The ISME Journal (2009) 3, 199–208

www.nature.com/ismej

ORIGINAL ARTICLE

A comparison of homologous recombination rates in bacteria and archaea

Michiel Vos[1] and Xavier Didelot[2]
[1]*Department of Zoology, University of Oxford, Oxford, UK and* [2]*Department of Statistics, University of Warwick, Coventry, UK*

It is a standard practice to test for the signature of homologous recombination in studies examining the genetic diversity of bacterial populations. Although it has emerged that homologous recombination rates can vary widely between species, comparing the results from different studies is made difficult by the diversity of estimation methods used. Here, Multi Locus Sequence Typing (MLST) datasets from a wide variety of bacteria and archaea are analyzed using the ClonalFrame method. This enables a direct comparison between species and allows for a first exploration of the question whether phylogeny or ecology is the primary determinant of homologous recombination rate.
The ISME Journal (2009) **3,** 199–208; doi:10.1038/ismej.2008.93; published online 2 October 2008
Subject Category: evolutionary genetics
Keywords: archaea; bacteria; ClonalFrame; homologous recombination; MLST

Introduction

The development of the Multi Locus Sequence Typing (MLST) method to genotype pathogenic bacteria (Maiden *et al.*, 1998) has not only benefited molecular epidemiology, but has also greatly improved our understanding of bacterial evolution (Feil *et al.*, 2001; Feil, 2004; Maiden, 2006). MLST consists of sequencing fragments of multiple housekeeping genes (genes encoding proteins essential for cell metabolism) spread around the chromosome. A clear advantage of this method over gel-based methods such as RFLP is that data are unambiguous and can be easily accessed and compared through online databases. Also, in contrast to many other genotyping methods, the resulting sequence data are readily amenable to population genetic analyses (Maiden *et al.*, 1998).

Tests for recombination are routinely performed in MLST-based studies and it has become clear that homologous recombination rates (HRR) vary widely between different species (for example, (Maynard Smith *et al.*, 1993; Feil *et al.*, 2001; Hanage *et al.*, 2006; Narra and Ochman, 2006; Perez-Losada *et al.*, 2006)). The underlying causes of this variation, however, are rarely addressed and not well understood. Calculating a measure of recombination rate, rather than simply detecting a significant presence or absence of homologous recombination events, enables an explicit comparison between species. This allows the variation in HRR to be reviewed in the light of phylogeny and ecology. Similar HRR among species having comparable ecologies but belonging to divergent taxonomic groups could indicate that recombination rates have evolved because of adaptive evolution. On the other hand, different HRR among species having comparable ecologies but belonging to divergent taxonomic groups could imply that recombination rates are evolutionarily constrained.

Why bacteria engage in homologous recombination is the subject of intense debate (Redfield, 2001; Narra and Ochman, 2006; Michod *et al.*, 2008). Three main hypotheses have been brought forward to explain the evolutionary benefits of homologous recombination. The DNA repair hypothesis states that foreign DNA serves as a template to repair double-stranded breaks (Bernstein *et al.*, 1981). According to the food hypothesis, incorporation of foreign DNA in the genome is a by-product of the uptake of DNA for metabolism (Redfield, 1993, 2001). Finally, the various hypotheses for the maintenance of sex in eukaryotes, that is, the removal of deleterious mutations and the combination of beneficial mutations, could be equally applied to bacteria (Narra and Ochman, 2006). Elevated HRR in certain groups thus could indicate increased need for DNA repair, increased importance

Correspondence: M Vos, Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, UK.
E-mail: michiel.vos@zoo.ox.ac.uk
Received 16 June 2008; revised 1 September 2008; accepted 2 September 2008; published online 2 October 2008

of DNA for metabolism or a role for recombination to increase the efficacy of natural selection.

Many approaches are available to identify homologous recombination events and rates from sequence data. However, different methods vary in their ability to detect recombination (Posada, 2002; Stumpf and McVean, 2003; Didelot and Falush, 2007), making comparisons of datasets from the literature difficult. Here, we reanalyzed MLST data from a wide variety of species using the coalescent-based method implemented in the computer package ClonalFrame (Didelot and Falush, 2007). This method estimates the relative probabilities that a nucleotide is changed as the result of recombination relative to point mutation (r/m), which is a direct measure of the relative impact of recombination on sequence diversification (Guttman and Dykhuizen, 1994).

Materials and methods

Measuring the impact of homologous recombination using ClonalFrame

A commonly used evolutionary-based measure for the prominence of recombination in bacteria is the ratio of the rates of occurrence of recombination and mutation, ρ/θ (Milkman and Bridges, 1990). A wide spectrum of methods exist to estimate this ratio, using either microevolutionary techniques (Falush *et al.*, 2001; Feil *et al.*, 2004) or population genetics methodology (Fearnhead *et al.*, 2005; Fraser *et al.*, 2005; Jolley *et al.*, 2005). The ratio ρ/θ is a measure of the frequency at which recombination occurs relative to mutation and therefore has an intuitive interpretation: if for example $\rho/\theta = 2$, recombination events occur two times as often as point mutation in the evolution of the population. However, since it ignores length and nucleotide diversity of imported fragments, it contains no information on the actual impact recombination has on evolutionary change.

To measure the relative effect of homologous recombination on the genetic diversification of populations, we decided to use the ratio r/m, or the ratio of rates at which nucleotides become substituted as a result of recombination and mutation (Guttman and Dykhuizen, 1994). For example, if $r/m = 10$, then recombination introduces 10 times more nucleotide substitutions than do point mutations during the evolution of the population. This is compatible with a value of $\rho/\theta = 2$ if each recombination event introduces five substitutions on average. r/m can be estimated using eBURST (Feil *et al.*, 2004; Spratt *et al.*, 2004), but this method has the disadvantage to be based only on the differences between close relatives within clonal complexes, and could therefore produce inflated results if the role of recombination has increased in recent time. Here, we calculated the values of r/m using ClonalFrame (Didelot and Falush, 2007) (freely available from http://bacteria.stats.ox.ac.uk).

