BMC Genomics

RESEARCH ARTICLE Open Access

Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*

Xavier Didelot[1]*, Guillaume Méric[2], Daniel Falush[3] and Aaron E Darling[4]

Abstract

Background: *Escherichia coli* is an important species of bacteria that can live as a harmless inhabitant of the guts of many animals, as a pathogen causing life-threatening conditions or freely in the non-host environment. This diversity of lifestyles has made it a particular focus of interest for studies of genetic variation, mainly with the aim to understand how a commensal can become a deadly pathogen. Many whole genomes of *E. coli* have been fully sequenced in the past few years, which offer helpful data to help understand how this important species evolved.

Results: We compared 27 whole genomes encompassing four phylogroups of *Escherichia coli* (A, B1, B2 and E). From the core-genome we established the clonal relationships between the isolates as well as the role played by homologous recombination during their evolution from a common ancestor. We found strong evidence for sexual isolation between three lineages (A+B1, B2, E), which could be explained by the ecological structuring of *E. coli* and may represent on-going speciation. We identified three hotspots of homologous recombination, one of which had not been previously described and contains the *aroC* gene, involved in the essential shikimate metabolic pathway. We also described the role played by non-homologous recombination in the pan-genome, and showed that this process was highly heterogeneous. Our analyses revealed in particular that the genomes of three enterohaemorrhagic (EHEC) strains within phylogroup B1 have converged from originally separate backgrounds as a result of both homologous and non-homologous recombination.

Conclusions: Recombination is an important force shaping the genomic evolution and diversification of *E. coli*, both by replacing fragments of genes with an homologous sequence and also by introducing new genes. In this study, several non-random patterns of these events were identified which correlated with important changes in the lifestyle of the bacteria, and therefore provide additional evidence to explain the relationship between genomic variation and ecological adaptation.

Background

Recombination is a fundamental process of bacterial evolution, capable of influencing the integrity of species [1-3]. Two types of recombination are typically distinguished: homologous recombination, where a fragment of a genome is replaced by the corresponding sequence from another genome [4], and non-homologous recombination, which causes genetic additions of new material and is also called lateral gene transfer (LGT) [5]. These two types of recombination may in fact often happen simultaneously, but they are usually studied separately because of the very different signatures they produce on the genomic sequences. Both homologous and non-homologous types of recombination are key elements of the evolution of bacteria and can be linked to variations in fitness, and thus ecologies and lifestyles. There is indeed an ecological component in bacterial recombination, in the sense that bacteria with overlapping living environments, reservoirs or hosts (i.e., "overlapping ecologies") will have more opportunities for genetic exchange than species or lineages living in drastically distinct environments. Recombination is therefore clearly conditioned by ecology, but conversely it is probable that recombination often drives

*Correspondence: x.didelot@imperial.ac.uk
[1] Department of Infectious Disease Epidemiology, Imperial College, Norfolk Place, London W2 1PG, UK
Full list of author information is available at the end of the article

ecological changes, for example by allowing favourable innovations to be exchanged by separate lineages adapting to a same lifestyle [3,6].

Escherichia coli is a good example of an environmentally versatile and adaptable bacterial species. It encompasses some strains able to live commensally with their host and others causing a relatively wide variety of disease symptoms, from diarrhoea or renal failure to meningitis [7]. On top of this commensal versus pathogen duality, which may not represent a strict categorization, *E. coli* can be found in a wide range of hosts, as well as secondary non-host environments such as water, soils or plants [8,9], in which it sometimes seems to maintain very well [10-13]. At the phylogenetic level, this plasticity is somewhat reflected by the population structure of *E. coli*, which is characterised by the presence of distinct phylogenetic groups (or "phylogroups") observable by phylogenetic reconstruction [14] or the use of specific markers [15]. Four major (A, B1, B2 and D) and two minor (E and F) phylogroups have so far been described [14,16]. Judging from the non-random isolation frequencies of different phylogroups in various hosts and environments [8,9,17], it seems that the fitness in different environments varies among *E. coli* isolates from different phylogroups, which raises the question of the evolutionary nature of these phylogroups. Are they the present reflection of *E. coli* subgroups undergoing speciation as a consequence of slightly variable ecologies? Or, the primary environment of any *E. coli* being the gastrointestinal tract of endotherms, is there a relative cohesion of these phylogroups within the *E. coli* species after all? An indirect but efficient method to answer these questions is to look at the patterns of recombination (homologous and non-homologous) between different strains and members of the different phylogroups. As mentioned above, recombination should be conditioned by existing ecological differences between lineages, and may even be partly responsible for them in which case this approach also has the potential to identify the genes that play a key role in the adaptation.

In this study, we contribute to the understanding of the association between genomic evolution and ecological adaptation by presenting bioinformatic analyses of recombination events (gene gain/loss and homologous recombination) between 27 publicly available genomes of *E. coli* from different phylogroups (A, B1, B2 and D) and ecological backgrounds (commensal and different pathotypes). More generally, our extensive knowledge about *E. coli* compared to other microbial species provides a unique opportunity to study the mechanisms of genomic evolution in its biological context. We used a genomic analytical pipeline (summarized in Figure 1) which combined progressiveMauve [18] for aligning the genomes, ClonalFrame [19] to establish their clonal relationships with one another, GenoPlast [20] to study non-homologous recombination and ClonalOrigin [21] to examine homologous recombination.

Methods

Genome sequences

A total of 30 genomes of *E. coli* were available from the NCBI reference sequence database [22] when this study was initiated. Three of these genomes (UMNO26 [23], IAI39 [23] and SMS-3-5 [24]) were described as members of phylogroup D but did not cluster together in our preliminary phylogenetic analysis (Additional file 1: Figure S1). Furthermore, these three genomes showed evidence of deviation in the molecular clock rate which could have confused the analyses presented here since the models in ClonalFrame [19] and ClonalOrigin [21] assume a constant clock rate (Additional file 1: Figure S1). These three genomes were therefore excluded so that we were left with a set of 27 genomes which is

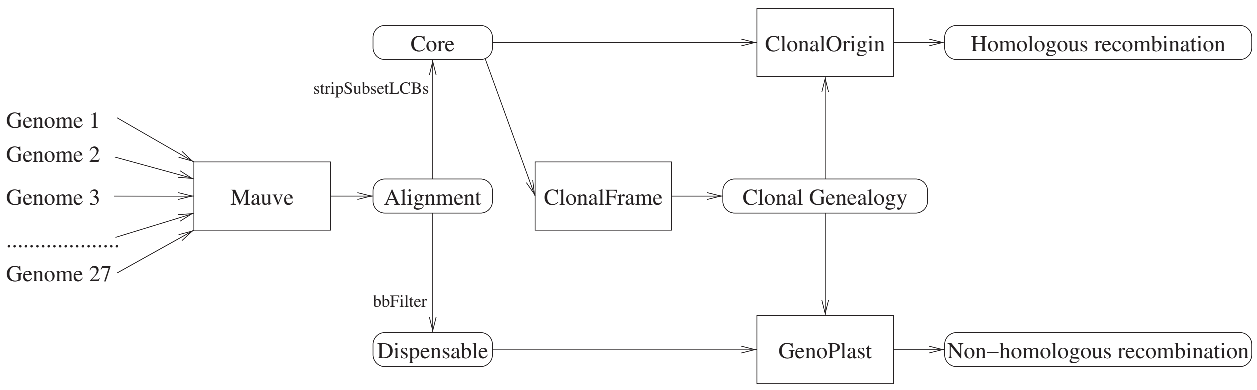

Figure 1 Genomic analytical pipeline used in this study. The genomes are first aligned using Mauve, and the core-genome is used to estimate the clonal genealogy using ClonalFrame. Non-core regions are then interpreted in terms of non-homologous recombination events on the branches of the clonal genealogy using GenoPlast, whereas core regions are analyzed using ClonalOrigin to infer homologous recombination events and their origins on the clonal genealogy.

summarized in Table 1. Several more genomes have recently become available on NCBI, but the complex analytical pipeline we used (Figure 1) could not easily accommodate them.

Multi-locus sequence typing data

To assess the representativeness of the 27 strains included in this study, we compared them with the isolates from the *E. coli* reference collection (ECOR) [44] which have been characterized by two independent Multi-Locus Sequence Typing [45,46] schemes. Fragments of 450-550bp from seven housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *recA* and *purA*) have been sequenced previously for a total concatenated length of 3423bp [47]. Additional fragments of 450-600bp from eight genes (*dinB*, *icdA*, *pabB*, *polB*, *putP*, *trpA*, *trpB* and *uidA*) have subsequently been sequenced for a total concatenated length of 4095bp [16]. To achieve maximum robustness, we combined the data from both studies to obtain 7518bp of sequence from each isolate. BLAST [48] was used to extract the sequences of each of the 15 gene fragments from each of the 27 genomes. A UPGMA dendrogram was then constructed to illustrate the phylogenetic relationship between the genomes and the ECOR collection (Figure 2).

