Published in final edited form as:

Trends Microbiol. 2010 July ; 18(7): 315–322. doi:10.1016/j.tim.2010.04.002.

Impact of recombination on bacterial evolution

Xavier Didelot¹ and Martin C.J. Maiden²

¹Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

²Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3SY, UK

Abstract

Genetic exchange plays a defining role in the evolution of many bacteria. The recent accumulation of nucleotide sequence data from multiple members of diverse bacterial genera has facilitated comparative studies that have revealed many features of this process. Here we focus on genetic exchange that has involved homologous recombination and illustrate how nucleotide sequence data have furthered our understanding of: (i) the frequency of recombination; (ii) the impact of recombination in different parts of the genome; and (iii) patterns of gene flow within bacterial populations. Summarizing the results obtained for a range of bacteria, we survey evidence indicating that the extent and nature of recombination vary widely among microbiological species and often among lineages assigned to the same microbiological species. These results have important implications in studies ranging from epidemiological investigations to examination of the bacterial species problem.

Recombination in bacteria

Genetic exchange, once thought to be uncommon in asexually reproducing bacteria, is now known to be a major driving force in the evolution of most prokaryotes. Indeed, the lack of genetic exchange among bacteria can now be regarded as an unusual situation, confined to a few lineages such as genetically monomorphic pathogens [1]. Gene transfer among and within bacterial populations is mediated by the three mechanisms of conjugation, transduction and transformation [2]. These processes promote the acquisition of novel genetic elements from the ‘accessory gene pool’, the impact of which has been extensively studied in human and animal pathogens and commensals, where they are often associated with the emergence of new phenotypes [3]. Bacteria also frequently import genes, or fragments of them, in place of existing homologous genetic material in their genome, a process that was first identified by the observation of mosaic genes at loci encoding antigens or antibiotic resistance [4,5]. This phenomenon, called homologous recombination, is widespread throughout the genomes of many bacteria and is usually a consequence of RecA-mediated homology-dependent recombination. Here we review the impact that studies of homologous recombination have had on our understanding of bacterial evolution.

Throughout this review, we only consider effective recombination events, defined as events that have a measurable effect on genetic material. Non-effective recombination, for example when imported DNA replaces an identical fragment in the recipient genome, cannot be directly observed, although this process is almost certainly very frequent when the donor and recipient are closely related and might be crucial as a method of DNA repair; indeed, this repair function might be the most important biological role of homologous recombination [6]. Measurement of recombination in terms of its capacity to promote

genetic change is especially important because selection often causes some recombination events to be quickly purged from the population and others to become widespread. It is, however, difficult to disentangle the recombination process and the selection exerted on it. For example, an increased rate of effective recombination might be caused by either an increase in the rate at which recombination events occur or by a relaxation of the negative selective pressure acting against the novel alleles generated.

A number of data sources have been used to investigate genetic exchange in bacteria. In particular, the widespread application of multilocus sequence typing (MLST) [7] to a large number of different bacteria has provided a wealth of data that has been used to investigate homologous recombination. This typing technique indexes nucleotide sequence of fragments of housekeeping genes, typically seven in number and each of ~450 bp in length, which are distributed around the genome so that a single recombination event would be unlikely to affect more than one fragment. Multilocus data can be used to assess the degree of recombination in a number of ways. First, they can be used to test predictions of a fully clonal population structure. In the absence of any form of genetic exchange, all genetic variation will be generated by *de novo* point mutations and will be characterized by the testable predictions of: (i) linkage disequilibrium; (ii) a tree-like phylogeny; and (iii) congruence – the property of exhibiting the same phylogeny at all loci (Figure 1). Second, multilocus sequence data can be used to quantify the extent of departure from the null expectation of clonality in terms of its frequency and nature [8].

More recently, studies based on comparisons of whole genome sequences have become possible [9]. These have the potential to reveal much more about the recombination process than studies based on seven-locus MLST, because a limited number of housekeeping genes might not be representative of the entire genome; however, at the time of writing whole genome studies are limited to a handful of isolates (typically between 5 and 20), which have often been chosen for specific reasons and are only poorly representative of any bacterial population (Box 1). In addition, a number of computational and bioinformatics challenges have to be met before the whole genomes of many bacteria can be analyzed comparatively. The current situation is therefore one of complementarity between MLST and whole genome studies.

In this review we consider three aspects of the recombination process of a given bacterial population in turn: (i) the frequency with which members of the population are affected by recombination events; (ii) those regions of their genomes that are exchanged; and (iii) the source of the imported genetic material.