ClonalFrame attempts to reconstruct the clonal genealogy of a sample of strains, as well as the mutation and recombination events that took place on the branches of this genealogy, based on a coalescent model. The coalescent is a population genetics model that tracks the ancestry of present day individuals back in time to their last common ancestor (Kingman, 1982). It approximates the expected genealogy of a sample of individuals within a large population evolving under the Wright–Fisher model (Fisher, 1930; Wright, 1931). Mutation and recombination are assumed to occur at constant rates $\theta/2$ and $\rho/2$ on the branches of the coalescent tree. When a mutation happens, it affects any nucleotide in the gene fragment with uniform probability and according to the Jukes–Cantor model of substitution (Jukes and Cantor, 1969). When a recombination event happens, it affects a stretch of DNA within which every nucleotide has an equal probability to be substituted (Didelot and Falush, 2007). By not attempting to reconstruct the origin of each recombination event within the population, ClonalFrame provides an accurate and efficient approximation of the computationally demanding coalescent with the recombination model (Hudson, 1983). ClonalFrame is capable of estimating a number of evolutionary parameters, including r/m. As it uses Bayesian statistics, a credibility interval can be computed for each parameter, which is a direct reflection of our uncertainty to infer the parameter based on the data.

All datasets analyzed in this study are listed in Table 1 in order of inferred mean r/m value. A brief description of each dataset is given in the Supplementary Information. In the main text, r/m values are referred to as low (<1), intermediate (1–2), high (2–10) or very high (>10). These boundaries are arbitrary but facilitate discussion and roughly correspond to interpretations of recombination rates in the literature. The values in Table 1 should be interpreted only as a general indication of HRR in a species. Results will vary when a different sample of strains is used. Loci vary in their recombination rate (Mau *et al.*, 2006), and so the choice of MLST loci will influence results. Some estimates will be imprecise because of suboptimal sampling from the natural population (see below). The more Sequence Types (unique combinations of MLST alleles) could be used in each analysis, the more statistical power was available to infer the genealogy and other parameters, resulting in tighter estimates of r/m. Finally, it has to be stressed that different populations belonging to the same species might have different HRR.

ClonalFrame settings

All values of r/m were computed with the scaled mutational rate θ set equal to Watterson's moment estimator (Watterson, 1978). For each dataset, two runs of the ClonalFrame MCMC were performed,

Table 1 The ratio of nucleotide changes as the result of recombination relative to point mutation (*r*/*m*) for different bacteria and archaea estimated from MLST data using ClonalFrame

Species	*Phylum/division*	*Ecology*	*n STs*	*n loci*	*r/m*	*95% CI*	*Reference*
Flavobacterium psychrophilum	Bacteroidetes	Obligate pathogen	33	7	63.6	32.8–82.8	Nicolas *et al.* (2008)
Pelagibacter ubique (SAR 11)	α-proteobacteria	Free-living, marine	9	8	63.1	47.6–81.8	Vergin *et al.* (2007)
Vibrio parahaemolyticus	γ-proteobacteria	Free-living, marine (OP)	20	7	39.8	27.4–48.2	Gonzalez-Escalona *et al.* (2008)
Salmonella enterica	γ-proteobacteria	Commensal (OP)	50	7	30.2	21.0–36.5	web.mpiib-berlin.mpg.de/mlst
Vibrio vulnificus	γ-proteobacteria	Free-living, marine (OP)	41	5	26.7	19.4–33.3	Bisharat *et al.* (2007)
Streptococcus pneumoniae	Firmicutes	Commensal (OP)	52	6	23.1	16.7–29.0	Hanage *et al.* (2005)
Microcystis aeruginosa	Cyanobacteria	Free-living, aquatic	79	7	18.3	13.7–21.2	Tanabe *et al.* (2007)
Streptococcus pyogenes	Firmicutes	Commensal (OP)	50	7	17.2	6.8–24.4	Enright *et al.* (2001)
Helicobacter pylori	ε-proteobacteria	Commensal (OP)	117	8	13.6	12.2–15.5	pubmlst.org
Moraxella catarrhalis	γ-proteobacteria	Commensal (OP)	50	8	10.1	4.5–18.6	web.mpiib-berlin.mpg.de/mlst
Neisseria meningitidis	β-proteobacteria	Commensal (OP)	83	7	7.1	5.1–9.5	Jolley *et al.* (2005)
Plesiomonas shigelloides	γ-proteobacteria	Free-living, aquatic	58	5	7.1	3.8–13.0	Salerno *et al.* (2007)
Neisseria lactamica	β-proteobacteria	Commensal	180	7	6.2	4.9–7.4	pubmlst.net
Myxococcus xanthus	δ-proteobacteria	Free-living, terrestrial	57	5	5.5	1.9–11.3	Vos and Velicer (2008)
Haemophilus influenzae	γ-proteobacteria	Commensal (OP)	50	7	3.7	2.6–5.4	Meats *et al.* (2003)
Wolbachia b complex	α-proteobacteria	Endosymbiont	16	5	3.5	1.8–6.3	Baldo *et al.* (2006)
Campylobacter insulaenigrae	ε-proteobacteria	Commensal (OP)	59	7	3.2	1.9–5.0	Stoddard *et al.* (2007)
Mycoplasma hyopneumoniae	Firmicutes	Commensal (OP)	33	7	3.0	1.1–5.8	Mayor *et al.* (2007)
Haemophilus parasuis	γ-proteobacteria	Commensal (OP)	79	7	2.7	2.1–3.6	Olvera *et al.* (2006)
Campylobacter jejuni	ε-proteobacteria	Commensal (OP)	110	7	2.2	1.7–2.8	pubmlst.org
Halorubrum sp.	Halobacteria (Archaea)	Halophile	28	4	2.1	1.2–3.3	Papke *et al.* (2004)
Pseudomonas viridiflava	γ-proteobacteria	Free-living, plant pathogen	92	3	2.0	1.2–2.9	Goss *et al.* (2005)
Bacillus weihenstephanensis	Firmicutes	Free-living, terrestrial	36	6	2.0	1.3–2.8	Sorokin *et al.* (2006)
Pseudomonas syringae	γ-proteobacteria	Free-living, plant pathogen	95	4	1.5	1.1–2.0	Sarkar and Guttman (2004)
Sulfolobus islandicus	Thermoprotei (Archaea)	Thermoacidophile	17	5	1.2	0.1–4.5	Whitaker *et al.* (2005)
Ralstonia solanacearum	β-proteobacteria	Plant pathogen	58	7	1.1	0.7–1.6	Castillo and Greenberg (2007)
Enterococcus faecium	Firmicutes	Commensal (OP)	15	7	1.1	0.3–2.5	Homan *et al.* (2002)
Mastigocladus laminosus	Cyanobacteria	Thermophile	34	4	0.9	0.5–1.5	Miller *et al.* (2007)
Legionella pneumophila	γ-proteobacteria	Protozoa pathogen	30	2	0.9	0.2–1.9	Coscolla and Gonzalez-Candelas (2007)
Microcoleus chthonoplastes	Cyanobacteria	Free-living, marine	22	2	0.8	0.2–1.9	Lodders *et al.* (2005)
Bacillus thuringiensis	Firmicutes	Insect pathogen	22	6	0.8	0.4–1.3	Sorokin *et al.* (2006)
Bacillus cereus	Firmicutes	Free-living, terrestrial (OP)	13	6	0.7	0.2–1.6	Sorokin *et al.* (2006)
Oenococcus oeni	Firmicutes	Free-living, terrestrial	17	5	0.7	0.2–1.7	de Las Rivas *et al.* (2004)
Escherichia coli ET-1 group	γ-proteobacteria	Commensal (free-living?)	44	7	0.7	0.03–2.0	Walk *et al.* (2007)
Listeria monocytogenes	Firmicutes	Free-living, terrestrial (OP)	34	7	0.7	0.4–1.1	Salcedo *et al.* (2003)
Enterococcus faecalis	Firmicutes	Commensal (OP)	37	7	0.6	0.0–3.2	Ruiz-Garbajosa *et al.* (2006)
Porphyromonas gingivalis	Bacteroidetes	Obligate pathogen	99	7	0.4	0.0–3.4	Enersen *et al.* (2006)
Yersinia pseudotuberculosis	γ-proteobacteria	Obligate pathogen	43	7	0.3	0.0–1.1	web.mpiib-berlin.mpg.de/mlst
Chlamydia trachomatis	Chlamydiae	Obligate pathogen	14	7	0.3	0.0–1.8	Pannekoek *et al.* (2008)
Klebsiella pneumoniae	γ-proteobacteria	Free-living, terrestrial (OP)	45	7	0.3	0.0–2.1	Diancourt *et al.* (2005)
Bordetella pertussis	β-proteobacteria	Obligate pathogen	32	7	0.2	0.0–0.7	Diavatopoulos *et al.* (2005)
Brachyspira sp.	Spirochaetes	Commensal (OP)	36	7	0.2	0.1–0.4	Rasback *et al.* (2007)
Clostridium difficile	Firmicutes	Commensal (OP)	34	6	0.2	0.0–0.5	Lemee *et al.* (2004)
Bartonella henselae	α-proteobacteria	Obligate pathogen	14	7	0.1	0.0–0.7	Arvand *et al.* (2007)
Lactobacillus casei	Firmicutes	Commensal	32	7	0.1	0.0–0.5	Diancourt *et al.* (2007)
Staphylococcus aureus	Firmicutes	Commensal (OP)	53	7	0.1	0.0–0.6	Enright *et al.* (2000)
Rhizobium gallicum	α-proteobacteria	Free-living, terrestrial	33	3	0.1	0.0–0.3	Silva *et al.* (2005)
Leptospira interrogans	Spirochaetes	Commensal (OP)	61	7	0.02	0.0–0.1	Thaipadungpanit *et al.* (2007)

