Why panmictic bacteria are rare

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

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

2 Department of Infectious Disease Epidemiology, Imperial College, Norfolk Place, London W2 1PG, UK

3 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

Bacterial populations are typically much more structured than those of higher eukaryotes but this difference is surprising given high recombination rates, enormous population sizes and effective dispersal in many bacterial species. We estimated the recombination scaled effective population size $N_e r$ in 21 bacterial species and find only two have estimates of 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 100 fold in *V. parahaemolyticus* if current values of $N_e r$ is maintained, to values much higher than that found in any species. We propose that panmixia is unstable in bacteria, and that long-lived environmental species are likely to have evolved some form of barriers to genetic exchange, which act to prevent a continuous increase in diversity by enhancing the rate of genetic drift. Our results highlight the dynamic nature of genetic structure in bacterial populations and imply that overall diversity levels found within the species are likely to be poorly predictive of the number of organisms.

Introduction

Bacteria are paragons of adaptability and make up more biomass than all organisms other than plants combined (Bar-On, Phillips, and Milo 2018). Many bacterial species have enormous population sizes, disperse effectively around the globe (Stoppe et al. 2017) and exhibit high rates of within-species homologous recombination (Vos and Didelot 2009). Recombination acts to progressively break down non-random associations between markers (linkage disequilibrium), so that in large populations a natural null expectation is for there to be linkage equilibrium between sites. Nevertheless, in most species for which data is available, substantial genome-wide structuring of variation is observed (Smith et al. 1993). 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

40 has undergone a recent increase in size. If this population size was maintained over evolutionary
41 timescales would lead to a greater than 100 fold accumulation of diversity, to levels higher than found
42 in any well characterized bacterial species. We propose that other environmental species with large
43 census population sizes, for example *Vibrio cholerae* or *Klebsiella pneumoniae*, may have been
44 through a similar stage before accumulating diversity and then evolving mechanisms that disrupted
45 free recombination.

46 Specifically, we use pairwise genetic distances between strains calculated from genome sequences to
47 estimate the recombination scaled effective population size $N_e r$, where N_e is the effective population
48 size and r is the recombination rate per site per generation, an approach briefly introduced by Cui et al
49 (Cui et al. 2015). Here, we test the method using simulations and find that it provides accurate
50 estimates in constant size populations for values of scaled recombination rate R (see below) of 5 or
51 more, and responds to changes in population size or structure far more quickly than estimates of N_e
52 based on nucleotide diversity. We apply the method to data from 21 species and find that some
53 common species, such as *Escherichia coli*, have strikingly low estimates of $N_e r$, implying that
54 recombination is inefficient in distributing diversity across the species, while for others the fit of
55 model to data is poor, highlighting deviations from a freely recombining population. Only datasets
56 from *V. parahaemolyticus* and *Helicobacter pylori* gives $N_e r$ estimates greater than 100. The former
57 has had a recent increase in population size, the latter experiences repeated bottlenecks associated
58 with geographical spread of its human host, implying neither population is close to demographic
59 equilibrium.

60 **Inference approach**

61 Population genetics is a century old discipline that provides a powerful set of theoretical and statistical
62 inference tools with which to interpret patterns of variation between closely related organisms.
63 Central to the theory is the concept of a population, which in outbreeding eukaryotes is set of
64 organisms that share a common gene pool. The simplest models assume panmixia which is random
65 mating of all individuals within a single closed population (Wright 1951). In most animals and plants
66 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, making random mating a reasonable first approximation for many species. Theory predicts that the level of genetic diversity in depends on the mutation rate and on the historical effective population size, which is the number of individuals that contribute to the gene pool in each generation (Wang 2005).

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 random samples from populations with more than a few hundred individuals, shared recent ancestry is rare and can be neglected in many types of analysis.

Bacteria reproduce by binary fission and only receive material from the gene pool or contribute to it via homologous recombination. Many cell divisions can take place between recombination events, most of which only affect a small proportion 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 estimate 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 programs like STRUCTURE (Pritchard, Stephens, and Donnelly 2000). For example, nucleotide diversity in a species like *E. coli* can be estimated for clonal lineages, phylogroups, species or for the *Escherichia* complex 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 justified. All methods lead to estimates of effective population size that are many orders of magnitude smaller than the census number of bacteria.

Here, we take a different approach which is to estimate a scaled version of the effective population size, $N_e r$. High values of $N_e r$ indicate that the population structure is similar to that of outbreeding eukaryotes. Informally, in eukaryotic population each individual is the product of a separate meiosis

94 and therefore genetically distinct. N_e is the number of genetically distinct individuals that contribute
95 in each generation, which is typically in the thousands or millions. $N_e r$ is designed to measure an
96 analogous quantity; the number of genetically distinct organisms that contribute to the future
97 bacterial gene pool. To this end, time is rescaled; N_e measures the rate of genetic drift per bacterial
98 generation, while $N_e r$ measures the genetic drift in the time it takes for strains to become similarly
99 distinct by importing half of their DNA by homologous recombination. The recombination rate of *V.*
100 *parahaemolyticus* is 1.7×10^{-4} per site per year (Yang et al. 2018), as expected number of differences
101 between two strains that affected by recombination through T years ago is e^{-rT}, it can be calculated
102 that it takes around 4,000 years on average for half of the genome to recombine. Half of this time is
103 approximately equivalent to a eukaryotic generation in the sense that two strains that shared a
104 common ancestor 2,000 years ago will be about as related as siblings. This represents a very different
105 time scale from that assumed based on bacterial generations. For example, in *V. parahaemolyticus*,
106 bacterial doubling times are likely to be similar to those in *V. cholerae*, which have been estimated at
107 around an hour (Gibson et al. 2018).