Rates of recombination

Several different approaches have been used to estimate recombination frequencies in natural populations of bacteria. The rate of recombination relative to that of mutation is a measure often used [10]. A relative rate of recombination of 5, for example, means that recombination has occurred five times as often as mutation during the evolution of the population investigated. Each recombination event is likely to introduce several substitutions, so a relative recombination rate of 1 will usually result in a greater per-site effect of recombination than mutation. If the mutation and recombination events that occurred could be reconstructed, then a straightforward estimate for the relative rate of recombination would be the ratio of numbers of recombination and mutation events reconstructed. Unfortunately, such reconstructions are difficult to perform except for the most recent events [8], which might not be representative of the entire history of the population. Relative rates of recombination are therefore typically estimated based on

summary statistics of the nucleotide sequence data (Box 2) or using reconstruction algorithms such as the ClonalFrame software (Box 3).

Comparison of results for analyses performed with different methodologies is problematic [11]; nevertheless, studies using the same methods across different genera have suggested wide variation in recombination rates (Figure 2). For example, a comparative study of MLST data from seven diverse bacteria suggested that recombination rates varied by two orders of magnitude from 0.05 for *Bacillus cereus* to 7.21 for *Streptococcus pyogenes* [12,13]. A different comparison of 16 bacterial pathogens, based on patterns of linkage disequilibrium, also revealed widely varying relative recombination rates, ranging from 0 for *Escherichia coli* to 29.3 for *Neisseria gonorrhoeae* [14]. A comparison of the relative effects of recombination and mutation across 46 bacterial species revealed variation in recombination rates of over three orders of magnitude [15]. Major differences are also observed between populations of very closely related bacteria, for example between different microbiological species of the genus *Francisella* [16]. The relative effects of recombination do not seem to correlate with the deep phylogeny of the bacterial kingdom either [15].

An issue that remains to be resolved is the poor consensus concerning the levels of recombination of some species, even among fairly recent studies. For example, several studies revealed that *B. cereus* recombines only rarely, with estimated relative rates approximately 0.05 [13], 0.3 [17] and 0.2 [18], but one study estimated that recombination occurs almost twice as often as mutation [14]. Conflicting levels of recombination have also been reported for *E. coli*, *Staphylococcus aureus* and *Haemophilus influenzae* [14]. Some of these discrepancies are likely to be caused by differences in analytical methodologies [11]. Another likely explanation is the fact that analytical methods are sensitive to the sampling strategy used to collect bacterial isolates: samples meant to represent a whole population often differ in the frequency of specific lineages, thus introducing substantial bias for the estimates obtained (Box 1).

Given the variation in relative recombination rates, even between closely related microbiological species, and the continuing debate over the definition of bacterial species [19,20], it is not surprising that rate variation has been reported among the lineages of a number of species. For example, the seroresistant clade of *Moraxella catarrhalis*, which includes most of the virulent strains of this organism, has a relative recombination rate six times higher than the serosensitive population [21]. Differences of the same order of magnitude have been found in favor of lineage II over lineage I in *Listeria monocytogenes* [22,23] as well as between the hyperinvasive lineages of *Neisseria meningitidis* [24]. Hints of inconstant recombination rates have also been uncovered between groups 1 and 2 of *S. aureus* [25] and between the three clades of the *B. cereus* group [18,26]. Again, non-random sampling of the natural population remains a possible explanation, but further study is required to determine the role of changes in recombination rates in the emergence of pathogenesis.

Closely related bacteria with distinct ecological characteristics can display very different rates of recombination. Several highly specialized pathogens show fewer signs of recombination than their non-specialized relatives [1]; examples of this include *Francisella tularensis* [16], *Yersinia pestis* [27], *Bacillus anthracis* [28] and *Mycobacterium tuberculosis* [29]. In these cases, recombination rates have decreased as a consequence of specialization. Nevertheless, increased virulence is sometimes accompanied by higher rates of recombination, for example in the pathogenic lineages of *E. coli* [30] or the seroresistant clade of *M. catarrhalis* [21]. *Salmonella enterica* Typhi, the cause of typhoid fever, is a particularly interesting case. This is a single clone that has become restricted to infection of

human hosts. During this process it exchanged a quarter of its genome with Paratyphi A, corresponding to a highly atypical rate of recombination for this species [31], but it seems that this process has ceased [32,33]. It is intriguing to speculate that this pattern of high recombination during adaptation, followed by low recombination once adapted, might be common in the evolution of many bacterial pathogens, but further examples are required to confirm this hypothesis.