Abbreviations: STs, sequence types; CI, credibility interval; OP, opportunistic pathogen.
See Supplementary Information for details on the datasets.

each consisting of 200 000 iterations. The first half of the chains was discarded, and the second half was sampled every hundred iterations. The Gelman–Rubin statistic (Gelman and Rubin, 1992) was then computed for *r/m* in each dataset to assess convergence and mixing properties of the MCMC. For the datasets in which we found a Gelman–Rubin statistic above 1.1, longer runs were performed consisting of 2 000 000 iterations. We then recomputed the Gelman–Rubin statistics and found all of them to be satisfactory (that is, below 1.1). Graphical comparisons of the traces of the likelihood and model parameters demonstrated that the runs were properly converged and mixed. For each dataset, the results of the two ClonalFrame runs were then concatenated, and the reported values of the mean and 95% credibility interval were computed based on the resulting posterior samples. The total computational cost of all ClonalFrame runs combined was approximately 1000 CPU hours.

Selection of loci

All datasets analyzed here are based on multiple, selectively constrained housekeeping loci. The use of multiple loci buffers against possible variation in HRR across the genome as well as against stochastic variation. Intergenic spacer regions, genes under diversifying selection and genes encoding ribosomal subunits were not included because of potential confounding effects of selection on the detection of HRR. The *r/m* values surveyed here are taken to be representative of HRR of the majority of selectively constrained protein-encoding loci located on the chromosome (the 'core genome').

Selection of strains

Representative sampling of bacterial populations is required to estimate recombination rates that are biologically meaningful. There are two main ways in which non-representative sampling can lead to an underestimation of the actual recombination rate: (1) when multiple distinct populations are lumped together and (2) when certain genotypes are over-represented in a sample (Figure 1). Avoiding these pitfalls requires a detailed knowledge on the biology of the species in question.

Distinct populations within a species may emerge because of differential local adaptation and/or genetic drift. These clusters of closely related genotypes within a named species are often termed ecotypes (Cohan, 2002). It is plausible that ecotypes could differ in their HRR because of adaptive evolution or environmental constraints. When a population sample contains different ecotypes inhabiting distinct, spatially separated micro-niches that preclude the close contact necessary for genetic exchange, HRR will be underestimated. Similarly, ecotypes inhabiting identical micro-niches in different locations are less likely to exchange DNA than

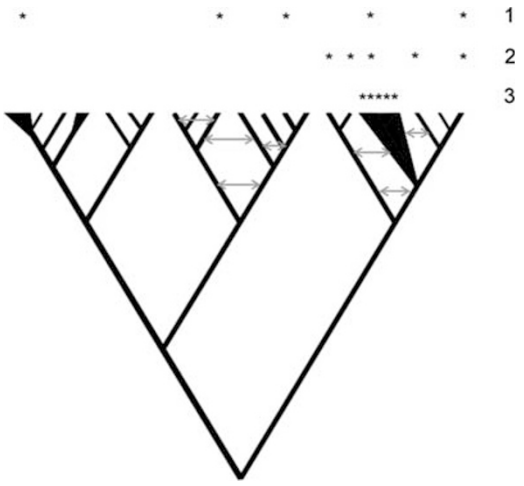

Figure 1 Sampling clones from a population. Homologous recombination events are depicted by arrows and take place primarily within separate evolutionary lineages (ecotypes). Asterisks represent sampled clones. Sampling scheme 1 is biased because it does not differentiate between distinct ecotypes. Sampling scheme 2 shows correct sampling from a distinct ecotype. Sampling scheme 3 is biased towards an epidemic clone within an ecotype (visualized by the increased width of the lineage).

clones from the same location. Evidence for this process has been found in the soil bacterium *Rhizobium leguminosarum*, where clonality was less pronounced at a regional scale than it was at a global scale (Souza *et al*., 1992).

