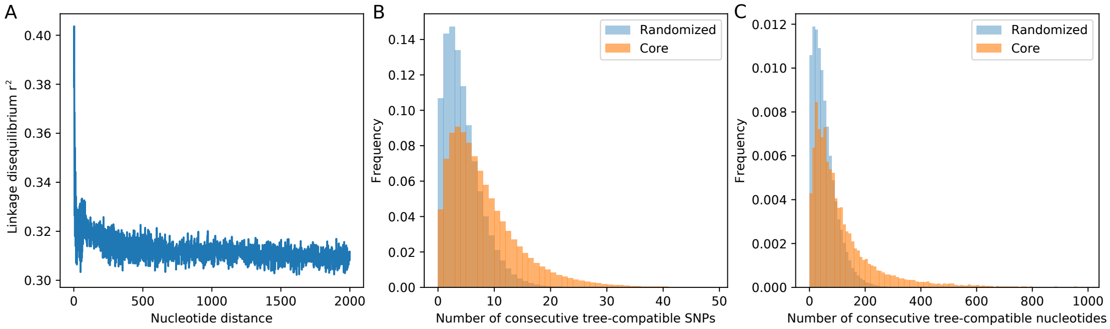

Figure 5. SNP compatibility along the core genome alignment. **A**: Linkage disequilibrium (squared correlation, see Methods) as a function of the separation of a pair of columns in the core genome alignment. **B**: Probability distribution of the number of consecutive SNP columns that are consistent with a common phylogeny for the core genome alignment (orange) and for an alignment in which the position of all columns has been randomized (blue). **C**: Probability distribution of the number of consecutive alignment columns consistent with a common phylogeny for both the real (orange) and randomized alignment (blue).

We first asssessed the length-scale over which phylogenies are correlated by calculating a standard linkage measure as a function of distance along the alignment (Fig. 5A and Methods). Linkage drops quickly over the first 100 base pairs and becomes approximately constant at distances beyond $200 - 300$ base pairs, indicating that segments of correlated phylogenies are much shorter than the typical length of a gene. Very short linkage profiles were recently also observed in thermophilic cyanobacteria isolated in Yellowstone National Park [22].

We next determined the lengths of segments consistent with a common phylogeny. Starting from each SNP s, we determined the number of consecutive SNPs n that are all mutually consistent with a common phylogeny. As shown in Fig. 5B, the distribution of tree-compatible stretches has a mode at $n = 4$, and stretches are very rarely longer than 20 consecutive SNPs. In terms of number of base pairs along the genome, tree-compatible segments are typically just a few tens of base pairs long, and very rarely more than 300 base pairs (Fig. 5C). Thus, stretches of tree-compatible segments are very short. For comparison, we also calculated the distribution of tree-compatible segment lengths in an alignment where the positions of all columns have been completely randomized and observe that these are still a bit shorter (blue distributions in Fig. 5). Thus, while there is some evidence that neighboring SNPs are more likely to be compatible than random pairs of SNPs, this compatibility is lost very quickly, typically within a handful of SNPs.

A lower bound on the ratio of recombination to substitution events

Every time inconsistent SNP columns are encountered as one moves along the core genome alignment, the local phylogeny must change. For example, somewhere between columns X and Z in Fig. 3 the phylogeny must change. This in turn implies that at least one recombination event must occur between columns X and Z. By going along the core genome, and determining the minimum number of times the phylogeny must change, one can thus derive a lower bound on the total number of recombination events [12] (see Methods). Using this we find that the phylogeny must change at least $C = 43'575$ times along the core phylogeny, i.e. there are at least C recombination events. If we denote by R the true total number of recombination events, then we can write $C = Rf$, where f can be thought of as the fraction of recombination events that are detected by SNP inconsistencies in the alignment. As we argued previously, almost all SNP columns correspond to a single substitution event, such that the total number of SNP columns M is a good estimate of the total number of substitutions in the alignment. Consequently, the ratio of phylogeny changes C to SNPs M provides a lower bound on the ratio of recombinations to mutations in the alignment, i.e.

$$\frac{C}{M} = f\frac{R}{M} < \frac{R}{M}. \tag{1}$$

Figure 6 shows the ratio C/M for random subsets of our 92 strains as a function of the number of strains in the subset.

We see that, for small subsets of strains, the recombination to mutation ratio C/M shows substantial fluctuations. For example, for subsets of $n = 10$ strains, the recombination to mutation ratio C/M ranges from 0.036 to 0.167, with a median of 0.1. However, as the number of strains in the subset increases, the recombination to mutation ratio converges to a value of $C/M \approx 0.155$. In particular, whenever there is a substantial fraction of the strains, i.e. $n \geq 50$, the ratio C/M is highly consistent across the subsets. Thus, the ratio C/M gives a highly informative summary statistic of the relative rate of recombination to mutation events along the alignment.

These results confirm that, also on the level of the entire alignment, the strains are in a regime where each position has been affected by recombination. For example, given the ratio $C/M = 0.155$, and the overall SNP rate of 0.1085, the average length of aligment segments between changes in phylogeny is $1/(0.1550.1085) \approx 59.4$ base pairs. From the analysis of close pairs we saw that the typical length of a recombined segment is about $20'000$ base pairs (Fig. 2J). Thus, as an order of magnitude estimate, a given position in the genome has been overwritten roughly $20'000/59.4 \approx 337$ times by recombination events. Moreover, since we only detect a fraction of the phylogeny changes across the alignment, the true number of times each locus has been overwritten by recombination is likely considerably higher.

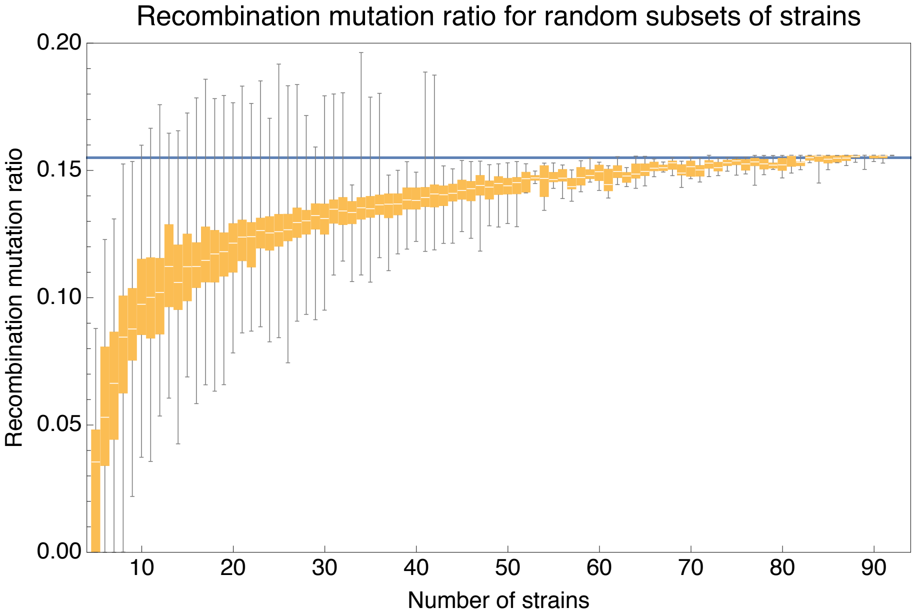

Figure 6. Ratio C/M of the minimal number of recombinations C to mutations M for random subsets of strains. For strain numbers ranging from $n = 4$ to $n = 92$, we collected random subsets of n strains and calculated the ratios C/M of phylogeny changes to mutations in the alignment. The figure shows box-whisker plots that indicate, for each strain number n, the 5th percentile, first quartile, median, third quartile, and 95th percentile of the distribution of C/M across subsets. The blue line shows $C/M = 0.155$.

Although, it is tempting to interpret the ratio C/M as an estimate of the relative rates of recombination and mutation in the evolution of the strains, this would require defining a specific evolutionary model, and even for simple models, e.g. a Kingman coalescent with a fixed rate of recombination [24], the relationship between the ratio of recombination and mutation rates, and the observed ratio C/M would be nontrivial. Moreover, we will see below that the data in fact suggests that recombination rates vary over a wide range across lineages.

Recombination rates across lineages follow scale-free distributions

The analyses above have shown that the core alignment consists of tens of thousands of short segments with different phylogenies. Thus, one approximate way of thinking about the core genome alignment is that the phylogeny at every genomic locus is drawn from some distribution of possible phylogenies. In a freely recombining population, every strain would be equally likely to recombine with any other, leading to a uniform distribution over all possible phylogenies. However, under such a model each pair of strains would become approximately equidistant and phylogenies build from large numbers of genomic loci would take a star-shape, which is clearly at odds with our observations. This suggests that the phylogenies along the genome are drawn from a highly non-uniform distribution in which some lineages are more likely to have recombined recently than others.

The distribution of observed SNP types in fact contains extensive information about the relative frequencies with which different lineages have recombined at different times in the past. For example, imagine a SNP where two strains share a nucleotide which differs from the nucleotide that all other strains possess. We will denote such SNPs as 2-SNPs or pair-SNPs. If, at some genomic locus g, we find a 2-SNP shared by strains s_1 and s_2, then it follows that, whatever the phylogeny is at locus g, the strains s_1 and s_2 must be nearest neighbors in the tree, and the SNP corresponds to a mutation that occurred on the branch connecting the ancestor of s_1 and s_2 to all other strains.

Thus, to quantify to which extent the lineage of a strain s has recently recombined with the lineages of the other strains, we can extract all 2-SNPs in which s shares a letter with one other strain s' and compare their frequencies. For example, Fig. 7A graphically shows the frequencies of all pair-SNPs $(A1, s)$ in which A1 shares a SNP with one other strain s. Note that, if there was a dominant clonal phylogeny, then A1 should essentially only have 2-SNPs with its nearest neighbor in this dominant phylogeny. However, we see that A1 shares 2-SNPs with 17 of the 92 strains in our collection. If, one the other hand, A1 were freely recombining with all other strains, then we would expect roughly equal frequencies of all possible 2-SNPs $(A1, s)$. However, we see that A1 shares 2-SNPs with some strains much more often than with others. For example, whereas 2-SNPs with strains A2, A11, and D8 are the most frequent and occur almost 200 times each, for 11 of the 17 strains the number of occurrences is 10 or less, and for 4 strains a 2-SNP with A1 is observed only once.

Figure 7B shows a graph representation of all observed pair-SNPs, with the thickness of the edges proportional to the logarithm of the frequency of occurrence of the 2-SNP type. We see that each strain is connected through 2-SNPs to a substantial number of other strains, indicating a high diversity of recent recombination events across the strains. At the same time, the large variability in the thickness of the edges indicates that some pairs occur much more frequently than others. Figure 7C shows the reverse cumulative distribution of the frequencies of all observed 2-SNPs, i.e. the distribution of the thickness of the edges in Fig. 7B (blue dots). Note that, if the strains were to recombine freely, each 2-SNP would be equally likely to occur, and the distribution of 2-SNPs would be peaked around a typical number of occurrences per type. Instead, we see that 2-SNP frequencies f vary over more than 3 orders of magnitude, i.e. from an occurrence of just $f = 1$ for many 2-SNPs to $f = 2965$ occurrences for the most common 2-SNP. Moreover, the reverse cumulative distribution of 2-SNP frequencies follows an approximate straight-line in a log-log plot. In other words, the distribution of frequencies $P(f)$ is approximately power-law, i.e $P(f) \propto f^{-\alpha}$. Fitting the 2-SNP data to a power-law (see Methods) we find that the exponent equals approximately $\alpha \approx 1.41$ (blue line in Fig. 7C). Importantly, this means that there is no clear most common 2-SNP

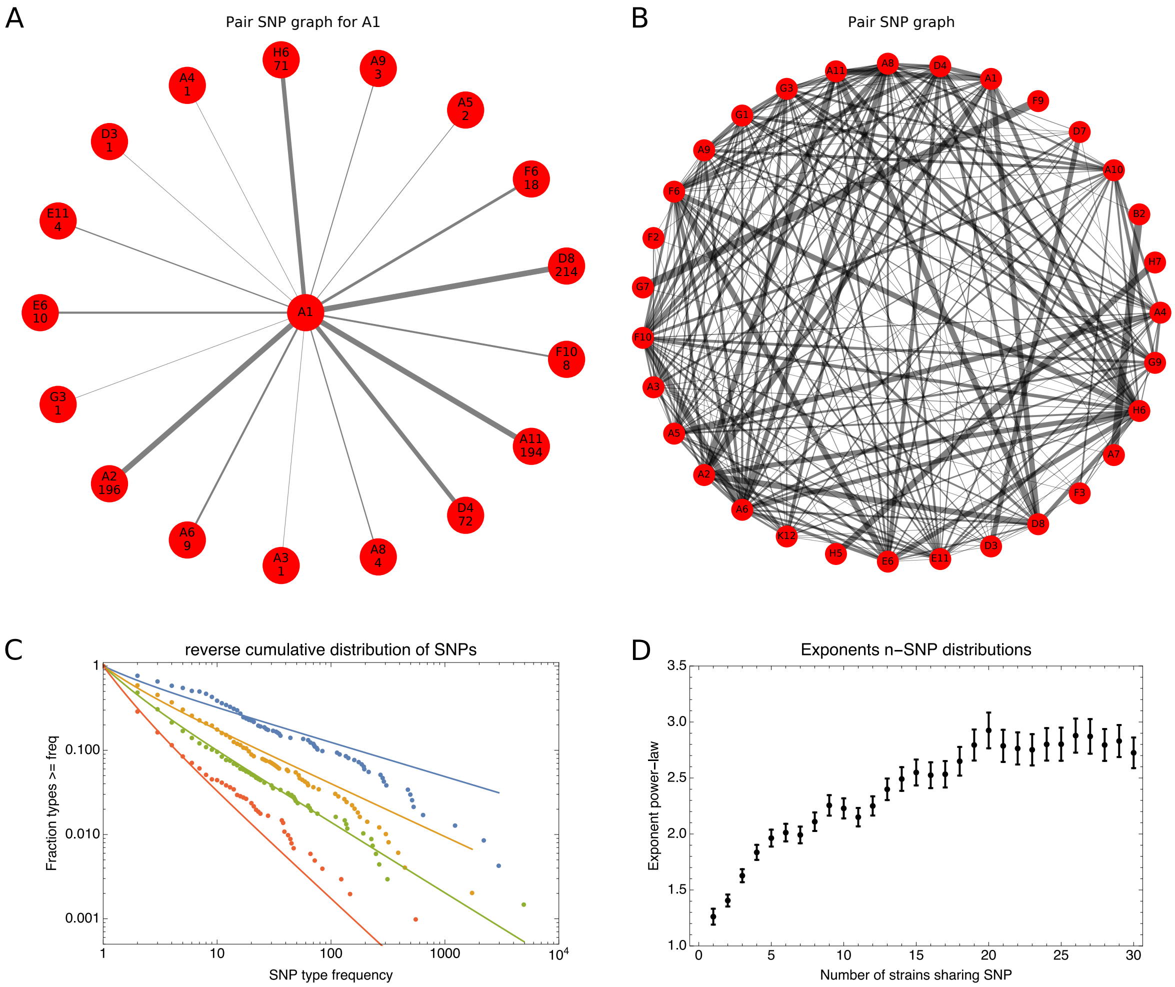

Figure 7. SNP-type frequencies follow approximately power-law distributions Distributions of the frequencies with which different SNP patterns occur. **A**: Frequencies of 2-SNPs of the type $(A1, s)$ in which a SNP is shared between strain A1 and one other strain s. Each edge corresponds to a 2-SNP $(A1, s)$ and the thickness of the edge is proportional to the logarithm of the number of occurrences of the 2-SNP. The frequency of each edge is also indicated at the corresponding outer node. **B**: A graph showing all 2-SNPS (s, s') that were observed in the core genome alignment. Each node corresponds to a strain and each edge to a 2-SNP, with the thickness of the edge proportional to the logarithm of the number of occurrences of the SNP. **C**: Reverse cumulative distributions of the frequencies of all observed 2-SNPs (blue dots), 3-SNPs (orange dots), 4-SNPs (green dots), and 12-SNPs (red dots). The solid lines in corresponding colors show power-law fits. Both axes are shown on a logarithmic scale. **D**: Exponents of the power-law fits to the n-SNP frequency distributions, as a function of the number of strains sharing a SNP n. Error bars correspond to 95% posterior probability intervals.

partner for each strain, and that one cannot naturally divide 2-SNPs into common and rare types. Instead, the distribution of 2-SNP frequencies is approximately scale-free.