Changes in the rate of recombination associated with adaptation can be explained by either neutral or selective arguments. As adaptation to a new lifestyle occurs, the opportunity for recombination with most other lineages decreases, whereas prospects for recombination with bacteria sharing the same lifestyle increase, for example during the adaptation of both Typhi and Paratyphi A to the human host [31]. The net result can be either an increase or a decrease in the rate of recombination on a change in lifestyle. However, the adaptive process is likely to shift the selective pressures exerted on any adapting lineage, which could result in a higher or lower rate of recombination and/or mutation, depending on the exact conditions [34]. Importantly, neutral and selective reasons for a shift in recombination rates are not mutually exclusive and might amplify each other. A full explanation of the emergence of bacterial pathogens will require disentanglement of these processes.

Genomic regions affected by recombination

When recombination occurs, a fragment of imported genetic material is integrated in the genome of the recipient. The imported fragment is usually assumed to be contiguous, although two recent laboratory results have shown that this is not always the case after transformation of *Helicobacter pylori* [35,36]. Large recombination events are thought to be rare in nature, although a few have been observed, most of which were probably caused by conjugative gene transfer followed by positive selection. For example, very large replacements have been observed in resistant lineages of *S. aureus* [37] and *Streptococcus pneumoniae* [38], and several large events were found in the genome of *Streptococcus agalactiae* [39]. The great majority of recombination events observed to date range from a few hundred to a few thousand base pairs, as estimated using MLST data for *N. meningitidis* [40], *Campylobacter jejuni* [41,42] and *B. cereus* [17,18], and whole genome data for *S. enterica* [17,31]. The length of imports occurring in nature is typically much smaller than those observed in *in vitro* studies, as noted for example in *H. pylori* [35], *C. jejuni* [41] and *E. coli* [43]. This difference might be due to the fact that in nature, larger events are more likely to reduce fitness and therefore be purged before observation [40].

Several studies have reported differences in the prevalence of recombination at different regions of the same bacterial genome. In *E. coli*, mismatch repair genes exhibit higher mosaicism than housekeeping genes [44]. Much variation was found in the relative recombination rates estimated on a gene-by-gene basis for housekeeping genes of *B. cereus*, *H. influenzae* and species of *Neisseria*, *Staphylococcus* and *Streptococcus* [14]. No evidence was found, however, of any specific locus with a systematic high or low rate across microbiological species. For example, the gene *gdh* encoding glutamate dehydrogenase has a relatively high recombination rate in *N. gonorrhoeae* compared to the same locus in *N. meningitidis* [14]. All these loci (but one) were found to be under negative selection at different rates from one species to another [14], which might explain the lack of consistency in the recombination rates observed.

A comparison of 26 whole genomes from the genus *Streptococcus* revealed that recombination occurs more often in regions under positive selection [45]. A similar result was found when comparing five genomes of *Listeria* [23]. In particular, genomic regions coding for proteins with a role in pathogenicity are often under positive selection [46] and

often exhibit higher rates of recombination. Examples include the *rlrA* islet of *S. pneumonia*, which encodes a pilus that increases virulence [47], the O-antigen *rfb* region and *fim* loci involved in adhesion to host cells in *E. coli* [43,48], and the internalin *inlA* and actin nucleator *actA* genes in *L. monocytogenes* [22,49]. Although recombination is relatively rare in *M. tuberculosis* and *Chlamydia trachomatis*, the MT0105 locus, which encodes a protein critical for host–pathogen interaction in *M. tuberculosis*, and the gene *ompA*, which encodes the major outer membrane protein in *C. trachomatis*, have both undergone significant recombination [29,50,51].

The fact that recombination is especially prevalent in positively selected regions of the bacterial genome can be explained if we consider that the only recombination we are likely to observe is that which brings together beneficial mutations and removes deleterious ones, thus allowing a much faster increase in fitness than would otherwise be possible by mutation alone [52]. This idea has been proposed as an alternative explanation for the existence of recombination in bacteria [52].

Patterns of gene flow

Several factors make recombination more likely to occur between closely related bacteria. The first factor is physical proximity, which is required for transformation, transduction or conjugation to take place. This proximity is more likely to occur between members of the same community, which are likely to be related because many species of bacteria show significant geographical and ecological structuring. The geographical structure of *H. pylori* has been described using the program STRUCTURE (Box 4), with admixture more frequent among types from the same part of the world [53–56]. For *C. jejuni*, an ecological structure has been described that depends on the infected host, which results in more frequent recombination among types sharing the same host [57–59]. A second factor reducing the rate of recombination among unrelated bacteria is the homology dependence of recombination as observed in the laboratory: in many bacterial species, the probability of acceptance of a recombination event decreases exponentially with genetic distance between the donor and recipient DNA [60]. Finally, in the absence of major environmental change, import of distantly related genetic material is more likely to reduce the fitness of the recipient and therefore be removed by negative selection.