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

117 To use these times to estimate the effective population size, we assume that the genealogy of clonal
118 relationships is generated by a coalescent model with a constant population size N_e. This model
119 generates expectations for the times in the past at which common ancestors of strains in a sample
120 lived. Specifically, in a sample of n strains, there are n-1 coalescence nodes. The age of the first node

corresponds to the common ancestor of the two most closely related strains in the sample, while the *n-1*th node corresponds to the common cellular ancestor of all the strains in the sample.

Coalescence theory implies that the expected time in the past at which the mth most recent coalescence occurs is $T_m = 2N_e(1/m - 1/n)$. These times can be converted into expected genetic distances d_m using the formula above. We use the UPGMA algorithm to obtain *n-1* coalescence distances from the *(n-1)* × *n* pairwise genetic distances between strains and find the values of $N_e r$ and $\hat{n}$ (effective sample size, see below) that gives the best fit between observed and expected distances for the *n-1* coalescence nodes.

Note that for bacteria with high recombination rate, the genome is likely to have been scrambled up sufficiently that strains will have inherited little or no DNA by direct descent from the common ancestor of the entire sample. This means that the genetic distances expected for the oldest coalescences plateau at $d_{unrelated}$. In plots, it is convenient to show the coalescences in chronological order, oldest first, to aid comparisons between datasets with different sample sizes (e.g. Figure 1).

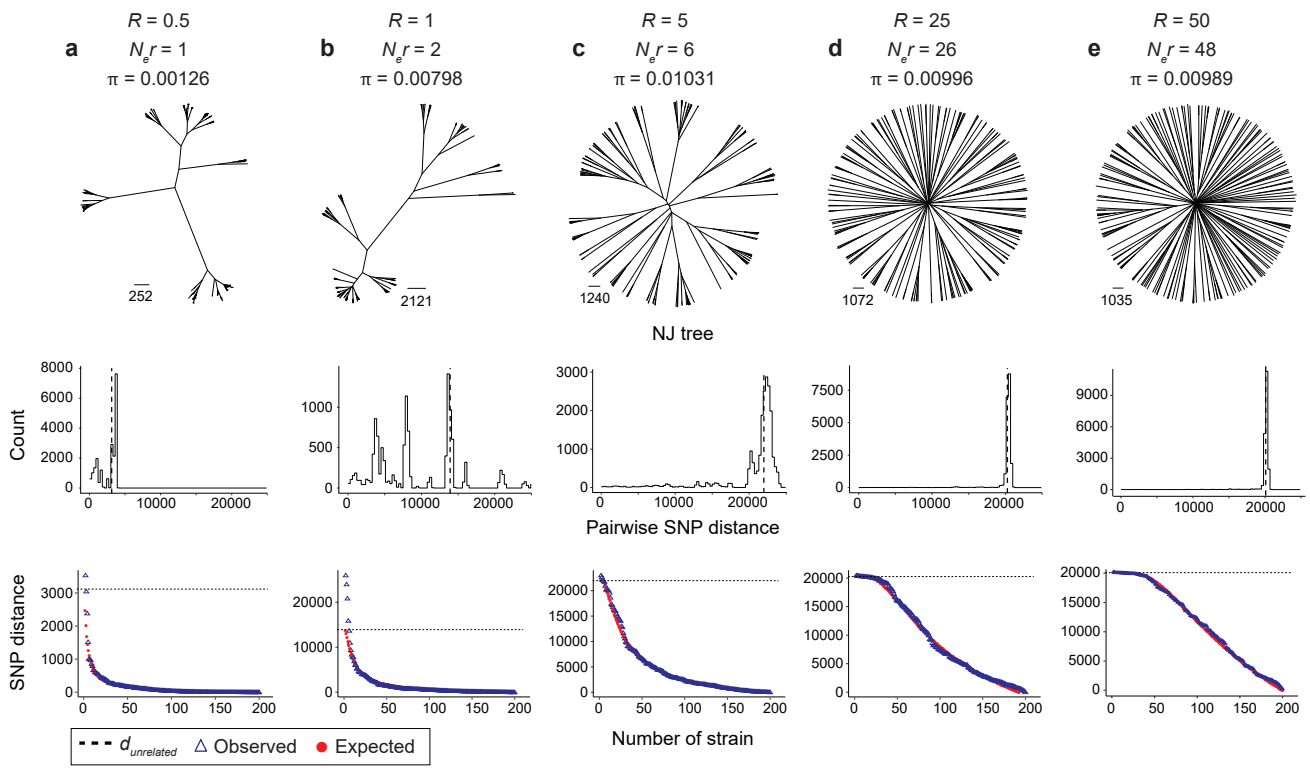

Figure 1. Recombination scaled effective population size ($N_e r$) estimation of simulated constant populations under different recombination effects. The panels indicate populations with recombination effects (*R*) of 0.5 (a), 1 (b), 5 (c), 25 (d), 50 (e). From top to bottom, indicating the NJ trees, 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.

