Why panmictic bacteria are rare

Chao Yang[1#], Yujun Cui[1#], Xavier Didelot[2,3], Ruifu Yang[1], Daniel Falush[4*]

1 State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China

2 School of Life Sciences, University of Warwick, Gibbet Hill Campus, Coventry CV4 7AL, United Kingdom

3 Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom

4 Milner Centre for Evolution, University of Bath, Bath, Somerset, United Kingdom

[#]These authors contributed equally to the article

* Corresponding authors: D. F. (danielfalush@googlemail.com)

Abstract

Bacteria typically have more structured populations than higher eukaryotes, but this difference is surprising given high recombination rates, enormous population sizes and effective geographical dispersal in many bacterial species. We estimated the recombination scaled effective population size $N_e r$ in 21 bacterial species and find that it does not correlate with synonymous nucleotide diversity as would be expected under neutral models of evolution. Only two species have estimates substantially over 100, consistent with approximate panmixia, namely *Helicobacter pylori* and *Vibrio parahaemolyticus*. Both species are far from demographic equilibrium, with diversity predicted to increase more than 30 fold in *V. parahaemolyticus* if the current value of $N_e r$ were maintained, to values much higher than found in any species. We propose that panmixia is unstable in bacteria, and that persistent environmental species are likely to evolve barriers to genetic exchange, which act to prevent a continuous increase in diversity by enhancing genetic drift. Our results highlight the dynamic nature of bacterial population structures and imply that overall diversity levels found within a species are poor indicators of its size.

Introduction

Bacteria are paragons of adaptability and make up more biomass than all organisms other than plants combined[1]. Many bacterial species have enormous population sizes, disperse effectively around the globe[2] and exhibit high rates of within-species homologous recombination[3]. Recombination progressively breaks down non-random associations between markers (linkage disequilibrium), so that in large populations, where pairs of individuals distantly related, linkage equilibrium is expected between all pairs of genomic sites. However, in most species for which data is available, there is substantial genome-wide structure variation[4]. Here we propose a resolution of this paradox, namely that large bacterial populations accumulate diversity progressively until that diversity acts as an effective barrier to genetic exchange between lineages.

We examine the population structure of 21 bacterial species and find that only the Asian population of *Vibrio parahaemolyticus* is close to genome-wide linkage equilibrium. We find that this population

has undergone a recent expansion. If the current population size was maintained over evolutionary timescales, it would lead to a greater than 30 fold increase in diversity, to levels higher than found in any well characterized bacterial species. We propose that other environmental species with large census population sizes, for example *Vibrio cholerae* or *Klebsiella pneumoniae*, may have been through a similar stage before accumulating diversity and then evolving mechanisms that disrupted free recombination.

Specifically, we use pairwise genetic distances between strains calculated from core genome sequences to estimate the recombination scaled effective population size parameter $N_e r$, where N_e is the effective population size and r is the recombination rate per site per generation, an approach briefly introduced by Cui et al[5]. Here, we test the method using simulated genomic datasets and find that it provides accurate estimates in constant size populations for values of the scaled recombination rate R (see below) of 5 or more, and responds to changes in population size or structure far more quickly than estimates of N_e based on nucleotide diversity. We apply the method to real genomic datasets from 21 species and find that some common species, such as *Escherichia coli*, have strikingly low estimates of $N_e r$, implying that recombination is inefficient in distributing diversity across the species, while for others the fit of model to data is poor, highlighting deviations from the assumption of a freely recombining population. Only datasets from *V. parahaemolyticus* and *Helicobacter pylori* give $N_e r$ estimates greater than 100. The former has had a recent increase in population size, whereas the latter experienced repeated bottlenecks associated with geographical spread of its human host, implying that neither population is close to demographic equilibrium.

Inference approach

Population genetics is a century-old discipline that provides a powerful set of theoretical and statistical inference tools with which to interpret patterns of genetic variation between closely related organisms. Central to the theory is the concept of a population, which in outbreeding eukaryotes is a set of organisms that share a common gene pool. The simplest models assume panmixia which means random mating of all individuals within a single closed population[6]. In most animals and plants mating is structured by geography, however even low levels of migration between locations, of the

order of one individual per generation, can prevent differences from accumulating between local populations, making random mating a reasonable first approximation for many species. Under neutral theory[7], the expected level of genetic diversity π depends on the per generation mutation rate μ and the effective population size N_e, which is the number of individuals contributing to the gene pool in each generation[8].

Outbreeding eukaryotes receive genetic material from the gene pool when they are born and contribute to it when they reproduce. Individuals are more similar to immediate relatives than they are to other members of the population, but the proportion of the genome shared by descent from particular ancestors decays rapidly with each successive generation, declining from 1/2 for full siblings to 1/8 for first cousins and 1/32 for second cousins. In small random samples from populations with more than a few hundred individuals, shared recent ancestry is rare and can be neglected for many types of analysis.

Bacteria reproduce by binary fission and only receive material from the gene pool or contribute to it via homologous recombination[9]. Many cell divisions can take place between consecutive recombination events, and typically recombination only affects a small fragment of the genome. These properties mean that the concept of a gene pool or a population is less straightforward to define than for outbreeding organisms. It is still possible to estimate the effective population size N_e from the average nucleotide diversity π, based on the standard assumptions of population genetic theory, but the assumptions are both less reasonable and harder to test for bacteria than for outbreeding eukaryotes, in which population boundaries can be delineated empirically using well established methods[10]. For example, nucleotide diversity in a species like *E. coli* can be estimated for clonal lineages, phylogroups, species or for the *Escherichia* genus as a whole. These choices lead to very different values for π and hence for N_e and it is not obvious a priori which is most meaningful. All methods lead to estimates of effective population size that are many orders of magnitude smaller than the census number of bacteria[11].