Beyond SNPs shared by pairs of strains, we can of course also look at SNPs shared by triplets, quartets, and so on. Besides the distribution of 2-SNP frequencies, Fig. 7C also shows the reverse cumulative distributions of 3-SNPs (orange dots), 4-SNPs (green dots), and 12-SNPs (red dots). We see that all these distributions follow approximately straight lines in a log-log plot and can be fitted with power-law distributions (solid lines). The n-SNP distributions drop more steeply as n increases. Figure 7D shows the exponent α of the n-SNP distribution as a function of n, showing that the exponents range from $\alpha \approx 1.25$ for singlets, i.e. $n = 1$, to $\alpha \approx 2.8$ for $n \geq 20$.

We find that essentially all n-SNP distributions are approximately scale-free, i.e. can be fitted with power-laws. Thus, while some subgroups of n strains share a common ancestor much more often than others subgroups of n strains, their frequencies fall along a scale-free continuum, so that there is no natural way of dividing the strains into groups of 'highly recombining' clades. Note also that each n-SNP corresponds to a mutation that occurred in the branch leading to the ancestor of a group of n strains. Therefore, n-SNPs for larger n typically correspond to mutational events that occurred further back in time. The fact that n-SNP distributions become more steep as n increases means that the average number of occurrences per n-SNP decreases as n increases. Thus, the diversity of n-SNPs tends to be larger further back in time (see Suppl Fig. S6).

Phylogenetic entropy profiles of individual strains

Another way to think about the structure evident in the n-SNP distributions is to quantify, for each strain s, how diverse the phylogenies are that s occurs in at different n. In particular, for a given n, all n-SNP types in which s is one of the strains sharing the minority nucleotide, are all mutually inconsistent with a common tree. For example, if the strain s occurs in 10 different quartets of strains, i.e. in 10 different 4-SNP types, then each of these 10 quartets must correspond to different phylogenies and the diversity of quartets among which s occurs can be quantified by the entropy of the frequency distribution of 4-SNP types in which s occurs. That is, for each strain s, and each n, we can extract all n-SNPs in which s occurs and calculate their relative frequencies, and then summarize the diversity of phylogenies by the *entropy* of this distribution across n-SNPs. In this way, for each strain s we can calculate an entropy profile $H_s(n)$ that contains the entropies of the n-SNP distributions in which strain s occurs, as a function of n (see Methods). Supplementary Fig. S7 shows the entropy profiles for 5 example strains, as well as the distribution of entropy profiles across all strains. We see that the entropy generally increases as n increases, again indicating that the diversity of phylogenies increases as one goes further back in time. The entropy profiles are highly diverse, e.g. for strains like A10 and H6 the entropy increases quickly to $5 - 6$ bits, while for the strain G8, which belongs to a cluster of 20 strains that are extremely closely related, the entropy only increases for $n > 20$. Most significantly, each strain s has an essentially unique entropy profile $H_s(n)$, showing that each strain has its own 'fingerprint' of the frequencies with which its lineage shares recent ancestors with the other strains. Finally, the entropy profiles become more similar as n increases, and for large n the entropy converges to roughly 7.5 bits, which corresponds to effectively 180 different possible ancestries per strain.

Other species of bacteria exhibit qualitatively similar statistics

To investigate to what extent the observations we made for *E. coli* generalize to other species of bacteria, we selected 5 additional species from different bacterial groups for which sufficiently many complete genome sequences of strains were available, and used REALPHY to obtain a core genome alignment of the strains for each species (see Table S1 for a list of the species, the number of strains, and other core genome statistics for each species). We then performed most of the analyses that we presented above for *E. coli* on each of these core alignments. Figure 8 presents a summary of the results that we observe across the species.

Figure 8A shows the cumulative distributions of pairwise divergences between strains for all species. We see that, while among our *E. coli* strains that were sampled from a common habitat there is a small percenage of very close pairs with divergence around 10^{-6}, for the strains of the other species the closest pairs are at divergence 10^{-5}. With the exception of *M. tuberculosis*, where the median pair divergence is around 10^{-4}, the median pairwise divergence in all other species is around 10^{-2} or larger. The vertical lines in Fig. 8A indicate the critical divergences, for each species, where half of the alignment is recombined. With the exception of *M. tuberculosis*, where all pairs are mostly clonal, the critical divergences lie in a fairly narrow range of $0.003 - 0.01$. Figure 8B shows the reverse cumulative distributions, across pairs of strains, of the fraction of the alignment that is clonally inherited, i.e. as for Fig. 2I for *E. coli*. Note that, for all species except *M. tuberculosis*, the large majority of the pairs is fully recombined. For *H. pylori* the fraction of pairs that still contain clonally inherited DNA is almost zero, whereas for *S. aureus* the fraction of pairs with a substantial fraction of clonally inherited DNA is largest. Thus, we see that for almost all species the situation is similar to what we observed in *E. coli*: for most pairs the distance to their common ancestor cannot be estimated from their alignment, because the entire alignment has been overwritten by recombination events. Note also that, for all species, there is only a relatively small fraction of pairs that lie in the partially recombined regime (yellow segment in Fig. 8B).

Figure 8C shows, for each species, the fraction of all SNPs that derive from recombination, for pairs of strains that are at the critical divergence where half of the alignment is recombined. Even though this critical divergence occurs for pairs that are relatively close compared to the typical distance between pairs, for all species more than 90% of the SNPs derive from recombination. That is, we also see that for all 5 species the divergence between close strains is dominated by SNPs that are introduced through recombination.

Figure 8D summarizes the distributions of support of the branches of the core tree as violin plots, i.e. as shown for *E. coli* as a cumulative distribution in Fig. 4. In *E. coli* most branches have many more SNPs that reject the split than support it, and even stronger rejection of the branches of the core tree are observed for *B. subtilis* and *H. pylori.* For the other three species, including *M. tuberculosis*, an almost uniform distribution of branch support is shown, i.e. for these species there are roughly as many branches that are strongly supported by the SNPs, strongly rejected by the SNPs, or supported and rejected by roughly equally many SNPs.

Figure 8E summarizes, for each species, the distribution of distances between SNPs along the core alignment as box-whisker plots (green) as well as the distribution of distances between phylogeny breakpoints (blue), i.e. as shown in Fig. 5C for *E. coli.* The figure shows that, with the exception of *M. tuberculosis*, the inter-SNP distances range from a few to a few dozen basepairs, with a median inter-SNP distance of 4 (*H. pylori*) to 15 (*S. aureus*) base pairs. For these 5 species, the median distances between phylogeny breakpoints range from around 10 (*H. pylori*) to about 100 base pairs for *S. aureus.* Note that, for all species, the tail of the distributions stretches to very short distances between breakpoints, whereas distances between breakpoints of more than 200 bps are very rare for all these 5 species. Thus, for these species the segments that are consistent with a single phylogeny are always much shorter than the typical length of a gene. In contrast, for *M. tuberculosis* both the distances between SNPs and the distances between breakpoints are almost two orders of magnitude larger.

Finally, Fig. 8F shows box-whisker plots for the distribution of the number of consecutive SNPs between breakpoints, as was shown for *E. coli* in Fig. 5B. We see that for all species, including *M. tuberculosis*, there are typically less than a handful of SNPs in a row before a phylogeny breakpoint occurs, and very rarely more than a dozen SNPs. The smallest number of SNPs per breakpoint is observed for *H. pylori*, i.e. typically less than 2 SNPs per breakpoint, but the range of SNPs per breakpoint is very similar across all species.

We next investigated whether the n-SNPs of the other species also exhibit approximately power-law

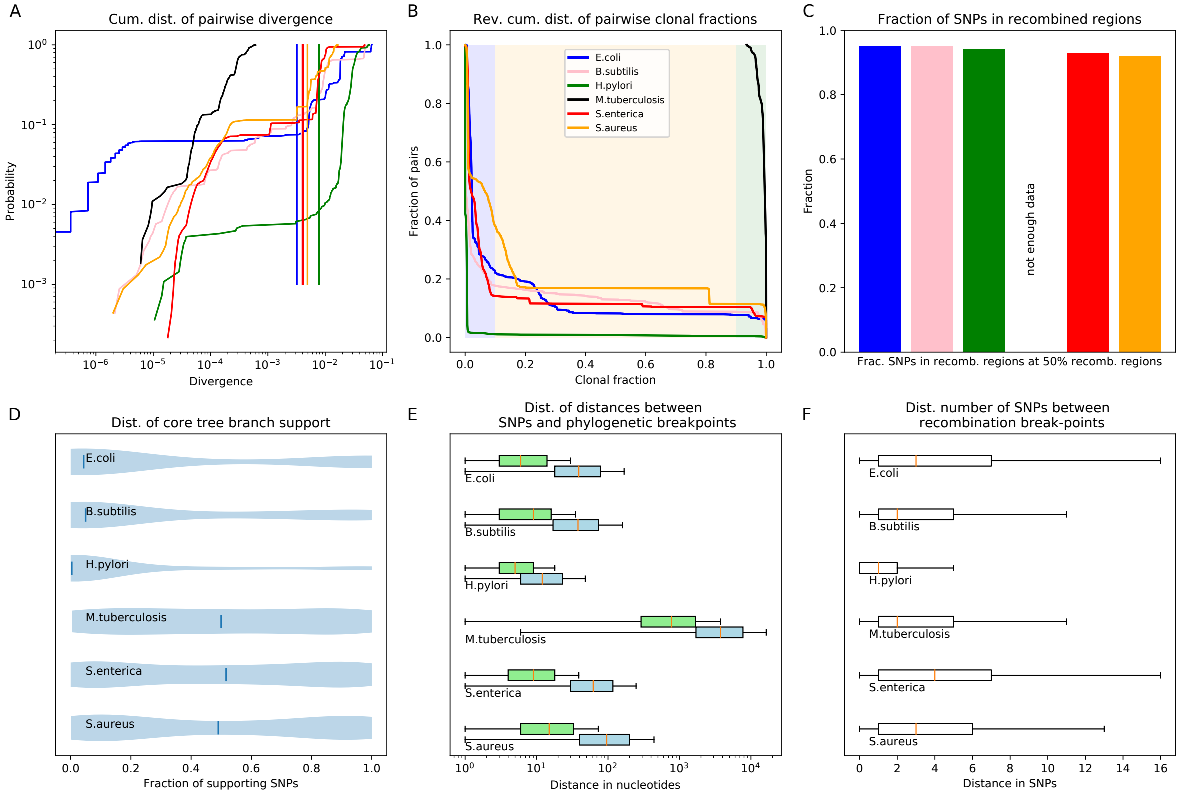

Figure 8. Summary of the results across species. **A**: The cumulative distribution of pairwise divergernces is shown as a different colored line for each species (see legend in panel). Both axes are shown on logarithmic scale. The vertical lines in corresponding colors show the critical divergence at which half of the genome is recombined for each species. **B**: Reverse cumulative distribution of clonal fractions across the pairs of strains of each species, with the green, yellow, and blue shaded regions indicating the mostly clonal, partly recombined, and fully recombined regimes, respectively, i.e. analogous to Fig. 2I **C**: For each species, the height of the bar shows the fraction of SNPs that fall in recombined regions for pairs of strains for which half of the genome is recombined, i.e. see Fig. 2H. **D**: The violin plots show, for each species, the distribution of branch support, i.e. the relative ratio of SNPs supporting or clashing with each branch split, analogous to the right panel of Fig. 4. The blue lines correspond to the medians of the distributions. **E**: Box-whisker plots showing the 5, 25, 50, 75, and 95 percentiles of the distributions of nucleotide distances between consecutive SNPs (green) and phylogeny breakpoints (blue, i.e. analogous to Fig. 5C), for each species. The axis is shown on a logarithmic scale. **F**: Box-whisker plot of the distribution of the number of consecutive SNPs in tree-compatible segments, i.e. analogous to Fig. 5B.

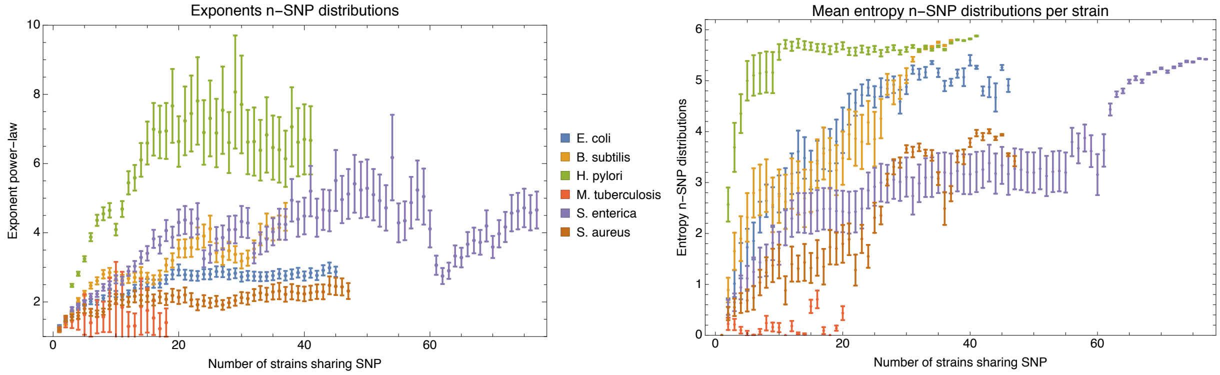

Figure 9. Left panel: Exponents of the power-law fits to the n-SNP frequency distributions, as a function of the number of strains sharing a SNP n for each of the species (different colors). Error bars correspond to 95% posterior probability intervals. **Right panel**: Mean entropy of the entropy profiles $H_n(s)$, averaged over all strains s, as a function of the number n of strains sharing the SNP, for each of the species (different colors). The error bars correspond to two standard-errors of the mean.

distributions, as observed in *E. coli*. Supplementary figure S8 shows the reverse cumulative distributions of 2-SNPs, 3-SNPs, 4-SNPs, and 12-SNPs across all 6 species together with power-law fits. Although the curves often deviate substantially from simple straight lines, they all exhibit long tails and range over several orders of magnitudes, i.e. up to 5 orders of magnitude for 2-SNPs in *S. enterica*. Note that, since in *M. tuberculosis* the total number of different n-SNP types is small, only the 2-SNP and 3-SNP distributions can be reasonably defined. Figure 9 (left panel) shows the fitted exponents of the power-law distributions of n-SNPs as a function of n for all species. With the exception of *M. tuberculosis*, for which the exponents are small for all n, we see that the exponents generally increase with n indicating that the phylogenetic diversity generally increases as one moves further back in time, i.e. to larger n. Consistent with other observations, *H. pylori* shows the highest exponents, i.e. the highest diversity, and *M. tuberculosis* the lowest. While the exponents become roughly constant for $n > 20$ for *E. coli*, *H. pylori* and *S. aureus*, *B. subtilis* and *S. enterica*, exhibit more complex patterns with sudden drops in exponent at particular values of n, suggesting more complex population structures for these species.