135 The model assumes that the strains are randomly sampled from a homogeneous population at a single
136 time point, but pathogenic clones, epidemic outbreaks or strains from specific locations are often
137 overrepresented in real data, leading to oversampling of very closely related isolates, relative to their
138 frequency in the global population. Therefore, in addition to $N_e r$ we estimate a second parameter $\hat{n}$,
139 indicated the effective sample size, which can be considered number of samples that are not epidemic
140 clones of each other. For simulated data, $\hat{n}$ is correctly estimated to be very close to number of strains
141 in the sample, so this additional parameter makes little difference to the inference. For real data,
142 estimating this additional parameter often improves the qualitative model fit considerably.

143 **Results**

144 **$N_e r$ can be estimated accurately for simulated data**

145 Figure 1 illustrates the effect of varying the recombination rate in genomes simulated using
146 FastSimBac (De Maio and Wilson 2017). FastSimBac assumes a coalescent model with initial
147 population size N_0, which remains constant in the simulations shown in this figure. We fix the
148 recombination tract length $\lambda = 1000$ and vary the rate of initiation of recombination events $r*$ per
149 $2N_0$ generations. We report the scaled recombination rate $R = \frac{\lambda r^*}{2N_0}$. For low recombination rates, the
150 phylogenetic tree is highly structured but becomes progressively bushier as R increases, and gives the
151 impression of being a nearly perfect star for $R = 50$. For high R, estimated $N_e r$ is close to R and the
152 observed genetic distances are well-fit by the model. For lower R, the fit is less good and the estimate
153 of R provided by $N_e r$ is also less accurate. For these parameter values, the size and shape of the
154 phylogenetic tree are highly variable between runs (not shown), and the poorer fit reflects this
155 stochasticity as well as greater inaccuracy in the approximations made in converting between genetic
156 distances and coalescence times.

157 Further simulations of more complex scenarios (Figure S1) shows that $N_e r$ estimates reflect the
158 current demography of the population that it is measured from more closely than estimates based on
159 π, which is more influenced by past demography and migration to and from different demes. Figure
160 S1a shows the effect of a population expansion on estimates of $N_e r$. At time 0, a single population

161 with scaled recombination rate R = 5 splits into two. The blue population maintains the ancestral
162 population size while the red population becomes 10 fold larger. 0.05 time units after the split the red
163 population is only marginally more diverse than the blue one but its estimated value of $N_e r$ is 6 fold
164 higher. The fit of observed and expected genetic distances is poor for the red population, reflecting the
165 inaccuracy of the modelling assumption that there is a single unchanging population size. 0.2 time
166 units after the split the estimate of $N_e r$ is close to its true post-split value for both populations and the
167 model fit is substantially improved. Merging data from the two populations gives intermediate
168 estimates of $N_e r$.

169 Figure S1b shows the effect of a population size reduction. A single population with R = 50 splits into
170 two, with the blue population remaining unchanged while the red population undergoes a 10 fold
171 reduction in size. As in Figure S1a, the model fit for the red population is poor in the immediate
172 aftermath of the split but quickly improves, with the estimate of $N_e r$ approaches the correct value of
173 5, while the nucleotide diversity of the population reduces much more slowly. Figure S1c shows the
174 effect of symmetric migration between two populations with 10 fold differences in effective
175 population size. Migration rates have a large effect on nucleotide diversity, especially for the smaller
176 population, but has little effect on estimates of $N_e r$. Overall, these results show that $N_e r$ provides
177 accurate estimates of the recombination-scaled effective population size for simulated data and that
178 deviations between observed and expected genetic distances can be informative about deviations from
179 model assumptions.

180 **Application to data from natural isolates**

181 We first applied the method to the 1,103 *V. parahaemolyticus* described in (Yang et al. 2018). For this
182 dataset, we obtained estimates of $N_e r$ of 484 and $\hat{n}$ of 471 (Figure 2a). $\hat{n}$ is less than half of total
183 number of samples principally because a large fraction of the isolates in the sample belong to
184 pandemic clonal lineages responsible for large numbers of human infections. These lineages arose
185 within the last few decades and are likely to represent a very small fraction of the global population of
186 *V. parahaemolyticus*. The genetic distances are well fit by the model except that the 33 oldest
187 coalescences, on the left hand side of the plot are larger than $d_{unrelated}$. The discrepancy is largely due

188 to population structure within the species, since the fit is better, although still not perfect when

189 analysis is restricted to the 944 VppAsia population isolates (Figure 2a).

