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

Background 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.

Results 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.

Conclusions 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.

Keywords

Panmixia, bacterial population structure, effective population size, recombination, genetic diversity

Background

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 [2-5] and exhibit high rates of within-species homologous recombination [6]. 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 linkage disequilibrium, indicating structuring of variation [7]. Here we

42 propose a resolution of this paradox, namely that large bacterial populations accumulate diversity
43 progressively until that diversity acts as an effective barrier to genetic exchange between lineages.
44
45 We examine the population structure of 21 bacterial species and find that only the Asian population of
46 *Vibrio parahaemolyticus* is close to genome-wide linkage equilibrium. We find that this population
47 has undergone a recent expansion. If the current population size was maintained over evolutionary
48 timescales, it would lead to a greater than 30 fold increase in diversity, to levels higher than found in
49 any well characterized bacterial species. We propose that other environmental species with large
50 census population sizes, for example *Vibrio cholerae* or *Klebsiella pneumoniae*, may have been
51 through a similar stage before accumulating diversity and that this diversity acted to generate barriers
52 to recombination, either directly or via selective pressure to reduce recombination rates between
53 genetically divergent lineages.
54
55 Specifically, we use pairwise genetic distances between strains calculated from core genome
56 sequences to estimate the recombination scaled effective population size parameter $N_e r$, where N_e is
57 the effective population size and r is the recombination rate per site per generation, an approach
58 briefly introduced by Cui et al [8]. Here, we test the method using simulated genomic datasets and
59 find that it provides accurate estimates in constant size populations for values of the scaled
60 recombination rate R (see below) of 5 or more, and responds to changes in population size or structure
61 far more quickly than estimates of N_e based on nucleotide diversity. We apply the method to real
62 genomic datasets from 21 species and find that some common species, such as *Escherichia coli*, have
63 strikingly low estimates of $N_e r$, implying that recombination is inefficient in distributing diversity
64 across the species, while for others the fit of model to data is poor, highlighting deviations from the
65 assumption of a freely recombining population. Only datasets from *V. parahaemolyticus* and
66 *Helicobacter pylori* give $N_e r$ estimates greater than 100. The former has had a recent increase in
67 population size, whereas the latter experienced repeated bottlenecks associated with geographical
68 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 [9]. 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 outbreeding species. Under neutral theory [10], 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 [11].

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 or non-homologous recombination [12]. 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 [13]. 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 [14].

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 this parameter r is the product of the per initiation site recombination rate ρ and mean tract length δ used in many bacterial recombination models [15, 16]. 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 and therefore genetically distinct. N_e is the number of genetically distinct individuals that contribute in each generation, which is typically in the thousands or millions. $N_e r$ is designed to measure an analogous quantity in bacteria, namely the number of genetically distinct organisms that contribute to the future bacterial gene pool. To this end, time is rescaled; N_e measures the rate of genetic drift per bacterial generation, while $N_e r$ measures the genetic drift in proportion to the time it takes for strains to become distinct from their ancestors by importing DNA by homologous recombination.

The recombination rate of *V. parahaemolyticus* is $r = 1.7 \times 10^{-4}$ per site per year [17]. After T years of evolution, the expected proportion of recombined genome is e^{-rT}, so that it takes around 4,000 years on average for half of the genome to recombine. Half of this time is approximately equivalent to a eukaryotic generation in the sense that two strains that shared a common ancestor 2,000 years ago will be about as related as siblings. This represents a very different time scale from that assumed based on

bacterial generations. For example, *V. parahaemolyticus* is capable of replicating in less than ten minutes in appropriate conditions [18].

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

To use these times to estimate the effective population size, we assume that the genealogy of clonal relationships is generated by a coalescent model with a constant population size N_e [19, 20]. This model generates expectations for the times in the past at which common ancestors of strains in a sample existed. Specifically, for a sample of n strains, there are $n - 1$ coalescent nodes. The age of the most recent node corresponds to the common ancestor of the two most closely related strains in the sample, while the $(n-1)^{\text{th}}$ node corresponds to the common ancestor of all the strains in the sample.