As an aside, we decided to investigate what the distribution of n-SNP frequencies looks like for a sexually reproducing organism with complex population structure such as human. We extracted SNP data for chromosome 21 for 2504 humans from the 1000 Genome project [25] and calculated the frequency distributions of n-SNP types. Supplementary Fig. S9 shows examples of the n-SNP distributions for human together with the fitted exponents for n ranging from 1 to 30. Interestingly, the n-SNP distributions in human are all well fit by power-law distributions but instead of exponents that systematically increase with n, as we observed for the bacteria, for human the exponent is slightly larger than 3 and independent of n.

Returning to the bacterial species, supplementary Figure S10 shows the entropy profiles $H_n(s)$ for all strains s in each of the species. As we observed in *E. coli*, essentially every strain s exhibits a unique entropy profile $H_n(s)$, showing that also in these other species each strain has a unique 'fingerprint' of frequencies with which its lineage shares ancestors with those of other lineages. Although the entropy rises quickly to values in the range $4-8$ for most strains, we also see strains for which the entropy only rises after n exceeds some fairly large value of n, e.g. at $n = 10$ for some strains in *H. pylori*, and at $n = 24$ and $n = 62$ for some *S. enterica* strains, suggesting that these strains are part of groups of closely-related strains. Note also that these events appear to correspond to the sudden drops in the exponents of the n-SNP distributions of those strains (Fig. 9, left panel), reiterating that these n-SNP statistics encode extensive information about the population structure of each species. To summarize the entropy profiles of each species, the right panel of Fig. 9 shows the mean and standard-error of the entropy profiles, averaged over all strains, as a function of n. As for most other statistics, *M. tuberculosis* is an outlier whose strains generally only show low phylogenetic entropy. For all other species, the average entropy clearly increases as n increase, indicating again that the phylogenetic diversity increases further back in the past. For 4 of the 6 species, the mean entropy at large n falls in a narrow range between 5 and 6, suggesting that the effective number of ancestries far back in the past is relatively similar for these species.

Discussion

In this work we have introduced new methods to analyze prokaryotic genome evolution from multiple alignments of the core genomes of strains from a species. In particular, showing that almost all bi-allelic SNPs in the core genome alignment correspond to single mutations in the history of that position in the alignment, we showed several new ways in which these SNPs can be used to quantify phylogenetic structures and the role of recombination in genome evolution within prokaryotic species.

Our analysis shows that, for the species studied here, evolution of the core genome is almost entirely driven by recombination. That is, even for very closely related pairs of strains, the large majority of mutations that separate them derive from recombination events. Moreover, for the large majority of pairs of strains, none of the DNA in their pairwise alignment derives from their common ancestor, and each position in the core alignment has been overwritten many times by recombination. Given this, it seems highly unlikely that the ancestral phylogeny of the strains can be reconstructed from the core genome alignment.

Although we cannot completely exclude that sufficient information about the ancestral phylogeny is still encoded in some way into the core alignment, it is clear that currently no method exists that is capable of extracting this information, and we suspect that it is in fact impossible, i.e. that recombination has destroyed the necessary information. However, even if it were possible to reconstruct the ancestral phylogeny, it is not clear how useful this clonal phylogeny would be for understanding core genome evolution. Our analysis of SNP compatibility along the core alignment shows that the phylogeny changes every few dozen basepairs (and every handful of SNPs), so that the core alignment fragments into many thousands of short segments with different phylogenies. Thus, modeling sequence evolution in the core genome as occurring along the branches of a fixed phylogenetic tree is clearly inappropriate.

One might infer from these statistics that bacterial species are quasi-sexual and recombining freely, but this is inconsistent with the observation that strains do not appear roughly equidistant and that phylogenies build from large numbers of genomic loci clearly converge to a well-defined average phylogeny. To understand how this phylogenetic structure emerges in the face of rampant recombination we developed several methods for using bi-allelic SNPs for quantifying population structure from the core genome alignment. In particular, although recombination is evident across the ancestral lineages of almost all strains, we find that some lineages recombine much more frequently than others, and that the relative rates with which different groups of strains share a recent common ancestor vary over $3-5$ orders of magnitude and follows roughly power-law distributions. Thus, the phylogeny build from the core genome alignment does not reflect the clonal history of the strains, but rather reflects the rates with which different lineages have recombined in the past. Notably, since the n-SNP distributions follow smooth long-tailed distributions that do not appear to have a characteristic scale, it is not possible to naturally subdivide a species into subspecies of freely recombining groups of strains. Rather, there is a large continuum of relative rates. As an aside, given that recombination rates vary over orders of magnitude across different lineages, the idea of an effective recombination rate for a species seems inherently misleading, and models that fit the data to a model that assumes a constant rate of recombination within a species, e.g. [26], seem inappropriate.

Essentially all population genetics and coalescent models start from assuming one or more populations of individuals that, for the purpose of the model, are exchangeable. However, using the entropy profiles of the n-SNP distributions of each strain, we observed that every strain has unique relative rates with which its lineage shares common ancestors with the lineages of other strains. That is, each strain has unique recombination statistics. These observations thus suggest that models that assume individuals are exchangeable are inappropriate by definition.

Given that models that assume either a single consensus tree, a fixed rate of recombination across strains, or even just exchangeable individuals, are all clearly at odds with the data on prokaryotic genome evolution, this raises the question of what would be an appropriate mathematical 'null model' that can capture the statistics that we observed here. In such a model, each lineage must have different rates of recombination with all other lineages, these rates must vary over multiple orders of magnitude, and the model should reproduce the roughly power-law distributions of n-SNP frequencies, ideally with exponents that can be tuned by parameters in the model. It is currently unclear how to construct such a model.

Given that recombination rates between different lineages appear to vary over several orders of magnitude, it also raises the question as to what sets these relative recombination rates. For example, it is not even clear whether these rates are shaped by natural selection, e.g. that due to epistatic interactions only recombinant segments from other strains with similar 'ecotypes' are not removed by purifying selection, or that recombination rates may rather be set by parameters such as the frequency with which lineages co-occur at the same geographical location. It is also conceivable that phages are a major source of transfer of DNA between strains, so that recombination rates may reflect the rates at which different lineages are infected by the same types of phages. It is also noteworthy that homologous recombination requires sufficient homology between the endpoints of the DNA fragment and the homologous segment in the host genome. Thus, recombination rates will intrinsically decrease with the nucleotide divergence between strains and previous studies have estimated that the rate of successful recombination decreases exponentially with nucleotide divergence [27,28]. In this regard it is interesting that the critical divergence at which half of the genome is recombined varies over a relatively small range, i.e. from $0.003 - 0.01$ (Fig. 8A). It is thus conceivable that a species is essentially defined by the collection of strains that are sufficiently close to allow efficient recombination [29]. However, the statistics reported here seem to suggest a much larger range of recombination rates than such a simple DNA-homology based model would predict.

While we here studied the frequency distribution of n-SNP types as well as the entropies $H_n(s)$ of the n-SNP distributions for each strain, it appears to us that this is just the tip of the iceberg of possible ways in which n-SNPs can be used to study the evolution of a set of strains from their core genome alignment. Our analyses indicate that prokaryotic genome evolution is driven by recombination that occurs at a very wide distribution of different rates between different lineages, and there is now a strong need for the development of new mathematical tools and models that can accurately describe this kind of genome evolution.

Methods

Data

The *E. coli* sequences analysed here can be accessed on NCBI Bioproject via the accession number PRJNA432505 [7,13]. In Table S2 strains names and details for the reference strains used for Figure S1 can be found.

Genome sequences for all other species were downloaded from `ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/`. All strain names and download dates are listed in Table S3.

Core genome alignment and core tree

To build a core genome alignment for the SC1 strains we used the Realphy tool [14] with default parameters and Bowtie 2 [30] for the alignments. Realphy used PhyML [15] with parameters -m GTR -b 0 to infer trees from the whole and parts of the core alignment. The tree visualizations were made using the Figtree software [31].

Analysis of core alignment blocks

For each 3 Kb block of the core alignment we used PhyML using the option -c 1 to infer a phylogeny while restricting the number of relative substitution rate categories to one. Furthermore, to calculate the log-likelihood of a given 3Kb block under the tree topologies of other blocks, we reran PhyML using the -o 'lr' option, which only optimizes the branch lengths as well as the substitution rate parameters but doesn't alter the topology of the phylogeny.

Pairwise analysis and mixture modeling

For each pair of strains we slide a 1Kb window over the core genome alignment of the pair, shifting by 100 bp at a time, and build a histogram of the number of SNPs per kilobase by counting the number of SNPs in each window. That is, we obtain the distribution P_n of the fraction of 1 Kb windows that have n SNPs. We then assumed that the one kilobase blocks can be separated into a fraction f_a of 'ancestral blocks', i.e. regions that were inherited from the clonal ancestor of the pair, and a fraction $(1 - f_a)$ that have been recombined since the pair diverged from a common ancestor. Although in previous work a simple *ad hoc* scheme was used in which it was assumed that blocks with less than a particular number of SNPs are ancestral and blocks with more SNPs are recombined [11], we found that this approach is not satisfactory and results significantly depend on the cut-off chosen.

We thus decided to employ a more principled mixture model approach. For the ancestral regions, the number of SNPs per kilobase should follow a simple Poisson distribution $P_n = \mu^n e^{-\mu}/n!$, with μ the expected number of mutations per block. For the recombined regions, we note that these regions themselves will consist of mosaics of subregions that have been recombined. Consequently, the recombined regions will consist of a mixture of Poisson distributions with different rates. It is well-known that mixtures of Poisson distributions with rates that are (close to) Gamma-distributed follow a negative binomial distribution and we found empirically that negative binomial distributions give excellent fits to the observed SNP distributions in our data. For the recombined regions we thus assume a negative binomial distribution of the form

$$P_n = \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)n!}\lambda^n(1-\lambda)^\alpha, \tag{2}$$

where $0 \leq \lambda \leq 1$ and $\alpha \geq 1$ are parameters of the distribution. We thus fit the observed distribution of SNPs per block P_n using the following mixture:

$$P_n = f_a \frac{\mu^n}{n!}e^{-\mu} + (1-f_a)\frac{\Gamma(n+\alpha)}{\Gamma(\alpha)n!}\lambda^n(1-\lambda)^\alpha, \tag{3}$$

where f_a is the fraction of the genome that is ancestral. Fits were obtained using maximum likelihood. While expectation maximization was used to fit the parameters f_a, μ, and λ, a grid search was employed to find the optimal dispersion parameter α.

Note that, in terms of the fitted parameters, the total number of mutations in ancestral blocks is μf_a, and the number of mutations in recombined blocks is $(1 - f_a)\alpha\lambda/(1 - \lambda)$.

To estimate the lengths of recombination events, we first extracted pairs that are sufficiently close (divergence less than 0.002) such that multiple overlapping recombination events are unlikely. We then used a two-state HMM with the same two components, i.e. a Poisson and a negative binomial component corresponding to ancenstral and recombined segments, and having fixed rates of transitioning from ancestral to recombined and vice versa, to parse the pairwise alignment into ancestral and recombined segments. We took the distribution of recombined segments in these alignments as the distribution of recombination events.

We define mostly clonal pairs as pairs with more than 90% of the alignment classified as ancestral, fully recombined pairs as pairs with less than 10% of the alignment classified as ancestral, and all other pairs as transition pairs. In order to estimate the critical divergence at which half of the genome is recombined we fit a linear model to the observed relationship between divergence and clonal fraction in all transition pairs, and define the critical divergence as the divergence at which the linear fit has a clonal fraction of 50%. To calculate the fraction of mutations that derive from recombined segments at the critical divergence we compute the fraction of mutations in recombined segments for all transition pairs (using the results from the mixture model) and fit a linear model to the observed dependence between the ancestral fraction an the fraction of mutations in recombined segments. We then define the fraction of mutations in recombined regions at the critical divergence as the fraction of mutations in the linear fit when the ancestral fraction is 50%.

Estimating the fraction of SNPs that correspond to single mutational events

The relatively low frequency of SNPs and the fact that almost all SNPs are bi-allelic strongly suggests that almost all bi-allelic SNPs correspond to single mutational events. Here we use a simple model to estimate the fraction of bi-allelic SNPs that correspond to single mutational events. To do this we will analyze the observed frequencies of columns with 1, 2, 3, and 4 different nucleotides under a simple model. To assess the effects of selection, we will consider these frequencies both for the subset of positions that should be under relatively little selection, i.e. third positions in fourfold degenerate codons, and positions that should be under relatively strong selection, i.e. second positions in codons. We will also do this separately for all strains, and all strains minus the 9 strains of the outgroup.

For a given position in the alignment, let μ denote the product of the mutation rate times the total length of the branches in the phylogeny at that position. The variable μ thus corresponds to the expected number of mutations at this position. The probability that n mutations took place at this position is given by a Poisson distribution:

$$P_n = \frac{\mu^n}{n!} e^{-\mu}. \tag{4}$$

We will assume that, every time a mutation occurs, each of the 3 possible target nucleotides is equally likely. Let d denote the number of different nucleotides in the column and let $T(d'|d)$ be the matrix of probabilities, that under a single mutation, the number of different nucleotides transitions from d to d'. We have $T(2|1) = 1$, $T(2|2) = 1/3$, $T(3|2) = 2/3$, $T(3|3) = 2/3$, $T(4|3) = 1/3$, $T(4|4) = 1$, and all other transition probabilities are zero. Starting from a single nucleotide in the column, the probability $P(d|n)$ to end up with d different nucleotides after n mutations is given by the n-th power of the transition matrix T, i.e. $P(d|n) = T^n(d|1)$. From this we can work out the probability $P(d|\mu)$ to end up with d different nucleotides as a function of the expected number of mutations μ as

$$P(d|\mu) = \sum_{n=0}^{\infty} T^n(d|1) \frac{\mu^n}{n!} e^{-\mu}. \tag{5}$$

The infinite sums can all be evaluated analytically and we find

$$P(1|\mu) = e^{-\mu}, \tag{6}$$

$$P(2|\mu) = 3e^{-\mu}\left(e^{\mu/3} - 1\right), \tag{7}$$

$$P(3|\mu) = 3e^{-\mu}\left(e^{\mu/3}-1\right)^2, \tag{8}$$

and

$$P(4|\mu) = e^{-\mu}\left(e^{\mu/3}-1\right)^3. \tag{9}$$

Assume we observe c_d columns with d different nucleotides, with d running from 1 to 4. The log-likelihood of this count data given μ is

$$L(\mu) = \sum_{d=1}^{4} c_d \log\left[P(d|\mu)\right]. \tag{10}$$

Maximizing the log-likelihood with respect to μ we find that the optimal value of μ given these counts as

$$\mu_* = 3\log\left[\frac{3(c_1+c_2+c_3+c_4)}{3c_1+2c_2+c_3}\right]. \tag{11}$$

Finally, given μ_*, the fraction f_{sm} of bi-allelic SNPs that correspond to single mutations is given by

$$f_{sm} = \frac{\mu_*}{3\left(e^{\mu_*/3}-1\right)}. \tag{12}$$

Table 1 shows the estimated expected number of mutations per column μ_* and the estimated fraction of bi-allelic SNPs that correspond to single mutations f_{sm} for the 5 different subsets of columns. We see that, for all 5 subsets, the fraction f_{sm} is over 95% and close to 100% for the second positions in codons.