Highly successful clones will become widespread in a population. Maynard Smith *et al*. (1993) first pointed out that the over-representation of closely related, high frequency (epidemic) clones in a sample will lead to an inflated estimate of clonality of the population as a whole. Oversampling of a single clone in an epidemic population structure will therefore result in an underestimation of HRR. Although pooling of distinct populations will generally result in an underestimation of HRR, it is possible to overestimate HRR of a given ecotype when it is lumped together with ecotypes that have higher HRR (Figure 1).

To avoid potential confounding effects of spatial population structure, local or regional strain collections were analyzed instead of global collections when possible. For studies where strains were found to cluster in multiple, deep-branching clades, only one such clade was analyzed to avoid possible pooling of distinct ecotypes (Cohan, 2002), each with possibly distinct HRR. Only one representative of each Sequence Type was included in the analysis to avoid possible effects of epidemic population structure.

Species classification

Bacterial species were classified according to ecology in the following broad groups: (1) extremophiles, (2) marine and aquatic bacteria, (3) terrestrial

bacteria, (4) commensals, that is, species that are part of the normal flora of humans or other animals, (5) obligate pathogens and (6) endosymbionts. Depending on their environment of origin, opportunistic pathogens are classified in group 2, 3 or 4. Within each group, species are divided according to phylum, with the proteobacteria further subdivided into the α-, β-, δ-, ε- and γ-divisions.

Results and discussion

Variation in HRR

Great variation in HRR was detected among species (Table 1). The lowest and highest *r/m* point estimates differ by three orders of magnitude. The upper 95% credibility interval of the species with lowest *r/m* and the lower 95% credibility interval of the species with highest *r/m* are over two orders of magnitude apart. It is obvious that homologous recombination is a powerful force in shaping the genetic diversity of a wide range of bacteria and archaea as its ability to change genomes exceed that of the process of mutation (that is, $r/m > 1$) in 56% (27/48) of the datasets analyzed. *Neisseria* and *Helicobacter* are frequently used as examples of bacteria with very high HRR, but lesser known species *Flavobacterium* and *Pelagibacter* were found to be even more recombinogenic.

Extremophiles

Extremophiles have attracted attention from microbial ecologists partly because the isolation of their habitats (such as geothermal vents, seeps, springs and salt lakes) results in potentially strongly structured populations, and therefore offer a special opportunity to study microbial biogeography. The hot spring inhabiting cyanobacterium *Mastigocladus laminosus* has a low-to-intermediate HRR. Two archaea, the thermoacidophile *Sulfolobus* and the halophile *Halorubrum*, have similar, intermediate HRR. Homologous recombination has been detected in the bacteria *Thermotoga* (Nesbo *et al.*, 2006) and *Leptospirillum* (Lo *et al.*, 2007) and the archaeon *Ferroplasma* (Tyson *et al.*, 2004; Eppley *et al.*, 2007), but these findings were based on non-MLST methods and so could not be included here.

Marine and aquatic bacteria (including opportunistic pathogens)

There has been a surge in sequence-based research on marine prokaryotes in recent years (for example, (Rusch *et al.*, 2007)). However, relatively few research efforts have focused at population level sequence variation. The oceanic species *Pelagibacter ubique* has very high HRR. HRR is also very high in the pelagic freshwater cyanobacterium *Microcystis* but low in the benthic marine cyanobacterium *Microcoleus*. MLST data on the marine cyanobacterium *Nodularia* were not reanalyzed as they were based on non-housekeeping loci but indicate high HRR (Hayes *et al.*, 2002). Environmental isolates of marine and estuarine *Vibrio parahaemolyticus* and *V. vulnificus* were found to have very high HRR. Disease-related lineages in both species show lowered HRR which is consistent with epidemic spread of a subset of virulent clones (Chowdhury *et al.*, 2004; Perez-Losada *et al.*, 2006; Bisharat *et al.*, 2007). HRR is high in the γ-proteobacterium *Plesiomonas shigelloides* found in freshwater and estuarine environments as well as in the gastrointestinal tracts of a wide variety of animals. It can cause gastrointestinal disease in humans after consumption of seafood or contact with untreated water (Salerno *et al.*, 2007).

Terrestrial bacteria (including opportunistic pathogens)

A number of MLST studies have been carried out for proteobacteria that live in soil, or are associated with plants. The α-proteobacterium *Rhizobium gallicum* has very low HRR, the β-proteobacterium *Ralstonia solanacearum* has intermediate HRR and the δ-proteobacterium *Myxococcus xanthus* has high HRR. The Pseudomonads are ubiquitous γ-proteobacteria in soil environments. HRR was found to be intermediate in both *Pseudomonas syringae* and *P. viridiflava*. Data on *P. stutzeri* (Cladera *et al.*, 2004) and Phi-producing Pseudomonads (Frapolli *et al.*, 2007) are indicative of similar HRR. The nitrogen fixing soil bacterium *Klebsiella pneumophila* is an important opportunistic pathogen for which we found a low HRR.

One of the first MLST studies on free-living bacteria investigated the Firmicutes *Bacillus cereus*, *B. thuringiensis* and *B. weihenstephanensis*, occurring sympatrically in soil (Sorokin *et al.*, 2006). In agreement with the original study, the first two species were found to have low HRR with *B. weihenstephanensis* having higher HRR. HRR was found to be higher in another local *B. thuringiensis* population isolated from clover ($r/m = 2.0$, credibility interval 1.2–3.1; see Supplementary Information). Firmicutes species for which less well-defined populations were sampled are *Listeria monocytogenes* and *Oenococcus oeni*.