Analysis of genomic content

The genomes of the 27 strains in Table 1 were aligned using progressiveMauve [18,49,50]. progressiveMauve does not use annotations to guide the alignment. Consequently, when there are multiple copies of a gene in the genome, progressiveMauve will usually align the copy that fits best in the context of surrounding sequence, unless the identity to a sequence in a different context scores so much better that it exceeds the breakpoint penalty. In general this will have the effect of aligning orthologous copies of genes unless the gene conversion rate among paralogs is

Table 1 Genomes used in this study

Strain	Pathotype	Phylogroup	Serotype	Length (Kbp)	Citation
ATCC8739	Commensal	A	O146	4743	[25]
HS	Commensal	A	O9:H4	4635	[26]
BL21	Commensal	A	O7	4568	[27]
REL606	Commensal	A	O7	4621	[27]
K-12/BW2952	Commensal	A	O16	4570	[28]
K-12/DH10B	Commensal	A	O16	4678	[29]
K-12/MG1655	Commensal	A	O16	4631	[30]
K-12/W3110	Commensal	A	O16	4638	[31]
IAI1	Commensal	B1	O8	4692	[23]
SE11	Commensal	B1	O158:H28	4879	[32]
55989	EAEC	B1	O128:H2	5154	[23]
12009	EHEC	B1	O103:H2	5441	[33]
E24377A	ETEC	B1	O139:H28	4971	[26]
11128	EHEC	B1	O111:H-	5363	[33]
11368	EHEC	B1	O26:H11	5689	[33]
EC4115	EHEC	E	O157:H7	5564	[34]
TW14359	EHEC	E	O157:H7	5520	[35]
EDL933	EHEC	E	O157:H7	5520	[36]
Sakai	EHEC	E	O157:H7	5490	[37]
CB9615	EPEC	E	O55:H7	5378	[38]
APEC01	ExPEC	B2	O1:K12:H7	5074	[39]
UTI89	ExPEC	B2	O18:K1:H7	5057	[40]
S88	ExPEC	B2	O45:K1	5024	[23]
CFT073	ExPEC	B2	O6:K2:H1	5223	[41]
ED1A	Commensal	B2	O81	5201	[23]
536	UPEC	B2	O6:K15:H31	4930	[42]
E2348/69	EPEC	B2	O45:K1	4957	[43]

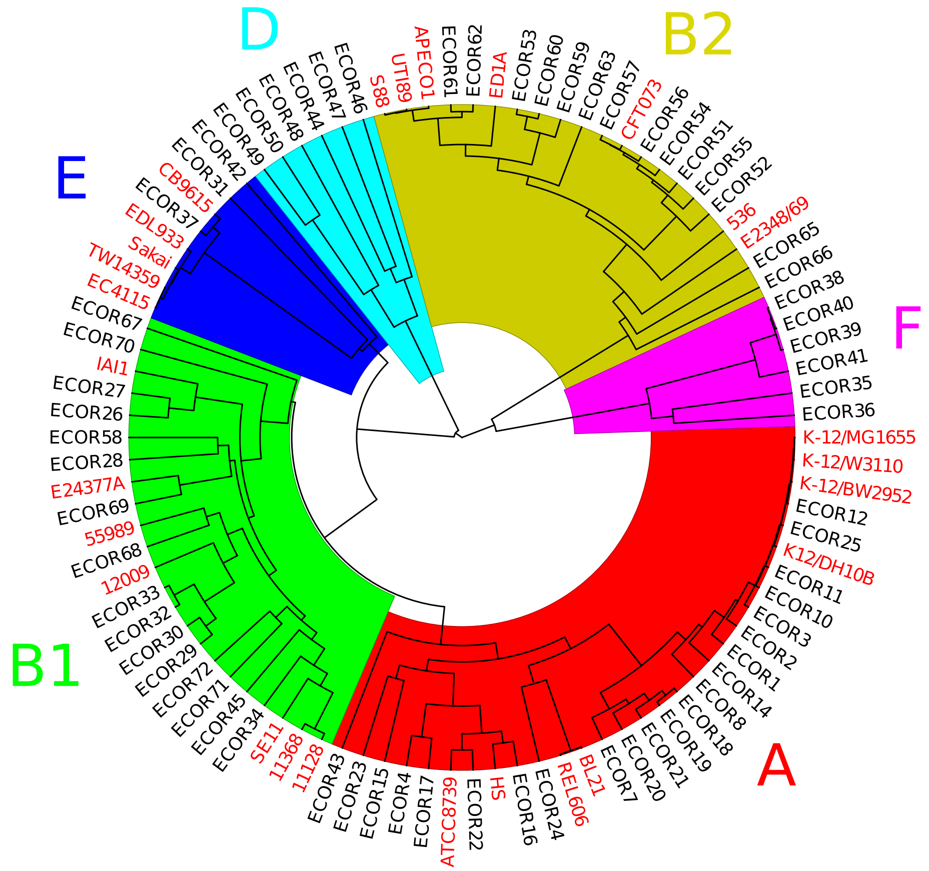

Figure 2 Representativeness of the genomes used. Phylogenetic relationships between the 27 genomes in this study (labels in red) and the ECOR reference collection (labels in black). Colors correspond to clade designations as follows: clade A in red, B1 in green, B2 in yellow, E in blue, D in cyan and F in mauve.

very high. The resulting alignment contained 2675 locally colinear blocks (LCBs). For all subsets of the genomes with cardinality ranging from 1 to 27, the concatenated size of the homologous regions found in all or a fraction of the subset was counted directly from the output of progressiveMauve. These values were used to generate Figure 3. Furthermore, for each pair of strains, a pairwise distance was computed representing the proportion of genome content that they have in common. This matrix of pairwise distances was then used to build the UPGMA tree in Figure 4B. The cophenetic correlation coefficient [51] for this tree was 0.89 indicating that it is a fairly good representation of the differences in genomic content between the genomes.

From the complete alignment of the 27 genomes, a matrix of feature presence/absence was computed using the bbFilter script distributed with Mauve, where each feature represented 50bp of unique sequence. This data was analyzed using GenoPlast [20] which infers how the genomic composition of the genomes evolved on the branches of the clonal genealogy (computed as explained in the next paragraph) assuming a model in which gain and loss of genetic material follow a relaxed molecular clock [52]. Briefly, GenoPlast explores the space of gain and loss events happening on branches that are compatible with the observed patterns of sharing of genomic regions observed in the genomes at the leaves of the tree. GenoPlast was run for 2,000,000 iterations with the first half discarded as burn-in. Good convergence and mixing properties were found by comparing different runs. The results of the GenoPlast analysis are shown in Figure 5.

Reconstruction of clonal genealogy

All regions of at least 500bp found in all 27 genomes were extracted from the progressiveMauve output using the stripSubsetLCBs script distributed with Mauve. A total of 765 such regions were found, ranging in size from 501bp to 27,115bp with a mean of 4322bp and a concatenated length of 3.3Mbp. These regions found in the 27 genomes represent the core-genome of *E. coli* (Figure 3). We applied ClonalFrame

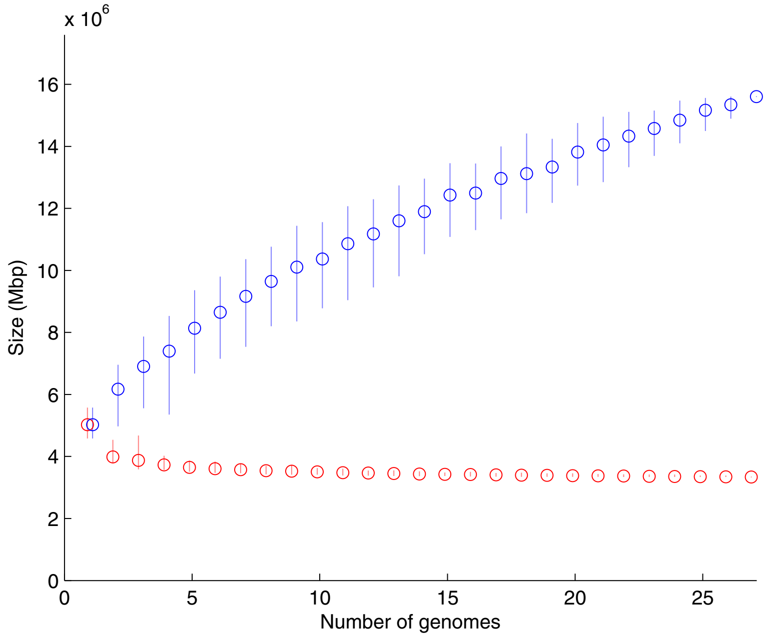

Figure 3 Core and pan genome cumulative plot. Concatenated length of the regions found in all (red) and at least one (blue) genome as more and more of the 27 genomes are aligned against altogether.

[19] to this core-genome in order to reconstruct the clonal relationships between the genomes. ClonalFrame is a Bayesian phylogenetic method which performs inference under an evolutionary model accounting for the effect of homologous recombination [19,53,54]. Five runs of ClonalFrame were performed independently each consisting of 100,000 iterations, the first half of which was discarded as burn-in. The results were compared between runs and found to be highly similar, indicating good convergence and mixing properties. The clonal genealogy inferred by ClonalFrame is shown in Figure 4A. The analyses of homologous and non-homologous recombination described below were performed conditionally on this clonal genealogy. Consequently, the fact that some genomes are more closely related to one another than others is fully accounted for in these analyses.