Column set	μ_*	f_{sm}
All columns	0.118	0.9804
Synonymous positions	0.287	0.953
Second positions in codons	0.0258	0.9957
Synom. pos. without outgroup	0.149	0.975
Sec. pos. without outgroup	0.0172	0.9971

Table 1. Estimated expected number of mutations per position μ_* and estimated fraction of bi-allelic SNPs that correspond to single mutation events f_{sm} for 5 different subsets of core alignment columns: all columns, all synonymous positions (third positions in fourfold degenerate codons), second positions in codons, synonymous positions excluding the outgroup, second positions in codons excluding the outgroup.

In addition, Supplementary Fig. S11 shows a comparison of the observed and predicted frequencies of columns with 1, 2, 3, and 4 letters. Since effects of selection are likely least for the synonymous positions, we expect the simple model to fit the data best and we indeed observe that, for the synonymous positions, the simple model can reasonably accurately fit the observed frequencies, and even for the set of all alignment columns the fits are quite accurate (Suppl. Fig. S11). In contrast, for the second positions in codons, we can see the effects of selection in that, from the larger fractions of columns without SNPs, the model infers a lower μ_*, and this leads to an underestimation of columns with 4 nucleotides. Thus, the true fraction f_{sm} is more likely close to the values inferred from the synonymous positions. Note that $f_{sm} = 0.953$ when including the outgroup and $f_{sm} = 0.975$ when the outgroup is excluded. The difference between these two estimates derives from the very high fraction of SNP columns in which the 9 strains of the outgroup have another nucleotide than all other strains. For this subset of SNPs the fraction of columns that have more than one mutation is much higher than for any other SNP column. Thus, for all other SNP columns, the estimate that 97.5% correspond to single mutations is likely the most accurate.

Constructing a tree that maximizes the number of compatible SNPs

We classify all SNPs in the core genome alignment into *SNP types* as follows. For each bi-allelic SNP, we map all letters with the majority nucleotide to a 0 and the minority nucleotide to a 1 and sort the bits according to the alphabetic order of the strain names. In this way, each SNP is mapped to a binary sequence of length 92. This binary sequence defines the SNP type. Note that a SNP type corresponds to a particular bi-partitition of the strains.

We next counted the number of occurrences n_t of each SNP type t and sorted the SNP types from most to least common. We then used the following greedy algorithm to a collect a subset T of mutually compatible SNP types that accounts for as many SNPs as possible. We seed T with the most common SNP type, i.e. the SNP type occurring at the top of the list. We then go down the list of SNP types, iteratively adding SNP types t to the set T that are compatible with all previous types in the set T.

Bottom up tree building

In this procedure we build phylogenies of subclades in a bottom-up manner, starting from the full set of 92 strains and iteratively fusing pairs, minimizing the number of incompatible SNPs at each step.

For any subset of strains S, we define the number of supporting SNPs n_S as the number of SNPs that fall on the branch between the subset S and the other strains, i.e. the number of SNPs in which all strains in S have one letter, and all other strains another letter. Similarly, we define the number of clashing SNPs c_S as the number of SNPs that are incompatible with the strains in S forming a subclade in the tree.

The iterative merging procedure is initiated with each of the 92 strains forming a subclade S. At each step of the iteration we calculate, for each pair of existing subclades S_1, S_2, the number of clashing SNPs c_S and supporting SNPs n_S for the set of strains $S = S_1 \cup S_2$ consisting of the union of the strains in S_1 and S_2. We then merge the pair (S_1, S_2) that minimizes the clashes c_S and, when their are ties, maximizes the number of supporting SNPs n_S. At each step of the calculation we keep track of the total number of SNPs on the branches of the subtrees build so far, as well as the total number of SNPs that are inconsistent with the subtrees build so far. In addition, we calculate the average pairwise divergences of the strains within the subclades. Supplementary Fig. S3 shows the ratio of clashing to supporting SNPs as a function of the divergence within the subclades.

Quartet analysis

Quartets were assembled in the following way. We construct a grid of target distances d starting at 0.00001 and having 50 points with 0.0005 sized distance. For every target distance d we scan the alignment for four strains which have all pairwise distances within 1.25 fold of distance d. Every target distance d for which no quartet can be found fulfilling these criteria is ignored.

For each quartet we extract all SNP columns where two strains have a specific nucleotide and the other two strains have another nucleotide. Every such SNP column unambiguously supports one out of three possible tree topologies for this quartet. For each quartet we determine which topology has the largest number of supporting SNPs, and what the fraction of SNPs is that support this topology.

Linkage Disequilibrium measure

A standard measure of linkage disequilibrium of SNPs at a given distance is given by the average squared-correlation of the genotypes at these positions [32]. For a pair of loci with bi-allelic SNPs there are 4 possible genotypes which we indicate as binary patterns 00, 01, 10, and 11. If the frequencies of these genotypes are f_{00}, f_{01}, f_{10}, and f_{11}, then the squared correlation is calculated as

$$r^2 = \frac{(f_{00}f_{11} - f_{01}f_{10})^2}{f_{1.}f_{0.}f_{.0}f_{.1}}, \tag{13}$$

where the variables with dots correspond to marginal probabilities, e.g. $f_{1.} = f_{10} + f_{11}$, $f_{.1} = f_{01} + f_{11}$, and so on.

Minimum number of phylogeny switches

We iterate over all SNP columns in order of the core genome and add the current SNP to a list if it is pairwise compatible with all SNPs currently in the list. If it is incompatible with at least one SNP in this list we empty the list, re-initialize the list with the current SNP, and increase the phylogeny counter by one.

Power-law fits of n-SNP distributions

We extract each n-SNP from the core genome alignment and count the frequency, i.e. the number of occurrences, f_t of each n-SNP type t as well as the total number T of n-SNP types that occur at least once. We assume the n-SNP type occurrences are drawn from a power-law of the form

$$P(f) = \frac{1}{\zeta(\alpha)} f^{-\alpha}, \tag{14}$$

where $\zeta(\alpha)$ is the Riemann zeta function defined by

$$\sum_{f=1}^{\infty} f^{-\alpha} = \zeta(\alpha). \tag{15}$$

The log-likelihood of the frequencies f_t as a function of α is given by

$$L(\alpha) = -T \log[\zeta(\alpha)] - \sum_t \alpha \log[f_t] = -T \left(\log[\zeta(\alpha)] + \alpha \langle \log[f] \rangle \right), \tag{16}$$

where $\langle \log[f] \rangle$ is the average of the logarithm of the SNP-type frequencies. Using a uniform prior on α, the posterior distribution of α is simply proportional to the likelihood function. The optimal exponent α_* is the solution of

$$\frac{\zeta'(\alpha_*)}{\zeta(\alpha_*)} = -\langle \log[f] \rangle. \tag{17}$$

To calculate error-bars on the fitted exponentials we approximate the posterior by a Gaussian by expanding the log-likelihood to second order around the optimal exponent α_*. We then find for the standard-devation of the posterior distribution:

$$\sigma(\alpha) = \frac{1}{\sqrt{T \left(\frac{\zeta''(\alpha_*)}{\zeta(\alpha_*)} - \frac{\zeta'(\alpha_*)^2}{\zeta(\alpha_*)^2} \right)}}. \tag{18}$$

Entropy profiles of n-SNP distributions

For a given strain X we first extract all SNP types t for which X is one of the strains that shares the minority nucleotide. We then further stratify these SNP types by the number of SNPs sharing the minority nucleotide. For each n we thus obtain a set $S(X, n)$ of n-SNPs in which strain X is one of the strains sharing the SNP. We denote the number of occurrences of a SNP of type t by f_t and the total number of n-SNPs within set $S(X, n)$ as $F(X, n)$, i.e.

$$F(X, n) = \sum_{t \in S(X,n)} f_t \tag{19}$$

The entropy $H(X,n)$ of the n-SNP distribution of strain X is then defined as

$$H(X,n) = -\sum_{t \in S(X,n)} \frac{f_t}{F(X,n)} \log_2 \left[\frac{f_t}{F(X,n)} \right]. \tag{20}$$

Acknowledgments

The authors thank Olin Silander and Frederic Bertels for useful discussions during the development of this work. This work was supported by the Swiss National Science Foundation grant No. 31003A_135397. In addition, this work was done in part while the authors were visiting the Simons Institute for the Theory of Computing and was thus supported in part by NSF Grant No. PHY17-48958, NIH Grant No. R25GM067110, and the Gordon and Betty Moore Foundation Grant No. 2919.01. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

Supplementary Figures

Species	Strains	Genome size	Core size	Informative SNPs
Escherichia coli	92	4929299	2756541 (56%)	247822
Bacillus subtilis	75	4155843	2341553 (56%)	182535
Helicobacter pylori	83	1655288	850827 (51%)	114993
Mycobacterium tuberculosis	40	4465985	4150139 (93%)	3502
Salmonella enterica	155	4810980	2846634 (59%)	192117
Staphylococcus aureus	95	2881899	2002833 (69%)	73756

Table S1. Summary statistics of the core genome alignments of the different bacterial species. For each species, the number of strains, the median genome size, the size of the core genome alignment, and the number of informative SNPs in the core alignment are listed.

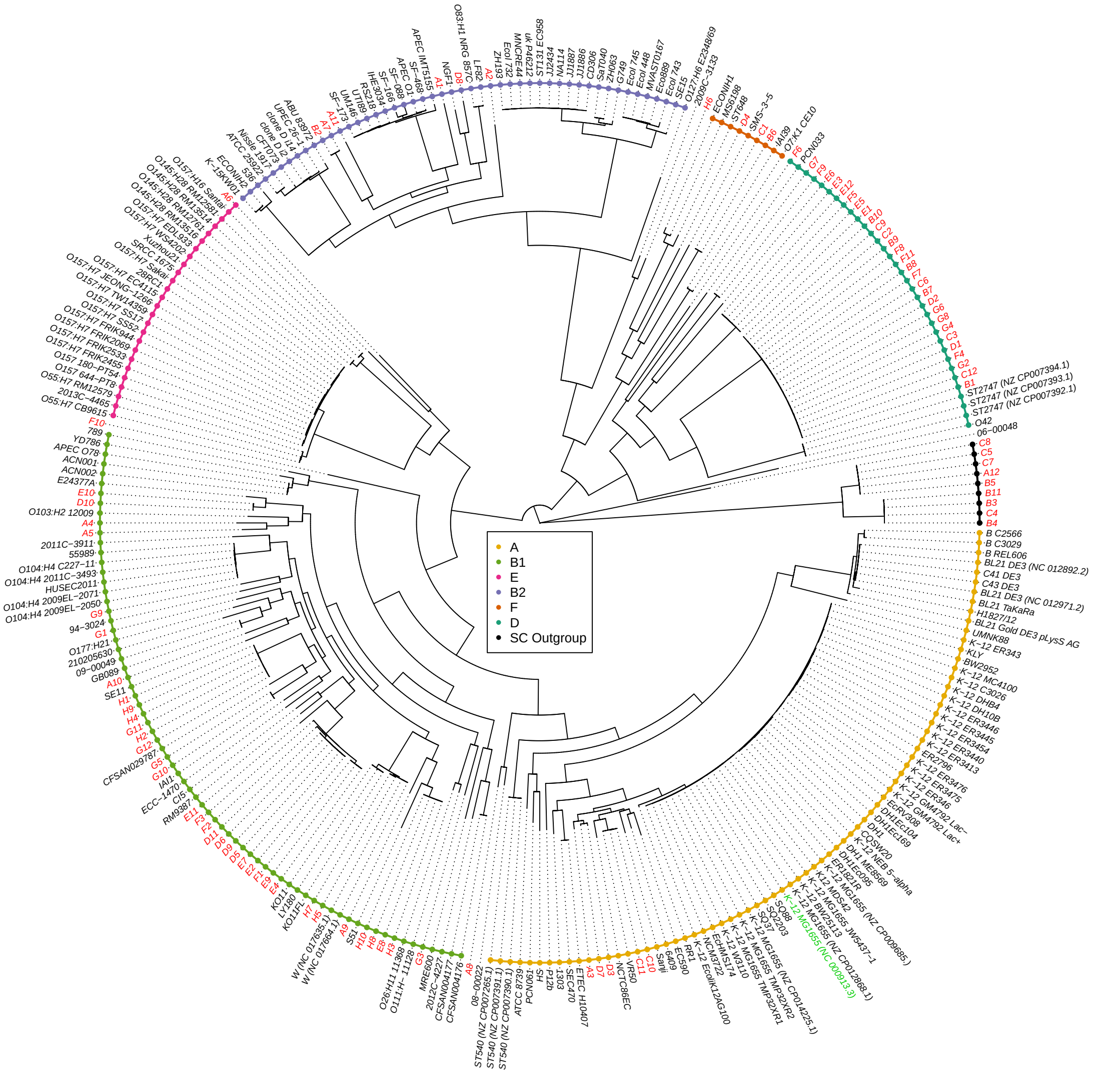

Figure S1. Maximum likelihood tree reconstructed from the core genome alignments of the SC1 strains (red font names), the K-12 lab strain (green font name), and 189 *E. coli* reference strains (black font names). The Known phylogroups are indicated as different colored leaf nodes. Note that the SC1 strains represent and are distributed across essentially all known phylogroups, and include some strains that cannot be easily assigned to a particular phylogroup. Also note the 'outgroup' of 9 SC1 strains shown as black leaf nodes, which have approximately 8% nucleotide divergence with the other strains (this branch is artificially shortened to fit into the figure), whereas all other strains are less than 3% diverged from each other.

