A putative bacterial ecocline in *Klebsiella pneumoniae*

Siqi Liu, Sarah L. Svensson, Daniel Falush*

Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai.

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

Abstract

The genetic structure of bacterial species is most often interpreted in terms of demographic processes such as clonal descent but can also reflect natural selection and hence give functional and ecological insight. *Klebsiella pneumoniae* (KP) disperses effectively around the world and has high recombination rates, which should result in the species having a well-mixed gene pool. Nevertheless, phylogenies based on diverse KP strains contain a "backbone". This structure reflects a component of variation where the first component in Principal Components Analysis (PCA), PC1, explains 16.8% of the total variation. We propose that the component reflects a “bacterial ecocline” generated by diversifying selection on a quantitative genetic trait. We simulated a model in which a trait is influenced by many genes, and strains with the most extreme trait values have a small advantage, which can recapitulate our KP PCA results and other features of its genetic diversity. As well as providing an explanation for the phylogenetic backbone, our results provide insight into how species such as KP can speciate, via stronger selection on the trait or a reduction in gene flow. Our hypothesis that there is a bacterial ecocline in KP raises two questions, namely what the trait is underlying it and why is the trait under diversifying selection? The genes that are most strongly associated with PC1 provide some hints, with the top locus encoding Kpa fimbriae. Identification of the trait, if it exists, should facilitate insight into selection on quantitative genetic traits in natural bacterial populations, which have largely been unstudied in microbiology, except in the atypical context of antibiotic resistance.

Introduction

What information can be gleaned from phylogenetic trees[1]? The shape of a tree reflects key demographic factors, such as clonal relationships, geographic structure shaped by barriers to dispersal and genetic exchange[2]. Deep splits within trees and an absence of intermediate strains can for example reflect long-standing divergence between groups that has arisen due to niche specialization and/or low levels of genetic exchange[3,4]. However, many species show a pattern of differentiation that is more continuous in character, which can be harder to explain.

Klebsiella pneumoniae (KP) is a major human pathogen capable of causing severe infections in immunocompromised individuals[5], but is found in a wide variety of environments[6], including soil[7], wastewater[8], and even the ocean[9]. Like many other bacterial species, KP has an enormous population size, exhibits high transmission efficiency and has high recombination rates between strains[10]. As a result, its global genetic diversity can be captured from samples collected in a single region, whether from Europe[11], Africa[12], Asia[13], or elsewhere. Despite these features, as we will describe in detail below, the phylogenetic tree of a broad sample of KP strains shows deep structure, with strains radiating from multiple points along a "backbone". Until now, to our knowledge, no explanations have been proposed for this shape.

We propose that the shape of the phylogenetic tree of KP reflects diversifying selection[14] acting on a quantitative trait[15]. Specifically, we simulate a population of bacteria with a single phenotypic trait. Genetic variants influence the trait additively, meaning that specific variants increasing or decrease trait values by a fixed amount that does not depend on the presence or absence of other variants. We show that if strains exhibiting extreme trait values have a fitness advantage, we can recapitulate patterns of diversity found in KP. We call the gradient in variant frequency values that arises a "bacterial ecocline". Quantitative traits are central to our understanding of genetic diversity in animals and plants but have hitherto been largely overlooked for bacteria. If we can identify the trait or traits under selection in KP, it should provide considerable insight into the genetic architecture of bacterial quantitative traits as well as the fitness trade-offs[16] that the species experiences.

Results

The phylogenetic backbone reflects a component of variation

Genetic relationships amongst bacteria are most often represented using a phylogenetic tree[1,17]. Trees based on core genome SNPs from global collections of *Klebsiella pneumoniae* (KP) isolates are "star-like" (Figure 1a & Supplementary Figure 1, non-redundant dataset with clonally related strains removed; Supplementary Figure 2, entire dataset), with many branches of approximately equal length. This property of the tree reflects extensive recombination, with the length of the branches reflecting the overall level of diversity in the gene pool[10,18]. However, the overall shape of the KP tree is oval rather than approximately circular as for example seen in *Vibrio parahaemolyticus*[19,20] and KP lineages, instead of radiating from a single point, are connected by a "backbone" (Figure 1a & Supplementary Figure 2). This backbone is present in any phylogeny drawn based on a diverse sample of KP genomes (Figure 1a and Supplementary Figure 2) but it may be less visually evident than in Figure 1a, depending on the strains sampled, the rooting of the tree, and the phylogenetic and plotting algorithms.

The backbone of the KP tree is a line from which multiple branches radiate at different points along its length. This structure is suggestive of a continuously varying trait. Informally, the backbone has a top and a bottom, which can be seen in Figure 1a. The strains that branch at the top could, in principle, be given maximum backbone scores and those at the bottom minimum ones, with intermediate strains given values according to how far along the line they branch. We have not defined scores in this way, because there is a more systematic way of describing this variation, namely Principal Components Analysis (PCA)[21,22].

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. When PCA was applied to the non-redundant global KP core genome SNP dataset used to construct the tree in Figure 1a, 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).

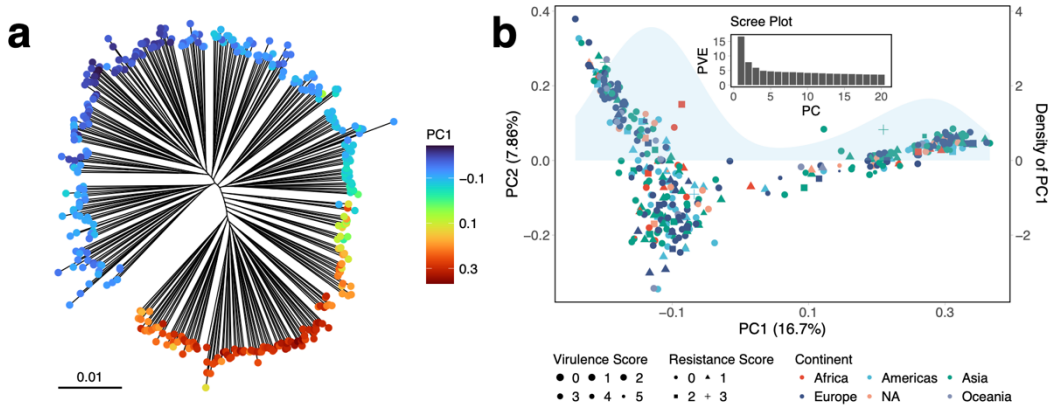


Figure 1. Population structure and genetic variation in a non-redundant KP dataset. (a) Maximum-likelihood phylogenetic tree of the non-redundant KP dataset based on core SNPs, with a color gradient showing PC1 variation. (b) PCA of the non-redundant KP dataset based on LD-pruned SNPs (see Methods). Colors indicate geographic origin, while marker size and shape represent resistance and virulence scores, respectively, determined by Kleborate[23]. The sky-blue shading represents the density of PC1 values. The inset scree plot shows the variance explained (PVE) by the first 20 principal components.

PCA is sensitive to all signals of relatedness, including clonal structure and can be prone to overfitting[24], especially where there are orders of magnitude more genetic markers than strains[24]. However, we found that PC1 was reproducible in different analyses of non-redundant datasets (Figure 2). PC1 strain loadings were consistent if components were defined by strains on the right-hand side of the phylogenetic tree (Supplementary Figure 3) or those on the left, implying that they are not generated by clonal structure (Figure 2a & b). 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 (Figure 2c & d). Specifically, they were also consistent when calculated using SNPs with extreme PC1 loadings or loadings that are closer to 0 (Figure 2e-h), implying that the signal affects genetic variation genome-wide.

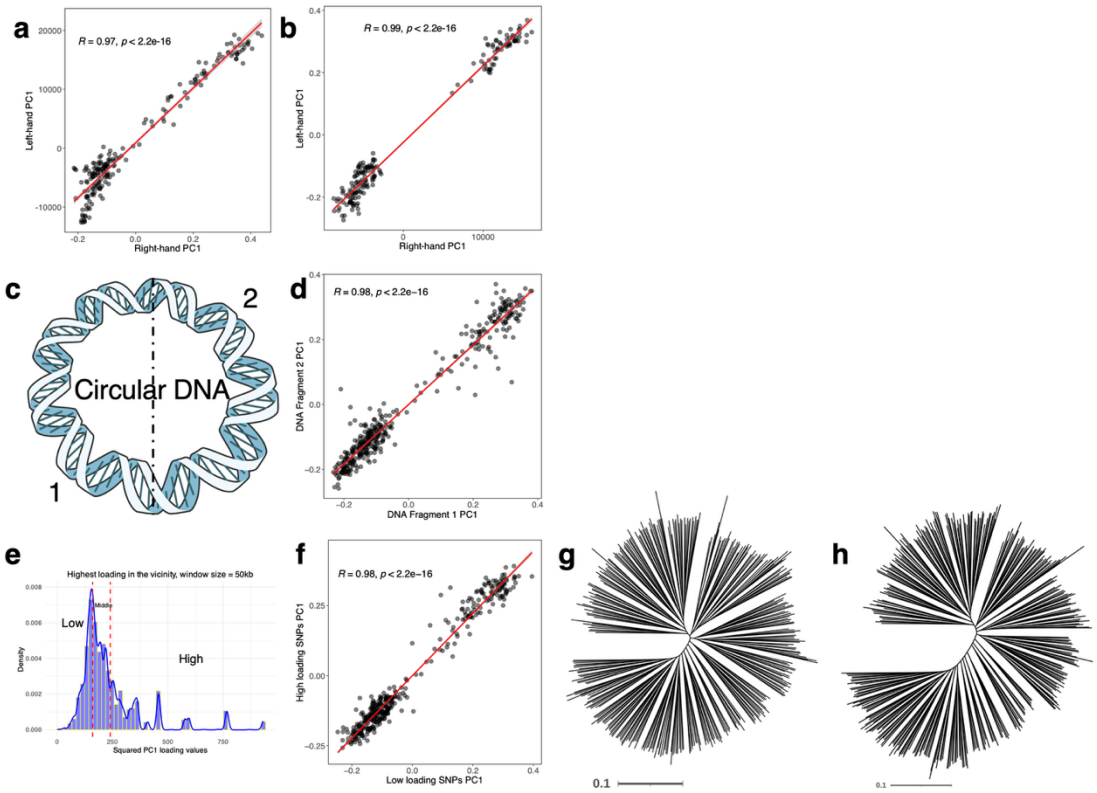
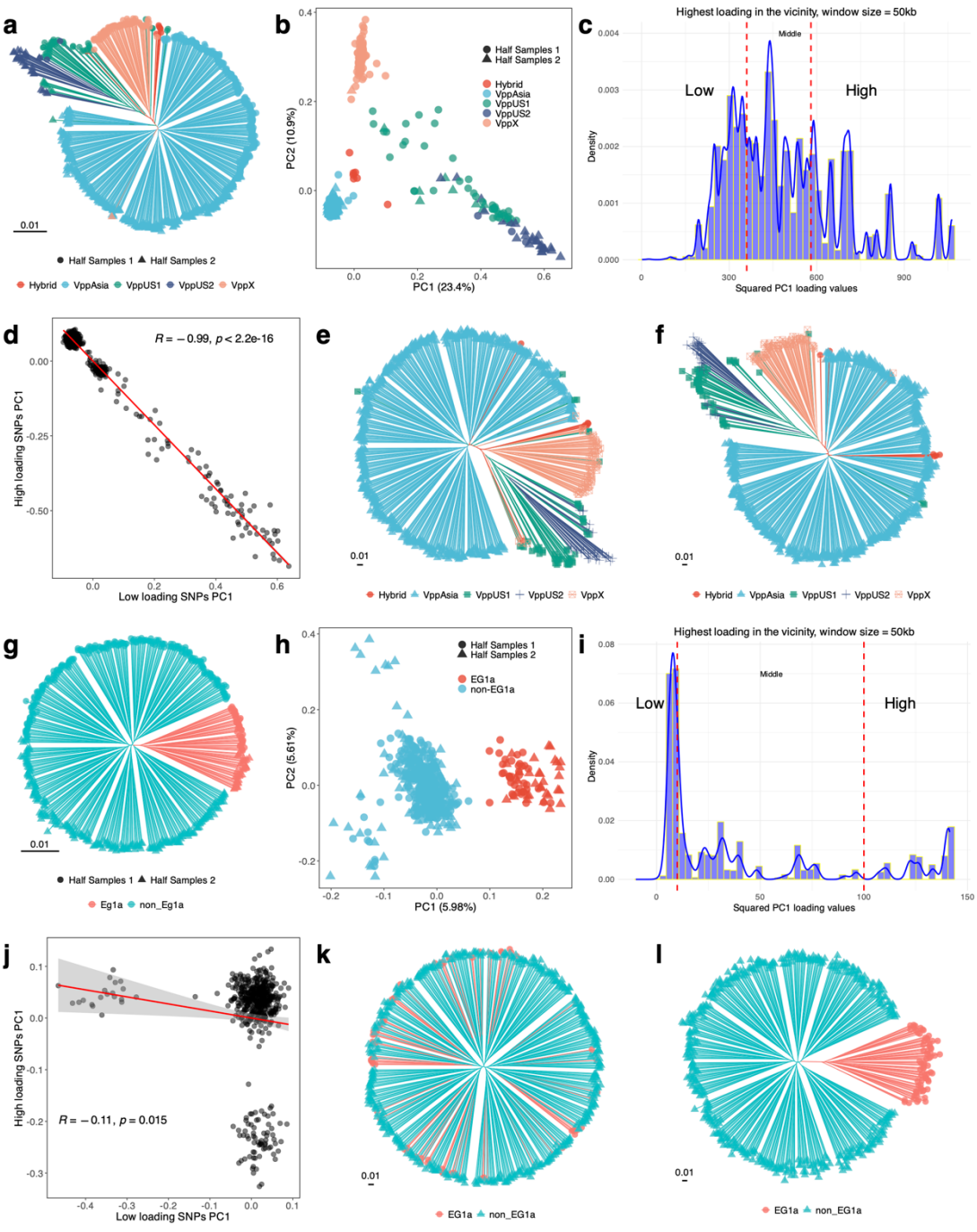


Figure 2. Three types of half-matching tests for the non-redundant KP dataset. (a–b) Half-strain matching: (a) Correlation of PC1 from strains on the right-hand of the tree projected onto left-hand group results. (b) Correlation of PC1 from the left-hand group mapped onto right-hand group results. (c–d) Half-genome Matching: (c) The genome is divided into two fragments: Fragment 1 (1 bp–2.6 Mb) and Fragment 2 (2.6 Mb–5.2 Mb). (d) Correlation of PC1 between the two fragments. (e–h) Partial-variants matching: (e) Histogram of SNP loadings on PC1, dividing SNPs into high- and low-loading groups. (f) Correlation of PC1 between high- and low-loading SNP groups. (g) Phylogenetic tree based on low-loading SNPs. (h) Phylogenetic tree based on high-loading SNPs.

To benchmark these results, we applied the same methods to data from a simulated neutrally evolving bacterial population (Supplementary Figure 4a & 4b), which will be discussed in more detail below. In the simulated data, PC1 did not explain as much of the variance as in the KP data (Supplementary Figure 4c). Furthermore, only one of three half matching tests gave significant correlation when applied to this data (Supplementary Figure 4c-i), contrasting with the results for KP. Thus, according to these criteria, we can falsify the simplest neutral models based on these results.

Vibrio parahaemolyticus (VP) also represents a useful benchmark for half-matching tests because its genetic structure is well characterized. The species has geographical population structure[19], which arises because the species lives in coastal waters and there have been

barriers to movement of bacteria between oceans, albeit disrupted by recent human activity. VP also has a differentiated ecospecies "Molassodon", which has nearly fixed differences with other "Typical" strains in approximately 40 core genes, while being undifferentiated in the rest of the genome[20,25]. This pattern is consistent with frequent recombination between ecospecies homogenizing variation in most of the genome, while natural selection has created and maintained differentiation at ecospecies loci.



136 **Figure 3. Partial-Variants Matching for *Vibrio parahaemolyticus* datasets.** (a) Phylogenetic tree of
137 the non-redundant VP dataset. Branch colors indicate populations defined by fineSTRUCTURE, and tip
138 nodes represent the half-matching-based dataset partition. The Half-Strain Matching and Half-
139 Chromosome Matching results are shown in Supplementary Figure 5 (b) PCA of the non-redundant VP
140 dataset, using the same legend as in panel a. (c) Distribution of SNP loadings, with SNPs categorized
141 into low, medium, and high groups based on the maximum loading value within 50 kb. (d) Correlation
142 of PC1 between high- and low-loading SNP groups for the non-redundant VP dataset. (e) Phylogenetic
143 tree constructed using low-loading SNPs for the non-redundant VP dataset. (f) Phylogenetic tree
144 constructed using high-loading SNPs for the non-redundant VP dataset. (g) Phylogenetic tree of the
145 VppAsia dataset. Branch colors indicate ecospecies, and tip nodes represent the half-matching-based
146 dataset partition (see Supplementary Figure 4). (h) PCA of the VppAsia dataset, using the same legend
147 as in panel g. (i) Distribution of SNP loadings, categorized into low, medium, and high based on loading
148 values. (j) Correlation of PC1 between high- and low-loading SNP groups for the VppAsia dataset. (k)
149 Phylogenetic tree constructed using low-loading SNPs. (l) Phylogenetic tree constructed using high-
150 loading SNPs.

151 When three different sorts of half-matching tests were applied to a global non-redundant
152 dataset of, all gave reproducible results that reflected the population structure in the dataset
153 (Figure 3a-f & Supplementary Figure 5a-e). However, when analysis was restricted to the
154 Asian population (VppAsia), the phylogenetic tree is close to being a perfect circle and
155 PC1 instead reflected ecospecies structure (Figure 3g-l & Supplementary Figure 5f-h). This
156 structure generated weaker correlations in two of the three tests (Figure 3j and
157 Supplementary Figure 5h), but there was no correlation between low loading and high
158 loading loci, consistent with a single randomly mixing gene pool for most the
159 genome[20,26]. Thus, the reproducibility of PC1 in all three tests in KP shows that the
160 structure is genome-wide and in this respect is more like geographic population structure
161 and less similar to ecospecies structure, which is only present in a fraction of the
162 genome[26].

163 **Loci most strongly associated with PC1 are localized in the genome**

164 In addition to strain loadings (Figure 1b), PCA generates SNP loadings, indicating which
165 genomic variants are most strongly associated with the component. SNP loadings for PC1
166 showed four clear peaks (Figure 4a, left). Each peak contained many SNPs within less than
167 10 kb (Supplementary Figure 6). We also calculated loadings for accessory gene presence-
168 absence, based on a Panaroo-generated[27] pangenome of the non-redundant strains, and
169 found many had high loadings (Figure 4a, right).

170 To investigate potential epistatic fitness interactions between the loci most strongly
171 correlated with PC1, we chose a cutoff of 300 for squared PC1 loadings and calculated

pairwise linkage disequilibrium (LD) values for all SNPs and genes above this cutoff. Hierarchical clustering of LD values organized the extreme loading variants into 44 blocks (Figure 4b). Notably, the top four blocks include a large fraction of all loci with PC1 values greater than 300, including the highest loading accessory genes. The organization and gene content of these blocks is shown in Figure 4c and Supplementary Figure 7.

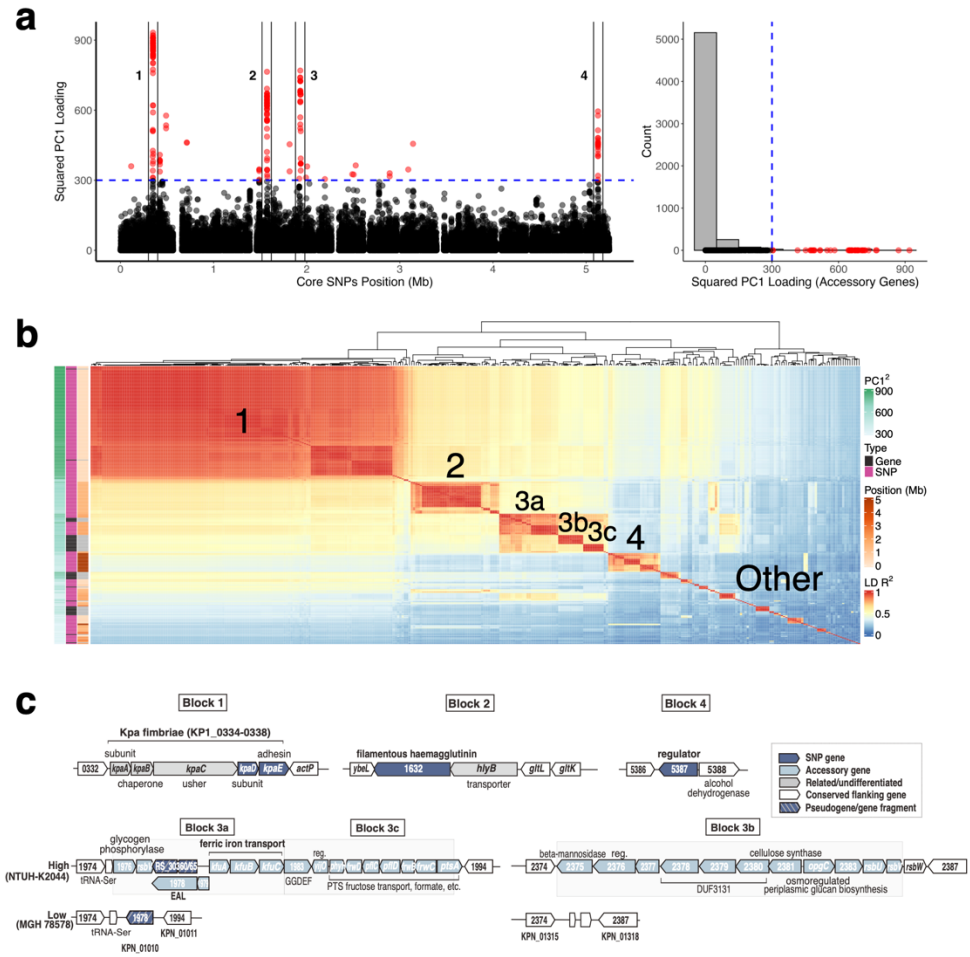


Figure 4. Location, Function, and Linkage Disequilibrium of PC1-Associated Loci. (a) Manhattan plot of squared core SNP (left) and accessory gene (right) loadings for PC1. Variants with values greater than 300 are defined as high loading variants and are marked in red. (b) Heatmap of pairwise LD R^2 values for high loading variants, with the first four blocks (including subblocks) labelled. (c) Functional annotation of blocks 1-4.