Commensals (including opportunistic pathogens)

This group is largely composed of species that inhabit the gastrointestinal tract, the respiratory tract and skin. The gastrointestinal lifestyle of some commensals means that they can also be common in the environment. The β-proteobacterium *Neisseria meningitidis* is one of the best-known examples of bacteria with high HRR. The related, but never pathogenic commensal *N. lactamica* also has high HRR. The microaerophilic ε-proteobacteria in the genus *Campylobacter* inhabit the gut and can cause intestinal infection. A *C. insulaenigrae* population

isolated from northern elephant seals displayed high HRR as did *Campylobacter jejuni* isolated from farm animals and the environment. We classified *Helicobacter pylori* as an opportunistic pathogens as it inhabits the stomachs of over half the global human population but only occasionally causes disease (Falush *et al.*, 2001). It is one of the best-known examples of bacteria with very high HRR (Suerbaum *et al.*, 1998; Falush *et al.*, 2001).

The γ-proteobacterium *E. coli* was the first model species in the study of bacterial population structure (Guttman, 1997). It is a ubiquitous commensal in the intestine of mammals and birds, but certain types are also known to persist in the environment (Walk *et al.*, 2007). Although usually harmless, *E. coli* also encompasses several important pathogenic lineages. Strains belonging to well-defined clade ET-1 are prevalent in freshwater environments (Walk *et al.*, 2007) and have low HRR. The intestinal γ-proteobacterium *Salmonella enterica* on the other hand was found to have very high HRR. The γ-proteobacterium *Haemophilus influenza* is a commensal in the upper respiratory tract of humans (Gilsdorf, 1998); HRR was found to be moderately high. *H. parasuis* is a commensal and opportunistic pathogens of the respiratory tract of pigs (Olvera *et al.*, 2006). Two divergent lineages, one consisting of mainly non-pathogenic isolates (Table 1) and one consisting of mainly pathogenic isolates (Supplementary Information) were analyzed with the latter having higher recombination rate. The γ-proteobacterium *Moraxella catarrhalis* resides in the upper respiratory tract where it can cause diseases; HRR was found to be very high.

Streptococci are Firmicutes found on the skin, in the intestine and in the upper respiratory tract and can cause a range of infections. HRR is very high in *S. pneumoniae* and in *S. pyogenes*. *Staphylococcus aureus* inhabits skin and nasal mucus and can cause a variety of infections. HRR was found to be very low. *Clostridium difficile* can be found in low numbers in the gastrointestinal tract, where it can cause diarrhoea, as well as in the environment; data indicate very low HRR. The gastrointestinal opportunistic pathogens *Enterococcus faecalis* and *E. faecium* were found to have low and intermediate HRR respectively. The gastrointestinal commensal *Lactobacillus casei* has low HRR. The only Firmicute belonging to the class mollicutes for which data were available, *Mycoplasma hyopneumoniae*, has a moderately high HRR. Finally, the intestinal Spirochaetes *Brachyspira* and *Leptospira* were found to have very low HRR.

Obligate pathogens

Obligate pathogens are specialized parasites that are primarily associated with disease. However, as the biology of most species is not well-known, it is possible that some species classified as obligate pathogens are actually opportunists from the environment or unrepresentative strains of commensals. The α-proteobacterium *Bartonella* and the β-proteobacterium *Bordetella* exhibit very low genetic diversity and have very low HRR. HRR is low as well in the γ-proteobacterium *Yersinia pseudotuberculosis*. The γ-proteobacterium *Legionella pneumophila* is an intracellular pathogen of protozoa. The relatively high optimum temperature it prefers permits it to occasionally thrive in spas and cooling towers where it can be transmitted to human airways and cause Legionnaire's disease. A local population was found to exhibit low HRR. *Porphyromonas gingivalis*, causing periodontal disease in humans and *Flavobacterium psychrophilus*, causing disease in salmonid fish, are the only representatives of the phylum bacteroidetes for which MLST data are available. In sharp contrast with low HRR in *Porphyromonas*, *Flavobacterium* was found to have the highest HRR of all species analyzed here. *Chlamydia trachomatis* is an obligate intracellular pathogen belonging to the phylum Chlamydiae; HRR was found to be low.

Endosymbionts

The *Wolbachia* α-proteobacteria form a peculiar group of intercellular arthropod and nematode symbionts with genomes profoundly shaped by the loss and inactivation of genes (Tamas *et al.*, 2002). *Wolbachia* has high HRR. No information is available on other endosymbionts.

Association between HRR and phylogeny

The number of species for which MLST data are available is small, especially given the astounding diversity of bacteria and archaea. It is therefore not possible to statistically test whether HRR is elevated in certain phylogenetic or ecological groups. However, this dataset clearly demonstrates wide variation in HRR among species belonging to the same phylum or division. Examples are the γ-proteobacteria *Klebsiella* and *Vibrio* and the Bacteroidetes *Porphyromonas* and *Flavobacterium* (Table 1). An even more striking instance is provided by *S. enterica* for which we estimated a HRR almost 50 times higher than for *E. coli* despite the fact that they belong to the same family. Figure 2 shows the r/m point estimates of all phyla (or divisions) for which three or more representatives were available. It is evident that variation within phyla is of the same order as variation among phyla (note the $\log_{10}$-scale).

Several genera are represented by two different species in this study. HRR is similar for the pairs of *Vibrio*, *Streptococcus*, *Neisseria*, *Haemophilus*, *Campylobacter* and *Pseudomonas* species analyzed (Table 1). HRR of the three *Bacillus* species are also quite similar. The variation in HRR thus seems to decrease with finer scales of taxonomic resolution, as expected when phylogeny is an important

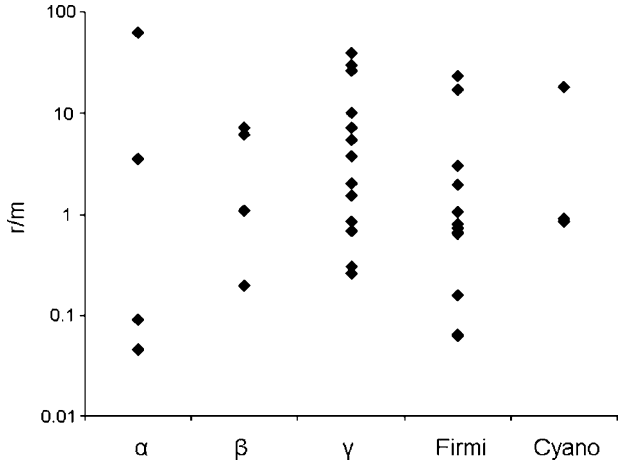

Figure 2 Range of inferred mean *r/m* values of the phylogenetic groupings for which three or more representative species were available. Greek letters refer to the Proteobacteria divisions, Firmi = Firmicutes, Cyano = Cyanobacteria. All *r/m* values (as well as credibility intervals) are listed in Table 1.

determinant of HRR. The role of phylogeny in determining HRR, however, is obscured by its strong correlation with ecology. For example, both *Neisseria* species are commensals of the nasopharynx and both *Pseudomonas* species are plant pathogens. When more data become available the hypothesis that the evolution of HRR is constrained at the genus level could be falsified.