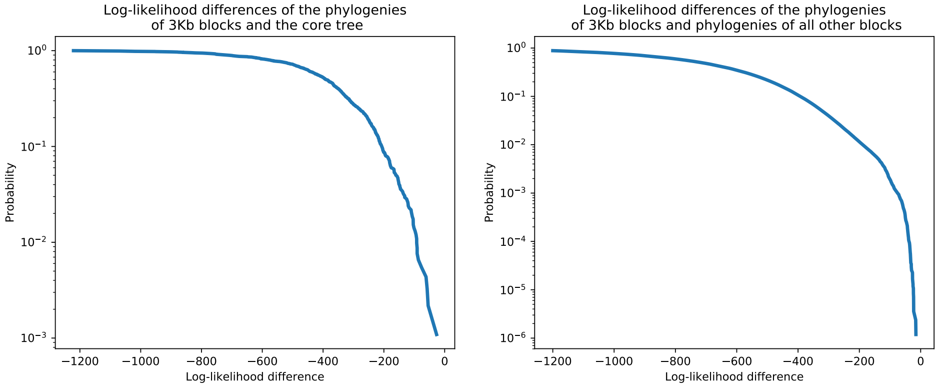

Figure S2. All 3kb alignment blocks reject the core tree topology as well as the topologies of the phylogenies reconstructed from all other blocks. **Left panel**: For each 3Kb block in the core alignment we used PhyML to reconstruct a phylogeny and then calculated the difference in the log-likelihood of the alignment block under the topology of the core tree and the log-likelihood of the reconstructed phylogeny. The figure shows the reverse cumulative distribution of these log-likelihood differences, with the vertical axis shown on a logarithmic scale. There are virtually no blocks for which the log-likelihood of the core tree topology is close to the log-likelihood under the block's own phylogeny. **Right panel**: As in the left panel, but now we calculated, for each 3kb block, the log-likelihood differences for the topologies of the phylogenies reconstructed from all other blocks. Each block attains a significantly higher log-likelihood when using its own topology than using the topology of any other alignment block.

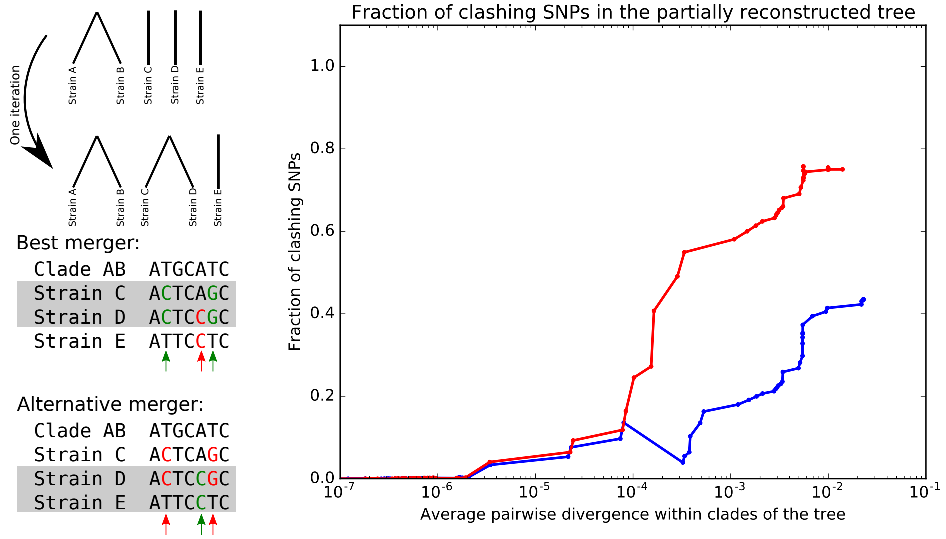

Figure S3. Bottom-up tree building, minimizing SNP clashes. **Left panel**: Illustration of the iterative bottom-up tree reconstruction. At each step the pair of clades is fused that minimizes the number of SNPs that clash with the fusion (red arrows). In case of multiple pairs that have the same number of clashing SNPs, the pair with the largest number of supporting SNPs (green arrows) is chosen. **Right panel**: Fraction of SNPs that support vs. clash with the partially reconstructed tree as a function of the average pairwise divergence of strains that occur within the same clade of the partially reconstructed tree. The blue curve corresponds to the full set of SC strains and the red curve to all SC strains except for the outgroup. The horizontal axis is shown on a logarithmic scale.

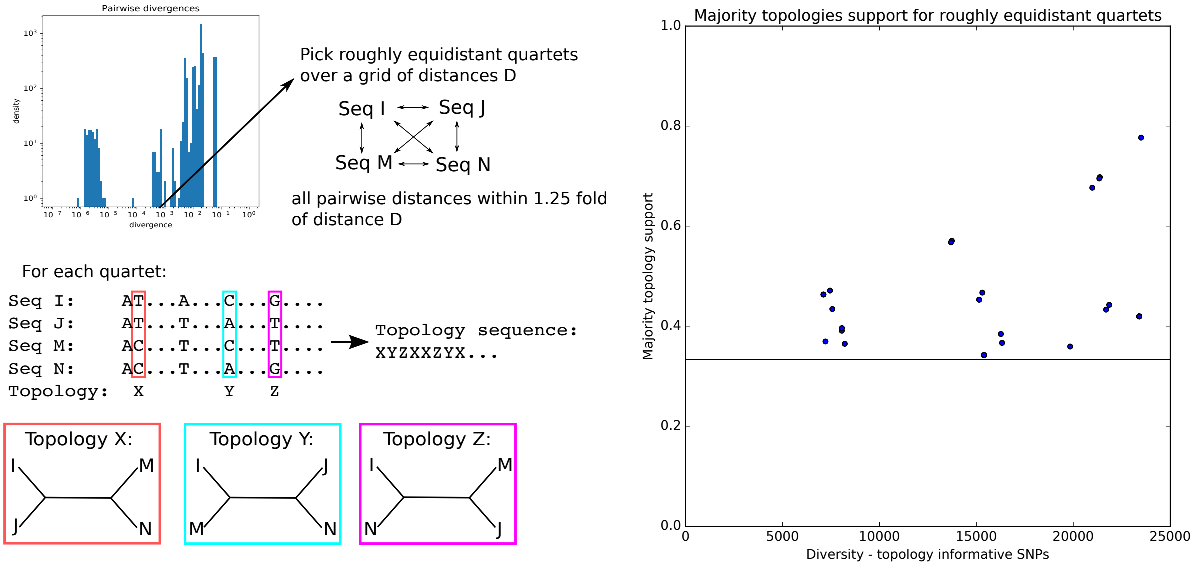

Figure S4. Quartets of roughly equidistant strains have no consensus phylogeny. **Left**: Using the distribution of pairwise distances (top panel) we select, for each pairwise distance D, quartets of strains whose pairwise distances are all within a factor 1.25 of D. SNPs for which two strains have one letter and two strains another are informative for the topology and each support one of the three possible topologies. **Right**: For each quartet we determined the topology that is supported by most SNPs and then calculated the fraction of topology-informative SNPs that supported the most common topology. The plot shows the fraction of SNPs supporting the most common topology (vertical axis) as a function of the total number of informative SNPs (horizontal axis). The horizontal line marks the minimal possible fraction, which is attained when all 3 topologies are supported by 1/3 of the SNPs. Note that, for the majority of quartets, the most common topology is supported by less than half of the informative SNPs.

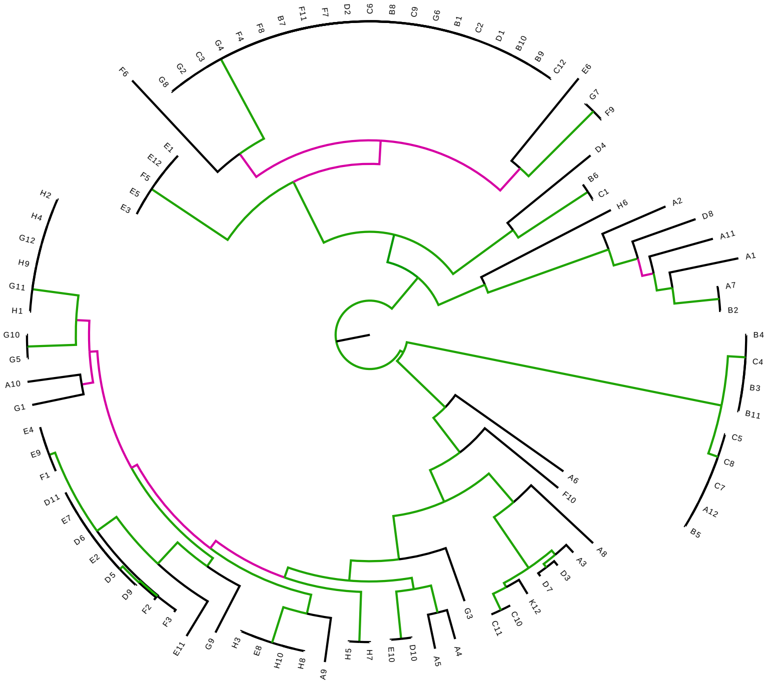

Figure S5. Differences between the core tree T and the tree T' reconstructed from the alignment from which all SNPs that fall on branches of the core tree have been removed. Each branch of the core tree T is colored green when the branch also occurs in T' and pink if it does not. The Robinson-Foulds distance between two trees is defined as the number of branches (i.e. bipartitions) that occur in only one of the two trees. For T and T' the Robinson-Foulds distance is 62 out of a maximal 178, i.e. a fraction 0.35 does not match. Note also that for tree T' only 3% of the SNPs fall on its branches.

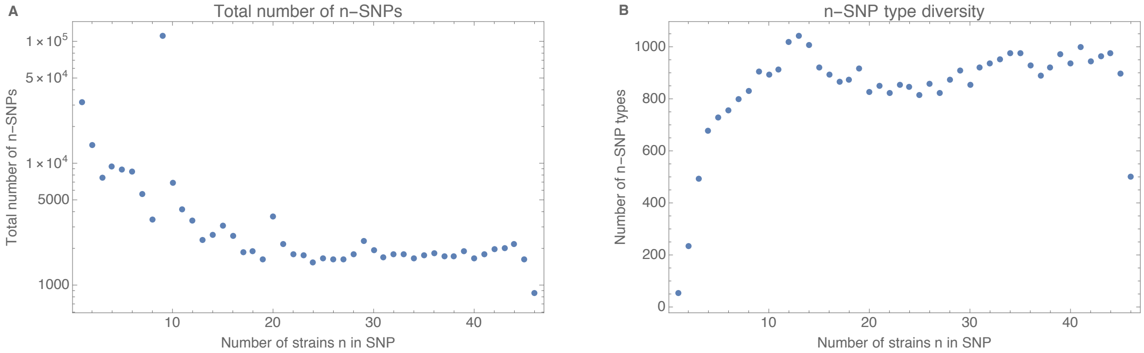

Figure S6. Overall frequency and diversity of n-SNPs. **A**: Total number of occurrences of n-SNPs, i.e. SNPs shared by n strains (vertical axis) as a function of n (horizontal axis). The vertical axis is shown on a logarithmic scale. Note the outlier at $n = 9$, which corresponds to the very large number of SNPs shared by the 9 strains of the outgroup. Note also that for $n \geq 10$ the number of n-SNPs is approximately constant around 2000. **B**: Number of unique n-SNP types as a function of n. The number of n-SNP types increases quickly with n and then saturates around a value of approximately $800 - 1000$ for $n \geq 10$.

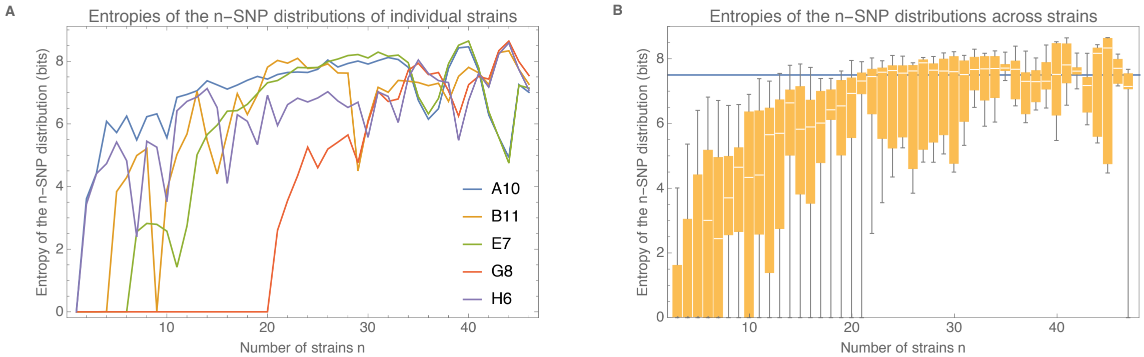

Figure S7. Entropy profiles of the n-SNP distributions across *E. coli* strains. **A**: Examples of the n-SNP entropy profiles for 5 different strains s (indicated in the legend). The entropy $H_s(n)$ of the distribution of n-SNPs in which a particular strain occurs is shown as a function of the number of strains n. **B**: Box-whisker plots showing the 5, 25, 50 (median), 75 and 95 percentiles of the distribution of the entropies of the n-SNP distributions as a function of the number of strains n. The blue line shows an entropy of 7.5 bits.

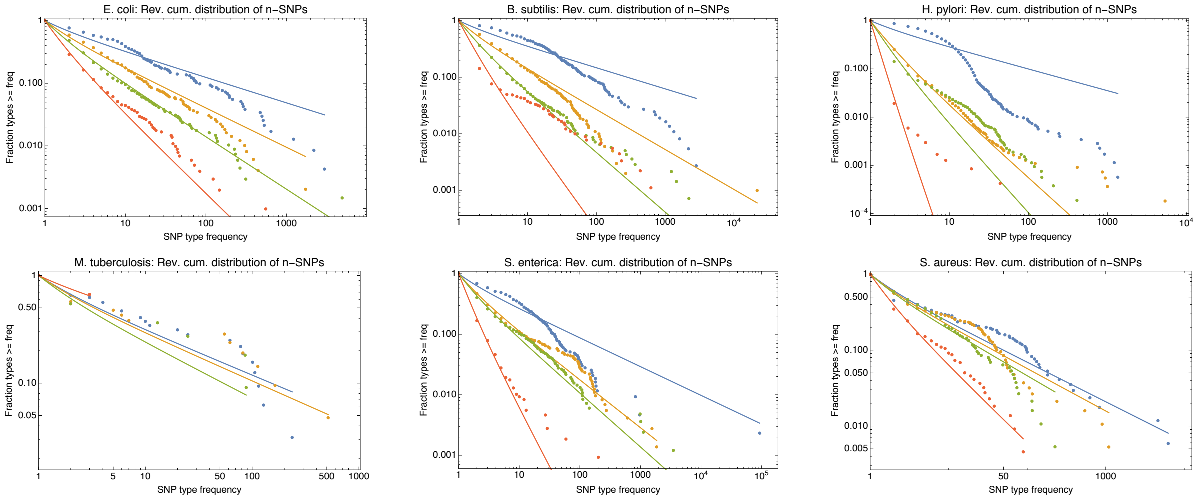

Figure S8. Power-law fits of the n-SNP distributions for all 6 species. Each panel shows the reverse cumulative distributions of the frequencies of all observed 2-SNPs (blue dots), 3-SNPs (orange dots), 4-SNPs (green dots), and 12-SNPs (red dots), with the solid lines in corresponding colors showing power-law fits. The species is indicated at the top of the panel.