Even when some of the above reasons do not apply, analyses of genetic diversity have shown that recombination occurs more often among members of the same microbiological species than members of different species, and more often within lineages of a species than among them. The plant pathogen *Pseudomonas viridiflava* has two distinct clades, which are not geographically differentiated, but recombination is still more frequent within clades than among them [61]. Despite the absence of a geographical structure for the cyanobacterium *Microcystis aeruginosa* and apparent neutrality at the *mcy* cluster encoding microcystin toxins, two phylogenetic clades of the cluster exist and recombinational exchange is more common within than between them [62]. Recombination in the *B. cereus* group is less constrained by homology dependence than in many other species, leading to frequent imports from external sources, but gene flow is still more important within each of the three clades than among them [18].

The overall picture of gene flow in most bacterial species is therefore dominated by exchanges between close relatives. This might even be more pronounced than reported, because imports from a distant origin are easier to detect as they introduce more nucleotide substitutions [11]. Yet recombination sometimes occurs at surprisingly high rates between members of separate lineages or species. In *Campylobacter*, for example, species *C. jejuni* and *C. coli* are separated by a genetic distance of ~3.5% and yet show significant signs of

recombination between one another, although at a lower rate than within each species [42,63]. These exchanges might be a result of recent changes in their ecology owing to human activity, and could even eventually lead to convergence or 'despeciation' of the two bacterial populations [64].

Several examples of interspecific recombination have been observed in the genus *Streptococcus*. Genetic exchanges are frequent between the zoonotic pathogen *S. zooepidemicus* and the pathogens *S. equi* and *S. pyogenes*, which infect horses and humans, respectively [65,66]. Recombination has also been detected between the two oral species *S. mitis* and *S. oralis* and the two pathogens *S. pneumoniae* and *S. pseudopneumoniae*, all of which are closely related [67-69]. Imports from the other species into *S. pneumoniae* have in fact occurred mostly in a lineage associated with antibiotic resistance [70], which might represent a novel lifestyle compared to other lineages. In summary, despite the many reasons why recombination should happen pre-eminently between close relatives, exchanges between members of separate lineages or species are occasionally observed at surprisingly high levels, which can often be explained by selective pressures.

Concluding remarks and future perspectives

The ever-increasing availability of contiguous nucleotide sequences from multiple regions of bacterial genomes has revealed the central role of homologous recombination in the evolution of most bacterial populations. Clonality now seems to be a special case that relies on extreme genetic isolation associated with very specialized ecologies, especially in obligate pathogens. Sequence data at the population level have led to a better understanding of the nature and rates of recombination in a variety of bacteria, but much remains to be learned concerning the impact of differing rates of recombination, both in terms of the causes of variations in recombination rates and the role that these differences play in the structuring of bacterial populations at all levels. This is of particular importance in resolving the issue of how to define bacterial taxonomic groups such as species [20,71,72].

Although MLST studies have brought a new understanding of the recombination process over the past decade, with thousands of isolates typed in some species [7], this method also has limitations in terms of the proportion of the genome analyzed and the nature of the variation observed. The increase in the number of genomes being sequenced has facilitated a number of recent comparative studies with the potential to reveal more about the recombination process and in particular its spatial properties along the genome. These whole genome comparisons are currently based on small numbers of distantly related genomes, which greatly limits their value, especially in terms of identifying patterns of gene flow between lineages.

New very high-throughput parallel sequencing technologies will soon facilitate comparative genomic studies based on hundreds of genomes [59,73]. Despite the opportunities offered by these new technologies, significant challenges remain. First, to make the most of these new large-scale data sets will require the development of new methods of analysis. All the methods currently used (Boxes 2, 3 and 4) have limitations and none can handle the very large amounts of data currently being generated, let alone the quantities that will become available in the immediate future. Explicit modeling of clonal genealogy using Clonal-Frame (Box 3) provides a basis for future developments because unlinked regions of the genome are approximately independent given the clonal genealogy [17,74]. Therefore, if the clonal genealogy can be reconstructed and the genome divided up into approximately unlinked regions, then recombination detection can be performed for each region separately, thus greatly reducing the overall computational cost; however, the genealogical depth to which the clonal signal persists in recombining populations remains an open question.

The second challenge is the difficult issue of adequate and consistent sampling of microbial populations and communities (Box 1). Although novel sequencing technologies and the emerging field of metagenomics will provide some solutions, there can be no substitute for careful structured sampling and archiving of bacterial specimens and isolates of known provenance. Such specimens will facilitate rigorous comparison of the rate and nature of recombination in different populations and of the association of any differences observed with differences in phenotype. Such data could lead to a more complete understanding of the role of recombination in the evolution of bacteria and the emergence of novel phenotypes.

