A putative ecocline in *Klebsiella pneumoniae*

Siqi Liu, Sarah L. Svensson, Daniel Falush*

Shanghai Institute of Immunity and Infection, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai.

*Correspondence to daniel.falush@ips.ac.cn

Abstract

Are quantitative traits important for bacteriology? The quantitative trait models used in animal and plant genetics assume that the trait takes continuous values and is influenced by many variants circulating in the population, each of which acts approximately independently. Amongst the many continuous-valued bacterial traits such as cell size or antibiotic minimum inhibitory concentration, we do not know of any for which these assumptions have been validated. We propose that such traits can be identified if they are the targets of selection and this selection structures overall genetic variation of the species. The phylogenetic trees of global strain collections of *Klebsiella pneumoniae* (KP) have a "backbone", which has not been explained. We propose that this backbone reflects a genome-wide component of variation, which is well-captured by Principal Components Analysis (PCA) applied to diverse KP genomes. We show that this structure can be recapitulated in simulations of diversifying selection on a quantitative trait and cannot be explained by neutral models. An implication is that variation in KP forms an ecocline – i.e. a cline generated by natural selection rather than geography but the trait underlying the cline, if it exists, remains to be determined.

Introduction

Klebsiella pneumoniae (KP) is notorious as a human pathogen, but lives in a great variety of habitats including soil, other animals and even the sea[1–4]. Like many other bacterial species, it has an enormous population size and disperses effectively, with the consequence that most of its global genetic variation can be captured from a single region in Europe[5], Africa[6], Asia[7], or elsewhere[8,9]. In large populations, there is ample opportunity for homologous recombination to assemble favourable combinations of variants, whose frequency can then be amplified by natural selection[10,11]. However, although it is well recognized that selection can amplify the frequency of specific clonal lineages [12] and of alleles at specific loci[13], less work has been done on its effect across an entire genome and species. We here propose that selection has organised variation in KP into an ecocline. If so, then characterising this cline should be informative about the diversity of ecological strategies used by lineages of KP.

Results

The phylogenetic backbone of *Klebsiella pneumoniae*

Genetic relationships amongst bacteria are most often represented using a phylogenetic tree[14]. Trees based on core genome SNPs from global collections *Klebsiella pneumoniae* (KP) isolates are “star-like” (Figure 1A & Supplementary Figure 2, non-redundant dataset; Supplementary Figure 1, entire dataset), with many branches of approximately equal length. The shape of the tree reflects extensive recombination, with the length of the branches reflecting the overall level of diversity in the gene pool[8,15]. However, the overall shape of the KP tree is oval rather than circular as for example seen in *Vibrio parahaemolyticus*[16,17] and KP lineages, instead of radiating from a single point, are connected by a "backbone". This backbone is present in any phylogeny drawn based on a diverse sample of KP genomes but it may be less visually evident than in Figure 1A, depending on the strains sampled (see for example Supplementary Figure 1, including redundant isolates), the rooting of the tree, and the phylogenetic and plotting algorithms.

Another common method for displaying genetic relationships is Principal Component Analysis (PCA)[18,19]. PCA identifies axes of variation that together explain the overall diversity within the species, with the first component, PC1, explaining the largest fraction of variation. Analysis using PCA implicitly assumes that variation is continuous, since each component can take arbitrary numerical values, but discrete structure is likely to result in strains forming clusters in PCA space. Furthermore, PCA is sensitive to all signals of relatedness, including clonal structure and can be prone to overfitting, especially where there are orders of magnitude more genetic markers than strains[20]. We reduced signals of recent clonal relationship from the data by applying the method to the non-redundant dataset shown in Figure 1A.

When PCA was applied to the global KP core genome SNP dataset, the first component (PC1) explained 16.7 % of the variation, with subsequent components each explaining approximately 8% or less (Figure 1B). PC1 captured the genetic variation underlying the phylogenetic backbone of the species, since strains with extreme PC1 values are found at either end of the backbone, while strains branching from the middle have intermediate values (Figure 1A). Strain loadings for PC1 are approximately continuous, albeit with a paucity of intermediate values (Figure 1B).

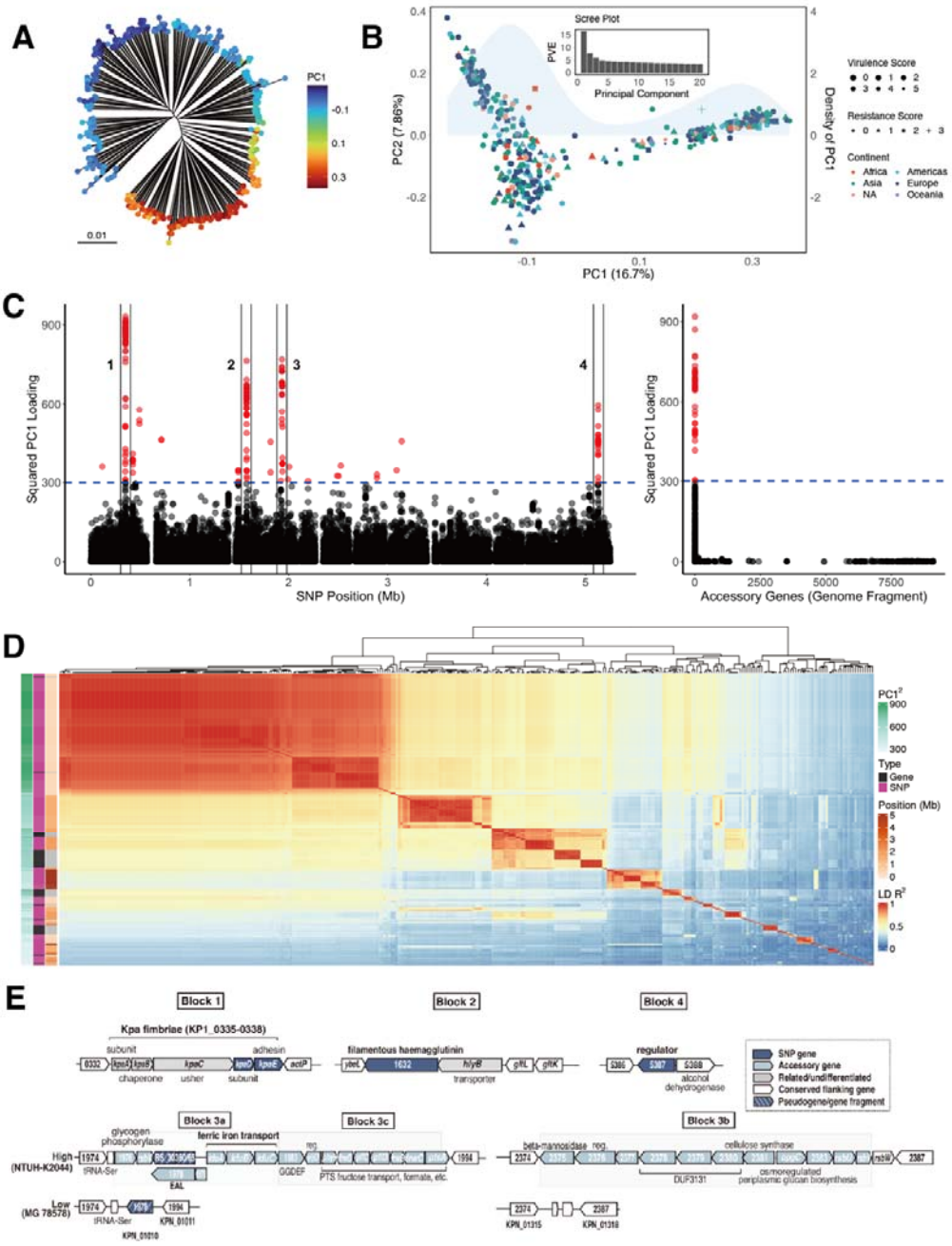


Figure 1. Population structure and genetic variation in a non-redundant *Klebsiella pneumoniae* dataset. (A) Maximum-likelihood phylogenetic tree of non-redundant KP based on core SNPs, with a color gradient showing PC1 variation. **(B)** PCA of non-redundant KP based on the pruned SNPs. Colors indicate geographic origin, while marker size and shape represent resistance and virulence scores, respectively, determined by Kleborate[21]. The inset scree plot shows the variance explained (PVE) by the first 20 principal components. Sky blue represents the density of PC1. **(C)** Manhattan plot of squared SNP (left) and gene (right, accessory genome) loadings for PC1. Variants with values greater than 300 are defined as high loading variants and are marked in red. For accessory genes, the ordering is based on the list from the Panaroo[22] output. Numbered peaks are present in Blocks displayed in panel D. For zoom-in of peaks, see Supplementary Figure 4B. **(D)** Heatmap of pairwise linkage disequilibrium (LD) R^2 values for high loading variants, with the first four blocks (including subblocks) labelled. **(E)** Functional annotation of blocks 1-4.

As well as explaining a high fraction of variation between strains, PC1 was reproducible in different analyses (Supplementary Figure 3). Specifically, PC1 values obtained when performing analysis on one half of the chromosome were correlated with those of PC1 calculated using the other half, implying that the signal is not due to specific genetic loci (Supplementary Figure 3 E and F). PC1 strain loadings were also consistent if components were defined by strains on the right-hand side of the oval or those on the left, implying that they are not generated by clonal structure (Supplementary Figure 3 A-D). They were also consistent when calculated using SNPs with extreme PC1 loadings or loadings that are closer to 0 (Supplementary Figure 3 G-J), implying that the signal affects genetic variation genome-wide.

The reproducibility of the cline in all three tests for the KP data contrasts with the irreproducibility of clines seen in a simulated neutrally evolving dataset, where no correlation is seen when PCA is applied separately to the strains on the right-hand side and left-hand side of the phylogenetic tree (Supplementary Figure 4).

Vibrio parahaemolyticus (VP) also represents a useful benchmark for half-matching tests as it has well characterized population and ecospecies structure[16,17,23]. When half matching tests were applied to a global non-redundant dataset of VP, all three half-matching tests gave reproducible results that reflected the population structure in the dataset (Supplementary Figure 5). However, when analysis was restricted to the Asian population (VppAsia), ecospecies structure gave weaker correlations in two of the three tests but there was no correlation between low loading and high loading loci, consistent with a single randomly mixing gene pool for most the genome[17,23] (Supplementary Figure 6). Thus, the reproducibility of PC1 in all three tests in KP shows that the structure is genome-wide and in this respect is more similar to geographic population structure and less similar to ecospecies structure, which is only present in a fraction of the genome[23].

In addition to strain loadings (Figure 1B), PCA generates SNP loadings, indicating which genomic variants are most strongly associated with the component. SNP loadings for PC1 showed four clear peaks (Figure 1C (*left*); Supplementary Figure 7A). Each peak contained many SNPs within less than 10kb (Supplementary Fig. 7B). We also calculated loadings for gene presence-absence, based on a Panaroo-generated[22] pangenome of the non-redundant strains, and found many had high loadings (Figure 1C, *right*).

To investigate the relationships between the loci most strongly correlated with PC1, we chose a cutoff of 300 for squared PC1 loadings and calculated pairwise LD values for all SNPs and genes above the cutoff. Hierarchical clustering of LD values organized the extreme loading variants into 44 blocks (Figure 1D). Four of these blocks include the SNPs with the most extreme loadings within the four peaks seen in Figure 1C. The organization and gene content of these blocks is shown in Figure 1E and Supplementary Figure 5.

Block 1 (Supplementary Figure 8A & B), which has the most extreme loading variants, includes two genes with SNPs (KP1_0337 & KP1_0338). These genes are part of the five-gene Kpa fimbriae locus, which is highly conserved in both KP and *K. variicola*[24,25]. While the Kpa fimbriae are less characterized than the *fim* locus, experiments in LM21 (a low-loading strain) found that absence of *kpaC*, encoding the usher, significantly reduced *in vitro* biofilm formation in microtitre plates. The Δ*kpaC* mutant did not affect binding to either a human cell line or *Arabidopsis* seedlings[26]. The most differentiated gene (KP1_0338/*kpaE*) encodes the adhesin subunit.

Block 2 (Supplementary Figure 8C & D) and Block 4 (Supplementary Figure 8G & H) also include SNPs in single genes (KP1_1632, filamentous haemagglutinin N-terminal

domain-containing protein and KP1_5387, winged helix-turn-helix transcriptional regulator, respectively) (Figure 1E). The former, as for the Kpa fimbral adhesin, suggests a difference in adhesion, but so far appears uncharacterized. The regulator encoded by KP1_5387 is also unstudied, but seems always encoded next to an alcohol dehydrogenase in KP, *K. variicola*, *K. quasipneumoniae*, *K. africana*, *K. quasivariicola* strains.

Block 3 (Supplementary Figure 8E & F) includes both SNPs and accessory genes and can be subdivided into three sub-blocks (3a, 3b, 3c) based on LD patterns. The majority of strains fall into two configurations (Figure 1E, Supplementary Figure 8F), with the high loading configuration including all accessory gene elements and the low loading configuration including none, but there are also strains with intermediate configurations.

Block 3 occurs in two different regions of the genome in the reference strain, with block 3b occurring in a different genome region from blocks 3a and 3c. Blocks 3a and 3c were previously highlighted as associated with invasive strains[27], as well as those implicated in bovine mastitis[28]. Block 3a includes SNPs in two genes, one of which is a c-di-GMP-related phosphodiesterase encoded by all KP. Key genes of 3a are *kfuABC*, encoding a ferric iron ABC transporter, which aids iron acquisition and impact mouse colonization[29]. The block also encodes two c-di-GMP-related proteins, a regulator, and glycogen/PTS sugar utilization/uptake genes. Block 3c occurs adjacent to Block 3a in the genome and contains genes related to sugar uptake, metabolism, and regulation (again via c-di-GMP). Block 3b contains accessory genes encoding a cellulose synthase, three DUF3131 proteins, OpgC (osmoregulated glucan production), and several regulatory proteins (hybrid sensor kinase, anti-sigma-factor and its antagonist).

The backbone is not explained by clonal relationships or geographic structure

The presence of this substantial component of variation begs the question: how has it arisen? Simple models of a bacterial population evolving with mutation and homologous recombination failed to recapitulate the patterns we observed in KP. Simulations with a population size of 2000 and one homologous recombination event per genome of size 10 kb approximately matched the overall LD (linkage disequilibrium) decay curve seen in the real data (Figure 2A). However, for these parameters, there was no phylogenetic backbone evident, and PC1 failed to explain as much variance as in the real data (Figure 2B & 2C). As noted above, PC1 was also not reproducible when applied to different halves of the tree in a half matching test (Supplementary Figure 4). More generally, we have never seen a phylogenetic backbone in any simulations of a single neutrally evolving population (*e.g.*, Supplementary Figure 9).

The model of evolution shown in Figure 2B assumes that recombination rates are uniform across the species. More complex models are possible, for example in which recombination rates are high between closely related strains but lower between more distant ones. Such a pattern can arise for example if homologous recombination is suppressed by mismatches between donor and recipient sequences[30]. Such models are thought to explain much higher recombination rates within *Escherichia coli* phylogroups than between them[31]. Many models with these properties are possible, but since they all reduce the effectiveness of recombination, they are likely to make it difficult to recapitulate the low levels of LD in KP at large genetic distances. Moreover, the models that have been investigated previously cause species to cluster into discrete groups[32,33], rather than generating an approximately continuous pattern of variation. We are unaware of any model in which variable recombination rates within a single population predict a continuous structure similar to that seen in KP (Supplementary Figures 9

& 10).

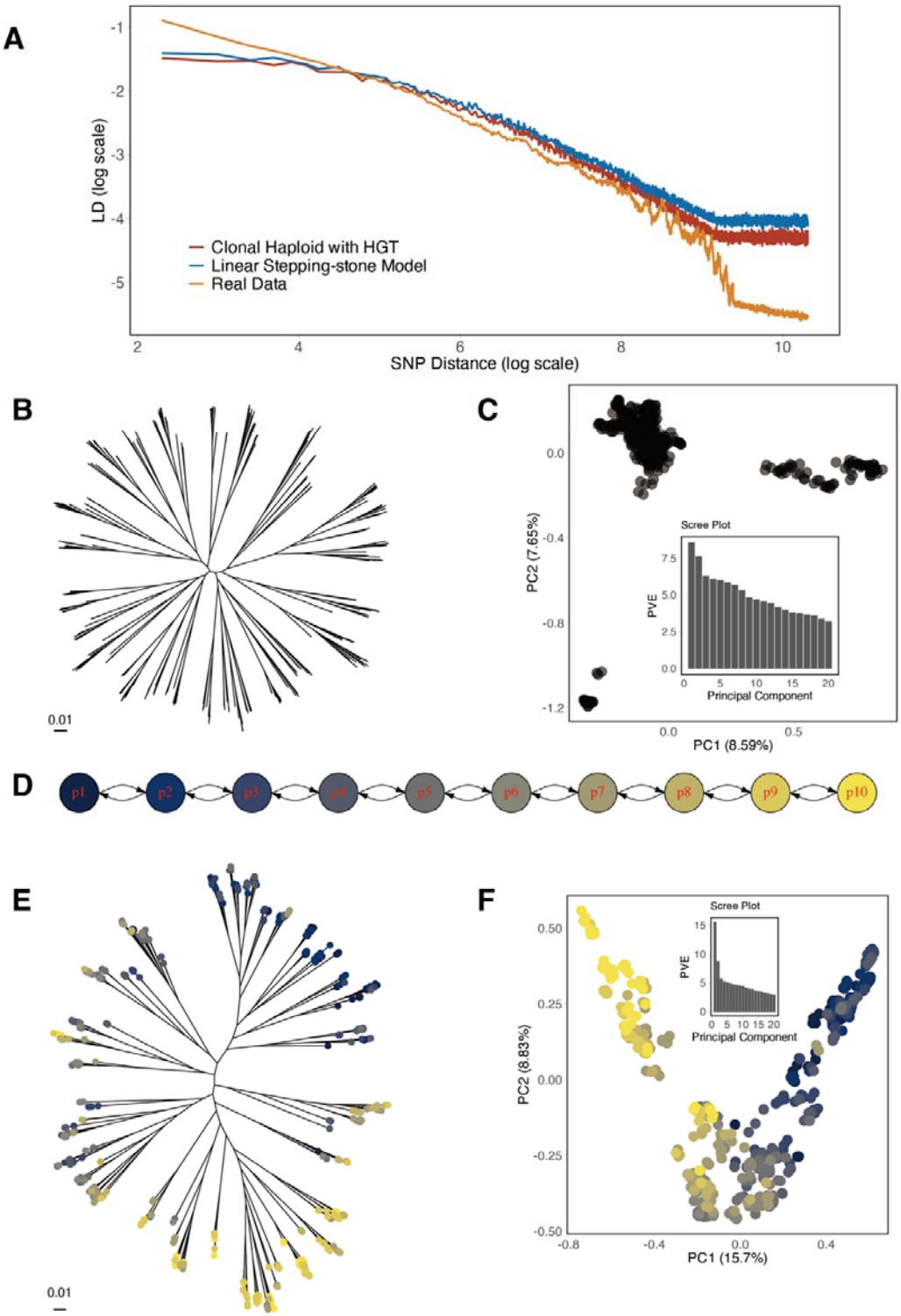


Figure 2. Simulated neutral models. (A) Linkage disequilibrium (LD) decay comparison to real data for best-fit neutral single population and 10 population stepping-stone models. Phylogenetic tree **(B)** and PCA results **(C)** for single population evolving neutrally. **(D)** Illustration of 10 population stepping-stone model. Phylogenetic tree **(E)** and PCA **(F)** for best fit stepping-stone model, coloured as shown in **(D)**.

Genetic differentiation can arise due to geographical structure, or other factors that restrict gene flow between particular subsets of the population[16,34]. To explore the properties of geographical population structure we next simulated a stepping-stone model with 10 demes (Figure 2D). The demes are organised along a line, with a small number of migrants moving between adjacent demes each generation. Data simulated according to this model gave a tree with a phylogenetic backbone similar to that seen in KP, with PC1 explaining a similar fraction of the overall variation (Figure 2E, F and Supplementary Figure 11) and also approximately matched the LD decay curve (Figure 2A and Supplementary Figure 12). However, the distribution of the SNP loadings for PC1 for these simulations did not match the real data, because there were many more SNPs with intermediate loading values (Supplementary Figure 13). Geographic population structure creates differentiation between populations through genetic drift, which is a random process that affects all of the genome but to differing degrees. The discrepancy in SNP loading values between simulated and real data implies that the variation in differentiation values in KP is higher than expected under a pure random drift model.

A longer-tailed distribution of drift levels, and hence PC1 values, could be generated by natural selection at a fraction of loci associated with adaptation to specific demes. However, even if a geographic model with some selection could be made to fit the data, this model seems to be poorly motivated in the case of KP. There is no evidence of KP being structured geographically (Figure 1B)[8,15] or of there being barriers to gene flow between ecological demes, since for example it is readily acquired by humans from a wide variety of ecological sources[35,36]. It is therefore hard to see how a model based on strong isolation between demes could be applicable to the species.