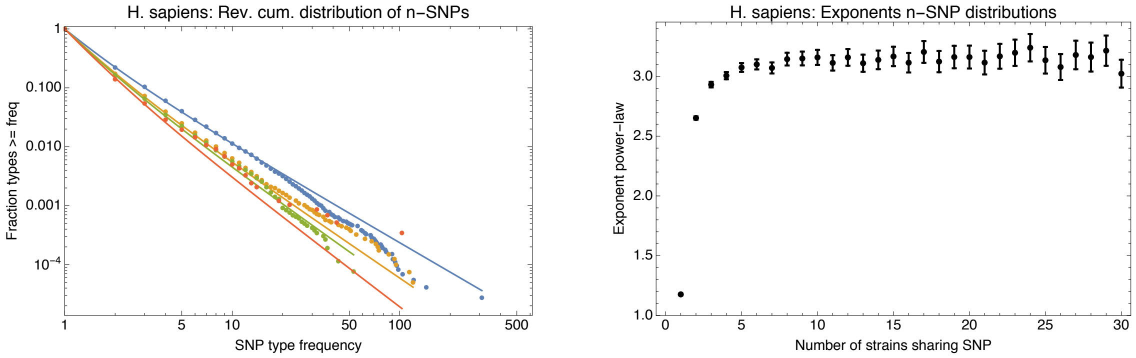

Figure S9. Human n-SNPs frequencies are also power-law distributed. **Left panel**: Reverse cumulative distributions of the frequencies of all observed 2-SNPs (blue dots), 3-SNPs (orange dots), 4-SNPs (green dots), and 12-SNPs (red dots), with the solid lines in corresponding colors showing power-law fits, for the human data. Both axes are shown on a logarithmic scale. **Right panel**: Fitted exponents for the power-law n-SNP distributions on the human data for n ranging from 1 to 30. The error bars correspond to 95% posterior probability intervals.

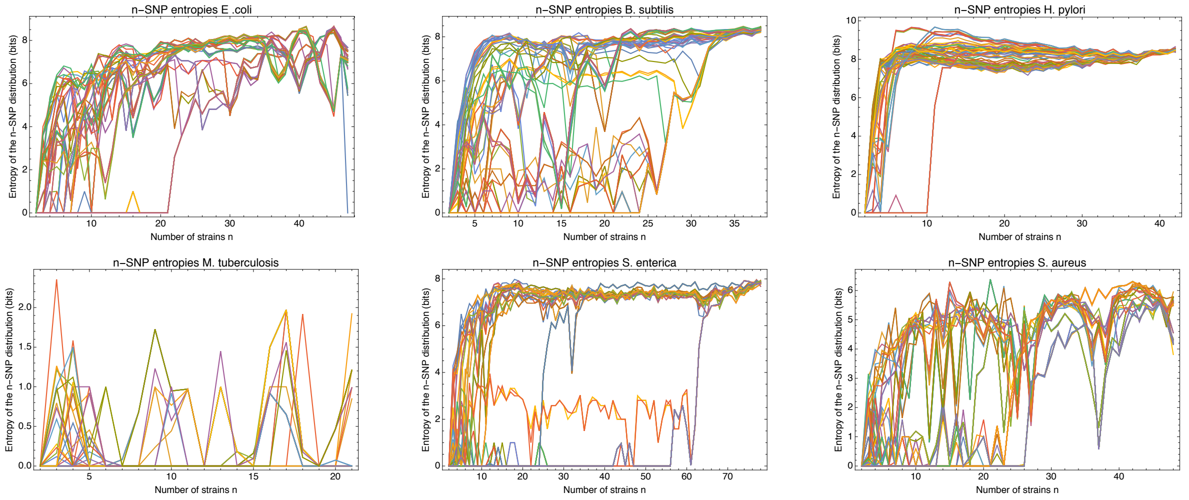

Figure S10. Entropy profiles of all the strains of each of the 6 species. Each panel corresponds to one species (indicated at the top) and shows the entropy profiles $H_s(n)$ of the distribution of n-SNPs in which a particular strain occurs is shown as a function of the number of strains n for each strain s (different colors).

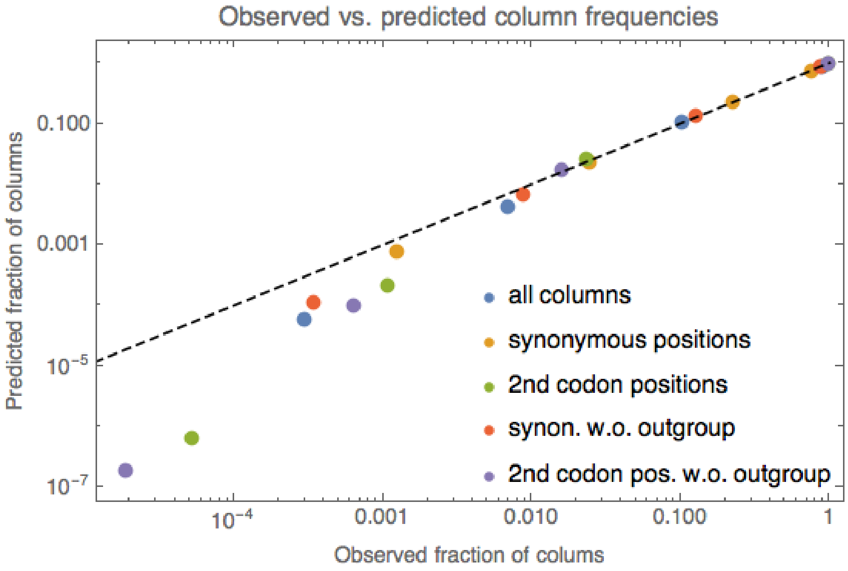

Figure S11. Comparison of the observed frequencies of columns with 1, 2, 3, and 4 different nucleotides under the simple model described in the methods. Different colored dots correspond to different subsets of columns, as indicated in the legend. For each color, 4 dots are shown corresponding to the observed frequencies of columns with 1, 2, 3, and 4 nucleotides (horizontal axis) and the predicted frequencies according to the simple model (vertical axis). The dashed line shows the identity $y = x$. Both axes are shown on logarithmic scales.

RefSeq ID	Short Name	Provenance
NC_000913.3	K-12 MG1655 (NC_000913.3)	Laboratory
NC_004431.1	CFT073	Natural
NC_002695.1	O157:H7 Sakai	Natural
NC_012892.2	BL21 DE3 (NC_012892.2)	Engineered
NC_007779.1	K-12 W3110	Laboratory
NC_011415.1	SE11	Natural
NC_013654.1	SE15	Natural
NC_013353.1	O103:H2 12009	Natural
NC_013364.1	O111:H- 11128	Natural
NC_007946.1	UTI89	Natural
NC_008253.1	536	Natural
NC_008563.1	APEC O1	Natural
NC_009801.1	E24377A	Natural
NC_009800.1	HS	Natural
NC_012967.1	B REL606	Engineered
NC_010468.1	ATCC 8739	Laboratory
NC_010473.1	K-12 DH10B	Engineered
NC_010498.1	SMS-3-5	Natural
NC_011353.1	O157:H7 EC4115	Natural
NC_013008.1	O157:H7 TW14359	Natural
NC_012759.1	BW2952	Engineered
NC_012971.2	BL21 DE3 (NC_012971.2)	Engineered
NC_017625.1	DH1	Engineered
NC_012947.1	BL21 Gold DE3 pLysS AG	Engineered
NC_013941.1	O55:H7 CB9615	Natural
NC_017628.1	IHE3034	Natural
NC_011748.1	55989	Natural
NC_011741.1	IAI1	Natural
NC_011750.1	IAI39	Natural
NC_011601.1	O127:H6 E2348/69	Natural
NC_017626.1	O42	Natural
NC_013361.1	O26:H11 11368	Natural
NC_016902.1	KO11	Engineered
NC_017631.1	ABU 83972	Natural
NC_017632.1	UM146	Natural
NC_017634.1	O83:H1 NRG 857C	Natural
NC_017635.1	W (NC_017635.1)	Laboratory
NC_017633.1	ETEC H10407	Natural
NC_017641.1	UMNK88	Natural
NC_017644.1	NA114	Natural
NZ_CP006632.1	PCN033	Natural
NC_017646.1	O7:K1 CE10	Natural
NC_017651.1	clone D i2	Natural
NC_017652.1	clone D i14	Natural
NC_017656.1	O55:H7 RM12579	Natural
NC_017663.1	P12b	Natural
NC_017660.1	KO11FL	Engineered
NC_017664.1	W (NC_017664.1)	Laboratory
NC_017906.1	Xuzhou21	Natural
NC_017638.1	DH1 ME8569	Engineered
NC_011993.1	LF82	Natural
NZ_HG941718.1	ST131 EC958	Natural
NC_018650.1	O104:H4 2009EL-2050	Natural
NC_018658.1	O104:H4 2011C-3493	Natural
NC_018661.1	O104:H4 2009EL-2071	Natural
NC_020163.1	APEC O78	Natural
NC_020518.1	K12 MDS42	Engineered
NC_022364.1	LY180	Engineered
NC_022648.1	JJ1886	Natural
NZ_HG738867.1	K-12 MC4100	Engineered
NZ_CP006027.1	O145:H28 RM13514	Natural
NZ_CP006262.1	O145:H28 RM13516	Natural
NZ_CP007265.1	ST540 (NZ_CP007265.1)	Natural
NZ_CP007390.1	ST540 (NZ_CP007390.1)	Natural
NZ_CP007391.1	ST540 (NZ_CP007391.1)	Natural
NZ_CP007392.1	ST2747 (NZ_CP007392.1)	Natural
NZ_CP007393.1	ST2747 (NZ_CP007393.1)	Natural
NZ_CP007394.1	ST2747 (NZ_CP007394.1)	Natural
NZ_CP007133.1	O145:H28 RM12761	Natural
NZ_CP007136.1	O145:H28 RM12581	Natural
NZ_CP007799.1	Nissle 1917	Natural
NZ_CP008801.1	KLY	Engineered

NZ_CP008805.1	O157:H7 SS17	Natural
NZ_CP008957.1	O157:H7 EDL933	Natural
NZ_CP009072.1	ATCC 25922	Natural
NZ_CP009273.1	K-12 BW25113	Laboratory
NZ_CP009859.1	ECONIH1	Natural
NZ_CP009644.1	ER2796	Laboratory
NZ_CP009789.1	K-12 ER3413	Engineered
NZ_CP007149.1	RS218	Natural
NZ_CP009104.1	RM9387	Natural
NZ_CP009106.2	94-3024	Natural
NZ_CP009685.1	K-12 MG1655 (NZ_CP009685.)	Laboratory
NZ_CP010304.1	O157:H7 SS52	Natural
NZ_CP005930.1	APEC IMT5155	Natural
NZ_CP010371.1	6409	Natural
NZ_CP010315.1	789	Natural
NZ_CP007592.1	O157:H16 Santai	Natural
NZ_CP009166.1	1303	Natural
NZ_CP010585.1	C41 DE3	Engineered
NZ_CP010344.1	ECC-1470	Natural
NZ_CP010816.1	BL21 TaKaRa	Engineered
NZ_CP010876.1	MNCRE44	Natural
NZ_LM995446.1	EcRV308	Engineered
NZ_LM993812.1	EcHMS174	Engineered
NZ_HF572917.1	HUSEC2011	Natural
NZ_CP011134.1	VR50	Natural
NZ_CP011018.1	CI5	Natural
NZ_CP010438.1	K-12 ER3454	Engineered
NZ_CP010439.1	K-12 ER3440	Engineered
NZ_CP010440.1	K-12 ER3476	Engineered
NZ_CP010441.1	K-12 ER3445	Engineered
NZ_CP010442.1	K-12 ER346	Engineered
NZ_CP010443.1	K-12 ER3446	Engineered
NZ_CP010444.1	K-12 ER3475	Engineered
NZ_CP010445.1	K-12 ER343	Engineered
NZ_LN832404.1	K-12 EcoliK12AG100	Engineered
NZ_CP011331.1	O104:H4 C227-11	Natural
NZ_CP007594.1	SEC470	Natural
NZ_CP011320.1	SQ37	Engineered
NZ_CP011321.1	SQ88	Engineered
NZ_CP011324.1	SQ2203	Engineered
NZ_CP011416.1	CFSAN029787	Natural
NZ_CP011342.2	K-12 GM4792 Lac+	Laboratory
NZ_CP011343.2	K-12 GM4792 Lac-	Laboratory
NZ_CP006636.1	PCN061	Natural
NZ_CP011938.1	C43 DE3	Engineered
NZ_CP011495.1	NCM3722	Engineered
NZ_CP007442.1	ACN001	Natural
NZ_CP012125.1	DH1Ec095	Engineered
NZ_CP012126.1	DH1Ec104	Engineered
NZ_CP012127.1	DH1Ec169	Engineered
NZ_CP011113.1	RR1	Engineered
NZ_CP012635.1	SF-088	Natural
NZ_CP012625.1	SF-468	Natural
NZ_CP012633.1	SF-166	Natural
NZ_CP012631.1	SF-173	Natural
NZ_CP012802.1	O157:H7 WS4202	Natural
NZ_CP012868.1	K-12 MG1655 (NZ_CP012868.1)	Laboratory
NZ_CP012869.1	K-12 MG1655 TMP32XR1	Laboratory
NZ_CP012870.1	K-12 MG1655 TMP32XR2	Laboratory
NZ_CP013029.1	2012C-4227	Natural
NZ_CP013025.1	2009C-3133	Natural
NZ_CP013112.1	YD786	Natural
NZ_CP013253.1	CQSW20	Natural
NZ_CP013658.1	uk_P46212	Natural
NZ_CP008697.1	ST648	Natural
NZ_CP013831.1	CD306	Natural
NZ_CP013835.1	JJ2434	Natural
NZ_CP007491.1	ACN002	Natural
NZ_CP014197.1	MRE600	Laboratory
NZ_CP014225.1	K-12 MG1655 (NZ_CP014225.1)	Laboratory
NZ_CP014314.1	O157:H7 JEONG-1266	Natural
NZ_CP014268.2	B C2566	Engineered
NZ_CP014269.1	B C3029	Engineered
NZ_CP014270.1	K-12 DHB4	Engineered

NZ_CP014272.1	K-12 C3026	Engineered
NZ_CP014348.1	K-12 MG1655 JW5437-1	Engineered
NZ_CP014495.1	SaT040	Natural
NZ_CP014488.1	G749	Natural
NZ_CP014492.1	MVAST0167	Natural
NZ_CP014497.1	ZH193	Natural
NZ_CP014522.1	ZH063	Natural
NZ_CP014316.1	JJ1887	Natural
NZ_CP011061.1	Sanji	Natural
NZ_CP015020.1	28RC1	Natural
NZ_CP015023.1	SRCC 1675	Natural
NZ_CP015138.1	Ecol_732	Natural
NZ_CP015069.1	Ecol_743	Natural
NZ_CP015074.2	Ecol_745	Natural
NZ_CP015076.1	Ecol_448	Natural
NZ_CP015240.1	2011C-3911	Natural
NZ_CP015241.1	2013C-4465	Natural
NZ_CP015832.1	O157 180-PT54	Natural
NZ_CP015831.1	O157 644-PT8	Natural
NZ_CP015846.1	O157:H7 FRIK2069	Natural
NZ_CP015842.1	O157:H7 FRIK2533	Natural
NZ_CP015843.1	O157:H7 FRIK2455	Natural
NZ_CP015995.1	S51	Natural
NZ_CP016007.1	NGF1	Engineered
NZ_CP016018.1	ER1821R	Engineered
NZ_CP015159.1	Eco889	Natural
NZ_CP014667.1	ECONIH2	Natural
NZ_CP015229.1	06-00048	Natural
NZ_CP015228.1	09-00049	Natural
NZ_CP013662.1	08-00022	Natural
NZ_CP013031.1	H1827/12	Natural
NZ_CP013663.1	GB089	Natural
NZ_CP015912.1	210205630	Natural
NZ_CP016182.1	EC590	Natural
NZ_CP016358.1	K-15KW01	Natural
NZ_CP016497.1	UPEC 26-1	Natural
NZ_CP016546.1	O177:H21	Natural
NZ_CP016625.1	O157:H7 FRIK944	Natural
NZ_CP014670.1	CFSAN004177	Natural
NZ_CP014583.1	CFSAN004176	Natural
NZ_CP015834.1	MS6198	Natural
NZ_CP017100.1	K-12 NEB 5-alpha	Engineered
NZ_LT601384.1	NCTC86EC	Laboratory

Table S2. The details of the reference strains used in this work. RefSeq IDs and short names are from the NCBI nucleotide database. Provenance refers to where the strain came from: a natural strain was sequenced directly after isolation, whereas a laboratory strain has passed many generations in artificial conditions, and engineered strains have had specific changes deliberately introduced to their genomes.