Block 1 (Supplementary Figure 7a-c), 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 part of the core genome in both KP and *K. variicola*[28,29], but found at lower frequency in other *Klebsiella* species. While the Kpa

187 fimbriae are less characterized than those encoded by the *fim* locus, experiments in LM21
188 (a low PC1-loading strain) found that absence of *kpaC*, encoding the usher, significantly
189 reduced *in vitro* biofilm formation[30]. The Δ*kpaC* mutant was not affected for binding to
190 either a human cell line or *Arabidopsis* seedlings[30]. The most differentiated gene
191 (KP1_0338/*kpaE*) encodes the adhesin subunit. The high and low variants are too distinct
192 to have evolved within KP and cluster separately on a multispecies tree of the locus
193 (Supplementary Figure 7c), implying that one or both variants was likely imported from
194 another species.

195 Block 2 (Supplementary Figure 7d & e) and block 4 (Supplementary Figure 7f & g) also
196 include SNPs in single genes (KP1_1632, filamentous haemagglutinin N-terminal domain-
197 containing protein and KP1_5387, winged helix-turn-helix transcriptional regulator,
198 respectively). The former, as for the *Kpa* fimbral adhesin, suggests a difference in adhesion,
199 but so far appears uncharacterized. The regulator encoded by KP1_5387 in block 4 is also
200 unstudied but seems always encoded next to an alcohol dehydrogenase in KP, *K. variicola*,
201 *K. quasipneumoniae*, *K. africana*, *K. quasivariicola* strains.

202 Block 3 (Supplementary Figure 7h & i) includes both SNPs and accessory genes and can
203 be subdivided into three sub-blocks (3a, 3b, 3c) based on LD patterns. The majority of
204 strains fall into two configurations (Supplementary Figure 7g), with the high-loading
205 configuration including all accessory gene elements and the low loading configuration
206 including none, but there are also strains with intermediate configurations.

207 Block 3 genes are located in two different genomic regions in the reference strain, with
208 block 3b genes encoded at a different location from block 3a and 3c genes (Figure 4c).
209 Blocks 3a and 3c were previously highlighted as associated with invasive strains[31], as
210 well as those implicated in bovine mastitis[32]. Key genes of block 3a are *kfuABC*,
211 encoding a ferric iron ABC transporter, which aids iron acquisition and impacts mouse
212 colonization[31,33]. The block also encodes two c-di-GMP-related proteins, a regulator,
213 and glycogen/PTS (phosphotransferase) sugar utilization/uptake genes. Block 3c is
214 adjacent to block 3a in the genome and contains genes related to sugar uptake, metabolism,
215 and regulation (again via c-di-GMP). Block 3b contains accessory genes encoding a
216 cellulose synthase, three DUF3131 proteins, OpgC (osmoregulated glucan production),
217 and several regulatory proteins (hybrid sensor kinase, anti-sigma-factor and its antagonist).

218 **Functional enrichment analysis**

219 To establish whether PC1 exerts a greater influence on core or accessory genome variation,
220 we compared the distribution of absolute PC1 loading values between core SNPs, core

genes and accessory genes. The dashed vertical lines indicate the mean loading value for each group, which are closely aligned, suggesting similar overall contributions to PC1 (Figure 5a). However, accessory genes display a longer right tail in the distribution, indicating the presence of more extreme loading values. This suggests that, although the average contribution is similar, a subset of accessory genes is more strongly influenced by PC1.

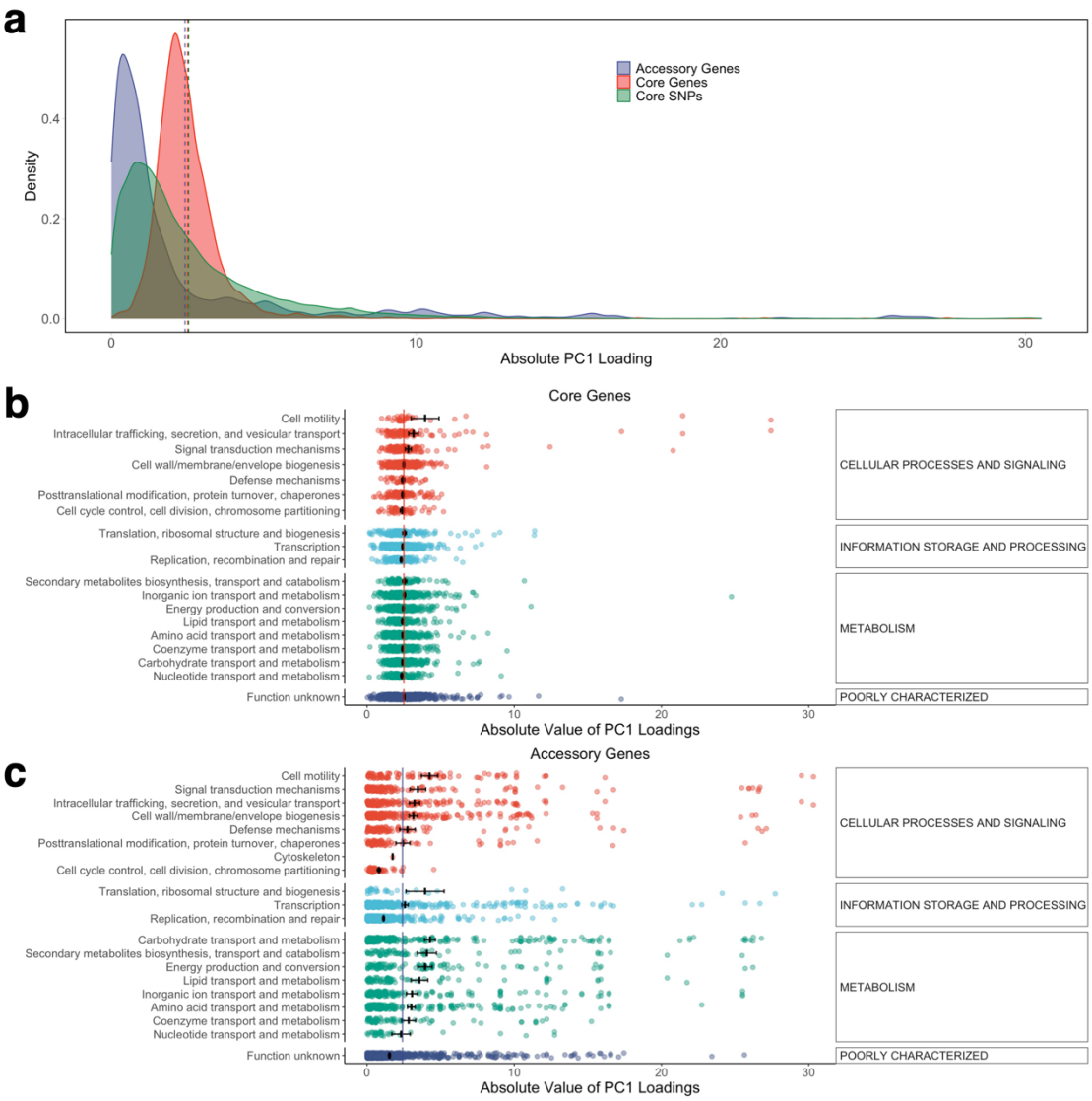


Figure 5. Overall and COG-specific distributions of absolute PC1 loading values for variants. (a) Density plot of PC1 loading values for core SNPs, core genes and accessory genes. Mean absolute PC1 loadings of core SNPs, (b) core genes and (c) accessory genes grouped by COG category. Bars represent the average absolute loading within each COG category, with error bars indicating the standard error. The solid (a) and dashed (b) lines mark the overall mean absolute loading across all features (genes or

233 SNPs, respectively).

234 To establish whether there is a systematic relationship between gene function and SNP
235 loadings, we calculated the mean absolute SNP loading for genes in each COG (Clusters
236 of Orthologous Groups) functional category (Figure 5b) and found almost no difference
237 between categories. However, larger and statistically significant differences were found for
238 accessory genes (Figure 5c), with metabolic genes showing consistently higher mean
239 absolute loading values than those involved in recombination and genome repair or
240 uncharacterized genes.

241 **Virulence, antibiotic and isolation source profiles provide few clues**

242 Although virulence and antibiotic resistance are important for the clinical assessment of
243 KP, our analysis suggests that they are not the primary determinants of its overall
244 population structure, which is not unexpected given that most lineages only cause human
245 disease sporadically. Figure 1b shows the isolate metadata and Kleborate[23]-assigned
246 resistance and virulence scores for the non-redundant dataset. Strains do not cluster based
247 on geography, resistance, or virulence.

248 For multi-host bacteria, host-imposed selective pressures are important drivers of their
249 evolution[34,35]. KP can be isolated from a variety of sources, including animals, plants,
250 humans (which represent the largest proportion in databases), and water, including the sea.
251 However, there is no strong evidence to support that host selective pressure is responsible
252 for shaping the backbone of the KP phylogenetic tree (Supplementary Table 1).

253 Based on KP1_0338, the gene most strongly associated with PC1, we classified the multi-
254 source strains into high and low variant groups (Supplementary Figure 8a). A Fisher's exact
255 test revealed a significant association between KP1_0338 variant group and host source (p
256 < 0.01), primarily driven by strains from water and horses (Supplementary Table 1).
257 However, we found that most water-isolated strains were from Manchester, where they
258 were collected from patient or hospital wastewater sources. According to the Kleborate[23]
259 annotation, most of these strains had a resistance score of 2 and a virulence score of 0 or 1,
260 indicating that they are typical hospital-associated clones. We think this signal is an artifact
261 of the strains mostly belonging to a handful of clonal lineages rather than evidence of true
262 host specificity.

263 The KP isolates from horses were predominantly collected in France and Iceland, from
264 various anatomical sites through both post-mortem examinations and live-animal sampling.
265 However, when we projected the combined dataset — including the non-redundant dataset,

266 along with strains isolated from water and horse sources — onto the results corresponding
267 to Figure 1b (Supplementary Figure 8b), the projection revealed that the distribution of
268 horse-derived strains was not significantly different from that of the non-redundant dataset.
269 Thus, the results provide preliminary evidence for an association of the KP1_0338 locus
270 with colonizing horses, but no evidence for a host association of PC1 as a whole.

271 In summary, our analysis finds no evidence that the observed population structure of KP is
272 shaped by simple host-specific selective pressures. Moreover, no clear associations were
273 observed between population structure and geographic origin, antimicrobial resistance, or
274 virulence. Instead, the distribution of this gene appears to be largely independent of host
275 origin, indicating that other ecological or evolutionary factors that are not captured by
276 current metadata may underlie the observed patterns.

277 **The backbone is not explained by simple neutral models**

278 The presence of this substantial component of variation raises the question: how has it
279 arisen? Simple models of a bacterial population evolving with mutation and homologous
280 recombination failed to recapitulate the patterns we observed in KP. For example,
281 simulations with a population size of 2000 and one homologous recombination event per
282 genome of size 10 kb approximately matched the overall LD decay curve seen in the real
283 data (Figure 6a). However, for these parameters, there was no phylogenetic backbone
284 evident, and PC1 failed to explain as much variance as in the real data (Figure 6b & c). As
285 noted above, PC1 was also not reproducible when applied to different halves of the tree in
286 a half matching test (Supplementary Figure 4b-e). More generally, we did not see a
287 phylogenetic backbone in any simulations of a single neutrally evolving population
288 (representative simulations shown in Supplementary Figure 9).

289 The model of evolution used to generate the data in Figure 6a assumes that recombination
290 rates are uniform across the species. More complex models are possible, for example in
291 which recombination rates are high between closely related strains but lower between more
292 distant ones. Such a pattern can arise for example if homologous recombination is
293 suppressed by mismatches between donor and recipient sequence[36]. Such models are
294 thought to explain much higher recombination rates within *Escherichia coli* phylogroups
295 than between them[37]. Many models with these properties are possible, but since they all
296 reduce the effectiveness of recombination, they are likely to make it difficult to recapitulate
297 the low levels of LD in KP at large genetic distances. Moreover, the models that have been
298 investigated previously cause species to cluster into discrete groups[38,39], rather than
299 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.

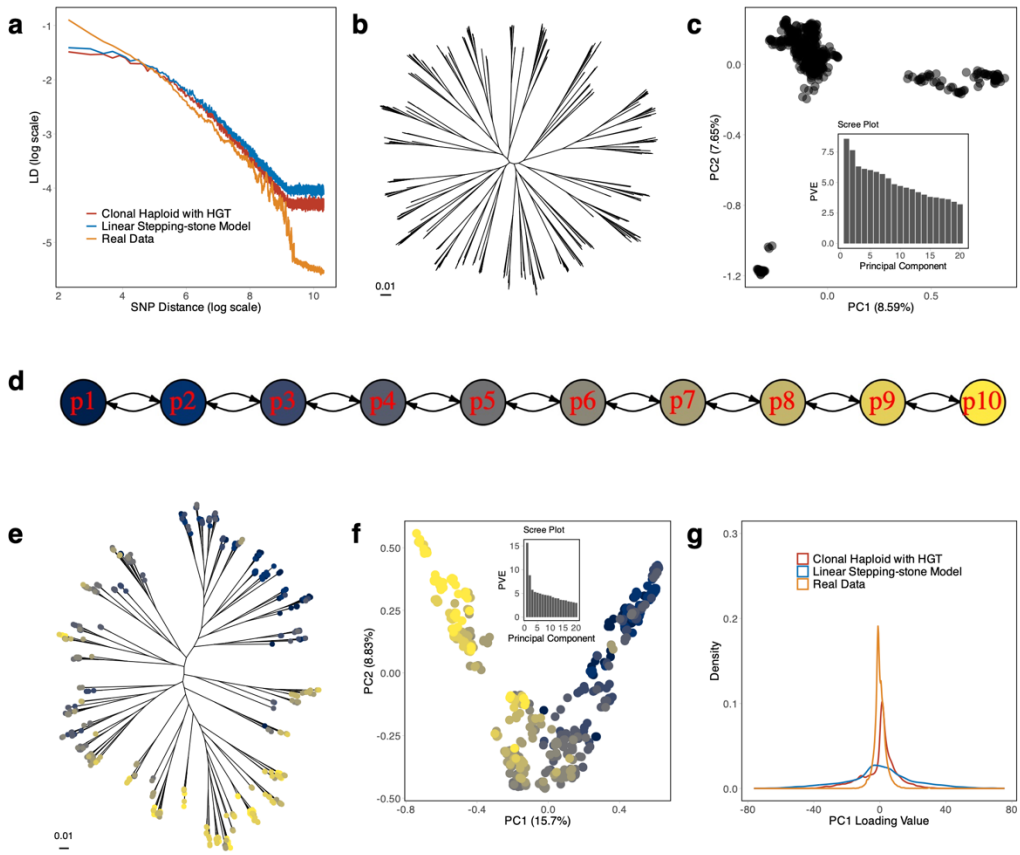


Figure 6. Simulated neutral models. (a) LD decay comparison to real data for best-fit neutral single population and ten-population stepping-stone models. Phylogenetic tree (b) and PCA results (c) for single population evolving neutrally. (d) Illustration of the ten-population stepping-stone model. Phylogenetic tree (e) and PCA (f) for the best fit stepping-stone model, coloured as shown in (d). (g) Density plot of SNP PC1 loading values illustrating the distribution of SNP loadings for real data and two neutral models.

Genetic differentiation can arise due to geographical structure, or other factors that restrict gene flow between particular subsets of the population[20,26,40]. To explore the properties of geographical population structure we next simulated a stepping-stone model with 10 demes (Figure 6d). 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 6e, f and Supplementary Figure 11) and also approximately matched the LD decay curve (Figure 6a and Supplementary Figure 12).

317 However, the distribution of the SNP loadings for PC1 for these simulations did not match
318 the real data, because there were many more SNPs with intermediate loading values (Figure
319 6g). Geographic population structure creates differentiation between populations through
320 genetic drift, which is a random process that affects all of the genome but to differing
321 degrees. The discrepancy in SNP loading values between simulated and real data implies
322 that the variation in differentiation values in KP is higher than expected under a pure
323 random drift model.

324 A longer-tailed distribution of drift levels, and hence PC1 values, could be generated by
325 natural selection at a fraction of loci associated with adaptation to specific demes. However,
326 even if a geographic model with some selection could be made to fit the data, this model
327 seems to be poorly motivated in the case of KP. There is no evidence of KP being structured
328 geographically (Figure 1b)[10,18] or of there being physical barriers to gene flow between
329 ecological demes, since for example it is readily acquired by humans from a wide variety
330 of ecological sources[41,42]. It is therefore hard to see how a model based on strong
331 isolation between demes could be applicable to the species.

332 **KP variation is consistent with diversifying selection on a quantitative trait**

333 Many traits are subject to diversifying selection[43]. For example, beak size has been
334 shown to have two distinct optima in a species of Darwin's finch, *Geospiza fortis,* which
335 seems to have arisen because birds can specialize either on small soft seeds or larger nuts.
336 Birds with intermediate beak sizes are not well suited to either type of food source and
337 have lower fitness during droughts[44].

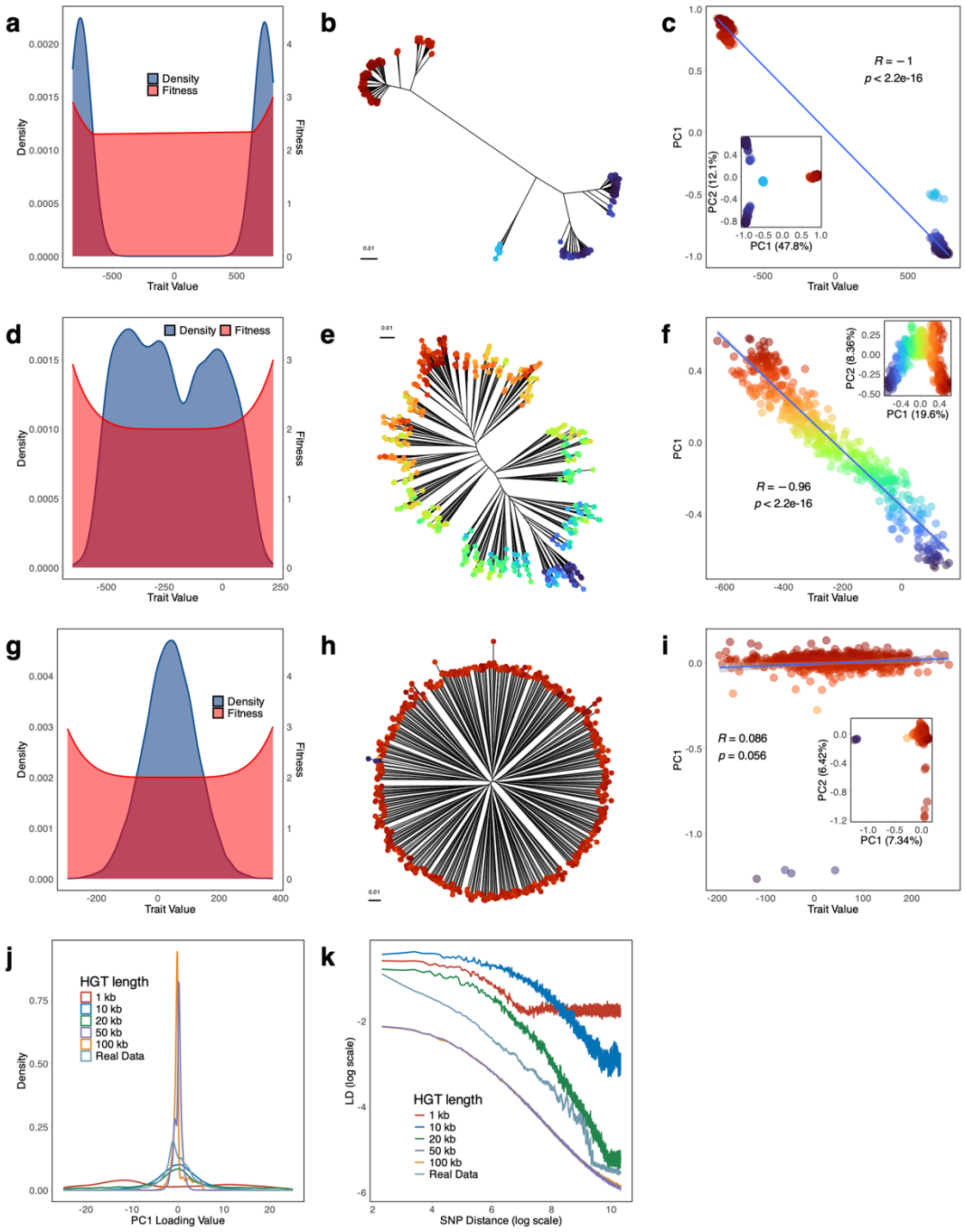


Figure 7. Simulation of diversifying selection on a polygenic trait at different recombination rates. (a, d, g) Trait value density distributions and fitness coefficients for recombination lengths of 1 kb, 20 kb, and 100 kb (Top to bottom) in final generation of simulation. (b, e, h) Phylogenetic trees depicting evolutionary relationships under each HGT length. (c, f, i) Correlations between PC1 values and trait values under each HGT length, with corresponding PC1 vs PC2 plots inserted in each panel. These simulations use 1Mb genomes. (j) Overall distribution of SNP loadings across all conditions. (k) Overall LD decay patterns across all conditions.