Acknowledgments

Xavier Didelot is a CRiSM Research Fellow. Martin Maiden is a Wellcome Trust Senior Fellow. We thank Mark Achtman, Daniel Falush, William Hanage and three anonymous reviewers for providing useful comments, ideas and discussions.

Glossary

Adaptation	evolutionary process whereby a lineage becomes better suited to its environment and potentially less suited to other environments
Admixture	results from recombination among members of distinct evolutionary lineages
Ecological structure	non-random association of evolutionary lineages and the ecological niches in which they are found
Geographical structure	non-random association of evolutionary lineages and the geographical location where they are found
Lineage	a subgroup of bacteria within a species with a common evolutionary origin
Linkage disequilibrium	non-random association of genotypes at different loci
Negative selection	evolutionary mechanism that tends to remove novel mutations because they reduce fitness. Negative selection is also called purifying or stabilizing selection and is the most frequent form of natural selection
Neutrality	the absence of either positive or negative selection
Positive selection	evolutionary mechanism that tends to favor novel mutations because they increase fitness. Positive selection is sometimes called diversifying selection
Relative rate of recombination	the rate of occurrence of recombination measured as a fraction of the mutational rate

References

1. Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* 2008; 62:53–70. [PubMed: 18785837]
2. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 2005; 3:711–721. [PubMed: 16138099]
3. Hacker J, Carniel E. Ecological fitness, genomic islands and bacterial pathogenicity: a Darwinian view of the evolution of microbes. *EMBO Rep.* 2001; 2:376–381. [PubMed: 11375927]

4. Feavers IM. Role of horizontal genetic exchange in the antigenic variation of the class 1 outer membrane protein of *Neisseria meningitidis*. *Mol. Microbiol.* 1992; 6:489–495. [PubMed: 1560777]
5. Spratt BG, et al. Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. *J. Mol. Evol.* 1992; 34:115–125. [PubMed: 1556747]
6. Michod RE, et al. Adaptive value of sex in microbial pathogens. *Infect. Genet. Evol.* 2008; 8:267–285. [PubMed: 18295550]
7. Maiden M. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* 2006; 60:561–588. [PubMed: 16774461]
8. Feil EJ, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U. S. A.* 2001; 98:182–187. [PubMed: 11136255]
9. Abby S, Daubin V. Comparative genomics and the evolution of prokaryotes. *Trends Microbiol.* 2007; 15:135–141. [PubMed: 17289390]
10. Milkman R, Bridges MM. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics.* 1990; 126:505–517. [PubMed: 1979037]
11. Didelot, X.; Falush, D. Horizontal Gene Transfer in the Evolution of Pathogenesis. Cambridge University Press; 2008. Bacterial recombination in vivo; p. 23–48.
12. Fraser C, et al. Neutral microepidemic evolution of bacterial pathogens. *Proc. Natl. Acad. Sci. U. S. A.* 2005; 102:1968–1973. [PubMed: 15684071]
13. Hanage WP, et al. The impact of homologous recombination on the generation of diversity in bacteria. *J. Theor. Biol.* 2006; 239:210–219. [PubMed: 16236325]
14. Perez-Losada M, et al. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect. Genet. Evol.* 2006; 6:97–112. [PubMed: 16503511]
15. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 2009; 3:199–208. [PubMed: 18830278]
16. Larsson P, et al. Molecular evolutionary consequences of niche restriction in *Francisella tularensis*, a facultative intracellular pathogen. *PLoS Pathog.* 2009; 5:e1000472. [PubMed: 19521508]
17. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. *Genetics.* 2007; 175:1251–1266. [PubMed: 17151252]
18. Didelot X, et al. Evolution of pathogenicity in the *Bacillus cereus* group. *Syst. Appl. Microbiol.* 2009; 32:81–90. [PubMed: 19200684]
19. Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* 2008; 6:431–440. [PubMed: 18461076]
20. Fraser C, et al. The bacterial species challenge: making sense of genetic and ecological diversity. *Science.* 2009; 323:741–746. [PubMed: 19197054]
21. Wirth T, et al. The rise and spread of a new pathogen: seroresistant *Moraxella catarrhalis*. *Genome Res.* 2007; 17:1647–1656. [PubMed: 17895425]
22. den Bakker HC, et al. Lineage specific recombination rates and microevolution in *Listeria monocytogenes*. *BMC Evol. Biol.* 2008; 8:277. [PubMed: 18842152]
23. Orsi R, et al. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. *BMC Evol. Biol.* 2008; 8:233. [PubMed: 18700032]
24. Didelot X, et al. Genealogical typing of *Neisseria meningitidis*. *Microbiology.* 2009; 155:3176–3186. [PubMed: 19643763]
25. Cooper JE, Feil EJ. The phylogeny of *Staphylococcus aureus* – which genes make the best intra-species markers? *Microbiology.* 2006; 152:1297–1305. [PubMed: 16622047]
26. Sorokin A, et al. 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.* 2006; 72:1569–1578. [PubMed: 16461712]
27. Achtman M, et al. Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc. Natl. Acad. Sci. U. S. A.* 2004; 101:17837–17842. [PubMed: 15598742]