Strain name	Dowload date
Bacillus subtilis	Feb 16, 2016
GCF_001043765.1_ASM104376v1_genomic	
GCF_000931835.1_G4C10_genomic	
GCF_000699465.1_ASM69946v1_genomic	
GCF_000740475.1_ASM74047v1_genomic	
GCF_000155325.1_ASM15532v1_genomic	
GCF_000338735.1_ASM33873v1_genomic	
GCF_000293765.1_ASM29376v1_genomic	
GCF_000340295.1_ASM34029v1_genomic	
GCF_000782835.1_ASM78283v1_genomic	
GCF_000699525.1_ASM69952v1_genomic	
GCF_000724125.1_BSUBE1_genomic	
GCF_000830695.1_ASM83069v1_genomic	
GCF_000816805.1_ASM81680v1_genomic	
GCF_000696635.1_ASM69663v1_genomic	
GCF_000227485.1_ASM22748v1_genomic	
GCF_000735115.1_ASM73511v1_genomic	
GCF_001015095.1_ASM101509v1_genomic	
GCF_000245035.1_ASM24503v2_genomic	
GCF_000706705.1_ASM70670v1_genomic	
GCF_000321395.1_ASM32139v1_genomic	
GCF_000830595.1_ASM83059v1_genomic	
GCF_000878265.1_ASM87826v1_genomic	
GCF_000827065.1_ASM82706v1_genomic	
GCF_000155375.1_ASM15537v1_genomic	
GCF_000934165.1_ASM93416v1_genomic	
GCF_000497365.1_MP11_genomic	
GCF_000830715.1_ASM83071v1_genomic	
GCF_000582885.1_QH-1_V1.0_genomic	
GCF_000959025.1_ASM95902v1_genomic	
GCF_000183765.1_ASM18376v2_genomic	
GCF_000696615.1_ASM69661v1_genomic	
GCF_000523045.1_ASM52304v1_genomic	
GCF_001465815.1_ASM146581v1_genomic	
GCF_000789295.1_ASM78929v1_genomic	
GCF_000691185.1_ASM69118v1_genomic	
GCF_001541905.1_ASM154190v1_genomic	
GCF_000341775.1_ASM34177v1_genomic	
GCF_000146565.1_ASM14656v1_genomic	
GCF_000973605.1_ASM97360v1_genomic	
GCF_000230755.1_ASM23075v2_genomic	
GCF_000186085.1_ASM18608v1_genomic	
GCF_000186745.1_ASM18674v1_genomic	
GCF_000830735.1_ASM83073v1_genomic	
GCF_000009045.1_ASM904v1_genomic	
GCF_000177595.1_ASM17759v1_genomic	
GCF_000209795.2_ASM20979v2_genomic	
GCF_000332645.1_BSI1.0_genomic	
GCF_000931825.1_G1A4_genomic	
GCF_000953615.1_BS49Ch_genomic	
GCF_000385985.1_Bacillus_subtilis_PS216_genomic	
GCF_000349795.1_ASM34979v1_genomic	
GCF_000507005.1_PTS394_genomic	
GCF_001037985.1_ASM103798v1_genomic	
GCF_000828495.1_ASM82849v1_genomic	
GCF_000971925.1_ASM97192v1_genomic	
GCF_000743215.1_BST_genomic	
GCF_000227465.1_ASM22746v1_genomic	
GCF_000830635.1_ASM83063v1_genomic	
GCF_000931815.1_G1A3_genomic	
GCF_000245295.1_ASM24529v1_genomic	
GCF_000830645.1_ASM83064v1_genomic	
GCF_001187765.1_BSMS1577_genomic	
GCF_000497345.1_MP9_genomic	
GCF_000830605.1_ASM83060v1_genomic	
GCF_000830675.1_ASM83067v1_genomic	
GCF_000789275.1_ASM78927v1_genomic	
GCF_000344745.1_ASM34474v1_genomic	
GCF_000409585.1_Hal1_genomic	
GCF_000497485.1_ASM49748v1_genomic	
GCF_000747645.1_ASM74764v1_genomic	
GCF_000155355.1_ASM15535v1_genomic	

GCF_000832195.1_ASM83219v1_genomic	
GCF_000740485.1_ASM74048v1_genomic	
GCF_000931845.1_G5B15_genomic	
GCF_001534785.1_ASM153478v1_genomic	
Helicobacter pylori	Oct 5, 2016
GCF_000008785.1_ASM878v1_genomic	
GCF_000498335.1_ASM49833v1_genomic	
GCF_000185245.1_ASM18524v1_genomic	
GCF_000307815.1_ASM30781v1_genomic	
GCF_000685745.1_ASM68574v1_genomic	
GCF_000270065.1_ASM27006v1_genomic	
GCF_000600205.1_ASM60020v1_genomic	
GCF_000600125.1_ASM60012v1_genomic	
GCF_000213135.1_ASM21313v1_genomic	
GCF_000091345.1_ASM9134v1_genomic	
GCF_000600185.1_ASM60018v1_genomic	
GCF_001653415.1_ASM165341v1_genomic	
GCF_000192315.1_ASM19231v1_genomic	
GCF_000600045.1_ASM60004v1_genomic	
GCF_000392455.3_ASM39245v3_genomic	
GCF_000307835.1_ASM30783v1_genomic	
GCF_000270025.1_ASM27002v1_genomic	
GCF_000826985.1_ASM82698v1_genomic	
GCF_000685665.1_ASM68566v1_genomic	
GCF_001433515.1_ASM143351v1_genomic	
GCF_000600145.1_ASM60014v1_genomic	
GCF_000315955.1_ASM31595v1_genomic	
GCF_000392515.3_ASM39251v3_genomic	
GCF_001653455.1_ASM165345v1_genomic	
GCF_000148915.1_ASM14891v1_genomic	
GCF_000590775.1_ASM59077v1_genomic	
GCF_000439295.2_ASM43929v2_genomic	
GCF_000277405.1_ASM27740v1_genomic	
GCF_000178935.2_ASM17893v2_genomic	
GCF_000392475.3_ASM39247v3_genomic	
GCF_000093185.1_ASM9318v1_genomic	
GCF_000270005.1_ASM27000v1_genomic	
GCF_000185205.1_ASM18520v1_genomic	
GCF_001653435.1_ASM165343v1_genomic	
GCF_000192335.1_ASM19233v1_genomic	
GCF_000828955.1_ASM82895v1_genomic	
GCF_001653395.1_ASM165339v1_genomic	
GCF_000259235.1_ASM25923v1_genomic	
GCF_001549875.1_ASM154987v1_genomic	
GCF_000020245.1_ASM2024v1_genomic	
GCF_000021165.1_ASM2116v1_genomic	
GCF_000008525.1_ASM852v1_genomic	
GCF_000600085.1_ASM60008v1_genomic	
GCF_000829135.1_ASM82913v1_genomic	
GCF_000600225.1_ASM60022v1_genomic	
GCF_000685625.1_ASM68562v1_genomic	
GCF_000011725.1_ASM1172v1_genomic	
GCF_000307795.1_ASM30779v1_genomic	
GCF_001653375.1_ASM165337v1_genomic	
GCF_000013245.1_ASM1324v1_genomic	
GCF_000317875.1_ASM31787v1_genomic	
GCF_001549715.1_ASM154971v1_genomic	
GCF_000023805.1_ASM2380v1_genomic	
GCF_000255955.1_ASM25595v1_genomic	
GCF_000224535.1_ASM22453v1_genomic	
GCF_000392535.3_ASM39253v3_genomic	
GCF_000348885.1_ASM34888v1_genomic	
GCF_000148895.1_ASM14889v1_genomic	
GCF_000348865.1_ASM34886v1_genomic	
GCF_000185185.1_ASM18518v1_genomic	
GCF_000148875.1_ASM14887v1_genomic	
GCF_000277425.1_ASM27742v1_genomic	
GCF_000817025.1_ASM81702v1_genomic	
GCF_000270045.1_ASM27004v1_genomic	
GCF_000224555.1_ASM22455v1_genomic	
GCF_000224575.1_ASM22457v1_genomic	
GCF_000600165.1_ASM60016v1_genomic	
GCF_000148855.1_ASM14885v1_genomic	
GCF_000498315.1_ASM49831v1_genomic	

GCF_001653475.1_ASM165347v1_genomic	
GCF_000982695.1_ASM98269v1_genomic	
GCF_000827025.1_ASM82702v1_genomic	
GCF_000277365.1_ASM27736v1_genomic	
GCF_000829115.1_ASM82911v1_genomic	
GCF_001433495.1_ASM143349v1_genomic	
GCF_000185225.1_ASM18522v1_genomic	
GCF_000685705.1_ASM68570v1_genomic	
GCF_000262655.1_ASM26265v1_genomic	
GCF_000829095.1_ASM82909v1_genomic	
GCF_000277385.1_ASM27738v1_genomic	
GCF_000021465.1_ASM2146v1_genomic	
GCF_000148665.1_ASM14866v1_genomic	
GCF_000196755.1_ASM19675v1_genomic	
Mycobacterium tuberculosis	Oct 5, 2016
GCF_000277735.2_ASM27773v2_genomic	
GCF_000756545.1_ASM75654v1_genomic	
GCF_000153685.2_ASM15368v2_genomic	
GCF_000831245.1_ASM83124v1_genomic	
GCF_000572155.1_ASM57215v1_genomic	
GCF_000008585.1_ASM858v1_genomic	
GCF_000154585.2_ASM15458v2_genomic	
GCF_001544705.1_ASM154470v1_genomic	
GCF_000828995.1_ASM82899v1_genomic	
GCF_001545015.1_ASM154501v1_genomic	
GCF_000023625.1_ASM2362v1_genomic	
GCF_000706665.1_ASM70666v1_genomic	
GCF_000016925.1_ASM1692v1_genomic	
GCF_001544985.1_ASM154498v1_genomic	
GCF_001702435.1_ASM170243v1_genomic	
GCF_000350205.1_ASM35020v1_genomic	
GCF_000422125.1_ASM42212v1_genomic	
GCF_000756525.1_ASM75652v1_genomic	
GCF_001545055.1_ASM154505v1_genomic	
GCF_000331445.1_ASM33144v1_genomic	
GCF_000827085.1_ASM82708v1_genomic	
GCF_000195955.2_ASM19595v2_genomic	
GCF_001275565.2_ASM127556v2_genomic	
GCF_000572175.1_ASM57217v1_genomic	
GCF_000016145.1_ASM1614v1_genomic	
GCF_000572125.1_ASM57212v1_genomic	
GCF_000193185.2_ASM19318v2_genomic	
GCF_000389945.1_ASM38994v1_genomic	
GCF_000224435.1_ASM22443v1_genomic	
GCF_000786505.1_MT49-02_genomic	
GCF_001708265.1_ASM170826v1_genomic	
GCF_000738475.1_ASM73847v1_genomic	
GCF_000270365.1_ASM27036v1_genomic	
GCF_000698475.1_ASM69847v1_genomic	
GCF_000154605.2_ASM15460v2_genomic	
GCF_000572195.1_ASM57219v1_genomic	
GCF_001544955.1_ASM154495v1_genomic	
GCF_000400615.1_ASM40061v1_genomic	
GCF_000364825.1_ASM36482v1_genomic	
GCF_000738445.1_ASM73844v1_genomic	
Salmonella enterica	Feb 26, 2016
GCF_000011885.1_ASM1188v1_genomic	
GCF_001185245.1_ASM118524v1_genomic	
GCF_000380325.1_ASM38032v1_genomic	
GCF_000188735.1_ASM18873v1_genomic	
GCF_000750395.2_ASM75039v2_genomic	
GCF_000486365.2_ASM48636v2_genomic	
GCF_001305235.1_ASM130523v1_genomic	
GCF_001441205.1_ASM144120v1_genomic	
GCF_000272715.3_ASM27271v3_genomic	
GCF_000623375.1_ASM62337v1_genomic	
GCF_000940935.1_ASM94093v1_genomic	
GCF_000953495.1_SINFA_genomic	
GCF_000940975.1_ASM94097v1_genomic	
GCF_000626335.1_ASM62633v1_genomic	
GCF_000636135.1_ASM63613v1_genomic	
GCF_000018625.1_ASM1862v1_genomic	
GCF_000020885.1_ASM2088v1_genomic	
GCF_000623315.1_ASM62331v1_genomic	