KP variation is consistent with diversifying selection on a quantitative trait

Many traits are subject to diversifying selection[37]. For example, the long necks of giraffes allow them to reach leaves that are high up in trees but are energetically costly to grow and maintain. In a population of giraffes, the individuals with the longest and shortest necks might therefore have a fitness advantage over intermediate individuals. The individuals with the longest necks can reach leaves that other individuals cannot, while those with the shortest necks have the lowest energetic costs[38].

We have performed simulations of selection on a single quantitative trait in a recombining bacterial population. We assumed that fitness is the same for all values of the trait, except that the bacteria with the highest and lowest values of the trait have a fitness advantage (Figure 3). Each mutation had a consistent effect on the magnitude of the trait. Simulation results showed that low recombination rates allow progressive divergence in trait values, while high rates homogenise the population, leading to a circular star-shaped phylogeny (Figure 3D). At intermediate recombination values, a phylogenetic backbone appeared, similar to the one seen in the KP data (Figure 1A). For low and intermediate recombination rates (1 kb, 10 kb), PC1 values were strongly correlated with the true simulated trait values (Supplementary Figure 14).

In the simulation shown in Figure 3 we assumed that every mutation in the genome has the same magnitude of effect on the trait, either increasing it or decreasing it by 1. However, these simulations failed to recapitulate the sharp peaks in SNP loadings seen in the KP data (Supplementary Figure 15 vs. Figure 1C). We therefore modified the simulations, by assuming that the distribution of the magnitudes of the effect of SNPs on the trait followed a gamma distribution (Supplementary Figure 16). High values of the gamma parameter

correspond to some mutations having a much larger effect on the trait than others.

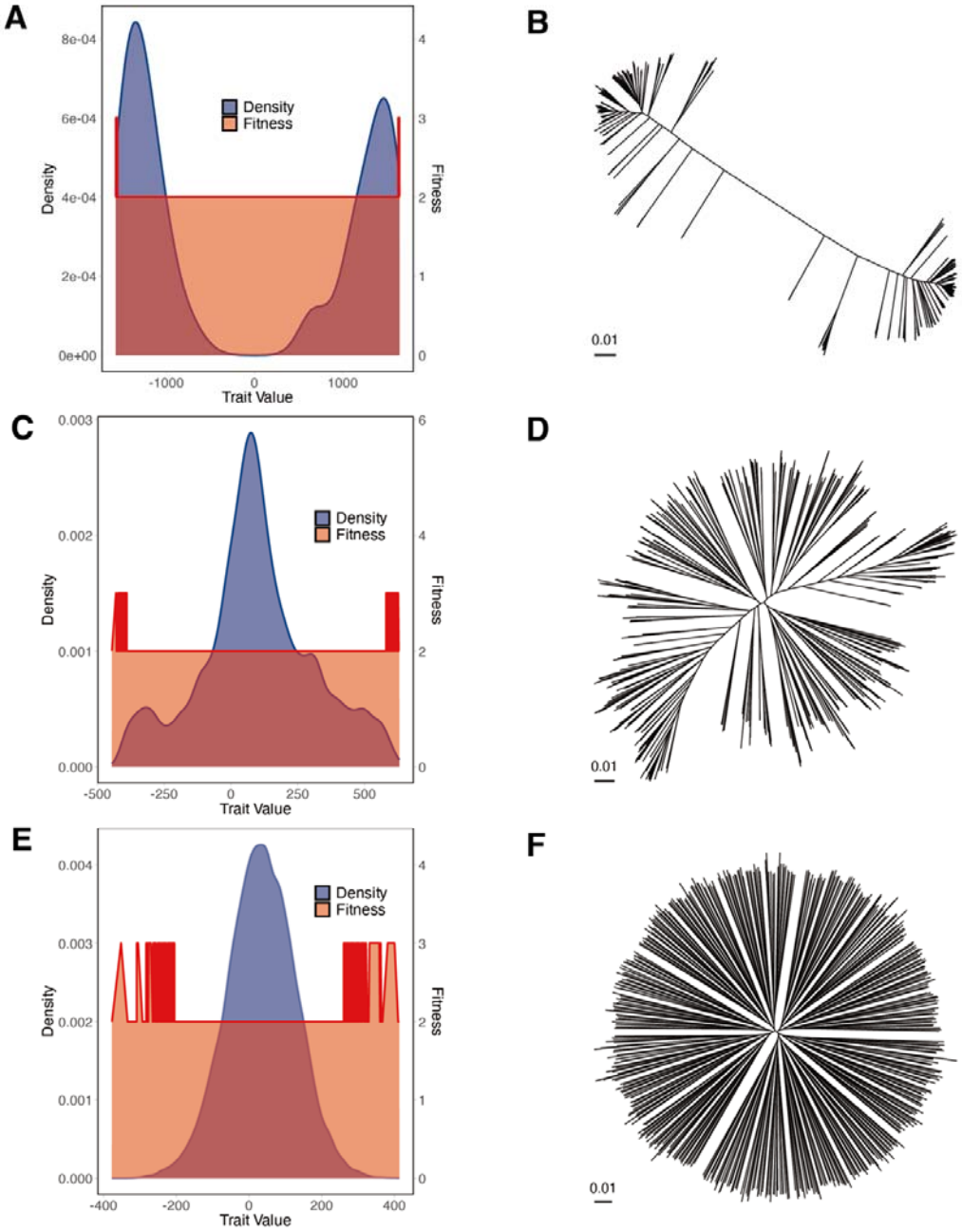


Figure 3. Simulated model of diversifying selection on a polygenic trait shows population structure depends on the balance between selection and recombination. (A, C, E) Trait value density distributions and fitness coefficients for HGT lengths of 1 kb, 10 kb, and 100 kb (Top to bottom) in final generation of simulation. (B, D, F) Phylogenetic trees depicting evolutionary relationships under each HGT length.

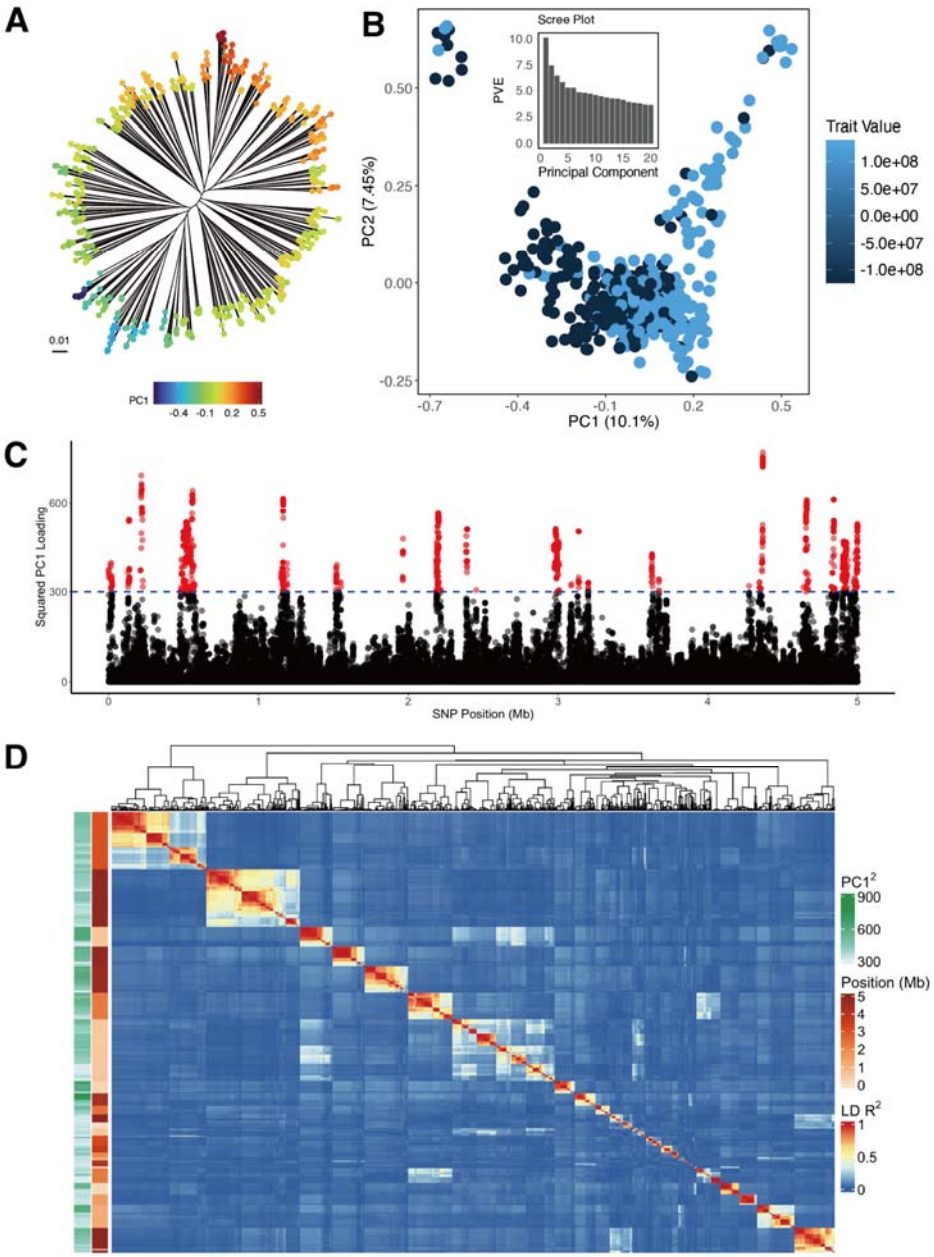


Figure 4. Best fit simulated diversifying selection model (5 Mb). (A) Phylogenetic tree. **(B)** PCA results. **(C)** Squared SNP loadings on PC1 using the same threshold with Fig 1C. **(D)** Heatmap of pairwise LD R^2 values for high loading SNPs.

Based on simulations with a 1MB chromosome length (Supplementary Figure 16), we found that the parameters of the gamma distribution had little impact on the structure of the phylogenetic tree and the results of the PCA. However, the parameters determining the SNP coefficients had a significant influence on the distribution of trait values and the distribution of SNP loading in the population. In real data, only a few peaks are observed, which may be due to the fact that only a small number of loci strongly influence the trait of individuals, while other loci, although contributing, have very small effects. We found that a reasonable qualitative fit to the KP data could be found by simulating a 5 Mb genome length (similar to the true length), assuming trait values for each SNP followed a gamma distribution with a

mean of 1 and a scale parameter of 5×10^{6} (shape of gamma distribution = 2×10^{-7}) (Figure 4). Specifically, we were able to recapitulate a phylogenetic backbone, a PC1 correlated with the trait value and a small number of high loading peaks (Figure 4C; Supplementary Figure 17), each of which is associated with SNPs with substantial influence on trait values.

Discussion

Do bacteria have quantitative traits?

Quantitative trait models are central to our understanding of human disease genetics and animal and plant biology[39–41]. These models are applicable to many morphological and disease susceptibility traits, which fit the assumptions of the model, namely that the genetic basis of variation of the trait in the population is highly polygenic, with most variants acting approximately independently of each other in their influence on the trait value[42,43]. These features make their evolution relatively predictable[44]. We could for example use quantitative genetics to predict the change in giraffe neck length in a given population over a few generations of artificial selection, based on the intensity of selection and the additive genetic variation present within the population before selection takes place[45].

In microbiology quantitative trait models have historically had less importance. Minimum inhibitory concentrations (MICs) of antibiotics are a well-studied trait that can be treated quantitatively[46], but the episodic nature of exposure to antibiotics and the strong selection coefficients involved accentuates the role of variants of large effect, as well as epistatic interactions[47]. Moreover, the population-level response to selection depends on specific lineages rising in frequency[48]. All of these factors reduce the predictive power of the additive models used in animal and plant biology[49,50]. Bacteria have many other traits that take continuous values[51], such as cell size and doubling time in rich media but for most traits and species, we lack a practical motivation for attempting to apply quantitative trait models such as being able to predict the response to selection[45].

We have here shown that there are scenarios in which selection on a quantitative trait in a bacterial population can play a substantial role in organizing variation. Specifically, we study the case of diversifying selection on a trait. In our simulations strains with the highest and lowest values of the trait have a small fitness advantage. Over time, this means that some strains systematically acquire variants that increase trait values, while others acquire variants that decrease trait values. However, this tendency is counteracted by recombination, which randomizes variation within the population. If the recombination rates at each locus are low, strains are likely to differentiate into two discrete groups with distinct trait values (Figure 3A). If recombination rates are higher, then intermediate strains can become common (Figure 3B, 3C). Nevertheless, selection for extreme values ensures that variants which increase trait values will be found together more often than expected under random assortment. These non-random associations represent a form of linkage disequilibrium, which can be detected with a variety of statistical approaches.

One approach to detecting non-random associations between variants is PCA, which is particularly suited to identifying factors that act genome-wide such as population structure[52,53]. When we perform PCA on a simulated population under diversifying selection, the values of the first component, PC1, for each strain can be strongly correlated with trait values (Supplementary Figure 14). We call the gradient in PC1 values an ecocline, since it does not require geographical or other barriers to gene flow and is generated simply by natural selection on a trait, which in real populations is likely to generated by ecological factors.

The ecocline we see in our simulations depends on a balance between selection and recombination. As such it is unlikely to arise in the organisms where quantitative traits have previously been studied, such as outbreeding animals and plants. In these species chromosomal segregation randomizes variants each generation, making recombination a stronger force than selection. As such, an ecocline in the sense we propose it is a novel form of genetic structure. The term ecocline is sometimes used by plant biologists, but in their examples they propose that geographical variation in selection coefficients has generated a geographical genetic cline[54,55]. Our usage of the term is more strictly accurate, since the clines in plants are fundamentally mediated by geography.

Reverse ecology of bacterial ecoclines

We do not yet have compelling evidence for any bacterial ecocline. To identify and understand them biologically, we propose a reverse ecology approach[56], in which we first identify that variation within a species is organised into a putative ecocline and then proceed to establish the phenotypic trait or traits that might underly that cline based on functional annotation and/or laboratory experiments, before finally attempting to understand its ecological significance and what maintains variation in the population. This approach entails making an ecological hypothesis – that diversifying selection on a trait has generated clinal variation - and a functional hypothesis - that the genetic variants showing clinal variation are modifying the selected trait. Demonstration of an ecocline requires validation of both hypotheses.

As our application to simulated data demonstrates, PCA is a potentially a useful tool for detecting possible ecoclines. However, considerable caution must be exercised in interpreting its results, because PCA will detect apparently clinal variation in essentially any dataset. The PC1 cline in KP is noteworthy because of the large fraction of variation it explains (Figure 1B), its reproducibility in analyses using subsets of the data (Supplementary Figure 3) and the difficulty in explaining these features using other models for generating genetic variation. Specifically, we have been unable to obtain similar patterns using simulated neutral models (Figure 2), with or without geographic structure. Models with geographical barriers to gene flow and geographical variation in selection coefficients might be able to generate an ecocline-like structure but they would also entail geographical differentiation, which there is no trace of in KP.

Fortunately, we do not have to rely on PCA alone to visualize and interpret the clinal variation that we see. In the simulations that best fit the KP data most of the variation in trait values in the population is generated by less than 20 sites in the genome. When PCA is applied to this data, the sites with the largest effect on trait values have extreme PC1 loadings (Supplementary Figure 16B), implying that if selection on a trait has generated PC1 then we can potentially identify the variants in the genome that have the largest effect on trait values based on their extreme PC1 loadings. We can visualize the linkage disequilibrium between sites with extreme loading values (Figure 1D) and see if the patterns are consistent with what we see in simulated ecoclines (Figure 4D).

In our diversifying selection model, linkage disequilibrium between sites arises because of the tendency of sites that increase trait values to accumulate on the same genetic backgrounds. There are no direct fitness interactions between the sites themselves. If there are no other forces generating LD, then this model implies that the sites with the largest effect on trait values should be in the strongest disequilibrium with all other sites in the genome. In practice, there are other forces generating LD, specifically including genetic linkage between sites that are physically close together on the chromosome. This means that in both simulated and real

data, the genome is organized into physically linked LD blocks, which can be seen as red squares on the diagonal in both Figure 1D and Figure 4D.

Once the block-like structure of LD in the chromosome created by physical linkage is taken into account, the diversifying selection model implies that most of the remaining LD should generated by selection of the trait. In the KP data, Block 1 has the most extreme PC1 loadings, so that this block is predicted to have the largest effect on the putative quantitative trait and concordant with this prediction, it has consistently high pairwise LD with loci in other blocks, while blocks 5-44, which have lower loading values, have consistently low pairwise LD values with other blocks. The LD patterns that we see thus seem broadly consistent with selection being mediated by the effect on a trait, rather than pairwise interactions between loci, which would generate a more patchwork-quilt like pattern.

There is one feature of the real data which is not matched by the simulations, namely that the four genome blocks with the most extreme PC1 loadings are also the genome blocks with the largest number of variants, with Block 1 having more than 100 SNPs (these 139 SNPs are in two genes, which are the same as the accessory genes). We think that this difference reflects a biologically unrealistic feature of our simulations, namely that each variant is a single site acting independently. However, functionally important variation can require changes at multiple sites. For example, Block 1 encodes the fimbria locus (Supplementary Figure 8A & 8B). Within the KP population, there are two highly distinct haplotypes, which differ on average at 139 sites, which potentially encodes for a substantially different biological function. This would be impossible to achieve using a single SNP.

The functions of the genes in the top four blocks are not obviously related. With the most differentiated gene being the adhesin component of Kpa fimbriae, this suggests the trait may be related to adhesion/diffusion. This is also supported by differentiated of several proteins related to c-di-GMP metabolism, a second messenger is thought to signal/control trade-offs between sessile/surface and motile/free-living behaviours[57]. The accessory *kfuABC* locus suggests there is some involvement of iron acquisition. While these genes were previously found to be over-represented in invasive clinical isolates and bovine mastitis strains, we could not find any evidence that the trait is directly and specifically involved in human infection. One thing that is clear is that virulent or resistant genotypes can assemble on all positions of the ecocline.

Despite the provisional nature of the current evidence, we have been emboldened to propose KP variation might be organized into an ecocline based on our previous finding of "ecospecies" in two other bacterial genera[17,23]. An ecospecies is similar to an ecocline, in that genetic differentiation is maintained by selection, in the absence of other barriers to genetic exchange. However, in an ecocline strains can fall anywhere along a gradient, while each strain must belong to one ecospecies or another and there are no intermediates. Because of their discrete character, ecospecies represent a particularly striking pattern of variation, which seem inconsistent with neutral models. Differentiation between Hardy and Ubiquitous *Helicobacter pylori* is largely restricted to 100 genes, which each have nearly fixed differences in at least 5 SNP and there are no intermediate genotypes[23]. In *V. parahaemolyticus* differentiation is restricted to 27 core genes, and once again there are no intermediate genotypes[17].

For ecospecies, we have also been able to make stronger functional predictions. Initial characterisation of ecospecies-specific genes led to the hypotheses that the *V. parahaemolyticus* ecospecies were related to swarming behaviour (lateral flagella) while the *Helicobacter* ecospecies are related to host diet (metal uptake genes) and interaction with the

host (outer membrane proteins)[17]. Furthermore, the Hardy ecospecies of *H. pylori* has a lineage that has made a host jump to big cats but has only been isolated from humans who have historically had little access to plant matter in their diet, providing an immediate functional hypothesis: that the Hardy ecospecies survives better in the stomachs of carnivores[23]. Further evidence for this hypothesis is that Hardy strains carry an additional urease gene, cofactored by iron instead of nickel, which is otherwise restricted to *Helicobacter* species from carnivorous mammals[23].

The absence of equivalent functional predictions means that our current designation of an ecocline in KP needs to be validated by phenotypic data. However, based on experience in human genetics, it is not unexpected that some traits which are closely linked with organismal fitness do not show simple functional associations. For example, height in humans appear to be close to "omnigenic". Height it is influenced to some degree by variants that influence many different aspects of human metabolism and development, while mutations that act directly on the trait by altering body morphology are often under strong negative selection and therefore only occur at low frequencies[43,58]. Core bacterial traits such as doubling time are likely to have similar characteristics, and if the trait itself is under some form of diversifying selection, it might be difficult to predict the functional category of the most important variants influencing the trait in natural populations.

If the ecological hypothesis is correct, and the genes with extreme loadings are influencing a single trait, then characterization of the effects of these variants should simultaneously provide considerable insight into gene function, ecology and evolution for a key bacterial fitness trait. This would likely represent a breakthrough as these aspects are hard to study individually, handsomely justifying the initial effort required to identify the trait.

It should be cautioned however that diversifying selection may not reflect a simple trait. An alternative is that there is some form of evolutionary trade off driving selection. For example, giraffes could reach leaves high up leaves in trees either by growing a long neck or by being able to jump high. Giraffes with longer necks are likely to be particularly unsuited to jumping, so that selection favoring either solution but not both could create linkage disequilibrium between jumping genes and neck length genes. Further, even if there is a simple trait in KP that is under diversifying selection, it may not be readily recreated under laboratory conditions.