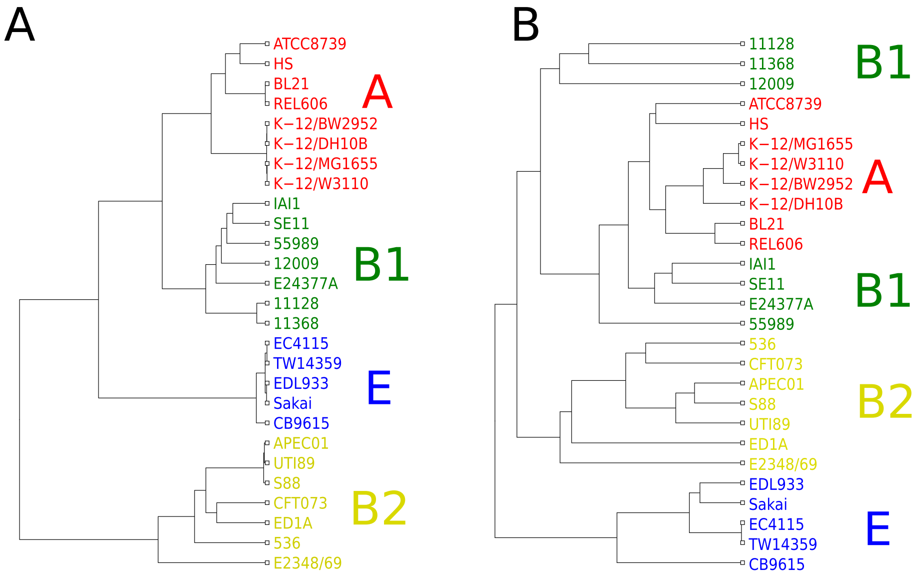

Figure 4 Genealogies based on homology and gene content. (**A**) ClonalFrame result based on core-genome. (**B**) UPGMA dendrogram based on similarity of genomic content. Colors correspond to clade designations as follows: clade A in red, B1 in green, B2 in yellow and E in blue.

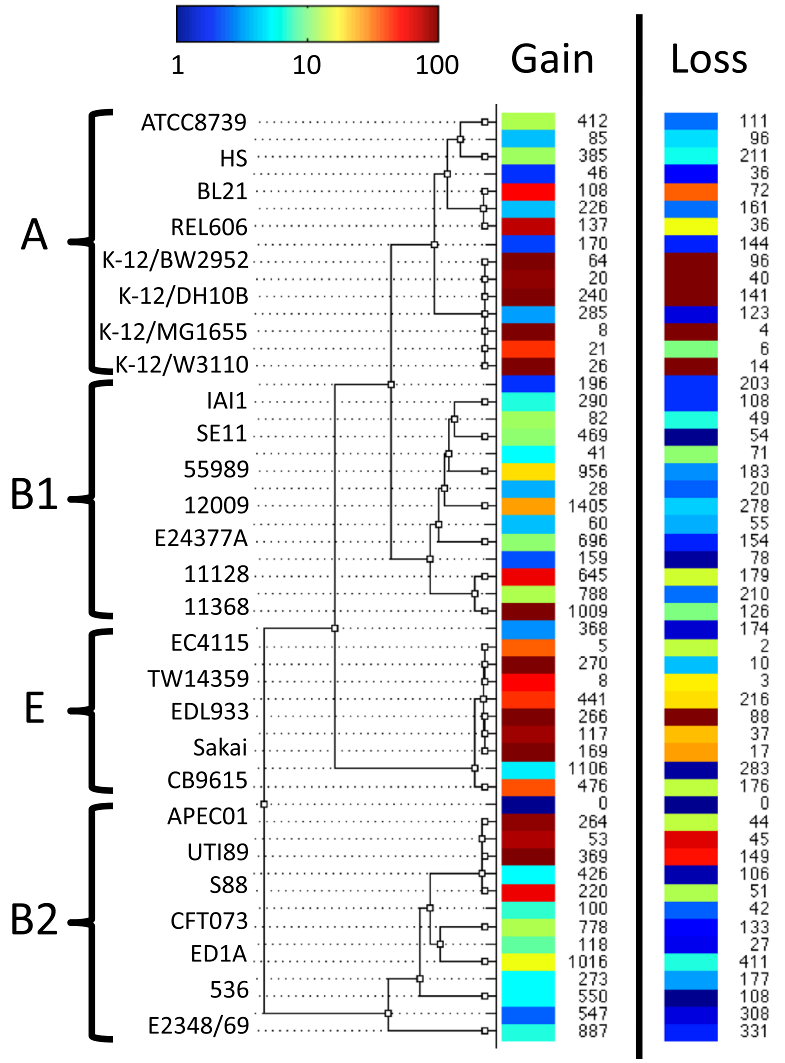

Figure 5 Patterns of non-homologous recombination. GenoPlast was used to infer how much material was gained and lost on each branch of the ClonalFrame tree. Each row corresponds to a branch of the tree as shown, gain is shown on the left and loss on the right. The numbers indicate the average amount of material gained and lost on each branch, measured in Kbp. The colors indicate how high the amounts gained or lost are relative to their expectations under a model of fixed gain and loss rates, where amounts gained and lost would be proportional to branch lengths. These colors are measured on a logarithmic relative scale as indicated at the top-left.

Analysis of homologous recombination

In order to further analyse the role played by homologous recombination during the diversification of *E. coli* from a common ancestor, we applied the computer software ClonalOrigin [21] which performs approximate inference under the coalescent with gene-conversion model [55,56]. ClonalOrigin detects recombination events, including their origin and destination on the clonal genealogy, and can therefore be used to reconstruct trends and patterns of homologous recombination [21,57]. The ClonalOrigin model rests on three global parameters which are the average length of recombination events δ and the scaled rates of mutation and recombination events respectively equal to $\theta_s = 2N_e\mu_s$ and $\rho_s = 2N_e r$ where N_e is the effective population size, μ is the per site per generation mutation frequency and r is the per site per generation recombination frequency. A first run of ClonalOrigin was performed for each of the 765 core regions where each region independently infers the three parameters (this phase is called "Step 2" in [21]). The median values of the three parameters across all regions were as follows: δ=542bp, θ_s=0.0125 and ρ_s=0.0128. ClonalOrigin was then rerun for each region with the three parameters set equal to these estimates (this phase is called "Step 3" in [21]). In both steps, ClonalOrigin was run for 2,000,000 iterations, the first half of which was discarded as burn-in.

Step 2 was only used to infer the values of the three global parameters, and all results presented here are based on the Step 3 results from ClonalOrigin. For instance, Figure 6 represents the number of recombination boundaries found in each of the 965 regions, with three hotspots (defined as contiguous regions of the genome in which the average recombination rate across alignment blocks is significantly higher than elsewhere in the genome) highlighted in grey. Figures 7 and 8 compare the number of inferred recombination events between different parts of the genealogy with the number expected under the prior model. These two figures are based on the numbers of the observed and expected recombination events computed by ClonalOrigin for all pairs of potential donor and recipient branches of clonal genealogy. These values are compiled in Additional file 2: Table S1.

Results and discussion

Representativeness of the strains used in this study

This study included 27 previously sequenced genomes of *Escherichia coli* (Table 1). To assess how representative these genomes are of the global diversity of the species, we compared them to the *Escherichia coli* reference collection (ECOR) [44] on the basis of two Multi-Locus Sequence Typing schemes which together spanned a total of 15 genes [16,47]. The resulting phylogeny (Figure 2) highlighted the six previously described lineages of *E. coli*, designated A, B1, B2, E, D and F [14,16]. Overall, the 27 strains covered much of the diversity of *E. coli*, with eight strains in clade A, seven in clade B1, five in clade E and seven in clade B2 (Figure 2; Table 1). In each of these four clades, the strains seem to represent much of the within-clade diversity rather than being closely related within the clade. However, two clades were not represented in this genomic panel: clade D and clade F. Three genomes from these phylogroups (IAI39 [23], SMS-3-5 [24] and UMN026 [23]) were initially intended to be included, but were removed because they showed evidence for significant deviation from the assumption of a fixed molecular clock (Additional file 1: Figure S1). Figure 2 indicates how the diversity of the genomes in this study relates with that of the ECOR strains, however, it should be noted that the issue of biased sampling of bacterial isolates is frequent and it is never possible to be sure of the representativeness of a sample [4,58].

Reconstruction of the clonal genealogy

Aligning the 27 genomes using progressiveMauve [18,49, 50] allowed us to compare their genomic content. As more genomes are considered in the analysis, the cumulative size of genomic regions shared by them decreased down to about 3.3Mbp, or about two thirds of each genome (Figure 3). Since this length is roughly constant whether 10, 15, 20 or all 27 genomes are aligned (Figure 3), these