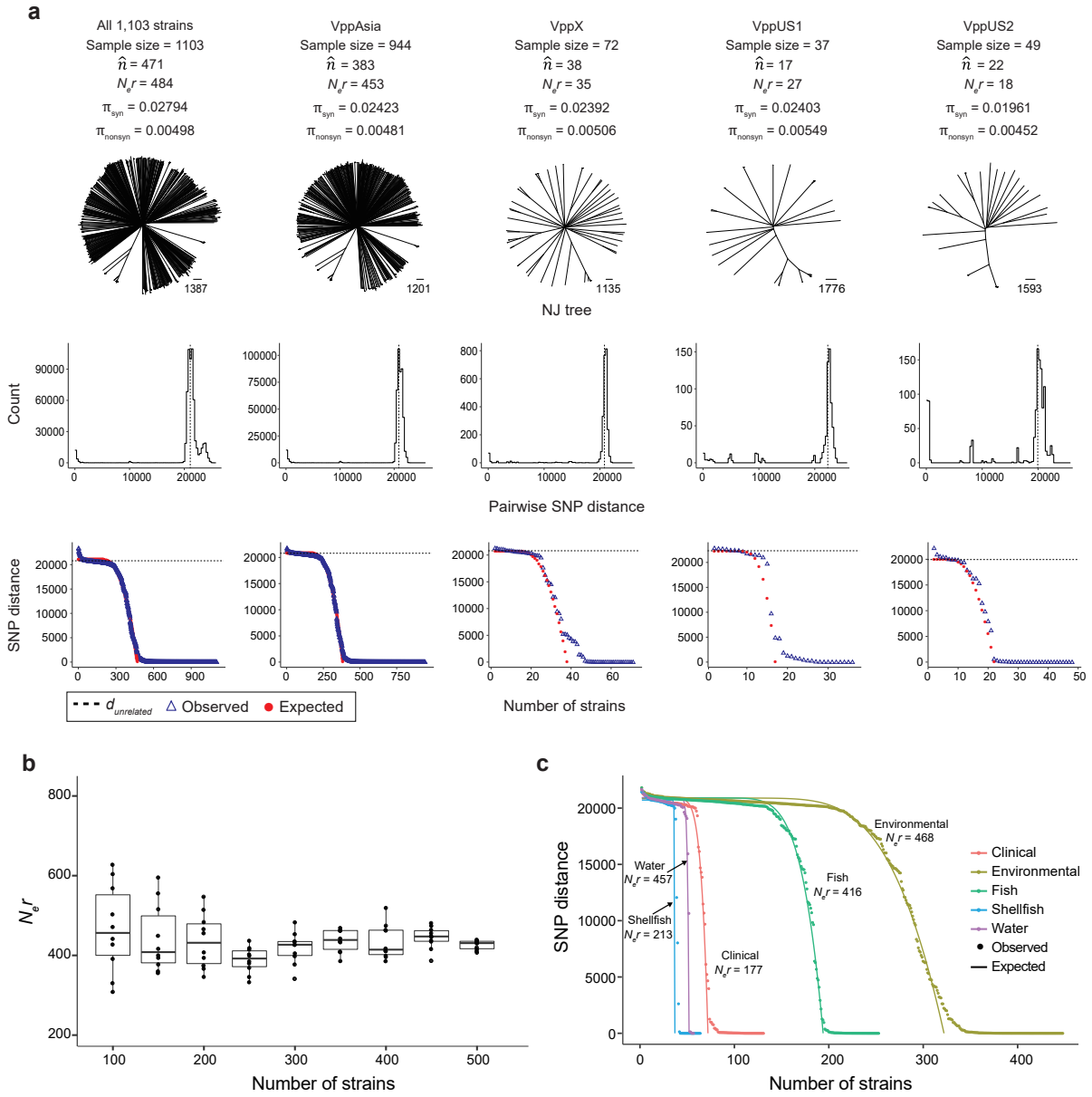

Figure 2. Recombination scaled effective population size ($N_e r$) estimation of *V. parahaemolyticus*. (a) $N_e r$ estimation of all the samples and four populations (VppAsia, VppX, VppUS1 and VppUS2) of *V. parahaemolyticus*. Layout and colors are the same as in Figure 1. (b) $N_e r$ estimation of 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.

191 $N_e r$ estimated for the VppAsia of 453 is similar to that for the dataset as a whole, while other

192 populations have substantially smaller values. These differences are not simply due to a larger sample

193 sizes since estimates based on subsets of the VppAsia data are reliably greater than 200 (Figure 2b).

194 Sampling strategy does make some difference, since estimates are lower for a dataset consisting only
195 of clinical strains than of shellfish, fish or all non-clinical isolates (Figure 2c). This difference
196 presumably reflects variation in disease causing potential, which results in samples of clinical isolates
197 not representing the full diversity of clonal lineages within the species.

198 We also applied the method to a survey of other bacteria for which large numbers of genomes are
199 publically available (Figure S2). *V. cholerae* has an estimate of $N_e r$ of 29, while *Vibrio vulnificus* has
200 a value of 43. Although the sample sizes available are relatively limited, in contrast to the Asian
201 population of *V. parahaemolyticus*, the genetic distances do not show a clear plateau corresponding to
202 a single value for $d_{unrelated}$ and while there are a continuous range of coalescence distances, the overall
203 fit between model and data is poor. Thus, although the other *Vibrio* species in our dataset recombine
204 frequently they are not close to being panmictic.

205 *H. pylori* has the largest estimated value of $N_e r$ of all data. *H. pylori* is characterized by extremely
206 high rates of recombination during mixed infection of the human stomach, with 10% or more of the
207 genome recombined during a single infection (Cao et al. 2015) and linkage disequilibrium falls off
208 much more rapidly as a function of genetic distance than for all other species, including *V.*
209 *parahaemolyticus* (Figure 3a). Although the tree is approximately star-like, the fit of the model is not
210 perfect, without a clear plateau for a single value of $d_{unrelated}$, due to the complex geographical
211 population structure of the species.