Coalescent theory implies that the expected time in the past at which the m^{th} most ancient coalescent 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 using the formula in the previous paragraph. We use the UPGMA algorithm to obtain $n - 1$ coalescent distances from the $n(n-1)/2$ 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 $n - 1$ coalescent 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 coalescent events plateau at $d_{\text{unrelated}}$. In graphical representations, it is convenient to show the coalescent events in chronological order with the oldest first, to aid comparisons between datasets with different sample sizes.

The model assumes that the strains are randomly sampled from a homogeneous population at a single time point, but pathogenic clones, epidemic outbreaks or strains from specific locations are often overrepresented in real data, leading to oversampling of very closely related isolates, relative to their frequency in the global population [7, 21]. Therefore, in addition to $N_e r$ we estimate a second parameter $\hat{n}$ called the effective sample size, which is an estimate of the number of strains remaining when over-sampled clonally related strains are removed. For simulated data where sampling is random, $\hat{n}$ is 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 [22]. 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 (0.001 to 0.1), so that the scaled recombination rate $R = \frac{\delta\rho}{2N_e}$ varies by two orders of magnitude, from 0.5 to 50. We also fix the rate of initiation of recombination event ρ (0.01) and vary the recombination tract length δ (100 to 10,000). Both methods give similar results, for low scaled recombination rates, the phylogenetic tree is highly

177 structured but becomes progressively bushier as R increases, and gives the impression of being a
178 nearly perfect star for R = 50. For high R, the estimated $N_e r$ is close to R and the observed genetic
179 distances are well-fit by the model. For lower R, the fit is less good and the estimate of R provided by
180 $N_e r$ is also less accurate, although it remains of the right order. For these parameter values, the size
181 and shape of the phylogenetic tree are highly variable between runs, and the poorer fit reflects this
182 stochasticity as well as greater inaccuracy in the approximations made in converting between genetic
183 distances and coalescent times.
184
185 Further simulations of more complex scenarios (Supplementary Fig. 1) show that $N_e r$ estimates
186 reflect the current demography of the population more closely than estimates based on π, which is
187 more influenced by past demography and migration to and from different demes. Supplementary Fig.
188 1a shows the effect of a population expansion on estimates of $N_e r$. At time t, a single population with
189 scaled recombination rate R = 5 splits into two. The blue population maintains the ancestral
190 population size while the red population becomes 10 times larger. At time t + 0.05 the red population
191 is only marginally more diverse than the blue one but its estimated value of $N_e r$ is 6 fold higher. The
192 fit of observed and expected genetic distances is poor for the red population, reflecting the inaccuracy
193 of the modelling assumption that there is a single unchanging population size. At time t + 0.2 the
194 estimate of $N_e r$ is close to its true post-split value for both populations and the model fit is
195 substantially improved. Merging data from the two populations gives intermediate estimates of $N_e r$.
196
197 Supplementary Fig. 1b shows the effect of a population size reduction. A single population with R =
198 50 splits into two, with the blue population remaining unchanged while the red population undergoes
199 a 10 fold reduction in size. As in Supplementary Fig. 1a, the model fit for the red population is poor
200 immediately after the split but quickly improves, with the estimate of $N_e r$ approaching the correct
201 value of 5, while the nucleotide diversity of the population reduces much more slowly. Supplementary
202 Fig. 1c shows the effect of symmetric migration between two populations with a 10 fold difference in
203 effective population size. Migration has a large effect on nucleotide diversity, especially for the

smaller population, but has little effect on estimates of $N_e r$. Overall, these results show that our inference approach provides accurate estimates of the recombination-scaled effective population size for simulated data and that deviations between observed and expected genetic distances can be informative about deviations from model assumptions.