Here, we take a different approach which is to estimate a scaled version of the effective population size, $N_e r$, where r is the per generation rate at which a given site recombines. Note in particular that

95 this parameter r is the product of the per initiation site recombination rate ρ and mean tract length δ
96 used in many bacterial recombination models[12,13]. High values of $N_e r$ indicate that the population
97 structure is similar to that of outbreeding eukaryotes. Informally, in eukaryotic population each
98 individual is the product of a separate meiosis and therefore genetically distinct. N_e is the number of
99 genetically distinct individuals that contribute in each generation, which is typically in the thousands
100 or millions. $N_e r$ is designed to measure an analogous quantity in bacteria, namely the number of
101 genetically distinct organisms that contribute to the future bacterial gene pool. To this end, time is
102 rescaled; N_e measures the rate of genetic drift per bacterial generation, while $N_e r$ measures the
103 genetic drift in proportion to the time it takes for strains to become distinct from their ancestors other
104 by importing DNA by homologous recombination. The recombination rate of *V. parahaemolyticus* is
105 $r = 1.7 \times 10^{-4}$ per site per year[14]. After T years of evolution, the expected proportion of recombined
106 genome is e^{-rT}, so that it takes around 4,000 years on average for half of the genome to recombine.
107 Half of this time is approximately equivalent to a eukaryotic generation in the sense that two strains
108 that shared a common ancestor 2,000 years ago will be about as related as siblings. This represents a
109 very different time scale from that assumed based on bacterial generations. For example, in *V.*
110 *parahaemolyticus*, bacterial doubling times are likely to be similar to those in *V. cholerae*, which have
111 been estimated at around an hour[15].

112 We estimate $N_e r$ using the pairwise genetic distances between strains, based on a number of
113 simplifying assumptions which are likely to hold true in freely recombining bacterial populations but
114 may break down in species where recombination rates are low or genetic exchange is structured by
115 geography or lineage. Specifically, the calculations assume that recombination introduces many more
116 substitutions than mutation, happens at the same rate throughout the genome and that each
117 recombination event introduces unrelated DNA from the population into the imported stretch. If
118 unrelated strains differ on average at $d_{\text{unrelated}}$ nucleotides throughout the alignment, then the
119 expected number of SNPs distinguishing strains with a common ancestor at time T in the past is
120 $d = d_{\text{unrelated}}\,(1 - \mathrm{e}^{-2rT})$.

121 To use these times to estimate the effective population size, we assume that the genealogy of clonal
122 relationships is generated by a coalescent model with a constant population size N_e[16,17]. This model
123 generates expectations for the times in the past at which common ancestors of strains in a sample
124 existed. Specifically, for a sample of n strains, there are $n-1$ coalescent nodes. The age of the most
125 recent node corresponds to the common ancestor of the two most closely related strains in the sample,
126 while the $(n-1)^{\text{th}}$ node corresponds to the common ancestor of all the strains in the sample.

127 Coalescent theory implies that the expected time in the past at which the mth most ancient coalescent
128 event occurs is $T_m = 2N_e\left(\frac{1}{m}-\frac{1}{n}\right)$. These times can be converted into expected genetic distances d_m
129 using the formula in the previous paragraph. We use the UPGMA algorithm to obtain $n-1$
130 coalescent distances from the $n(n-1)/2$ pairwise genetic distances between strains and find the
131 values of $N_e r$ and $\hat{n}$ (effective sample size, see below) that gives the best fit between observed and
132 expected distances for the $n-1$ coalescent nodes.

133 Note that for bacteria with high recombination rate, the genome is likely to have been scrambled up
134 sufficiently that strains will have inherited little or no DNA by direct descent from the common
135 ancestor of the entire sample. This means that the genetic distances expected for the oldest coalescent
136 events plateau at $d_{\text{unrelated}}$. In graphical representations, it is convenient to show the coalescent events
137 in chronological order with the oldest first, to aid comparisons between datasets with different sample
138 sizes.

139 The model assumes that the strains are randomly sampled from a homogeneous population at a single
140 time point, but pathogenic clones, epidemic outbreaks or strains from specific locations are often
141 overrepresented in real data, leading to oversampling of very closely related isolates, relative to their
142 frequency in the global population[4,18]. Therefore, in addition to $N_e r$ we estimate a second parameter $\hat{n}$
143 called the effective sample size, which is an estimate of the number of strains remaining when over-
144 sampled clonally related strains are removed. For simulated data where sampling is random, $\hat{n}$ is
145 correctly estimated to be very close to number of strains in the sample, so this additional parameter

makes little difference to the inference. For real data, estimating this additional parameter often improves the qualitative model fit considerably.

Results

$N_e r$ can be estimated accurately for simulated data

Fig. 1 illustrates the effect of varying the recombination rate in genomes simulated using FastSimBac[19]. The simulations include 200 genomes of length 2 Mb under a coalescent model with constant effective population size N_e. We fix the recombination tract length δ = 1,000 and vary the rate of initiation of recombination events ρ per $2N_e$ generations, so that the scaled recombination rate $R = \frac{\delta\rho}{2N_e}$ varies by two orders of magnitude, from 0.5 to 50. For low recombination rates, the phylogenetic tree is highly structured but becomes progressively bushier as R increases, and gives the impression of being a nearly perfect star for R = 50. For high R, the estimated $N_e r$ is close to R and the observed genetic distances are well-fit by the model. For lower R, the fit is less good and the estimate of R provided by $N_e r$ is also less accurate, although it remains of the right order. For these parameter values, the size and shape of the phylogenetic tree are highly variable between runs, and the poorer fit reflects this stochasticity as well as greater inaccuracy in the approximations made in converting between genetic distances and coalescent times.

