Recombinant transfer in the basic genome of *Escherichia coli*

Purushottam D. Dixit[1,2], Tin Yau Pang[1,3], F. William Studier[4], and Sergei Maslov[4,5]

Biological, Environmental and Climate Sciences Department, Brookhaven National Laboratory, Upton, NY 11973

Contributed by F. William Studier, June 3, 2015 (sent for review February 20, 2015; reviewed by Richard E. Lenski and Lise Raleigh)

An approximation to the ∼4-Mbp basic genome shared by 32 strains of *Escherichia coli* representing six evolutionary groups has been derived and analyzed computationally. A multiple alignment of the 32 complete genome sequences was filtered to remove mobile elements and identify the most reliable ∼90% of the aligned length of each of the resulting 496 basic-genome pairs. Patterns of single base-pair mutations (SNPs) in aligned pairs distinguish clonally inherited regions from regions where either genome has acquired DNA fragments from diverged genomes by homologous recombination since their last common ancestor. Such recombinant transfer is pervasive across the basic genome, mostly between genomes in the same evolutionary group, and generates many unique mosaic patterns. The six least-diverged genome pairs have one or two recombinant transfers of length ∼40–115 kbp (and few if any other transfers), each containing one or more gene clusters known to confer strong selective advantage in some environments. Moderately diverged genome pairs (0.4–1% SNPs) show mosaic patterns of interspersed clonal and recombinant regions of varying lengths throughout the basic genome, whereas more highly diverged pairs within an evolutionary group or pairs between evolutionary groups having >1.3% SNPs have few clonal matches longer than a few kilobase pairs. Many recombinant transfers appear to incorporate fragments of the entering DNA produced by restriction systems of the recipient cell. A simple computational model can closely fit the data. Most recombinant transfers seem likely to be due to generalized transduction by coevolving populations of phages, which could efficiently distribute variability throughout bacterial genomes.

E. coli evolution | basic genome | core genome | recombinant transfer | generalized transduction

The increasing availability of complete genome sequences of many different bacterial and archaeal species, as well as metagenomic sequencing of mixed populations from natural environments, has stimulated theoretical and computational approaches to understand mechanisms of speciation and how prokaryotic species should be defined (1–8). Much genome analysis and comparison has been at the level of gene content, identifying core genomes (the set of genes found in most or all genomes in a group) and the continually expanding pan-genome. Population genomics of *Escherichia coli* has been particularly well studied because of its long history in laboratory research and because many pathogenic strains have been isolated and completely sequenced (9–14). Proposed models of how related groups or species form and evolve include isolation by ecological niche (7–9, 11, 15), decreased homologous recombination as divergence between isolated populations increases (2–4, 8, 14, 16), and coevolving phage and bacterial populations (6).

E. coli genomes are highly variable, containing an array of phage-related mobile elements integrated at many different sites (17), random insertions of multiple transposable elements (18), and idiosyncratic genome rearrangements that include inversions, translocations, duplications, and deletions. Although *E. coli* grows by binary cell division, genetic exchange by homologous recombination has come to be recognized as a significant factor in adaptation and genome evolution (9, 10, 19). Of particular interest has been the relative contribution to genome variability of random mutations (single base-pair differences referred to as SNPs) and replacement of genome regions by homologous recombination with fragments imported from other genomes (here referred to as recombinant transfers or transferred regions). Estimates of the rate, extent, and average lengths of recombinant transfers in the core genome vary widely, as do methods for detecting transferred regions and assessing their impact on phylogenetic relationships (12–14, 20, 21).

In a previous comparison of complete genome sequences of the K-12 reference strain MG1655 and the reconstructed genome of the B strain of Delbrück and Luria referred to here as B-DL, we observed that SNPs are not randomly distributed among 3,620 perfectly matched pairs of coding sequences but rather have two distinct regimes: sharply decreasing numbers of genes having 0, 1, 2, or 3 SNPs, and an abrupt transition to a much broader exponential distribution in which decreasing numbers of genes contain increasing numbers of SNPs from 4 to 102 SNPs per gene (22). Genes in the two regimes of the distribution are interspersed in clusters of variable lengths throughout what we referred to as the basic genome, namely, the ∼4 Mbp shared by the two genomes after eliminating mobile elements. We speculated that genes having 0 to 3 SNPs may primarily have been inherited clonally from the last common ancestor, whereas genes comprising the exponential tail may primarily have been acquired by horizontal transfer from diverged members of the population.

Significance

A significant fraction of the length of *Escherichia coli* genomes comprises mobile elements integrated at various sites in a ∼4-Mbp basic genome shared by the species. We find that the entire basic genome is continually exchanged by homologous recombination with genome fragments acquired from other genomes in the population. Evolutionary groups appear to exchange DNA preferentially within the same group but also with other groups to different extents. Entering DNA is often fragmented by restriction systems of the recipient cell, with surviving pieces replacing homologous parts of the recipient chromosome. Coevolving populations of phages that package genome fragments and deliver them to cells that have appropriate receptors are likely mediators of most DNA transfers, distributing variability throughout the species.

Author contributions: F.W.S. and S.M. designed research; P.D.D., T.Y.P., F.W.S., and S.M. performed research; P.D.D., T.Y.P., F.W.S., and S.M. contributed new reagents/analytic tools; P.D.D., T.Y.P., F.W.S., and S.M. analyzed data; and P.D.D., F.W.S., and S.M. wrote the paper.

Reviewers: R.E.L., Michigan State University; and L.R., New England Biolabs.

The authors declare no conflict of interest.