346 To investigate the signatures of diversifying selection, we performed simulations of
347 selection on a single quantitative trait in a recombining bacterial population, in which
348 strains with extreme trait values have higher fitness, while the fitness function is flat for
349 intermediate trait values (see Methods). Simulation results in which half of all new
350 mutations increased fitness by one and half decreased fitness by one showed that low
351 recombination rates allow progressive divergence in trait values (Figure 7a-c). At
352 intermediate recombination values, a phylogenetic backbone appeared that is similar to the
353 one seen in the KP data (Figure 7 d-f), while high rates homogenize the population, leading
354 to a circular star-shaped phylogeny (Figure 7g-i). For low and intermediate recombination
355 rates (1 kb, 20 kb), PC1 values were strongly correlated with the true simulated trait values
356 (Figure 7c, f). As expected, linkage disequilibrium values are strongly dependent on
357 recombination rates (Figure 7k), with intermediate rates giving LD curves closest to the
358 KP data.

359 These simulations failed to recapitulate the sharp peaks in SNP loadings seen in the KP
360 data (Figure 7j). We therefore modified the simulations by assuming that the distribution
361 of the magnitudes of the effect of SNPs on the trait followed a Gamma distribution (Figure
362 8). High values of the Gamma shape parameter correspond to some mutations having a
363 much larger effect on the trait than others. In these simulations, each PC1 loading peak
364 corresponds to or is immediately adjacent to a SNP with a large effect on trait values. For
365 high Gamma values, the number of peaks is smaller (Figure 8c, f & i).

366 We found that simulations with a very large Gamma value could recapitulate the sharp
367 peaks seen in the KP data (Figure 8j). For this parameter values, outside the sharp peaks,
368 there is a great deal of variation in PC1 loadings, which is however not influenced by the
369 SNP coefficients (Figure 8f & i) and therefore represents noise, that is generated by genetic
370 drift. For high Gamma shape parameters, the correlation between strain PC1 values and the
371 trait values accordingly becomes weaker (compare Figure 8h with Figure 8b&e) , including
372 if PC1 values are calculated only using top-loading SNPs (not shown). For intermediate
373 Gamma shape parameters, the distribution of SNP loadings looked similar to that obtained
374 assuming each SNP has the same effect on trait values (Figure 8j), but the distribution
375 becomes more leptokurtic (with many SNPs with low values and long thin tails) as Gamma
376 increases.

377

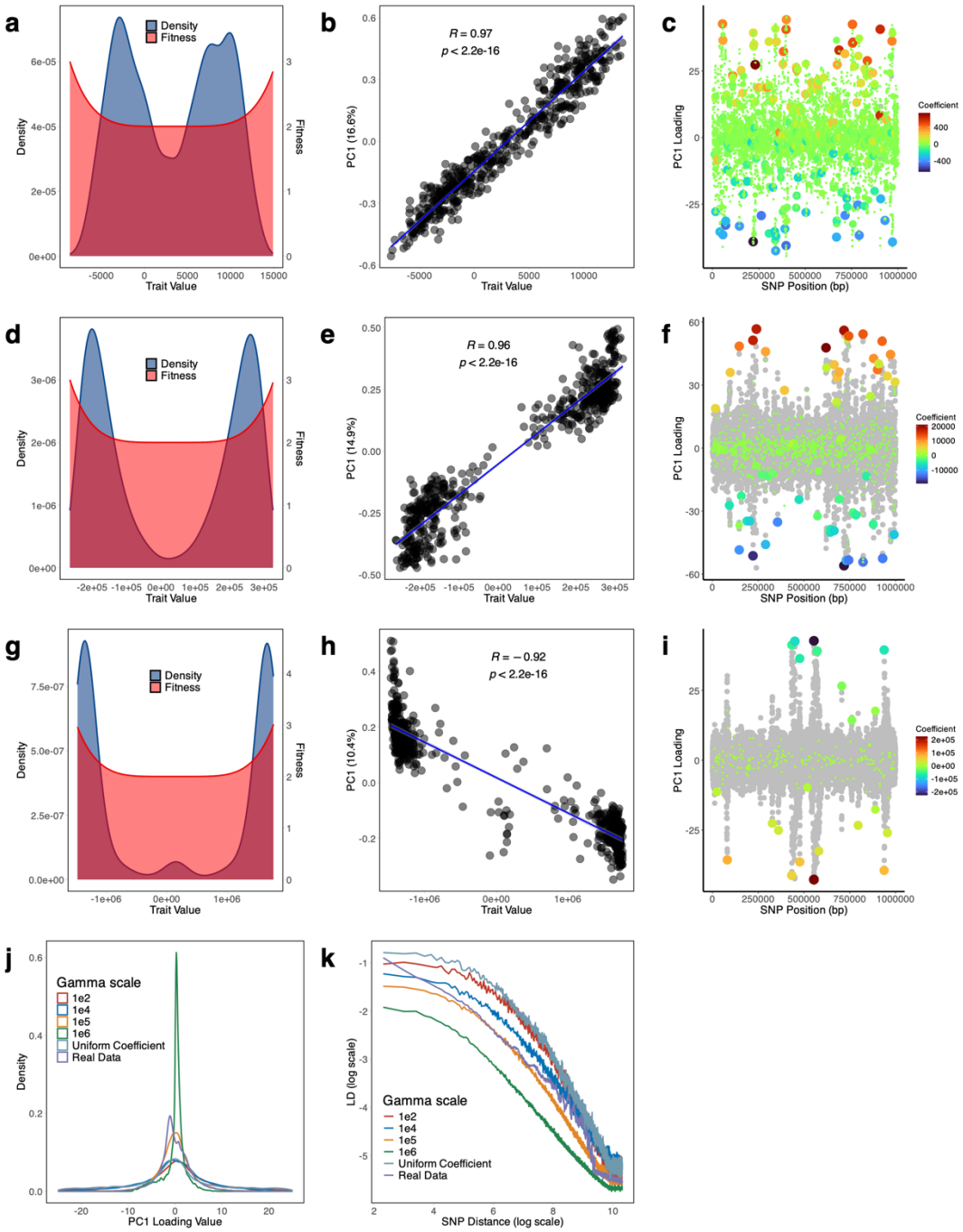


Figure 8. Effects of different selection coefficient on diversifying selection. (a) Distribution of fitness values under a gamma-distributed SNP effect size model (Gamma scale = $1e^2$), plotted against trait values. (b) Correlation between trait values and PC1 scores. (c) SNP loadings along PC1, with point colors indicating the magnitude of SNP coefficients. (d–f) Same as panels A–C, but under a gamma distribution with Gamma scale = $1e^4$. (g–i) Same as above, but with Gamma scale = $1e^5$. (j) Overall distribution of SNP loadings across all conditions. (k) Linkage disequilibrium (LD) decay pattern.

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) and 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 9). Specifically, we were able to recapitulate a phylogenetic backbone (Figure 9a & b), a PC1 correlated with the trait value (Figure 9e) and a small number of high loading peaks (Figure 9c), each of which is associated with SNPs with substantial influence on trait values. These simulations have slightly weaker LD than is seen in the real data (Figure 9h), with a smaller proportion of the variance explained by PC1. However, the pairwise LD between the highest loading SNPs showed a very similar pattern to that seen in real data (Figure 9d).

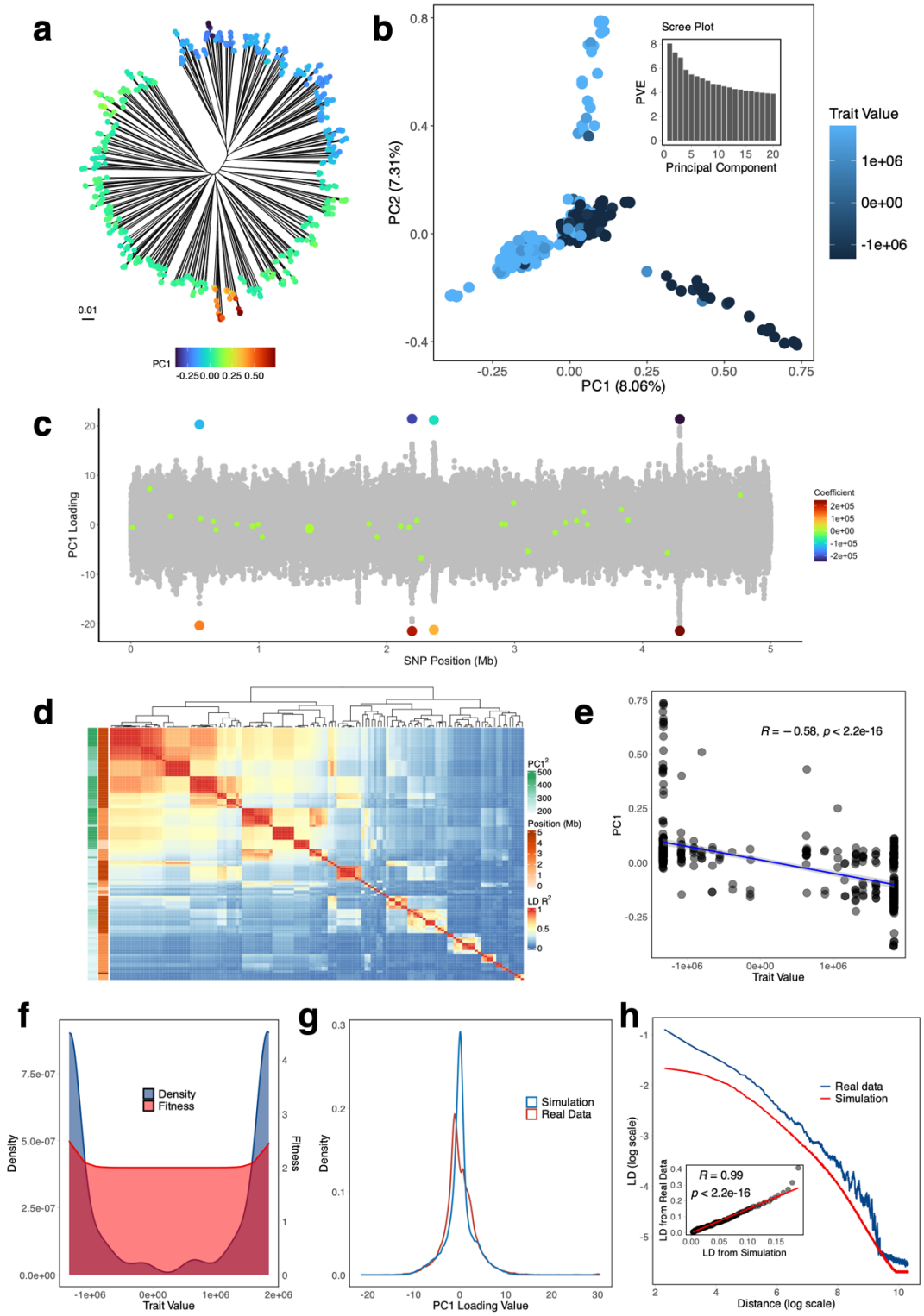


Figure 9. Best-fit simulated diversifying selection model. (a) Phylogenetic tree of simulated 5Mb genomes. (b) PCA results. (c) SNP loadings for PC1. (d) Heatmap of pairwise LD R^2 values for high loading SNPs. (e) Trait value density distributions and fitness function. (f) Correlation between trait

values and strain PC1 values. (g) Distribution of PC1 SNP loadings for the 5 Mb simulation compared to the real data. (h) LD decay curves for the simulated and real datasets. An inset scatter plot compares the LD values between the two datasets, highlighting their correlation.

Discussion

Are quantitative trait models relevant for microbiology?

Quantitative trait models are central to our understanding of human disease genetics[45] and animal[46] and plant[47] biology. These models are applicable to many morphological and disease susceptibility traits, which fit the assumptions of the model, namely that the trait value is dependent on many variants, none of which explain a substantial fraction of the overall variation. Furthermore, each variants acts approximately independently of the others in its influence on the trait value[48,49]. These features allow the traits of offspring to be predicted based on the values found in their parents. The predictability inherent in these models has found several compelling practical applications, including predicting disease risk in humans using Polygenic Risk Scores[50] and to assess the genetic quality of livestock, based on estimated breeding values[51]. The same rules also make short-term evolution relatively predictable[52]. Quantitative genetics can be used for example to predict the change in beak sizes in populations of Darwin's finches based on the intensity of selection and the additive genetic variation present within the population before selection takes place[44].

Bacteria do not undergo sexual reproduction in each generation, but instead reproduce clonally, while occasionally importing tracts of DNA from other strains[53]. The population-level response to selection depends on specific lineages rising in frequency[54]. As a result, the most straightforward applications of quantitative trait models in eukaryotes, such as predicting offspring trait values do not apply[55]. Despite these differences, bacteria do have many continuous-valued traits[56], such as cell size, growth rate and antibiotic minimum inhibitory concentration (MIC). It is of interest to understand the genetic variation of each trait circulating within species and how the trait responds to selection.

Laboratory selection experiments have provided important evidence on the genetic architecture of the response to selection in a novel environment[57]. Bacterial fitness in a particular environment is a continuous valued trait[58]. In asexual laboratory populations, mutations of large effect dominate the initial response, but diminishing returns epistasis sets in rapidly with later variants having progressively smaller effect[59]. Although these laboratory experiments provide excellent evidence on responses to novel environments for

434 clonal populations, we currently have neither data nor theoretical models applicable to
435 natural recombining populations of bacteria. For example, if a lineage is selected for
436 increased cell size, we do not know whether the response is likely to be driven by variants
437 of small or large effect, and whether the variants will principally arise *de novo* or be
438 imported from other members of the population by homologous or non-homologous
439 recombination.

440 To our knowledge, the only continuous valued trait whose variation has been studied
441 extensively in bacterial populations is antibiotic MIC, which can be treated
442 quantitively[60]. Antibiotic resistance is however an unusual trait, because of the episodic
443 nature of exposure to antibiotics and the strong selection coefficients involved, which
444 accentuates the role of variants of large effect[61]. The most important sources of variation
445 are mobile genetic elements carrying antibiotic resistance islands and *de novo* point
446 mutations[62,63]. These variants do not fit the assumptions of quantitative trait models
447 because individual variants have a large effect and moreover, there are strong epistatic
448 interactions between them[64,65]. In any case, when antibiotic selection pressure is applied,
449 instead of increasing MIC across the population, the outcome is some lineages become
450 highly resistant, while others remain susceptible[66].

451 One factor that is likely to be critical in determining the genetic architecture of a trait is the
452 past selection history[67]. For example, if cell size has previously been consistently
453 selected to be smaller within the natural environment, then we might expect the response
454 to selection for larger cell size to be like that seen in laboratory selection
455 environments[68,69], with *de novo* mutation dominating and the initial response generated
456 by mutations of large effect. By contrast, if there has been fluctuating or diversifying
457 selection in the population, then many variants affecting size may already be segregating.
458 These existing variants will have already been tested by selection and will tend to be less
459 disruptive to other features of the organism's biology. Therefore, if we want to find traits
460 that are polygenic within bacterial populations, where several variants behaving additively
461 and most variants having small effect, we might focus on traits that have been under such
462 selection regimes.

463 Here we have used simulation to demonstrate that there are circumstances in which
464 selection on a quantitative trait might structure the genetic variation across an entire
465 bacterial species. Specifically, we have performed simulations of an additive trait in which
466 strains with the highest and lowest values of the trait have a small fitness advantage. In the
467 simulations this tendency is counteracted by recombination, which randomises variation
468 within the population. If the recombination rates at each locus are low, strains are likely to

469 differentiate into two discrete groups with distinct trait values (Figure 7a-c). If
470 recombination rates are higher, then intermediate strains can become common (Figure 7d-
471 f) and variation becomes clinal. Because this cline is generated by selection on a trait, we
472 call it a bacterial ecocline.

473 The clinal variation we see in our simulations depends on a balance between selection and
474 recombination and can potentially be seen throughout the geographical range of the species.
475 As such it is distinct from the clinal variation commonly seen in outbreeding animals and
476 plants, which frequently has a geographical component. Eukaryotic biologists use the term
477 ecocline to describe a situation where there is a single key ecological variable, such as
478 altitude above sea levels for terrestrial animals or saltwater concentration for marine ones,
479 which generates a strong geographical gene frequency cline in the loci that determine
480 adaptation to that variable[70,71]. However, in their context, the gene frequency cline is
481 dependent on limited dispersal between locations, which reduces the rate of recombination
482 between individuals with high and low trait values. We think that our use of the term
483 ecocline is more accurate, while theirs should strictly be called eco-geoclines. Their clines
484 are, however, similar to the one we have simulated in being dependent on a key ecological
485 trait.

486 **Clinal variation in KP**

487 Fitness tradeoffs, such as between fast growth and starvation resistance are important for
488 all bacterial species. However, given our limited understanding of selection in natural
489 bacterial populations, it can be difficult to predict which traits might be subject to
490 diversifying selection. We therefore suggest a "reverse ecology" approach [20,25,72,73] in
491 which we first identify clinal genetic variation in bacterial species, before seeking to
492 identify the trait or traits associated with the cline that are under diversifying selection.

493 Specifically, we noticed that the KP phylogenetic tree is structured by a backbone, which
494 has been noticed informally by many researchers, but resists immediate explanation. To
495 characterize the underlying variation in more detail, we used PCA. When PCA is applied
496 to a non-redundant dataset, PC1 is strongly correlated with the backbone. However, caution
497 must be exercised in interpreting PCA results, because the method will detect apparently
498 clinal variation in essentially any dataset. The PC1 cline in KP is noteworthy because of
499 the large fraction of variation it explains (Figure 1b), its reproducibility in analyses using
500 subsets of the data (Figure 2) and the absence of geographic differentiation in the species.

501 Our PCA results are particularly useful because they give us the first clues about the genetic
502 architecture of the clinal variation. Specifically, PC1 variant loadings have four prominent

503 peaks, corresponding to LD blocks in the genome. In the simulations of selection on a
504 quantitative trait that best fit these data, the Gamma parameter determining variation in
505 trait values is large, meaning that most of the variation is generated by a small number of
506 loci. In these simulations, we found that the highest PC1 loading peaks corresponded to the
507 loci with the largest effect on trait values, but that remaining variation in PC1 appears to
508 be approximately independent of trait values. That is to say loci with small effects on the
509 trait had randomly distributed PC1 values, while most intermediate PC1 peaks correspond
510 to neutral sites (Figure 8).

511 We have also visualized the linkage disequilibrium between these sites (Figure 1d) and
512 found a pattern that is broadly consistent with the patterns seen in simulated ecoclines
513 (Figure 7d-f). In our simulations of diversifying selection, aside from physical linkage
514 blocks, most of the LD between sites is generated by selection on the trait, with pairwise
515 LD values between unlinked blocks dependent on the amount by which each of the markers
516 influences the trait. In the KP data, block 1 has the most extreme PC1 loadings, so that this
517 block is predicted to have the largest effect on the putative quantitative trait. Concordant
518 with this prediction, it has consistently high pairwise LD with loci in other blocks, while
519 blocks 5-44, which have lower loading values, have consistently low pairwise LD values
520 with other blocks. Very similar patterns are seen in the simulated data (Figure 7k). The LD
521 patterns that we see thus seem broadly consistent with selection being mediated by the
522 effect on a trait, rather than direct pairwise interactions between loci, which would generate
523 a more patchwork-quilt like pattern.

524 **Phenotypic and ecological hypotheses about KP variation**

525 Our hypothesis that selection is organizing KP variation into a bacterial ecocline entails
526 two sub-hypotheses. The first is an ecological hypothesis — that diversifying selection on
527 a trait has generated clinal variation — and a functional hypothesis — that the genetic
528 variants showing clinal variation are modifying the selected trait. Demonstration of an
529 ecocline requires validation of both hypotheses.

530 Existing functional annotations of the genes in the top four blocks provide hints, but not
531 strong evidence that the genes might influence the same trait. The most differentiated gene
532 is the adhesin component of Kpa fimbriae, which suggests the trait may be related to
533 adhesion/diffusion. Block 2 encodes a filamentous hemagglutinin, which may also be an
534 adhesin. The EAL- and GGDEF-domain proteins ecnoded in blocks 3a and 3c, respectively
535 are related to c-di-GMP metabolism, a second messenger that controls trade-offs between
536 sessile/surface and motile/free-living behaviours[74]. The accessory *kfuABC* genes in

537 block 3a suggests there is some involvement of iron acquisition[33]. While these genes
538 were previously found to be over-represented in invasive clinical isolates and bovine
539 mastitis strains[32], we could not find any evidence that the trait is directly and specifically
540 involved in human infection. One thing that is clear is that virulent or resistant genotypes
541 can assemble on all positions of the ecocline.