GCF_000828595.1_ASM82859v1_genomic
GCF_000750435.1_ASM75043v1_genomic
GCF_000626355.1_ASM62635v1_genomic
GCF_000626115.1_ASM62611v1_genomic
GCF_000626155.1_ASM62615v1_genomic
GCF_000623355.1_ASM62335v1_genomic
GCF_000335875.2_ASM33587v2_genomic
GCF_001484025.1_ASM148402v1_genomic
GCF_000210855.2_ASM21085v2_genomic
GCF_000272755.3_ASM27275v3_genomic
GCF_000020745.1_ASM2074v1_genomic
GCF_000750215.1_ASM75021v1_genomic
GCF_000195995.1_ASM19599v1_genomic
GCF_000473275.1_ASM47327v1_genomic
GCF_000503845.1_ASM50384v1_genomic
GCF_000462995.1_ASM46299v1_genomic
GCF_000624155.1_ASM62415v1_genomic
GCF_000486405.2_ASM48640v2_genomic
GCF_000612325.1_ASM61232v1_genomic
GCF_000486445.2_ASM48644v2_genomic
GCF_000626235.1_ASM62623v1_genomic
GCF_001006525.1_ASM100652v1_genomic
GCF_000487575.2_ASM48757v2_genomic
GCF_000016045.1_ASM1604v1_genomic
GCF_001457675.1_NCTC10384_genomic
GCF_000940895.1_ASM94089v1_genomic
GCF_000988525.1_ASM98852v1_genomic
GCF_001454965.1_ASM145496v1_genomic
GCF_000245535.1_ASM24553v1_genomic
GCF_000742815.1_ASM74281v1_genomic
GCF_000487915.2_ASM48791v2_genomic
GCF_000626195.1_ASM62619v1_genomic
GCF_000487295.2_ASM48729v2_genomic
GCF_000750475.1_ASM75047v1_genomic
GCF_000623295.1_ASM62329v1_genomic
GCF_000020705.1_ASM2070v1_genomic
GCF_000626175.1_ASM62617v1_genomic
GCF_001558355.1_ASM155835v1_genomic
GCF_001409175.1_99_3134_genomic
GCF_000818075.1_ASM81807v1_genomic
GCF_000623115.2_ASM62311v2_genomic
GCF_000213635.1_ASM21363v1_genomic
GCF_000993725.1_ASM99372v1_genomic
GCF_000623095.1_ASM62309v2_genomic
GCF_000430085.2_ASM43008v2_genomic
GCF_000272775.3_ASM27277v3_genomic
GCF_000754375.1_ASM75437v1_genomic
GCF_001441245.1_ASM144124v1_genomic
GCF_000385905.1_ASM38590v1_genomic
GCF_000444445.1_ASM44444v1_genomic
GCF_001302625.1_ASM130262v1_genomic
GCF_000007545.1_ASM754v1_genomic
GCF_000623735.2_ASM62373v2_genomic
GCF_000750335.1_ASM75033v1_genomic
GCF_000272735.3_ASM27273v3_genomic
GCF_000018385.1_ASM1838v1_genomic
GCF_000623195.2_ASM62319v2_genomic
GCF_000022165.1_ASM2216v1_genomic
GCF_001409195.1_C2346_genomic
GCF_001447095.1_ASM144709v1_genomic
GCF_000009525.1_ASM952v1_genomic
GCF_000750415.2_ASM75041v2_genomic
GCF_001293505.1_ASM129350v1_genomic
GCF_001409135.1_10259_genomic
GCF_001540845.1_SO4698_09_genomic
GCF_000009505.1_ASM950v1_genomic
GCF_000963535.1_ASM96353v1_genomic
GCF_001409155.1_98_11262_genomic
GCF_000715155.1_ASM71515v2_genomic
GCF_000258365.1_ASM25836v1_genomic
GCF_000750375.1_ASM75037v1_genomic
GCF_000623135.1_ASM62313v2_genomic
GCF_000020925.1_ASM2092v1_genomic
GCF_000626275.2_ASM62627v2_genomic

GCF_000505365.2_ASM50536v2_genomic	
GCF_000487775.1_ASM48777v2_genomic	
GCF_000626315.1_ASM62631v1_genomic	
GCF_000623055.1_ASM62305v2_genomic	
GCF_000280315.2_ASM28031v2_genomic	
GCF_000624395.2_ASM62439v2_genomic	
GCF_000831045.1_ASM83104v1_genomic	
GCF_000430165.1_ASM43016v1_genomic	
GCF_000626295.1_ASM62629v1_genomic	
GCF_000623275.1_ASM62327v1_genomic	
GCF_000623475.1_ASM62347v2_genomic	
GCF_000430125.1_ASM43012v1_genomic	
GCF_000750455.1_ASM75045v1_genomic	
GCF_000626135.1_ASM62613v1_genomic	
GCF_000272835.3_ASM27283v3_genomic	
GCF_000626415.1_ASM62641v1_genomic	
GCF_000430145.2_ASM43014v3_genomic	
GCF_000623335.1_ASM62333v1_genomic	
GCF_000330485.2_ASM33048v2_genomic	
GCF_000973665.1_ASM97366v1_genomic	
GCF_000188955.2_ASM18895v5_genomic	
GCF_000626255.1_ASM62625v1_genomic	
GCF_001441225.1_ASM144122v1_genomic	
GCF_000430105.1_ASM43010v1_genomic	
GCF_001185215.1_ASM118521v1_genomic	
GCF_000941015.1_ASM94101v1_genomic	
GCF_000008105.1_ASM810v1_genomic	
GCF_000973645.1_ASM97364v1_genomic	
GCF_001302605.1_ASM130260v1_genomic	
GCF_000623395.2_ASM62339v2_genomic	
GCF_000018705.1_ASM1870v1_genomic	
GCF_000750295.1_ASM75029v1_genomic	
GCF_001447115.1_ASM144711v1_genomic	
GCF_000272895.2_ASM27289v3_genomic	
GCF_000973685.1_ASM97368v1_genomic	
GCF_000235545.1_ASM23554v1_genomic	
GCF_000027025.1_ASM2702v1_genomic	
GCF_000283735.1_ASM28373v1_genomic	
GCF_000329365.2_ASM32936v2_genomic	
GCF_000750255.1_ASM75025v1_genomic	
GCF_000006945.1_ASM694v1_genomic	
GCF_000493535.1_DT2_genomic	
GCF_000493675.1_DT104_genomic	
GCF_001305815.1_ASM130581v1_genomic	
GCF_000442415.1_ASM44241v1_genomic	
GCF_000487615.2_ASM48761v2_genomic	
GCF_000626375.1_ASM62637v1_genomic	
GCF_001305835.1_ASM130583v1_genomic	
GCF_000750495.1_ASM75049v1_genomic	
GCF_000272815.2_ASM27281v2_genomic	
GCF_000505705.1_ASM50570v1_genomic	
GCF_000626215.1_ASM62621v1_genomic	
GCF_000831025.1_ASM83102v1_genomic	
GCF_000026565.1_ASM2656v1_genomic	
GCF_000439415.1_ASM43941v1_genomic	
GCF_000743055.1_ASM74305v1_genomic	
GCF_000341425.1_ASM34142v1_genomic	
GCF_000818115.1_ASM81811v1_genomic	
GCF_000486765.2_ASM48676v2_genomic	
GCF_000252875.1_ASM25287v1_genomic	
GCF_000756465.1_ASM75646v1_genomic	
GCF_000484195.2_ASM48419v2_genomic	
GCF_000623455.2_ASM62345v2_genomic	
Staphylococcus aureus	Feb 23, 2016
GCF_000382985.1_ASM38298v1_genomic	
GCF_000382965.1_ASM38296v1_genomic	
GCF_001465635.1_ASM146563v1_genomic	
GCF_001281145.1_ASM128114v1_genomic	
GCF_000967345.1_ASM96734v1_genomic	
GCF_000463055.1_ASM46305v1_genomic	
GCF_001307235.1_ASM130723v1_genomic	
GCF_000769575.1_ASM76957v1_genomic	
GCF_000967325.1_ASM96732v1_genomic	
GCF_001515665.1_ASM151566v1_genomic	

GCF_001183705.1_ASM118370v1_genomic
GCF_001515705.1_ASM151570v1_genomic
GCF_001444345.1_ASM144434v1_genomic
GCF_000010445.1_ASM1044v1_genomic
GCF_000967365.1_ASM96736v1_genomic
GCF_001457495.1_NCTC13435_genomic
GCF_000967385.1_ASM96738v1_genomic
GCF_001046095.2_ASM104609v2_genomic
GCF_000626615.1_ASM62661v1_genomic
GCF_000969225.1_ASM96922v1_genomic
GCF_001183725.1_ASM118372v1_genomic
GCF_000011265.1_ASM1126v1_genomic
GCF_001465755.1_ASM146575v1_genomic
GCF_000815085.1_ASM81508v1_genomic
GCF_000695215.1_ASM69521v1_genomic
GCF_000009585.1_ASM958v1_genomic
GCF_001021875.1_ASM102187v1_genomic
GCF_000025145.1_ASM2514v1_genomic
GCF_000462955.1_ASM46295v1_genomic
GCF_001515685.1_ASM151568v1_genomic
GCF_000756205.1_ASM75620v1_genomic
GCF_000009665.1_ASM966v1_genomic
GCF_000016805.1_ASM1680v1_genomic
GCF_001515745.1_ASM151574v1_genomic
GCF_000237125.1_ASM23712v1_genomic
GCF_000597965.1_ASM59796v1_genomic
GCF_000010465.1_ASM1046v1_genomic
GCF_000013465.1_ASM1346v1_genomic
GCF_000296595.1_ASM29659v1_genomic
GCF_001558795.1_ASM155879v1_genomic
GCF_000237265.1_ASM23726v1_genomic
GCF_000815205.1_ASM81520v1_genomic
GCF_001278745.1_ASM127874v1_genomic
GCF_000253135.1_ASM25313v1_genomic
GCF_000245495.1_ASM24549v1_genomic
GCF_001457515.1_NCTC8532_genomic
GCF_000746505.1_ASM74650v1_genomic
GCF_001549675.1_ASM154967v1_genomic
GCF_000145595.1_ASM14559v1_genomic
GCF_000485885.1_ASM48588v1_genomic
GCF_000412775.1_ASM41277v1_genomic
GCF_000011525.1_ASM1152v1_genomic
GCF_000011505.1_ASM1150v1_genomic
GCF_000027045.1_ASM2704v1_genomic
GCF_000017085.1_ASM1708v1_genomic
GCF_000013425.1_ASM1342v1_genomic
GCF_000159535.2_ASM15953v2_genomic
GCF_000009645.1_ASM964v1_genomic
GCF_000239235.1_ASM23923v1_genomic
GCF_000160335.2_ASM16033v2_genomic
GCF_000204665.1_ASM20466v1_genomic
GCF_000144955.1_ASM14495v1_genomic
GCF_000828035.1_ASM82803v1_genomic
GCF_000383005.1_ASM38300v1_genomic
GCF_000568455.1_ASM56845v1_genomic
GCF_001456215.1_ASM145621v1_genomic
GCF_001045995.2_ASM104599v2_genomic
GCF_001548295.1_ASM154829v1_genomic
GCF_000772025.1_ASM77202v1_genomic
GCF_000418345.1_ASM41834v1_genomic
GCF_001548415.1_ASM154841v1_genomic
GCF_000815125.1_ASM81512v1_genomic
GCF_001021895.1_ASM102189v1_genomic
GCF_000815165.1_ASM81516v1_genomic
GCF_000953255.1_Staphylococcus_aureus_Sa_ILRI_217_genomic
GCF_000815045.1_ASM81504v1_genomic
GCF_000967405.1_ASM96740v1_genomic
GCF_001515765.1_ASM151576v1_genomic
GCF_001027045.1_ASM102704v1_genomic
GCF_001296985.1_ASM129698v1_genomic
GCF_000284535.1_ASM28453v1_genomic
GCF_000009005.1_ASM900v1_genomic
GCF_000012045.1_ASM1204v1_genomic
GCF_001549655.1_ASM154965v1_genomic

GCF_001465675.1_ASM146567v1_genomic	
GCF_000024585.1_ASM2458v1_genomic	
GCF_001027105.1_ASM102710v1_genomic	
GCF_000017125.1_ASM1712v1_genomic	
GCF_000737615.1_ASM73761v1_genomic	
GCF_001045795.2_ASM104579v2_genomic	
GCF_000470865.1_ASM47086v1_genomic	
GCF_000210315.1_ASM21031v1_genomic	
GCF_000695875.1_ASM69587v1_genomic	
GCF_000470845.1_ASM47084v1_genomic	
GCF_000815245.1_ASM81524v1_genomic	

Table S3. For each of the other 5 species, the table lists the strain names and that date on which the sequences were donwloaded from the database.

References

1. Darwin C. On the Origin of Species by Means of Natural Selection. London: Murray; 1859. Or the Preservation of Favored Races in the Struggle for Life.

2. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 1981 Nov;17(6):368–376. Available from: https://doi.org/10.1007/BF01734359.

3. Page RDM, Holmes EC. Molecular Evolution. A Phylognetetic Approach. Blackwell Science; 1998.

4. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. Evolving genes and proteins. 1965;97:97–166.

5. Hartl DL, Clark AG. Principles of Population Genetics. 4th ed. Sinauer Associates; 2006.

6. Wakeley J. Coalescent theory: An Introduction. W. H. Freeman; 2008.

7. Ishii S, Hansen DL, Hicks RE, Sadowsky MJ. Beach Sand and Sediments are Temporal Sinks and Sources of Escherichia coli in Lake Superior. Environmental Science & Technology. 2007;41(7):2203–2209. PMID: 17438764.

8. Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. Trends Genet. 2002 Sep;18(9):472–479.

9. Lerat E, Daubin V, Moran NA. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol. 2003 Oct;1(1):E19.

10. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. PLOS Genetics. 2009 01;5(1):1–25. Available from: https://doi.org/10.1371/journal.pgen.1000344.

11. Dixit PD, Pang TY, Studier FW, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proceedings of the National Academy of Sciences. 2015;112(29):9070–9075. Available from: https://www.pnas.org/content/112/29/9070.

12. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985 Sep;111(1):147–164.

13. Field CM, Sakoparnig T, van Nimwegen E. Recombination Drives Gene Content and Phenotype Evolution in Wild Type *E. coli* Strains; 2019. In preparation.

14. Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E. Automated reconstruction of whole genome phylogenies from short sequence reads. Molecular Biology and Evolution. 2014;31(5):1077–1088.

15. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010;59(3):307–321.

16. Guttman DS, Dykhuizen DE. Clonal divergence in Escherichia coli as a result of recombination, not mutation. Science. 1994 Nov;266(5189):1380–1383.

17. Lawrence JG, Ochman H. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA. 1998 Aug;95(16):9413–9417.

18. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999 Jun;284(5423):2124–2129.

19. Bobay LM, Traverse CC, Ochman H. Impermanence of bacterial clones. Proc Natl Acad Sci USA. 2015 Jul;112(29):8893–8900.

20. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Computational Biology. 2015 02;11(2):1–18. Available from: https://doi.org/10.1371/journal.pcbi.1004041.

21. Croucher NJ, Mostowy R, Andam CP, Hanage WP, Corander J, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. Molecular Biology and Evolution. 2017 02;34(5):1167–1182. Available from: https://dx.doi.org/10.1093/molbev/msx066.

22. Rosen MJ, Davison M, Bhaya D, Fisher DS. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015;348(6238):1019–1023. Available from: http://science.sciencemag.org/content/348/6238/1019.

23. Lai YP, Ioerger TR. A statistical method to identify recombination in bacterial genomes based on SNP incompatibility. BMC Bioinformatics. 2018 Nov;19(1):450.

24. McVean G, Awadalla P, Fearnhead P. A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. Genetics. 2002;160(3):1231–1241. Available from: http://www.genetics.org/content/160/3/1231.

25. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

26. Lin M, Kussell E. Inferring bacterial recombination rates from large-scale sequencing datasets. Nature Methods. 2019;16(2):199–204.

27. Vulić M, Dionisio F, Taddei F, Radman M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci USA. 1997 Sep;94(18):9763–9767.

28. Oliveira PH, Touchon M, Rocha EP. Regulation of genetic flux between bacteria by restriction-modification systems. Proc Natl Acad Sci USA. 2016 May;113(20):5658–5663.

29. Dixit PD, Pang TY, Maslov S. Recombination-Driven Genome Evolution and Stability of Bacterial Species. Genetics. 2017 09;207(1):281–295.

30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357–359.

31. Rambaut A. FigTree software package;. http://tree.bio.ed.ac.uk/software/figtree.

32. Lewontin RC. On measures of gametic disequilibrium. Genetics. 1988;120(3):849–852. Available from: http://www.genetics.org/content/120/3/849.