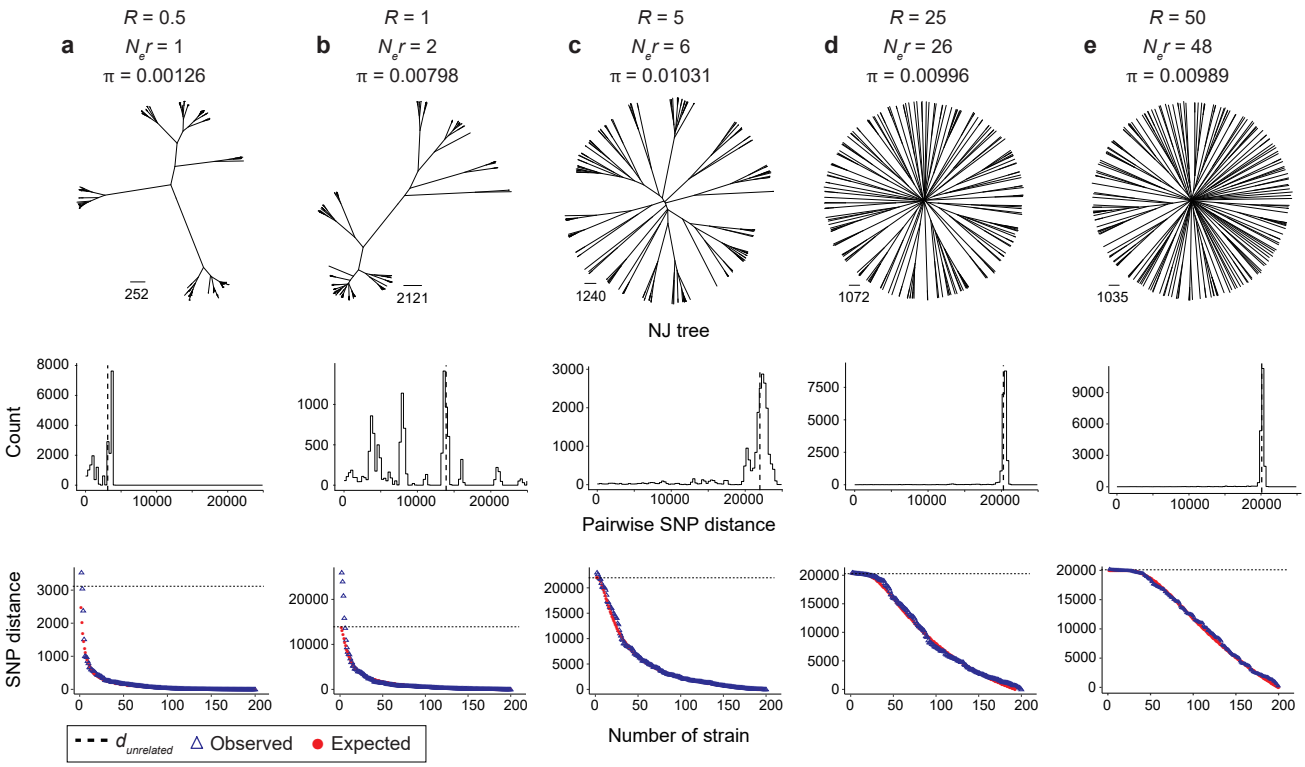

Figure 1. Recombination scaled effective population size ($N_e r$) estimation of simulated constant populations under different recombination effects. The panels indicate populations with recombination effects (R) of 0.5 (a), 1 (b), 5 (c), 25 (d), 50 (e). From top to bottom, indicating the NJ trees, distribution of pairwise SNP distance and $N_e r$ estimation of each population. Blue triangles and red points show observed and expected coalescence curves, respectively. Dotted lines show SNP distance of unrelated strains ($d_{unrelated}$) in coalescence formula.

Further simulations of more complex scenarios (Supplementary Fig. 1) show that $N_e r$ estimates reflect the current demography of the population more closely than estimates based on π, which is more influenced by past demography and migration to and from different demes. Supplementary Fig. 1a shows the effect of a population expansion on estimates of $N_e r$. At time t, a single population with scaled recombination rate $R = 5$ splits into two. The blue population maintains the ancestral population size while the red population becomes 10 times larger. At time $t + 0.05$ the red population is only marginally more diverse than the blue one but its estimated value of $N_e r$ is 6 fold higher. The fit of observed and expected genetic distances is poor for the red population, reflecting the inaccuracy of the modelling assumption that there is a single unchanging population size. At time $t + 0.2$ the estimate of $N_e r$ is close to its true post-split value for both populations and the model fit is substantially improved. Merging data from the two populations gives intermediate estimates of $N_e r$.

174 Supplementary Fig. 1b shows the effect of a population size reduction. A single population with R =
175 50 splits into two, with the blue population remaining unchanged while the red population undergoes
176 a 10 fold reduction in size. As in Supplementary Fig. 1a, the model fit for the red population is poor
177 immediately after the split but quickly improves, with the estimate of $N_e r$ approaching the correct
178 value of 5, while the nucleotide diversity of the population reduces much more slowly. Supplementary
179 Fig. 1c shows the effect of symmetric migration between two populations with a 10 fold difference in
180 effective population size. Migration has a large effect on nucleotide diversity, especially for the
181 smaller population, but has little effect on estimates of $N_e r$. Overall, these results show that our
182 inference approach provides accurate estimates of the recombination-scaled effective population size
183 for simulated data and that deviations between observed and expected genetic distances can be
184 informative about deviations from model assumptions.