Application to *V. parahaemolyticus* genomes

We first applied the method to the 1,103 *V. parahaemolyticus* genomes described in Yang et al [17]. For this dataset, we obtained estimates of $N_e r$ = 484 and $\hat{n}$ = 471 (Fig. 2a). $\hat{n}$ is less than half of the sample size principally because a large fraction of the isolates in the sample belong to pandemic clonal lineages responsible for large numbers of human infections. Both lineages have most recent common ancestors within the last few decades and are likely to represent a very small fraction of the global population of *V. parahaemolyticus*. The genetic distances fit the model well except that the 33 oldest coalescent events, on the left hand side of the plot are larger than $d_{\text{unrelated}}$. The discrepancy is largely due to population structure within the species, since the fit is better, although still not perfect when analysis is restricted to the 944 isolates from the VppAsia population (Fig. 2a).

As described by Yang et al., the sample of *V. parahaemolyticus* is subdivided into 4 modestly differentiated populations, likely due to historical barriers to migration between oceans [17]. 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 for a dataset consisting only of clinical strains than of shellfish, fish or all non-clinical isolates (Fig. 2c). This difference presumably reflects variation in disease causing potential, which results in samples of clinical isolates not representing the full diversity of clonal lineages within the species.

Application to multiple bacterial species

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

H. pylori has the largest estimated value of $N_e r$ of all datasets we analysed. *H. pylori* is characterized by extremely high rates of recombination during mixed infection of the human stomach, with 10% or more of the genome recombined during a single infection [23] and linkage disequilibrium decreases much more rapidly as a function of genetic distance than for all other species, including *V. parahaemolyticus* (Fig. 3a). Although the tree is approximately star-like, the fit of the model is not perfect, with no clear plateau for a single value of $d_{\text{unrelated}}$, due to the complex geographical population structure of the species [24].

Amongst the other species, $N_e r$ varies from 1 for *Chlamydia trachomatis* to 88 for *Salmonella enterica*. The overall fit of the model varies considerably between species (Supplementary Fig. 2) and is typically worst for the oldest coalescent events, as examined in more detail above for *V. parahaemolyticus*. Our estimated $N_e r$ values are correlated with the linkage disequilibrium statistic r^2 measured between distant makers at pairwise distance of 3 kb (Fig. 3b). However, $N_e r$ shows no correlation with nucleotide diversity of synonymous sites π_{syn} (Fig. 3c) and r^2 at pairwise distance of 3 kb also shows no correlation with π_{syn} (Fig. 3d).

Discussion

Structure of genetic diversity in 21 bacterial species

In bacteria, adaptation to diverse environmental challenges should be most effective in species where realized recombination rates are high enough to thoroughly mix up genetic variation (panmixia), since this creates the largest possible pool of genotypes on which natural selection can act. High recombination can also make it easier for researchers to detect the imprint of natural selection, a feature exploited by Cui et al. 2015, 2018 [8, 25] in investigating coadaptation in *Vibrio parahaemolyticus*. However, despite the utility of well-mixed gene pools, both to the species that have it and to the researchers studying it, it appears to be rare in bacteria for which large numbers of genomes are currently available.

The genetic structure of bacteria species depends on the interplay of many different processes, including changes in population size over time, geographical and ecological subdivision and the complex biology of genetic exchange which takes place via conjugation, transformation and transduction. As a result no single parameter summarizes the effect of recombination in breaking down linkage disequilibrium. Furthermore, available genomes rarely come close to being a random sample of the bacterial population from which they are taken.

Here we have used two summary statistics of the effectiveness of recombination; a non-parametric measure of long-range linkage disequilibrium, r^2 between markers 3 kb apart on the genome, and a parametric approach estimating the composite population genetic parameter $N_e r$. Informally, high values of $N_e r$ implies that recombination has generated new clonal complexes quickly compared to the rate at which genetic drift (proportional to $1/N_e$) removes them, with the result there are many distinct clonal-complexes segregating in the population. In data simulated according to a coalescent model with recombination of short tracts, estimated values of $N_e r$ are close to the true values for $N_e r > 2$.