[1]P.D.D. and T.Y.P. contributed equally to this work.

[2]Present address: Department of Systems Biology, Columbia University, New York, NY 10032.

[3]Present address: Institute for Bioinformatics, Heinrich-Heine-Universität Düsseldorf, 40221 Düsseldorf, Germany.

[4]To whom correspondence may be addressed. Email: studier@bnl.gov or ssmaslov@gmail.com.

[5]Present address: Department of Bioengineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1510839112/-/DCSupplemental.

The current study was undertaken to extend these observations to a diverse set of 32 completely sequenced *E. coli* genomes and to analyze how SNP distributions in the basic genome change as a function of evolutionary divergence between the 496 pairs of strains in this set. We have taken a simpler approach than those of Touchon et al. (13), Didelot et al. (14), and McNally et al. (21), who previously analyzed multiple alignments of complete genomes of *E. coli* strains. The appreciably larger basic genome derived here is not restricted to protein-coding sequences and retains positional information.

Results and Discussion

Deriving Basic Genomes. We selected for analysis the completely sequenced chromosomal genomes of 32 independently isolated *E. coli* strains from six previously defined evolutionary groups: A, B1, B2, D1, D2, and E (Fig. 1). Whole-genome multiple alignment of all 32 genomes was produced by the Mauve program (23). Computational filters and procedures described in *Materials and Methods* generated a 3,955,192-bp alignment that has eliminated essentially all mobile elements and approximates the basic genomes of these 32 strains. The organization of 21 of these 32 basic genomes is that of the comprehensively annotated K-12 laboratory strain MG1655. The remaining 11 genomes contained idiosyncratic inversions and translocations, most of which were reconfigured manually to align with the consensus organization in the multiple alignment (*SI Materials and Methods*).

Reliable Genomewide SNP Densities. The filtered multiple alignment contains 105 ordered alignment blocks that were arbitrarily divided into tandem strings of 1-kbp segments, starting at the left end of each block. This process generated 3,903 segments of 1 kbp in 88 strings from 1 to 370 segments long, separated by 105 shorter right-end segments covering ~1.3% of the multiple alignment. The segments are numbered 1–4,008 from left to right. Segment ends are indexed to both the basic- and complete-genome sequences of each strain to allow easy comparisons and link to annotations. SNPs between each of the 496 pairs of basic genomes and cumulatively across all 32 strains are extracted directly from the filtered multiple alignment.

SNPs are accurately identified in regions where all 32 basic genomes are unambiguously aligned, but highly variable regions, particularly where alignment lengths differ, can be problematic. To minimize erroneous SNPs due to such multiple-alignment difficulties, we focused most of our analyses on a set of 3,769 segments of 1 kbp that have an approximately normal distribution over cumulative SNP densities of 0.3–18.0% (averaging 7.5%) and cover 95.3% of the filtered 32-genome multiple alignment (Fig. S1). The 134 most-diverged 1-kbp segments in the scattered tail of the distribution extend to 60.2% cumulative SNP density and are primarily in known regions of high variability and subject to known selective pressures, including genes for making O-antigens, lipopolysaccharides, flagella, fimbrial-adhesins, DNA modification and restriction enzymes, and surface receptors for phages and colicins. The most variable parts of such exchangeable regions did not pass the computational filters and are not present in the aligned basic genomes analyzed here.

Even in regions where cumulative SNP densities are in the normal range, alignment problems involving group-specific or individual deletions, misplaced remnants of insertion sequence (IS) elements, variable numbers of repeats, or other idiosyncrasies can generate false SNPs in some genome pairs. To minimize such problems, most of our analyses were limited to perfectly aligned 1-kbp segments having no indels. Perfectly aligned segments in the set of 3,769 usually cover ~90% of an aligned basic-genome pair. This additional filter reduces average SNP density by 5–15% (but as much as 30% between closely related genomes with few total SNPs).

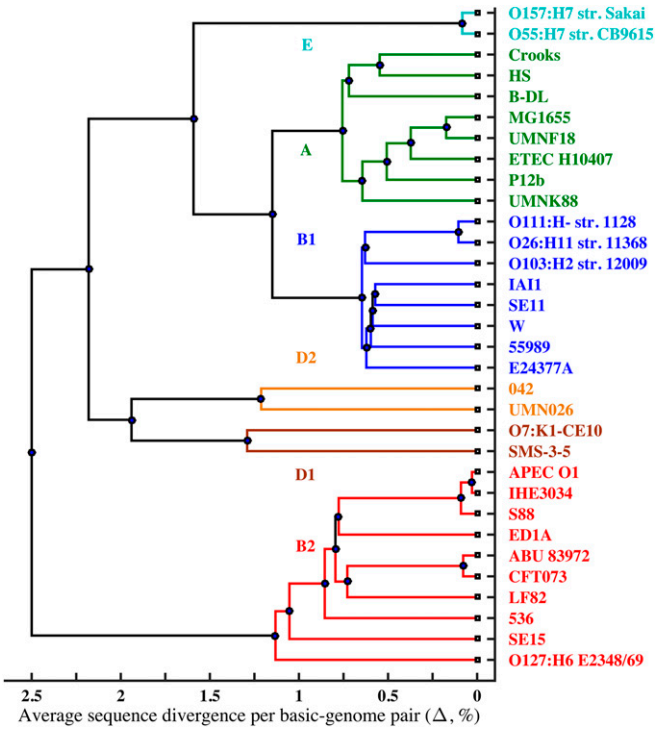

Fig. 1. Phylogenetic tree derived from filtered genome-wide average SNP densities (Δ) between 496 pairs of 32 basic genomes. Previously recognized phylogenetic groups: E (light blue), A (green), B1 (blue), D2 (yellow), D1 (brown), and B2 (red). The tree was calculated using UPGMA algorithm. Dots in the lines connecting pairs of evolutionary groups are placed approximately at average SNP densities between them. The groups are ordered to fit the relative divergences among them summarized in Table S1. GenBank accession numbers are given in Table S3.