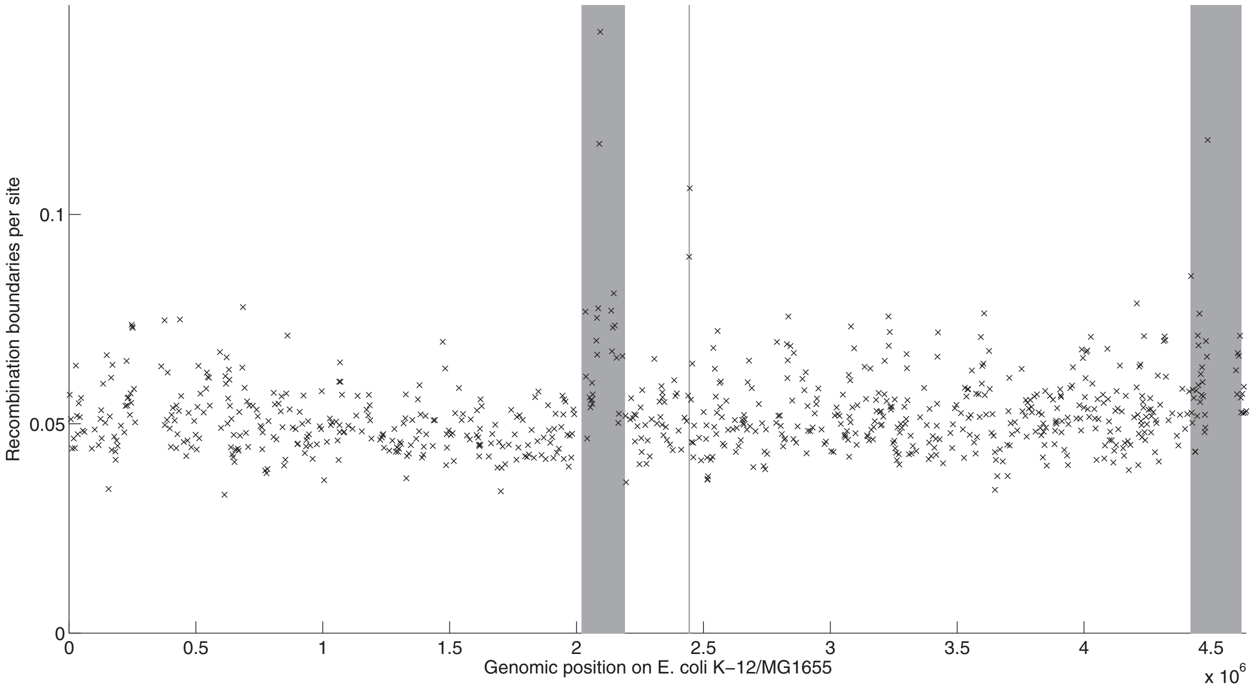

Figure 6 Intensity of recombination along the genome. Scatter plot where each cross represents a genomic region found in all 27 genomes. The X-axis indicates the positionof the region in the reference genome K-12 MG1655 [30] and the Y-axis is a measure of the intensity of recombination inferred by ClonalOrigin. Three hotspots of recombination are highlighted in grey.

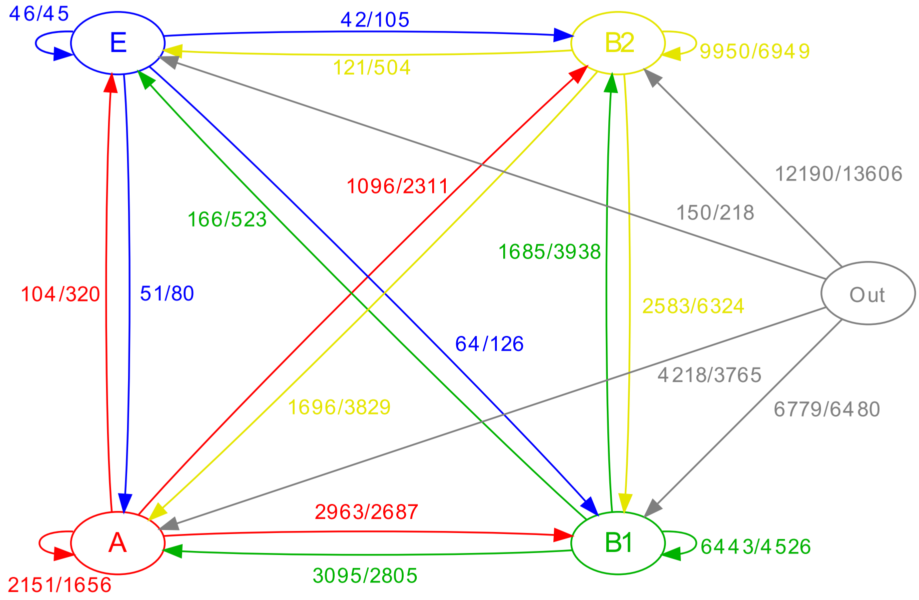

Figure 7 Flux of recombination between clades. Graph where the four clades of *Escherichia coli* A, B1, B2 and E are shown as nodes and edges represent recombination within and between them. Each edge corresponds to imports from one of the four clades (or external) into one of the four clades, and is labelled by two values separated by a forward slash. The first value is the number of recombination events inferred by ClonalOrigin. The second value is the number of recombination events expected under the ClonalOrigin model. With the exception of the flux from clade E into clade E, all observed values are outside of their expected 99% credibility intervals.

regions are likely to represent the core-genome of *E. coli*, which means that homologs of these regions would be found in virtually any sequenced genome. We found 765 core regions present in the genomes of all 27 strains, with total length 3,306,899bp. These core regions were input into ClonalFrame [19] in order to estimate the clonal genealogy in a way that accounts for homologous recombination which can confuse the signal of clonal inheritance [59]. This aspect is important because recombination has been reported to be frequent in *E. coli* by a large number of previous studies [14,16,47,60,61]. The inferred clonal genealogy (Figure 4A) consisted of four clades corresponding to A, B1, B2 and E. The relationships between genomes within clades were fully resolved, which is typically not achievable with MLST (eg. [14,16]). The relationships between clades were unbalanced, with clade A and B1 most closely related to each other, and clade B2 most distant from any other clade. The stemminess (ie. the ratio of internal to external branch lengths) of this tree was compatible with expectation under the standard coalescent model (Additional file 3: Figure S2), suggesting no evidence for population size variation during the evolution of *E. coli* [62-64].

Analysis of the dispensable genome

In contrast to the core regions described above, non-core regions are found only in a strict subset of the genomes. The set of non-core regions is called the dispensable genome and together with the core genome forms the pan-genome [65-67]. The cumulative length of the non-core regions continues to increase up to the 27th genome, showing no sign of flattening, with each new genome adding about 250Kbp of previously unobserved sequence (Figure 3). This distribution has been observed before, including in *E. coli*, and its pan-genome has consequently been called "open" [23,65-67]. However, an important difference between these previous studies and ours is that in Figure 3 the lengths of genomic material are measured directly whereas previous studies counted the number of genes. Our analysis is therefore robust to the problem of identifying homologous families of genes. Nevertheless, this result indicates that the pan-genome of species with a high diversity and ecological plasticity such as *E. coli* draws from a large repertoire of genes that can be gained and lost through lateral gene transfer [5,67].

The similarity of the genomes in terms of genomic content was calculated from the patterns of presence and absence of non-core regions (Figure 4B). Compared with the clonal genealogy (Figure 4A), the clade structure is only partly preserved in this tree of genome content, with clades B2 and E intact but clades A and B1 intermingled. Clade B1 was split into three parts which were perfectly congruent with pathotypes. The three EHEC strains 12009, 11368 and 11128 [33] formed one separate cluster. The two commensal strains IAI1 [23] and SE11 [32] and

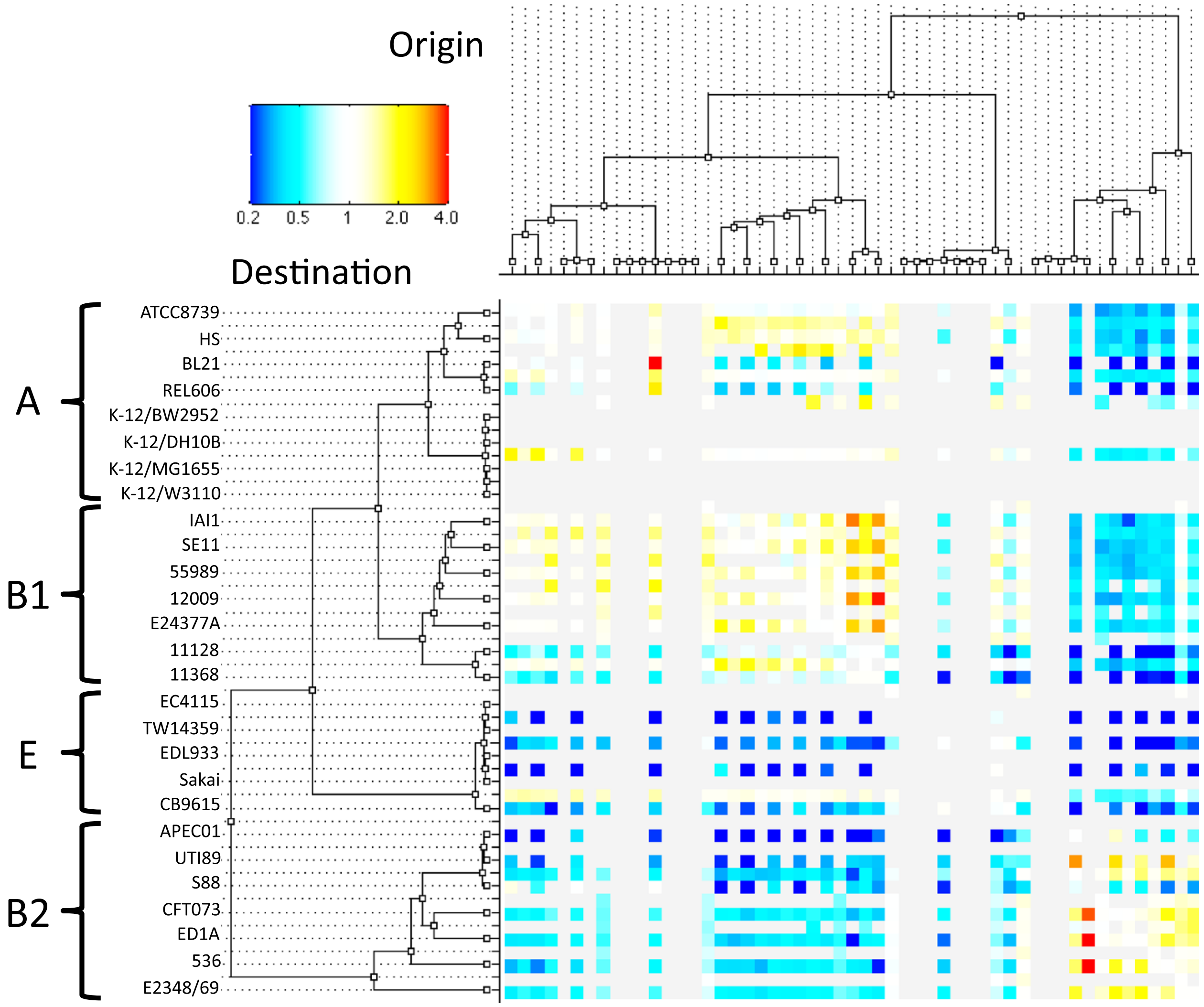

Figure 8 Detailed flux of recombination. Heat map showing the number of recombination events inferred by ClonalOrigin relative to its expectation under the prior model for each donor/recipient pair of branches. Cells for which both the number of observed and expected events are less or equal than three are shown in light gray.