A challenge and virtue of the reverse ecology approach is that it confronts us with the gulf between laboratory conditions and the natural environment that we seek to understand, and more specifically our difficulty in empathizing with bacteria and predicting the challenges they are likely to face. In any case, we propose that investigation of phenotypic differences associated with the PC1 cline has the potential to deliver key insights into KP ecology, as well as on the biology of quantitative traits in microbes.

Methods

Genome collection and Sequence alignment

We collected 1,421 whole genome sequences of *Klebsiella pneumoniae* from the NCBI database, selected based on their high quality and availability, along with a reference genome (*K. pneumoniae* strain NTUH-K2044, NC_012731.1). Additionally, we obtained genome metadata using the NCBI Datasets command-line tools (v16.12.1)[59] and used Kleborate (v2.4.0)[21] to extract clinically-relevant genotypic information on strains (Table S1). All genomes were aligned to the NTUH-K2044 (NC_012731.1) reference genome using Nucmer (v3.1)[21] for reference-based alignment.

In total, 8,331 global Vibrio parahaemolyticus genomes (Table S2) were downloaded from the NCBI database, based on the dataset from the study 'Why panmictic bacteria are rare' [61] and aligned to the *V. parahaemolyticus* RIMD 2210633 reference genome (RefSeq NC_004603.1 and NC_004605.1) using Nucmer (v3.1) as previously described[62], following the same method used for *K. pneumoniae*.

Core genomic variant calling and Definition of Non-redundant Dataset

SNP-sites (v2.5.2)[63] was used to call all the variants from the *K. pneumoniae* genome alignment. Core SNPs were defined as the variants present in at least 99% of the genomes across the whole genome alignment, resulting in 850,158 core SNPs after extraction. Using these core SNPs, pairwise distances were calculated across all genomes with snp-dists (v0.8.2)[64]. To define a non-redundant dataset, genomes with pairwise SNP distances greater than 20 kb were selected, resulting in 369 non-redundant genomes (Supplementary Figure 2A).

We applied the same method to the *V. parahaemolyticus* dataset. Genomes with pairwise SNP distances greater than 50 kb were considered non-redundant, a threshold chosen to significantly shrink the dataset, resulting in a final dataset of 667 *V. parahaemolyticus* genomes with 1,619,597 core SNPs.

fineSTRUCTURE and Population assignment

To further investigate the population structure of the non-redundant 369 *K. pneumoniae* strains described above, we performed chromosome painting and fineSTRUCTURE analysis. Using ChromoPainter mode in fineSTRUCTURE (v4.1.0)[65], we inferred chunks of DNA donated from a donor to a recipient for each recipient haplotype and summarized the results into a co-ancestry matrix. Subsequently, fineSTRUCTURE clustered 369 individuals based on the co-ancestry matrix, with 100,000 iterations for both burn-in and Markov Chain Monte Carlo (MCMC) (Supplementary Figure 2B). In this analysis, only variants present in all non-redundant strains were considered, as fineSTRUCTURE requires input data without missing values.

To assign a population to the non-redundant *V. parahaemolyticus* dataset, we followed the same fineSTRUCTURE approach using haplotype data from core SNPs (Supplementary Figure 5A). Based on the fineSTRUCTURE results, the genomes were classified into five populations: VppAsia (506 strains), Hybrid (5 strains), VppUS1 (49 strains), VppUS2 (23 strains), and VppX (84 strains). Within the VppAsia dataset, strains were further categorized into ecospecies types (EG1a and non-EG1a) based on 26 differentiating genes[17].

Construction of the Phylogenetic Tree

Based on the core SNPs defined above, we constructed nucleotide sequence alignment files for both the entire *K. pneumoniae* dataset and the non-redundant *K. pneumoniae* dataset. Phylogenetic trees were constructed using FastTree (v2.1.10)[66] with the "-nt" option under the default model. The trees were visualized and compared using the ggtree package (v3.12.0)[67] in R, highlighting the phylogenetic relationships within the whole *K. pneumoniae* dataset (Supplementary Figure 1) and the non-redundant dataset (Figure 1A).

Similarly, we used the same software and parameters to construct phylogenetic trees for the non-redundant and VppAsia *V. parahaemolyticus* datasets, as well as the simulated datasets.

Principal Component Analysis

Before performing PCA (Principal Component Analysis) on the non-redundant *K. pneumoniae* dataset, PLINK (v1.9)[68] was used to filter out loci with a minimum allele frequency (MAF) below 2% from the core SNPs, leaving 102,226 variants. Linkage disequilibrium (LD) pruning was subsequently performed to remove linked SNPs, where one SNP from each pair was removed if the LD (R^2) exceeded 0.1 within a 50 kb window. The window was then shifted forward by 10 bp, and the procedure repeated to ensure comprehensive coverage. A total of 33,589 pruned variants were obtained and used for conducting PCA on the non-redundant dataset with PLINK (v1.9)[68], extracting and reporting the first 20 principal components by default.

Similarly, we used the same software and parameters to perform PCA on the non-redundant *V. parahaemolyticus* dataset, which included 220,737 SNPs after MAF filtering and 69,537 SNPs after pruning. We also applied PCA to the VppAsia dataset, containing 207,292 SNPs after MAF filtering and 65,058 SNPs after pruning, as well as to the simulated datasets independently.

Variant loadings calculation

We first calculated the correlation between the original variables and the principal components, which is specifically referred to as the PCA loading for strains. The loadings can be determined as the coordinates of the variables divided by the square root of the eigenvalue associated with that principal component (Equation 1). Next, we transform the MAF-filtered SNP matrix into a binary matrix, assigning a value of 1 to minor alleles and 0 to all others. Finally, through matrix multiplication, we computed the product of the core SNP matrix and the PCA loading matrix to obtain the loading values for the MAF-filtered SNPs (Equation 2).

$$strain\ loadings = eigen\ vectors \cdot \sqrt{eigen\ values} \quad \text{(Equation 1)}$$

$$SNP\ loadings = strain\ loadings \cdot SNP\ matrix \quad \text{(Equation 2)}$$

To better highlight the significant peaks in the non-redundant *K. pneumoniae* result, we plotted the squared PC1 SNP loading values and marked the dots with squared loading values greater than 300 (Figure 1C). Supplementary Figure 3A shows the original SNP loading values for PC1, and Supplementary Figure 3B presents the zoomed-in views of the four peak regions with locus tags.

Similarly, we used the same method to calculate SNP loadings for the non-redundant *V. parahaemolyticus* dataset, the VppAsia dataset, and the simulated datasets.

Pangenome analysis

To estimate the gene content, we annotated all *K. pneumoniae* genomes with prokka (v1.13)[69] software, using a curated version of the NC_012731.1 (NTUH_K2044) proteome as the primary annotation source. The Generic Feature Format (GFF) outputs generated by prokka were then used as the input files for panaroo (v1.2.8)[22] with the strict model. The "--remove-invalid-genes" and "--merge_paralogs" options were also employed during the analysis. Next, we extracted the non-redundant dataset from the overall results and defined 5,558 accessory genes as those present in between 2% and 98% of the genomes in the non-redundant dataset. Using the same approach (*e.g.*, Equation 2), we also calculated the gene loadings and defined genes with squared PC1 loadings greater than 300 as high-loading genes.

Calculation of Linkage Disequilibrium Decay

To analyze the decay of linkage disequilibrium (LD) in *K. pneumoniae*, we included 102,226 MAF-filtered SNPs and used the Haploview (v4.2)[70] software to compute R^2 values. When running Haploview, we set "-maxdistance 30" to ensure that the maximum marker distance for R^2 calculation did not exceed 30 kb. Additionally, we set "-minMAF 0" and "-hwcutoff 0" to exclude any filtering based on minor allele frequency and Hardy-Weinberg equilibrium testing, ensuring that all input SNPs were included in the calculations. The average r^2 value was calculated using a 10 bp step size.

Similarly, the same software and parameters were used for the simulated datasets to obtain the LD decay results for the simulated data.

We also calculated the LD between the high-loading variants defined above (340 variants in total, including SNPs and accessory genes) used Equation 3 in R. Each element R_{ij}^2 of the matrix quantifies the degree of linkage disequilibrium between variant *i* and variant *j*, providing a comprehensive view of the genomic associations among core SNPs and accessory genes. Based on manual inspection, we removed variants that represented fragmented versions of KP1_1982 (339 variants left), which is visualized in the heatmap Figure 1D , generated using the ComplexHeatmap package[71].

$$R^2 = \frac{\left(f_{gene_i,gene_j} - f_{gene_i} f_{gene_j}\right)^2}{f_{gene_i} f_{gene_i'} f_{gene_j} f_{gene_j'}} \qquad \text{(Equation 3)}$$

Based on the LD heatmap of all high-loading variants, we performed hierarchical clustering to divide the heatmap into 44 blocks. Among these, 4 larger blocks contain multiple variants, while the remaining 40 smaller blocks each contain only a few variants. The details are annotated in Figure 1D. Notably, block 3 was further subdivided into 3a, 3b, and 3c based on the physical locations of the mutations on the reference genome. We observed that some SNPs in block 3 are located within the KP1_RS30360 and KP1_RS30365 genes in the reference genome. According to the annotation file of the reference genome, KP1_RS30365 is classified as a pseudogene. However, we do not concur with this interpretation. We believe that KP1_RS30360 and KP1_RS30365 together represent a complete gene, with some strains (*e.g.*, MGH 78578) containing only a fragment of this gene. the differentiation at this locus also

likely represents a functional presence/absence rather than functional differentiation via SNPs.

As mentioned above, we converted the variants into binary, where 0 indicates that the core SNP is a major SNP and the accessory gene is absent in the strain, and 1 indicates that the core SNP is a minor SNP, and the accessory gene is present in the strain. We then generated heatmaps of the variant states according to the blocks, with rows following the same order as the LD heatmap and columns subjected to hierarchical clustering. The details are shown in the supplementary figure 4.

Functional Annotation

To further characterize the functional content in *K. pneumoniae*, we annotated both the MAF-filtered SNPs (MAF > 0.02) with SnpEff (v4.3t)[72] , using the reference genome NTUH-K2044 (NC_012731.1), and applied custom filtering to focus on coding region variants by disabling annotation for upstream, downstream, and UTR (untranslated) regions. We then obtained the types of the SNPs and determined whether they were synonymous mutations. For accessory genes, we used the eggNOG-mapper[73] online tool with its default settings to ensure a comprehensive and standardized annotation.

For the identified high-loading variants (squared PC1 SNP loading > 300), we further mapped these variants to the reference genomes of *K. pneumoniae* strains MGH 78578 (NC_009648.1) using BLASTn[74]. The mapping aimed to identify their genomic locations and potential functional implications in these strains. Additionally, we searched the UniProt[75] database to retrieve gene names and associated functional annotations for the high-loading variants.

Half-matching testing

To assess the stability and repeatability of the first principal component (PC1) in *K. pneumoniae*, we used three half-matching tests, each designed to evaluate the consistency of PC1 under varying conditions.

1. Half-Strain Matching

Firstly, we divided the non-redundant *K. pneumoniae* dataset into two groups based on the backbone in the phylogenetic tree (right-hand side and left-hand side). We then performed PCA on the first group using the pruned SNPs. Afterward, we calculated the MAF-filtered SNP loadings using the above steps. Subsequently, we obtained the product for the second group through matrix multiplication, representing the mapping of the PCA results from the first group onto the second group. We then calculated the correlation for the PC1 between the mapped results of the first group and the PCA results of the second group. Conversely, the same matching analysis was performed on the data from the second group.

2. Half- Segment Matching

K. pneumoniae has a circular chromosome with a length of approximately 5 to 6 Mb, and the reference genome we used is approximately 5.2 Mb in length. We divided the circular chromosome into two DNA fragments (Fragment 1 (1 bp–2.6 Mb) and Fragment 2 (2.6 Mb–5.2 Mb)), each containing approximately half of the chromosome. Specifically, the first dataset comprised the pruned SNPs in the first half of the chromosome, while the second dataset encompassed the unlinked SNPs from the second half. We separately performed PCA on these two datasets using PLINK. Finally, we calculated the correlation between the PC1 of the two fragments.

3. Partial-Variants Matching

First, we assigned maximum loading values to the MAF-filtered SNPs based on their squared

PC1 loadings, using a 50 kb window size with a 1 bp step size. This process identified the maximum squared PC1 loadings within each 50 kb vicinity. Based on the distribution of maximum squared PC1 loadings, the SNPs were categorized into three types: low-loading SNPs, medium-loading SNPs, and high-loading SNPs (*e.g.*, Supplementary Figure 3G). We separately conducted PCA on the high-loading SNP dataset and low-loading SNP dataset, by extracting unlinked SNPs from each. Subsequently, we calculated the correlation between the PC1 of the two datasets. In this test, we ignored SNPs with intermediate loading values.

Similarly, the three types of half-matching tests were applied to the non-redundant *V. parahaemolyticus* dataset (Supplementary Figure 6), the VppAsia dataset (Supplementary Figure 6), and a non-redundant clonal population simulation (Simulation 1, SNP distance > 2 kb, shown in Supplementary Figure 4A). However, in the half-strain matching test, although we still performed splits along the phylogenetic tree, these datasets were not always divided into left-hand and right-hand subsets in the trees. Therefore, in these datasets, we referred to the two subsets as half samples 1 and half samples 2. Additionally, in *V. parahaemolyticus*, which inherently possesses two chromosomes of different sizes, we directly divided the data into two parts based on chromosome numbers.

Simulation of bacterial populations

We used SLiM (v4.0.1)[76] to simulate the evolutionary dynamics of bacterial populations under different scenarios using a non-Wright–Fisher (nonWF) framework[77]. All scenarios started with identical initial conditions: each bacterial genome consisted of a single 1 Mb haploid chromosome, with a mutation rate of 10□ □ per site per generation. Horizontal gene transfer (HGT) occurred at a rate of 1.0, ensuring that every individual experienced HGT in each generation. The simulations ran for 10,000 generations. Two mutation types, "m1" and "m2", occurred at each site with equal probability (1:1) and were assumed to be neutral. However, in the final simulation, we adjusted the chromosome length to closely match the length of a real *K. pneumoniae* genome.

1. Modeling clonal haploid bacteria with horizontal gene transfer

In the first simulation scenario, we modeled the evolutionary dynamics of a single bacterial population, focusing on the spread of a single type of neutral mutation and the genetic exchange within population via HGT. The HGT tract length was set to 100 bp, 1 kb, 10 kb, or 100 kb, representing a range of recombination sizes to accurately capture genetic variation within the population. Each bacterium produced offspring according to a Poisson distribution with $\lambda = 2$, meaning that, on average, each bacterium produced two offspring per generation. However, the number of offspring per bacterium in any given generation could vary, with some bacteria producing more and others producing fewer, reflecting the probabilistic nature of the Poisson distribution. We conducted three trials with population sizes of 1,000, 2,000, and 10,000.

2. A linear stepping-stone model

In the second simulation scenario, we constructed a linear stepping-stone model with 10 subpopulations of clonal haploid bacteria, where each subpopulation was connected by migration only to its nearest neighbors in the chain. The average recombination tract length within each subpopulation was set to 10 kb. Migration occurred with varying numbers of migrants—1, 10, 20, or 50 individuals—moving to neighboring subpopulations each generation to simulate gene flow driven by geographical factor or other isolating barriers. Each individual produced offspring according to a Poisson distribution with $\lambda = 2$. We conducted three trials with total population sizes of 1,000, 2,000, and 10,000, corresponding to 100, 200, and 1,000 individuals per subpopulation, respectively.

3. Selective model

In the third simulation scenario, we focused on a population of 10,000 clonal haploid bacteria, introducing selective pressure into the model. Horizontal gene transfer (HGT) tract lengths were set to 1 kb, 10 kb, and 100 kb. Each individual was assigned a score, termed a quantitative trait value, based on the difference in the number of "m1" and "m2" ($\text{score} = \sum \text{m1} - \sum \text{m2}$) and ranked accordingly. In this model, the coefficient for any locus on the chromosome was 1. Depending on the mutation type, if it was m1, the value at that locus was 1; if it was m2, the value at that locus became -1. From the top and bottom 160 individuals, 80 were selected to have increased offspring production, following a Poisson distribution with $\lambda = 3$, while the rest reproduced with $\lambda = 2$. The $\lambda = 3$ Poisson distribution meant that these selected individuals, on average, produced 3 offspring per generation, resulting in higher offspring production compared to the rest of the bacterial strains. This represents a simple selective pressure model.

4. Refined selective model

In the fourth simulation scenario, we used a selective model similar to the third scenario, with the addition of locus-specific selection coefficients. We created a dataset following a gamma distribution to serve as the selection coefficients, with the size of the dataset corresponding to the simulated chromosome length. The scale parameter in the gamma distribution controls the magnitude of the selection coefficients: larger scale values result in fewer loci having larger selection coefficients, while the majority of loci have very low selection coefficients (close to 0). Since each locus had a specific selection coefficient, the assignment of values for each simulated strain was modified accordingly. The quantitative trait value was calculated as follows:

$$\text{score} = \sum_{i=1}^{N_{m1}} (m1_i \cdot \text{coefficient}_i) - \sum_{j=1}^{N_{m2}} \left(m2_j \cdot \text{coefficient}_j\right) \qquad \text{(Equation 4)}$$

In this formula, N_{m1} and N_{m2} represent the number of m1 and m2 mutations in the simulation, respectively. This calculation differs from the previous simulations, where the trait value was simply the difference between the numbers of "m1" and "m2" mutations. We first used the previous parameters (*e.g.*, Chromosome = 1 Mb, HGT = 10 kb, and population size = 10,000) and tested three sets of experiments with a gamma-distributed dataset, where the mean value was set to 1. The scale parameters for these experiments were set to 1×10^2, 1×10^4, and 1×10^6 (with the final scale corresponding to the simulated chromosome length). From the top and bottom 300 individuals, 150 were selected to have increased offspring production, following a Poisson distribution with $\lambda = 3$, while the rest reproduced with $\lambda = 2$. At 10,000 generations, we obtained reliable results, and the extremity of the SNP loading values was found to be associated with SNP coefficients.

To better approximate the real K. pneumoniae genome, we adjusted the chromosome length to 5 Mb and increased the HGT tract length, setting the length of HGT between each bacterium and others in the population to 100 kb per generation. Each locus on the chromosome was assigned a selection coefficient drawn from a gamma distribution, which included five million values, one for each locus on the chromosome, regardless of whether a mutation occurred at that locus. The scale for the gamma distribution was set to 5×10^6 for this simulation. Similarly, the mean value of the gamma distribution was set to 1 (shape of gamma distribution = 2×10^{-7}). From the top and bottom 600 individuals, 300 from each group were selected to have increased offspring production, following a Poisson distribution with $\lambda = 3$, while the remaining individuals reproduced with $\lambda = 2$.

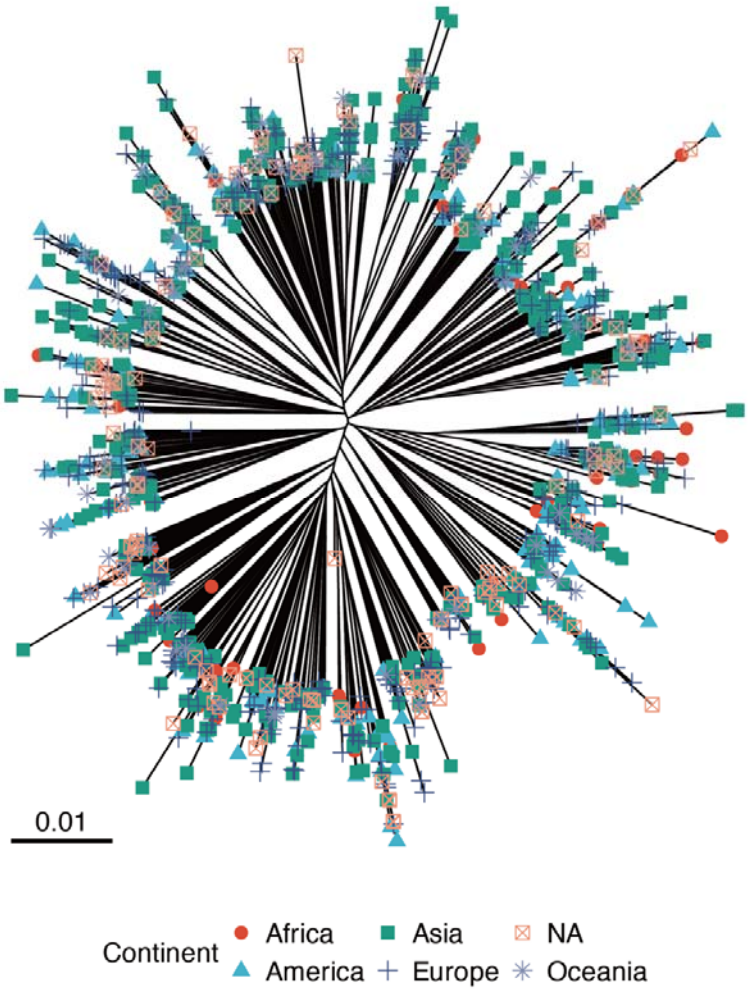
Compared to the Selective Model, the Refined Selective Model introduces locus-specific selection coefficients, better simulating the different selective pressures at each locus, and thus more closely approximating the complexity of real-world genomes.