Association between HRR and ecology

As found in earlier reviews (Feil *et al.*, 2001; Hanage *et al.*, 2006; Perez-Losada *et al.*, 2006), bacteria that cause disease vary widely in HRR. This is true for obligate pathogens, commensals and opportunistic pathogens from the environment. This is unsurprising, as great variation exists in pathogenic lifestyles, for example, in host species, host range, virulence, site of infection, mechanisms of immune evasion and host-to-host transmission. When considering truly free-living, non-animal associated species, one particular trend seems to emerge. HRR is high or very high in all marine and aquatic species examined, with the exception of *Microcoleus* (Table 1). Interestingly, the fish pathogen *Flavobacterium psychrophilum* also has very high HRR. In contrast, HRR of terrestrial bacteria analyzed is low or intermediate across all phyla/divisions analyzed, with the exception of *Myxococcus* (Table 1). Data for only three, widely divergent, extremophile species were analyzed; all three had similar HRR. Figure 3 shows all *r/m* values of the three types of free-living bacteria.

Conclusion

The comparative method is the most general way to approach patterns of evolutionary change (Harvey and Pagel, 1991). Most of the species for which data are currently available are (opportunistic) pathogens

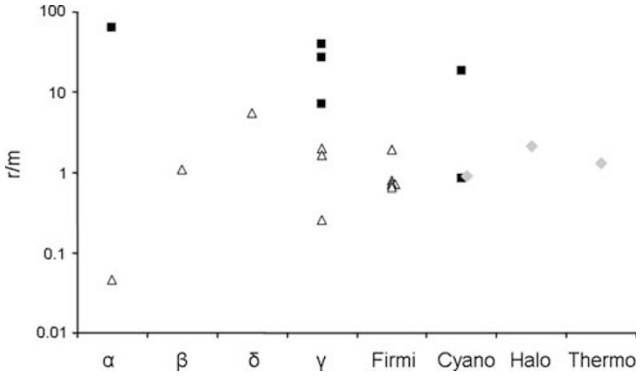

Figure 3 All inferred mean *r/m* values of terrestrial species (open triangles), marine/aquatic species (black squares) and extremophiles (grey diamonds). Therm = Thermoprotei, Halo = Halobacteria, other abbreviations as in Figure 1. Species included are: *B. cereus*, *B. thuringiensis*, *B. weihenstephanensis*, *K. pneumoniae*, *L. monocytogenes*, *M. xanthus*, *O. oeni*, *P. syringae*, *P. viridiflava*, *R. solanacearum*, *R. gallicum* (terrestrial), *M. aeruginosa*, *M. chtonoplastes*, *P. shigelloides*, *P. ubique*, *V. parahaemolyticus*, *V. vulnificus* (marine/aquatic) and *Halorubrum*, *M. laminosum* and *S. islandicus* (extremophile). All *r/m* values (as well as credibility intervals) are listed in Table 1.

or agronomically important bacteria. However, the MLST method is gaining popularity among microbial ecologists, and more data on the population structures of the overwhelming majority of free-living, non-pathogenic bacteria are expected in the near future. The accumulation of sequence-based, population-level studies will enable more systematic testing whether certain ecological variables correlate with a particularly high homologous recombination rate.

Acknowledgements

We thank Angus Buckling, Gabriel Perron and Martin Maiden for helpful comments. We thank all scientists who have contributed to the MLST databases: http://www.mlst.net, http://pubmlst.org, http://www.pasteur.fr/mlst and http://web.mpiib-berlin.mpg.de/mlst. We especially thank colleagues who have sent us their data. This work was supported by a Rubicon grant awarded by the Netherlands Organisation for Scientific Research (NWO) to MV.

References

Arvand M, Feil EJ, Giladi M, Boulouis HJ, Viezens J. (2007). Multi-locus sequence typing of *Bartonella henselae* isolates from three continents reveals hypervirulent and feline-associated clones. *PLoS ONE* **2**: e1346.

Baldo L, Dunning Hotopp JC, Jolley KA, Bordenstein SR, Biber SA, Choudhury RR *et al.* (2006). Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Appl Environ Microbiol* **72**: 7098–7110.

Bernstein H, Byers GS, Michod RE. (1981). Evolution of sexual reproduction: importance of DNA repair,

complementation, and variation. *Am Nat* **117**: 537–549.

Bisharat N, Cohen DI, Maiden MC, Crook DW, Peto T, Harding RM. (2007). The evolution of genetic structure in the marine pathogen, *Vibrio vulnificus*. *Infect Genet Evol* **7**: 685–693.

Castillo JA, Greenberg JT. (2007). Evolutionary dynamics of *Ralstonia solanacearum*. *Appl Environ Microbiol* **73**: 1225–1238.

Chowdhury NR, Stine OC, Morris JG, Nair GB. (2004). Assessment of evolution of pandemic *Vibrio parahaemolyticus* by multilocus sequence typing. *J Clin Microbiol* **42**: 1280–1282.

Cladera AM, Bennasar A, Barcelo M, Lalucat J, Garcia-Valdes E. (2004). Comparative genetic diversity of *Pseudomonas stutzeri* genomovars, clonal structure, and phylogeny of the species. *J Bacteriol* **186**: 5239–5248.

Cohan FM. (2002). What are bacterial species? *Annu Rev Microbiol* **56**: 457–487.