542 We also attempted to identify specific gene function categories that were particularly
543 associated with the backbone. For core genes, there was no difference between COG
544 categories in mean absolute PC1 loading values (Figure 5). These results are consistent
545 with our simulation results, which imply that if a trait is determined principally by a small
546 number of loci, then genes that modify the trait by smaller amounts cannot be distinguished
547 from neutral loci based on PC1 values. Therefore, it seems likely that outside the four peaks,
548 most of the variation in PC1 values in core genome SNPs is generated by genetic drift.
549 However, for accessory gene presence-absence, metabolic and cellular process-associated
550 genes have systematically higher absolute loadings than genes of unknown function or
551 those associated with replication, recombination or transcription, which provides
552 suggestive evidence that these are under stronger selection.

553 Despite the provisional nature of the current evidence, we have been emboldened to
554 propose KP variation might be organized into an ecocline based on our previous finding of
555 "ecospecies" in two others bacterial genera[20,25,26] An ecospecies is like an ecocline, in
556 that genetic differentiation is maintained by selection, in the absence of other barriers to
557 genetic exchange. However, in contrast to bacterial ecoclines, each strain belongs to one
558 ecospecies or another and there are no intermediates. For example, differentiation between
559 Hardy and Ubiquitous *Helicobacter pylori* is largely restricted to 100 genes, which each
560 have nearly fixed differences in at least 5 SNPs and there are no intermediate genotypes[26].
561 In VP differentiation has taken place at 40 core genes, and once again there are no
562 intermediate genotypes[20].

563 For ecospecies, we have also been able to make stronger functional predictions, relating
564 the VP ecospecies to motility[20] and the *H. pylori* ecospecies to host diet[26]. We have
565 also recently used laboratory experiments to confirm and extend the original functional
566 predictions and show that the differentiation between VP Typical and Molassodon
567 ecospecies is functionally coherent, with the Molassodon having evolved a hunt, kill and
568 devour phenotype that it employs in viscous liquids[25]. These results show that the reverse
569 ecology approach can identify loci that act together in natural populations.

570 The absence of equivalent functional predictions means that our current designation of a

571 bacterial ecocline in KP is provisional and needs to be validated by phenotypic data. If the
572 ecological hypothesis is correct, and the genes with extreme loadings are influencing a
573 single trait, then characterization of the effects of these variants should simultaneously
574 provide considerable insight into gene function, ecology and evolution for a key bacterial
575 fitness trait. This would likely represent a breakthrough as these aspects are hard to study
576 individually, handsomely justifying the initial effort required to identify the trait.

577 It should be cautioned however that diversifying selection may not reflect a single
578 measurable trait. For example, distinct foraging strategies in Darwin's finches might
579 require different beak sizes but also different behaviors, which may also be genetically
580 determined. This could create linkage disequilibrium between genes determining beak size
581 and foraging behavior. This ecologically-generated linkage disequilibrium is also evident
582 in the Molassodon ecospecies example[25], where genes responsible for motility, bacterial
583 killing and nutrient uptake have coevolved. In these cases, it might be hard to find a
584 laboratory condition that links together the function of the different genes involved in the
585 ecocline.

586 A challenge and virtue of the reverse ecology approach[25,72,73] is that it confronts us
587 with the gulf between laboratory conditions and the natural environment that we seek to
588 understand, and more specifically our difficulty in putting ourselves in the same shoes as
589 bacteria and predicting the ecological challenges they are likely to face. In any case, we
590 propose that investigation of phenotypic differences associated with the PC1 cline has the
591 potential to deliver key insights into KP ecology, as well as on the biology of quantitative
592 traits in microbes.

593

Methods

Genome collection and sequence alignment

We collected 1,421 KP genomes from the NCBI database, selected according to iterative-PopPUNK test dataset[78], along with a reference genome (KP strain NTUH-K2044, NC_012731.1). Additionally, we obtained genome metadata using the NCBI Datasets command-line tools (v16.12.1)[76] and used Kleborate (v2.4.0)[23] to extract clinically relevant genotypic information on strains. All genomes were aligned to the NTUH-K2044 (NC_012731.1) reference genome using Nucmer (v3.1)[61] for reference-based alignment.

In total, 8,331 global VP genomes were downloaded from the NCBI database, based on the dataset from the study 'Why panmictic bacteria are rare'[77] and aligned to the VP RIMD 2210633 (RefSeq NC_004603.1 and NC_004605.1)[78] using Nucmer (v3.1)[79], following the approach described in previous VP studies[19,20].

Core genomic variant calling and definition of non-redundant dataset

SNP-sites (v2.5.2)[80] was used to call all the variants from the whole KP dataset alignment. Core SNPs were defined as the SNPs 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)[81]. A non-redundant dataset of 369 genomes was defined by selecting genomes with pairwise SNP distances >20 kb (Supplementary Table 2).

The same method to the whole VP dataset, using a 50 kb SNP distance threshold to reduce redundancy, yielding 667 genomes (Supplementary Table 3) with 1,619,597 core SNPs.

fineSTRUCTURE and population assignment

To further investigate the population structure of the non-redundant 369 KP strains described above, we performed chromosome painting and fineSTRUCTURE analysis. Using ChromoPainter mode in fineSTRUCTURE (v4.1.0)[82], 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). 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 VP dataset, we followed the same

625 fineSTRUCTURE approach using haplotype data from core SNPs. Based on the
626 fineSTRUCTURE results, the genomes were classified into five populations: VppAsia
627 (506 strains), Hybrid (7 strains), VppUS1 (50 strains), VppUS2 (30 strains), and VppX (74
628 strains). Within the VppAsia dataset, strains were further categorized into ecospecies types
629 (EG1a and non-EG1a) based on 27 differentiating genes[20].

630 **Construction of phylogenetic trees**

631 Based on the core SNPs defined above, we constructed nucleotide sequence alignment files
632 for both the entire KP dataset and the non-redundant KP dataset. Phylogenetic trees were
633 constructed using FastTree (v2.1.10)[83] with the "-nt" option under the default model. The
634 trees were visualized and compared using the ggtree package (v3.12.0)[84] in R.

635 Similarly, we used the same software and parameters to construct phylogenetic trees for
636 the non-redundant VP and VppAsia datasets, as well as the simulated datasets. During
637 Partial-Variants Matching, we also constructed phylogenetic trees based on a subset of
638 SNPs from the real datasets.

639 **Principal Component Analysis**

640 Before performing PCA (Principal Component Analysis) on the non-redundant KP dataset,
641 PLINK (v1.9)[85] was used to filter out loci with a minimum allele frequency (MAF)
642 below 2% from the core SNPs, leaving 102,226 variants. Linkage disequilibrium (LD)
643 pruning was subsequently performed to remove linked SNPs, where one SNP from each
644 pair was removed if the LD (R^2) exceeded 0.1 within a 50 kb window. The window was
645 then shifted forward by 10 bp, and the procedure repeated to ensure comprehensive
646 coverage. A total of 33,589 pruned variants were obtained and used for conducting PCA on
647 the non-redundant dataset with PLINK[85], extracting and reporting the first 20 principal
648 components by default.

649 Similarly, we used the same software and parameters to perform PCA on the non-redundant
650 VP dataset, which included 220,737 SNPs after MAF filtering and 69,537 SNPs after
651 pruning. We also applied PCA to the VppAsia dataset, containing 207,292 SNPs after MAF
652 filtering and 65,058 SNPs after pruning. For the simulated datasets, we only applied MAF
653 filtering to retain enough SNPs, followed directly by PCA without pruning.

654 **Variant loadings calculation**

655 We first calculated the correlation between the original variables and the principal
656 components, which is specifically referred to as the PCA loading for strains. The loadings

657 can be determined as the coordinates of the variables (eigenvectors) multiplied by the
658 square root of the eigenvalues associated with that principal component (Equation 1).
659 When performing PCA, PLINK[85] generates output files ending with eigenval and
660 eigenvec. Next, we transform the MAF-filtered SNP matrix into a binary matrix, assigning
661 a value of 1 to minor alleles and 0 to all others (including the third and fourth most frequent
662 alleles at non-biallelic sites). Finally, through matrix multiplication, we computed the
663 product of the core SNP matrix and the PCA loading matrix to obtain the loading values
664 for the MAF-filtered SNPs (Equation 2).

665 $$strainloadings = eigenvectors \cdot \sqrt{eigenvalues} \quad \text{(Equation 1)}$$

666 $$SNPloadings = strainloadings \cdot SNPmatrix \quad \text{(Equation 2)}$$

667 To better highlight the significant peaks in the non-redundant KP result, we plotted the
668 squared PC1 SNP loading values and marked the dots with squared loading values greater
669 than 300.

670 Similarly, we used the same method to calculate SNP loadings for the non-redundant VP
671 dataset, the VppAsia dataset, and the simulated datasets.

672 **Pangenome analysis**

673 To estimate the gene content, we annotated all KP genomes with prokka (v1.13)[86]
674 software, using a curated version of the NC_012731.1 (NTUH_K2044) proteome as the
675 primary annotation source. The Generic Feature Format (GFF) outputs generated by
676 prokka were then used as the input files for panaroo (v1.2.8)[27] with the strict model. The
677 "--remove-invalid-genes" and "--merge_paralogs" options were also employed during the
678 analysis. Next, we extracted the non-redundant dataset from the overall results and defined
679 5,558 accessory genes as those present in between 2% and 98% of the genomes in the non-
680 redundant dataset. Using the same approach (Equation 2), we also calculated the gene
681 loadings and defined genes with squared PC1 loadings greater than 300 as high-loading
682 genes.

683 **Calculation of linkage disequilibrium decay**

684 To analyze the decay of linkage disequilibrium (LD) in KP, we included 102,226 MAF-
685 filtered SNPs and used the Haploview (v4.2)[87] software to compute R^2 values. When
686 running Haploview, we set "-maxdistance 30" to ensure that the maximum marker distance
687 for R^2 calculation did not exceed 30 kb. Additionally, we set "-minMAF 0" and "-hwcutoff

688 0" to exclude any filtering based on minor allele frequency and Hardy-Weinberg
689 equilibrium testing, ensuring that all input SNPs were included in the calculations. The
690 average R^2 value was calculated using a 10 bp step size.

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

693 We also calculated the LD between the high-loading variants defined above (340 variants
694 in total, including core SNPs and accessory genes) used Equation 3 in R. Each element R_{ij}^2
695 of the matrix quantifies the degree of linkage disequilibrium between variant *i* and variant
696 *j*, providing a comprehensive view of the genomic associations among core SNPs and
697 accessory genes. Based on manual inspection, we removed an accessory gene (group_205)
698 that represented fragmented versions of KP1_1982 (339 variants left), which is visualized
699 in a heatmap, generated using the ComplexHeatmap package[88].

700 $$R^2 = \frac{\left(f_{gene_i,gene_j} - f_{gene_i} f_{gene_j}\right)^2}{f_{gene_i} f_{gene_i'} f_{gene_j} f_{gene_j'}} \quad \text{(Equation 3)}$$

701 Based on the LD heatmap of all high-loading variants, we performed hierarchical clustering
702 to divide the heatmap into 44 blocks. Among these, four larger blocks contain multiple
703 variants, while the remaining 40 smaller blocks each contain only a few variants. Notably,
704 block 3 was further subdivided into 3a, 3b, and 3c based on the physical locations of the
705 mutations on the reference genome. We observed that some SNPs in block 3 are located
706 within the KP1_RS30360 and KP1_RS30365 genes in the reference genome. According to
707 the annotation file of the reference genome, KP1_RS30365 is classified as a pseudogene.
708 However, we do not concur with this interpretation. We believe that KP1_RS30360 and
709 KP1_RS30365 together represent a complete gene, with some strains (*e.g.*, MGH 78578)
710 containing only a fragment of this gene. the differentiation at this locus also likely
711 represents a functional presence/absence rather than functional differentiation via SNPs.

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

717 **Phylogenetic reconstruction of a gene across multiple *Klebsiella* subspecies**

718 Genomes were retrieved from the Pasteur Institute Klebsiella MLST Database by applying
719 the following filtering criteria: (i) the taxonomic designation must be one of the following
720 subspecies: *K. quasipneumoniae subsp. quasipneumoniae*, *K. quasipneumoniae subsp.*
721 *similipneumoniae*, *K. variicola subsp. variicola*, *K. variicola subsp. tropica*, *K.*
722 *quasivariicola*, or *K. africana*; (ii) the annotation status for MLST, ribosomal MLST, and
723 scgMLST629_S must all be labeled as good; and (iii) assembly quality checks including
724 the number of contigs, assembly size, minimum N50, and %GC must all be pass.

725 To confirm species identity, the resulting genomes were further screened using
726 Kleborate[23]. The gene KP1_0338 from the reference strain was used as a query in a
727 BLAST search, and MGH 78578 was included as an additional representative strain. To
728 avoid overrepresentation of any subspecies, 10 genomes were randomly selected from each
729 subspecies identified in the BLAST results.

730 These selected genomes, together with the reference strain, MGH 78578, and the 5 strains
731 with the highest and 5 strains with lowest PC1 loadings from the previously defined non-
732 redundant dataset, were aligned using MAFFT[89]. The resulting multiple sequence
733 alignment was then used to construct a phylogenetic tree using FastTree[83].

734 **Functional annotation**

735 To further characterize the functional content in KP, we annotated both the MAF-filtered
736 SNPs (MAF > 0.02) with SnpEff (v4.3t)[90], using the reference genome NTUH-K2044
737 (NC_012731.1), and applied custom filtering to focus on coding region variants by
738 disabling annotation for upstream, downstream, and UTR (untranslated) regions. We then
739 obtained the types of the SNPs and determined whether they were synonymous mutations.
740 For accessory genes, we used the eggNOG-mapper[91] online tool with its default settings
741 to ensure a comprehensive and standardized annotation.

742 For the identified high-loading variants (squared PC1 SNP loading > 300), we further
743 mapped these variants to the reference genomes of KP strains MGH 78578 (NC_009648.1)
744 using BLASTn[92]. The mapping aimed to identify their genomic locations and potential
745 functional implications in these strains. Additionally, we searched the UniProt[93] database
746 to retrieve gene names and associated functional annotations for the high-loading variants.

747 We calculated the mean absolute PC1 loading for accessory genes within each COG
748 category, as well as the overall mean absolute loading across all accessory genes. For core
749 SNPs, we first computed the mean absolute PC1 loading of SNPs belonging to each core
750 gene, then calculated the average of these values for each COG category. In addition, we

751 determined the overall mean absolute loading across all core genes.

752 We visualized the results using bar plots, where each bar represents the mean absolute PC1
753 loading of genes (core or accessory) within a given COG category, and the error bars
754 indicate the standard error of the mean. A solid line denotes the overall mean absolute PC1
755 loading across all accessory genes (or core genes), serving as a reference.

756 **Isolation background characteristics**

757 To investigate potential host specificity in KP, and to address the overrepresentation of
758 human-derived isolates in our primary dataset, we incorporated a multi-host dataset from
759 the Pasteur Institute *Klebsiella* MLST Database (https://bigsdb.pasteur.fr/cgi-
760 bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_isolates). Isolates were selected based on the
761 following quality criteria: chromosomal assembly size between 5 and 6 Mb, fewer than
762 100 contigs, an N50 greater than 20,000 bp, fewer than 1,000 ambiguous bases (Ns), and
763 a GC content between 55% and 60%. Additionally, only genomes with "good" annotation
764 status for MLST, ribosomal MLST, and SCGMLST629_s schemes were included.
765 Kleborate[23] was applied to refine species classification, and *Klebsiella pneumoniae*
766 species complex (KpSC) isolates were excluded. To minimize overrepresentation of
767 human-associated strains, genomes with host information explicitly annotated as "human"
768 were excluded. However, a small number of human-derived isolates may still be present
769 due to inconsistent or alternative host labels (e.g., *Homo sapiens*). (Supplementary Table
770 4).

771 We used the KP1_0338 gene from the reference strain as the query sequence and included
772 MGH 78578 as an additional representative. These two strains represent the extremes along
773 the PC1 axis in our population structure analysis, with the reference strain showing an
774 extremely high PC1 value and MGH 78578 a low one, thus capturing the two major
775 variants of KP1_0338 present in the KP population. Using BLASTn[92], we identified
776 KP1_0338 homologs in the multi-host dataset. The retrieved sequences were aligned using
777 MAFFT[89], and a phylogenetic tree was constructed with FastTree[83] to examine
778 sequence diversity. Based on the topology of the KP1_0338 phylogeny, isolates were
779 grouped into two distinct clusters, designated as High and Low types.

780 To assess host-specific enrichment of these gene variants, we performed a one-vs-rest
781 Fisher's exact test for each host. This test compared the counts of High and Low types
782 within each host to those in all other hosts combined.

783 We projected the combined dataset—including the non-redundant dataset as well as strains

784 isolated from water and horse sources—onto the principal component space defined by the
785 non-redundant dataset. For each group (non-redundant, water-derived, and horse-derived),
786 we calculated the mean PC1 loading values and visualized them with standard deviations
787 as error bars. Pairwise comparisons between groups were performed using two-sample t-
788 tests to assess the statistical significance of differences in PC1 loadings.

789 **Half-matching testing**

790 To assess the stability and repeatability of PC1 in KP, we used three half-matching tests,
791 each designed to evaluate the consistency of PC1 under varying conditions.

792 **1. Half-Strain Matching**

793 Firstly, we divided the non-redundant KP dataset into two groups based on the backbone
794 in the phylogenetic tree. We then performed PCA on the first group using the pruned SNPs.
795 Afterward, we calculated the MAF-filtered SNP loadings using the above steps.
796 Subsequently, we obtained the product for the second group through matrix multiplication,
797 representing the mapping of the PCA results from the first group onto the second group.
798 We then calculated the correlation for the PC1 between the mapped results of the first group
799 and the PCA results of the second group. Conversely, the same matching analysis was
800 performed on the data from the second group.

801 **2. Half-Genome Matching**

802 KP has a circular chromosome with a length of approximately 5 to 6 Mb, and the reference
803 genome we used is approximately 5.2 Mb in length. We divided the circular chromosome
804 into two DNA fragments (Fragment 1 (1 bp–2.6 Mb) and Fragment 2 (2.6 Mb–5.2 Mb),
805 each containing approximately half of the chromosome. Specifically, the first dataset
806 comprised the pruned SNPs in the first half of the chromosome, while the second dataset
807 encompassed the unlinked SNPs from the second half. We separately performed PCA on
808 these two datasets using PLINK[85]. Finally, we calculated the correlation between the
809 PC1 of the two fragments.

810 **3. Partial-Variants Matching**

811 First, we assigned maximum loading values to the MAF-filtered SNPs based on their
812 squared PC1 loadings, using a 50 kb window size with a 1 bp step size. This process
813 identified the maximum squared PC1 loadings within each 50 kb vicinity. Based on the
814 distribution of maximum squared PC1 loadings, the SNPs were categorized into three types:
815 low-loading SNPs, medium-loading SNPs, and high-loading SNPs. We separately

816 conducted PCA on the high-loading SNP dataset and low-loading SNP dataset, by
817 extracting unlinked SNPs from each. Subsequently, we calculated the correlation between
818 the PC1 of the 2 datasets and constructed the phylogenetic tree for the 2 datasets. In this
819 test, we ignored SNPs with intermediate loading values.

820 Similarly, the three types of half-matching tests were applied to the non-redundant clonal
821 population simulation (Clonal Haploid Bacteria Model with Horizontal Gene Transfer,
822 SNP distance > 2 kb, shown in Supplementary Figure 4a), the non-redundant VP dataset
823 and the VppAsia dataset. However, in the half-strain matching test, although we still
824 performed splits along the phylogenetic tree, these datasets were not always divided into
825 left-hand and right-hand subsets in the trees. Therefore, in these datasets, we referred to the
826 two subsets as half samples 1 and half samples 2. Additionally, in VP, which contains two
827 chromosomes, the half-genome matching is implemented by separating the data according
828 to chromosome number rather than splitting each chromosome into two segments.

829 **Simulation of bacterial populations**

830 We used SLiM (v4.0.1)[94] to simulate the evolutionary dynamics of bacterial populations
831 under different scenarios using a non-Wright–Fisher (nonWF) framework[95]. All
832 scenarios started with identical initial conditions: each bacterial genome consisted of a
833 single 1 Mb haploid chromosome, with a mutation rate of 10^{-6} per site per generation,
834 approximately ten times higher than the estimated mutation rate for KP[92]. Horizontal gene
835 transfer (HGT) occurred at a rate of 1.0, ensuring that every individual experienced HGT
836 in each generation. The simulations ran for 10,000 generations. Two mutation types, "m1"
837 and "m2", occurred at each site with equal probability (1:1) and were assumed to be neutral.
838 However, in the final simulation, we adjusted the chromosome length to closely match the
839 length of a real KP genome.

840 **Modeling clonal haploid bacteria with horizontal gene transfer**

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

sizes of 1,000, 2,000, and 10,000.

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.

Selective model

In the third simulation scenario, we focused on a population of 10,000 clonal haploid bacteria, where generations 1–5000 evolved neutrally, and selective pressure was introduced from generations 5001–10,000. Horizontal gene transfer (HGT) tract lengths were set to 1 kb, 10 kb, 20 kb, 50 kb, and 100 kb. 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. Each strain was assigned a score, termed a quantitative trait value, based on the difference in the number of "m1" and "m2" (score $= \sum \mathrm{m1} - \sum \mathrm{m2}$), and fitness was determined by a normalized power function of the deviation from the mean score ($fitness = 2.0 + 1.0 \times (|score - \mu|^5)/(max(|score - \mu|^5) + 10^{-6})$, where μ is the mean score across all individuals.). Based on the trait value–derived fitness (ranging from 2.0 to 3.0), individuals with more extreme trait values had higher fitness. The number of offspring produced by each individual followed a Poisson distribution with λ = fitness, such that individuals with extreme trait values tended to produce more offspring on average. This setting imposed a diversifying selection pressure, favoring individuals at the tails of the trait value distribution.