28. Van Ert M, et al. Global genetic population structure of *Bacillus anthracis*. PLoS ONE. 2007; 2:e461. [PubMed: 17520020]
29. Liu X, et al. Evidence for recombination in *Mycobacterium tuberculosis*. J. Bacteriol. 2006; 188:8169–8177. [PubMed: 16997954]
30. Wirth T, et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. 2006; 60:1136–1151. [PubMed: 16689791]
31. Didelot X, et al. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res. 2007; 17:61–68. [PubMed: 17090663]
32. Roumagnac P, et al. Evolutionary history of *Salmonella* Typhi. Science. 2006; 314:1301–1304. [PubMed: 17124322]
33. Holt K, et al. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat. Genet. 2008; 40:987–993. [PubMed: 18660809]
34. Denamur E, Matic I. Evolution of mutation rates in bacteria. Mol. Microbiol. 2006; 60:820–827. [PubMed: 16677295]
35. Kulick S, et al. Mosaic DNA imports with interspersions of recipient sequence after natural transformation of *Helicobacter pylori*. PLoS ONE. 2008; 3:e3797. [PubMed: 19030104]
36. Lin EA, et al. Natural transformation of *Helicobacter pylori* involves the integration of short DNA fragments interrupted by gaps of variable size. PLoS Pathog. 2009; 5:e1000337. [PubMed: 19282979]
37. Robinson DA, Enright MC. Evolution of *Staphylococcus aureus* by large chromosomal replacements. J. Bacteriol. 2004; 186:1060–1064. [PubMed: 14762000]
38. Brueggemann A, et al. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. PLoS Pathog. 2007; 3:e168. [PubMed: 18020702]
39. Brochet M, et al. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. Proc. Natl. Acad. Sci. U. S. A. 2008; 105:15961–15966. [PubMed: 18832470]
40. Jolley KA, et al. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol. Biol. Evol. 2005; 22:562–569. [PubMed: 15537808]
41. Fearnhead P, et al. Analysis of recombination in *Campylobacter jejuni* from MLST population data. J. Mol. Evol. 2005; 61:333–340. [PubMed: 16044246]
42. Wilson DJ, et al. Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. Mol. Biol. Evol. 2009; 26:385–397. [PubMed: 19008526]
43. Touchon M, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 2009; 5:e1000344. [PubMed: 19165319]
44. Denamur E, et al. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell. 2000; 103:711–721. [PubMed: 11114328]
45. Lefebure T, Stanhope M. Evolution of the core and pangenome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 2007; 8:R71. [PubMed: 17475002]
46. Petersen L, et al. Genes under positive selection in *Escherichia coli*. Genome Res. 2007; 17:1336–1343. [PubMed: 17675366]
47. Muzzi A, et al. Pilus operon evolution in *Streptococcus pneumoniae* is driven by positive selection and recombination. PLoS One. 2008; 3:e3660. [PubMed: 18987753]
48. Milkman R, et al. Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. Genetics. 2003; 163:475–483. [PubMed: 12618387]
49. Ragon M, et al. A new perspective on *Listeria monocytogenes* evolution. PLoS Pathog. 2008; 4:e1000146. [PubMed: 18773117]
50. Brunelle B, Sensabaugh G. The *ompA* gene in *Chlamydia trachomatis* differs in phylogeny and rate of evolution from other regions of the genome. Infect. Immun. 2006; 74:578–585. [PubMed: 16369014]