All the scripts can be found at https://github.com/sql647/KpBACKBONE.

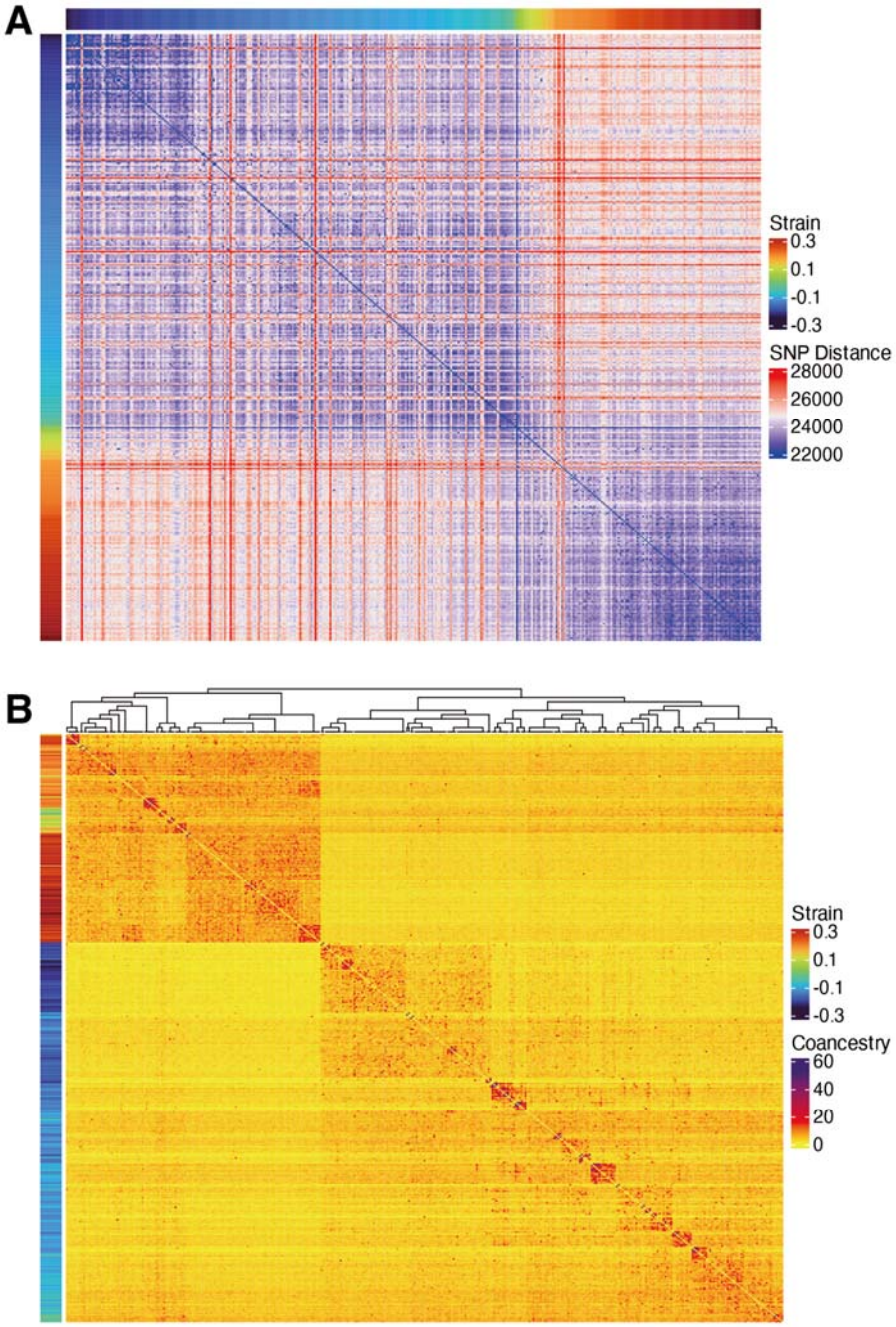
Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) grant 32350710791 awarded to D.F.

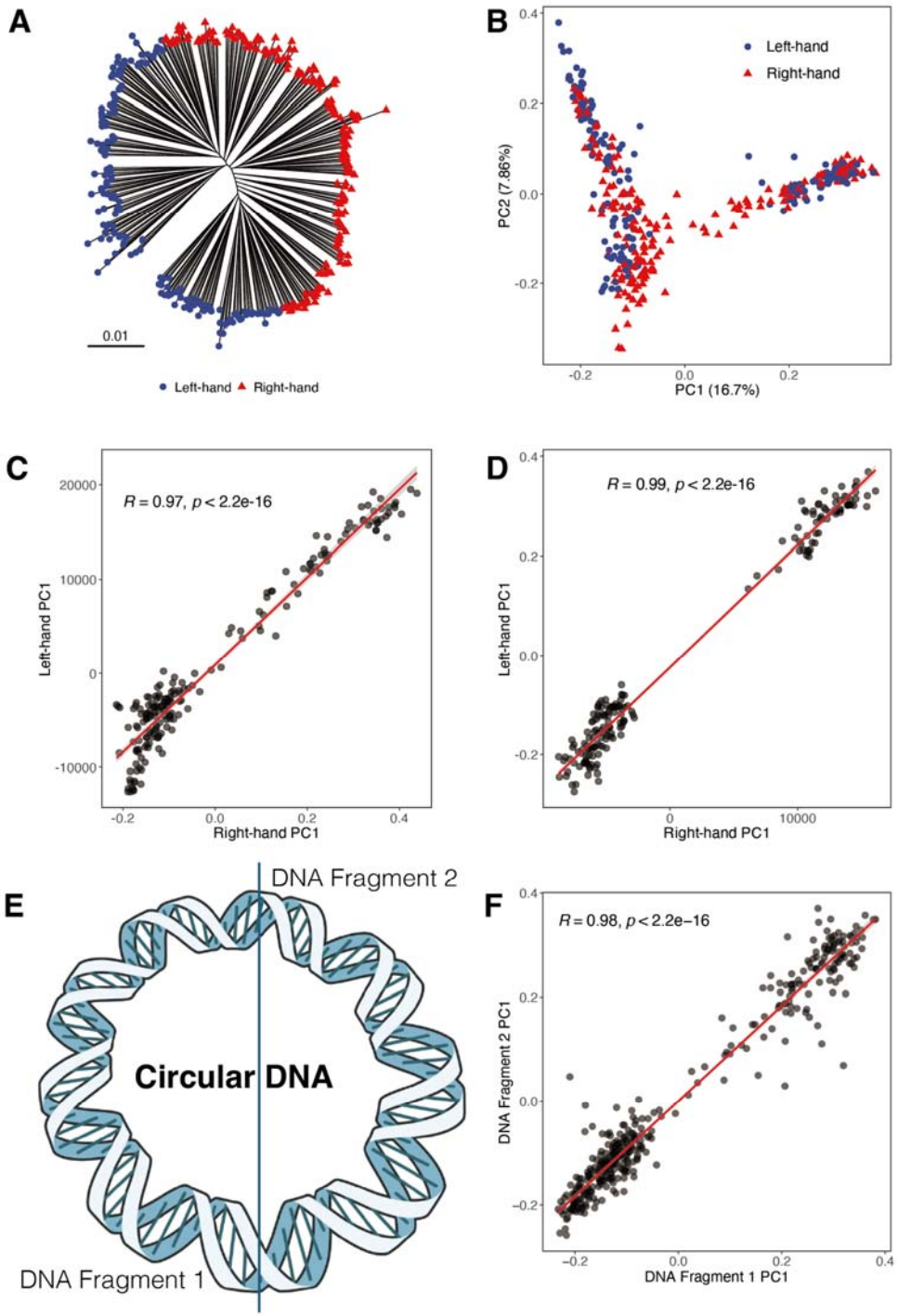
Supplementary Figures



Supplementary Figure 1. Phylogenetic tree of the complete *Klebsiella pneumoniae* dataset. A maximum-likelihood phylogenetic tree constructed based on core SNPs from the full *K. pneumoniae* dataset, with branches color-coded by continental origin to illustrate the global distribution of isolates.



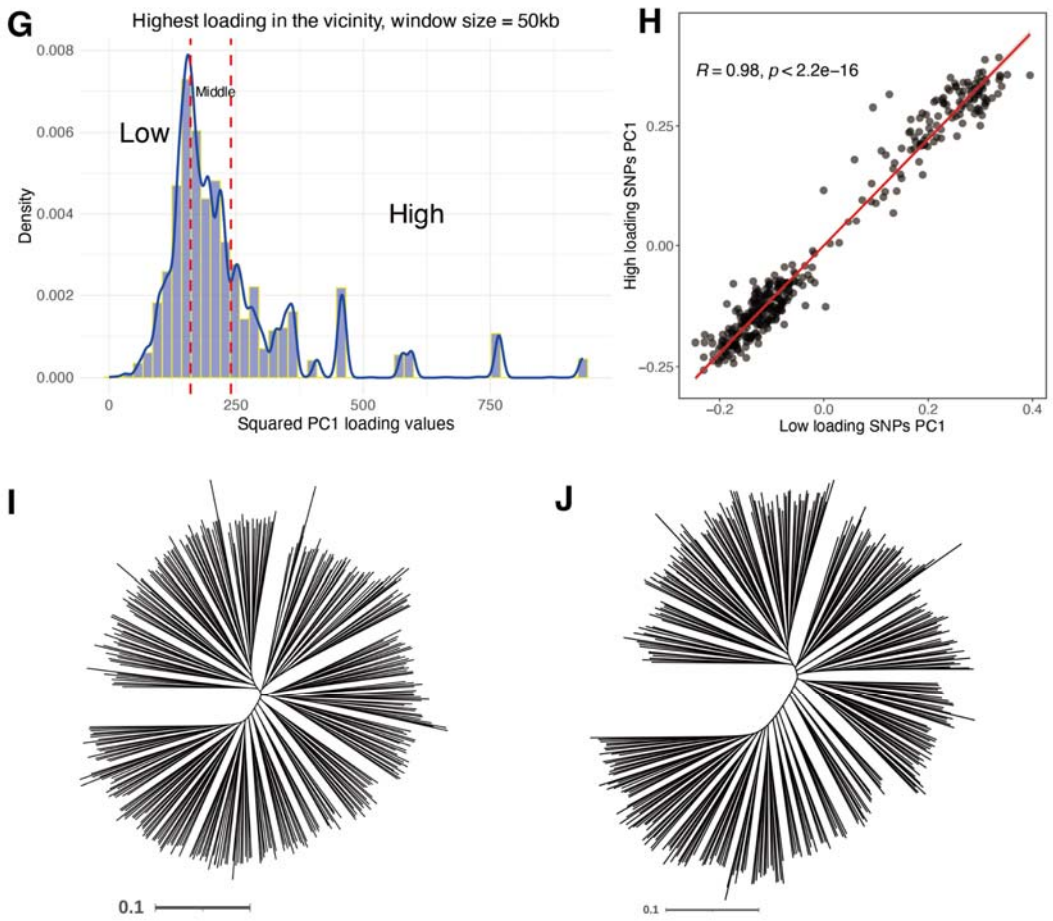
Supplementary Figure 2. Genomic characteristics of the non-redundant *Klebsiella pneumoniae* dataset. (A) SNP distance matrix illustrating the genetic distances between strains in the non-redundant dataset, with strains ordered based on their PC1 values. The minimum SNP distance observed is 20,000 SNPs. **(B)** FineSTRUCTURE analysis of the non-redundant KP dataset, highlighting population structure and clustering patterns. The annotation bar indicates the PC1 values of the strains.



Supplementary Figure 3. Three types of half-matching tests for the non-redundant *Klebsiella pneumoniae* dataset.

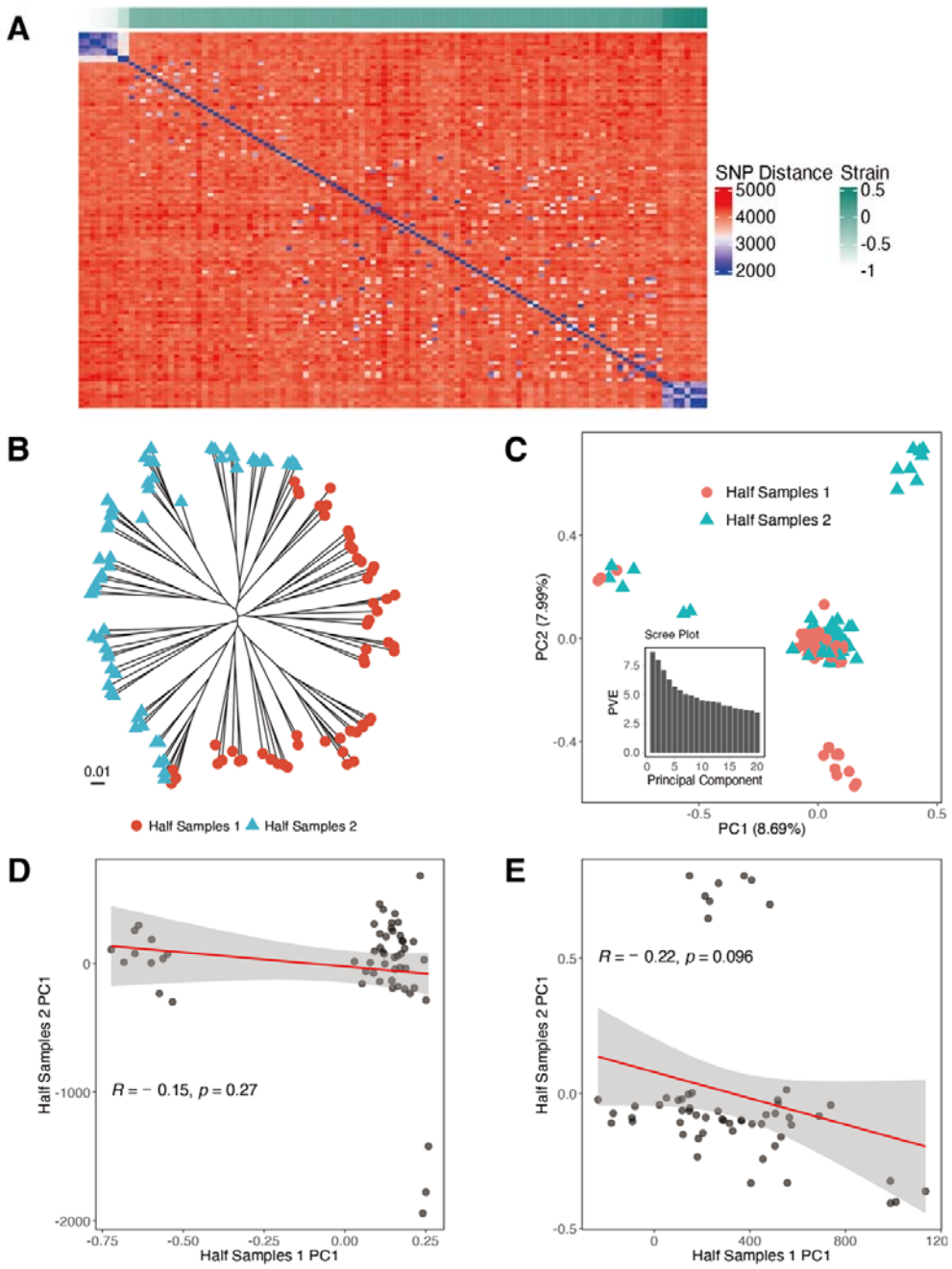
(A–D) Half-Strain Matching: (A) Phylogenetic tree of the non-redundant *K. pneumoniae* dataset divided into two subsets based on the backbone. **(B)** PCA of the non-redundant *K. pneumoniae* dataset divided into two groups. **(C)** Correlation of PC1 from right-hand group mapped onto left-hand group results. **(D)** Correlation of PC1 from left-hand group mapped onto right-hand group results.

(E–H) Half-Segment Matching: (E) The genome is divided into two fragments: Fragment 1 (1 bp–2.6 Mb) and Fragment 2 (2.6 Mb–5.2 Mb). **(H)** Correlation of PC1 between the two fragments.



Supplementary Figure 3. Three types of half-matching tests for the non-redundant *Klebsiella pneumoniae* dataset.

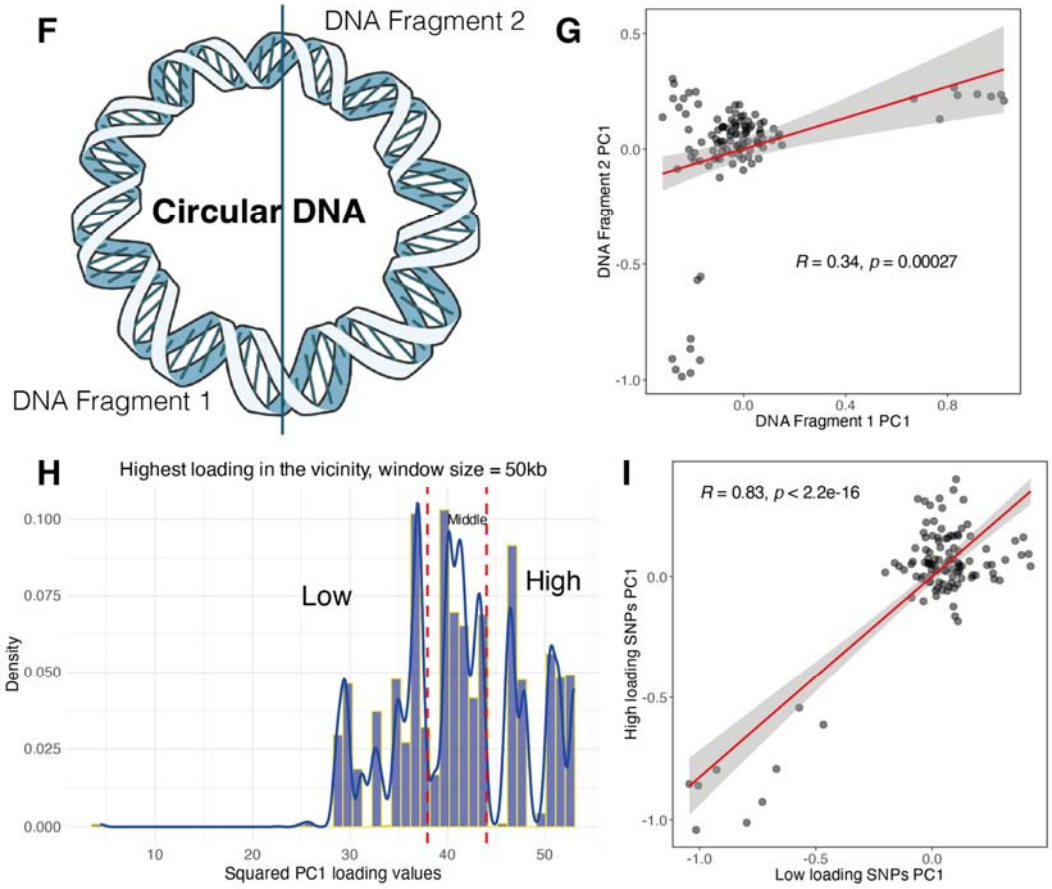
(G–J) Partial-Variants Matching: (G) Histogram of SNP loadings on PC1, dividing SNPs into high- and low-loading groups. **(H)** Correlation of PC1 between high- and low-loading SNP groups. **(I)** Phylogenetic tree based on low-loading SNPs. **(J)** Phylogenetic tree based on high-loading SNPs.



Supplementary Figure 4. Three types of half-matching tests for the clonal simulation.

(A) SNP distance matrix for the clonal simulation. The minimum SNP distance observed is 2,000 SNPs.