In our survey of 21 species from which large numbers of genomes are available, our two summary statistics are strongly but incompletely correlated ($R^2 = 0.69$, $P < 0.01$, Fig. 3b). According to both

284 statistics, two species are clear outliers, with estimates of $N_e r = 453$ for the Asian population of *V.*
285 *parahaemolyticus* and 1,976 for the hpEurope population of *H. pylori,* and the estimated value of $N_e r$
286 was lower than 100 in the other 19 species, which are therefore far from being panmictic.
287
288 Estimated $N_e r$ or r^2 are uncorrelated with the nucleotide diversity of synonymous sites π_{syn} (Fig. 3c,
289 Fig. 3d). For example, *V. parahaemolyticus* has similar diversity to *V. cholerae* and much lower
290 diversity than *V. vulnificus*, both of which have estimated $N_e r$ lower than 50. Furthermore, estimates
291 of $N_e r$ vary over a factor of 1,000, while π varies only by a factor of 10.
292
293 **Mechanisms by which high diversity can create barriers to recombination**
294 Since panmixia is possible within bacterial populations and should facilitate genetic adaptation to the
295 widest possible range of niches available to the species, it raises the question why it is not widespread.
296 There are several bacterial species in our sample which survive well in the environment, have
297 effective global dispersal, enormous census population sizes and high recombination rates, such as *E.*
298 *coli, Campylobacter jejuni* and *K. pneumoniae*. Species with these characteristics are the most
299 obvious candidates to be panmictic, whereas *V. parahaemolyticus* thrives only in warm brackish
300 waters and has oceanic gene pools, implying historical limits on its dispersal [17], and *H. pylori* only
301 survives in human stomachs and shows geographic differentiation associated with historical
302 migrations of its host [26].
303
304 Neutral population genetic models imply that, all else being equal, bacteria with higher census
305 population sizes should have higher $N_e r$ as well as higher nucleotide diversity, which is proportional
306 to $2N_e\mu$ at equilibrium. However, across our 21 species, there is no correlation between the two
307 statistics. We propose that the absence of a correlation occurs because high diversity tends to
308 supresses recombination. There are a number of mechanisms by which this suppression can occur that
309 have been described in the literature and we do not attempt to reach a conclusion about which are in
310 fact most important in nature.

For example, in Cui et al [25], we observe nascent boundaries to genetic exchange between lineages of *V. parahaemolyticus* associated with differentiation into ecological types. This kind of genetic differentiation is more likely to be common, and more likely to lead to barriers to exchange in more diverse bacterial populations, which tend to have more diverse accessory genomes [27], and might also have a larger number of distinct ecological niches. Such barriers can result in speciation [21, 28] but can also lead to genetic structuring within a single species, such as host-specific gene pools found in *C. jejuni* [29].

Another mechanism is changes in the pattern or rate of recombination. Recombination requires homology between donor and recipient DNA to be recognized by the cellular machinery and, in addition, most species have a mismatch repair system, which aborts the process if there are too many differences. Therefore, as homology decreases, so will recombination [23, 28, 30, 31]. Mismatch dependent recombination is likely to be the main reason why *E. coli* has an estimated value of $N_e r$ of only 12, implying that recombination is ineffective in reassorting variation across the whole species [32].

Finally, a genetic dependence of recombination rates on diversity might arise from effects of epistasis in constraining realized recombination. Simulations of facultatively sexual organisms have shown that selection on multiple interacting loci can lead to populations being dominated by small numbers of clones, even in the presence of frequent recombination [33]. These simulations also show that a phase transition from panmixia can take place due to a decrease in recombination rate or even an increase in population size.

Vibrio parahaemolyticus and _Helicobacter pylori_ are far from demographic equilibrium

If, as we propose, suppression of recombination due to high diversity is a general phenomenon in bacteria, then the existence of populations with high $N_e r$ might seem paradoxical, because high N_e leads to high diversity which should suppress r. However this argument only holds in populations

where N_e has been high for long enough for diversity to accumulate and do not apply if high N_e is a recent phenomenon on an evolutionary timescale.