SNP Distributions in Genome Pairs. Our measure of evolutionary divergence is SNP density, the average percentage of SNPs between perfectly aligned 1-kbp segments in the set of 3,769, referred to here as Δ for an entire basic-genome pair and δ for individual segments. Distributions of SNP densities among individual segments are shown in Fig. 2*A* for five basic-genome pairs over the range of Δ = 0.38–2.54%. The alignments are between the K-12 reference genome MG1655 and two other group A genomes, and between MG1655 and one genome each from groups B1, E, and B2. The two group A pairs show a sharply decreasing number of segments with increasing numbers of SNPs from 0 to 3 per 1-kbp segment (0–0.3% SNP density), which we refer to as the clonal peak on the assumption that most of the segments in that peak are likely to have been clonally inherited from a common ancestor. Consistent with our previous observations using matched protein-coding sequences between MG1655 and B-DL (22), the more highly diverged segments are distributed in a roughly exponential tail extending from the clonal peak.

A clonal peak is apparent only when both of the paired genomes are in the same evolutionary group, but not all genome pairs within a group show a clonal peak: the 28 pairs between the 8 genomes of group A, the 28 pairs between the 8 genomes of group B1, and the single pairs in D2 and E all show at least a small peak; however, only 4 of the 45 pairs between the 10 genomes of B2 show a clonal peak, and neither the single pair in D1 nor any of the 392 pairs between genomes from different groups show a pronounced clonal peak (Dataset S1). As the clonal peak decreases, the increasing number of segments in the exponential tail maintain approximately the same slope, and the most highly diverged genome pairs have a broad maximum around δ = 1–2% (Fig. 2*A*). Our computational model of genome divergence, summarized in a later section and detailed in *Supporting Information*, fits the observed distributions quite well over a broad range of Δ (Fig. 2*B*).

EVOLUTION

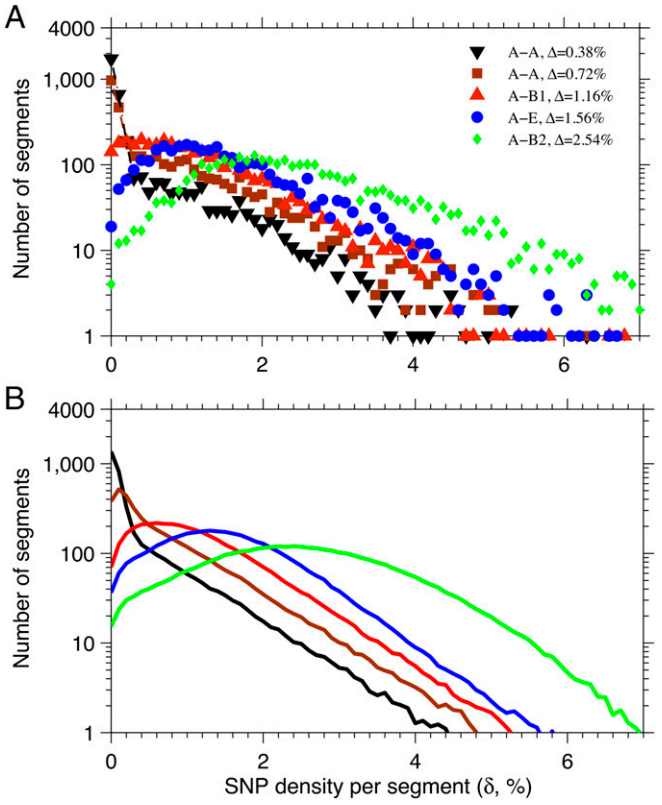

Fig. 2. Distributions of SNP densities between basic genomes. (*A*) Distribution of perfectly aligned 1-kbp segments from the set of 3,769 as a function of average SNP density δ for five basic-genome pairs: group A strain MG1655 aligned with ETEC (A–A, $\Delta = 0.38\%$, black triangles); B–DL (A–A, $\Delta = 0.72\%$, brown squares); SE11 (A–B1, $\Delta = 1.16\%$, red triangles); O157 (A–E, $\Delta = 1.56\%$, blue circles); and IHE (A–B2, $\Delta = 2.54\%$, green diamonds). (*B*) Distributions of SNP density as a function of δ as predicted by the computational model given in *Supporting Information* for pairs of genomes having the same average SNP densities Δ as the five genome pairs in *A*.

Recombinant Transfers and Mosaic Genomes. For simplicity, we refer to all perfectly aligned 1-kbp segments having 0 to 3 SNPs as clonal, because distinguishing whether such segments were inherited vertically from a common ancestor or represent incidental matches to mosaic regions acquired by recombinant transfer is not always unambiguous. We also refer to segments having more than three SNPs as transferred, meaning that the SNPs were acquired at least in part by recombinant transfer from a diverged genome, even though we recognize that some isolated segments containing four to six SNPs are likely to be clonal. Using these designations, we calculated four quantities for each of the 496 basic-genome pairs: Δ, the average SNP density for all perfectly aligned 1-kbp segments from the set of 3,769; f_c, the fraction of these segments containing 0 to 3 SNPs, referred to as the clonal fraction; Δ_c, the average SNP density in the clonal fraction; and Δ_t, the average SNP density in the remaining, putatively transferred segments (Dataset S1). The values of f_c, Δ_c, and Δ_t are plotted as a function of Δ for all 496 genome pairs in Fig. 3 *A*–*C* and summarized for all combinations between evolutionary groups in Table S1.