51. Gomes JP, et al. Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots. *Genome Res.* 2007; 17:50–60. [PubMed: 17090662]
52. Vos M. Why do bacteria engage in homologous recombination? *Trends Microbiol.* 2009; 17:226–232. [PubMed: 19464181]
53. Falush D, et al. Traces of human migrations in *Helicobacter pylori* populations. *Science.* 2003; 299:1582–1585. [PubMed: 12624269]
54. Wirth T, et al. Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. *Proc. Natl. Acad. Sci.* 2004; 101:4746–4751. [PubMed: 15051885]
55. Linz B, et al. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature.* 2007; 445:915–918. [PubMed: 17287725]
56. Moodley Y, et al. The peopling of the Pacific from a bacterial perspective. *Science.* 2009; 323:527–530. [PubMed: 19164753]
57. McCarthy N, et al. Host-associated genetic import in *Campylobacter jejuni*. *Emerg. Infect. Dis.* 2007; 13:267–272. [PubMed: 17479890]
58. Colles FM, et al. Comparison of *Campylobacter* populations in wild geese with those in starlings and free-range poultry on the same farm. *Appl. Environ. Microbiol.* 2008; 74:3583–3590. [PubMed: 18390684]
59. Falush D. Toward the use of genomics to study microevolutionary change in bacteria. *PLoS Genet.* 2009; 5:e1000627. [PubMed: 19855823]
60. Majewski J. Sexual isolation in bacteria. *FEMS Microbiol. Lett.* 2001; 199:161–169. [PubMed: 11377861]
61. Goss E, et al. Genetic diversity, recombination and cryptic clades in *Pseudomonas viridiflava* infecting natural populations of *Arabidopsis thaliana*. *Genetics.* 2005; 169:21–35. [PubMed: 15489535]
62. Tanabe Y, et al. Recombination, cryptic clades and neutral molecular divergence of the microcystin synthetase (*mcy*) genes of toxic cyanobacterium *Microcystis aeruginosa*. *BMC Evol. Biol.* 2009; 9:115. [PubMed: 19463155]
63. Dingle KE, et al. Sequence typing and comparison of population biology of *Campylobacter coli* and *Campylobacter jejuni*. *J. Clin. Microbiol.* 2005; 43:340–347. [PubMed: 15634992]
64. Sheppard SK, et al. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science.* 2008; 320:237–239. [PubMed: 18403712]
65. Webb K, et al. Development of an unambiguous and discriminatory multilocus sequence typing scheme for the *Streptococcus zooepidemicus* group. *Microbiology.* 2008; 154:3016–3024. [PubMed: 18832307]
66. Holden MT, et al. Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathog.* 2009; 5:e1000346. [PubMed: 19325880]
67. King SJ, et al. NanA, a neuraminidase from *Streptococcus pneumoniae*, shows high levels of sequence diversity, at least in part through recombination with *Streptococcus oralis*. *J. Bacteriol.* 2005; 187:5376–5386. [PubMed: 16030232]
68. Do T, et al. Population structure of *Streptococcus oralis*. *Microbiology.* 2009; 155:2593–2602. [PubMed: 19423627]
69. Bishop CJ, et al. Assigning strains to bacterial species via the internet. *BMC Biol.* 2009; 7:3. [PubMed: 19171050]
70. Hanage WP, et al. Hyper-recombination, diversity, and antibiotic resistance in *Pneumococcus*. *Science.* 2009; 324:1454–1457. [PubMed: 19520963]
71. Falush D, et al. Mismatch induced speciation in *Salmonella*: model and data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 2006; 361:2045–2053. [PubMed: 17062419]
72. Fraser C, et al. Recombination and the nature of bacterial speciation. *Science.* 2007; 315:476–480. [PubMed: 17255503]
73. Ahmed N. A flood of microbial genomes – do we need more? *PLoS ONE.* 2009; 4:e5831. [PubMed: 19513110]

74. Didelot X, et al. SimMLST: simulation of multi-locus sequence typing data under a neutral model. *Bioinformatics*. 2009; 25:1442–1444. [PubMed: 19286834]
75. Caugant DA, Maiden MCJ. Meningococcal carriage and disease – population biology and evolution. *Vaccine*. 2009; 27:B64–B70. [PubMed: 19464092]
76. McVean G, et al. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*. 2002; 160:1231–1241. [PubMed: 11901136]
77. Falush D, et al. Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies. *Genetics*. 2003; 164:1567–1587. [PubMed: 12930761]

Box 1. Samples of bacterial isolates

A sample of bacteria needs to be representative of a population to be informative about the underlying recombination process [15]. Unfortunately, many (indeed most) bacterial population samples are not fully representative because some strains, particularly the most virulent members of pathogen populations, attract more attention than others. For example, a disease isolate collection of *N. meningitidis* can over-represent hypervirulent lineages by two orders of magnitude compared to an asymptomatic carriage collection [75]. There are several reasons why incorrect sampling can affect estimates of recombination rates and/or patterns of gene flow. First, evolutionary models typically assume that the sample is uniformly distributed from a population, with the notable exception of the microepidemic model [12,13]. Second, a constant rate of recombination is usually assumed, which raises an issue, for example, if the population comprises two evolutionary lineages with different recombination rates. If the two lineages are represented in the sample according to their frequency in nature, the result can be interpreted as the true average recombination rate for the population, but if one lineage is over-represented in the sample then the result is meaningless. Finally, most methods of analysis can only detect recombination when descendants of both the recipient and donor strains are present in the sample. One exception is ClonalFrame [17] (Box 3), which can reconstruct imports from external sources.