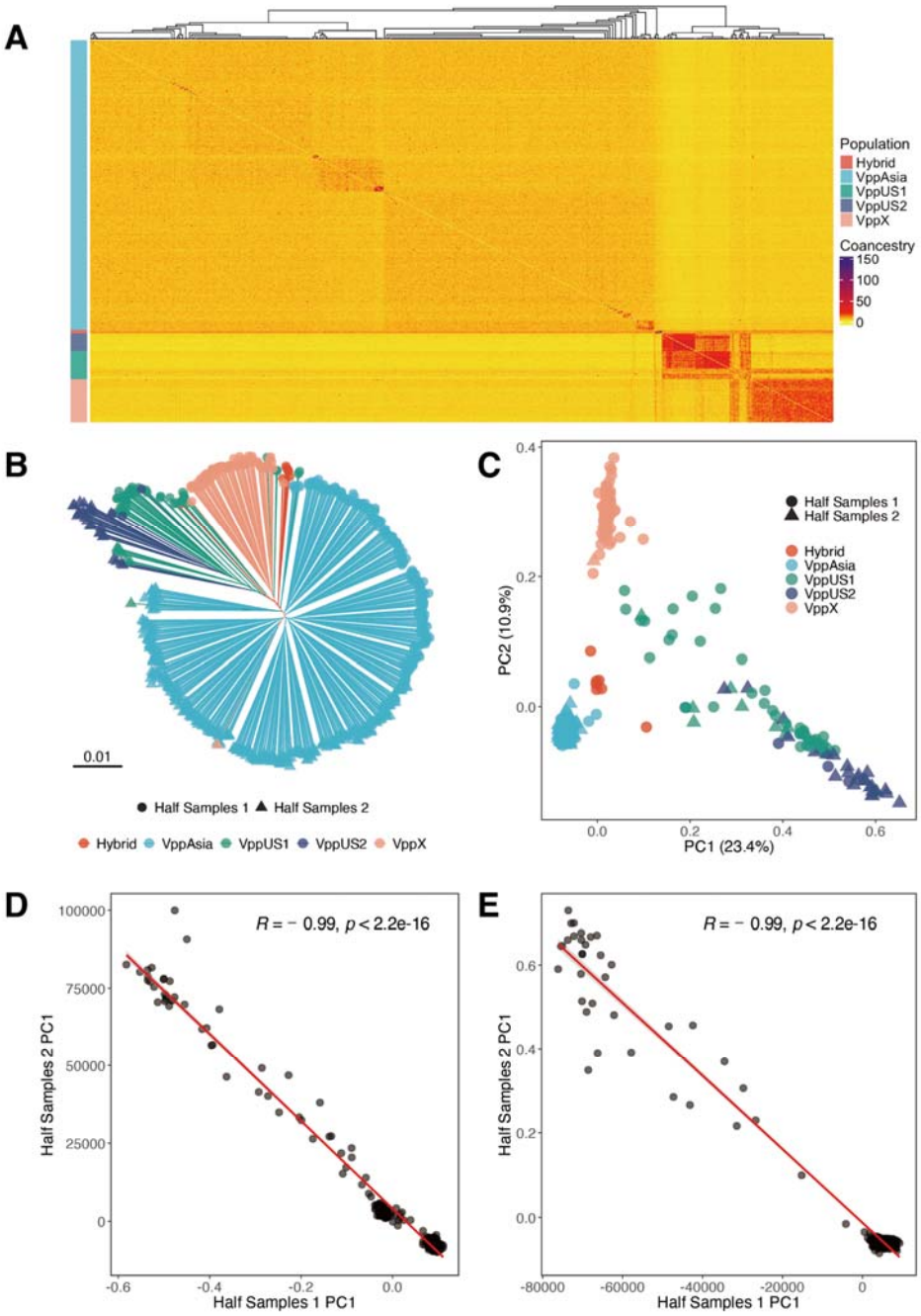
(B–E) Half-Strain Matching: (B) Phylogenetic tree of the non-redundant clonal simulation divided into two subsets. **(C)** PCA of the non-redundant clonal simulation divided into two groups. **(D)** Correlation of PC1 from group 2 mapped onto group 1 results. **(E)** Correlation of PC1 from group 1 mapped onto group 2 results.



Supplementary Figure 4. Three types of half-matching tests for the clonal simulation.

(F–G) Half-Segment Matching: (F) Division of the genome into two fragments (Fragment 1 and Fragment 2). Adapted from "DNA (circular, large)", by BioRender.com (2024). **(G)** Correlation of PC1 between the two fragments.

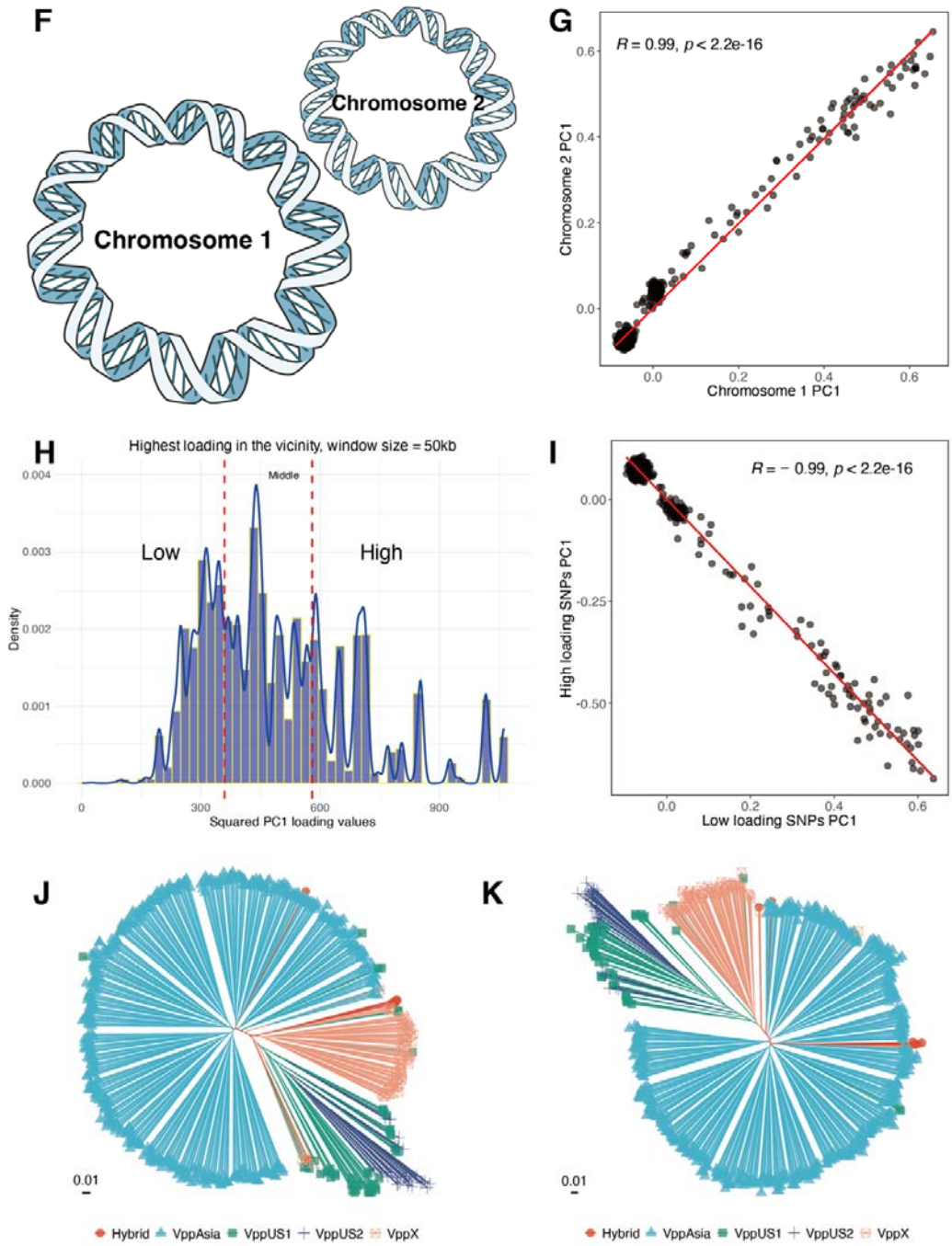
(H–I) Partial-Variants Matching: (H) Histogram of SNP loadings on PC1, dividing SNPs into high- and low-loading groups. **(I)** Correlation of PC1 between high- and low-loading SNP groups.



Supplementary Figure 5. Three types of half-matching tests for the non-redundant *Vibrio parahaemolyticus* dataset.

(A) FineSTRUCTURE analysis assigning populations for the non-redundant *V. parahaemolyticus* dataset.

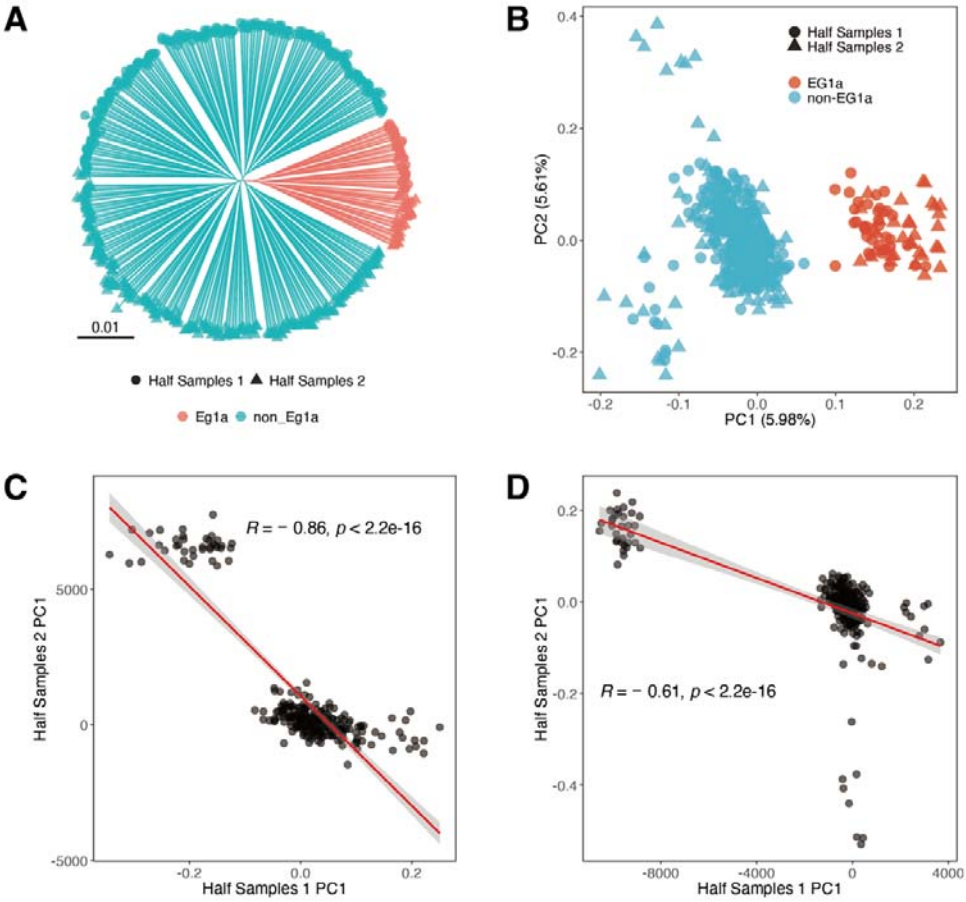
(B–E) Half-Strain Matching: (B) Phylogenetic tree of the non-redundant *V. parahaemolyticus* dataset divided into two subsets, with different shapes representing the groups and colors indicating different populations. **(C)** PCA of the non-redundant *V. parahaemolyticus dataset*, divided into two groups, with colors and point shapes consistent with panel A. **(D)** Correlation of PC1 from group 2 mapped onto group 1 results. **(E)** Correlation of PC1 from group 1 mapped onto group 2 results.



Supplementary Figure 5. Three types of half-matching tests for the non-redundant *Vibrio parahaemolyticus* dataset.

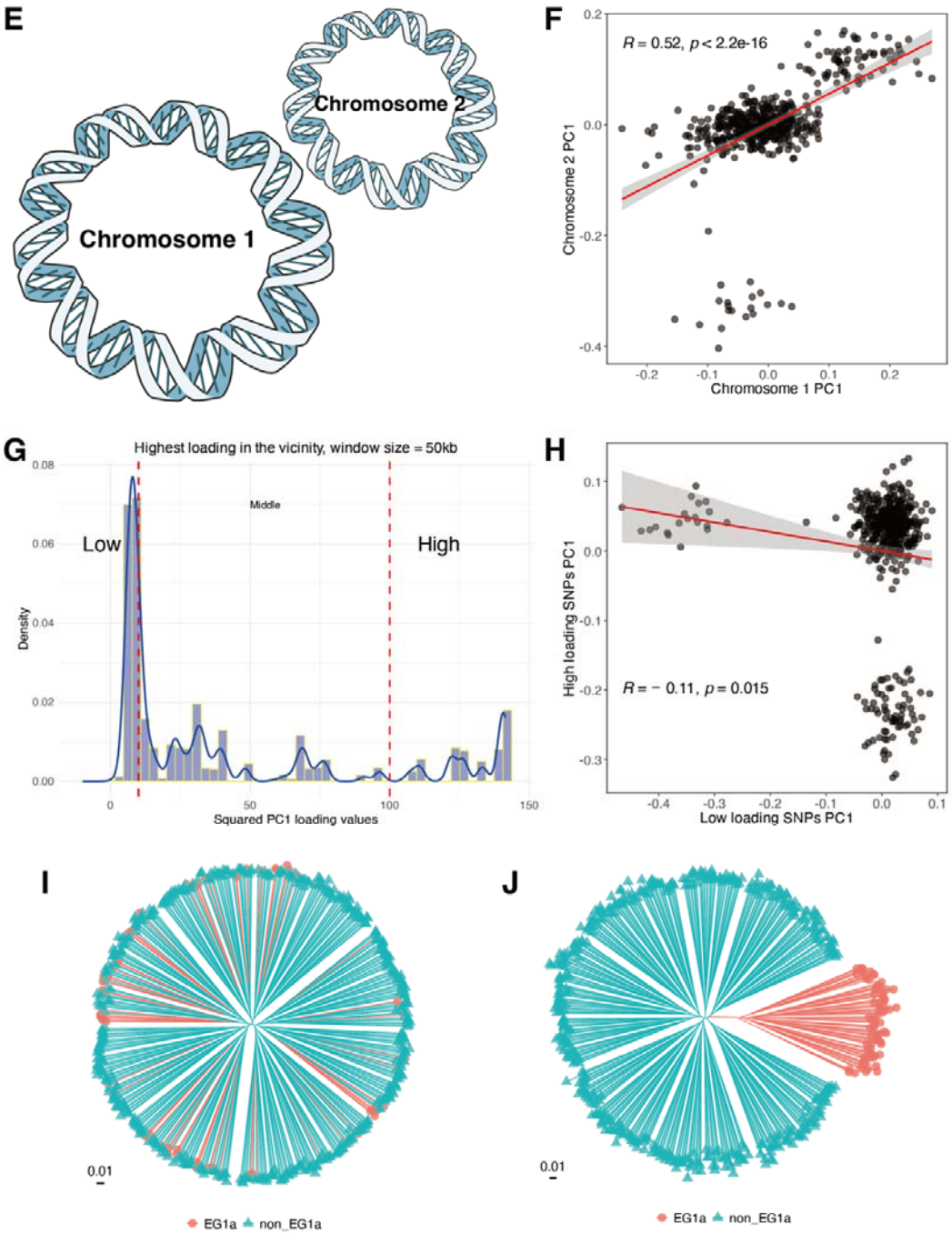
(F–G) Half-Segment Matching: (F) Representation of V. parahaemolyticus Chromosome 1 and Chromosome 2, with proportions reflecting their actual sizes. Adapted from "DNA (circular, large)", by BioRender.com (2024). **(G)** Correlation of PC1 between the two chromosomes.

(H–K) Partial-Variants Matching: (H) Histogram of SNP loadings on PC1, dividing SNPs into high- and low-loading groups. **(I)** Correlation of PC1 between high- and low-loading SNP groups. **(J)** Phylogenetic tree based on low-loading SNPs. **(K)** Phylogenetic tree based on high-loading SNPs.



Supplementary Figure 6. Three types of half-matching tests for the VppAsia dataset.

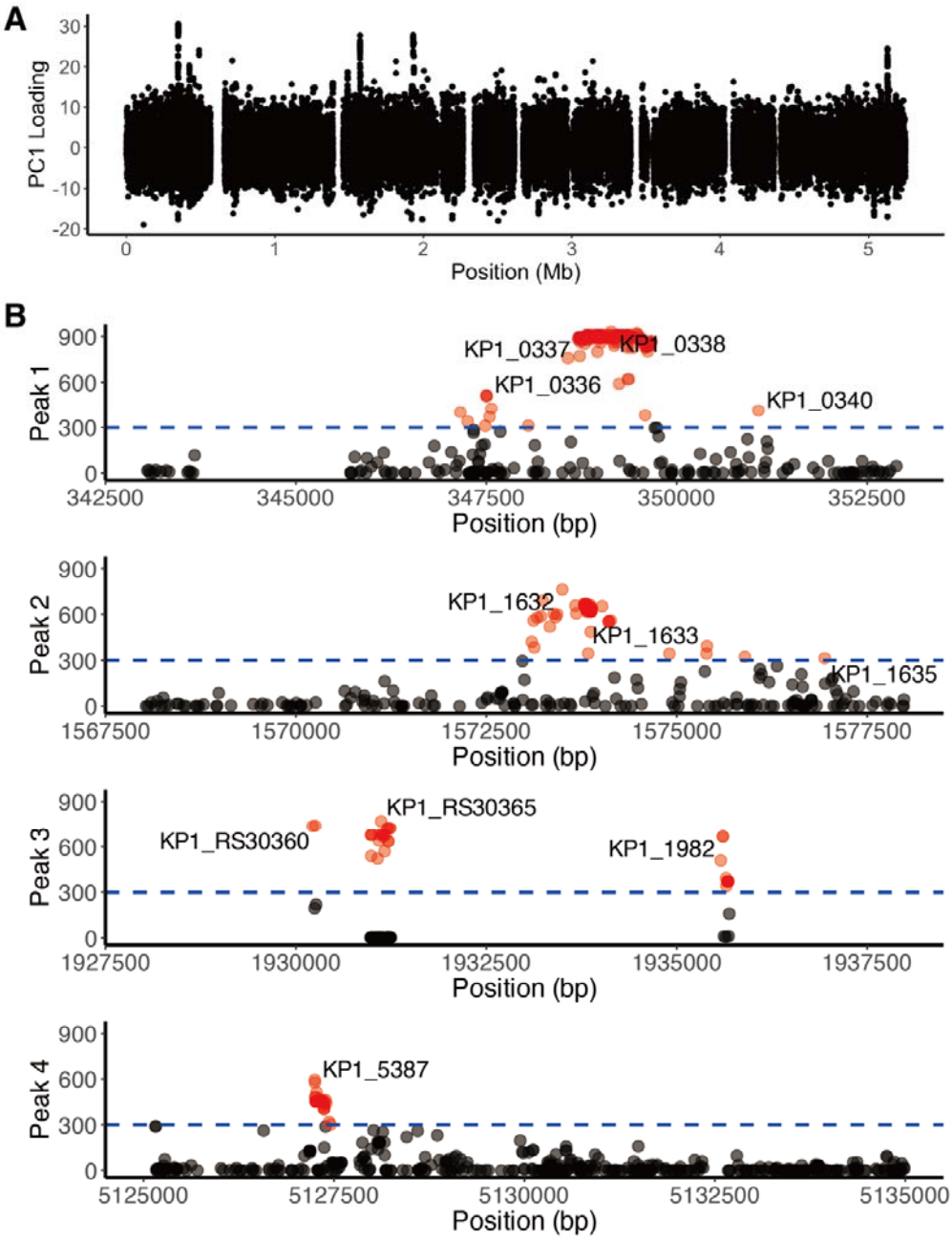
(A–D) Half-Strain Matching: (A) Phylogenetic tree of the VppAsia dataset divided into two subsets, with different shapes representing the groups and colors indicating different ecospecies. **(B)** PCA of the non-redundant V. parahaemolyticus Asia dataset, divided into two groups, with colors and point shapes consistent with panel A. **(C)** Correlation of PC1 from group 2 mapped onto group 1 results. **(D)** Correlation of PC1 from group 1 mapped onto group 2 results.