212 Amongst other species, $N_e r$ varies between 1 for *Chlamydia trachomatis* and 88 for *Salmonella*
213 *enterica*. The overall fit of the model varies considerably between species (Figure S2) and is typically
214 worst for the oldest coalescences, as examined in more detail for *V. parahaemolyticus* above. $N_e r$
215 values are correlated with linkage disequilibrium statistics in the chromosome, but show no
216 correlation with nucleotide diversity π (Figure 3).

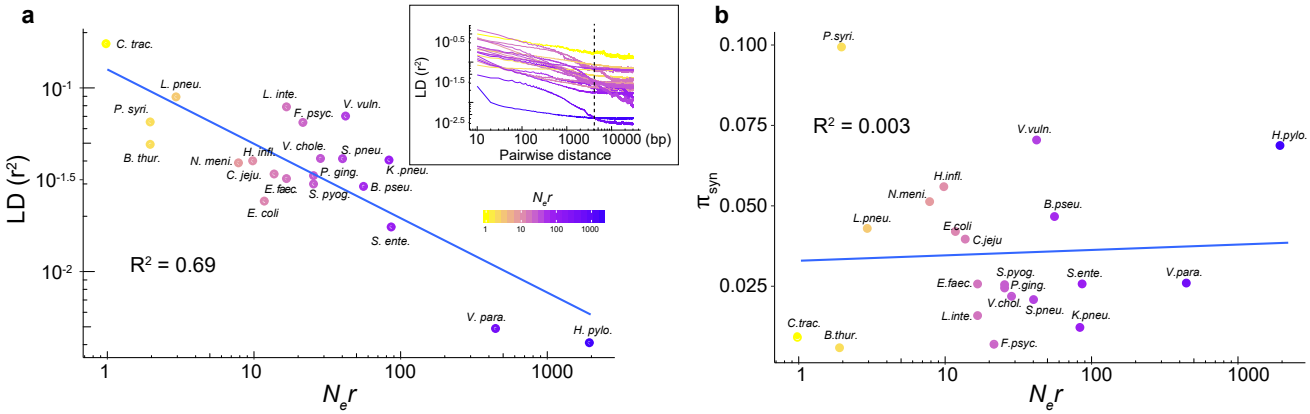

Figure 3. Correlation between $N_e r$ and short-range linkage disequilibrium (a) and between $N_e r$ and nucleotide diversity of synonymous sites (b) in 21 bacterial species. The r^2 values at pairwise distance of 3,000 bp were used in (a).

Discussion

We have estimated the recombination-scaled effective population size $N_e r$ in simulated data, where it provides accurate estimates, and from 21 bacterial species for which large numbers of genomes were available. High values of $N_e r$ imply that there are many unrelated clones 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 1976 for hpEurope population of *H. pylori* but the other 19 species all had estimated $N_e r$ lower than 100.

Given that panmixia is possible within bacterial populations, it becomes puzzling 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, but *V. parahaemolyticus* thrives only in warm brackish waters and has oceanic gene pools, implying historical limits on its dispersal (Yang et al. 2018), while *H. pylori* is only survives in human stomachs and shows geographic differentiation associated with historical migrations of its host (Linz et al. 2007).

A further puzzle is there is no correlation evident between estimated $N_e r$ and nucleotide diversity π (Figure 3b), which at equilibrium in a neutrally evolving population is predicted to equal $2N_e\mu$. For example, *V. parahaemolyticus* has similar diversity to *V. cholerae* and much lower diversity than *V.*

236 *vulnificus*, both of which have estimated $N_e r$ lower than 50. Furthermore, estimated $N_e r$ varies over a
237 factor of 1000, while π, varies only by a factor of 10, consistent with previous observations of much
238 lower variation in π between species than is expected based on variation in census population size in
239 many domains of life (Lynch 2010), which remains poorly understood, despite decades of attempts at
240 an explanation (Leffler et al. 2012).

241 We propose that panmixia is rare in bacteria because it is unstable and does not persist over
242 evolutionary time. Specifically, high $N_e r$, if sustained, leads to an accumulation of diversity to levels
243 higher than that observed in any species. We propose that an accumulation of diversity leads to the
244 evolution of barriers to genetic exchange within the species that reduce $N_e r$. It follows that when
245 populations with high values of $N_e r$ are found, this should have arisen recently on an evolutionary
246 timescale.

247 As proposed in (Yang et al. 2018), the Asian population of *V. parahaemolyticus*, VppAsia, has spread
248 within the last decades due to human activity but has an ancestral range that spans coastal waters from
249 India to Japan. The other *V. parahaemolyticus* populations that we have sampled, with different
250 ancestral ranges have smaller estimated $N_e r$, perhaps because they have smaller or less fecund ranges
251 or experience greater competition or more frequent demographic bottlenecks.