These data show that recombinant transfers are pervasive throughout the basic genome. The 104 genome pairs in which both genomes are from any one of the six evolutionary groups have clonal fractions that decrease approximately linearly from >0.90 to ∼0.15 as Δ increases from <0.1% to ∼1.3% (Fig. 3*A*). Over the same range, Δ_c increases from <0.02% to ∼0.2% (Fig. 3*B*). Clearly, most of the divergence between basic-genome pairs within an evolutionary group is due to accumulating recombinant transfers from diverged genomes since their last common ancestor. Assuming random recombinant transfers, a recombining population will generate over time a steady-state population of mosaic genomes with many uniquely different patterns of interspersed clonal and recombinant regions. Examples of mosaic patterns between genome pairs of different divergence can be visualized in Dataset S2 by scrolling through the spreadsheet, which has color coding and information at the top and bottom to help locate specific features.

Clonal Fraction (f_c) and Average SNP Density in Transferred Segments (Δ_t) Within and Between Groups. As clonal fraction decreases with accumulating recombinant transfers, Δ_t remains relatively constant, averaging 1.39%, 1.18%, and 1.28% in groups A, B1, and B2 in relatively narrow distributions with slight overlaps (except for considerable scatter in the least-diverged pairs, which have few recombinant transfers) (Fig. 3*C*, Table S1, and Dataset S1). The average Δ_t within a group should reflect the average divergence in the recombining population, and the relatively narrow distributions and slight overlaps suggest that each group may be exchanging genome fragments by recombinant transfers primarily within its own recombining population, but occasionally with genomes in other recombining groups. The D1, D2, and E groups are represented by only one genome pair each. The D1 pair ($f_c = 0.20$) and D2 pair ($f_c = 0.28$) each have more than 2,000 recombinant 1-kbp segments, and their Δ_t of 1.58% and 1.66% may be approximately representative of the average divergences in their recombining populations. However, the group E pair ($f_c = 0.98$) has fewer than 100 recombinant segments and its Δ_t may be far from representative.

The only intergroup genome pairs with significant clonal fractions are the 64 pairings between group A and B1 genomes (f_c = 0.27–0.18 and Δ = 1.03–1.19%), similar to the 17 most diverged genome pairs within the B2 group (those involving SE15 or O127:H6 have f_c = 0.22–0.15 and Δ = 1.03–1.22%). The average of Δ_t over the 64 A–B1 genome pairs is 1.41% ± 0.02 (SD) and that of the 17 most diverged B2–B2 pairs is 1.33% ± 0.06. The appreciable clonal fractions in A–B1 genome pairs suggest that the two groups diverged relatively recently.

The 328 intergroup genome pairs in the other 14 of the 15 possible pairings between the six evolutionary groups assort into sets of 4–80 genome pairs per combination (Table S1 and Dataset S1). The average values of Δ_t of all genome pairs in any single set range from 1.7% to 2.6% and each combination has a very narrow distribution (SD usually less than 0.02). These narrow distributions support the interpretation that the different mosaic genomes in each recombining population are well equilibrated to the average diversity in the population throughout their lengths. The clonal fraction is negligible in all 14 of these sets of intergroup genome pairs, decreasing from 0.07 to 0.01 as Δ increases from ∼1.6% to 2.6%. As a consequence, Δ_t is approximately the same as Δ throughout this range (Fig. 3*C*). The few 1-kbp segments that appear clonal in the most diverged intergroup genome pairs are usually highly conserved across the 32 genomes (Dataset S2, set 4) and thus would appear to be clonal whether or not they had been transferred. The longest of these highly conserved regions contains the cluster of 26 ribosomal protein genes *rplQ* to *rpsJ* in 13 tandem 1-kbp segments.

Six Slightly Diverged Genome Pairs Reveal an Important Mode of Recombinant Transfer. The six least-diverged pairs of basic genomes, all of which have a clonal fraction >0.95, have a striking pattern of recombinant transfers. Instead of mosaic transferred regions of various lengths dispersed throughout the basic genome in moderately diverged genome pairs, each genome pair has only one or two long transferred regions, each region extending across 42–107 kbp of basic-genome sequence and together containing the vast majority of SNPs attributable to recombinant transfer (Table S2). End points of these transferred regions are even farther apart in the complete-genome sequences, ∼85–240 kbp, due mostly to mobile elements, which may either have been in the transferred fragment or inserted after acquisition.