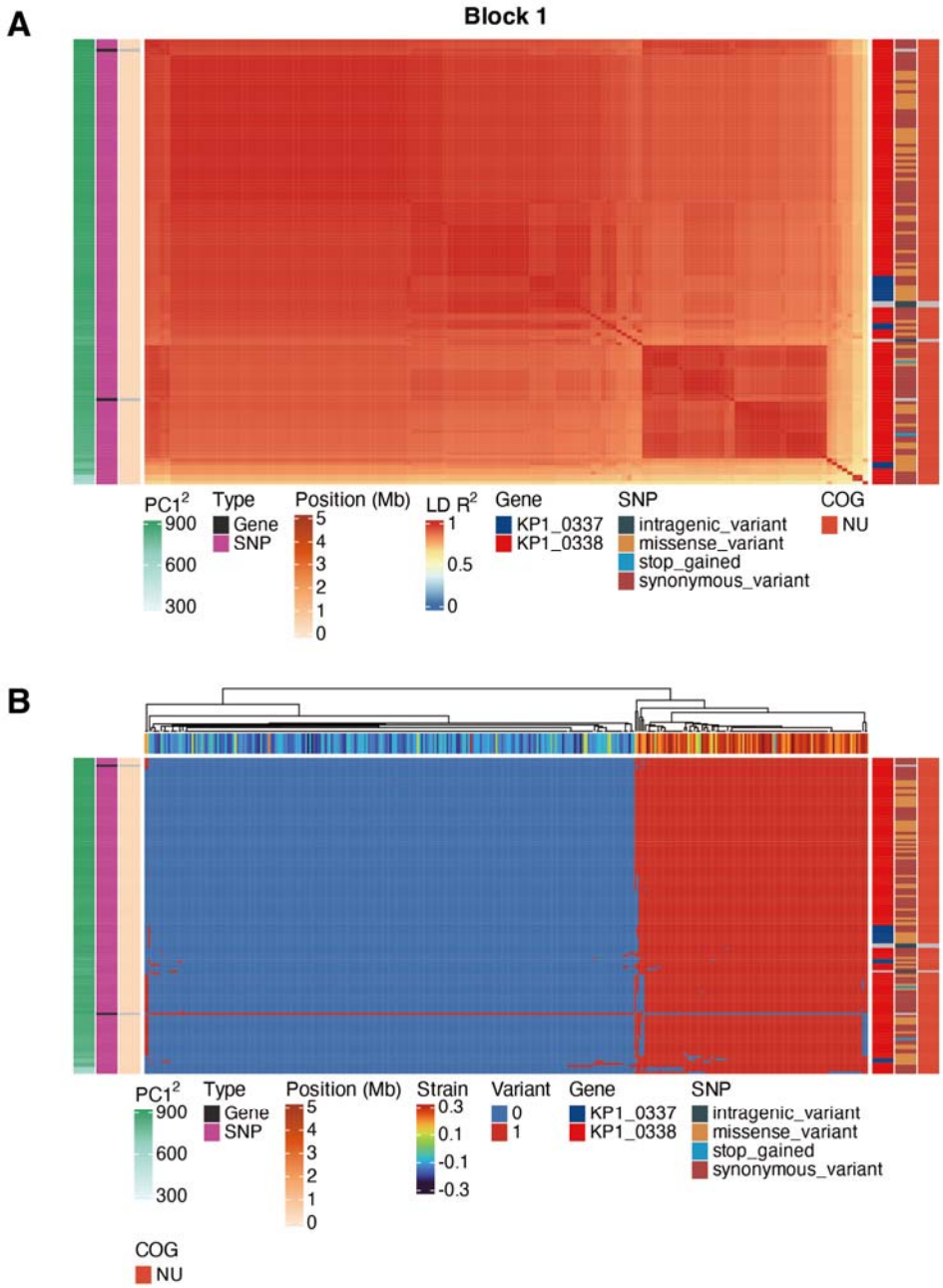
Supplementary Figure 6. Three types of half-matching tests for the VppAsia dataset.

(E–H) Half-Segment Matching: (E) Representation of V. parahaemolyticus Chromosome 1 and Chromosome 2, with proportions reflecting their actual sizes. Adapted from "DNA (circular, large)", by BioRender.com (2024). **(H)** Correlation of PC1 between the two chromosomes.

(G–H) Partial-Variants Matching: (G) Histogram of SNP loadings on PC1, dividing SNPs into high- and low-loading groups. **(H)** Correlation of PC1 between high- and low-loading SNP groups. **(I)** Phylogenetic tree based on low-loading SNPs. **(J)** Phylogenetic tree based on high-loading SNPs.

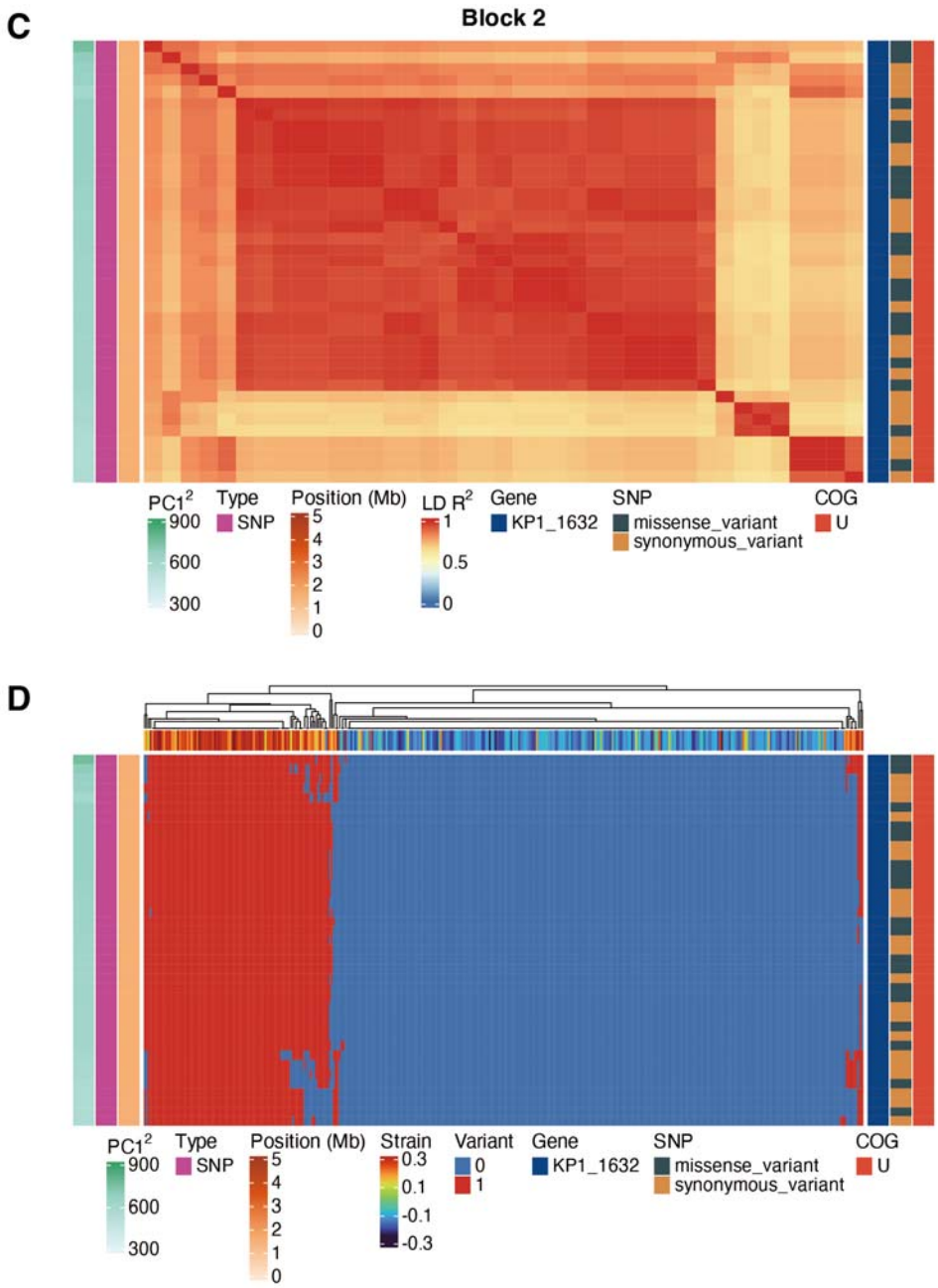


Supplementary Figure 7. Details of PC1 SNP loadings across the genome. (A) Manhattan plots showing SNP loading values for the original PC1 across the genome. **(B)** Zoomed-in views of four high-loading SNP peaks identified across the genome.



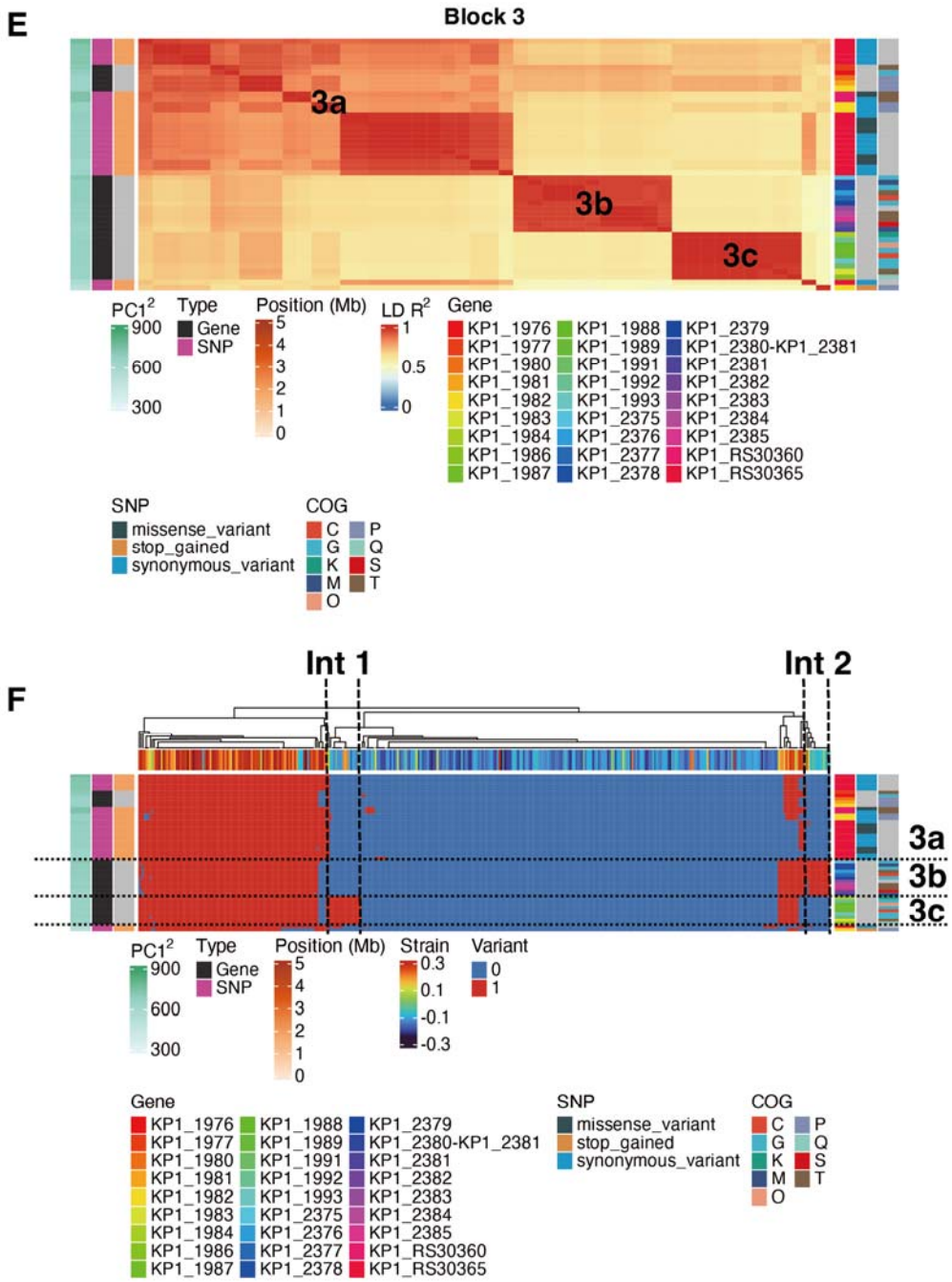
Supplementary Figure 8: Overview of blocks 1–4 from LD analysis with variant status heatmaps.

Block 1 (A, B): (A) SNPs in KP1_0337, KP1_0338, and associated accessory genes form a high LD block. **(B)** Variant (1/0) heatmap shows clear differentiation in gene presence/absence or SNP minor/major alleles across isolates.



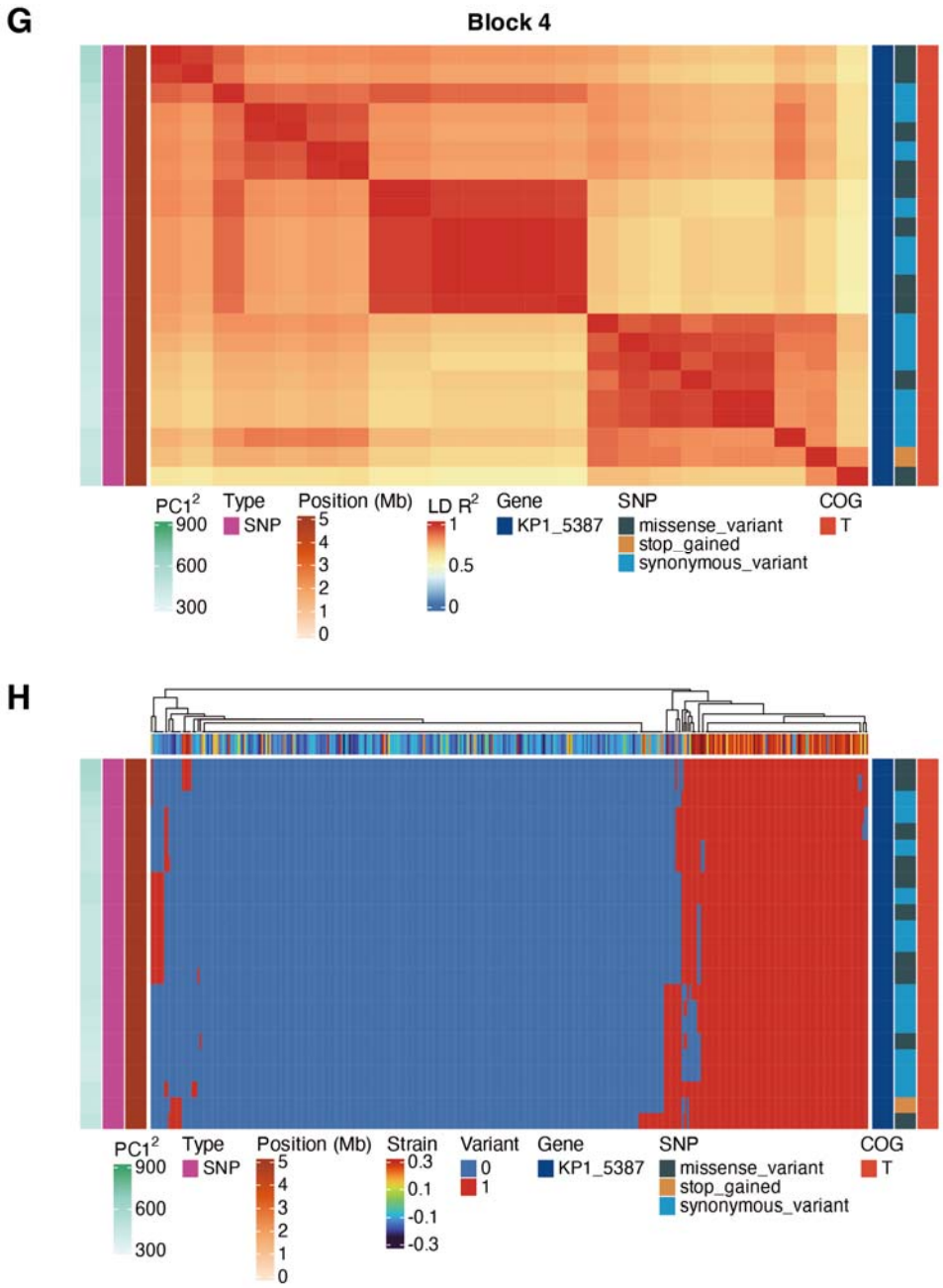
Supplementary Figure 8: Overview of blocks 1–4 from LD analysis with variant status heatmaps.

Block 2 (C, D): (C) High LD observed for SNPs in KP1_1632. **(D)** Variant (1/0) heatmap highlights distinct patterns of gene presence/absence or SNP minor/major alleles across isolates.



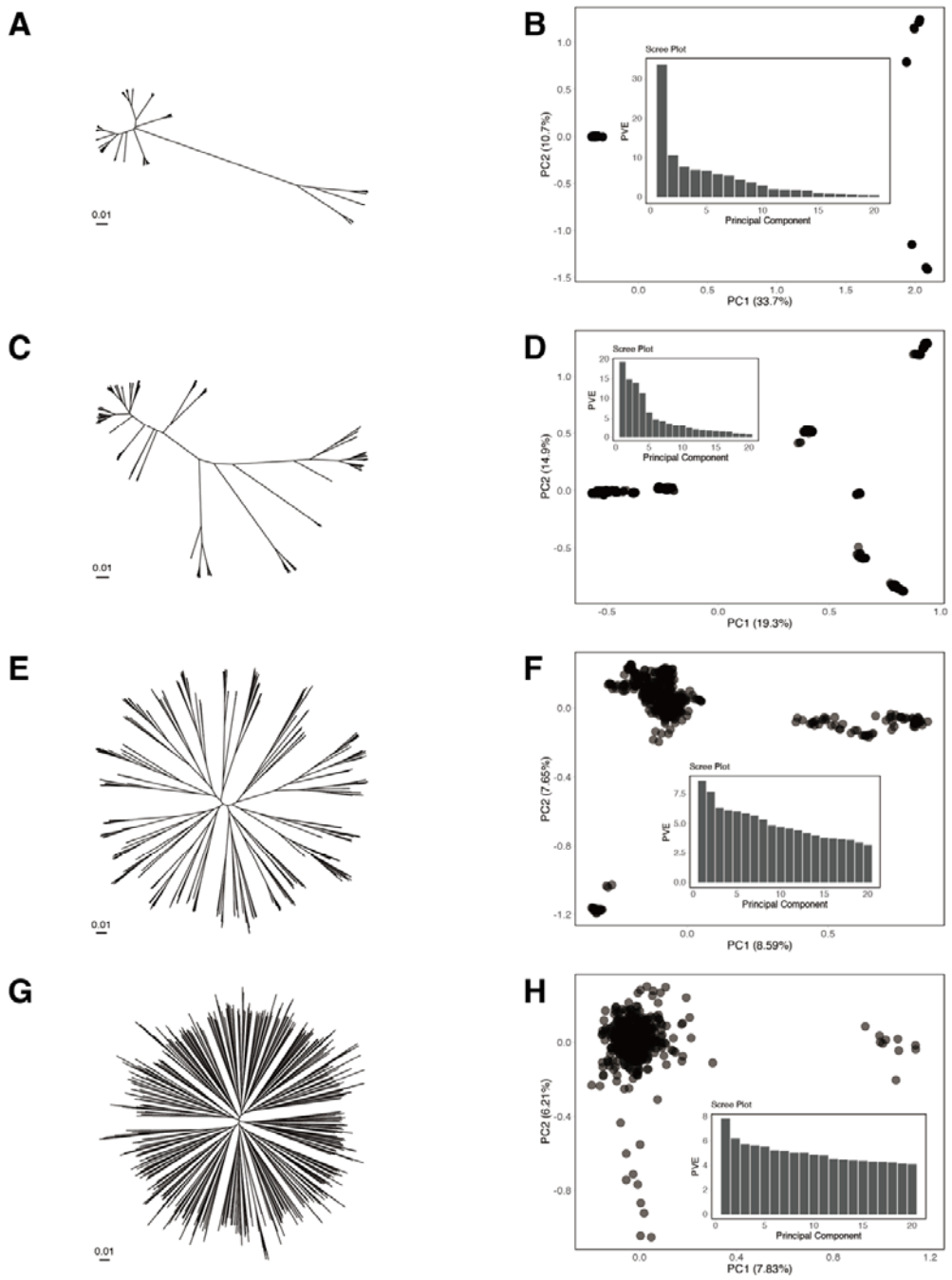
Supplementary Figure 8: Overview of blocks 1–4 from LD analysis with variant status heatmaps.

Block 3 (E, F): (E) SNPs in KP1_RS09345, KP1_RS30360, KP1_1982, and accessory genes form tightly linked clusters. **(F)** Variant (1/0) heatmap reveals clear differentiation in gene presence/absence or SNP minor/major alleles across isolates. Int 1 represents an intermediate case that includes Block 3b but excludes 3c, while Int 2 represents an intermediate case that includes 3c but excludes 3b.