the only ETEC strain E24377A [26] constituted another separate cluster, in which the two commensal strains were closest to each other. Finally, the EAEC strain 55989 [23] was on a separate branch in spite of its close relationship with the commensal strains IAI1 and SE11 in the clonal genealogy (Figure 4A). This subdivision of B1 in terms of genomic content has been partially hinted at before [68] and the fact that it is congruent with pathotypes suggests that it is linked with differences in ecological and pathogenic lifestyles. The presence or absence of genomic regions in the 27 observed genomes is the result of a process of gain and loss of content by the ancestors of the genomes since their evolution from a common ancestor. If gain and loss happened randomly and at constant rates, the tree based on genomic content (Figure 4B) would be expected to to be very similar to the tree based on homology of the core-genome (Figure 4A) since the evolution of both core and pan genomes would then follow the same molecular clock. The two trees were however highly different, indicating that the non-homologous recombination process (gain and loss of regions) did not follow a strict molecular clock. GenoPlast [20] was used to infer the non-homologous recombination events that happened in the context of the clonal genealogy inferred by ClonalFrame (Figure 4A) under a model where the rates of gain and loss are allowed to change. The results of the GenoPlast analysis are shown in Figure 5, with differences in the rates of gain and loss on specific branches spanning two orders of magnitude. The rates of gain and (to a lesser extent) loss of genomic material were found to be higher on the short recent branches within clades A, E and B2 than on older and longer branches, which explained the higher stemminess of the genomic content tree (Figure 4B) compared with the clonal genealogy (Figure 4A).

The branch directly above EHEC strain 12009 had the largest amount of gain of any branch (1405 Kbp) whereas the branch above the common ancestor of the other two B1 EHEC strains 11368 and 11128 was the highest amount of gain for an internal branch (788 Kbp; with the exception of the very long branch above clade E). Amongst the genomic material gained on these two branches, 265 Kbp were shared by the three genomes, which explained why they clustered together in Figure 4B. The distribution of this gain on the three genomes (Additional file 4: Figure S3) indicated that their convergence in genomic content happened as a result of multiple gain events that happened both on the branch above 12009 and on the branch above the common ancestor of 11368 and 11128. The convergence in genomic content of the three EHEC B1 strains was therefore reciprocal rather than unidirectional. Few convergence events were found on the branches directly above 11368 and 11128 (Additional file 4: Figure S3) in spite of considerable gain on these branches (1009Kbp and 645Kbp respectively), which could indicate that the convergence in gene content with 12009 is not on-going. Unsurprisingly, this convergence involved several genes known to be EHEC determinants, including Shiga toxins [69] and all genes from the locus of enterocyte effacement (or LEE [70]). However, it also included additional genes, such as flagellar genes (*fli* [71]) and a few metabolic clusters (*frl* [72] and *gal* [73]) with a notable presence of genes involved in aromatic compounds metabolism (*hpa*, *hpc*, *mhp* and *mhp* [74]). These genes were not present in the other B1 strains examined in this study, which may indicate that acquiring EHEC determinants via HGT is an important means of *E. coli* adaptation, possibly enhanced by the differences in host-associated selective pressures on EHEC compared to commensals or more opportunistic pathotypes.

Homologous recombination hotspots in *Escherichia coli*

To quantify the propensity, genomic distribution and directionality of homologous recombination during the evolution of *E. coli*, we applied ClonalOrigin [21] to the 765 core regions and assuming the clonal relationships between genomes estimated by ClonalFrame [19] in Figure 4A. The average length of fragments involved in homologous recombination was estimated at δ=542bp. This is almost ten times higher than a previous estimate in *E. coli* [23], but is of the same order as recent whole-genome estimates in *Bacillus cereus* [21], *Helicobacter pylori* [75] or *Chlamydia trachomatis* [76]. The relative rate of occurrence of recombination and mutation [77] was estimated at $\rho_s/\theta_s = 0.0128/0.0125 = 1.024$ which means that overall recombination happened just as frequently as mutation. The estimated rate of homologous recombination was fairly constant throughout the genome (Figure 6), with the exception of three clear hotspots (highlighted in grey) in which recombination rates were significantly higher. This included two large regions around the *rfb* operon involved in synthesis of the O antigen (positions 2,020 to 2,190 Kbp in the reference genome K-12/MG1655 [30]) and around the *fimA* gene (positions 4,420 to 4,620 Kbp). These two regions had been reported previously as hotspots of diversity and recombination [23,78].

A smaller recombination hotspot was also detected, made of just two nearly adjacent core regions (between positions 2,442 and 2,447 Kbp). This region had a similarly high recombination rate as the two regions above, but had not previously been detected as a hotspot, perhaps because of its small size (around 5 Kbp). This hotspot contained genes *yfcL*, *yfcM*, *yfcA*, *mepA*, *aroC*, *prmB* and *smrB*. The gene *mepA* encodes for a murein endopeptidase [79] whose role is presumably to restructure the bacterial cell wall during elongation or stabilise the peptidoglycan. Mutational analyses on *mepA* [79,80] do not provide enough information to explain why recombination should be high for this gene. In the bacterial cell, *aroC* governs the synthesis of chorismate, a key precursor to the biosynthesis of aromatic compounds including the amino acids tryptophan and phenylalanine but also the siderophore enterobactin. The positive maintenance of a functional allele of *aroC* is arguably crucial for the cell to maintain appropriate levels of these amino acids and siderophores in natural conditions. In *Salmonella* [81], as well as in *Brucella suis* [82], *aroC* is required for virulence. Incidentally, *aroC* is a common target to produce knocked-out attenuated vaccine strains [83], for instance in *Salmonella* serovars Typhi [84-86] and Typhimurium [81,85,87,88], pathogenic *E. coli* [89], *Brucella suis* [82], *Burkholderia pseudomallei* [90] or *Edwardsiella tarda* [91]. To our knowledge, this is the first mention of *aroC* being part of a recombination hotspot, giving additional clues on evolutionary dynamics at this locus. Depending on how *aroC* is involved with virulence in *E. coli*, it may be under selective pressure from the immune system of the host, which could explain the observed peak in recombination rate [4], but this hypothesis will need further work to be fully assessed.

Flux of homologous recombination

The numbers of recombination events inferred by ClonalOrigin were counted for every combination of clades receiving and donating, and these values were compared with their expectation under the ClonalOrigin model which represents a close approximation to the coalescent model with gene conversion [21,55,56]. This comparison revealed significant non-uniformity in the homologous recombination flux within and between clades (Figure 7). The three clades A, B1 and B2 had higher numbers of within-clade recombination than expected, whereas clade E had almost exactly the expected number. On the other hand, the number of recombination events detected between clades was almost systematically below its expected value, with the only exception being recombination from clade A to B1 and vice-versa which had slightly higher than expected values. Clades A and B1 are the two most closely related phylogroups (Figure 4A) which may contribute to explain this observation. Overall, inter-phylogroup recombination fluxes were lower than intra-phylogroup ones, which is compatible with the hypothesis that there is a preferred way of gene sharing within phylogroups [92]. This preferred exchange among strains of the same phylogroups could be explained by the possibility that the different *E. coli* phylogroups have slightly distinct ecological overlaps, which makes the likelihood of gene transfer higher among them than between them.

A similar analysis as above was performed on a branch-by-branch basis rather than a clade-by-clade basis (Figure 8), the only added difficulty being that some donor/recipient pairs of branches have too low numbers of expected and observed recombination events for the comparison to be meaningful (represented in grey in Figure 8).This analysis confirmed the general pattern described above, with more recombination than expected within-clades and between A and B1, and less recombination between all other clades. However, it also allowed the comparison of the individual behaviours of strains belonging to the same clade. For instance, strains BL21 and REL606 [27] showed little history of importing recombination from clade B1, contrasting with ATCC8739 [25] or HS [26] even though all four strains belong to clade A. This may be explained by the fact that these two strains are laboratory-adapted derived from *E. coli* strain B [27,93], so that they would have had little or no opportunity for recent encounter and recombination with B1 strains.The four K-12 strains in this study [28-31] were also laboratory-adapted, but had terminal branches too small to reliably estimate deviations in the number of recombination events. These four strains all originated from bioengineering manipulation on K-12 lineages over the last century and therefore harbour a very limited number of differences between them compared to what would be observed in natural populations.