185 **Application to *V. parahaemolyticus* genomes**

186 We first applied the method to the 1,103 *V. parahaemolyticus* genomes described in Yang et al[14]. For
187 this dataset, we obtained estimates of $N_e r$ = 484 and $\hat{n}$ = 471 (Fig. 2a). $\hat{n}$ is less than half of the
188 sample size principally because a large fraction of the isolates in the sample belong to pandemic
189 clonal lineages responsible for large numbers of human infections. Both lineages have most recent
190 common ancestors within the last few decades and are likely to represent a very small fraction of the
191 global population of *V. parahaemolyticus*. The genetic distances fit the model well except that the 33
192 oldest coalescent events, on the left hand side of the plot are larger than $d_{\text{unrelated}}$. The discrepancy is
193 largely due to population structure within the species, since the fit is better, although still not perfect
194 when analysis is restricted to the 944 isolates from the VppAsia population (Fig. 2a).

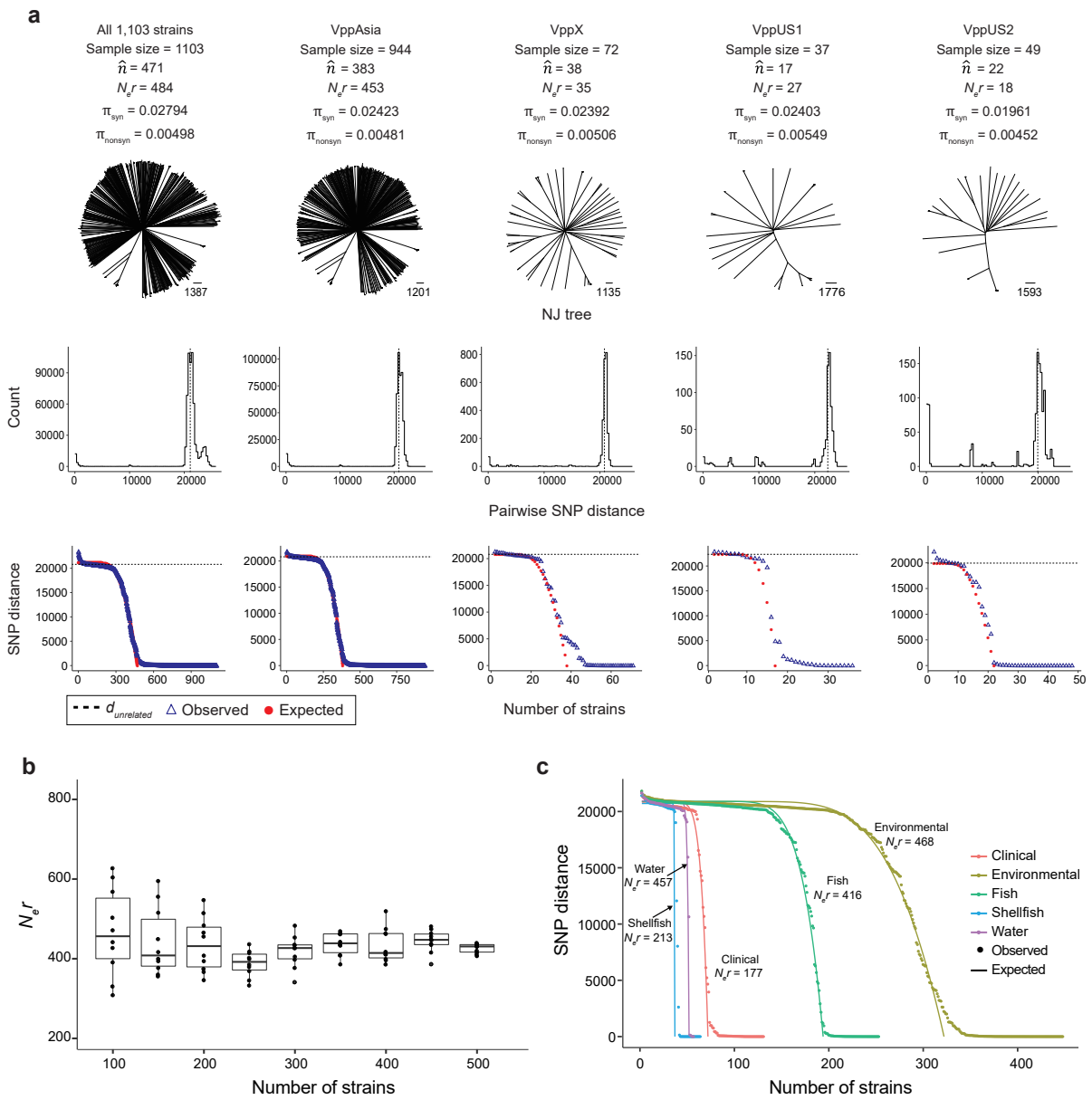

Figure 2. Recombination scaled effective population size ($N_e r$) estimation of *V. parahaemolyticus*. (a) $N_e r$ estimation of all the samples and four populations (VppAsia, VppX, VppUS1 and VppUS2) of *V. parahaemolyticus*. Layout and colors are the same as in Figure 1. (b) $N_e r$ estimation of VppAsia population based on different sample sizes. 100-500 genomes were randomly selected from total 944 VppAsia genomes, 10 repeats were performed for each sample size to create the boxplot. (c) $N_e r$ estimation of VppAsia population based on different types of samples. Points and lines show observed and expected coalescence curves, respectively.