Coscolla M, Gonzalez-Candelas F. (2007). Population structure and recombination in environmental isolates of *Legionella pneumophila*. *Environ Microbiol* **9**: 643–656.

de Las Rivas B, Marcobal A, Munoz R. (2004). Allelic diversity and population structure in *Oenococcus oeni* as determined from sequence analysis of housekeeping genes. *Appl Environ Microbiol* **70**: 7210–7219.

Diancourt L, Passet V, Chervaux C, Garault P, Smokvina T, Brisse S. (2007). Multilocus multilocus sequence typing of *Lactobacillus casei* (*L. paracasei*) reveals a clonal population structure with low levels of homologous recombination. *Appl Environ Microbiol* **73**: 6601–6611.

Diancourt L, Passet V, Verhoef J, Grimont PA, Brisse S. (2005). Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J Clin Microbiol* **43**: 4178–4182.

Diavatopoulos DA, Cummings CA, Schouls LM, Brinig MM, Relman DA, Mooi FR. (2005). *Bordetella pertussis*, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of B. bronchiseptica. *PLoS Pathog* **1**: e45.

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

Enersen M, Olsen I, van Winkelhoff AJ, Caugant DA. (2006). Multilocus sequence typing of *Porphyromonas gingivalis* strains from different geographic origins. *J Clin Microbiol* **44**: 35–41.

Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J Clin Microbiol* **38**: 1008–1015.

Enright MC, Spratt BG, Kalia A, Cross JH, Bessen DE. (2001). Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between emm type and clone. *Infect Immun* **69**: 2416–2427.

Eppley JM, Tyson GW, Getz WM, Banfield JF. (2007). Genetic exchange across a species boundary in the archaeal genus ferroplasma. *Genetics* **177**: 407–416.

Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M *et al.* (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci USA* **98**: 15056–15061.

Fearnhead P, Smith NG, Barrigas M, Fox A, French N. (2005). Analysis of recombination in *Campylobacter jejuni* from MLST population data. *J Mol Evol* **61**: 333–340.

Feil EJ. (2004). Small change: keeping pace with microevolution. *Nature Rev Microbiol* **2**: 483–495.

Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC *et al.* (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci USA* **98**: 182–187.

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**: 1518–1530.

Fisher RA. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press: Oxford.

Frapolli M, Defago G, Moenne-Loccoz Y. (2007). Multilocus sequence analysis of biocontrol fluorescent Pseudomonas spp. producing the antifungal compound 2,4-diacetylphloroglucinol. *Environ Microbiol* **9**: 1939–1955.

Fraser C, Hanage WP, Spratt BG. (2005). Neutral microepidemic evolution of bacterial pathogens. *Proc Natl Acad Sci USA* **102**: 1968–1973.

Gelman A, Rubin DB. (1992). Inference from iterative simulation using multiple sequences. *Stat Sci* **7**: 457–472.

Gilsdorf JR. (1998). Antigenic diversity and gene polymorphisms in *Haemophilus influenzae*. *Infect Immun* **66**: 5053–5059.

Gonzalez-Escalona N, Martinez-Urtaza J, Romero J, Espejo RT, Jaykus LA, Depaola A. (2008). Determination of molecular phylogenetics of *Vibrio parahaemolyticus* strains by multilocus sequence typing. *J Bacteriol* **190**: 2831–2840.

Goss EM, Kreitman M, Bergelson J. (2005). Genetic diversity, recombination and cryptic clades in *Pseudomonas viridiflava* infecting natural populations of Arabidopsis thaliana. *Genetics* **169**: 21–35.

Guttman DS. (1997). Recombination and clonality in natural populations of *Escherichia coli*. *Trends Ecol Evol* **12**: 16–22.

Guttman DS, Dykhuizen DE. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**: 1380–1383.

Hanage WP, Fraser C, Spratt BG. (2006). The impact of homologous recombination on the generation of diversity in bacteria. *J Theor Biol* **239**: 210–219.

Hanage WP, Kaijalainen T, Herva E, Saukkoriipi A, Syrjanen R, Spratt BG. (2005). Using multilocus sequence data to define the pneumococcus. *J Bacteriol* **187**: 6223–6230.

Harvey PH, Pagel MD. (1991). *The Comparative Method in Evolutionary Biology*. Oxford University Press: Oxford.

Hayes PK, Barker GL, Batley J, Beard SJ, Handley BA, Vacharapiyasophon P *et al.* (2002). Genetic diversity within populations of cyanobacteria assessed by analysis of single filaments. *Antonie Leeuwenhoek* **81**: 197–202.

Homan WL, Tribe D, Poznanski S, Li M, Hogg G, Spalburg E *et al.* (2002). Multilocus sequence typing scheme for *Enterococcus faecium*. *J Clin Microbiol* **40**: 1963–1971.

Hudson RR. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* **23**: 183–201.

Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MC. (2005). The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Mol Biol Evol* **22**: 562–569.

Jukes TH, Cantor CR. (1969). Evolution of protein molecules. In: Munro HN (ed). *Mammalian Protein Metabolism III*. Academic Press: New York, pp 21–132.

Kingman JFC. (1982). The coalescent. *Stochastic Processes and their Applications* **13**: 235–248.

Lemee L, Dhalluin A, Pestel-Caron M, Lemeland JF, Pons JL. (2004). Multilocus sequence typing analysis of human and animal *Clostridium difficile* isolates of various toxigenic types. *J Clin Microbiol* **42**: 2609–2617.

Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G *et al.* (2007). Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**: 537–541.

Lodders N, Stackebrandt E, Nubel U. (2005). Frequent genetic recombination in natural populations of the marine cyanobacterium *Microcoleus chthonoplastes*. *Environ Microbiol* **7**: 434–442.

Maiden MCJ. (2006). Multilocus sequence typing of bacteria. *Annu Rev Microbiol* **60**: 561–588.

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R *et al.* (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* **95**: 3140–3145.

Mau B, Glasner JD, Darling AE, Perna NT. (2006). Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol* **7**: R44.

Maynard Smith J, Smith NH, O'Rourke M, Spratt BG. (1993). How clonal are bacteria? *Proc Natl Acad Sci USA* **90**: 4384–4388.

Mayor D, Zeeh F, Frey J, Kuhnert P. (2007). Diversity of *Mycoplasma hyopneumoniae* in pig farms revealed by direct molecular typing of clinical material. *Vet Res* **38**: 391–398.