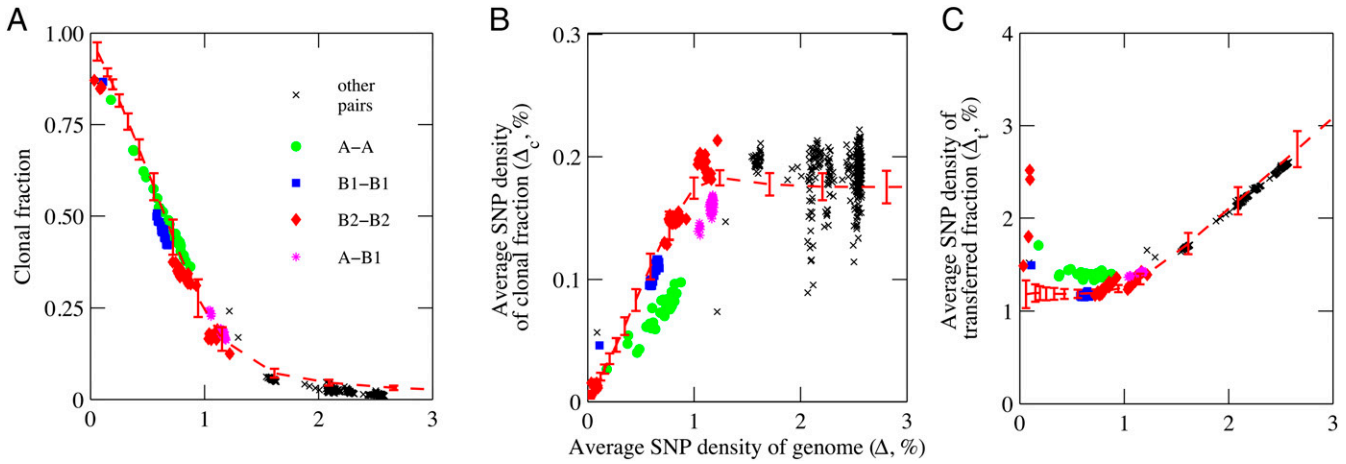

Fig. 3. Values of clonal fraction, Δ_c, and Δ_t as a function of overall divergence Δ in 496 basic-genome pairs. (*A*) Clonal fraction f_c; (*B*) average SNP density in the clonal fraction Δ_c; (*C*) average SNP density in the transferred fraction Δ_t. Data points for genome pairs within evolutionary groups A are given by solid green circles; B1 by blue squares; B2 by red diamonds; between A and B1 by violet asterisks; and other pairs by a black X. Dashed red lines with error bars are values predicted by the computational model given in *Supporting Information*. Error bars correspond to SD of 100 runs of the model.

These long transferred regions replace the mosaic pattern of the recipient genome region with the mosaic pattern of the homologous region of the donor genome. The number of incidental clonal matches between them and the percentage of each transferred region occupied by clonal matches should decrease with increasing average SNP density in the transferred region. Indeed, the least diverged of the 10 transferred regions in Table S2 (1.37% SNP density) has 15 clonal matches of one to four segments covering 23% of the transferred length; the next (1.67% SNP density) has 11 clonal matches constituting 11% of the transferred length; the six transferred regions in the range of 2.23–3.57% SNP density have only one or two clonal matches constituting 2–7% of the transferred length; the transferred region in S88 with 4.05% SNP density has no clonal matches; and the transferred region with $\Delta = 5.27\%$ SNP density has a single clonal segment constituting 1% of its length. The transferred region in the B2 strain S88 came from a group A genome, as evidenced by tandem strings of clonal segments in this region when the S88 genome is paired with different group A genomes (Dataset S2, set 4).

Initially, it appeared that short recombinant transfers were present elsewhere in the six genome pairs containing long transfers. However, examination of the candidate SNP clusters provided other explanations for almost all of them. The most interesting proved to be the result of shuffling of variants in ribosomal RNA operons by internal recombination (gene conversion) among the seven operons characteristic of *E. coli* (Dataset S2). These internally generated SNPs in rRNA operons can be a significant fraction of putative recombinant SNPs in the six least-diverged genome pairs. Most of the other high-density SNP clusters can be attributed to multiple-alignment difficulties. Internally generated and erroneous SNP clusters become a negligible fraction of transferred SNPs in even moderately diverged genome pairs, and corrections for them were not made in the figures and SI datasets. Three isolated clusters of 28–38% SNPs in 88–197 bp in APEC but not in any of the other 31 genomes are the most likely candidates in the six least-diverged genome pairs to have resulted from short recombinant transfers (Dataset S2, set 2).

Each of the long transferred regions contains gene clusters known to have exchangeable variants that provide a strong selective advantage in some situations. These variable gene clusters can nonetheless be exchanged by homologous recombination because the sequences flanking them are much less variable. The O-antigen and DNA restriction genes were previously identified as having many variants that are frequently exchanged and efficiently retained in *E. coli* populations because they can confer significant selective advantage (24). An O-antigen gene cluster was transferred into at least one of the genomes in all six pairs, the DNA restriction cluster into two genomes, and a gene cluster for capsule formation in one (Table S2). It seems likely that each of these long recombinant transfers was the initial step in divergence of a new lineage in a population under stress, and that the selective advantage it conferred fixed not only the advantageous gene cluster but also the unique mosaic pattern of the recipient genome as the ancestral sequence of the new lineage.

The last of these selective transfers in each of the nine different genomes in the least-diverged genome pairs was apparently recent enough that no subsequent relatively neutral recombinant transfers have become fixed in the population since the last common ancestor (with the possible exception of the three short candidates in APEC). Rough estimates of the number of generations since the mosaic patterns in the >95% of basic-genome length in these six genome pairs were fixed can be made by dividing their corrected numbers of SNPs per base pair by twice the estimated mutation rate of 8.9×10^{-11} mutations per base pair per generation (25) (arbitrarily assigning one-half of the SNPs to each genome in a pair). This calculation gives estimates of $\sim 8 \times 10^5$ generations for the five B2 genomes and $\sim 3 \times 10^6$ for the two B1 and two E genomes.