Refined selective model

We refined the selective model by introducing 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. 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:

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

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 = 20 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, 1×10^5, and 1×10^6. The timing of selection and the fitness function were the same as those used in the third simulation scenario.

To better approximate the real KP genome, we adjusted the chromosome length to 5 Mb, set the mutation rate to 10^{-6}, and increased the number of HGT events to 5, , with each event transferring 20 kb, resulting in a total of 100 kb HGT per generation between each bacterium and others in the population. 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. Fitness was determined by a normalized power function of the deviation from the mean score ($fitness = 2.0 + 0.5 \times (|score - \mu|^{10})/(max(|score - \mu|^{10}) + 10^{-6})$, where μ is the mean score across all individuals.). Based on the trait value–derived fitness (ranging from 2.0 to 2.5), individuals with more extreme trait values had higher fitness.

All the simulation 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. We thank many colleagues for feedback on a previous version of the manuscript.

References

1. St. John K. Review Paper: The Shape of Phylogenetic Treespace. Syst Biol. 2017;66: e83–e94. doi:10.1093/sysbio/syw025
2. Mooers AØ, Heard SB. Using Tree Shape. Syst Biol. 2002;51: 833–834. doi:10.1080/10635150290102528
3. Feil EJ, Holmes EC, Bessen DE, Chan M-S, Day NPJ, Enright MC, et al. Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci. 2001;98: 182–187. doi:10.1073/pnas.98.1.182
4. Fahimipour AK, Gross T. Mapping the bacterial metabolic niche space. Nat Commun. 2020;11: 4887. doi:10.1038/s41467-020-18695-z
5. Liu Y, Huang L, Cai J, Zhu H, Li J, Yu Y, et al. Clinical characteristics of respiratory tract infection caused by Klebsiella pneumoniae in immunocompromised patients: a retrospective cohort study. Front Cell Infect Microbiol. 2023;13: 1137664. doi:10.3389/fcimb.2023.1137664
6. Brown C, Seidler RJ. Potential Pathogens in the Environment: Klebsiella pneumoniae, a Taxonomic and Ecological Enigma. Appl Microbiol. 1973;25: 900–904. doi:10.1128/am.25.6.900-904.1973
7. Zhang Q, Yang J, Zhou X, Ding Y, Wang Y, Zhu X, et al. Soilborne bacterium Klebsiella pneumoniae promotes cluster root formation in white lupin through ethylene mediation. New Phytol. 2023;237: 1320–1332. doi:10.1111/nph.18600
8. Di Cesare A, Cornacchia A, Sbaffi T, Sabatino R, Corno G, Cammà C, et al. Treated wastewater: A hotspot for multidrug- and colistin-resistant Klebsiella pneumoniae. Environ Pollut Barking Essex 1987. 2024;359: 124598. doi:10.1016/j.envpol.2024.124598
9. Håkonsholm F, Hetland MAK, Svanevik CS, Lunestad BT, Löhr IH, Marathe NP. Insights into the genetic diversity, antibiotic resistance and pathogenic potential of *Klebsiella pneumoniae* from the Norwegian marine environment using whole-genome analysis. Int J Hyg Environ Health. 2022;242: 113967. doi:10.1016/j.ijheh.2022.113967
10. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad Sci U S A. 2015;112: E3574–E3581. doi:10.1073/pnas.1501049112
11. Budia-Silva M, Kostyanev T, Ayala-Montaño S, Bravo-Ferrer Acosta J, Garcia-Castillo M, Cantón R, et al. International and regional spread of carbapenem-resistant Klebsiella pneumoniae in Europe. Nat Commun. 2024;15: 5092. doi:10.1038/s41467-024-49349-z
12. Madni O, Amoako DG, Abia ALK, Rout J, Essack SY. Genomic Investigation of Carbapenem-Resistant Klebsiella pneumonia Colonization in an Intensive Care Unit in South Africa. Genes. 2021;12: 951. doi:10.3390/genes12070951
13. Wyres KL, Nguyen TNT, Lam MMC, Judd LM, van Vinh Chau N, Dance DAB, et al. Genomic surveillance for hypervirulence and multi-drug resistance in invasive Klebsiella pneumoniae from South and Southeast Asia. Genome Med. 2020;12: 11. doi:10.1186/s13073-019-0706-y
14. Futuyma DJ. Evolution. Third edition. Sunderland, Massachusetts U.S.A: Sinauer

956 Associates, Inc. Publishers; 2013.
957 15. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland, Mass:
958 Sinauer Assoc; 1998.
959 16. Mauro AA, Ghalambor CK. Trade-offs, Pleiotropy, and Shared Molecular Pathways:
960 A Unified View of Constraints on Adaptation. Integr Comp Biol. 2020;60: 332–347.
961 doi:10.1093/icb/icaa056
962 17. Abaza S. What is and why do we have to know the phylogenetic tree? Parasitol
963 United J. 2020 [cited 2 Dec 2024]. doi:10.21608/puj.2020.35843.1082
964 18. Wyres KL, Lam MMC, Holt KE. Population genomics of Klebsiella pneumoniae.
965 Nat Rev Microbiol. 2020;18: 344–359. doi:10.1038/s41579-019-0315-1
966 19. Yang C, Pei X, Wu Y, Yan L, Yan Y, Song Y, et al. Recent mixing of Vibrio
967 parahaemolyticus populations. ISME J. 2019;13: 2578–2588. doi:10.1038/s41396-019-
968 0461-5
969 20. Cui Y, Yang C, Qiu H, Wang H, Yang R, Falush D. The landscape of coadaptation
970 in Vibrio parahaemolyticus. Rainey PB, Soldati-Favre D, editors. eLife. 2020;9: e54136.
971 doi:10.7554/eLife.54136
972 21. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26: 303–
973 304. doi:10.1038/nbt0308-303
974 22. Jolliffe IT, Cadima J. Principal component analysis: a review and recent
975 developments. Philos Trans R Soc Math Phys Eng Sci. 2016;374: 20150202.
976 doi:10.1098/rsta.2015.0202
977 23. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. A genomic
978 surveillance framework and genotyping tool for Klebsiella pneumoniae and its related
979 species complex. Nat Commun. 2021;12: 4188. doi:10.1038/s41467-021-24448-3
980 24. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic
981 studies are highly biased and must be reevaluated. Sci Rep. 2022;12: 14683.
982 doi:10.1038/s41598-022-14395-4
983 25. Svensson SL, Wen H, Jia Z, Xie Z, Xu W, Fu H, et al. Evolution of hunt, kill, devour
984 in a Vibrio ecospecies. bioRxiv; 2025. p. 2025.06.27.661923.
985 doi:10.1101/2025.06.27.661923
986 26. Tourrette E, Torres RC, Svensson SL, Matsumoto T, Miftahussurur M, Fauzia KA,
987 et al. An ancient ecospecies of Helicobacter pylori. Nature. 2024;635: 178–185.
988 doi:10.1038/s41586-024-07991-z
989 27. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al.
990 Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol.
991 2020;21: 180. doi:10.1186/s13059-020-02090-4
992 28. Wu C-C, Huang Y-J, Fung C-P, Peng H-L. Regulation of the Klebsiella pneumoniae
993 Kpc fimbriae by the site-specific recombinase KpcI. Microbiol Read Engl. 2010;156:
994 1983–1992. doi:10.1099/mic.0.038158-0
995 29. Potter RF, Lainhart W, Twentyman J, Wallace MA, Wang B, Burnham C-AD, et al.
996 Population Structure, Antibiotic Resistance, and Uropathogenicity of Klebsiella variicola.
997 mBio. 2018;9: 10.1128/mbio.02481-18. doi:10.1128/mbio.02481-18
998 30. Khater F, Balestrino D, Charbonnel N, Dufayard JF, Brisse S, Forestier C. In silico
999 analysis of usher encoding genes in Klebsiella pneumoniae and characterization of their
1000 role in adhesion and colonization. PloS One. 2015;10: e0116215.
1001 doi:10.1371/journal.pone.0116215

1002 31. Ma L-C, Fang C-T, Lee C-Z, Shun C-T, Wang J-T. Genomic Heterogeneity in
1003 Klebsiella pneumoniae Strains Is Associated with Primary Pyogenic Liver Abscess and
1004 Metastatic Infection. J Infect Dis. 2005;192: 117–128. doi:10.1086/430619
1005 32. Yang Y, Higgins CH, Rehman I, Galvao KN, Brito IL, Bicalho ML, et al. Genomic
1006 Diversity, Virulence, and Antimicrobial Resistance of Klebsiella pneumoniae Strains
1007 from Cows and Humans. Appl Environ Microbiol. 2019;85: e02654-18.
1008 doi:10.1128/AEM.02654-18
1009 33. Zhao Q, Yan J, Wang J, Liu R, Bartlam M. Structural analysis of the ferric-binding
1010 protein KfuA from Klebsiella pneumoniae. Biochem Biophys Res Commun. 2023;679:
1011 52–57. doi:10.1016/j.bbrc.2023.08.068
1012 34. Richardson EJ, Bacigalupe R, Harrison EM, Weinert LA, Lycett S, Vrieling M, et al.
1013 Gene exchange drives the ecological success of a multi-host bacterial pathogen. Nat Ecol
1014 Evol. 2018;2: 1468–1478. doi:10.1038/s41559-018-0617-0
1015 35. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host
1016 adaptation. Nat Rev Genet. 2018;19: 549–565. doi:10.1038/s41576-018-0032-z
1017 36. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M. Mismatch
1018 induced speciation in Salmonella: model and data. Philos Trans R Soc Lond B Biol Sci.
1019 2006;361: 2045–2053. doi:10.1098/rstb.2006.1925
1020 37. Didelot X, Méric G, Falush D, Darling AE. Impact of homologous and non-
1021 homologous recombination in the genomic evolution of Escherichia coli. BMC
1022 Genomics. 2012;13: 256. doi:10.1186/1471-2164-13-256
1023 38. Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, et al. Identifying
1024 the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into
1025 bacterial systematics. Proc Natl Acad Sci U S A. 2008;105: 2504–2509.
1026 doi:10.1073/pnas.0712205105
1027 39. Brown T, Didelot X, Wilson DJ, Maio ND. SimBac: simulation of whole bacterial
1028 genomes with homologous recombination. Microb Genomics. 2016;2: e000044.
1029 doi:10.1099/mgen.0.000044
1030 40. Thorell K, Muñoz-Ramírez ZY, Wang D, Sandoval-Motta S, Boscolo Agostini R,
1031 Ghirotto S, et al. The Helicobacter pylori Genome Project: insights into H. pylori
1032 population structure from analysis of a worldwide collection of complete genomes. Nat
1033 Commun. 2023;14: 8184. doi:10.1038/s41467-023-43562-y
1034 41. Yang J, Li Y, Tang N, Li J, Zhou J, Lu S, et al. The human gut serves as a reservoir
1035 of hypervirulent Klebsiella pneumoniae. Gut Microbes. 2022;14: 2114739.
1036 doi:10.1080/19490976.2022.2114739
1037 42. Heng H, Yang X, Ye L, Tang Y, Guo Z, Li J, et al. Global genomic profiling of
1038 Klebsiella pneumoniae: A spatio-temporal population structure analysis. Int J Antimicrob
1039 Agents. 2024;63: 107055. doi:10.1016/j.ijantimicag.2023.107055
1040 43. Anderson JT, Wagner MR, Rushworth CA, Prasad KVSK, Mitchell-Olds T. The
1041 evolution of quantitative traits in complex environments. Heredity. 2014;112: 4–12.
1042 doi:10.1038/hdy.2013.33
1043 44. Hendry AP, Huber SK, De León LF, Herrel A, Podos J. Disruptive selection in a
1044 bimodal population of Darwin's finches. Proc R Soc B Biol Sci. 2009;276: 753–759.
1045 doi:10.1098/rspb.2008.1321
1046 45. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery:
1047 Realizing the promise. Am J Hum Genet. 2023;110: 179–194.

1048 doi:10.1016/j.ajhg.2022.12.011
1049 46. Biscarini F, Nicolazzi EL, Stella A, Boettcher PJ, Gandini G. Challenges and
1050 opportunities in genetic improvement of local livestock breeds. Front Genet. 2015;6: 33.
1051 doi:10.3389/fgene.2015.00033
1052 47. Bernardo R. Reinventing quantitative genetics for plant breeding: something old,
1053 something new, something borrowed, something BLUE. Heredity. 2020;125: 375–385.
1054 doi:10.1038/s41437-020-0312-1
1055 48. Mackay TF, Moore JH. Why epistasis is important for tackling complex human
1056 disease genetics. Genome Med. 2014;6: 42. doi:10.1186/gm561
1057 49. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al.
1058 Common SNPs explain a large proportion of the heritability for human height. Nat Genet.
1059 2010;42: 565–569. doi:10.1038/ng.608
1060 50. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical
1061 instruments. Genome Med. 2020;12: 44. doi:10.1186/s13073-020-00742-5
1062 51. Sharko FS, Khatib A, Prokhortchouk EB. Genomic Estimated Breeding Value of
1063 Milk Performance and Fertility Traits in the Russian Black-and-White Cattle Population.
1064 Acta Naturae. 2022;14: 109–122. doi:10.32607/actanaturae.11648
1065 52. Nosil P, Villoutreix R, de Carvalho CF, Farkas TE, Soria-Carrasco V, Feder JL, et al.
1066 Natural selection and the predictability of evolution in Timema stick insects. Science.
1067 2018;359: 765–770. doi:10.1126/science.aap9125
1068 53. Shapiro BJ. How clonal are bacteria over time? Curr Opin Microbiol. 2016;31: 116–
1069 123. doi:10.1016/j.mib.2016.03.013
1070 54. Keenan JD, Klugman KP, McGee L, Vidal JE, Chochua S, Hawkins P, et al.
1071 Evidence for Clonal Expansion After Antibiotic Selection Pressure: Pneumococcal
1072 Multilocus Sequence Types Before and After Mass Azithromycin Treatments. J Infect
1073 Dis. 2015;211: 988–994. doi:10.1093/infdis/jiu552
1074 55. Lang A, Thomas Beatty J, Rice PA. Guest editorial: Mobile genetic elements and
1075 horizontal gene transfer in prokaryotes. Curr Opin Microbiol. 2017;38: v–vii.
1076 doi:10.1016/j.mib.2017.09.018
1077 56. Westoby M, Nielsen DA, Gillings MR, Gumerov VM, Madin JS, Paulsen IT, et al.
1078 Strategic traits of bacteria and archaea vary widely within substrate-use groups. FEMS
1079 Microbiol Ecol. 2021;97: fiab142. doi:10.1093/femsec/fiab142
1080 57. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, et al.
1081 Tempo and mode of genome evolution in a 50,000-generation experiment. Nature.
1082 2016;536: 165–170. doi:10.1038/nature18959
1083 58. Alonso-del Valle A, León-Sampedro R, Rodríguez-Beltrán J, DelaFuente J,
1084 Hernández-García M, Ruiz-Garbajosa P, et al. Variability of plasmid fitness effects
1085 contributes to plasmid persistence in bacterial communities. Nat Commun. 2021;12:
1086 2653. doi:10.1038/s41467-021-22849-y
1087 59. Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, et al.
1088 Weak Epistasis May Drive Adaptation in Recombining Bacteria. Genetics. 2018;208:
1089 1247–1260. doi:10.1534/genetics.117.300662
1090 60. Mai TT, Lees JA, Gladstone RA, Corander J. Inferring the heritability of bacterial
1091 traits in the era of machine learning. Bioinforma Adv. 2023;3: vbad027.
1092 doi:10.1093/bioadv/vbad027
1093 61. Lehtinen S, Blanquart F, Croucher NJ, Turner P, Lipsitch M, Fraser C. Evolution of

1094 antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of
1095 carriage. Proc Natl Acad Sci. 2017;114: 1075–1080. doi:10.1073/pnas.1617849114
1096 62. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated
1097 with Antimicrobial Resistance. Clin Microbiol Rev. 2018;31: e00088-17.
1098 doi:10.1128/CMR.00088-17
1099 63. Alam MK, Alhhazmi A, DeCoteau JF, Luo Y, Geyer CR. RecA Inhibitors Potentiate
1100 Antibiotic Activity and Block Evolution of Antibiotic Resistance. Cell Chem Biol.
1101 2016;23: 381–391. doi:10.1016/j.chembiol.2016.02.010
1102 64. Vogwill T, Kojadinovic M, MacLean RC. Epistasis between antibiotic resistance
1103 mutations and genetic background shape the fitness effect of resistance across species of
1104 Pseudomonas. Proc Biol Sci. 2016;283: 20160151. doi:10.1098/rspb.2016.0151
1105 65. Schubert B, Maddamsetti R, Nyman J, Farhat MR, Marks DS. Genome-wide
1106 discovery of epistatic loci affecting antibiotic resistance in Neisseria gonorrhoeae using
1107 evolutionary couplings. Nat Microbiol. 2019;4: 328–338. doi:10.1038/s41564-018-0309-
1108 1
1109 66. Colijn C, Cohen T, Fraser C, Hanage W, Goldstein E, Givon-Lavi N, et al. What is
1110 the mechanism for persistent coexistence of drug-susceptible and drug-resistant strains of
1111 Streptococcus pneumoniae? J R Soc Interface. 2010;7: 905–919.
1112 doi:10.1098/rsif.2009.0400
1113 67. Rajon E, Plotkin JB. The evolution of genetic architectures underlying quantitative
1114 traits. Proc Biol Sci. 2013;280: 20131552. doi:10.1098/rspb.2013.1552
1115 68. Yoshida M, Tsuru S, Hirata N, Seno S, Matsuda H, Ying B-W, et al. Directed
1116 evolution of cell size in Escherichia coli. BMC Evol Biol. 2014;14: 257.
1117 doi:10.1186/s12862-014-0257-1
1118 69. F P-G, Jm S, P F. Evolution of cell size control is canalized towards adders or sizers
1119 by cell cycle structure and selective pressures. eLife. 2022;11. doi:10.7554/eLife.79919
1120 70. Barletta M, Lima ARA, Costa MF, Dantas DV. Estuarine Ecoclines and the
1121 Associated Fauna: Ecological Information as the Basis for Ecosystem Conservation. In:
1122 Finkl CW, Makowski C, editors. Coastal Wetlands: Alteration and Remediation. Cham:
1123 Springer International Publishing; 2017. pp. 479–512. doi:10.1007/978-3-319-56179-
1124 0_16
1125 71. Kent M, Gill WJ, Weaver RE, Armitage RP. Landscape and plant community
1126 boundaries in biogeography. Prog Phys Geogr Earth Environ. 1997;21: 315–353.
1127 doi:10.1177/030913339702100301
1128 72. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. A Reverse Ecology
1129 Approach Based on a Biological Definition of Microbial Populations. Cell. 2019;178:
1130 820-834.e14. doi:10.1016/j.cell.2019.06.033
1131 73. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically
1132 cohesive units. Trends Microbiol. 2014;22: 235–247. doi:10.1016/j.tim.2014.02.006
1133 74. Jenal U, Malone J. Mechanisms of cyclic-di-GMP signaling in bacteria. Annu Rev
1134 Genet. 2006;40: 385–407. doi:10.1146/annurev.genet.40.110405.090423
1135 75. Zhao B, Lees JA, Wu H, Yang C, Falush D. Genealogical inference and more
1136 flexible sequence clustering using iterative-PopPUNK. Genome Res. 2023;33: 988–998.
1137 doi:10.1101/gr.277395.122
1138 76. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, et al. Exploring
1139 and retrieving sequence and metadata for species across the tree of life with NCBI

Datasets. Sci Data. 2024;11: 732. doi:10.1038/s41597-024-03571-y
77. Yang C, Cui Y, Didelot X, Yang R, Falush D. Why panmictic bacteria are rare. bioRxiv; 2019. p. 385336. doi:10.1101/385336
78. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, et al. Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae. The Lancet. 2003;361: 743–749. doi:10.1016/S0140-6736(03)12659-1
79. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. Darling AE, editor. PLOS Comput Biol. 2018;14: e1005944. doi:10.1371/journal.pcbi.1005944
80. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genomics. 2016;2: e000056. doi:10.1099/mgen.0.000056
81. Seemann T. Source code for snp-dists software. Zenodo; 2018. doi:10.5281/ZENODO.1411986
82. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. PLOS Genet. 2012;8: e1002453. doi:10.1371/journal.pgen.1002453
83. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010;5: e9490. doi:10.1371/journal.pone.0009490
84. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8: 28–36. doi:10.1111/2041-210X.12628
85. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81: 559–575.
86. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma Oxf Engl. 2014;30: 2068–2069. doi:10.1093/bioinformatics/btu153
87. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21: 263–265. doi:10.1093/bioinformatics/bth457
88. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32: 2847–2849. doi:10.1093/bioinformatics/btw313
89. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010
90. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118} ; iso-2; iso-3. Fly (Austin). 2012;6: 80–92. doi:10.4161/fly.19695
91. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38: 5825–5829. doi:10.1093/molbev/msab293
92. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped

1186 BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic
1187 Acids Res. 1997;25: 3389–3402. doi:10.1093/nar/25.17.3389
1188 93. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt:
1189 the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32: D115–D119.
1190 doi:10.1093/nar/gkh131
1191 94. Haller BC, Messer PW. SLiM 4: Multispecies Eco-Evolutionary Modeling. Am Nat.
1192 2023;201: E127–E139. doi:10.1086/723601
1193 95. Cury J, Haller BC, Achaz G, Jay F. Simulation of bacterial populations with SLiM.
1194 Peer Community J. 2022;2. doi:10.24072/pcjournal.72
1195 96. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-
1196 scale rates of evolutionary change in bacteria. Microb Genomics. 2016;2: e000094.
1197 doi:10.1099/mgen.0.000094

1198

1199

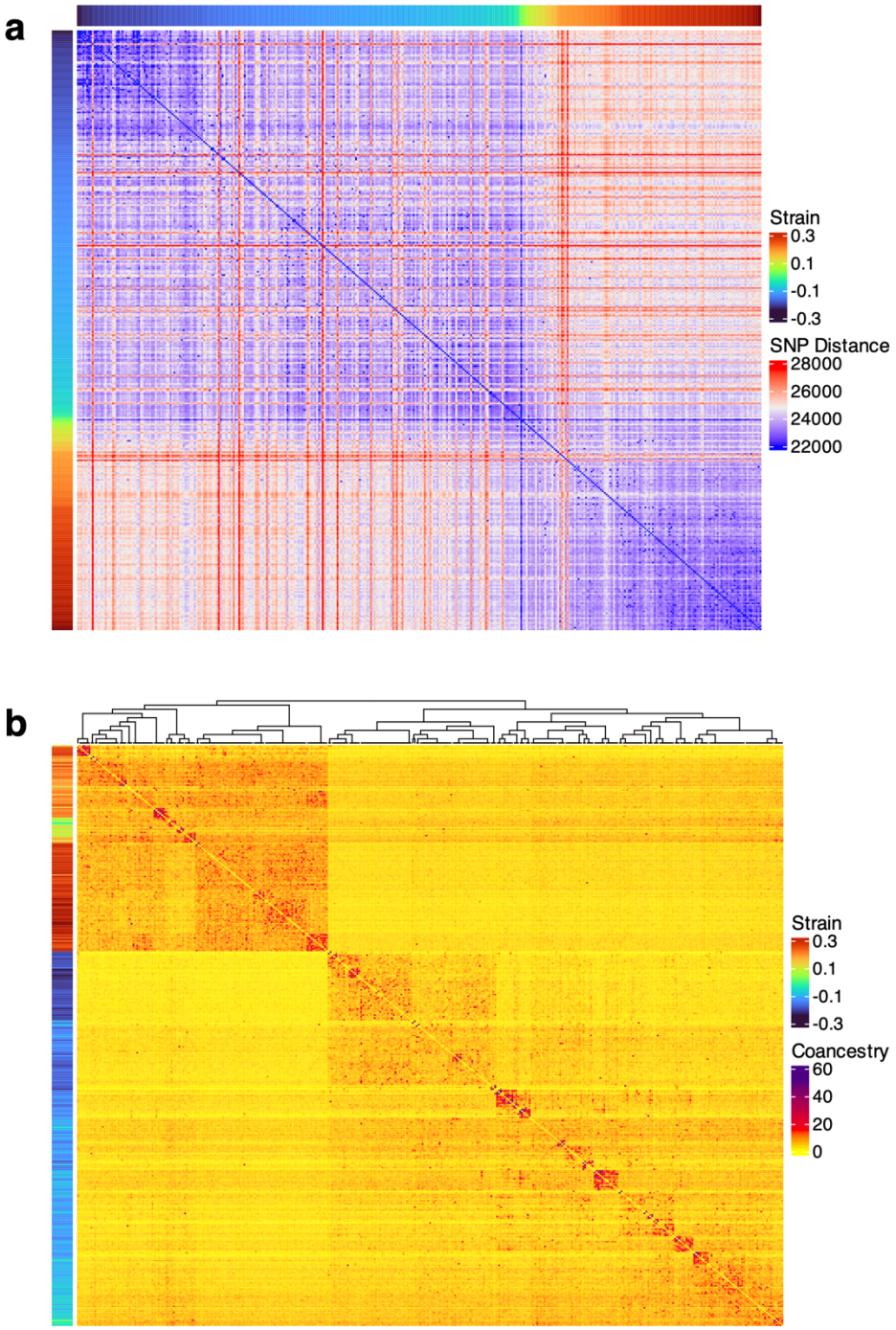
1200 # Supplementary Table

1201 Supplementary Table 1. Enrichment of KP1_0338 high and low gene types among different host sources
1202 based on Fisher's exact test. For each host, a one-vs-rest Fisher's exact test was performed to evaluate
1203 whether the gene type distribution differed significantly from that of all other sources combined.

Source	**High**	**Low**	***p*-value**
Animal	10	23	6.907177e-01
Bird	6	8	2.196240e-01
Cat	9	35	4.892939e-01
Cow	153	376	1.523384e-01
Dog	27	75	1.000000e+00
Environment	60	198	2.272826e-01
Food	12	21	2.313088e-01
Horse	58	67	9.531088e-07
Mouse	1	16	5.529954e-02
Pig	28	112	7.438013e-02
Poultry	9	59	1.128966e-02
Water	8	79	5.870439e-05

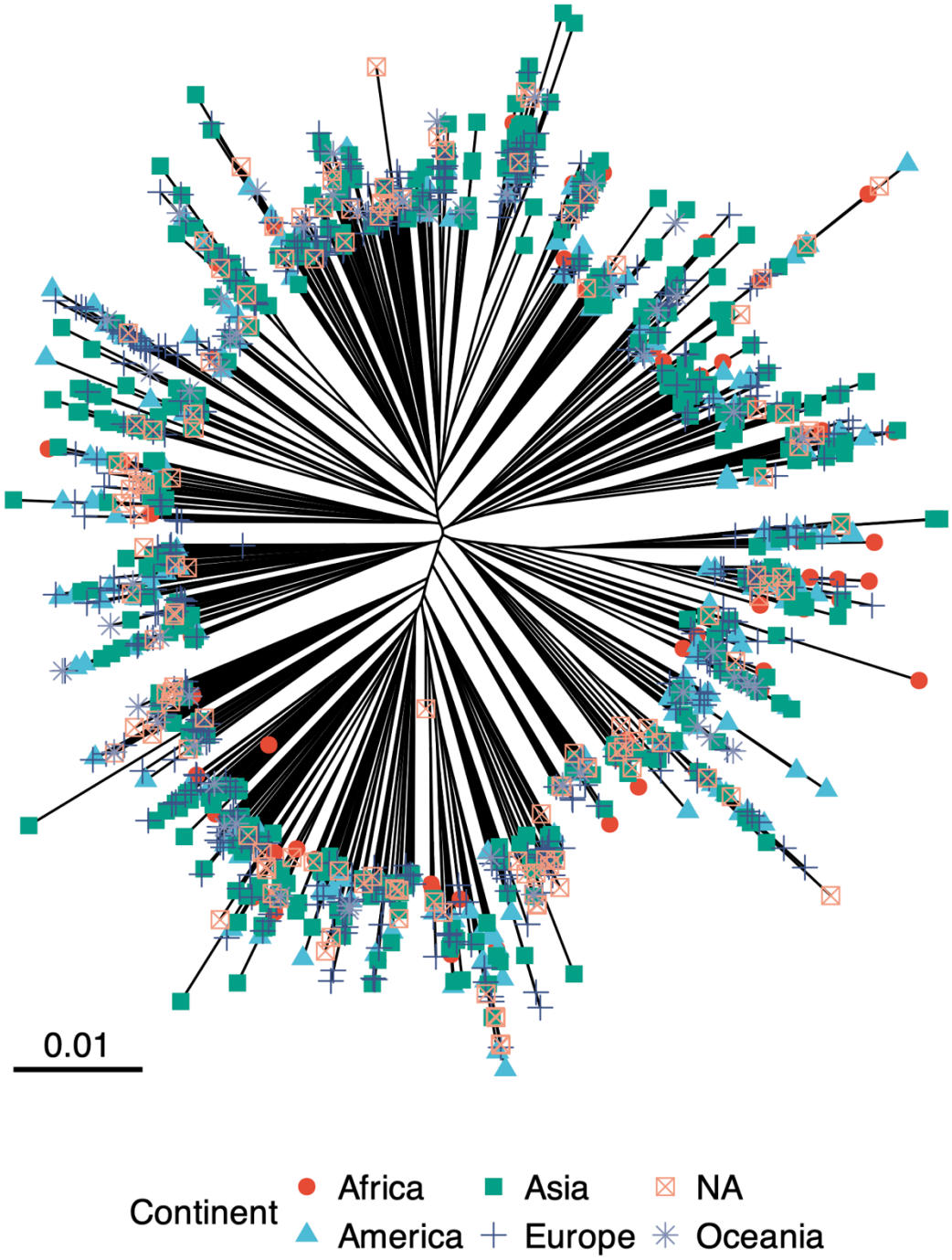
1204
1205

1206 **Supplementary Figures**



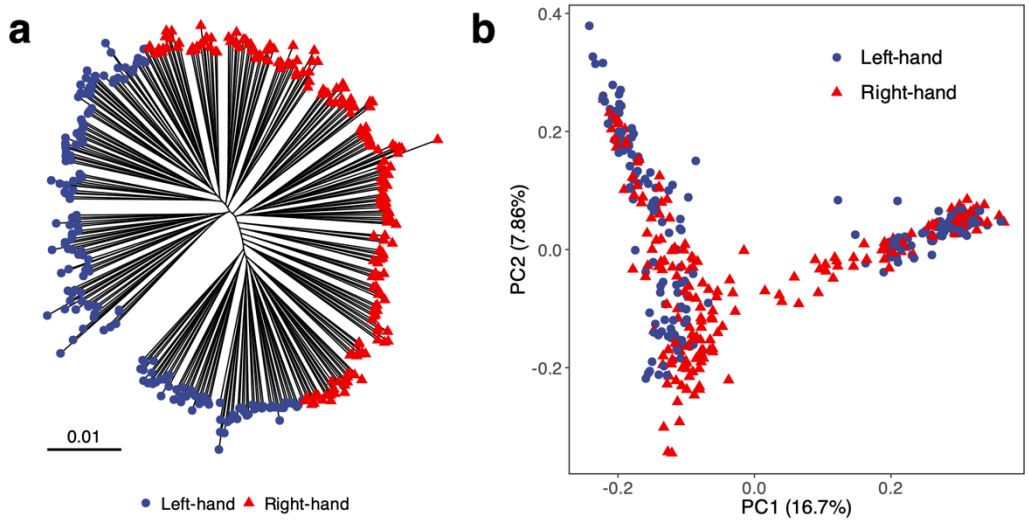
1207

1208 **Supplementary Figure 1. Genomic characteristics of the non-redundant KP dataset.** (a) SNP
1209 distance matrix illustrating the genetic distances between strains in the non-redundant dataset, with
1210 strains ordered based on their PC1 values. The minimum SNP distance observed is 20,000 SNPs. (b)
1211 fineSTRUCTURE analysis of the non-redundant KP dataset, highlighting population structure and
1212 clustering patterns. The "strain" bar on the left represents the PC1 values of the respective strains.

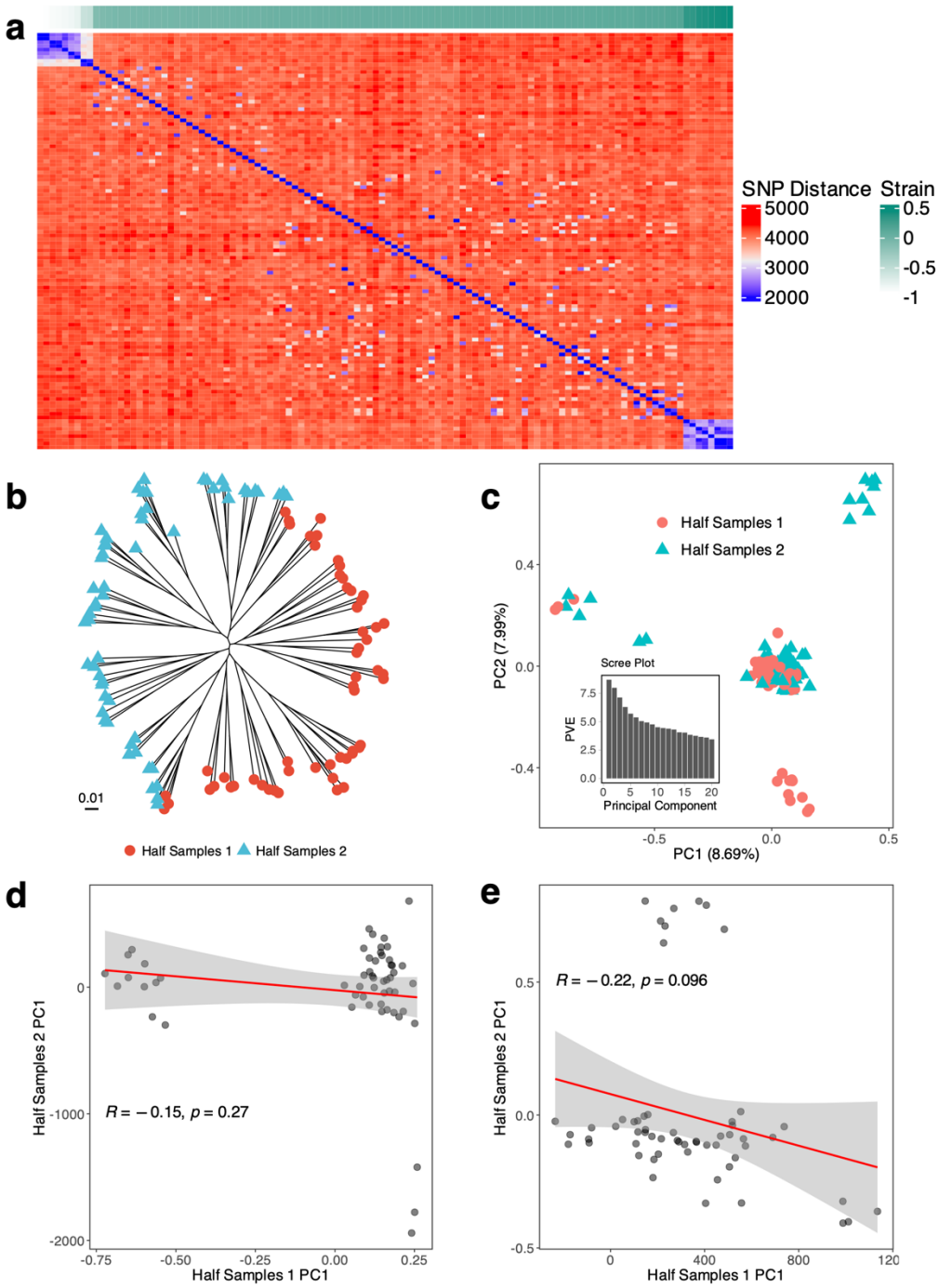


1213

1214 **Supplementary Figure 2.** Phylogenetic tree of the complete KP dataset. A maximum-likelihood
1215 phylogenetic tree constructed based on core SNPs from the full KP dataset, with branches color-coded
1216 by continental origin to illustrate the global distribution of isolates. NA: no information available.

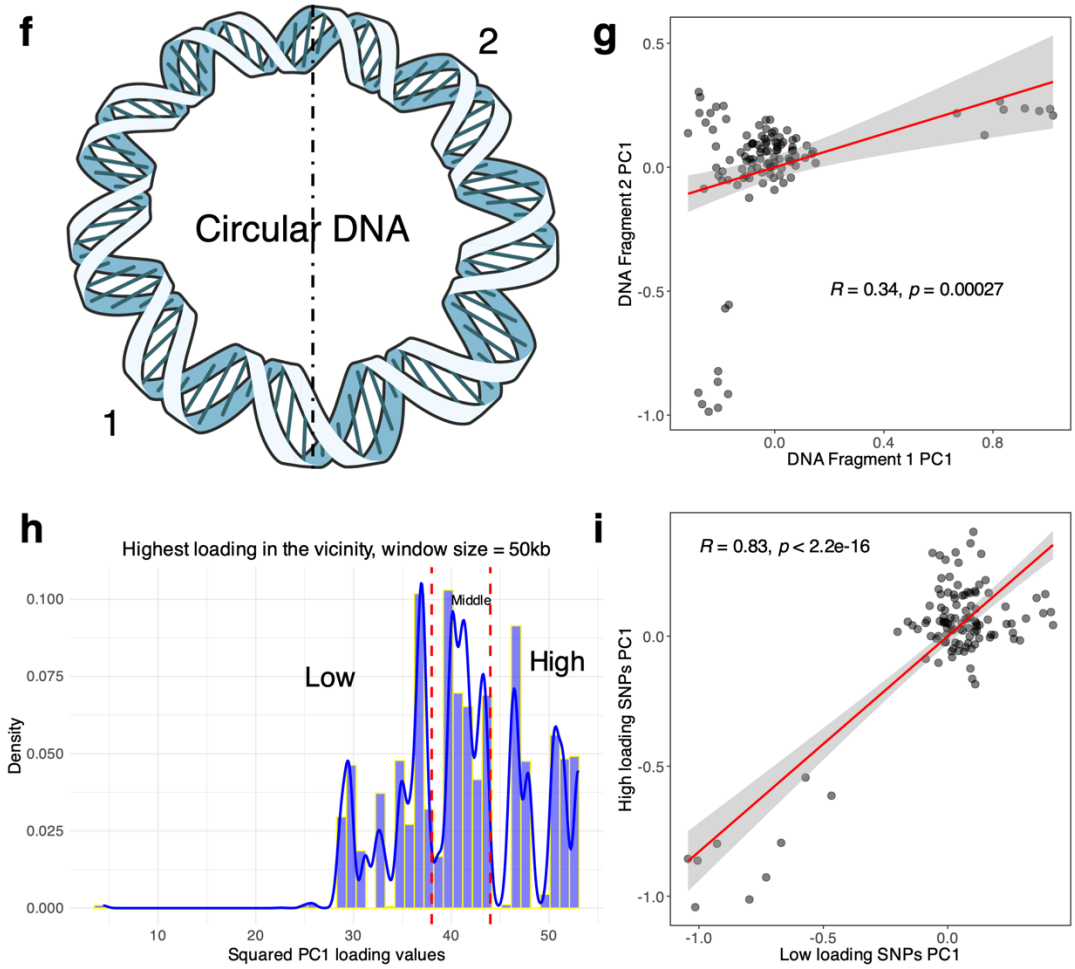


Supplementary Figure 3. (a) Phylogenetic tree of the non-redundant KP dataset divided into two subsets based on the backbone. (b) PCA of the non-redundant KP dataset colored according to two sub-groups.

