Impermanence of bacterial clones

Louis-Marie Bobay, Charles C. Traverse, and Howard Ochman

Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712

Edited by John C. Avise, University of California, Irvine, CA, and approved March 18, 2015 (received for review February 3, 2015)

Bacteria reproduce asexually and pass on a single genome copied from the parent, a reproductive mode that assures the clonal descent of progeny; however, a truly clonal bacterial species is extremely rare. The signal of clonality can be interrupted by gene uptake and exchange, initiating homologous recombination that results in the unique sequence of one clone being incorporated into another. Because recombination occurs sporadically and on local scales, these events are often difficult to recognize, even when considering large samples of completely sequenced genomes. Moreover, several processes can produce the appearance of clonality in populations that undergo frequent recombination. The rates and consequences of recombination have been studied in Escherichia coli for over 40 y, and, during this time, there have been several shifting views of its clonal status, population structure, and rates of gene exchange. We reexamine the studies and retrace the evolution of the methods that have assessed the extent of DNA flux, largely focusing on its impact on the E. coli genome.

Discovering Clonality in Natural Populations

By the 1950s, the numerous mechanisms by which bacteria could obtain new DNA sequences—conjugation (8), transformation (9), and transduction (10)—had been characterized, but the incidence of these processes and the extent of their effects on the diversification of bacterial clones remained unknown. Multilocus enzyme electrophoresis (MLEE), applied to bacteria by the 1970s (11), was able to supply quantifiable information about the forces that shape the allelic and genotypic variation in natural populations. The first large-scale population genetic survey of Escherichia coli concluded that the strain variation within hosts was generated by the “regular” occurrence of recombination and that the species as a whole was in linkage equilibrium (12). This supposition led the author (12) to construe that selective forces caused the preponderance of certain alleles and deviations from the random assortment of alleles over loci.

A more refined analysis incorporating several additional loci revealed that E. coli was, in fact, essentially clonal, with recombination rates perhaps on the order of that of mutation rates (13). In that study, evidence indicating that E. coli had a clonal population structure came from several sources. First, despite extensive allelic diversity at each of the 20 loci assayed, only a small number of genotypes were recovered, reflecting the infrequent reassortment of alleles. Second, strains of the same (or very similar) multilocus genotype were present in unrelated and geographically distant hosts (and, in one case, an infant from Massachusetts harbored a strain identical to the laboratory type specimen E. coli K-12, originally isolated in California in 1922), attesting to the long-term stability and wide geographic distribution of individual clones. Additionally, single locus variants (SLVs; strains identical at all but one locus) usually differed by the presence of a unique allele, suggesting that these polymorphisms arose by mutation rather than by recombination. Expanded studies on E. coli from diverse sets of hosts reported almost the same results (14, 15), and, around the same time, Ørskov and Ørskov devised the “clone concept” for E. coli pathogens to explain their findings that certain serotypic combinations were recovered repeatedly from temporally and geographically unassociated hosts (16). Although serological classification is based on cell surface factors whose variation could result from selection caused by the interactions between bacteria and hosts, the concordance between the population structures defined by serotyping and by MLEE provided strong support for the view that E. coli is predominantly clonal (17).

Reproduction by binary fission virtually guarantees the clonality of a bacterial lineage. Apart from mutations and other rare events that might modify chromosome integrity during replication, the primary sequence of DNA in all daughter and descendent cells remains identical, generation after generation. Unlike animals, in which parthenogenetic forms are ecologically constrained and relatively short-lived over evolutionary timescales (1–3), asexually reproducing bacteria have persisted since the origin of cellular life and represent the most diverse and widespread organisms on the planet. Naturally, the vast diversity present in bacteria could have arisen solely by asexual means—there has certainly been sufficient time and large enough population sizes to allow for enormous numbers of mutations (and combinations of mutations) to be experienced. Moreover, it seems as though some of the most extraordinary innovations in the history of life have occurred without intervention of the sexual process (4).

Bacteria as Clonal Organisms

Despite their obligatory asexual mode of reproduction, the clonality of bacterial lineages can be disrupted by sex, or at least by what we refer to as sex. In bacteria, sex is the inheritance of genetic material from any source aside from their own parent cell and can occur by any of several processes. Foreign DNA can be introduced by cell-to-cell contact, transmitted to the cell by an infectious agent, or acquired directly from the environment; and, therefore, genes can be obtained from organisms representing any domain of life, and even from entities (i.e., viruses and phages) that are not classified to any domain of life. Moreover, events of sex in bacteria occur without known regularity and usually constitute a very small portion of the genome. In fact, sexually acquired DNA need not involve recombination at all but can persist as a heritable extrachromosomal element—yielding a situation where the genome has changed but clonality is preserved. Taken together, sex in bacteria shares few features with those normally associated with sex in eukaryotes: it is simply the uptake of any genetic material that might eventually be vertically or horizontally transmitted (5–7).

www.pnas.org/cgi/doi/10.1073/pnas.1501724112

PNAS | July 21, 2015 | vol. 112 | no. 29 | 8893–8900

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, “In the Light of Evolution IX: Clonal Reproduction: Alternatives to Sex,” held January 9–10, 2015, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/IX_Clonal_Reproduction.