For the VppAsia population $N_e r$ was estimated to be 453 which is similar to that for the dataset as a whole, while other populations have substantially smaller values. These differences are not simply due to a larger sample size since estimates based on subsets of the VppAsia data are consistently greater than 200 (Fig. 2b). Sampling strategy does make some difference, since estimates are lower

200 for a dataset consisting only of clinical strains than of shellfish, fish or all non-clinical isolates (Fig.
201 2c). This difference presumably reflects variation in disease causing potential, which results in
202 samples of clinical isolates not representing the full diversity of clonal lineages within the species.

203 **Application to multiple bacterial species**

204 We also applied the method to a survey of other bacteria for which large numbers of genomes are
205 publically available (Supplementary Fig. 2). *V. cholerae* has an estimate of $N_e r$ of 29, while *Vibrio*
206 *vulnificus* has a value of 43. Although the sample sizes available are relatively limited, in contrast to
207 the Asian population of *V. parahaemolyticus*, the genetic distances do not show a clear plateau
208 corresponding to a single value for $d_{\text{unrelated}}$ and while there are a continuous range of coalescent
209 distances, the overall fit between model and data is poor. Thus, although the other *Vibrio* species in
210 our dataset recombine frequently they are far from being panmictic.

211 *H. pylori* has the largest estimated value of $N_e r$ of all datasets we analysed. *H. pylori* is characterized
212 by extremely high rates of recombination during mixed infection of the human stomach, with 10% or
213 more of the genome recombined during a single infection[20] and linkage disequilibrium decreases
214 much more rapidly as a function of genetic distance than for all other species, including *V.*
215 *parahaemolyticus* (Fig. 3a). Although the tree is approximately star-like, the fit of the model is not
216 perfect, with no clear plateau for a single value of $d_{\text{unrelated}}$, due to the complex geographical
217 population structure of the species[21].

218 Amongst the other species, $N_e r$ varies from 1 for *Chlamydia trachomatis* to 88 for *Salmonella*
219 *enterica*. The overall fit of the model varies considerably between species (Supplementary Fig. 2) and
220 is typically worst for the oldest coalescent events, as examined in more detail above for *V.*
221 *parahaemolyticus*. $N_e r$ values are correlated with linkage disequilibrium statistics in the chromosome,
222 but show no correlation with nucleotide diversity π (Fig. 3).

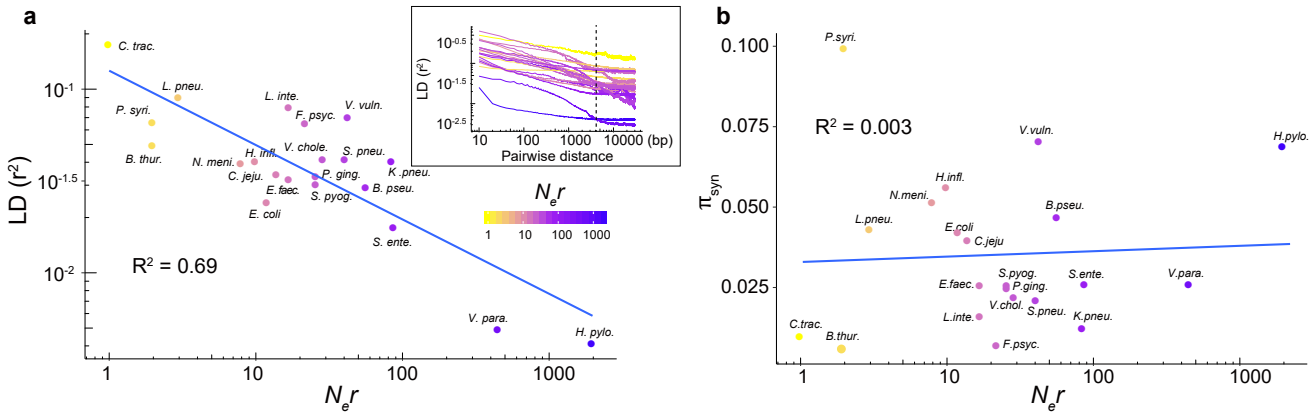

Figure 3. Correlation between $N_e r$ and short-range linkage disequilibrium (a) and between $N_e r$ and nucleotide diversity of synonymous sites (b) in 21 bacterial species. The r^2 values at pairwise distance of 3,000 bp were used in (a).

Discussion

We have estimated the recombination-scaled effective population size $N_e r$ in simulated data, where accurate estimates were obtained, and in 21 bacterial species for which large numbers of genomes were available. High values of $N_e r$ imply that there are many unrelated members in the population and low genome-wide LD and thus is consistent with panmixia. We obtained estimates of 453 for the Asian population of *V. parahaemolyticus* and 1,976 for the hpEurope population of *H. pylori,* but the estimated value of $N_e r$ was lower than 100 in the other 19 species.

Since panmixia is possible within bacterial populations, it raises the question why it is not widespread. There are several bacterial species in our sample which survive well in the environment, have effective global dispersal, for example by birds, enormous census population sizes and high recombination rates, such as *E. coli, Campylobacter jejuni* and *K. pneumoniae*. Species with these characteristics are the most obvious candidates to be panmictic, whereas *V. parahaemolyticus* thrives only in warm brackish waters and has oceanic gene pools, implying historical limits on its dispersal[14], and *H. pylori* only survives in human stomachs and shows geographic differentiation associated with historical migrations of its host[22].