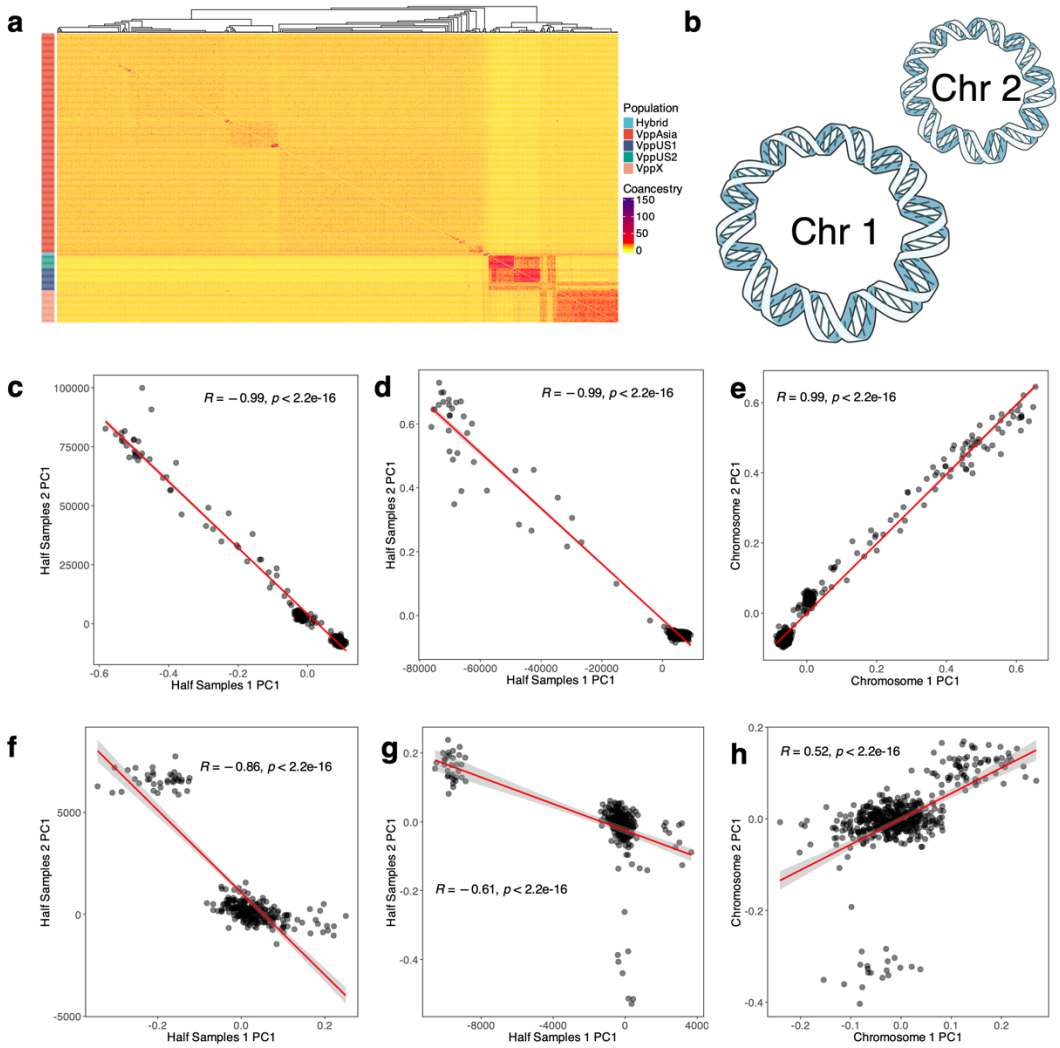


1221

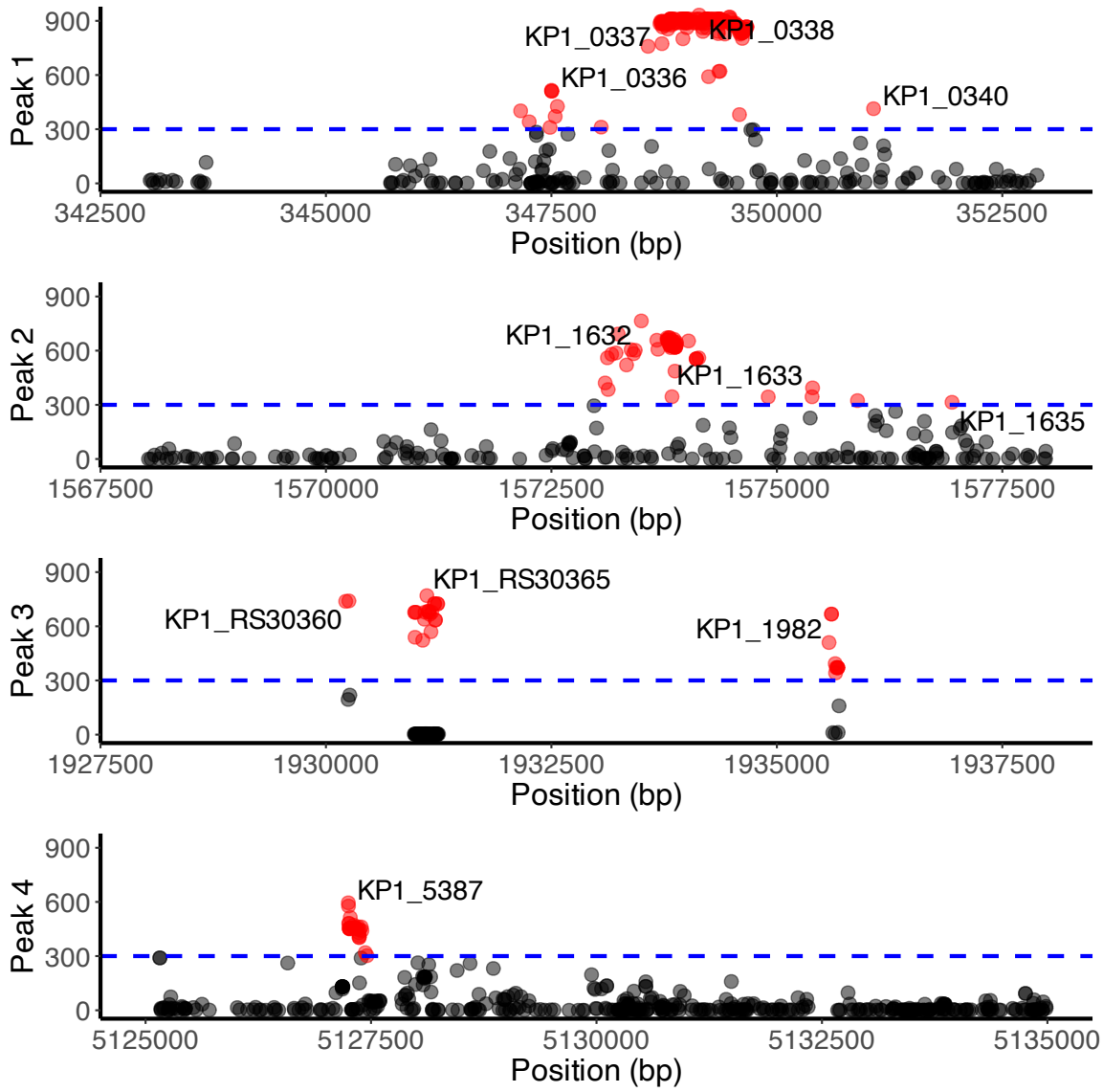
1222 **Supplementary Figure 4. Three types of half-matching tests for the clonal simulation.** (a) SNP
1223 distance matrix for the clonal simulation. The minimum SNP distance observed is 2,000 SNPs. (b–e)
1224 Half-strain matching: (b) Phylogenetic tree of the non-redundant clonal simulation divided into 2
1225 subsets. (c) PCA of the non-redundant clonal simulation data. (d) Correlation of PC1 from group 2
1226 projected onto group 1 results. (e) Correlation of PC1 from group 1 projected onto group 2 results.



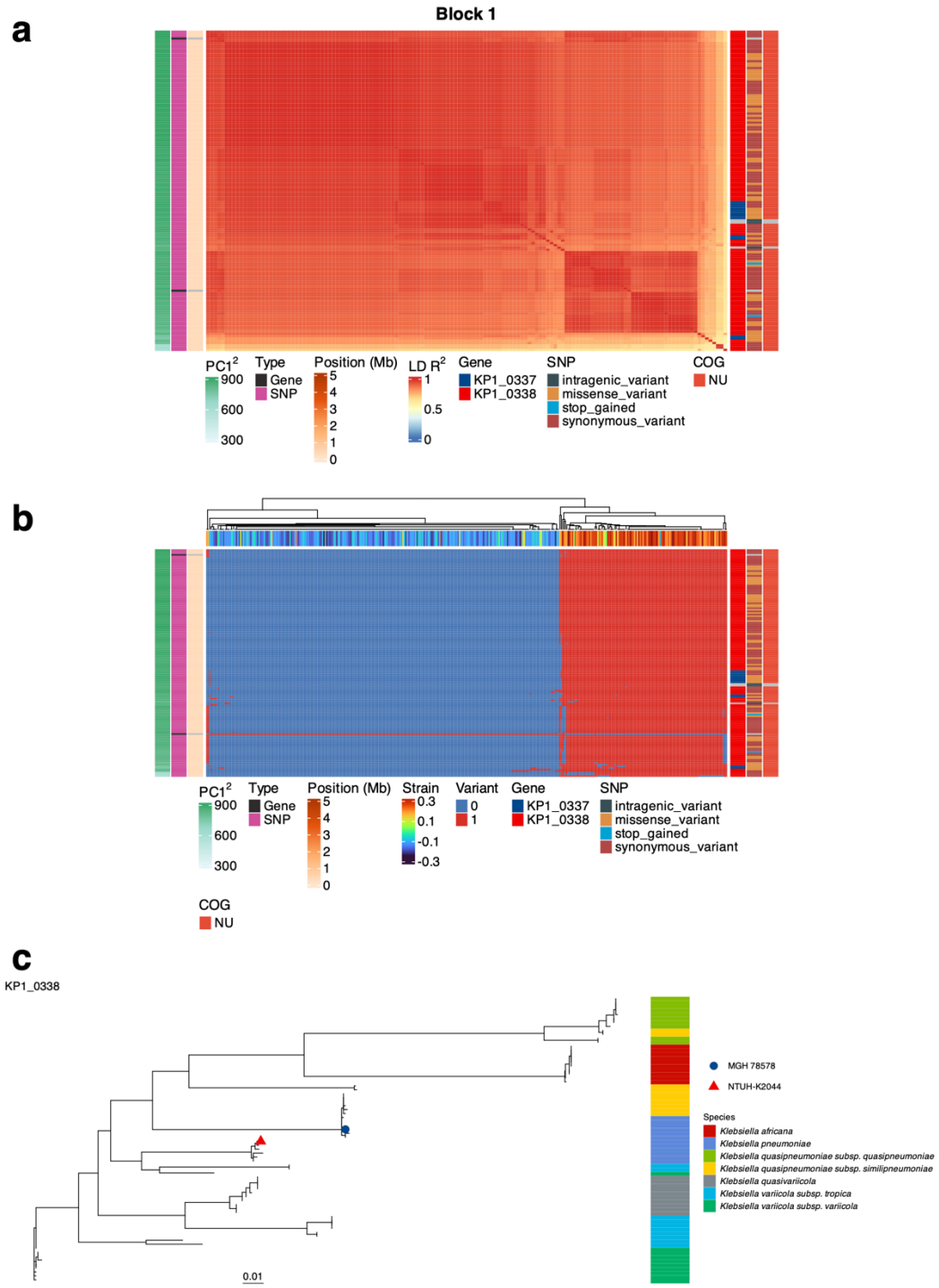
Supplementary Figure 4 (continued). Three types of half-matching tests for the clonal simulation. (f–g) Half-genome matching: (f) Division of the genome into two fragments (Fragment 1 and Fragment 2). (g) Correlation of PC1 between the two fragments. (h–i) Partial-variants matching: (h) Histogram of SNP loadings on PC1, dividing SNPs into high- and low-loading groups. (i) Correlation of PC1 between high- and low-loading SNP groups.



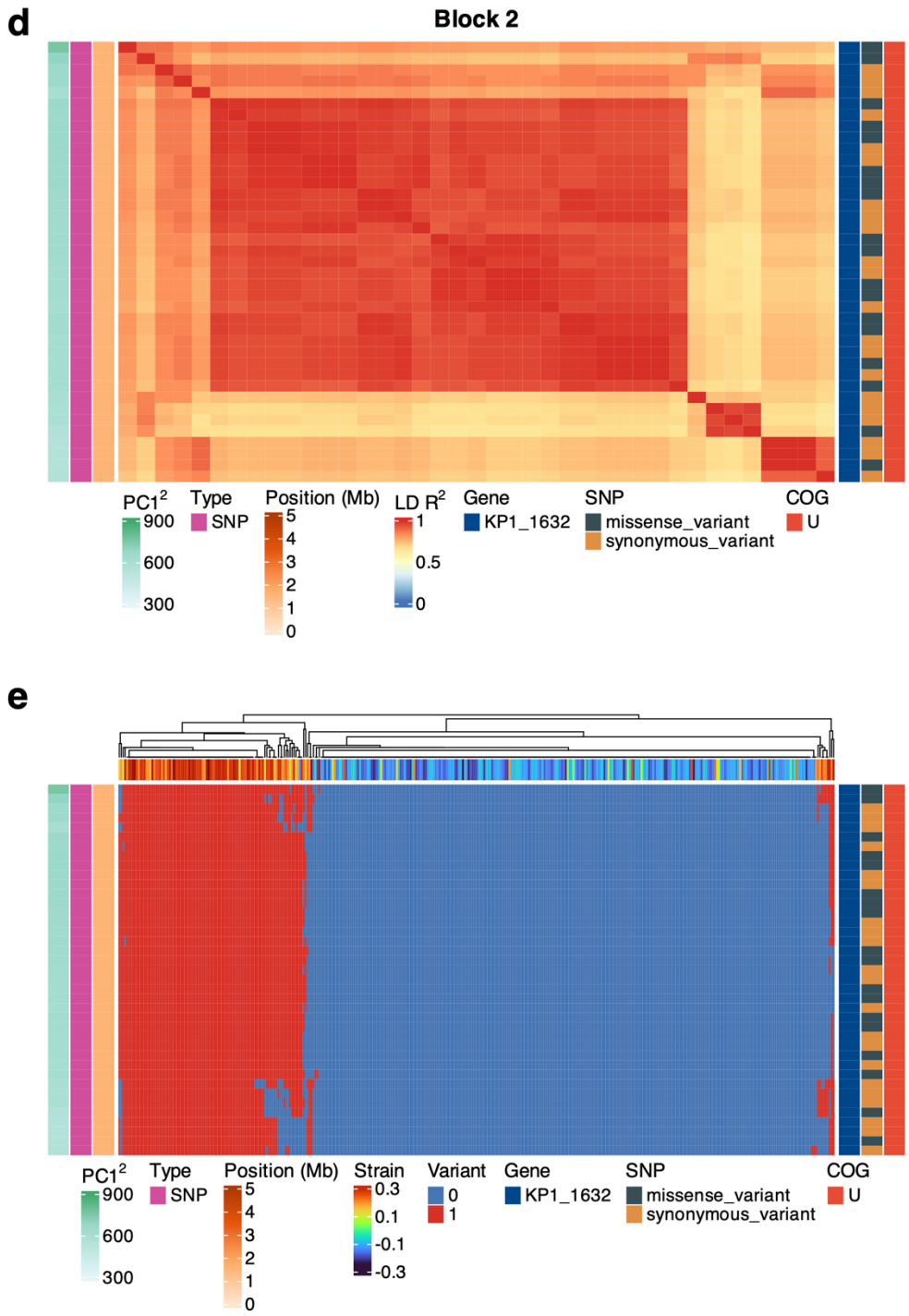
Supplementary Figure 5. Half-matching tests for VP datasets. (a) fineSTRUCTURE analysis assigning populations for the non-redundant VP dataset. (b) Representation of VP Chromosome 1 and Chromosome 2, with proportions reflecting their actual sizes. (c-e) Half-strain matching for the non-redundant VP dataset: (c) Correlation of PC1 from group 2 projected onto group 1 results; (d) Correlation of PC1 from group 1 projected onto group 2 results. Half-chromosome matching for the non-redundant VP dataset: (e) Correlation of PC1 between the two chromosomes. (f-h) Half-matching for VppAsia dataset: (f) Correlation of PC1 from group 2 projected onto group 1 results; (g) Correlation of PC1 from group 1 projected onto group 2 results. Half-chromosome matching for VppAsia dataset: (h) Correlation of PC1 between the two chromosomes.



Supplementary Figure 6. Zoomed-in views of four high-loading SNP peaks identified across the genome.

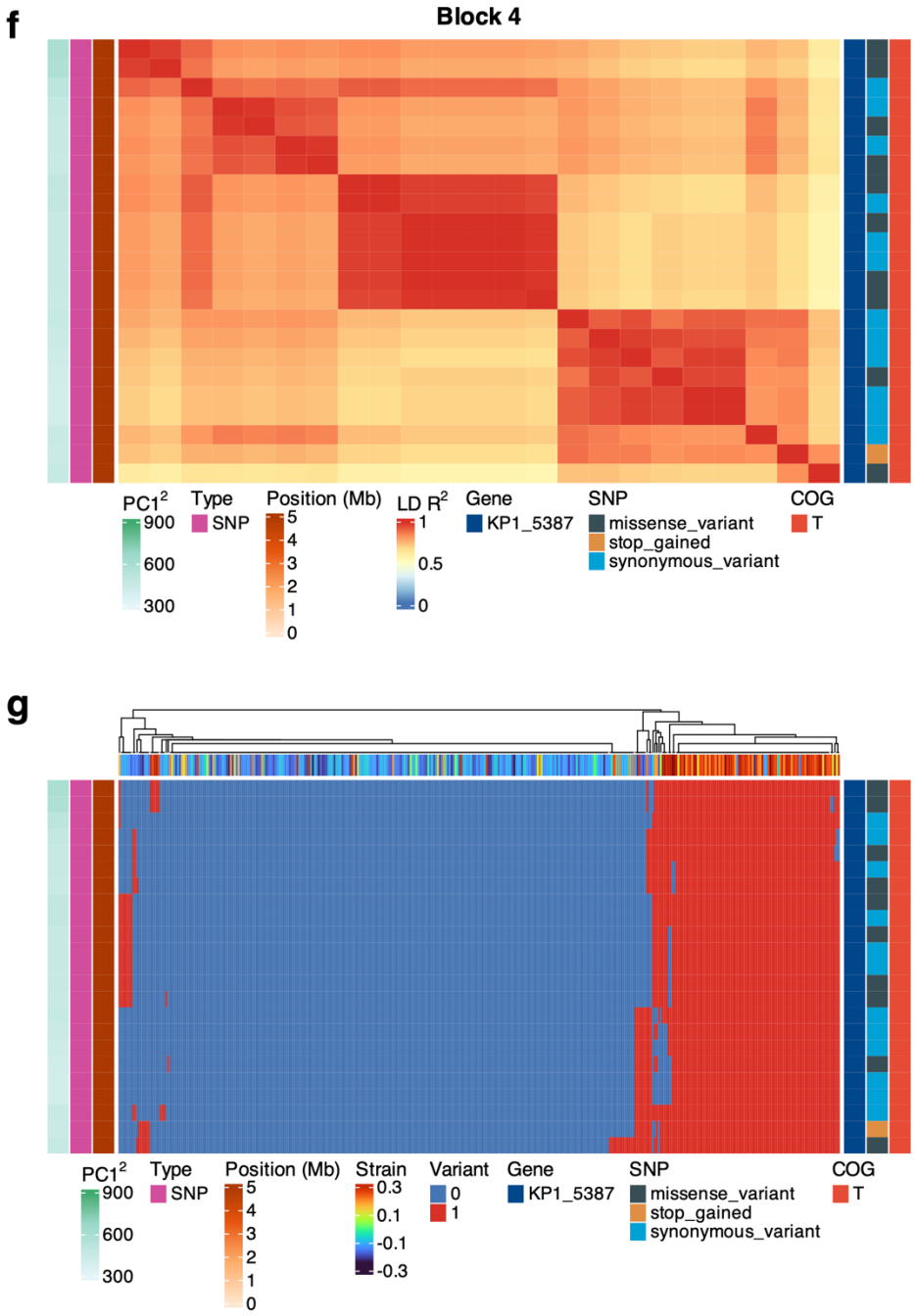


Supplementary Figure 7. Overview of blocks 1–4 from LD analysis with variant status heatmaps. Block 1 (a-b): (a) SNPs in KP1_0337, KP1_0338, and associated accessory genes form a high LD block. (b) Variant (1/0) heatmap shows clear differentiation in gene presence/absence or SNP minor/major alleles across isolates. (c) Multi-species phylogenetic tree of the KP1_0338 gene. The annotation strip indicates the species of each strain.



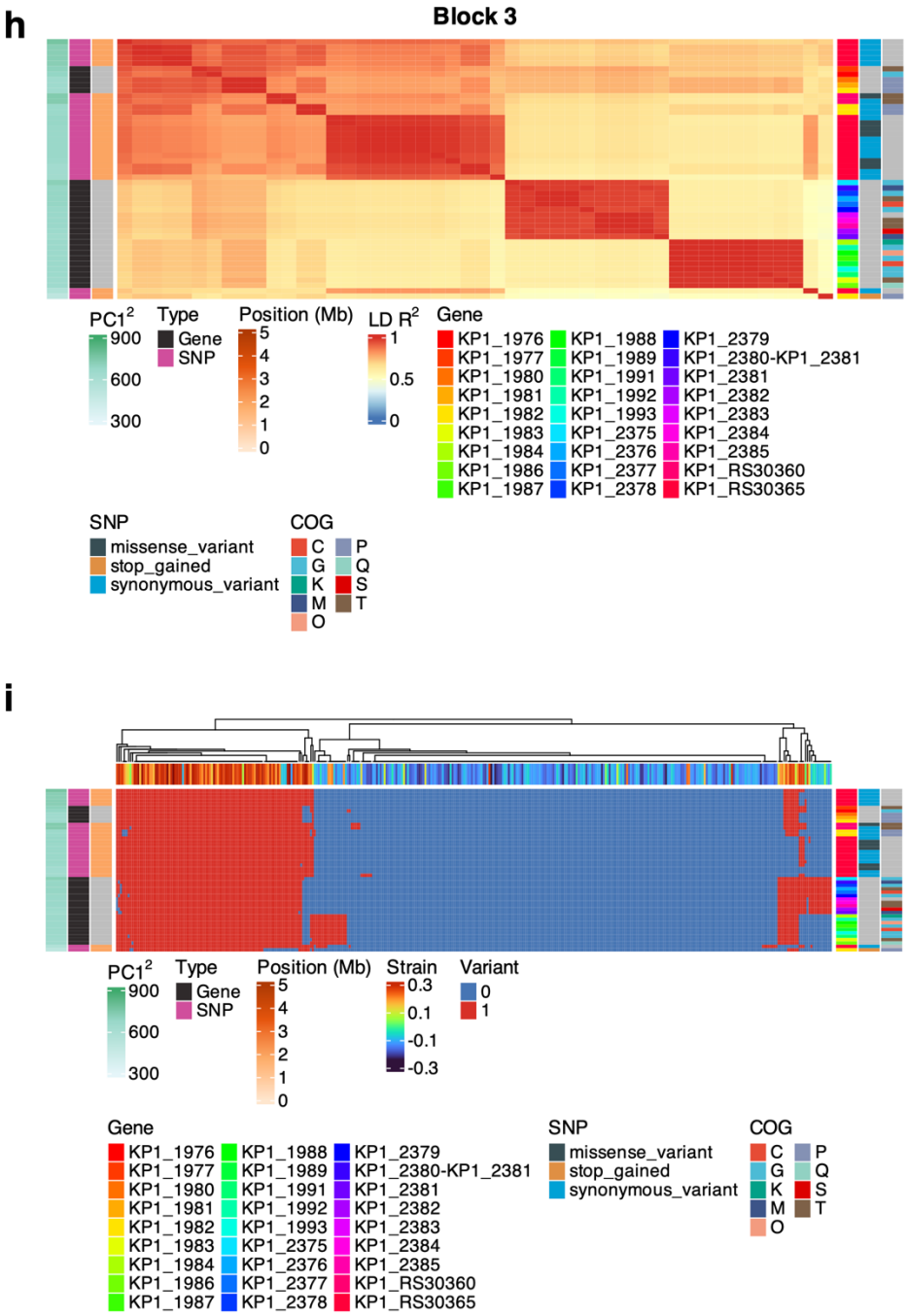
1252

1253 **Supplementary Figure 7 (continued). Overview of blocks 1–4 from LD analysis with variant status**
1254 **heatmaps.** Block 2 (d-e): (d) High LD observed for SNPs in KP1_1632. (e) Variant (1/0) heatmap
1255 highlights distinct patterns of gene presence/absence or SNP minor/major alleles across isolates.