Author contributions: L.-M.B., C.C.T., and H.O. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. Email: howard.ochman@austin.utexas.edu.
**Just How Clonal Are Bacteria?**

Broad application of MLEE illuminated the clonal nature of the bacterial populations (18, 19). Discovering that the majority of species displayed a nonrandom association of alleles among loci (linkage disequilibrium) led to the view that rates of recombination are typically low in bacteria isolated from their natural habitats—but, unfortunately, there is trouble with this interpretation. Whereas the occurrence of linkage equilibrium can be attributed to recombination, the converse need not be true: i.e., linkage disequilibrium is not always indicative of a lack of recombination. Numerous factors, quite apart from the lack of recombination, can cause linkage disequilibrium, and the possibility that the clonality observed in most bacterial populations derives from sources other than the lack of recombination was brought to the forefront in a PNAS paper by Maynard Smith et al. (20) that asked (and was entitled, as is this subsection): “How clonal are bacteria?”

Some bacterial species may be truly clonal: i.e., they experience no recombination. However, several circumstances will give the appearance of clonality, even in species that undergo regular bouts of recombination. Scenarios particularly relevant to bacteria in which this circumstance is evident are when recombination proceeds within genetically or geographically isolated subpopulations and when there has been epidemic expansion (or periodic selection) of a particular strain. In these cases, sampling a mixture of strains from multiple subpopulations, or only the progeny of the ephemeral epidemic strain, will both yield evidence of strong disequilibrium, and the challenge is to distinguish such cases from true clonality. Maynard Smith et al. (20) tried to differentiate these cases both by partitioning the samples into subgroups to determine how the observed extent of recombination changed and by confining analyses to individual genotypes (as opposed to the entire sample, which may contain multiple isolates of the epidemic clone). Despite its vagaries, reanalysis of MLEE data using their approach (20) yielded species that were completely clonal (e.g., *Salmonella enterica*) and others that were panmictic (e.g., *Neisseria gonorrhoeae*), as well as some with intermediate population structures. Surprisingly, they did not apply their methods to *E. coli*, the bacterial species for which the most comprehensive MLEE data were then available (~5,000 isolates); however, its population structure was thought to most closely resemble that of its sister group *Salmonella*.

**Entering the Sequencing Era**

MLEE, by assaying allelic variants in a handful of loci scattered around the genome, is limited to the detection of rather large-scale events of recombination, typically those involving regions much larger than a bacterial gene, leaving events occurring on a much smaller scale unnoticed. The advent of sequence-based analyses remedied this situation by offering resolution of allelic variation at the level of the individual nucleotide. Once it became possible to generate nucleotide sequences for homologs in multiple isolates, the question became a matter of how best to proceed within genetically or geographically isolated subpopulations. However, the sequence information generated for *E. coli*, Milkman and Bridges (21) enumerated 10 recombinational replacement events in regions estimated to have incurred 50 nucleotide substitutions, yielding a rate of retained replacements to retained substitutions of 0.2 (and again implying that *E. coli* is largely clonal and that its genetic diversity was generated chiefly by mutations). Despite these results, it was argued that simply enumerating events of recombination and mutation does not convey their actual contributions because recombination can change multiple nucleotides—so, in order for the processes to be compared directly, both need to be expressed on a per-nucleotide-site basis. Examining the branching orders for four genes in 12 strains of *E. coli*, Guttmann and Dykhuizen (22) detected inconsistencies attributable to three recombinational events within a group of strains that had no polymorphisms attributable to mutation, resulting in a rate of nucleotide change by recombination that was an estimated 50 times that of mutation. Note, however, that they found that the phylogenies for all genes were largely congruent, such that the few recombinational events, although affecting a disproportionate number of nucleotides, did not disrupt the clonal nature of *E. coli* populations.

**Moving from MLEE to MLST**

Nucleotide sequencing methods added new dimensions to analysis of bacterial populations and led to the widespread use of a multilocus sequence typing (MLST) approach, in which six or seven gene fragments (of lengths suitable for Sanger sequencing) were PCR-amplified and sequenced for each bacterial strain (23–25). MLST is, in many ways, an extension of MLEE, in that it indexes the allelic variation at multiple housekeeping genes in each strain. Naturally, MLST had advantages over MLEE, the most prominent of which was its high level of resolution, its reproducibility, and its portability, allowing any researchers to generate data that could be easily processed and compared across laboratories.