252 Crucially, it seems that the ancestral population size of *V. parahaemolyticus* as a whole was smaller
253 than currently found in VppAsia and more similar to that in the other populations. First, the site
254 frequency spectrum of VppAsia, but not the other populations, is out of equilibrium and is
255 approximately consistent with a demographic scenario in which the effective population size
256 increased by a factor of 10 15,000 years ago (Figure S3). One possible scenario is that the end of the
257 last ice age around 11,000 years ago created a habitat that existed till the present and that the
258 population expansion implied by the site frequency spectrum is due to a demographic expansion at the
259 end of the ice age. Second, r/μ, the number of bases imported for each mutation event, for the species
260 has been estimated at 313 (Yang et al. 2018) implying that if nucleotide diversity reached an
261 equilibrium, consistent with the current value of $N_e r$ of VppAsia, this would predict a value of $\pi =$

262 $2N_e\mu = 2N_e r\left(\frac{\mu}{r}\right) = 2 \times 453 \times \left(\frac{1}{313}\right) = 2.9$ implying that most synonymous sites should be
263 variable and is far from the current value of 0.024 or indeed diversity levels in organisms in this study
264 or indeed in any well-characterised bacterial species. Our simulation results show that estimates of
265 $N_e r$ respond much more quickly to changes in demography than π. The equivalent argument is more
266 complex in *H. pylori* due to its geographic population structure but levels of variation are clearly far
267 from equilibrium, as illustrated by the substantial variation between populations that reflects
268 bottlenecks associated with migration rather than current population sizes (Linz et al. 2007).

269 Simulations of facultatively sexual organisms have shown that selection on multiple interacting loci
270 can lead to populations being dominated by small numbers of clones, even in the presence of frequent
271 recombination (Neher and Shraiman 2009). These simulations also show that a phase transmission
272 from panmixia can take place due to a decrease in recombination rate or even an increase in
273 population size.

274 There are several mechanisms which might lead to similar phase transition like behaviour in bacterial
275 populations as diversity increases. In (Cui et al. 2018), we observe nascent boundaries to genetic
276 exchange between lineages of *V. parahaemolyticus* associated with differentiation into ecological
277 types. This type of genetic differentiation is likely to be more common and more likely to lead to
278 barriers to exchange in more diverse bacterial populations, which tend to have more diverse accessory
279 genome complement (Andreani, Hesse, and Vos 2017). Such barriers can result in speciation but
280 might also lead to genetic structuring within a single species, such as host specific gene pools found in
281 *C. jejuni* (Sheppard et al. 2014). Another class of mechanism is changes in the pattern or rate of
282 recombination. Recombination requires homology between donor and recipient DNA to be
283 recognized by the cellular machinery and, in addition, most species have a mismatch repair system,
284 which aborts the process if there are too many differences. Therefore, as homology decreases, so will
285 recombination (Falush et al. 2006). Mismatch dependent recombination is likely to be the main reason
286 why *E. coli* has an estimated value of $N_e r$ of only 12, implying that recombination is ineffective in
287 reasserting variation across the whole species (Didelot et al. 2012).

288 Our arguments imply that many environmental organisms with large census population sizes are
289 likely to have gone through a stage of high $N_e r$ values, similar to that currently observed in the Asian
290 population of *V. parahaemolyticus,* before a phase-transition reduced the number of lineages
291 circulating in the population, which also prevented a continuous increase in diversity. Support for this
292 hypothesis could be provided by identifying species that have been through such a transition and
293 characterizing the factors that currently mitigate against recombination between lineages.

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

299

300 **Material and methods**

301 **Genomic datasets used in analysis**

302 The genome sequences of 21 bacterial species were used in this analysis, including *Bacillus*
303 *thuringiensis, Burkholderia pseudomallei, Campylobacter jejuni, Chlamydia trachomatis,*
304 *Enterococcus faecalis, Escherichia coli, Flavobacterium psychrophilum, Haemophilus influenzae,*
305 *Helicobacter pylori, Klebsiella pneumoniae, Legionella pneumophila, Leptospira interrogans,*
306 *Neisseria meningitidis, Porphyromonas gingivalis, Pseudomonas syringae, Salmonella enterica,*
307 *Streptococcus pneumoniae, Streptococcus pyogenes, Vibrio cholerae, Vibrio parahaemolyticus,*
308 *Vibrio vulnificus.* The SNP information of 278 *H. pylori* and 1,103 *V. parahaemolyticus* strains were
309 acquired from our other papers (Berthenet et al. 2018, Yang et al. 2018). Other genome sequences
310 were publicly available and were downloaded from NCBI database. For species that have more than
311 500 genomes in NCBI database, such as *E. coli*, 500 genomes were randomly selected and were used
312 for Neighbour-joining tree reconstruction firstly. Strains located in the extreme long branches and
313 belonged to clonal groups were manually removed, finally resulting in a dataset of 6,355 genomes.

314 The accessions of genomes used (excluding *H. pylori* and *V. parahaemolyticus*) were listed in Table
315 S1, and the whole-genome alignment files were available in the figshare data repository
316 (https://figshare.com/s/3f9d04a8229f30dd785b).

317 **Simulation**

318 The software FastSimBac (De Maio and Wilson 2017) was used to generate the simulated bacterial
319 populations under different hypothetical evolution scenarios, including a constant population with
320 different recombination effects (Figure 1), and the changing populations with population growth
321 (Figure S1a), bottleneck (Figure S1b) and migration (Figure S1c). All the simulated genome length
322 was set to 2 Mb, mutation rate was fixed at 0.01 (per site per 2N generations). The detailed parameters
323 used in simulation were listed in Table S2.