DNA Restriction Has a Prominent Role in Recombinant Transfer. Many types and specificities of DNA restriction and modification are widely distributed in *E. coli* and about one-half of completely sequenced *E. coli* genomes contain one or more of the many variants of *hsd* genes, which specify type I restriction/modification systems, and often other restriction enzymes as well (26, 27). Milkman and colleagues (28, 29) used restriction fragment length polymorphism to show that genome fragments introduced into *E. coli* by transduction or conjugation are fragmented and reduced in length by restriction systems, a process they postulated is responsible for generating the mosaic structure of *E. coli* genomes. More recently, we analyzed three completely sequenced genomes to deduce the patterns of recombinant transfer by P1 transduction of genomic DNA from the K-12 strain W3110 across a type I restriction barrier into two different B genomes (22). P1 is a generalized transducing phage that delivers genome fragments as long as ~115 kbp without accompanying phage DNA (30). One of these recombinant transfers delivered a single fragment of 6.0–10.6 kbp, but the second delivered six DNA fragments totaling 44.9–55.1 kbp across 71.3–77.0 kbp of genome. The average length of the six transferred W3110 fragments is ~8.3 kbp and that of the five clonal B intervals is ~4.8 kbp, but the lengths of individual fragments have very wide possible ranges, 0.3–26.5 kbp for

transferred fragments and 0.1–13.9 kbp for clonal intervals. We examined nucleotide sequences of all 32 genomes in the primary locus of genes specifying type I and other types of restrictions enzymes, referred to as the immigration control region (31). Eighteen of the 32 genomes appear to have the intact *hsd* genes needed for type I restriction, with several different types represented (details in *Supporting Information*).

Lengths of Interspersed Clonal and Recombinant Regions. As random recombinant transfers accumulate in genomes that have a recent common ancestor, average lengths of clonal regions decrease rapidly with the corresponding increases in Δ (Fig. 4*A* and Dataset S3). The seven least-diverged genome pairs (adding the group A pair MG–F18) all have clonal fractions >0.90, uninterrupted strings of clonal segments hundreds of kilobase pairs long, and one, two, or possibly five recombinant transfer events consistent with entering DNA longer than 40 kbp (Table S2 and Datasets S2 and S3). The next levels of divergence are seen in a set of six group A genome pairs having clonal fractions between 0.76 and 0.62 (Δ = 0.38–0.59%), five of which have 20–50% of their aligned lengths in uninterrupted strings of clonal segments longer than 50 kbp (Dataset S3). Including all 4,008 segments in the analyses (and allowing clonal regions to extend through the high-SNP segments in rRNA operons, segments containing obviously erroneous SNPs, and rare isolated segments containing four to six SNPs per kilobase pair), these moderately diverged genome pairs have as many as seven clonal regions longer than 100 kbp, the longest covering 967 kbp of aligned length (Dataset S2, set 3).

These long clonal regions are distributed across the paired genomes at intervals consistent with clonal inheritance from a recent common ancestor interrupted by random recombinant transfers of indeterminate numbers and lengths from genomes with different mosaic patterns. However, the P12b genome uniquely has long clonal matches with two different genomes: five clonal regions of 184–715 kbp are evident when P12b is aligned with MG, and a single 756-kbp clonal region when aligned with Crooks (in a region that has a divergent mosaic pattern relative to MG; Dataset S2, set 3). The long clonal region between P12b and Crooks must almost certainly be due to an uninterrupted recombinant exchange of ∼20% of the basic genome of P12b with DNA delivered by conjugation from a close relative of Crooks.

At higher divergence, average lengths of clonal matches between genomes within an evolutionary group decrease to 1.5–3 kbp, and the longest clonal regions in moderately to highly diverged pairs within a group are typically less than 1% of the aligned basic-genome length (Fig. 4*A* and Dataset S3). The lengths and distributions of clonal and transferred regions reflect the uniquely different mosaic patterns generated by random recombinant transfers in the two lineages, and incidental clonal matches in transferred regions should decrease with increasing distance from the last common ancestor of donor and recipient. Fragmentation of the entering DNA in a significant fraction of transfers can also reduce average lengths of both clonal and transferred regions. Average lengths of transferred regions are ∼2.6–4.6 kbp in the range of Δ = 0.38–1.0% before increasing at higher divergence as newly acquired recombinant transfers overlap those acquired previously (Fig. 4*B*). The average number of interspersed regions reaches a maximum around 500 in group A genome pairs and around 600 in B1 and B2 genome pairs before decreasing somewhat in the B2 pairs having Δ > 1.0% due to overlaps (Fig. 4*C*).

Computational Model of Divergence. We developed a simple computational model of divergence in a steady-state population of genomes assumed to comprise 4,000 segments of 1 kbp and to be accumulating random point mutations and acquiring random tandem segments of variable lengths by recombinant transfer from other genomes in the population at fixed rates. Analytical derivation of the probabilities of possible outcomes, combined with Markov chain simulations, generated probable SNP distributions across 100 pairs of genomes evolving to any given Δ. The model (details in *Supporting Information*) fits the data quite well (Figs. 2*C*, 3, and 4) using a mutation rate $\mu = 8.9 \times 10^{-11}$ SNPs per base pair per generation (25), an average length of recombinant transfer of 3 kbp (Fig. 4*B* and Dataset S3), and a ratio of rate of accumulation of SNPs by recombinant transfer relative to random mutations, ρ/μ, of 0.31, slightly higher than the 0.14–0.21 calculated in Dataset S1 for group A, B1, and B2 genome pairs. The value of r/m, a measure of the ratio of the number of mutations in transferred regions relative to random mutations throughout the genome, is 11.2 when calculated from the parameters used in the model, slightly higher than the 8.9 for group A and 5.4 for groups B1 and B2 in Dataset S1 (overall average = 6.5; SD, 0.2). These values for r/m may be compared with the original estimate of 50 by Guttman and Dykhuizen (20), from limited data, and 0.34 by McNally et al. (21), 1.5 by Touchon et al. (13), and 7 by Didelot et al. (14) from complete-genome sequences.