Similar to MLEE, most applications of MLST assign a unique number to each allelic variant (regardless of its number of nucleotide differences from a nonidentical allele), and each strain is designated by its multilocus genotype: i.e., its allelic profile across loci. However, the sequence information generated for MLST proved extremely useful for examining the role of mutation and recombination in the divergence of bacterial lineages (26–28). Focusing on SLVs (i.e., allelic profiles that differed at only one locus), Feil et al. (29) tabulated those in which the allelic variants differed at single sites, indicating an SLV generated by mutation, or at multiple sites, taken as evidence of an SLV generated by recombination. (Actually, their complementary analysis based on homoplasy revealed that perhaps half of allelic variants differing at a single site also arose through recombination.) Their calculations of *riw* (the ratio of substitutions introduced by recombination relative to mutation) for *Streptococcus pneumoniae* and *Neisseria meningitidis* ranged from 50 to 100, on the order of what Guttmann and Dykhuizen (22) estimated in *E. coli*.

Current practice is to use *r* and *m* to denote per-site rates of recombination and mutation, and *ρ* and *θ* to denote events of recombination and mutation, respectively; however, these notations have been applied somewhat indiscriminately and their values derived by disparate methods, often hindering comparisons across studies. Vos and Didelot (30) revisited the MLST datasets...
for scores of bacterial taxa and recalculated \( r \) and \( m \) in a single framework, thereby allowing direct comparisons of the extent of recombination in generating the clonal divergence within species. The \( r/m \) values ranged over three orders of magnitude, and there was no clear association between recombination rates and bacterial lifestyle or phylogenetic division. Additionally, there were several cases where the values that they obtained were clearly at odds with previous studies; for example, they found \textit{S. enterica}—the most clonal species based on MLEE—to have among the highest \( r/m \) ratios, even higher than that of \textit{Helicobacter pylori}, which is essentially panmictic. Contrarily, \( r/m \) of \textit{E. coli} was only 0.7, substantially lower than some previous estimates. Such discrepancies are likely due to the methods used to identify recombinant sites, the specific datasets that were analyzed, and the effects of sampling on recognition of recombination.

The population structure of \textit{E. coli} was viewed as largely clonal because recombination was either limited to particular genes and to particular groups of strains. A broad MLST survey involving 100s of \textit{E. coli} strains looked at the incidence of recombination within the well-established subgroups (clades) that were originally defined by MLEE (31). Although the mutation rates were similar for all seven genes across all subgroups, recombination rates differed substantially. Moreover, that study found a link between recombination and virulence, such that subgroups comprising pathogenic strains of \textit{E. coli} displayed increased rates of recombination.

**Clonality in the Genomic Era**

Even if recombination occurs infrequently and affects small regions of the chromosome, the clonal status of the lineage will erode, making it difficult to establish the degree of clonality without sequences of entire genomes. Complete genome sequences now offer the opportunity to decipher the impact of recombination on bacterial evolution; but, admittedly, comparing sets of whole genomes is much more computationally challenging than analyzing the sequences from a few MLST loci and still suffers from many of the same biases. Although many of the same analytical problems arise when examining any set of sequences, the advantages of using full genome sequences are that they show the full scale of recombination events occurring through the genome, that they are better for defining recombinant segments and that they can reveal how recombination might be related to certain functional features of genes or structural features of genomes.

The first comprehensive analysis of recombinational events occurring throughout the \textit{E. coli} genome, conducted by Mau et al. (32), considered the complete sequences of six strains and used phylogenetic and clustering methods to identify recombinant segments within regions that were conserved in all strains. (32). Although they inferred one long (~100-kb) stretch of the chromosome that underwent a recombination event in these strains, they reported that the typical length of recombinant segments was only about 1 kb in length, which was much shorter than that reported in studies based in more limited portions of the genome; and furthermore, they estimated that the extent of recombination was higher than previous estimates. The short size of recombinant fragments suggested that recombination occurred primarily by events of gene conversion rather than crossing-over, as is typical in eukaryotes, and by transduction and conjugation, which usually involve much larger pieces of DNA. Shorter segments of DNA could result from the partial degradation of longer sequences or could directly enter the cell through transformation, but \textit{E. coli} is not naturally transformable, and its occurrence has been reported only under specific conditions (33, 34).

A second study on \textit{E. coli} (35) focused on a diverse set of 20 complete genomes and used population-genetics approaches (36, 37) to detect recombinant fragments. In this analysis, the length of recombinant segments was much shorter than previous estimates (only 50 bp) although the relative impact of recombination and mutation on the introduction of nucleotide polymorphism was very close to that estimated with MLST data (\( r/m \approx 0.9 \)) (30). The study (35) also asked how the effects of recombination differed along the chromosome and identified several (and confirmed some) recombination hotspots, most notably, two centering on the \textit{rfb} and the \textit{fmn} operons (38, 39). These two loci are involved in O-antigen synthesis (\textit{rfb}) and adhesion to host cells (\textit{fmn}), and, because these two cellular features are exposed to phages, protists, or the host immune system, they are thought to evolve quickly by diversifying selection (40).