In fact, neither the Asian population of *V. parahaemolyticus* nor *H. pylori* are close to demographic equilibrium. As proposed in Yang et al [17], the Asian population of *V. parahaemolyticus*, VppAsia, has spread within the last few decades due to human activity but has an ancestral range restricted to coastal waters from India to Japan. The other *V. parahaemolyticus* populations that we have sampled, with different ancestral ranges, have smaller estimated $N_e r$, perhaps because they have smaller or less fecund ranges or experience greater competition or more frequent demographic bottlenecks.

Crucially, it seems that the ancestral population size of *V. parahaemolyticus* as a whole was smaller than currently found in VppAsia and more similar to that in the other populations. First, the site frequency spectrum of VppAsia, but not the other populations, is out of equilibrium and is approximately consistent with a demographic scenario in which the effective population size increased by a factor of ten 15,000 years ago (Supplementary Fig. 3). One possible scenario is that the end of the last ice age around 11,000 years ago created a habitat that existed till the present and that the population expansion implied by the site frequency spectrum is due to a demographic expansion at the end of the ice age.

Secondly the level of synonymous nucleotide diversity in the population is 30-fold lower than expected at demographic equilibrium. r/μ, the number of recombinant sites for each mutant site, has been estimated for the species to be around 313 [17]. Note that this parameter r/μ is not the same as the parameter r/m often used in the literature [6] because the former considers all recombined sites whereas the latter considers only recombinant sites that are substituted. If nucleotide diversity reached an equilibrium, consistent with the current value of $N_e r$ of VppAsia, this would predict a value of $2N_e\mu = 2N_e r \left(\frac{\mu}{r}\right) = 2 \times 453 \times \left(\frac{1}{313}\right) = 2.9$. At equilibrium in a neutrally evolving population, with sites evolving according to the Jukes-Cantor model, π is predicted to be equal to $\frac{3}{4} \times$

366 $\left(1 - e^{-\frac{8N_e\mu}{3}}\right) = 0.73$, and therefore most synonymous sites should differ between individuals. This is
367 far from the current value of 0.024 or indeed diversity levels in any well-characterised bacterial
368 species. Our simulation results show that estimates of $N_e r$ respond much more quickly to changes in
369 demography than π, implying that diversity levels would be likely to continue a slow but steady rise if
370 the effective population size remains constant (ignoring the changes in population structure caused by
371 human activity in recent decades).
372
373 The equivalent argument is more complex in *H. pylori* due to its geographic population structure but
374 levels of variation are clearly far from demographic equilibrium, as illustrated by the substantial
375 variation between populations that reflects bottlenecks associated with historical human migration
376 rather than current population sizes [26]. These results therefore suggest that if effective population
377 sizes remain continuously high in either species, then diversity would increase. We propose that in
378 this case, barriers to recombination would evolve, perhaps by one of the mechanisms described above.
379
380 Despite decades of study by population geneticists, the factors determining the amount of diversity
381 within species are still poorly understood, for example with much lower variation in nucleotide
382 variation π between species than is expected based on variation in census population size in many
383 domains of life [34, 35]. Our results suggest that patterns of genome variation in bacteria are likely to
384 be dynamic, with ecological and genomic differentiation, barriers to gene flow and demographic
385 factors interacting with each other in complex ways to alter both the level of genetic diversity within
386 the species and the way it is partitioned, to produce the wide variety of bacterial population structures
387 that are observed.
388
389 **Material and methods**
390 **Genomic datasets used in analysis**
391 Because our inference approach of $N_e r$ hypothesizes that recombination drives the genetic variation
392 other than mutation, we firstly selected 27 bacterial species, in which r/m values (ratio of nucleotide