Coevolving Transducing Phages as Primary Vectors of Transfer. How are the continuing, pervasive, and primarily group-specific transfers of genome fragments of 40–115 kbp (or even larger) accomplished routinely in recombining populations of *E. coli*? Of the three well-studied mechanisms of DNA transfer (conjugation, transformation, and transduction), generalized transduction seems to us likely to be the primary mode of routine transfer. Generalized transducing phages such as coliphage P1 (30) are widely distributed and can have packaging capacity as high as 300 kbp of DNA (32–34).

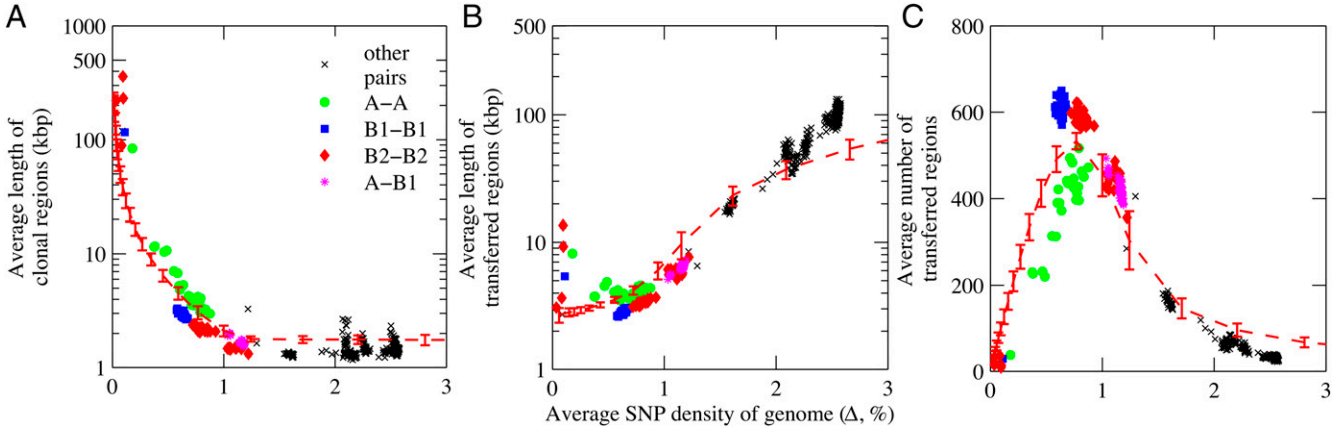

Fig. 4. Average lengths of clonal and transferred regions, and numbers of each as a function of overall divergence Δ in 496 basic-genome pairs. (*A*) Average lengths of clonal regions (in kilobase pairs); (*B*) average lengths of transferred regions (in kilobase pairs); (*C*) number of transferred regions (equal numbers of interspersed clonal and transferred regions). Data points, dashed red lines, and error bars are as in Fig. 3.

Coevolving populations of phages and bacteria (6) would maintain specificity for delivery of genome fragments primarily to members of the recombining population of bacteria, but host-range variability could provide occasional delivery of genome fragments to cells of other populations. Transducing phage particles potentially deliver random, well-protected genome fragments throughout a coevolving population much more widely and efficiently than conjugation or conjugative plasmids, which require specialized transfer mechanisms and cell-to-cell contact. That is not to argue that conjugation or transfer of conjugative plasmids does not occur or is not important, and we did detect one obvious example of conjugation in the P12b genome. However, phages are ideally suited to mediate the pervasive and continuing recombinant transfer documented by our analyses, and coevolution of phage and bacterial populations provides a simple explanation for a large body of previous work on genome variability, recombinant transfer, and speciation.

Basic Genome: A Platform for Annotation. Further development of the basic-genome platform and computational methodology developed here could simplify and facilitate classification and annotation of the current and anticipated flood of complete, draft, and metagenome sequences of *E. coli*. Consensus basic-genome sequences of all 32 strains and of the group A, B1, and B2 strains are given as FASTA files in Datasets S4–S7. Alignment to these consensus sequences can help to classify newly sequenced *E. coli* genomes, identify orthologs, and distinguish types and locations of mobile elements. Standardized basic-genome annotations and catalogs of group-specific features, exchangeable gene clusters, and other features could accelerate and improve uniformity and reliability of annotation. The methodology should be applicable to any bacterial or archaeal species or evolutionary group.

Materials and Methods

Extracting Basic Genomes. We used the Mauve program with default parameters (23) to produce a multiple alignment of the complete genome sequences of 32 independently isolated strains of *E. coli*. Their names, GenBank accession numbers, complete genome lengths, and derived basic-genome lengths are given in Table S3, and an annotated complete genome sequence of B–DL is given as Dataset S8. Sequences of 11 genomes were reconfigured to simplify the multiple alignment (*SI Materials and Methods*).