Aside from these hotspots, smoother fluctuations of the recombination rate are apparent over broader scales. Chromosome scanning revealed a decrease in the recombination rate in the ∼1-Mb region surrounding the replication terminus (35). Several hypotheses have been proposed to account for this transition in recombination rate along the chromosome, including: (i) a replication-associated dosage effect, which leads to a higher copy number and increased recombination rate (due to this increased availability of homologous strands) proximate to the replication origin; (ii) a higher mutation rate nearer to the terminus, resulting in an effectively lower value \( r/m \) ratio (41); and (iii) the macrodomain structure of the \textit{E. coli} chromosome, in which the broad region spanning the replication terminus is the most tightly packed and has a reduced ability to recombine due to physical constraints (42). (An alternate hypothesis, combining features of i and ii posits that the homogenizing effect of recombination serves to reduce the rate of evolution of conserved housekeeping genes, which are disproportionately located near the replication origin.) In fact, each of the hypotheses that attempt to account for the variation in \( r/m \) values along the chromosome remain blurred by the tight association of mutation, selection, and recombination; therefore, caution is needed when interpreting this metric.

A more recent study involving 27 complete \textit{E. coli} genomes used a Bayesian approach, implemented in ClonalFrame (43), to detect recombination events (44). Again, the \( r/m \) ratio was near unity; however, recombination tracts were estimated to be an order of magnitude longer than the previous based on many of the same genomes (542 bp vs. 50 bp), but still shorter than original estimates of the size of recombinant regions. That study (44) defined a third hotspot around the \textit{aroC} gene, which could be involved in host interactions and virulence.

These analyses, all based on full genome sequences, estimated similar recombination rates for \textit{E. coli}, confirming previous observations that, on average, recombination introduces as many nucleotide substitutions as mutations. Despite rather frequent recombination, this amount of DNA flux does not blur the signal of vertical descent for genes conserved among all strains (i.e., the “core genome”) (35). Unfortunately, the delineation of recombination breakpoints is still imprecise and highly dependent on the particular method and the dataset used to recognize recombination events. In all cases, similar sets of genes were overly affected by recombination, particularly fast-evolving loci that encoded proteins that were exposed to the environment, involved in stress response, or considered virulence factors. And, in fact, the recombination hotspots in \textit{Pseudomonas aeruginosa} have been recently shown to include similar gene categories (45).

**How to Best Assess the Impact of Recombination on \textit{E. coli} Evolution**

The large variety of methods that have been developed to detect recombination (46) reflects the fact that there are numerous technical and conceptual difficulties associated with the identification of the specific tracts of DNA that have been involved in gene exchange. As might be expected, the power and accuracy of these algorithms are maximized when a donor sequence is included (imparting the source of homology between unrelated
lineages) and when the recombinant sequence introduces many polymorphic nucleotides (43, 46). Therefore, homoplasies—characters that are inferred to be shared by, but not present in, the common ancestor of lineages—represent robust signals of recombination and provide a very fine (i.e., per nucleotide site) resolution of recombination maps, as have been performed recently for sequenced strains of *Staphylococcus aureus* (47). Homoplastic sites allow detection of internal recombination events (i.e., recombinant polymorphic sites that are included in the dataset) but ignore polymorphic sites that were introduced by external, unsampled sources. Unsampled polymorphism can be introduced by closely related lineages (that acquired new mutations and would go undetected because they mimic vertical inheritance) or by divergent unsampled lineages. Although approaches based on homoplasies could miss the latter cases of recombination—virtually all approaches overlook the former—the increasing number of sequenced genomes and the long history of MLEE and MLST analyses suggest that current sampling of *E. coli* genomes is ample. However, it remains possible that several new major lineages have yet to be discovered (48, 49).

To obtain the recombinational landscape of the *E. coli* genome (Fig. 1), we focused our analysis on the detection of homoplasies.

![Recombinational landscape of E. coli. Genomic features and recombination parameters were calculated for the entire core genome of concatenated orthologs shared by 19 fully sequenced strains of E. coli. (A) Proportion of homoplasic sites, inferred by a distance metric that uses fixed branch lengths (see How to Best Assess the Impact of Recombination on E. coli Evolution) in nonoverlapping 10-kb windows spanning the core genome. (B) Proportion of homoplasic sites, inferred by a phylogenetic topology method (implemented in recHMM) (54), in nonoverlapping 10-kb windows spanning the core genome. (C) Proportion of the total variation (polymorphism) present in nonoverlapping 10-kb windows (gray) that is attributable to homoplasic sites (red). (D) Relative impact of recombination ($\pi_r$) and mutation ($\pi_m$) on the polymorphism in nonoverlapping 10-kb windows spanning the core genome based on the $\pi_r/m$ ($\pi_r/\pi_m$) statistic. (Polymorphisms identified with the distance metric, as in A.) (E) Detecting recombination within core fragments (orthologous gene contigs) spanning the core genome with PHI (10,000 permutations; $P < 0.001$) (57). Because this method detects recombination on a gene-by-gene (not site-by-site) scale, results are presented for 50-kb sliding windows with 10-kb steps. (F) G+C contents in nonoverlapping 10-kb windows spanning the core genome.](image-url)
among the genes common to a subset of the complete genomes [www.ncbi.nlm.nih.gov/genome (June 2014 version)]. The complete set of orthologous genes was defined by best bidirectional hit with USEARCH (global) with 80% identity (50) among the annotated genes present in the 65 sequenced strains. Adjacent core genes were merged into 1,131 core fragments that conserved gene order and intergenic regions, aligned with MAFFT v7 (51), and concatenated into a single large contig. The maximum likelihood distances $D$ between the core gene sets were computed with TREE-PUZZLE (52), allowing us to select a subset of 19 genomes in a manner that maximized the nucleotide diversity of their core genomes and to retain only a single representative of nearly identical strains.