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

325 The SNPs of publicly available genome sequences were called as previously described (Cui et al.
326 2015, Yang et al. 2018). Briefly, the individual genome was aligned against reference genome of the
327 corresponding species (Table S1) using MUMmer (Delcher, Salzberg, and Phillippy 2003), then the
328 detected bi-allelic SNPs in core genome were integrated to a SNP matrix for each species. The
329 Neighbour-joining trees were constructed using software Treebest
330 (http://treesoft.sourceforge.net/treebest.shtml) based on concatenated SNPs. Haploview (Barrett 2009)
331 was used to calculate the r^2 value based on previous and new identified SNP sets, maximum
332 comparison distance was set to 30 kb. Sliding window method was used to summarize the r^2 value
333 with window size = 10 bp.

334 **Recombination scaled effective population size estimation**

335 For a freely mixing population in which recombination drives diversification, theory predicts that $N_e r$
336 (the effective population size N_e times recombination rate r) is in proportion to genealogical
337 coalescence rate, and the expected coalescence curves can be estimated based on the formula $\mathrm{d} =$
338 $d_{\mathrm{unrelated}}(1 - \exp(-4\mathrm{N}_e \mathrm{r}\left(\frac{1}{\mathrm{m}} - \frac{1}{\mathrm{n}}\right))$ defined in our previous work (Cui et al. 2015). In the formula, d

339 is the expected pairwise genetic divergence, $d_{unrelated}$ is the median pairwise SNP distance of a
340 population, n is the number of individual strains and m is number of ancestral node of n strains along
341 coalescent tree. Observed coalescence curves can be estimated with UPGMA algorithm based on the
342 pairwise SNP distance. By fitting the expected and observed coalescence curves with least square
343 method to searching for the optimal parameters, we acquired the best fit of $\hat{n}$ and $N_e r$ value that used
344 in further analysis.

345 **Site frequency spectrum estimation**

346 The minor allele frequency (MAF) of each SNP locus in a SNP matrix was calculated and then the
347 distribution of MAF was counted to generate the site frequency spectrum (SFS) for each dataset. To
348 get a comparable result, the SFS showed in each Figure/panel was calculated based on same sample
349 size. In FigureS3a, the merged population was generated by arbitrarily selecting 100 genomes from
350 Pop1 and Pop2, respectively. In Figure S3b, same number of genomes compared with natural genome
351 dataset were simulated to show populations with different evolution scenarios.

352 **Acknowledgements**

353 This work is supported by the National Key Research & Development Program of China (No.
354 2016YFC1200100, 2017YFC1200800, and 2017YFC1601503), Sanming Project of Medicine in
355 Shenzhen (No. SZSM201811071), the National Natural Science Foundation of China (No. 31770001)
356 and the National Key Program for Infectious Diseases of China (No. 2017ZX10104002). D.F. is
357 funded by a Medical Research Council Fellowship as part of the MRC CLIMB consortium for
358 microbial bioinformatics (grant number MR/M501608/1).