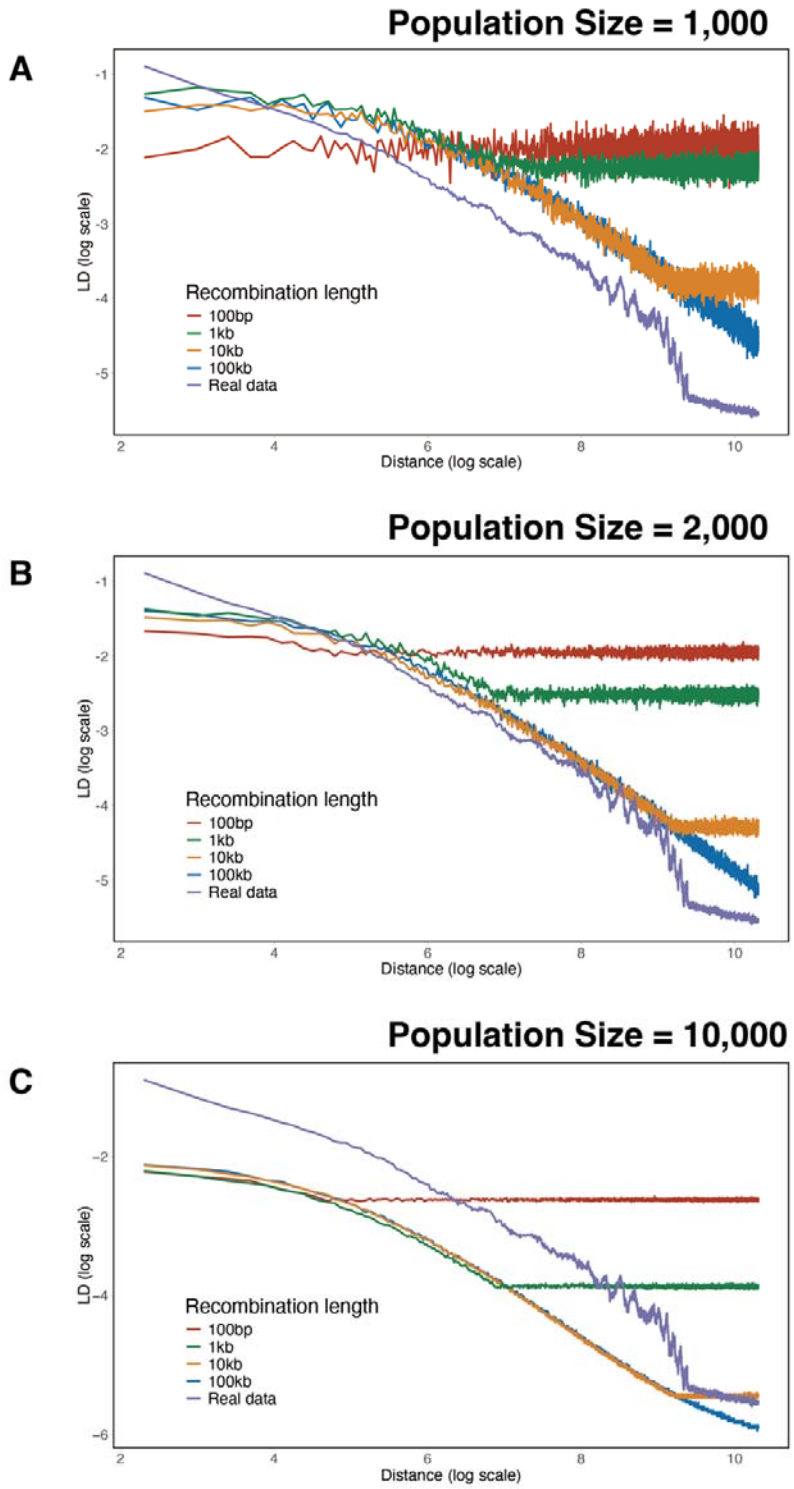


Supplementary Figure 8: Overview of blocks 1–4 from LD analysis with variant status heatmaps.

Block 4 (G, H): (G) High LD observed for variants in KP1_5387, forming a small, highly linked cluster. **(H)** Variant (1/0) heatmap shows distinct variation in gene presence/absence or SNP minor/major alleles across isolates.

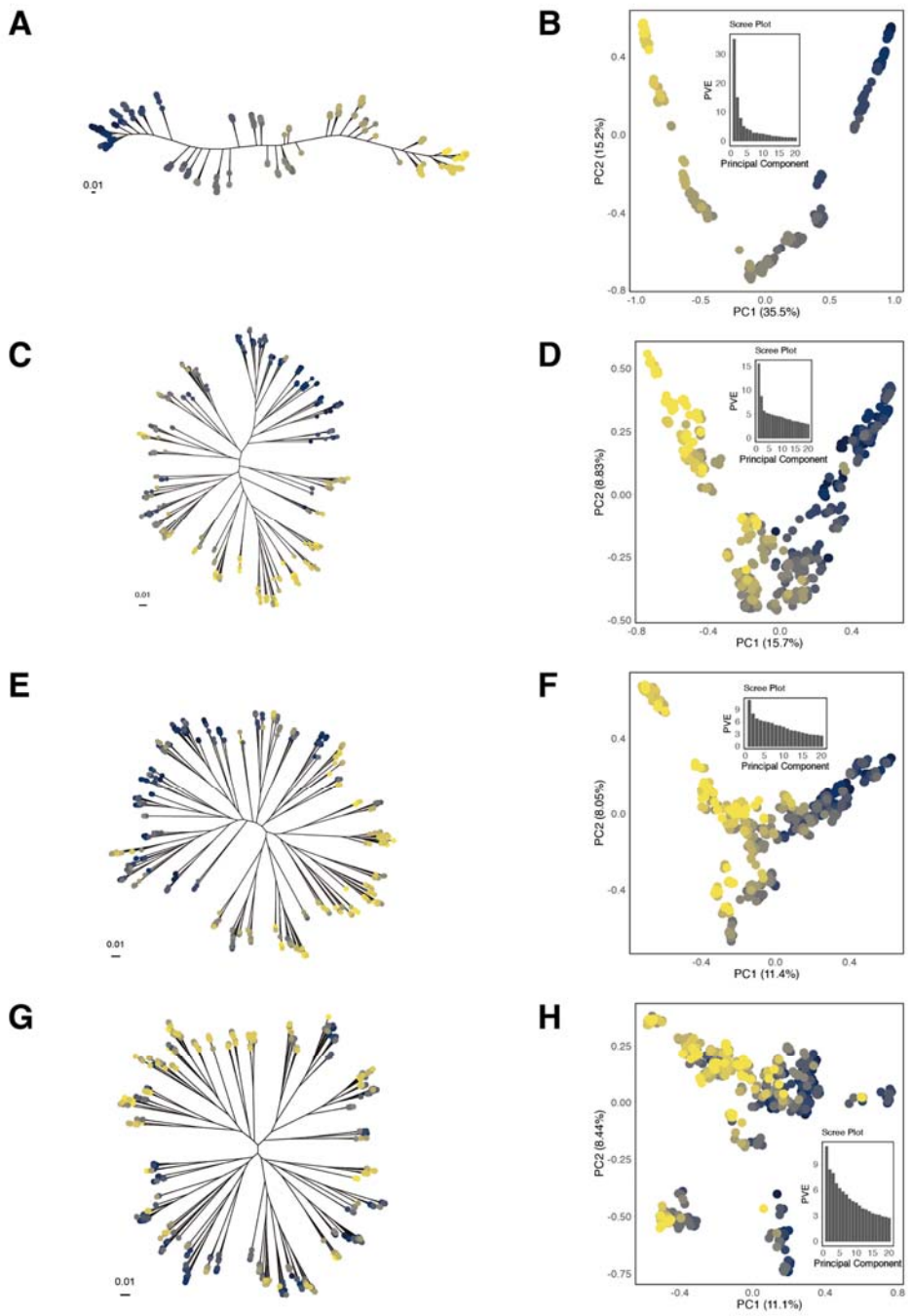


Supplementary Figure 9. Clonal haploid bacteria model with varying horizontal gene transfer (HGT) lengths and a population size of 2000. Phylogenetic trees, principal component analysis (PCA) plots, and scree plots for a population of 2000 individuals under different HGT lengths. Panels (A–H) show results for: **(A-B)** 100 bp, **(C-D)** 1 kb, **(E-F)** 10 kb, **(G-H)** 100 kb.

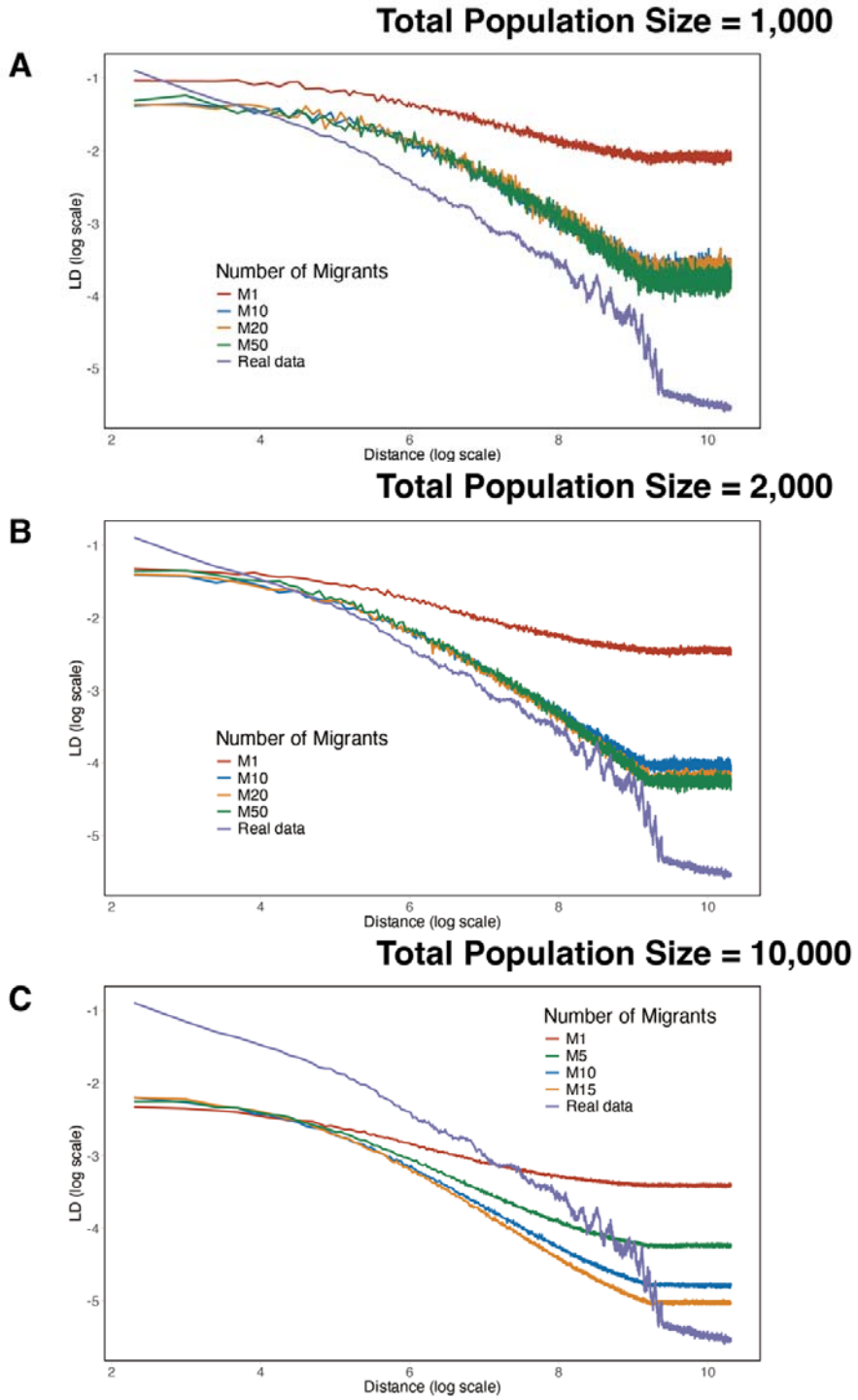


Supplementary Figure 10. Linkage disequilibrium (LD) decay under different horizontal gene transfer (HGT) lengths in clonal haploid bacteria models.

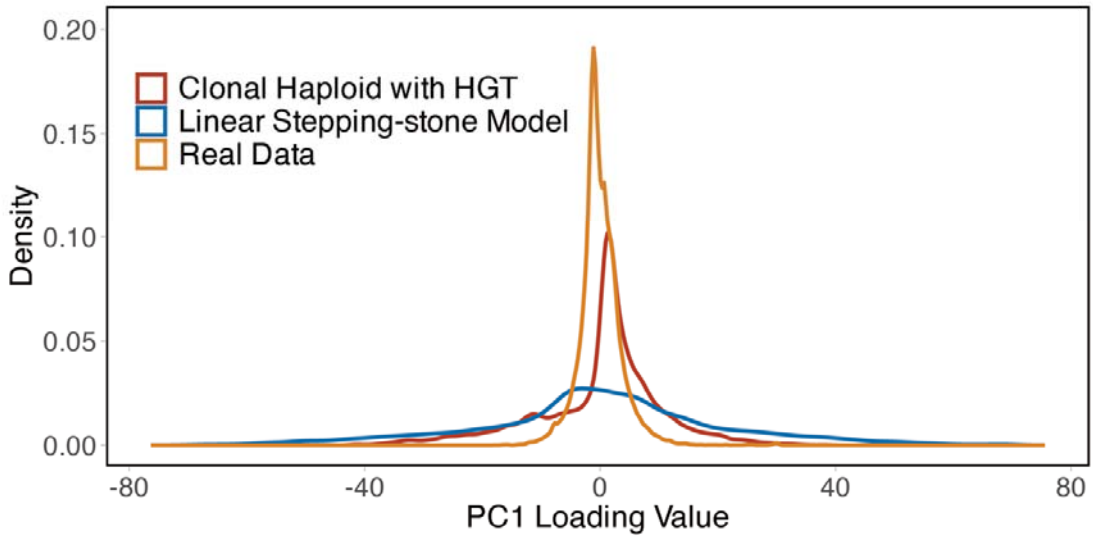
(A-C) Comparison of LD decay between different HGT lengths (100 bp, 1 kb, 10 kb, and 100 kb) and the non-redundant Klebsiella pneumoniae dataset. **(A)** Population size of 1,000. **(B)** Population size of 2,000. **(C)** Population size of 10,000. The x-axis represents the SNP distance (log scale), and the y-axis represents the LD value (R^2) on a log scale.



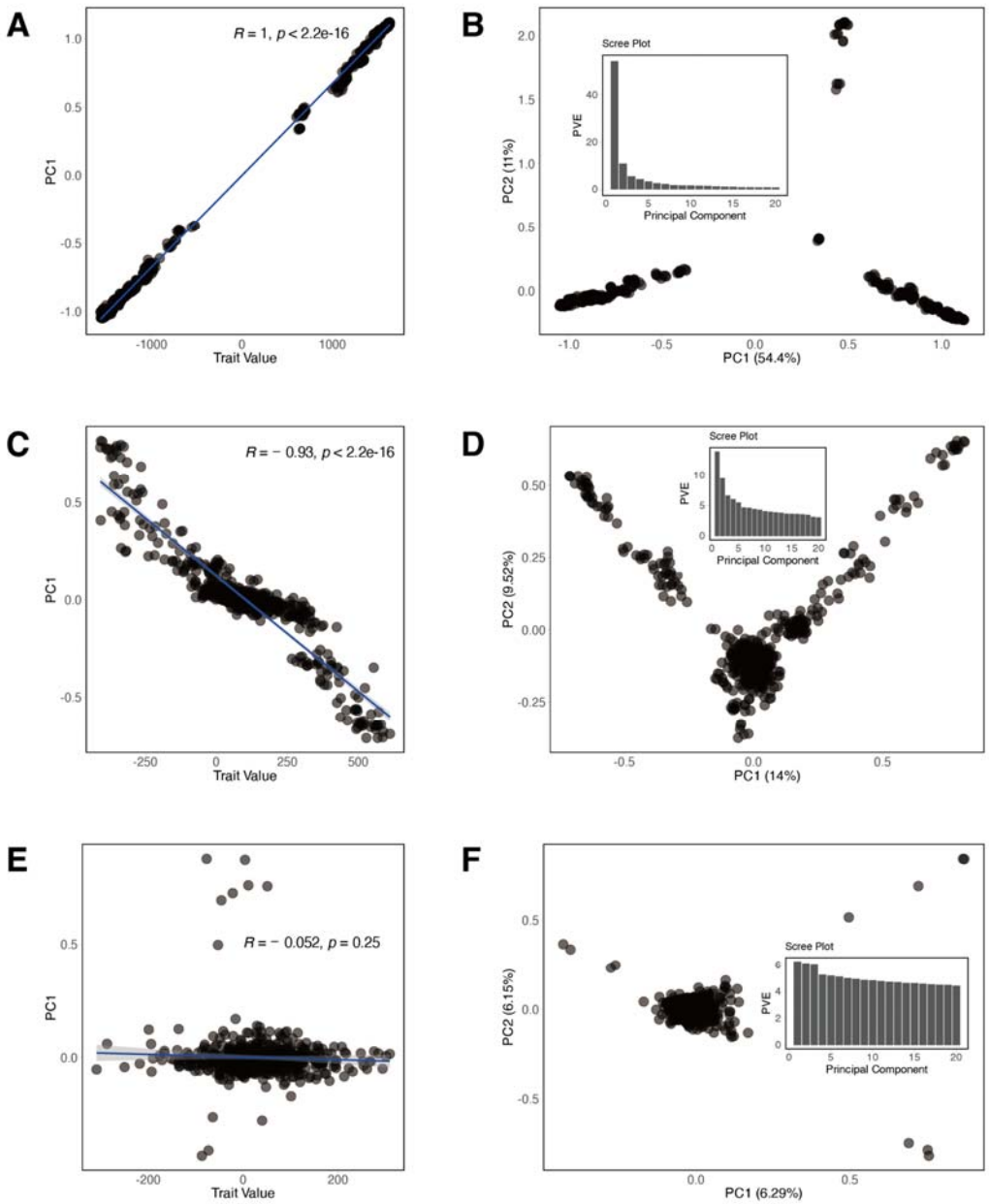
Supplementary Figure 11. Linear stepping-stone model with varying numbers of migrants and a population size of 2000. Phylogenetic trees, principal component analysis (PCA) plots, and scree plots for a model with horizontal gene transfer (HGT) of 10 kb. The population consists of 10 subpopulations (200 individuals each). Panels (A–J) show results for different numbers of migrants exchanged between neighbor subpopulations: **(A-B)** 1 migrant, **(C-D)** 10 migrants, **(E-F)** 20 migrants, **(G-H)** 50 migrants.



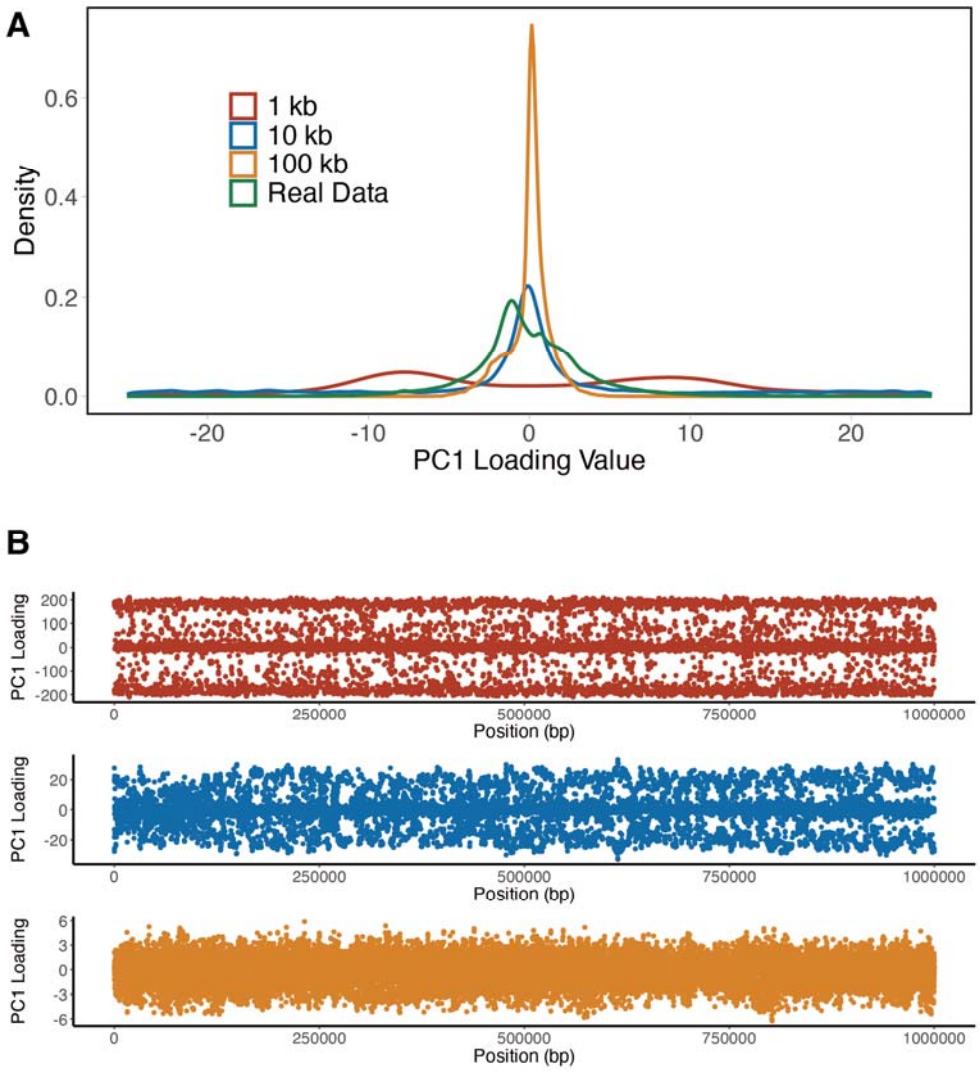
Supplementary Figure 12. Linkage disequilibrium (LD) decay in linear stepping-stone models compared to the non-redundant *K. pneumoniae* data. Linkage disequilibrium (LD) decay profiles are compared between linear stepping-stone models with varying numbers of migrants (1, 10, 20, 50) and the non-redundant *K. pneumoniae* dataset. Panels illustrate results for different total population sizes: **(A)** 2000 individuals, **(B)** 10,000 individuals, **(C)** 100,000 individuals. Each model includes horizontal gene transfer (HGT) of 10 kb. The x-axis represents the SNP distance (log scale), and the y-axis represents the LD value (R^2) on a log scale.