Homoplases appearing in the concatenation of core fragments (∼3.2 Mb, representing 3,022 genes) were detected by two methods: (i) Distance method: Based on the distance $D$ between genomes, a polymorphic site composed of two alleles $N_0$ and $N_1$ was considered homoplasic when max ($D_{N_0N_1}$) > min ($D_{N_0N_1}$), with $N_0$ and $N_1$ the major and minor alleles, respectively. (Triallelic sites were subjected to the same pairwise comparison, but tetraallelic sites ($n = 357$) were not considered.); and (ii) Topology method: A maximum likelihood tree was inferred from the concatenation of core fragments for all 19 genomes using RAxML v7.2 (53), and polymorphic sites were considered homoplasic when incongruent with the topology of the core genome tree using recHMM (54, 55). Both the distance and the topology methods yielded similar results (Fig. 1 A and B) and detected 116,288 and 112,400 homoplases, respectively, of which 110,143 were common to both methods. We confined subsequent analyses to homoplase sites identified by the distance method because it is faster and less sensitive to unsolved internal branches.

Homoplases arise from recombination but can also result from mutations that occur independently in the lineages in question. Fortunately, the two processes can often be distinguished because a single recombination event is likely to introduce multiple homoplases that display the same incongruent pattern (i.e., clusters of polymorphic sites that have the same distribution among lineages). To establish whether homoplases arose from recombination or from convergent mutations, we looked for the signatures of congruent homoplases in 1-kb windows across the entire concatenation. Almost half (46%) of the homoplasic sites have a nearby (within 500-bp) homoplasic site displaying the same distribution among strains, suggesting that they were introduced in the same recombination event, not by convergent mutations. By simulating the accumulation of the current polymorphism in the E. coli genome, and assuming that it was introduced exclusively by random mutations, we estimate that only 2.4% of polymorphic sites would be homoplasic due to independent mutations, indicating that convergent mutations have a negligible contribution relative to recombination in the introduction of homoplases.

Using homoplasic sites, we mapped the incidence of recombination across the entire E. coli genome. This method could possibly underestimate the frequency of recombination events by failing to recognize recombinant segments from unsampled lineages; however, it clearly illustrates the extent to which recombination varies in its effects across the chromosome. These results confirm the existence of the recombination hotspot near the fim gene and suggest the presence of several additional recombination hotspots. The previously identified $rfb$ hotspot is situated in a region of high polymorphism (Fig. 1C); however, the relative frequencies of homoplasic to nonhomoplasic polymorphisms is not particularly high (Fig. 1D), which suggests that this hotspot might in fact be attributable to diversifying selection as opposed to recombination. Our analysis, based solely on homoplasic sites, also recognized the region surrounding the replication terminus as having a reduced rate of recombination. Overall, the relative frequency of homoplasic and nonhomoplasic polymorphisms suggests a near equal contribution of recombination and mutation in the introduction of polymorphic sites ($r/m = 0.92, SD ± 0.26$), which is consistent with previous analyses based on full genome data (Table 1).

Aside from contributing to the variation of individual genes, recombination also seems to affect how the chromosome itself evolves. At the terminus of replication, the lower recombination rate coincides with a reduction in the G+C content (35), as is observed in other species (56) (Fig. 1F). This effect becomes even more noticeable when detecting recombination at larger scales, as with the computational method PHI (pairwise homoplasy index) (Fig. 1E) (57). In that mutations are universally biased toward A and T (58, 59) and recombination influences the distribution of A and T-biased), which suggests that this hotspot might in fact be attributable to diversifying selection as opposed to recombination. Our analysis, based solely on homoplasic sites, also recognized the region surrounding the replication terminus as having a reduced rate of recombination. Overall, the relative frequency of homoplasic and nonhomoplasic polymorphisms suggests a near equal contribution of recombination and mutation in the introduction of polymorphic sites ($r/m = 0.92, SD ± 0.26$), which is consistent with previous analyses based on full genome data (Table 1).