A key step in deriving basic genomes from the multiple alignment was to apply a simple computational filter designed to eliminate mobile elements, idiosyncratic insertions or duplications, and highly diverged regions that do not align well. The filter applied in the present analysis removed every base pair position in which fewer than 22 of the 32 genomes have an aligned base pair, which retained idiosyncratic and group-specific deletions. This filter reduced the initial 3,044 Mauve alignment blocks to 105 ordered blocks and the total aligned sequence within these blocks from 20.5 Mbp to the 3,955,192-bp approximation to the basic genome analyzed here. We determined that all mobile elements (prophages, *rhs* elements, IS elements) annotated in the extensively analyzed complete-genome sequences of MG1655 and B–DL (22) were removed by the filter, but at least a few other genomes retained remnants due to difficulties in multiple alignment at sites where mobile elements are integrated. Our procedures appear to have captured a reasonable approximation to the basic genome shared by these 32 strains and provide a platform for further analysis and refinement.

Consensus Sequences. Majority rule at each position where at least 22 genomes are represented generated a consensus basic-genome sequence for all 32 genomes. At least six genomes were required at each position in the group A and B1 consensus sequences, and at least seven genomes in B2.

ACKNOWLEDGMENTS. This work was supported by Grants PM-031 and ELS165 from the Office of Biological and Environmental Research of the US Department of Energy and internal research funding from Brookhaven National Laboratory.

1. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.
2. Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 3(9):679–687.
3. Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science* 315(5811):476–480.
4. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: Making sense of genetic and ecological diversity. *Science* 323(5915):741–746.
5. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. *Trends Genet* 25(3):107–110.
6. Rodriguez-Valera F, et al. (2009) Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* 7(11):828–836.
7. Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* 35(5):957–976.
8. Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet* 29(3):170–175.
9. Tenaillon O, Skurnik D, Picard B, Denamur E (2010) The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 8(3):207–217.
10. Milkman R (1997) Recombination and population structure in *Escherichia coli*. *Genetics* 146(3):745–750.
11. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 19(12):2226–2238.
12. Mau B, Glasner JD, Darling AE, Perna NT (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol* 7(5):R44.
13. Touchon M, et al. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5(1):e1000344.
14. Didelot X, Méric G, Falush D, Darling AE (2012) Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* 13:256.
15. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad Sci USA* 105(23):7899–7906.
16. Vulić M, Lenski RE, Radman M (1999) Mutation, recombination, and incipient speciation of bacteria in the laboratory. *Proc Natl Acad Sci USA* 96(13):7348–7351.
17. Bobay LM, Touchon M, Rocha EP (2014) Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci USA* 111(33):12127–12132.
18. Mahillon J, Chandler M (1998) Insertion sequences. *Microbiol Mol Biol Rev* 62(3):725–774.
19. Dykhuizen DE, Green L (1991) Recombination in *Escherichia coli* and the definition of biological species. *J Bacteriol* 173(22):7257–7268.
20. Guttman DS, Dykhuizen DE (1994) Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266(5189):1380–1383.
21. McNally A, Cheng L, Harris SR, Corander J (2013) The evolutionary path to extra-intestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. *Genome Biol Evol* 5(4):699–710.
22. Studier FW, Daegelen P, Lenski RE, Maslov S, Kim JF (2009) Understanding the differences between genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3) and comparison of the *E. coli* B and K-12 genomes. *J Mol Biol* 394(4):653–680.
23. Darling AE, Mau B, Perna NT (2010) progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5(6):e11147.
24. Milkman 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.
25. Wielgoss S, et al. (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* 1(3):183–186.
26. Loenen WAM, Dryden DTF, Raleigh EA, Wilson GG (2014) Type I restriction enzymes and their relatives. *Nucleic Acids Res* 42(1):20–44.
27. Loenen WAM, Dryden DTF, Raleigh EA, Wilson GG, Murray NE (2014) Highlights of the DNA cutters: A short history of the restriction enzymes. *Nucleic Acids Res* 42(1):3–19.
28. McKane M, Milkman R (1995) Transduction, restriction and recombination patterns in *Escherichia coli*. *Genetics* 139(1):35–43.
29. Milkman R, et al. (1999) Molecular evolution of the *Escherichia coli* chromosome. V. Recombination patterns among strains of diverse origin. *Genetics* 153(2):539–554.
30. Sternberg NL, Maurer R (1991) Bacteriophage-mediated generalized transduction in *Escherichia coli* and *Salmonella typhimurium*. *Methods Enzymol* 204:18–43.
31. Sibley MH, Raleigh EA (2004) Cassette-like variation of restriction enzyme genes in *Escherichia coli* C and relatives. *Nucleic Acids Res* 32(2):522–534.
32. Muniesa M, Imamovic L, Jofre J (2011) Bacteriophages and genetic mobilization in sewage and faecally polluted environments. *Microb Biotechnol* 4(6):725–734.
33. Battaglioli EJ, et al. (2011) Isolation of generalized transducing bacteriophages for uropathogenic strains of *Escherichia coli*. *Appl Environ Microbiol* 77(18):6630–6635.
34. Petty NK, et al. (2007) A generalized transducing phage for the murine pathogen *Citrobacter rodentium*. *Microbiology* 153(Pt 9):2984–2988.
35. Vulić M, Dionisio F, Taddei F, Radman M (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci USA* 94(18):9763–9767.
36. Majewski J, Zawadzki P, Pickerill P, Cohan FM, Dowson CG (2000) Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J Bacteriol* 182(4):1016–1023.
37. Zawadzki P, Roberts MS, Cohan FM (1995) The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. *Genetics* 140(3):917–932.
38. Ansari MA, Didelot X (2014) Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* 196(1):253–265.
39. Kingman JF (2000) Origins of the coalescent. 1974–1982. *Genetics* 156(4):1461–1463.
40. Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145(2):505–518.