393 changes resulted from recombination relative to point mutation) of them were greater than one based
394 on previous multi-locus sequence typing data estimation [6]. We then counted the number of
395 assembled genomes of them in NCBI database, and found that the genome numbers of 21 bacterial
396 species were greater than 100, which were used in further analysis, including *Bacillus thuringiensis,*
397 *Burkholderia pseudomallei, Campylobacter jejuni, Chlamydia trachomatis, Enterococcus faecalis,*
398 *Escherichia coli, Flavobacterium psychrophilum, Haemophilus influenzae, Helicobacter pylori,*
399 *Klebsiella pneumoniae, Legionella pneumophila, Leptospira interrogans, Neisseria meningitidis,*
400 *Porphyromonas gingivalis, Pseudomonas syringae, Salmonella enterica, Streptococcus pneumoniae,*
401 *Streptococcus pyogenes, Vibrio cholerae, Vibrio parahaemolyticus* and *Vibrio vulnificus.*
402
403 For species with more than 500 genomes in NCBI database, such as *E. coli* (>14,000 assembled
404 genomes), 500 genomes were randomly selected to reduce the amount of calculation. Each genome
405 was aligned against the reference genome of the corresponding species (Supplementary Table 1) using
406 MUMmer [36] to generate the whole genome alignments and identify SNPs in core genomes (regions
407 presented in all isolates) as previously described [8, 17]. Only genomes with genome-wide coverage >
408 70% (compared to reference genome) were used in further analysis. SNPs located in repetitive regions
409 were removed, and the filtered bi-allelic SNPs sets were used to construct the Neighbour-joining (NJ)
410 tree of each species. Strains located on the extremely long branches of the NJ tree and strains
411 belonged to clonal groups were manually removed, finally resulting in a dataset of 6,355 genomes
412 (56-1,103 genomes for each species). The accession numbers of genomes used (excluding *H. pylori*
413 and *V. parahaemolyticus*) were listed in Supplementary Table 1, and the whole-genome alignments
414 were available in the figshare data repository (https://figshare.com/s/3f9d04a8229f30dd785b).
415
416 **SNP calling, phylogeny reconstruction and LD decay calculation**
417 The SNP dataset of 278 *H. pylori* (hpEurope population) and 1,103 *V. parahaemolyticus* genomes
418 were reused from previous studies [17, 37]. The SNPs of the left 4,974 genomes were recalled using
419 same pipelines as described above. Totally 17,875-462,214 bi-allelic SNPs were identified separately
420 for 21 bacterial species, which were used in further analysis, including Neighbour-joining tree

421 construction, pairwise SNP distance calculation, LD r^2 value calculation, and $N_e r$ estimation as
422 previously described [8]. The Neighbour-joining trees were constructed using the software Treebest
423 (http://treesoft.sourceforge.net/treebest.shtml) based on concatenated SNPs. Haploview [38] was used
424 to calculate the LD r^2, and the maximum comparison distance was set to 30 kb.
425
426 **Simulation**
427 The software FastSimBac [22] was used to generate the simulated bacterial populations under
428 different hypothetical evolution scenarios, including a constant population with different scaled
429 recombination rate *R* (Fig. 1), and the changing populations with population expansion
430 (Supplementary Fig. 1a), reduction (Supplementary Fig. 1b) and migration (Supplementary Fig. 1c).
431 All the simulated genome length was set to 2 Mb, mutation rate was fixed at 0.01 (per site per $2N_e$
432 generations). The detailed parameters used in simulation were listed in Supplementary Table 2.
433
434 **Recombination scaled effective population size estimation**
435 For a freely mixing population in which recombination drives diversification, neutral theory predicts
436 that $N_e r$ is in proportion to genealogical coalescent rate, and the expected coalescent curves can be
437 estimated based on the formula $d = d_{\text{unrelated}}(1 - \exp(-4N_e r\left(\frac{1}{m} - \frac{1}{n}\right))$ as mentioned in previous
438 work [8]. In the formula, d is the expected pairwise genetic divergence, $d_{\text{unrelated}}$ is the median
439 pairwise SNP distance of a population, n is the number of individual strains and m is the index of the
440 ancestral node along the coalescent tree of n strains. Coalescent curves can be estimated using the
441 UPGMA algorithm based on the pairwise SNP distance. By fitting the expected and observed
442 coalescent curves with least square method to search for the optimal parameters, we found the optimal
443 values of $\hat{n}$ (effective sample size) and $N_e r$ that were used in further analysis. $\hat{n}$ was an estimate of
444 the number of strains remaining when over-sampled clonally related strains are removed.
445
446 **Site frequency spectrum estimation**

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

Acknowledgements

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

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 YM, Phillips R, Milo R: **The biomass distribution on Earth.** *Proc Natl Acad Sci U S A* 2018, **115:**6506-6511.