A further puzzle lies in the lack of correlation between the estimated $N_e r$ and the nucleotide diversity π (Fig. 3b). For example, *V. parahaemolyticus* has similar diversity to *V. cholerae* and much lower

241 diversity than *V. vulnificus*, both of which have estimated $N_e r$ lower than 50. Furthermore, estimates
242 of $N_e r$ vary over a factor of 1,000, while π varies only by a factor of 10, consistent with previous
243 observations of much lower variation in π between species than is expected based on variation in
244 census population size in many domains of life[23]. This phenomenon remains poorly understood,
245 despite decades of attempts at an explanation[24].

246 We propose that panmixia is rare in bacteria because it is unstable and does not persist over
247 evolutionary time. Specifically, high $N_e r$, if sustained, leads to an accumulation of diversity to levels
248 higher than observed in any species. We propose that an accumulation of diversity leads to the
249 evolution of barriers to genetic exchange within the species that reduce $N_e r$[25]. It follows that the
250 populations with high values of $N_e r$ we found have arisen recently on an evolutionary timescale.

251 As proposed in Yang et al[14], the Asian population of *V. parahaemolyticus*, VppAsia, has spread
252 within the last few decades due to human activity but has an ancestral range restricted to coastal
253 waters from India to Japan. The other *V. parahaemolyticus* populations that we have sampled, with
254 different ancestral ranges, have smaller estimated $N_e r$, perhaps because they have smaller or less
255 fecund ranges or experience greater competition or more frequent demographic bottlenecks.

256 Crucially, it seems that the ancestral population size of *V. parahaemolyticus* as a whole was smaller
257 than currently found in VppAsia and more similar to that in the other populations. First, the site
258 frequency spectrum of VppAsia, but not the other populations, is out of equilibrium and is
259 approximately consistent with a demographic scenario in which the effective population size
260 increased by a factor of ten 15,000 years ago (Supplementary Fig. 3). One possible scenario is that the
261 end of the last ice age around 11,000 years ago created a habitat that existed till the present and that
262 the population expansion implied by the site frequency spectrum is due to a demographic expansion at
263 the end of the ice age. Second, r/μ, the number of recombinant sites for each mutant site, has been
264 estimated for the species to be around 313[14]. Note that this parameter r/μ is not the same as the
265 parameter r/m often used in the literature[3] because the former considers all recombined sites whereas
266 the latter considers only recombinant sites that are substituted. If nucleotide diversity reached an

267 equilibrium, consistent with the current value of $N_e r$ of VppAsia, this would predict a value of

268 $2N_e\mu = 2N_e r\left(\frac{\mu}{r}\right) = 2 \times 453 \times \left(\frac{1}{313}\right) = 2.9.$ At equilibrium in a neutrally evolving population,

269 with sites evolving according to the Jukes-Cantor model, π is predicted to be equal to $\frac{3}{4} \times$

270 $\left(1 - e^{-\frac{8N_e\mu}{3}}\right) = 0.73$ and therefore most synonymous sites should differ between individuals. This is

271 far from the current value of 0.024 or indeed diversity levels in any well-characterised bacterial

272 species. Our simulation results show that estimates of $N_e r$ respond much more quickly to changes in

273 demography than π, implying that diversity levels would be likely to continue a slow but steady rise if

274 the effective population size remains constant (ignoring the changes in population structure caused by

275 human activity in recent decades).

276 The equivalent argument is more complex in *H. pylori* due to its geographic population structure but

277 levels of variation are clearly far from demographic equilibrium, as illustrated by the substantial

278 variation between populations that reflects bottlenecks associated with historical human migration

279 rather than current population sizes[22].

280 Simulations of facultatively sexual organisms have shown that selection on multiple interacting loci

281 can lead to populations being dominated by small numbers of clones, even in the presence of frequent

282 recombination[26]. These simulations also show that a phase transition from panmixia can take place

283 due to a decrease in recombination rate or even an increase in population size.

284 There are several mechanisms which might lead to similar phase transition behaviour in bacterial

285 populations as diversity increases. In Cui et al[27], we observe nascent boundaries to genetic exchange

286 between lineages of *V. parahaemolyticus* associated with differentiation into ecological types. This

287 kind of genetic differentiation is more likely to be common, and more likely to lead to barriers to

288 exchange in more diverse bacterial populations, which tend to have more diverse accessory

289 genomes[28]. Such barriers can result in speciation[18,29] but can also lead to genetic structuring within a

290 single species, such as host-specific gene pools found in *C. jejuni*[30]. Another mechanism is changes in

291 the pattern or rate of recombination. Recombination requires homology between donor and recipient

292 DNA to be recognized by the cellular machinery and, in addition, most species have a mismatch

293 repair system, which aborts the process if there are too many differences. Therefore, as homology
294 decreases, so will recombination[20,29]. Mismatch dependent recombination is likely to be the main
295 reason why *E. coli* has an estimated value of $N_e r$ of only 12, implying that recombination is
296 ineffective in reassorting variation across the whole species[31].

297 Our arguments imply that many environmental organisms with large census population sizes are
298 likely to have gone through a stage of high $N_e r$ values, similar to that currently observed in the Asian
299 population of *V. parahaemolyticus,* followed by a phase transition which reduced the number of
300 lineages circulating in the population, which also acted to prevent a continuous increase in nucleotide
301 diversity by increasing the rate of genetic drift and led either to long-term stability in population
302 structure or to occasional speciation events such as described in *Salmonella*[29]. Further support for this
303 hypothesis could be provided by identifying species that have endured such a transition and
304 characterizing the factors that currently mitigate against recombination between lineages.