The B1 strains 11128 and 11368 [33] showed significantly less sign of import from clade A (and to a lesser extent from B1) than other strains of B1. This observation implies that these EHEC strains have stopped recombining with strains of clade A (which are all commensals) as they adapted to this new pathogenic lifestyle. Two of the highest values throughout Figure 8 were the ones corresponding to imports from strains 11128 and 11386 into strain 12009 [33]. As previously noted, these are the only three EHEC strains in B1, and these three genomes have been converging in genomic content due to numerous non-homologous recombination events. This result indicates that the three genomes also have an extensive history of convergence through homologous recombination, which may have occurred at the same time as the gain of new shared genes. The evolutionary history of these three genomes seems therefore analogous to that of *Salmonella* serovars Typhi and Paratyphi A, for which both core and pan genomes converged through recombination as they were progressively adapting to exclusive infection of the human host [6].

Speciation in *E. coli*

In the analysis of homologous recombination described above, three groups corresponding respectively to phylogroups E, B2, and A+B1 exhibited more recombination within than between one another (Figures 7 and 8). This pattern is compatible with a definition of speciation in bacteria in which recombination plays the role of a cohesive force counterbalancing divergence by genetic drift and population structure, and where species appear when this force is weakened between lineages [1,2,94]. Under such a model, patterns of genetic diversity can be generated *in silico* similar to those observed for example in *Salmonella enterica* [95,96]. The three groups might therefore represent lineages that, because of slightly distinct ecologies or notable variations in the species life cycle, have gradually diverged too far from one another for recombination to play its cohesive role, so that they might eventually become separate species, should these variations remain or increase. In other words, all *E. coli* phylogenetic backgrounds are found in the gut of endotherms [14] which is their primary environment, and to some extent in nonhost secondary environments [8,9] but it sounds plausible that phylogroup-associated variations in ecological fitness in different hosts or secondary environments could gradually decrease the physical and ecological overlap of strains from different phylogroups through time, and therefore the genetic flux between them. A number of studies seem to support this hypothesis, as different proportions of the different phylogroups are found in different environments and hosts [8,9,97,98]. Additionally, some phylogroups seem to harbour strains that have either host-restricted or more generalist lifestyles [99], as well as strains that are either resident or transients in their ability to colonize the gut [100]. Our study contributes to highlight that such variations in ecology could potentially have an impact on genetic exchange in *E. coli*.

An additional number of factors can be evoked to explain why the three groups would be diverging, including differences in their geographic distribution, adaptative selection, or simply as a result of the dependence of recombination on homology between donor and recipient [4,94,96,101,102]. The three groups are clearly separate in terms of genomic content (Figure 4B) which could explain why they recombine less with each other and why clades A and B1 still recombine frequently since they are not differentiated in terms of genomic content. To test this hypothesis, we compared the distribution of the number of recombination events found in the middle and at the edge of core regions (Additional file 5: Figure S4). We found that recombination happened more often in the middle of core regions at a small but highly significant level (Kolmogorov-Smirnov test; p-value=8.8e-09). If the variable genomic content is not just a random process, then homologous recombination would be expected to happen less often around these genes, a concept sometimes called fragmented speciation or "species in pieces" [103-106] as it would predict that speciation can apply differentially across the genome. Our results therefore demonstrate that fragmented speciation applies to *E. coli*, and that difference in genomic content is at least one of the factors driving the divergence of the three lineages.

Conclusions

We applied a pipeline of statistical analyses in order to compare the sequences of 27 *E. coli* genomes and reveal the ancestral history of clonal relationships, homologous recombination events and non-homologous recombination events that has led the ancestor of *E. coli* to diversify into the genomes we see today. The overall picture was one of divergence between three lineages (A+B1, B2, E) which were well differentiated on the basis of both genomic content and preference for homologous recombination, with the former apparently driving the latter as expected under a fragmented speciation scenario. However, against this divergence background, we observed the convergence of three EHEC strains within B1 in both their core- and pan-genomes. These observations were correlated with the diversity of ecology and pathogenicity of the *E. coli* strains, and provide hypotheses for which genes and evolutionary processes are adaptively important.

Additional files

Additional file 1: Figure S1. Test of molecular clock assumption. Neighbour-joining phylogenetic reconstruction based on all 30 genomes available from NCBI and which shows that three of them (UMNO26, IAI39 and SMS-3-5) showed significant deviation from the assumption of constant molecular clock.

Additional file 2: Table S1. Detailed results of the ClonalOrigin analysis. This table contains all expected and observed values of the number of recombination events for all pairs of donor and recipient branches, as computed by ClonalOrigin. This is the data on which Figure 7 is based. There is a cell for each donor/recipient combination, and the cells are ordered vertically and horizontally in the same way as in . In each cell, two values are given separated by a semi-colon: the first one is the observed value and the second one is the expected value.

Additional file 3: Figure S2. Test of ancestral population size dynamics. Distribution of expected values of stemminess under the coalescent model. The observed value for the clonal genealogy estimated by ClonalFrame is shown as a vertical line and falls within the expected values.

Additional file 4: Figure S3. Gain in the three genomes 12009, 11368 and 11128. The genomic regions gained by the three genomes 12009, 11368 and 11128 are colored. The regions in red are the ones that are uniquely shared by the three genomes, whereas the regions in green are not. For genome 12009, only the gain happening on the branch directly above is shown. For genomes 11368 and 11128, the gain on the branches directly above are shown using lighter green and red, and the gain that happened on the branch above the common ancestor of 11368 and 11128 is shown using darker green and red.

Additional file 5: Figure S4. Test of the fragmented speciation model. Boxplots of the distributions of the numbers of recombination events found in the middle (left) and at the edge of core regions (right). To generate the distribution on the left, the number of recombination events affecting the middle position was counted for each of the 765 core regions. To generate the distribution on the right, the number of recombination events affecting the site 10bp after the beginning of each core region was counted, as well as the number of recombination events affecting the site 10bp before the end of each core region.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

XD, DF and AED conceived and designed the study. XD, GM and AED analyzed the data. All authors contributed to the writing of the paper and approved the final manuscript.

Acknowledgements

XD was supported by a fellowship from the UKCRC Modernising Medical Microbiology Consortium, which is funded under the UKCRC Translational Infection Research Initiative supported by the Wellcome Trust (Grant 087646/Z/08/Z) and by the Medical Research Council, Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the Department of Health (Grant G0800778). AED was supported in part by National Science Foundation award
ER 0949453.

Author details

[1]Department of Infectious Disease Epidemiology, Imperial College, Norfolk Place, London W2 1PG, UK. [2]College of Medicine, Swansea University, Swansea, SA2 8PP, UK. [3]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103 , Germany. [4]Genome Center, University of California, Davis, CA 95616, USA.

Received: 1 February 2012 Accepted: 30 May 2012
Published: 19 June 2012

References

1. Achtman M, Wagner M: **Microbial diversity and the genetic nature of microbial species.** *Nat Rev Microbiol* 2008, **6:**431–440.
2. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP: **The bacterial species challenge: making sense of genetic and ecological diversity.** *Science* 2009, **323**(5915):741–746.
3. Sheppard S, McCarthy N, Falush D, Maiden M: **Convergence of Campylobacter species: implications for bacterial evolution.** *Science* 2008, **320**(5873):237–239.
4. Didelot X, Maiden MC: **Impact of recombination on bacterial evolution.** *Trends Microbiol* 2010, **18:**315–322.

5. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405**(6784):299–304.
6. Didelot X, Achtman M, Parkhill J, Thomson N, Falush D: **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.
7. Croxen MA, Finlay BB: **Molecular mechanisms of Escherichia coli pathogenicity.** *Nat Rev Microbiol* 2010, **8:**26–38.
8. Walk S, Alm E, Calhoun L, Mladonicky J, Whittam T: **Genetic diversity and population structure of Escherichia coli isolated from freshwater beaches.** *Environ Microbiol* 2007, **9**(9):2274–2288.
9. Bergholz PW, Noar JD, Buckley DH: **Environmental patterns are imposed on the population structure of Escherichia coli after fecal deposition.** *Appl Environ Microbiol* 2011, **77:**211–219.
10. Ishii S, Ksoll WB, Hicks RE, Sadowsky MJ: **Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds.** *Appl Environ Microbiol* 2006, **72:**612–621.
11. Texier S, Prigent-Combaret C, Gourdon MH, Poirier MA, Faivre P, Dorioz JM, Poulenard J, Jocteur-Monrozier L, Moënne-Loccoz Y, Trevisan D: **Persistence of culturable Escherichia coli fecal contaminants in dairy alpine grassland soils.** *J Environ Qual* 2008, **37**(6): 2299–2310.
12. Brennan FP, Abram F, Chinalia FA, Richards KG, O'Flaherty V: **Characterization of environmentally persistent Escherichia coli isolates leached from an Irish soil.** *Appl Environ Microbiol* 2010, **76**(7):2175–2180.
13. Brennan FP, O'Flaherty V, Kramers G, Grant J, Richards KG: **Long-term persistence and leaching of Escherichia coli in temperate maritime soils.** *Appl Environ Microbiol* 2010, **76**(5):1449–1455.
14. Tenaillon O, Skurnik D, Picard B, Denamur E: **The population genetics of commensal Escherichia coli.** *Nat Rev Microbiol* 2010, **8**(3):207–217.
15. Clermont O, Bonacorsi S, Bingen E: **Rapid and simple determination of the Escherichia coli phylogenetic Group.** *Appl Environ Microbiol* 2000, **66**(10):4555–4558.
16. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E, Lortholary O, Clermont O, Denamur E, Picard B, Nassif X, Brisse S: **Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains.** *BMC Genomics* 2008, **9:**560–560.
17. Skurnik D, Bonnet D, Bernède-Bauduin C, Michel R, Guette C, Becker JM, Balaire C, Chau F, Mohler J, Jarlier V, Boutin JP, Moreau B, Guillemot D, Denamur E, Andremont A, Ruimy R: **Characteristics of human intestinal Escherichia coli with changing environments.** *Environ Microbiol* 2008, **10**(8):2132–2137.
18. Darling A, Mau B, Perna N: **progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement.** *PLoS one* 2010, **5**(6):e11147.
19. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data.** *Genetics* 2007, **175**(3):1251–1266.
20. Didelot X, Darling A, Falush D: **Inferring genomic flux in bacteria.** *Genome Res* 2009, **19:**306–317.
21. Didelot X, Lawson D, Darling A, Falush D: **Inference of homologous recombination in bacteria using whole-genome sequences.** *Genetics* 2010, **186**(4):1435–1449.
22. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, **37**(Database issue):32–36.
23. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EP, Denamur E: **Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths.** *PLoS Genet* 2009, **5:**e1000344.
24. Fricke WF, Wright MS, Lindell AH, Harkins DM, Baker-Austin C, Ravel J, Stepanauskas R: **Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate Escherichia coli SMS-3-5.** *J Bacteriol* 2008, **190**(20):6779–6794.
25. Copeland A, Lucas S, Lapidus A, Glavina del Rio T, Dalin E, Tice H, Bruce D, Goodwin L, Pitluck S, Kiss H, Brettin T, Detter J, Han C, Kuske C, Schmutz J, Larimer F, Land M, Hauser L, Kyrpides N, Mikhailova N, Ingram L, Richardson P: **Complete sequence of *Escherichia coli* C str. ATCC 8739.** [http://www.ncbi.nlm.nih.gov/nucleotide/169752989].
26. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel: **The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates.** *J Bacteriol* 2008, **190**(20):6881–6893.
27. Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH, Couloux A, Lee SW, Yoon SH, Cattolico L, Hur CG, Park HS, Ségurens B, Kim SC, Oh TK, Lenski RE, Studier FW, Daegelen P, Kim JF: **Genome sequences of Escherichia coli B strains REL606 and BL21(DE3).** *J Mol Biol* 2009, **394**(4):644–652.
28. Ferenci T, Zhou Z, Betteridge T, Ren Y, Liu Y, Feng L, Reeves PR, Wang L: **Genomic sequencing reveals regulatory mutations and recombinational events in the widely used MC4100 lineage of Escherichia coli K-12.** *J Bacteriol* 2009, **191**(12):4025–4029.
29. Durfee T, Nelson R, Baldwin S, Plunkett G, Bxurland V, Mau B, Petrosino JF, Qin X, Muzny DM, Ayele M, Gibbs RA, Csörgo B, Pósfai G, Weinstock GM, Blattner FR: **The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse.** *J Bacteriol* 2008, **190**(7):2597–2606.
30. Blattner FR, Plunkett G, Bloch CA, Perna NT, Bxurland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of Escherichia coli K-12.** *Science* 1997, **277**(5331):1453–1474.
31. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL: **Escherichia coli K-12: a cooperatively developed annotation snapshot.** *Nucleic Acids Res* 2006, **34:**1–9.
32. Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, Park SH, Ooka T, Iyoda S, Taylor TD, Hayashi T, Itoh K, Hattori M: **Complete genome sequence and comparative analysis of the wild-type commensal Escherichia coli strain SE11 isolated from a healthy adult.** *DNA Res* 2008, **15**(6):375–386.
33. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, Tobe T, Hattori M, Hayashi T: **Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli.** *Proc Natl Acad Sci U S A* 2009, **106**(42):17939–17944.
34. Eppinger M, Mammel MK, Leclerc JE, Ravel J: **Cebula TA : Genomic anatomy of Escherichia coli O157:H7 outbreaks.** *Proc Natl Acad Sci U S A* 2011, **108**(50):20142–20147.
35. Kulasekara BR, Jacobs M, Zhou Y, Wu Z, Sims E, Saenphimmachak C, Rohmer L, Ritchie JM, Radey M, McKevitt M, Freeman TL, Hayden H, Haugen E, Gillett W, Fong C, Chang J, Beskhlebnaya V, Waldor MK, Samadpour M, Whittam TS, Kaul R, Brittnacher M, Miller SI: **Analysis of the genome of the Escherichia coli O157:H7 2006 spinach-associated outbreak isolate indicates candidate genes that may enhance virulence.** *Infect Immun* 2009, **77**(9):3713–3721.
36. Perna NT, Plunkett G, Bxurland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis WN, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR: **Genome sequence of enterohaemorrhagic Escherichia coli O157:H7.** *Nature* 2001, **409**(6819):529–533.
37. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: **Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12.** *DNA Res* 2001, **8:**11–22.
38. Zhou Z, Li X, Liu B, Beutin L, Xu J, Ren Y, Feng L, Lan R, Reeves PR, Wang L: **Derivation of Escherichia coli O157:H7 from its O55:H7 precursor.** *PLoS One* 2010, **5:**e8700.
39. Johnson TJ, Kariyawasam S, Wannemuehler Y, Mangiamele P, Johnson SJ, Doetkott C, Skyberg JA, Lynne AM, Johnson JR, Nolan LK: **The genome sequence of avian pathogenic Escherichia coli strain O1:K1:H7 shares strong similarities with human extraintestinal**

pathogenic E. coli genomes. *J Bacteriol* 2007, **189**(8): 3228–3236.