Aside from contributing to the variation of individual genes, recombination also seems to affect how the chromosome itself evolves. At the terminus of replication, the lower recombination rate coincides with a reduction in the G+C content (35), as is observed in other species (56) (Fig. 1F). This effect becomes even more noticeable when detecting recombination at larger scales, as with the computational method PHI (pairwise homoplasy index) (Fig. 1E) (57). In that mutations are universally biased toward A and T (58, 59) and recombination influences the distribution of A and T-biased), which suggests that this hotspot might in fact be attributable to diversifying selection as opposed to recombination. Our analysis, based solely on homoplasic sites, also recognized the region surrounding the replication terminus as having a reduced rate of recombination. Overall, the relative frequency of homoplasic and nonhomoplasic polymorphisms suggests a near equal contribution of recombination and mutation in the introduction of polymorphic sites ($r/m = 0.92, SD ± 0.26$), which is consistent with previous analyses based on full genome data (Table 1).

Aside from contributing to the variation of individual genes, recombination also seems to affect how the chromosome itself evolves. At the terminus of replication, the lower recombination rate coincides with a reduction in the G+C content (35), as is observed in other species (56) (Fig. 1F). This effect becomes even more noticeable when detecting recombination at larger scales, as with the computational method PHI (pairwise homoplasy index) (Fig. 1E) (57). In that mutations are universally biased toward A and T (58, 59) and recombination influences the distribution of A and T-biased), which suggests that this hotspot might in fact be attributable to diversifying selection as opposed to recombination. Our analysis, based solely on homoplasic sites, also recognized the region surrounding the replication terminus as having a reduced rate of recombination. Overall, the relative frequency of homoplasic and nonhomoplasic polymorphisms suggests a near equal contribution of recombination and mutation in the introduction of polymorphic sites ($r/m = 0.92, SD ± 0.26$), which is consistent with previous analyses based on full genome data (Table 1).
nonsynonymous sites near the terminus (35). Moreover, there is additional evidence that selection serves to elevate genomic G+C contents in bacteria (61, 62). Alternatively, a lower recombination rate near the replication terminus could lower the G+C content of the region by minimizing the G+C-biased repair of recombination-induced mismatches by biased gene conversion (63).

**Beyond the Core Genome**

Most genome-wide analyses of recombination have been limited to the regions constituting the core genome, but this approach ignores the accessory genes—those that are not ubiquitous among strains—and their neighboring intergenic regions. Such regions are just as prone to recombination events; however, their sporadic distributions make their identification and analysis somewhat more difficult. There are many classes of accessory genes, such as mobile elements (e.g., prophages, transposons), which are known to be associated with elevated rates of recombination. In both *E. coli* and *S. aureus*, it was recently shown that core genes in the vicinity of accessory genes or mobile elements experience higher recombination rates (44, 47). Chromosome loci with the highest homologous recombination rates (recombination hotspots) have also been associated with nonmobilizable genomic islands in *E. coli* (e.g., the *fim* locus). These heightened rates of recombination could be due to selection—elements can encode adaptive traits that confer an advantage to the acquisition (64)—and the absence of site-specific integrases or transposons within many of these elements suggests that many rely on recombination to propagate in the population. Additionally, many recombination hotspots in *E. coli* seem to be evolving under diversifying selection, supporting a general role of homologous exchange in spreading both beneficial alleles and beneficial accessory genes (35).

The ability of recombination to spread beneficial alleles (and purge deleterious alleles) has been known for some time (65); however, its effect on the dynamics of bacterial genes and genomes remains obscure. Studies on *Vibrio cyclitrophicus* and *Burkholderia pseudomallei* both suggest that genes, rather than genomes, reach fixation into the population (66, 67), but these species undergo much higher recombination rates than *E. coli* (30). The population structure of *E. coli*, in which certain genotypes dominate the population, would indicate that periodic selection (selective sweeps) lead to occasional epidemic structures in *E. coli* and other species that experience local or low rates of recombination.

**Genomic Determinants of Bacterial Clonality**

What determines whether a bacterial population is clonal or panmictic? Several genomic features have been linked to the ability of bacteria to modulate the amount of DNA uptake and exchange within and between populations.

Firstly, recombination efficiency is connected to the extent of sequence identity. *mutS* mutants of *E. coli* demonstrate lower levels of sexual isolation, suggesting that mismatch repair plays a central role in the frequency of recombination (68). Recombination initiation requires minimal substrate lengths of 23–27 identical nucleotides, termed "minimal efficient processing segments" (MEPS) (69). The frequency of MEPS decreases exponentially with sequence divergence, suggesting that the clonal or panmictic status of a species depends on its level of polymorphism and its population structure. Moreover, this requirement would imply that more divergent strains display lower frequencies of DNA exchange, compatible with clonal evolution, whereas closely related strains recombine more frequently. As highlighted previously (in *Just How Clonal Are Bacteria?*), frequent recombination, when confined to close relatives, would produce populations that possess all of the hallmarks of clonality, making it difficult to determine the actual clonal status of the species.