305 Despite decades of study by population geneticists, the factors determining the amount of diversity
306 within species are still poorly understood. Our results suggest that patterns of variation in bacteria are
307 likely to be dynamic, with ecological and genomic differentiation, barriers to gene flow and
308 demographic factors interacting with each other in complex ways to produce the wide variety of
309 bacterial population structures that are observed.

310

311 **Material and methods**

312 **Genomic datasets used in analysis**

313 Multiple genome sequences from 21 bacterial species were used in this analysis, including *Bacillus*
314 *thuringiensis, Burkholderia pseudomallei, Campylobacter jejuni, Chlamydia trachomatis,*
315 *Enterococcus faecalis, Escherichia coli, Flavobacterium psychrophilum, Haemophilus influenzae,*
316 *Helicobacter pylori, Klebsiella pneumoniae, Legionella pneumophila, Leptospira interrogans,*
317 *Neisseria meningitidis, Porphyromonas gingivalis, Pseudomonas syringae, Salmonella enterica,*
318 *Streptococcus pneumoniae, Streptococcus pyogenes, Vibrio cholerae, Vibrio parahaemolyticus* and

319 *Vibrio vulnificus.* The SNP data from 278 *H. pylori* and 1,103 *V. parahaemolyticus* genomes were
320 reused from previous studies[14,32]. Other genome sequences were publicly available and were
321 downloaded from NCBI database. For species that have more than 500 genomes in NCBI database,
322 such as *E. coli*, 500 genomes were randomly selected and were used first to reconstruct a Neighbour-
323 joining tree. Strains located on extremely long branches and belonging to clonal groups were
324 manually removed, finally resulting in a dataset of 6,355 genomes. The accessions of genomes used
325 (excluding *H. pylori* and *V. parahaemolyticus*) are listed in Supplementary Table 1, and the whole-
326 genome alignment files are available in the figshare data repository
327 (https://figshare.com/s/3f9d04a8229f30dd785b).

328 **Simulation**

329 The software FastSimBac[19] was used to generate the simulated bacterial populations under different
330 hypothetical evolution scenarios, including a constant population with different recombination effects
331 (Fig. 1), and the changing populations with population growth (Supplementary Fig. 1a), decline
332 (Supplementary Fig. 1b) and migration (Supplementary Fig. 1c). All the simulated genome length was
333 set to 2 Mb, mutation rate was fixed at 0.01 (per site per $2N_e$ generations). The detailed parameters
334 used in simulation are listed in Supplementary Table 2.

335 **SNP calling, phylogeny reconstruction and LD decay calculation**

336 The SNPs of publicly available genome sequences were called as previously described[5,14]. Briefly, the
337 individual genomes were aligned against a reference genome of the corresponding species
338 (Supplementary Table 1) using MUMmer[33], then the detected core bi-allelic SNPs were summarised
339 as a SNP matrix for each species. The Neighbour-joining trees were constructed using the software
340 Treebest (http://treesoft.sourceforge.net/treebest.shtml) based on concatenated SNPs. Haploview[34]
341 was used to calculate the r^2 value based on previous and new identified SNP sets, with the maximum
342 comparison distance set to 30 kb. A sliding window method was used to summarize the r^2 value with
343 a window of size 10 bp.

344 **Recombination scaled effective population size estimation**

345 For a freely mixing population in which recombination drives diversification, neutral theory predicts
346 that $N_e r$ (the product of the effective population size N_e and the recombination rate r) is in proportion
347 to genealogical coalescent rate, and the expected coalescent curves can be estimated based on the
348 formula $d = d_{\text{unrelated}}(1 - \exp(-4N_e r\left(\frac{1}{m} - \frac{1}{n}\right))$ as mentioned in previous work[5]. In the formula, d
349 is the expected pairwise genetic divergence, $d_{\text{unrelated}}$ is the median pairwise SNP distance of a
350 population, n is the number of individual strains and m is the index of the ancestral node along the
351 coalescent tree of n strains. Coalescent curves can be estimated using the UPGMA algorithm based on
352 the pairwise SNP distance. By fitting the expected and observed coalescent curves with least square
353 method to search for the optimal parameters, we found the optimal values of $\hat{n}$ and $N_e r$ that were used
354 in further analysis.

355 **Site frequency spectrum estimation**

356 The minor allele frequency (MAF) of each SNP locus in a SNP matrix was calculated and then the
357 distribution of MAF was counted to generate the site frequency spectrum (SFS) for each dataset. To
358 get a comparable result, the SFS showed in each Figure or panel was calculated based on same sample
359 size. In Supplementary Fig. 3a, the merged population was generated by randomly selecting 100
360 genomes from each of Pop1 and Pop2. In Supplementary Fig. 3b, the same number of genomes as in
361 the real genome dataset were simulated to show populations with different evolution scenarios.

362 **Acknowledgements**

363 This work is supported by the National Key Research & Development Program of China (No.
364 2017YFC1200800, 2017YFC1601503, and 2016YFC1200100), the National Key Program for
365 Infectious Diseases of China (No. 2017ZX10104002), the National Natural Science Foundation of
366 China (No. 31770001) and Sanming Project of Medicine in Shenzhen (No. SZSM201811071). D.F. is
367 funded by a Medical Research Council Fellowship as part of the MRC CLIMB consortium for
368 microbial bioinformatics (grant number MR/M501608/1).

369 **Author Contributions**

D. F., Y. C., and R. Y. designed the study and coordinated the project; C. Y., Y. C., X. D., and D. F. analyzed the data; D. F. wrote the manuscript. All authors approved the final version of the manuscript.