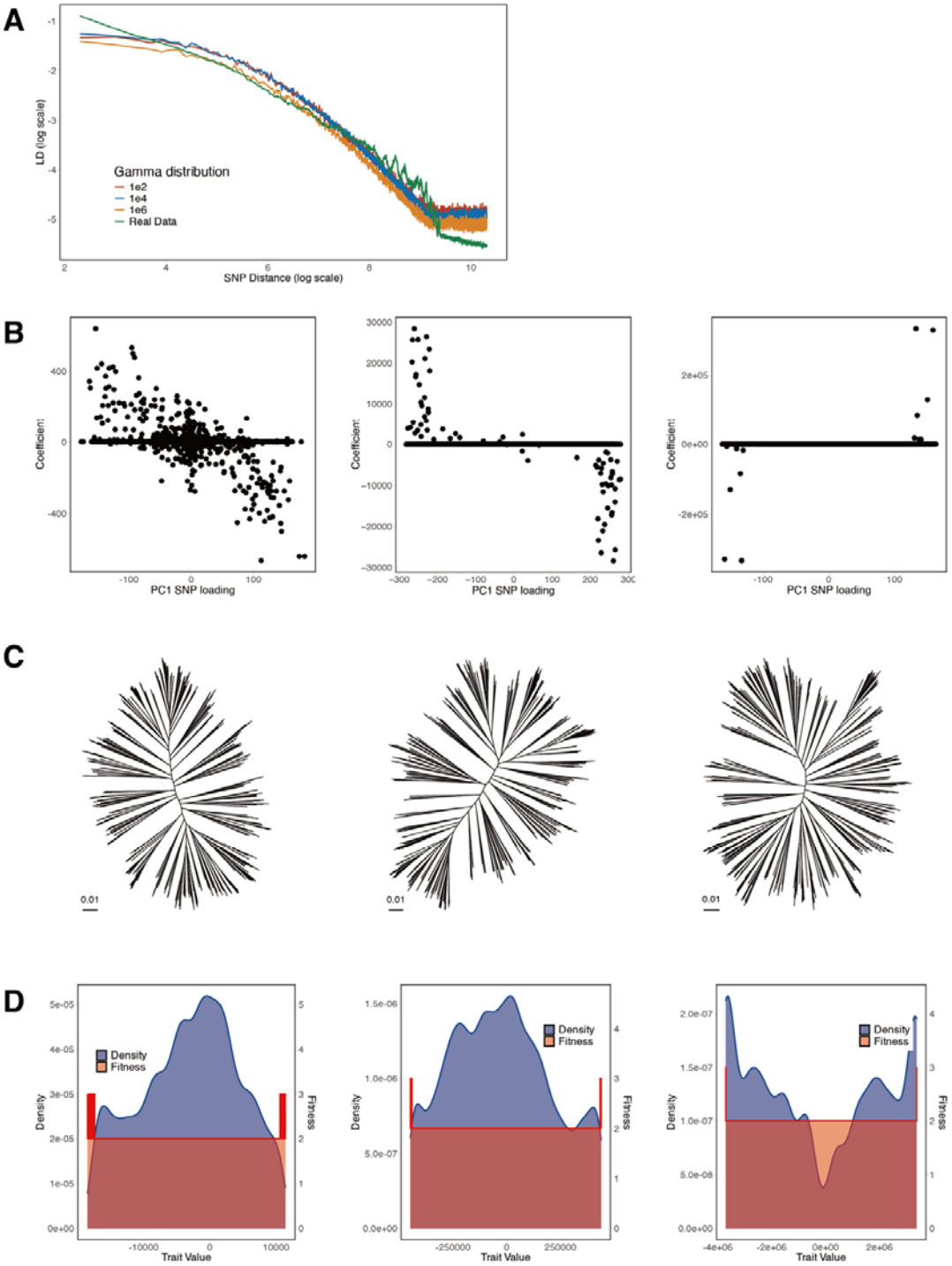
Supplementary Figure 13. Density plot of SNP PC1 loading values illustrating the distribution of SNP loadings for real data and two simulated models.



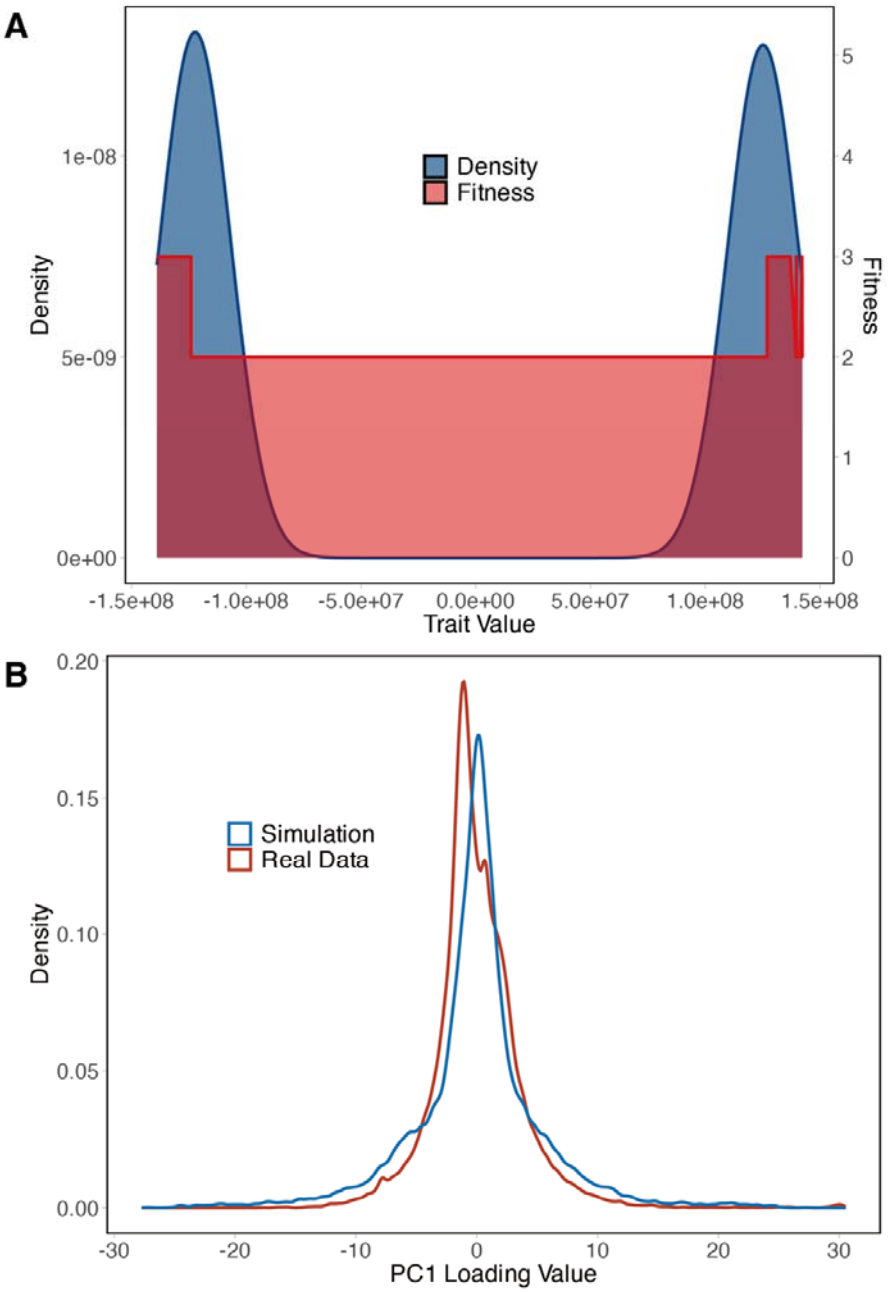
Supplementary Figure 14. PCA results and trait-PC1 correlation in the selection model with varying HGT lengths. (A, C, E) Principal component analysis (PCA) plots corresponding to HGT lengths of 1 kb, 10 kb, and 100 kb, respectively, in the selection model described in Figure 3. **(B, D, F)** Correlation between the trait value (m1–m2) and PC1 for each individual across different HGT lengths. Each plot represents the correlation for HGT lengths of 1 kb, 10 kb, and 100 kb, with correlation coefficients (R) and associated p-values indicated.



Supplementary Figure 15. SNP loadings distribution and genomic positioning under diversifying selection models with varying HGT lengths. (A) Distributions of SNP loadings (-25~25) for horizontal gene transfer (HGT) with different lengths (1 kb, 10 kb, and 100 kb) compared to real data. **(B)** Genomic positions and PC1 loading values of SNPs under selection models with varying HGT lengths.



Supplementary Figure 16: Comparison of LD decay, phylogenetic trees, and trait-fitness relationships under the 1 Mb optimized selection model with varying gamma distribution scales for locus-specific selection coefficients. (A) LD decay curves for SNPs generated under varying gamma distribution scales (10^2, 10^4, 10^6), compared to real data. **(B)** Relationship between SNP coefficients and PC1 SNP loadings for the three gamma scales. **(C)** Phylogenetic trees reconstructed from SNPs with gamma distribution scales of 10^2, 10^4, and 10^6. **(D)** Distribution of trait values and fitness for SNPs under the three gamma scales, with fitness modeled using Poisson distributions for offspring numbers.



Supplementary Figure 17. 5 Mb selection model with locus-specific selection coefficients. (A) Trait value density distributions and fitness function. **(B)** Distribution of PC1 SNP loadings for the 5 Mb simulation compared to the real data.

References

1. Bagley, S. T. Habitat Association of Klebsiella Species. *Infection Control & Hospital Epidemiology* **6**, 52–58 (1985).

2. Håkonsholm, F. *et al.* Insights into the genetic diversity, antibiotic resistance and pathogenic potential of *Klebsiella pneumoniae* from the Norwegian marine environment using whole-genome analysis. *International Journal of Hygiene and Environmental Health* **242**, 113967 (2022).

3. Calland, J. K. *et al.* Population structure and antimicrobial resistance among Klebsiella isolates sampled from human, animal, and environmental sources in Ghana: a cross-sectional genomic One Health study. *The Lancet Microbe* **4**, e943–e952 (2023).

4. Thorpe, H. A. *et al.* A large-scale genomic snapshot of Klebsiella spp. isolates in Northern Italy reveals limited transmission between clinical and non-clinical settings. *Nat Microbiol* **7**, 2054–2067 (2022).

5. Budia-Silva, M. *et al.* International and regional spread of carbapenem-resistant Klebsiella pneumoniae in Europe. *Nat Commun* **15**, 5092 (2024).

6. Madni, O., Amoako, D. G., Abia, A. L. K., Rout, J. & Essack, S. Y. Genomic Investigation of Carbapenem-Resistant Klebsiella pneumonia Colonization in an Intensive Care Unit in South Africa. *Genes (Basel)* **12**, 951 (2021).

7. Wyres, K. L. *et al.* Genomic surveillance for hypervirulence and multi-drug resistance in invasive Klebsiella pneumoniae from South and Southeast Asia. *Genome Med* **12**, 11 (2020).

8. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proceedings of the National Academy of Sciences* **112**, E3574–E3581 (2015).

9. Wyres, K. L. & Holt, K. E. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Current Opinion in Microbiology* **45**, 131–139 (2018).

10. Yahara, K. *et al.* The Landscape of Realized Homologous Recombination in Pathogenic Bacteria. *Mol Biol Evol* **33**, 456–471 (2016).

11. Levin, B. R. & Cornejo, O. E. The Population and Evolutionary Dynamics of Homologous Gene Recombination in Bacteria. *PLOS Genetics* **5**, e1000601 (2009).

12. Smith, J. M., Feil, E. J. & Smith, N. H. Population structure and evolutionary dynamics of pathogenic bacteria. *BioEssays* **22**, 1115–1122 (2000).

13. Guttman, D. S. & Dykhuizen, D. E. Detecting Selective Sweeps in Naturally Occurring Escherichia Coli. *Genetics* **138**, 993–1003 (1994).

14. Abaza, S. What is and why do we have to know the phylogenetic tree? *Parasitologists United Journal* (2020) doi:10.21608/puj.2020.35843.1082.

15. Wyres, K. L., Lam, M. M. C. & Holt, K. E. Population genomics of Klebsiella pneumoniae. *Nat Rev Microbiol* **18**, 344–359 (2020).

16. Yang, C. *et al.* Recent mixing of Vibrio parahaemolyticus populations. *ISME J* **13**, 2578–2588 (2019).

17. Cui, Y. *et al.* The landscape of coadaptation in Vibrio parahaemolyticus. *eLife* **9**, e54136 (2020).

18. Ringnér, M. What is principal component analysis? *Nat Biotechnol* **26**, 303–304 (2008).

19. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A.* **374**, 20150202 (2016).

20. Elhaik, E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Sci Rep* **12**, 14683 (2022).

21. Lam, M. M. C. *et al.* A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. *Nat Commun* **12**, 4188 (2021).

22. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).

23. Tourrette, E. *et al.* An ancient ecospecies of Helicobacter pylori. *Nature* **635**, 178–185 (2024).

24. Wu, C.-C., Huang, Y.-J., Fung, C.-P. & Peng, H.-L. Regulation of the Klebsiella pneumoniae Kpc fimbriae by the site-specific recombinase KpcI. *Microbiology (Reading)* **156**, 1983–1992 (2010).

25. Potter, R. F. *et al.* Population Structure, Antibiotic Resistance, and Uropathogenicity of Klebsiella variicola. *mBio* **9**, 10.1128/mbio.02481-18 (2018).

26. Khater, F. *et al.* In silico analysis of usher encoding genes in Klebsiella pneumoniae and characterization of their role in adhesion and colonization. *PLoS One* **10**, e0116215 (2015).

27. Ma, L.-C., Fang, C.-T., Lee, C.-Z., Shun, C.-T. & Wang, J.-T. Genomic heterogeneity in Klebsiella pneumoniae strains is associated with primary pyogenic liver abscess and metastatic infection. *J Infect Dis* **192**, 117–128 (2005).

28. Yang, Y. *et al.* Genomic Diversity, Virulence, and Antimicrobial Resistance of Klebsiella pneumoniae Strains from Cows and Humans. *Appl Environ Microbiol* **85**, e02654-18 (2019).

29. Ma, L.-C., Fang, C.-T., Lee, C.-Z., Shun, C.-T. & Wang, J.-T. Genomic Heterogeneity in Klebsiella pneumoniae Strains Is Associated with Primary Pyogenic Liver Abscess and Metastatic Infection. *The Journal of Infectious Diseases* **192**, 117–128 (2005).

30. Falush, D. *et al.* Mismatch induced speciation in Salmonella: model and data. *Philos Trans R Soc Lond B Biol Sci* **361**, 2045–2053 (2006).

31. Didelot, X., Méric, G., Falush, D. & Darling, A. E. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics* **13**, 256 (2012).

32. Koeppel, A. *et al.* Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci U S A* **105**, 2504–2509 (2008).

33. Brown, T., Didelot, X., Wilson, D. J. & Maio, N. D. SimBac: simulation of whole bacterial genomes with homologous recombination. *Microb Genom* **2**, e000044 (2016).

34. Thorell, K. *et al.* The Helicobacter pylori Genome Project: insights into H. pylori population structure from analysis of a worldwide collection of complete genomes. *Nat Commun* **14**, 8184 (2023).

35. Yang, J. *et al.* The human gut serves as a reservoir of hypervirulent Klebsiella pneumoniae. *Gut Microbes* **14**, 2114739.

36. Heng, H. *et al.* Global genomic profiling of Klebsiella pneumoniae: A spatio-temporal population structure analysis. *International Journal of Antimicrobial Agents* **63**, 107055 (2024).

37. Anderson, J. T., Wagner, M. R., Rushworth, C. A., Prasad, K. V. S. K. & Mitchell-Olds, T. The evolution of quantitative traits in complex environments. *Heredity* **112**, 4–12 (2014).

38. Darwin, C. & Darwin, C. *The Origin of Species and the Descent of Man in Relation to Sex*. (John Murray, London, 1871). doi:10.5962/bhl.title.110063.

39. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics* **110**, 179–194 (2023).

40. Biscarini, F., Nicolazzi, E. L., Stella, A., Boettcher, P. J. & Gandini, G. Challenges and opportunities in genetic improvement of local livestock breeds. *Front Genet* **6**, 33 (2015).

41. Bernardo, R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity* **125**, 375–385 (2020).

42. Mackay, T. F. & Moore, J. H. Why epistasis is important for tackling complex human disease genetics. *Genome Medicine* **6**, 42 (2014).

43. Yang, J. *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565–569 (2010).

44. Nosil, P. *et al.* Natural selection and the predictability of evolution in Timema stick insects. *Science* **359**, 765–770 (2018).

45. Walsh, B. & Lynch, M. *Evolution and Selection of Quantitative Traits*. (Oxford University Press, Oxford, 2018). doi:10.1093/oso/9780198830870.001.0001.

46. Mai, T. T., Lees, J. A., Gladstone, R. A. & Corander, J. Inferring the heritability of bacterial traits in the era of machine learning. *Bioinformatics Advances* **3**, vbad027 (2023).

47. Lehtinen, S. *et al*. Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proceedings of the National Academy of Sciences* **114**, 1075–1080 (2017).

48. Keenan, J. D. *et al.* Evidence for Clonal Expansion After Antibiotic Selection Pressure: Pneumococcal Multilocus Sequence Types Before and After Mass Azithromycin Treatments. *J Infect Dis* **211**, 988–994 (2015).

49. Varona, L., Legarra, A., Toro, M. A. & Vitezica, Z. G. Non-additive Effects in Genomic Selection. *Frontiers in Genetics* **9**, (2018).

50. Mackay, T., Stone, E. & Ayroles, J. The genetics of quantitative traits: Challenges and Prospects. *Nature reviews. Genetics* **10**, 565–77 (2009).

51. Westoby, M. *et al.* Strategic traits of bacteria and archaea vary widely within substrate-use groups. *FEMS Microbiology Ecology* **97**, fiab142 (2021).

52. Ma, J. & Amos, C. I. Principal Components Analysis of Population Admixture. *PLOS ONE* **7**, e40115 (2012).

53. van Waaij, J., Li, S., Garcia-Erill, G., Albrechtsen, A. & Wiuf, C. Evaluation of population structure inferred by principal component analysis or the admixture model. *Genetics* **225**, iyad157 (2023).

54. Barletta, M., Lima, A. R. A., Costa, M. F. & Dantas, D. V. Estuarine Ecoclines and the Associated Fauna: Ecological Information as the Basis for Ecosystem Conservation. in *Coastal Wetlands: Alteration and Remediation* (eds. Finkl, C. W. & Makowski, C.) 479–512 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-56179-0_16.

55. Kent, M., Gill, W. J., Weaver, R. E. & Armitage, R. P. Landscape and plant community boundaries in biogeography. *Progress in Physical Geography: Earth and Environment* **21**, 315–353 (1997).

56. *Population Genomics: Microorganisms*. (Springer, Cham, 2019).

57. Jenal, U. & Malone, J. Mechanisms of cyclic-di-GMP signaling in bacteria. *Annu Rev*

Genet **40**, 385–407 (2006).

58. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

59. O'Leary, N. A. *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci Data* **11**, 732 (2024).

60. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944 (2018).

61. Yang, C., Cui, Y., Didelot, X., Yang, R. & Falush, D. Why panmictic bacteria are rare. 385336 Preprint at https://doi.org/10.1101/385336 (2019).

62. Cui, Y. *et al.* Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus. *Molecular Biology and Evolution* **32**, 1396–1410 (2015).

63. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056 (2016).

64. Seemann & Torsten. Pairwise SNP distance matrix from a FASTA sequence alignment. (2018).

65. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet* **8**, e1002453 (2012).

66. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

67. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28–36 (2017).

68. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).

69. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

70. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).

71. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

72. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

73. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).

74. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).

75. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research* **32**, D115–D119 (2004).

76. Haller, B. C. & Messer, P. W. SLiM 4: Multispecies Eco-Evolutionary Modeling. *The American Naturalist* **201**, E127–E139 (2023).

77. Cury, J., Haller, B. C., Achaz, G. & Jay, F. Simulation of bacterial populations with SLiM.