Secondly, several additional barriers to DNA acquisition and exchange occur in bacteria (70); and among them, restriction-modification (R-M) systems vary substantially among species and strains (71). By selectively degrading incoming DNA according to their sequence and methylation patterns, these systems can influence the range and extent of DNA exchange between cells and populations, and a recent study highlighted the role of R-M systems in regulating sequence exchange within *B. pseudomallei* (67).

Third, the mobile element repertoires, which can be highly variable among strains (72, 73), are likely to determine the capacity for DNA transfer by mediating transduction and conjugation, and by providing templates for homologous exchange. Additionally, mobile elements integrated into the *E. coli* genome sometimes encode enzymes catalyzing homologous exchange (74, 75): For example, the defective prophage *rac* encodes the RecT recombinase, which can supplement recombination functions in RecBCD mutants (76), and is typically more promiscuous than the RecBCD pathway (77, 78). Moreover, there is wide variation among *E. coli* strains in the repertoires of complete or partially degraded prophages, implying that strains can acquire and lose recombinase genes depending on their particular set of mobile elements. This dynamic reservoir of ready-to-use recombination enzymes might serve to promote fluctuations in recombination rates within and among lineages.

Finally, there can be counterselection against recombination in some genomes arising from the epistatic interactions among alleles at different loci (79, 80). In this scenario, genes whose products are involved in multiprotein complexes or depend on specific protein–protein interactions would sustain fewer nonsynonymous substitutions introduced by recombination (analogous to barriers to gene exchange proposed in the "complexity hypothesis" (81), in which highly interacting proteins are not susceptible to horizontal acquisition).