Meats E, Feil EJ, Stringer S, Cody AJ, Goldstein R, Kroll JS *et al.* (2003). Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. *J Clin Microbiol* **41**: 1623–1636.

Michod RE, Bernstein H, Nedelcu AM. (2008). Adaptive value of sex in microbial pathogens. *Infect Genet Evol* **8**: 267–285.

Milkman R, Bridges MM. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* **126**: 505–517.

Miller SR, Castenholz RW, Pedersen D. (2007). Phylogeography of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Appl Environ Microbiol* **73**: 4751–4759.

Narra HP, Ochman H. (2006). Of what use is sex to bacteria? *Curr Biol* **16**: R705–R710.

Nesbo CL, Dlutek M, Doolittle WF. (2006). Recombination in Thermotoga: implications for species concepts and biogeography. *Genetics* **172**: 759–769.

Nicolas P, Mondot S, Achaz G, Bouchenot C, Bernardet JF, Duchaud E. (2008). Population structure of the fish-pathogenic bacterium *Flavobacterium psychrophilum*. *Appl Environ Microbiol* **74**: 3702–3709.

Olvera A, Cerda-Cuellar M, Aragon V. (2006). Study of the population structure of *Haemophilus parasuis* by multilocus sequence typing. *Microbiol* **152**: 3683–3690.

Pannekoek Y, Morelli G, Kusecek B, Morre SA, Ossewaarde JM, Langerak AA *et al.* (2008). Multi locus sequence typing of Chlamydiales: clonal groupings within the obligate intracellular bacteria *Chlamydia trachomatis*. *BMC Microbiol* **8**: 42.

Papke RT, Koenig JE, Rodriguez-Valera F, Doolittle WF. (2004). Frequent recombination in a saltern population of Halorubrum. *Science* **306**: 1928–1929.

Perez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. (2006). Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol* **6**: 97–112.

Posada D. (2002). Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol Biol Evol* **19**: 708–717.

Rasback T, Johansson KE, Jansson DS, Fellstrom C, Alikhani MY, La T *et al.* (2007). Development of a multilocus sequence typing scheme for intestinal spirochaetes within the genus Brachyspira. *Microbiol* **153**: 4074–4087.

Redfield RJ. (1993). Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. *J Hered* **84**: 400–404.

Redfield RJ. (2001). Do bacteria have sex? *Nature Rev Genet* **2**: 634–639.

Ruiz-Garbajosa P, Bonten MJ, Robinson DA, Top J, Nallapareddy SR, Torres C *et al.* (2006). Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. *J Clin Microbiol* **44**: 2220–2228.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Salcedo C, Arreaza L, Alcala B, de la Fuente L, Vazquez JA. (2003). Development of a multilocus sequence typing method for analysis of *Listeria monocytogenes* clones. *J Clin Microbiol* **41**: 757–762.

Salerno A, Deletoile A, Lefevre M, Ciznar I, Krovacek K, Grimont P *et al.* (2007). Recombining population structure of the Enterobacteriaceae *Plesiomonas shigelloides* revealed by multilocus sequence typing. *J Bacteriol* **189**: 7808–7818.

Sarkar SF, Guttman DS. (2004). Evolution of the core genome of Pseudomonas syringae, a highly clonal, endemic plant pathogen. *Appl Environ Microbiol* **70**: 1999–2012.

Silva C, Vinuesa P, Eguiarte LE, Souza V, Martinez-Romero E. (2005). Evolutionary genetics and biogeographic structure of *Rhizobium gallicum* sensu lato, a widely distributed bacterial symbiont of diverse legumes. *Mol Ecol* **14**: 4033–4050.

Sorokin A, Candelon B, Guilloux K, Galleron N, Wackerow-Kouzova N, Ehrlich SD *et al.* (2006). Multiple-locus sequence typing analysis of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct population structure of psychrotrophic strains. *Appl Environ Microbiol* **72**: 1569–1578.

Souza V, Nguyen TT, Hudson RR, Pinero D, Lenski RE. (1992). Hierarchical analysis of linkage disequilibrium in rhizobium populations—evidence for sex. *Proc Natl Acad Sci USA* **89**: 8389–8393.

Spratt BG, Hanage WP, Li B, Aanensen DM, Feil EJ. (2004). Displaying the relatedness among isolates of bacterial species—the eBURST approach. *FEMS Microbiol Lett* **241**: 129–134.

Stoddard RA, Miller WG, Foley JE, Lawrence J, Gulland FM, Conrad PA *et al.* (2007). *Campylobacter insulaenigrae* isolates from northern elephant seals (*Mirounga angustirostris*) in California. *Appl Environ Microbiol* **73**: 1729–1735.

Stumpf MP, McVean GA. (2003). Estimating recombination rates from population-genetic data. *Nature Rev Genet* **4**: 959–968.

Suerbaum S, Smith JM, Bapumia K, Morelli G, Smith NH, Kunstmann E *et al.* (1998). Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci USA* **95**: 12619–12624.

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ *et al.* (2002). 50 million years of genomic stasis in Endosymbiotic Bacteria. *Science* **296**: 2376–2379.

Tanabe Y, Kasai F, Watanabe MM. (2007). Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *Microcystis aeruginosa*. *Microbiol* **153**: 3695–3703.

Thaipadungpanit J, Wuthiekanun V, Chierakul W, Smythe LD, Petkanchanapong W, Limpaiboon R *et al.* (2007). A dominant clone of *Leptospira interrogans* associated with an outbreak of human leptospirosis in Thailand. *PLoS Negl Trop Dis* **1**: e56.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappe MS, Giovannoni SJ. (2007). High intraspecific recombination rate in a native population of Candidatus *pelagibacter ubique* (SAR11). *Environ Microbiol* **9**: 2430–2440.

Vos M, Velicer GJ. (2008). Isolation by distance in the spore-forming soil bacterium *Myxococcus xanthus*. *Curr Biol* **18**: 386–391.

Walk ST, Alm EW, Calhoun LM, Mladonicky JM, Whittam TS. (2007). Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environ Microbiol* **9**: 2274–2288.

Watterson GA. (1978). Homozygosity test of neutrality. *Genetics* **88**: 405–417.

Whitaker RJ, Grogan DW, Taylor JW. (2005). Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol Evol* **22**: 2354–2361.

Wright S. (1931). Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)