40. Chen SLL, Hung CSS, Xu J, Reigstad CSS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RRR, Ozersky P, Armstrong JRR, Fulton RSS, Latreille JPP, Spieth J, Hooton TMM, Mardis ERR, Hultgren SJJ, Gordon JII: **Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: A comparative genomics approach.** *Proc Natl Acad Sci U S A* 2006, **103**(15):5977–5982.
41. Welch RA, Bxurland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli .** *Proc Natl Acad Sci U S A* 2002, **99**(26):17020–17024.
42. Brzuszkiewicz E, Brüggemann H, Liesegang H, Emmerth M, Olschläger T, Nagy G, Albermann K, Wagner C, Buchrieser C, Emody L, Gottschalk G, Hacker J, Dobrindt U: **How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains.** *Proc Natl Acad Sci U S A* 2006, **103**(34): 12879–12884.
43. Iguchi A, Thomson NR, Ogura Y, Saunders D, Ooka T, Henderson IR, Harris D, Asadulghani M, Kurokawa K, Dean P, Kenny B, Quail MA, Thurston S, Dougan G, Hayashi T, Parkhill J, Frankel G: **Complete genome sequence and comparative genome analysis of enteropathogenic Escherichia coli O127:H6 strain E2348/69.** *J Bacteriol* 2009, **191:**347–354.
44. Ochman H, Selander RK: **Standard reference strains of Escherichia coli from natural populations.** *J Bacteriol* 1984, **157**(2):690–693.
45. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG: **Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms.** *PNAS* 1998, **95**(6):3140–3145.
46. Maiden MC: **Multilocus sequence typing of bacteria.** *Annu Rev Microbiol* 2006, **60:**561–588.
47. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M: **Sex and virulence in Escherichia coli: an evolutionary perspective.** *Mol Microbiol* 2006, **60**(5):1136–1151.
48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
49. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14**(7):1394–1403.
50. Darling AE, Treangen TJ, Messeguer X, Perna NT: **Analyzing patterns of microbial evolution using the mauve genome alignment system.** *Methods Mol Biol (Clifton, N.J.)* 2007, **396:**135–152.
51. Sokal R, Rohlf F: **The comparison of dendrograms by objective methods.** *Taxon* 1962, **11**(2):33–40.
52. Huelsenbeck JP, Larget B, Swofford D: **A compound poisson process for relaxing the molecular clock.** *Genetics* 2000, **154**(4):1879–1892.
53. Didelot X, Falush D: *Bacterial Recombination in vivo*. Horizontal Gene Transfer in the Evolution of Pathogenesis : Cambridge University Press; 2008.
54. Didelot X: *Sequence-based analysis of bacterial population structure*. Bacterial Population Genetics in Infectious Disease : Wiley Press; 2010.
55. Wiuf C, Hein J: **The coalescent with gene conversion.** *Genetics* 2000, **155:**451–462.
56. Didelot X, Lawson D, Falush D: **SimMLST: simulation of multi-locus sequence typing data under a neutral model.** *Bioinformatics* 2009, **25**(11):1442–1444.
57. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ: **Patterns of gene flow define species of thermophilic Archaea.** *PLoS Biol* 2012, **10**(2):e1001265.
58. Fraser C, Hanage WP, Spratt BG: **Neutral microepidemic evolution of bacterial pathogens.** *Proc Natl Acad Sci U S A* 2005, **102**(6):1968–1973.
59. Schierup MH, Hein J: **Consequences of recombination on traditional phylogenetic analysis.** *Genetics* 2000, **156**(2):879–891.
60. Guttman DS, Dykhuizen DE: **Clonal divergence in Escherichia coli as a result of recombination, not mutation.** *Science* 1994, **266**(5189):1380–1383.
61. Vos M, Didelot X: **A comparison of homologous recombination rates in bacteria and archaea.** *ISME J* 2009, **3**(2):199–208.
62. Fiala KL, Sokal RR: **Factors determining the accuracy of cladogram estimation – evaluation using computer-simulation.** *Evolution* 1985, **39:**609–622.
63. den Bakker, H, Didelot X, Fortes E, Nightingale K, Wiedmann M: **Lineage specific recombination rates and microevolution in Listeria monocytogenes.** *BMC Evolutionary Biol* 2008, **8:**277.
64. Didelot X, Urwin R, Maiden MCJ, Falush D: **Genealogical typing of Neisseria meningitidis.** *Microbiology* 2009, **155:**3176–3186.
65. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y, Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelsonm WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102**(39):13950–13955.
66. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15**(6):589–594.
67. Tettelin H, Riley D, Cattuto C, Medini D: **Comparative genomics: the bacterial pan-genome.** *Curr Opin Microbiol* 2008, **11**(5):472–477.
68. Sims GE, Kim SH: **Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs).** *Proc Natl Acad Sci U S A* 2011, **108**(20):8329–8334.
69. Boerlin P, McEwen SA, Boerlin-Petzold F, Wilson JB, Johnson RP, Gyles CL: **Associations between virulence factors of Shiga toxin-producing Escherichia coli and disease in humans.** *J Clin Microbiol* 1999, **37**(3):497–503.
70. McDaniel TK, Jarvis KG, Donnenberg MS, Kaper JB: **A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens.** *Proc Natl Acad Sci U S A* 1995, **92**(5):1664–1668.
71. Malakooti J, Ely B, Matsumura P: **Molecular characterization, nucleotide sequence, and expression of the fliO, fliP, fliQ, and fliR genes of Escherichia coli.** *J Bacteriol* 1994, **176:**189–197.
72. Wiame E, Delpierre G, Collard F, Van Schaftingen E: **Identification of a pathway for the utilization of the Amadori product fructoselysine in Escherichia coli.** *J Biol Chem* 2002, **277**(45):42523–42529.
73. Weickert MJ, Adhya S: **The galactose regulon of Escherichia coli.** *Mol Microbiol* 1993, **10**(2):245–251.
74. Díaz E, Ferrández A: **Biodegradation of aromatic compounds by Escherichia coli.** *Microbiol Mol Biol Rev* 2001, **65**(4):523–569.
75. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S: **Helicobacter pylori genome evolution during human infection.** *Proc Natl Acad Sci U S A* 2011, **108**(12):5033–5038.
76. Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD: **Interplay of recombination and selection in the genomes of Chlamydia trachomatis.** *Biol Direct* 2011, **6:**28–28.
77. Milkman R, Bridges MM: **Molecular evolution of the Escherichia coli Chromosome. III. Clonal Frames.** *Genetics* 1990, **126:**505–517.
78. Milkman R, Jaeger E: **Molecular evolution of the Escherichia coli chromosome. VI. Two regions of high effective recombination.** *Genetics* 2003, **163**(2):475–483.
79. Keck W, van Leeuwen AM: **Cloning and characterization of mepA, the structural gene of the penicillin-insensitive murein endopeptidase from Escherichia coli.** *Mol Microbiol* 1990, **4**(2):209–219.
80. Iida K: **Mutants of Escherichia coli defective in penicillin-insensitive murein DD-endopeptidase.** *Mol Gen Genet* 1983, **189**(2):215–221.
81. Dougan G, Chatfield S, Pickard D, Bester J: **Construction and characterization of vaccine strains of Salmonella harboring mutations in two different aro genes.** *J Infect Dis* 1988, **158**(6):1329–1335.
82. Foulongne V, Walravens K, Bourg G, Boschiroli ML, Godfroid J: **Aromatic compound-dependent Brucella suis is attenuated in both cultured cells and mouse models.** *Infect Immun* 2001, **69:**547–550.
83. Roberts CW, Leroux MM, Fleming MD, Orkin SH: **Highly penetrant, rapid tumorigenesis through conditional inversion of the tumor suppressor gene Snf5.** *Cancer Cell* 2002, **2**(5):415–425.

84. Tacket CO, Levine MM: **CVD 908, CVD 908-htrA, and CVD 909 live oral typhoid vaccines: a logical progression.** *Clin Infect Dis* 2007, **45**(Suppl 1):20–23.
85. Chatfield SN, Fairweather N, Charles I, Pickard D, Levine M, Hone D, Posada M: **Construction of a genetically defined Salmonella typhi Ty2 aroA, aroC mutant for the engineering of a candidate oral typhoid-tetanus vaccine.** *Vaccine* 1992, **10:**53–60.
86. Tacket CO, Sztein MB, Losonsky GA, Wasserman SS, Nataro JP, Edelman R, Pickard D, Dougan G: **Safety of live oral Salmonella typhi vaccine strains with deletions in htrA and aroC aroD and immune response in humans.** *Infect Immun* 1997, **65**(2):452–456.
87. Khan LA, Khan SA, Al-Hateeti HS, Bhat AR, Bhat KS, Sheikh FS: **Clinical profile and outcome of poisoning in Najran.** *Ann Saudi Med* 2003, **23**(3–4):205–207.
88. Hindle Z, Chatfield SN, Phillimore J, Bentley M, Johnson J, Cosgrove CA, Ghaem-Maghami M, Sexton A, Khan M, Brennan FR, Everest P, Wu T, Pickard D, Holden DW, Dougan G, Griffin GE, House D, Santangelo JD, Khan SA, Shea JE: **Characterization of Salmonella enterica derivatives harboring defined aroC and Salmonella pathogenicity island 2 type III secretion system (ssaV) mutations by immunization of healthy volunteers.** *Infect Immun* 2002, **70**(7):3457–3467.
89. Daley A, Randall R, Darsley M, Choudhry N, Thomas N, Sanderson IR: **Genetically modified enterotoxigenic Escherichia coli vaccines induce mucosal immune responses without inflammation.** *Gut* 2007, **56**(11):1550–1556.
90. Srilunchang T, Proungvitaya T, Wongratanacheewin S: **Construction and characterization of an unmarked aroC deletion mutant of Burkholderia pseudomallei strain A2.** *Southeast Asian J Trop Med Public Health* 2009, **40:**123–130.
91. Xiao J, Chen T, Wang Q, Liu Q, Wang X, Lv Y: **Search for live attenuated vaccine candidate against edwardsiellosis by mutating virulence-related genes of fish pathogen Edwardsiella tarda.** *Lett Appl Microbiol* 2011, **53**(4):430–437.
92. Leopold S, Sawyer S: **Obscured Phylogeny and Recombinational Dormancy in Escherichia coli.** *BMC Evolutionary Biol* 2011, **11:**183.
93. Daegelen P, Studier FW, Lenski RE, Cure S, Kim JF: **Tracing ancestors and relatives of Escherichia coli B , and the derivation of B strains REL606 and BL21(DE3).** *J Mol Biol* 2009, **394**(4):634–643.
94. Fraser C, Hanage W, Spratt B: **Recombination and the nature of bacterial speciation.** *Science* 2007, **315**(5811):476–480.
95. Falush D, Torpdahl M, Didelot X, Conrad DF: **Mismatch induced speciation in Salmonella: model and data.** *Phil Trans R Soc B* 2006, **361:**2045–53.
96. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D, Donnelly P: **Recombination and population structure in Salmonella enterica.** *PLoS Genet* 2011, **7**(7):e1002191.
97. Gordon DM: **The genetic structure of Escherichia coli populations in primary and secondary habitats.** *Microbiology* 2002, **148**(Pt 5):1513–1522.
98. Gordon DM, Cowling A: **The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects.** *Microbiology* 2003, **149**(Pt 12):3575–3586.
99. White AP, Arnold PM, Norvell DC, Ecker E, Fehlings MG: **Pharmacologic management of chronic low back pain: synthesis of the evidence.** *Spine (Phila Pa 1976)* 2011, **36**(Suppl 21):131–143.
100. Nowrouzian FL: **Escherichia coli strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants.** *J Infect Dis* 2005, **191**(7):1078–1083.
101. Vulic M, Dionisio F, Taddei F, Radman M: **Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria.** *Proc Natl Acad Sci U S A* 1997, **94**(18):9763–9767.
102. Majewski J: **Sexual isolation in bacteria.** *FEMS Microbiol Lett* 2001, **199**(2):161–169.
103. Lawrence JG: **Gene transfer in bacteria: speciation without species?** *Theor Popul Biol* 2002, **61**(4):449–460.
104. Retchless AC, Lawrence JG: **Temporal fragmentation of speciation in bacteria.** *Science* 2007, **317**(5841):1093–1096.
105. Lawrence JG, Retchless AC: **The interplay of homologous recombination and horizontal gene transfer in bacterial speciation.** *Methods Mol Biol* 2009, **532:**29–53.
106. Retchless AC, Lawrence JG: **Phylogenetic incongruence arising from fragmented speciation in enteric bacteria.** *Proc Natl Acad Sci U S A* 2010, **107**(25):11453–11458.

doi:10.1186/1471-2164-13-256
Cite this article as: Didelot *et al.*: Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* 2012 **13**:256.