359 **Author Contributions**

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

363 **Competing Financial Interests statement**

364 No

365 **References**

366 Andreani, N. A., E. Hesse, and M. Vos. 2017. "Prokaryote genome fluidity is dependent on effective
367 population size." *ISME J* no. 11 (7):1719-1721. doi: 10.1038/ismej.2017.36.
368 Bar-On, Y. M., R. Phillips, and R. Milo. 2018. "The biomass distribution on Earth." *Proc Natl Acad Sci*
369 *U S A* no. 115 (25):6506-6511. doi: 10.1073/pnas.1711842115.
370 Barrett, J. C. 2009. "Haploview: Visualization and analysis of SNP genotype data." *Cold Spring Harb*
371 *Protoc* no. 2009 (10):pdb ip71. doi: 10.1101/pdb.ip71.
372 Berthenet, E., K. Yahara, K. Thorell, B. Pascoe, G. Meric, J. M. Mikhail, L. Engstrand, H. Enroth, A.
373 Burette, F. Megraud, C. Varon, J. C. Atherton, S. Smith, T. S. Wilkinson, M. D. Hitchings, D.
374 Falush, and S. K. Sheppard. 2018. "A GWAS on Helicobacter pylori strains points to genetic
375 variants associated with gastric cancer risk." *BMC Biol* no. 16 (1):84. doi: 10.1186/s12915-
376 018-0550-3.
377 Cao, Q., X. Didelot, Z. Wu, Z. Li, L. He, Y. Li, M. Ni, Y. You, X. Lin, Z. Li, Y. Gong, M. Zheng, M. Zhang, J.
378 Liu, W. Wang, X. Bo, D. Falush, S. Wang, and J. Zhang. 2015. "Progressive genomic
379 convergence of two Helicobacter pylori strains during mixed infection of a patient with
380 chronic gastritis." *Gut* no. 64 (4):554-61. doi: 10.1136/gutjnl-2014-307345.
381 Cui, Y., X. Yang, X. Didelot, C. Guo, D. Li, Y. Yan, Y. Zhang, Y. Yuan, H. Yang, J. Wang, J. Wang, Y. Song,
382 D. Zhou, D. Falush, and R. Yang. 2015. "Epidemic Clones, Oceanic Gene Pools, and Eco-LD in
383 the Free Living Marine Pathogen Vibrio parahaemolyticus." *Mol Biol Evol* no. 32 (6):1396-
384 410. doi: 10.1093/molbev/msv009.
385 Cui, Yujun, Chao Yang, Hongling Qiu, Hui Wang, Ruifu Yang, and Daniel Falush. 2018. "The landscape
386 of coadaptation in Vibrio parahaemolyticus." *bioRxiv*. doi: 10.1101/373936.
387 De Maio, N., and D. J. Wilson. 2017. "The Bacterial Sequential Markov Coalescent." *Genetics* no. 206
388 (1):333-343. doi: 10.1534/genetics.116.198796.
389 Delcher, A. L., S. L. Salzberg, and A. M. Phillippy. 2003. "Using MUMmer to identify similar regions in
390 large sequence sets." *Curr Protoc Bioinformatics* no. Chapter 10:Unit 10 3. doi:
391 10.1002/0471250953.bi1003s00.
392 Didelot, X., G. Meric, D. Falush, and A. E. Darling. 2012. "Impact of homologous and non-homologous
393 recombination in the genomic evolution of Escherichia coli." *BMC Genomics* no. 13:256. doi:
394 10.1186/1471-2164-13-256.
395 Falush, D., M. Torpdahl, X. Didelot, D. F. Conrad, D. J. Wilson, and M. Achtman. 2006. "Mismatch
396 induced speciation in Salmonella: model and data." *Philos Trans R Soc Lond B Biol Sci* no. 361
397 (1475):2045-53. doi: 10.1098/rstb.2006.1925.
398 Gibson, B., D. J. Wilson, E. Feil, and A. Eyre-Walker. 2018. "The distribution of bacterial doubling
399 times in the wild." *Proc Biol Sci* no. 285 (1880). doi: 10.1098/rspb.2018.0789.
400 Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Segurel, A. Venkat, P. Andolfatto, and M.
401 Przeworski. 2012. "Revisiting an old riddle: what determines genetic diversity levels within
402 species?" *PLoS Biol* no. 10 (9):e1001388. doi: 10.1371/journal.pbio.1001388.
403 Linz, B., F. Balloux, Y. Moodley, A. Manica, H. Liu, P. Roumagnac, D. Falush, C. Stamer, F. Prugnolle, S.
404 W. van der Merwe, Y. Yamaoka, D. Y. Graham, E. Perez-Trallero, T. Wadstrom, S. Suerbaum,
405 and M. Achtman. 2007. "An African origin for the intimate association between humans and
406 Helicobacter pylori." *Nature* no. 445 (7130):915-918. doi: 10.1038/nature05562.
407 Lynch, M. 2010. "Evolution of the mutation rate." *Trends Genet* no. 26 (8):345-52. doi:
408 10.1016/j.tig.2010.05.003.
409 Neher, R. A., and B. I. Shraiman. 2009. "Competition between recombination and epistasis can cause
410 a transition from allele to genotype selection." *Proc Natl Acad Sci U S A* no. 106 (16):6866-
411 71. doi: 10.1073/pnas.0812560106.
412 Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. "Inference of population structure using
413 multilocus genotype data." *Genetics* no. 155 (2):945-59.

Sheppard, S. K., L. Cheng, G. Meric, C. P. de Haan, A. K. Llarena, P. Marttinen, A. Vidal, A. Ridley, F. Clifton-Hadley, T. R. Connor, N. J. Strachan, K. Forbes, F. M. Colles, K. A. Jolley, S. D. Bentley, M. C. Maiden, M. L. Hanninen, J. Parkhill, W. P. Hanage, and J. Corander. 2014. "Cryptic ecology among host generalist Campylobacter jejuni in domestic animals." *Mol Ecol* no. 23 (10):2442-51. doi: 10.1111/mec.12742.

Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt. 1993. "How clonal are bacteria?" *Proc Natl Acad Sci U S A* no. 90 (10):4384-8.

Stoppe, N. C., J. S. Silva, C. Carlos, M. I. Z. Sato, A. M. Saraiva, L. M. M. Ottoboni, and T. T. Torres. 2017. "Worldwide Phylogenetic Group Patterns of Escherichia coli from Commensal Human and Wastewater Treatment Plant Isolates." *Front Microbiol* no. 8:2512. doi: 10.3389/fmicb.2017.02512.

Vos, M., and X. Didelot. 2009. "A comparison of homologous recombination rates in bacteria and archaea." *ISME J* no. 3 (2):199-208. doi: 10.1038/ismej.2008.93.

Wang, J. 2005. "Estimation of effective population sizes from data on genetic markers." *Philos Trans R Soc Lond B Biol Sci* no. 360 (1459):1395-409. doi: 10.1098/rstb.2005.1682.

Wright, S. 1951. "The genetical structure of populations." *Ann Eugen* no. 15 (4):323-54.

Yang, Chao, Xiaoyan Pei, Yarong Wu, Lin Yan, Yanfeng Yan, Yuqin Song, Nicola Coyle, Jaime Martinez-Urtaza, Christopher Quince, Qinghua Hu, Min Jiang, Edward Feil, Dajin Yang, Yajun Song, Dongsheng Zhou, Ruifu Yang, Daniel Falush, and Yujun Cui. 2018. "Recent mixing of Vibrio parahaemolyticus populations." *bioRxiv*. doi: 10.1101/354761.