Box 2. Estimation of recombination rates using summary statistics

Because of the statistical difficulty in inferring a recombination rate directly from genetic data, summary statistics of the data are often used. This has the advantage of facilitating inference, but this process loses some of the information contained in the full genetic (usually sequence) data and can result in inaccurate estimates. Linkage disequilibrium is a commonly used summary statistic, which is reduced between sites inside and outside of the imported region when recombination occurs. Patterns of linkage disequilibrium can therefore be used to estimate the prevalence of recombination in a genetic data set. The program LDhat [76], for example, is based on this principle and has been applied to bacteria for both MLST data [21,40,62] and whole genome data [43]. Another example of a summary statistic used to estimate relative recombination rates from MLST data is allelic mismatch distribution, which is the distribution of the number of allelic differences between pairs of isolates [12,13]. More complex summary statistics can be considered using the approximate Bayesian computation (ABC), which facilitates inference even when the statistical link between parameters and summary statistics is not established [42]. The only requirement for this method is the ability to quickly simulate large amounts of data, which can be achieved, for example, using SimMLST [74].

Box 3. Using the ClonalFrame software

Under a clonal evolutionary scenario, the relationships between members of a sample from a population can be represented as a tree. Conventional phylogenetic methods use a model of successive, usually point, mutations to reconstruct a tree, but these are confused by recombination. ClonalFrame [17] is an extension of this approach that also takes recombination into account (Figure I). It reconstructs a tree that represents not only the clonal genealogy of a sample as a whole, but also the genetic location of the mutation and recombination events that occurred on each branch of this genealogy. The frequency of occurrence of these reconstructed events allows the program to estimate the relative rate of recombination. ClonalFrame can be applied to either MLST data[22,56,64] or a limited number of whole genomes [16,17]. It does not model the origin of imports, which has the advantage of making the reconstruction faster and more robust to extra-population imports, at the cost of reducing the sensitivity of recombination detection, because searching for a putative origin is one of the best indications of whether recombination took place [17]. The lack of modeling of import origin also means that ClonalFrame cannot be used on its own to study gene flow among lineages. Instead, this requires processing of the output of ClonalFrame to find the most likely origin of detected imports by comparing them with other sequences [18], but this approach is not as effective as detecting imports and their origins simultaneously.

Figure I. Illustration of the ClonalFrame model.

The top part shows the clonal genealogy in black and recombination events in distinct colors. The bottom part shows the genotypes of the isolates, with fragments unaffected by recombination shown in black and fragments affected by recombination colored according to the recombination events whereby they originated.

Box 4. Using the STRUCTURE software

The linkage model of STRUCTURE [77] has been used to identify patterns of gene flow in bacteria [21,53,54]. STRUCTURE assumes the existence of a number of ancestral populations and under the linkage model, the genotype of each individual is made of blocks originating from the ancestral populations (Figure I). The output of the program therefore indicates the population that each isolate for each site is likely to have come from. In a typical output, some isolates are purely from one population whereas others are admixed. Patterns of admixture are indicative of the gene flow that occurred between or among the populations. STRUCTURE does not model the clonal genealogy underlying the population structure of a bacterial population. This implies that when two isolates share a fragment from the same ancestral population, it is not possible to say whether this is the result of shared clonal ancestry or recent recombination affecting one or the other. Consequently, STRUCTURE does not produce an estimate of the recombination rate. Not modeling the clonal genealogy is well suited for the study of highly recombinogenic species such as *H. pylori*, for which the approach was originally designed [53,77], because the clonal signal is likely to have been fully erased by recombination in such species. It might, however, be less appropriate for the analysis of bacteria with more clonal population structures [71].

Figure I. Illustration of the STRUCTURE model.

The top part shows the ancestral populations in distinct colors. The bottom part shows the genotypes of the isolates, with fragments colored according to their ancestral population of origin.

Figure 1. Effect of recombination on the discrepancy of phylogenetic signals.

Three isolates (1, 2 and 3) and two genes (A and B) are considered. In the first scenario, no recombination occurred so that the phylogenies of genes A and B are identical to the clonal genealogy and they exhibit congruence. In the second scenario, a recombination event in gene B of isolate 3 resulted in a different phylogeny for that gene and the phylogenies of genes A and B are incongruent.

Figure 2. Comparison of three inter-species studies on recombination rates.

The results of Hanage *et al.* [13], Perez-Losada *et al.* [14] and Vos and Didelot [15] are shown for the six species considered by all three studies, namely *S. aureus*, *B. cereus*, *N. meningitidis*, *H. pylori*, *S. pneumoniae* and *S. pyogenes*.