**Conclusions**

The perception of the clonal status of *E. coli* has changed substantially over the past four decades, mostly due to the maturation of allele-typing methods, to increased awareness of effects of strain and locus sampling, and to the development of new recombination-detection algorithms. Genome sequencing should overcome many of the remaining obstacles to evaluating recombination occurring in bacterial populations, and recombination-detection methods have been improved to provide better resolution of sequence history and population structures. Given the abundance of lineages that have been sampled in *E. coli* and other species, the use of homoplasies as signals of recombination should help to increase the application and accuracy of these methods. When considering *E. coli*, several studies have converged on the view that recombination contributes almost as much as mutations to the introduction of nucleotide polymorphism. However, despite its occurrence, the rate and scale of recombination does not destroy the clonal status of *E. coli*. Estimations of the size of recombinant segments in *E. coli* are imprecise but mostly range from 50–500 bp, indicating that DNA exchange acts primarily by gene conversion. Moreover, recombination rates are highly variable among genes and along the chromosome, and it remains difficult to tease apart the relative contributions of mutations, recombination, and selection to the observed variation.
7. Levin BR, Cornejo OE (2009) The population and evolutionary dynamics of homo- 
geneous gene recombination in bacterial populations. PLoS Genet 5(8):e1000601.
8. Lederberg J, Tatum EL (1946) Gene recombination in Escherichia coli. Nature 158(4016):558.
9. Avery OT, Macleod CM, McCarty M (1944) Studies on the chemical nature of the 
substrate inducing transformation of pneumococcal types: Induction of toxin formation 
by a deoxyribonuclease fraction isolated from pneumococcus type III. J Exp Med 79(2):137–158.
10. Zinder ND, Lederberg J (1952) Genetic exchange in Salmonella. J Bacteriol 64(3): 
679–689.
11. Baptist JN, Shaw CR, Mandel M (1971) Comparative zone electrophoresis of enzymes 
of Pseudomonas solanacearum and Pseudomonas cepacia. J Bacteriol 102(8):799–803.
12. Millman R (1973) Electrophoretic variation in Escherichia coli from natural sources. 
Science 184(4149):1204–1206.
13. Selander RK, Levin BR (1980) Genetic diversity and structure in Escherichia coli pop-
ulations. Science 210(4469):545–547.
14. Whittam TS, Ochman H, Selander RK (1983) Multilocus genetic structure in natural 
populations of Escherichia coli. Proc Natl Acad Sci USA 80(6):1751–1755.
15. Caugant DA, Levin BR, Selander RK (1984) Distribution of multilocus genotypes of 
Escherichia coli within and between host families. J Hyg (Lond) 92(3):377–384.
16. Ørskov F, Ørskov I (1983) From the National Institutes of Health: Summary of a 
workshop on the clone concept in the epidemiology, taxonomy, and evolution of the 
enterobacteriaceae and other bacteria. J Infect Dis 148(2):346–357.
17. Achtman M, et al. (1986) Clonal analysis of Escherichia coli O2K1 isolated from dis-
 eased humans and animals. Infect Immun 51(1):268–276.
18. Selander RK, et al. (1986) Methods of multilocus enzyme electrophoresis for bacterial 
population genetics and systems. Ann Appl Microbiol 51(3):873–884.
19. Selander RK, Musser JM, Caugant DA, Gilmour MN, Whittam TS (1987) Population 
genetics of pathogenic bacteria. Microb Pathog 3(1):1–17.
20. Smith JM, Smith NH, O’Toole GA, Spratt BG (1993) How clonal are bacteria? Proc 
Acad Sci USA 90(10):4384–4388.
21. Millman R, Bridges MM (1990) Molecular evolution of the Escherichia coli chromo-
some. III. Clonal frames. Genetics 126(3):505–517.
22. Gutman DS, Dykhuizen DE (1994) Clonal divergence in Escherichia coli as a result of 
recombination, not mutation. Science 265(5189):1380–1383.
23. Maiden MC, 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(9):3410–3415.
24. Feil EJ, Spratt BG (2001) Recombination and the population structures of bacterial 
pathogens. Annu Rev Microbiol 55:561–590.
25. Maiden MC, et al. (2003) MLST revisited: The gene-by-gene approach to bacterial 
genomics. Nat Rev Microbiol 11(10):728–736.
26. Spratt BG, Harris D (2001) The relative contributions of recombination and point 
mutation to the diversification of bacterial clones. Curr Opin Microbiol 4(5): 
602–606.
27. Hanage WP, Fraser C, Spratt BG (2006) Sequences, sequence clusters and bacterial 
plasmid evolution. Trends Microb 14(7):359–362.
28. Didelet X, Maiden MC (2010) Impact of recombination on bacterial evolution. Trends 
Microb 18(7):315–322.
29. Feil EJ, Smith JM, Enright MC, Spratt BG (2000) Estimating recombinational param-
eters in Streptococcus pneumoniae from multilocus sequence typing data. Genetics 154(4): 
1439–1450.
30. Vos M, Didelet X (2009) A comparison of homologous recombination rates in bacteria 
and archaea. ISME J 3(2):199–208.
31. Wirth T, et al. (2006) Sex and virulence in Escherichia coli: an evolutionary perspec-
tive. Mol Microbiol 60(5):1136–1151.
32. Mau B, Glaser JD, Darling AE, Perna NT (2006) Genome-wide detection and analysis 
of homologous recombination among sequenced strains of Escherichia coli. Genome 
Biol 7(4):R84.
33. Baur B, Hanselmann K, Schlimme W, Jenni B (1996) Genetic transformation in 
freshwater: Escherichia coli is able to develop natural competence. Appl Environ 
Microbiol 62(10):3673–3678.
34. Tsen SD, et al. (2002) Natural transformation of Escherichia coli. J Biomed Sci 9(3): 
246–252.
35. Tousson M, et al. (2009) Organised genome dynamics in the Escherichia coli species 
results in highly diverse adaptive paths. PLoS Genet 5(1):e1000344.
36. Padhukasahasram B, Wall JD, Marinam P, Nordborg M (2006) Estimating re-
combination rates from single-nucleotide polymorphisms using summary statistics. 
Genetics 174(3):1517–1528.
37. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in 
population genetics. Genetics 162(4):2025–2035.
38. Milinkov R, Jaeger E, McBride RD (2003) Molecular evolution of the Escherichia coli 
chromosome. VI. Two regions of high effective recombination. Genetics 163(2): 
475–483.
39. Weinreich SJ, et al. (2006) Clonal analysis reveals high rate of structural mutations in 
fimbrial adhesins of extraintestinal pathogenic Escherichia coli. Mol Microbiol 59(3): 
975–988.
40. Seed KD, et al. (2014) Evolutionary consequences of intra-host phage predation on 
plasmid populations. eLife 3:e02745.
41. Mira A, Ochman H (2002) Gene location and bacterial sequence divergence. Mol Biol 
Evol 19(8):1350–1358.
42. Valens M, Penaud S, Rossignol M, Cornet F, Boccard F (2004) Macromolecular organi-
zation of the Escherichia coli chromosome. EMBO J 23(12):4320–4341.
43. Didelet X, Falush D (2007) Inference of bacterial microevolution using multilocus 
sequence data. Genetics 175(3):1251–1266.
44. Didelet X, Méric G, Falush D, Darling AE (2012) Impact of homologous and non-ho-
moLOGous recombination in the genomic evolution of Escherichia coli. BMC Genomics 13:256.
80. Friedman J, Alm EJ, Shapiro BJ (2013) Sympatric speciation: When is it possible in bacteria? PLoS ONE 8(1):e53539.

81. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci USA 96(7):3801–3806.

82. Dykhuizen DE, Green L (1991) Recombination in Escherichia coli and the definition of biological species. J Bacteriol 173(22):7257–7268.

83. Milkman R, Bridges MM (1993) Molecular evolution of the Escherichia coli chromosome. IV. Sequence comparisons. Genetics 133(3):455–468.

84. Whittam TS, Ake SE (1993) Genetic polymorphisms and recombination in natural populations of Escherichia coli. Mechanisms of Molecular Evolution, eds Takahata N, Clark AG (Sinauer, Sunderland, MA), pp 223–246.

85. Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS (2000) Parallel evolution of virulence in pathogenic Escherichia coli. Nature 406(6791):64–67.

86. Pérez-Losada M, et al. (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol 6(2):97–112.