Competing Financial Interests statement

None

References

1 Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc Natl Acad Sci U S A* **115**, 6506-6511, doi:10.1073/pnas.1711842115 (2018).

2 Stoppe, N. C. *et al.* Worldwide Phylogenetic Group Patterns of Escherichia coli from Commensal Human and Wastewater Treatment Plant Isolates. *Frontiers in microbiology* **8**, 2512, doi:10.3389/fmicb.2017.02512 (2017).

3 Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**, 199-208, doi:10.1038/ismej.2008.93 (2009).

4 Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**, 4384-4388 (1993).

5 Cui, Y. *et al.* Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus. *Molecular biology and evolution* **32**, 1396-1410, doi:10.1093/molbev/msv009 (2015).

6 Wright, S. The genetical structure of populations. *Annals of eugenics* **15**, 323-354 (1951).

7 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

8 Wang, J. Estimation of effective population sizes from data on genetic markers. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **360**, 1395-1409, doi:10.1098/rstb.2005.1682 (2005).

9 Didelot, X. & Maiden, M. C. Impact of recombination on bacterial evolution. *Trends in microbiology* **18**, 315-322, doi:10.1016/j.tim.2010.04.002 (2010).

10 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

11 Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial speciation. *Science* **315**, 476-480, doi:10.1126/science.1127573 (2007).

12 Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251-1266, doi:10.1534/genetics.106.063305 (2007).

13 Didelot, X., Lawson, D., Darling, A. & Falush, D. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435-1449, doi:10.1534/genetics.110.120121 (2010).

14 Yang, C. *et al.* Recent mixing of Vibrio parahaemolyticus populations. *bioRxiv*, doi:10.1101/354761 (2018).

15 Gibson, B., Wilson, D. J., Feil, E. & Eyre-Walker, A. The distribution of bacterial doubling times in the wild. *Proceedings. Biological sciences* **285**, doi:10.1098/rspb.2018.0789 (2018).

16 Kingman, J. F. C. The coalescent. *Stochastic processes and their applications* **13**, 235-248 (1982).

17 Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature reviews. Genetics* **3**, 380-390, doi:10.1038/nrg795 (2002).

18 Fraser, C., Hanage, W. P. & Spratt, B. G. Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci U S A* **102**, 1968-1973, doi:10.1073/pnas.0406993102 (2005).

415 19 De Maio, N. & Wilson, D. J. The Bacterial Sequential Markov Coalescent. *Genetics* **206**, 333-
416 343, doi:10.1534/genetics.116.198796 (2017).
417 20 Cao, Q. *et al.* Progressive genomic convergence of two Helicobacter pylori strains during
418 mixed infection of a patient with chronic gastritis. *Gut* **64**, 554-561, doi:10.1136/gutjnl-2014-
419 307345 (2015).
420 21 Falush, D. *et al.* Traces of human migrations in Helicobacter pylori populations. *Science* **299**,
421 1582-1585, doi:10.1126/science.1080857 (2003).
422 22 Linz, B. *et al.* An African origin for the intimate association between humans and
423 Helicobacter pylori. *Nature* **445**, 915-918, doi:10.1038/nature05562 (2007).
424 23 Lynch, M. Evolution of the mutation rate. *Trends in genetics : TIG* **26**, 345-352,
425 doi:10.1016/j.tig.2010.05.003 (2010).
426 24 Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within
427 species? *PLoS biology* **10**, e1001388, doi:10.1371/journal.pbio.1001388 (2012).
428 25 Ansari, M. A. & Didelot, X. Inference of the properties of the recombination process from
429 whole bacterial genomes. *Genetics* **196**, 253-265 (2014).
430 26 Neher, R. A. & Shraiman, B. I. Competition between recombination and epistasis can cause a
431 transition from allele to genotype selection. *Proc Natl Acad Sci U S A* **106**, 6866-6871,
432 doi:10.1073/pnas.0812560106 (2009).
433 27 Cui, Y. *et al.* The landscape of coadaptation in Vibrio parahaemolyticus. *bioRxiv*,
434 doi:10.1101/373936 (2018).
435 28 Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective
436 population size. *ISME J* **11**, 1719-1721, doi:10.1038/ismej.2017.36 (2017).
437 29 Falush, D. *et al.* Mismatch induced speciation in Salmonella: model and data. *Philosophical*
438 *transactions of the Royal Society of London. Series B, Biological sciences* **361**, 2045-2053,
439 doi:10.1098/rstb.2006.1925 (2006).
440 30 Sheppard, S. K. *et al.* Cryptic ecology among host generalist Campylobacter jejuni in
441 domestic animals. *Molecular ecology* **23**, 2442-2451, doi:10.1111/mec.12742 (2014).
442 31 Didelot, X., Meric, G., Falush, D. & Darling, A. E. Impact of homologous and non-homologous
443 recombination in the genomic evolution of Escherichia coli. *BMC genomics* **13**, 256,
444 doi:10.1186/1471-2164-13-256 (2012).
445 32 Berthenet, E. *et al.* A GWAS on Helicobacter pylori strains points to genetic variants
446 associated with gastric cancer risk. *BMC biology* **16**, 84, doi:10.1186/s12915-018-0550-3
447 (2018).
448 33 Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in
449 large sequence sets. *Current protocols in bioinformatics* **Chapter 10**, Unit 10 13,
450 doi:10.1002/0471250953.bi1003s00 (2003).
451 34 Barrett, J. C. Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harbor*
452 *protocols* **2009**, pdb ip71, doi:10.1101/pdb.ip71 (2009).

453