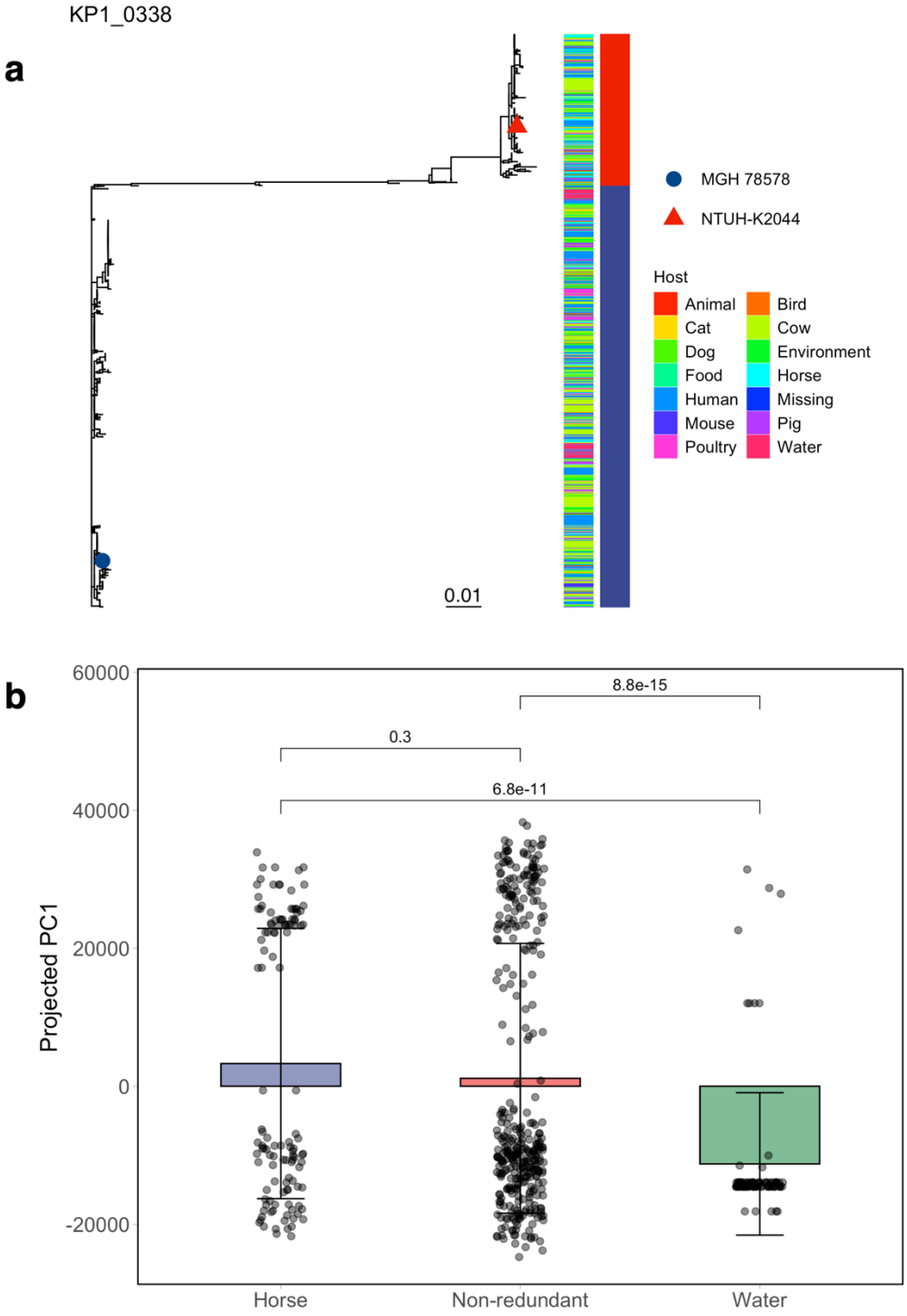


1256

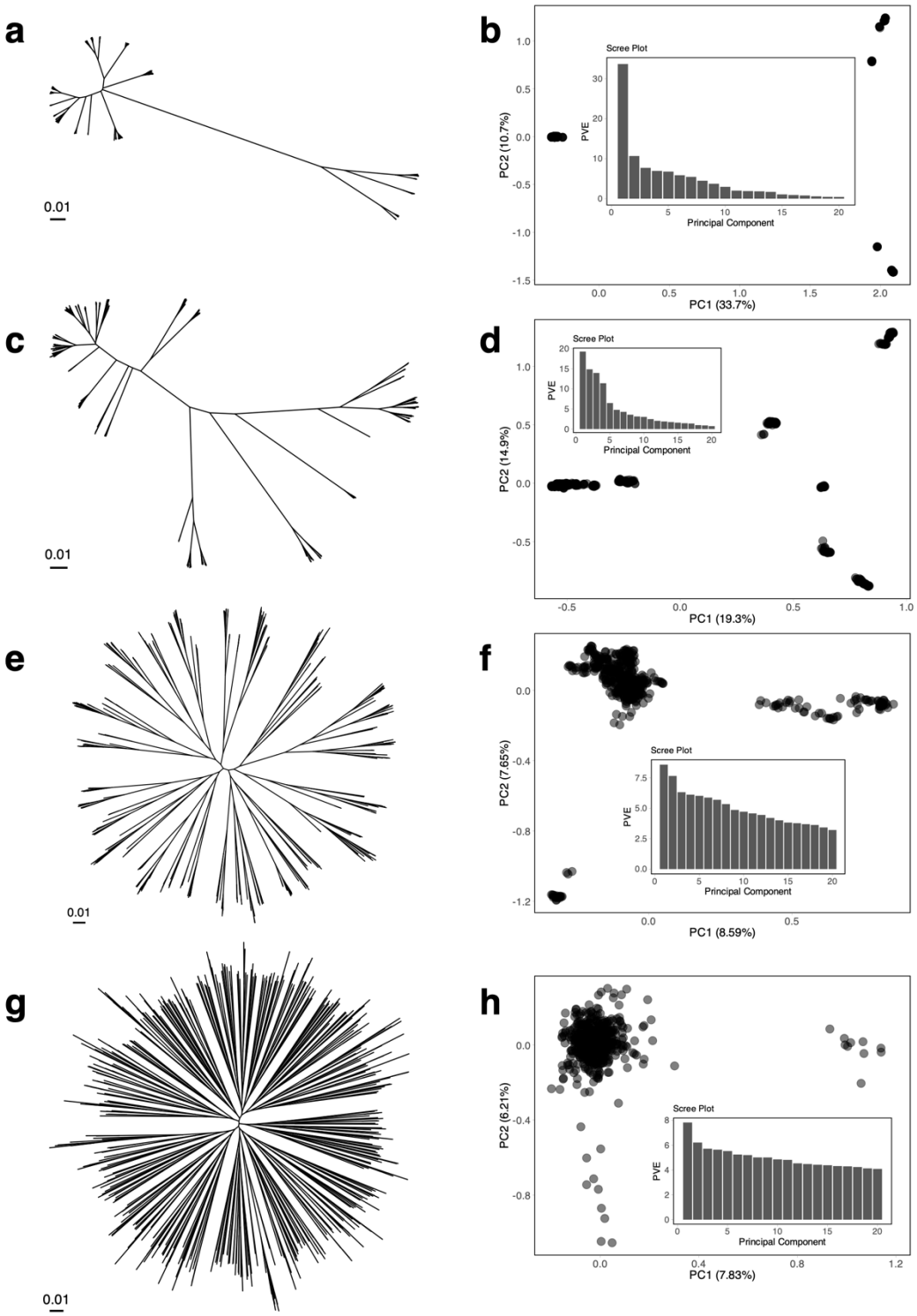
1257 **Supplementary Figure 7 (continued). Overview of blocks 1–4 from LD analysis with variant status**
1258 **heatmaps.** Block 4 (h-i): (h) High LD observed for variants in KP1_5387, forming a small, highly
1259 linked cluster. (i) Variant (1/0) heatmap shows distinct variation in gene presence/absence or SNP
1260 minor/major alleles across isolates.



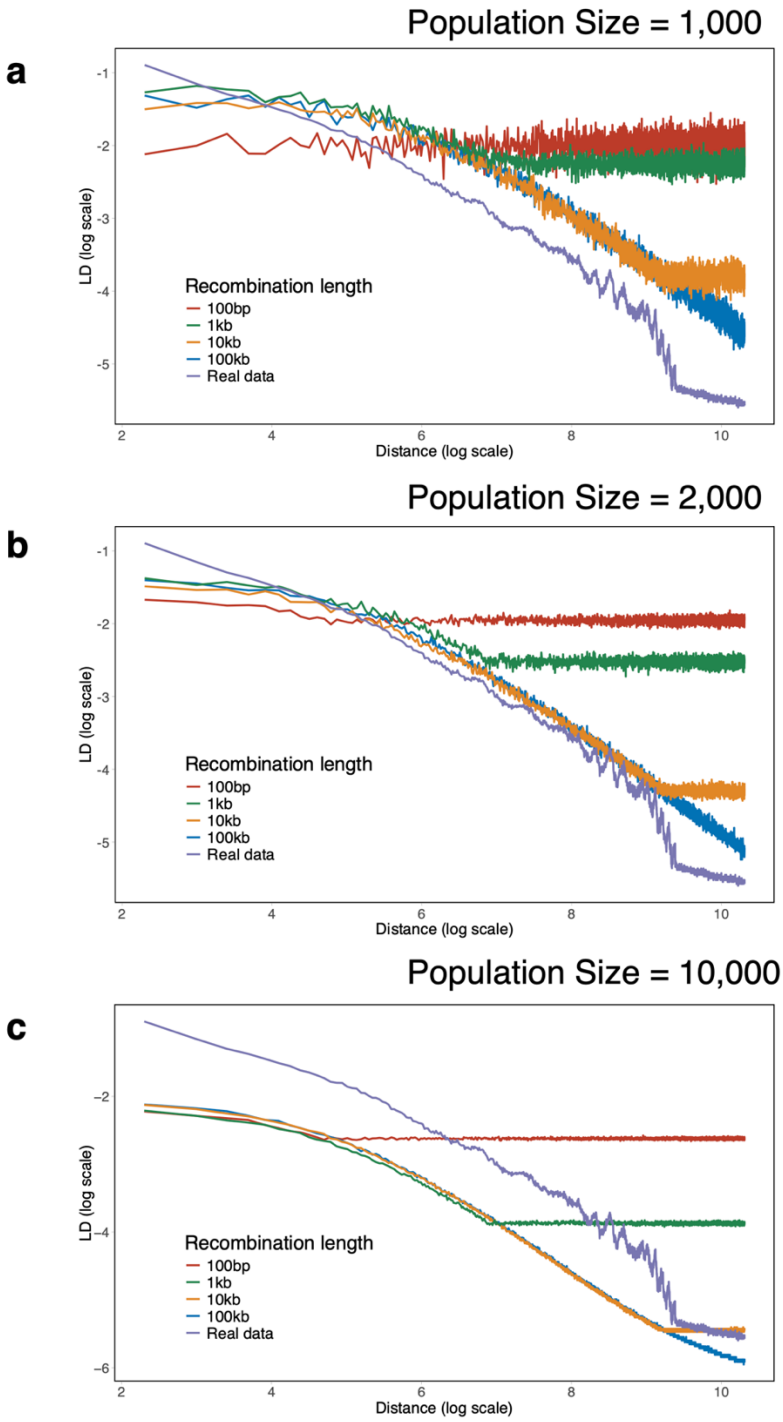
Supplementary Figure 7 (continued). Overview of blocks 1–4 from LD analysis with variant status heatmaps. Block 3 (f-g): (f) SNPs in KP1_RS09345, KP1_RS30360, KP1_1982, and accessory genes form tightly linked clusters. (g) Variant (1/0) heatmap reveals clear differentiation in gene presence/absence or SNP minor/major alleles across isolates.



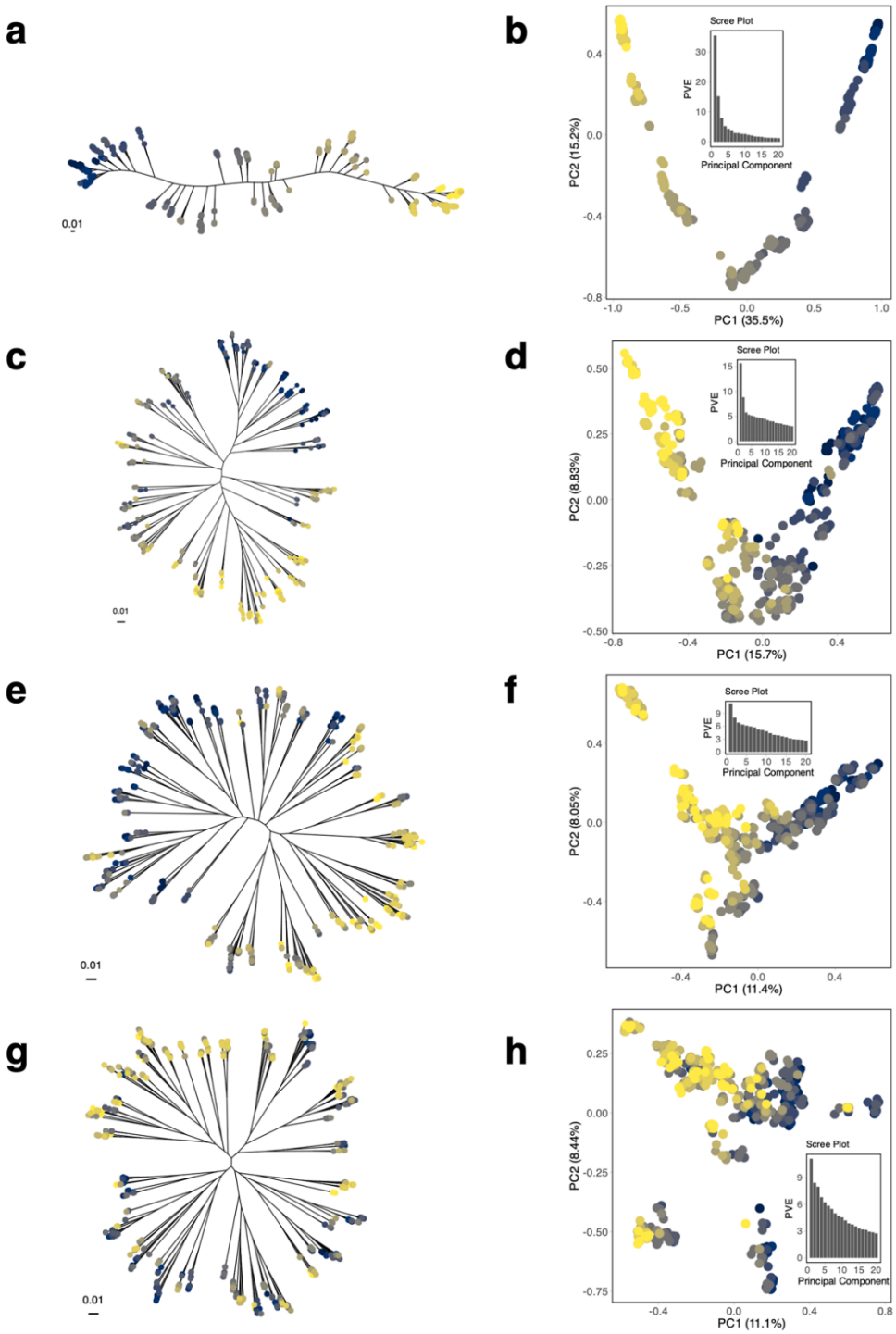
Supplementary Figure 8. Phylogenetic tree of KP1_0338 within KP and host origin profiles. (a) Phylogenetic tree of the KP1_0338 gene based on sequences from the multi-host KP dataset. Two annotation strips indicate host origin and KP1_0338 classification (High vs. Low) based on phylogenetic clustering. (b) Projected the multi-host dataset onto the non-redundant dataset. Each bar represents the mean PC1 value for isolates from a given host, with error bars indicating the standard deviation.



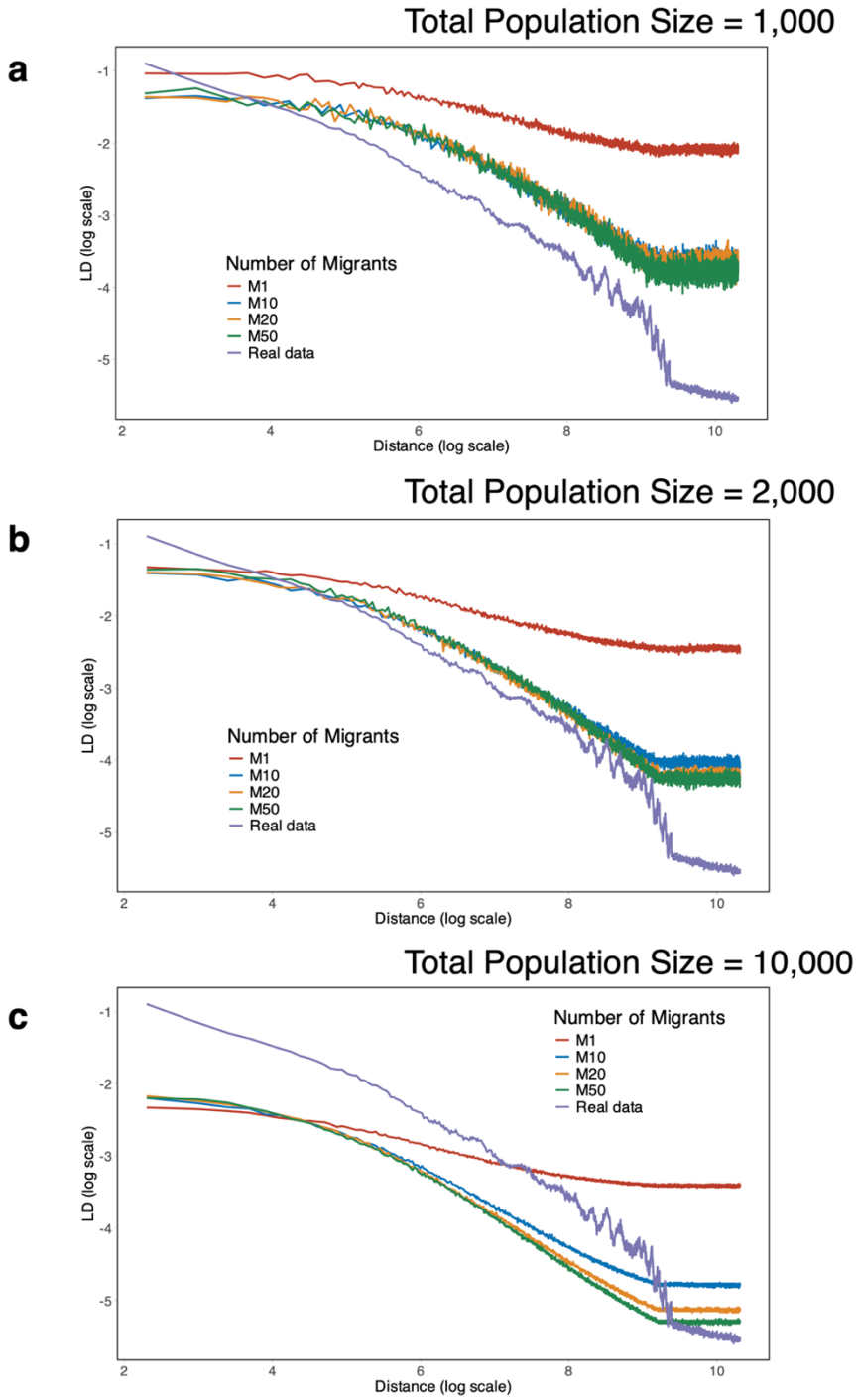
Supplementary Figure 9. Clonal haploid bacteria model with varying HGT lengths and a population size of 2000. Phylogenetic trees, PCA plots, and scree plots for a population of 2000 individuals under different HGT lengths. Panels (a-h) show results for: (a-b) 100 bp, (c-d) 1 kb, (e-f) 10 kb, (g-h) 100 kb.



Supplementary Figure 10. LD decay under different 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 KP 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, PCA plots, and scree plots for a model with HGT of 10 kb. The population consists of 10 subpopulations (200 individuals each). Panels (a–h) 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. LD decay profiles are compared between linear stepping-stone models with varying numbers of migrants (1, 10, 20, 50) and the non-redundant KP dataset. Panels illustrate results for different total population sizes: (a) 2000 individuals, (b) 10,000 individuals, (c) 100,000 individuals. Each model includes 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.