2. Hendriksen RS, Vieira AR, Karlsmose S, Lo Fo Wong DM, Jensen AB, Wegener HC, Aarestrup FM: **Global monitoring of Salmonella serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007.** *Foodborne pathogens and disease* 2011, **8:**887-900.

3. Pike BL, Guerry P, Poly F: **Global distribution of Campylobacter jejuni Penner serotypes: a systematic review.** *PloS one* 2013, **8:**e67375.

4. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR, Hsu LY, Severin J: **Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health.** *Proceedings of the National Academy of Sciences* 2015, **112:**E3574-E3581.

5. Stoppe NC, Silva JS, Carlos C, Sato MIZ, Saraiva AM, Ottoboni LMM, Torres TT: **Worldwide Phylogenetic Group Patterns of Escherichia coli from Commensal Human and Wastewater Treatment Plant Isolates.** *Front Microbiol* 2017, **8:**2512.

6. Vos M, Didelot X: **A comparison of homologous recombination rates in bacteria and archaea.** *ISME J* 2009, **3:**199-208.

7. Smith JM, Smith NH, O'Rourke M, Spratt BG: **How clonal are bacteria?** *Proc Natl Acad Sci U S A* 1993, **90:**4384-4388.

8. Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, Zhang Y, Yuan Y, Yang H, Wang J, et al: **Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus.** *Mol Biol Evol* 2015, **32:**1396-1410.

9. Wright S: **The genetical structure of populations.** *Ann Eugen* 1951, **15:**323-354.

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

11. Wang J: **Estimation of effective population sizes from data on genetic markers.** *Philos Trans R Soc Lond B Biol Sci* 2005, **360:**1395-1409.

12. Didelot X, Maiden MC: **Impact of recombination on bacterial evolution.** *Trends Microbiol* 2010, **18:**315-322.

13. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155:**945-959.

14. Fraser C, Hanage WP, Spratt BG: **Recombination and the nature of bacterial speciation.** *Science* 2007, **315:**476-480.

15. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data.** *Genetics* 2007, **175:**1251-1266.

16. Didelot X, Lawson D, Darling A, Falush D: **Inference of homologous recombination in bacteria using whole-genome sequences.** *Genetics* 2010, **186:**1435-1449.

17. Yang C, Pei X, Wu Y, Yan L, Yan Y, Song Y, Coyle N, Martinez-Urtaza J, Quince C, Hu Q, et al: **Recent mixing of Vibrio parahaemolyticus populations.** *bioRxiv* 2018.

18. Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M, Nakano M, Yamashita A, et al: **Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of V cholerae.** *Lancet* 2003, **361:**743-749.

19. Kingman JFC: **The coalescent.** *Stochastic processes and their applications* 1982, **13:**235-248.

20. Rosenberg NA, Nordborg M: **Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.** *Nat Rev Genet* 2002, **3:**380-390.

21. Fraser C, Hanage WP, Spratt BG: **Neutral microepidemic evolution of bacterial pathogens.** *Proc Natl Acad Sci U S A* 2005, **102:**1968-1973.

22. De Maio N, Wilson DJ: **The Bacterial Sequential Markov Coalescent.** *Genetics* 2017, **206:**333-343.

23. Cao Q, Didelot X, Wu Z, Li Z, He L, Li Y, Ni M, You Y, Lin X, Li Z, et al: **Progressive genomic convergence of two Helicobacter pylori strains during mixed infection of a patient with chronic gastritis.** *Gut* 2015, **64:**554-561.

24. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, et al: **Traces of human migrations in Helicobacter pylori populations.** *Science* 2003, **299:**1582-1585.

25. Cui Y, Yang C, Qiu H, Wang H, Yang R, Falush D: **The landscape of coadaptation in Vibrio parahaemolyticus.** *bioRxiv* 2018.

26. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, et al: **An African origin for the intimate association between humans and Helicobacter pylori.** *Nature* 2007, **445:**915-918.

27. Andreani NA, Hesse E, Vos M: **Prokaryote genome fluidity is dependent on effective population size.** *ISME J* 2017, **11:**1719-1721.

28. Falush D, Torpdahl M, Didelot X, Conrad DF, Wilson DJ, Achtman M: **Mismatch induced speciation in Salmonella: model and data.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**2045-2053.

29. Sheppard SK, Cheng L, Meric G, de Haan CP, Llarena AK, Marttinen P, Vidal A, Ridley A, Clifton-Hadley F, Connor TR, et al: **Cryptic ecology among host generalist Campylobacter jejuni in domestic animals.** *Mol Ecol* 2014, **23:**2442-2451.

30. Retchless AC, Lawrence JG: **Temporal fragmentation of speciation in bacteria.** *Science* 2007, **317:**1093-1096.

31. Stephan W, Langley CH: **Evolutionary consequences of DNA mismatch inhibited repair opportunity.** *Genetics* 1992, **132:**567-574.

32. Didelot X, Meric G, Falush D, Darling AE: **Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli.** *BMC Genomics* 2012, **13:**256.

33. Neher RA, Shraiman BI: **Competition between recombination and epistasis can cause a transition from allele to genotype selection.** *Proc Natl Acad Sci U S A* 2009, **106:**6866-6871.

34. Lynch M: **Evolution of the mutation rate.** *Trends Genet* 2010, **26:**345-352.

35. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M: **Revisiting an old riddle: what determines genetic diversity levels within species?** *PLoS Biol* 2012, **10:**e1001388.

36. Delcher AL, Salzberg SL, Phillippy AM: **Using MUMmer to identify similar regions in large sequence sets.** *Curr Protoc Bioinformatics* 2003, **Chapter 10:**Unit 10 13.

37. Berthenet E, Yahara K, Thorell K, Pascoe B, Meric G, Mikhail JM, Engstrand L, Enroth H, Burette A, Megraud F, et al: **A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk.** *BMC Biol* 2018, **16:**84.

38. Barrett JC: **Haploview: Visualization and analysis of SNP genotype data.** *Cold Spring Harb Protoc* 2009, **2009:**pdb ip71.

Figure legends

Figure 1. Recombination scaled effective population size ($N_e r$) estimation of simulated constant populations under different scaled recombination rate. The panels indicated populations with scaled recombination rate *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 between individuals and observed and expected coalescence curves. The dashed line of middle and bottom panels indicated the median SNP distance between individuals. The red points in the bottom panel indicated the expected distances between $n-1$ coalescent nodes, the blue triangles indicated the observed distances estimated from pairwise SNP distances using the UPGMA algorithm.

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

576 to create the boxplot. (c) $N_e r$ estimation of VppAsia population based on different types of samples.
577 Points and lines show observed and expected coalescence curves, respectively.
578
579 **Figure 3. Correlation between $N_e r$, linkage disequilibrium (LD) statistic r^2 and nucleotide**
580 **diversity of synonymous sites.** (a) LD decay of 21 bacterial species. The maximum comparison
581 distance was set to 30 kb. The vertical dashed line indicated the LD r^2 values at pairwise distance of 3
582 kb, which were used in panel b and d. Line colours indicated the estimated $N_e r$ values. (b).
583 Correlation between $N_e r$ and LD r^2 values. (c) Correlation between $N_e r$ and nucleotide diversity of
584 synonymous sites π_{syn}. (d) Correlation between LD r^2 values and nucleotide diversity of synonymous
585 sites π_{syn}. Point colours in panel b-d indicated the estimated $N_e r$ values.