The Evolutionary Path to Extraintestinal Pathogenic, Drug-Resistant *Escherichia coli* Is Marked by Drastic Reduction in Detectable Recombination within the Core Genome

Alan McNally[1,*], Lu Cheng[2], Simon R. Harris[3], and Jukka Corander[2]

[1]Pathogen Research Group, Nottingham Trent University, United Kingdom

[2]Department of Mathematics and Statistics, University of Helsinki, Finland

[3]Pathogen Genomics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom

*Corresponding author: E-mail: alan.mcnally@ntu.ac.uk.

Accepted: March 5, 2013

Abstract

Escherichia coli is a highly diverse group of pathogens ranging from commensal of the intestinal tract, through to intestinal pathogen, and extraintestinal pathogen. Here, we present data on the population diversity of *E. coli*, using Bayesian analysis to identify 13 distinct clusters within the population from multilocus sequence typing data, which map onto a whole-genome-derived phylogeny based on 62 genome sequences. Bayesian analysis of recombination within the core genome identified reduction in detectable core genome recombination as one moves from the commensals, through the intestinal pathogens down to the multidrug-resistant extraintestinal pathogenic clone *E. coli* ST131. Our data show that the emergence of a multidrug-resistant, extraintestinal pathogenic lineage of *E. coli* is marked by substantial reduction in detectable core genome recombination, resulting in a lineage which is phylogenetically distinct and sexually isolated in terms of core genome recombination.

Key words: *E. coli*, recombination, population, diversity.

Introduction

Escherichia coli is a bacterial species of enormous diversity, ranging from a harmless intestinal commensal organism of a myriad of animals, through environmental organism, to zoonotic intestinal pathogen, and causative agent of extraintestinal infections such as urinary tract infection (UTI) and bacteremia (Croxen and Finlay 2010). Classical attempts at classifying *E. coli* have centered on simplistic methods such as pathotype where strains have been classified as commensals, EHEC, EPEC, ETEC, EIEC, EAEC, or ExPEC based on the disease pathology they are most associated with (Kaper et al. 2004). Phylogenetically *E. coli* can be separated into phylogroups based on a small number of discrete genetic markers (Clermont et al. 2000), which show a degree of correlation with isolation from host or niche. More advanced genotyping techniques such as multilocus sequence typing (MLST [Wirth et al. 2006]) have highlighted the shortcomings of pathotype distinction with sequence types (STs) of *E. coli* often spanning pathotypes (Olesen et al. 2012). The higher discriminatory power of MLST also identified more phylogroups within *E. coli,* the borders of which are clouded due to recombination (Wirth et al. 2006).

Indeed, recombination has played a central role in the well-described diversity observed within *E. coli*. The majority of recombination studies have focused on the vast level of horizontal gene transfer and genetic acquisition across *E. coli,* which is often intrinsic to the pathogenic lifestyle of the organism (Dobrindt et al. 2004). Such are the levels of horizontal exchange of mobile genetic elements across *E. coli* that the accessory genome of the organism is essentially open (Rasko et al. 2008). Neither is the genetic exchange of accessory elements limited within subgroups of *E. coli* such as ST or phylogroup with toxigenic phage, pathogenicity islands, and antimicrobial resistance plasmids transcending across all subgroup boundaries as exemplified by the mosaic genomics of the *E. coli* O104 outbreak strain (Rasko et al. 2011). This recombination-derived mosaicism has presented a problem in untangling the population structure of *E. coli* and the evolutionary relationship between the various pathogenic variants. Furthermore, because most studies of recombination in *E. coli*

have focused on the transfer of accessory elements between pathotypes, very little is known on how recombination in the core genome of *E. coli* varies across the population or how that variation is related to pathogenesis or niche. Creating a better understanding of core genome recombination has recently been shown to provide evolutionary insights into the important human pathogens *Enterococcus faecium, Streptococcus pneumonia,* and *Neisseria* spp. (Hanage et al. 2009; Corander et al. 2012; Willems et al. 2012), and it has been shown that differences in levels of recombination across a population are closely linked with ecological factors. Studies based on the diversity of *E. coli* using MLST suggest that recombination has played a key role in the evolution of virulence and the emergence of strains with increased pathogenesis (Wirth et al. 2006), whereas studies based on a limited number of genome sequences have suggested that both homologous and nonhomologous recombination have played a role in evolution of pathogenesis, though there is sexual isolation between phylogroups A/B1, B2, and E (Didelot et al. 2012).

In this study, we utilize algorithms designed for estimating recombination and population structure in large genome data sets, namely BratNextGen (Marttinen et al. 2012) and Bayesian population genetics software (BAPS) (Corander et al. 2008) to analyze the population structure of *E. coli* and determine how recombination correlates with pathogenesis. We analyzed the entire *E. coli* MLST data set (mlst.ucc.ie) and genome data for 62 *E. coli* strains representing the sequenced diversity of the organism. The genomes used range from commensal K12 laboratory strains, to intestinal pathogenic strains, through to strains associated with extraintestinal infections such as UTI, and culminating in a number of strains of *E. coli* ST131. ST131 has emerged over the last decade to become the globally dominant strain type associated with extraintestinal disease and dissemination of multidrug resistance, leading to it being termed the pandemic *E. coli* (Rogers et al. 2011). By utilizing the most comprehensive set of data and analytical tools to date, we provide new insights into recombination and population structure in *E. coli*. Whole-genome phylogeny shows concordance with traditional phylogroups, with advanced Bayesian population analysis of the MLST data set for *E. coli* suggesting the presence of 13 separated population clusters, which exhibit admixture throughout. Detailed analysis of core genome recombination suggests an evolutionary pattern from ubiquitous intestinal commensal exhibiting relatively frequent core genome recombination, through to highly specialized extraintestinal pathogen marked by a drastic decrease in detectable core genome recombination, and in the case of the newly emerged multidrug-resistant ST131, an almost stable core genome that is sexually isolated from the rest of the species including the most closely related phylogroup B2 ExPEC strains. These findings further our understanding of the processes involved in evolution of pathogenesis within the enterobacteriaceae, illustrating how core genome recombination levels correlate to environmental niche and pathogenesis in *E. coli* and provide new avenues of research in understanding the emergence of global pathogens.

Materials and Methods

Genome Data

A total of 62 publically available *E. coli* genome sequences were used in this analysis (table 1). Fifty are reference genome sequences available from NCBI, whereas 12 are ST131 genomes produced during previous studies in our group (Clark et al. 2012).

MLST Data

In an attempt to provide a higher level of resolution to the population, we performed BAPS analysis using the data available on the entire *E. coli* MLST database as of September 2012 (supplementary table S1, Supplementary Material online). The database contained (accessed 1 September 2012) 2,880 STs for which public and nonaberrant allele sequences were available.

Whole-Genome-Based Phylogeny

Genome sequences were aligned using Mugsy (Sahl et al. 2011) and the core genome extracted using Mothur (Schloss et al. 2009) with the default settings of the methods. The resulting alignment was used to determine a maximum likelihood (ML) phylogeny using RAxML (Stamatakis et al. 2005) implementing the rapid bootstrap function and the general time reversible (GTR) model with Gamma correction, with 100 bootstraps performed. The best tree was imported into Figtree (http://tree.bio.ed.ac.uk/software/figtree/, last accessed March 31, 2013) for graphical annotation.

Bayesian Population Genetics Software

BAPS (Tang et al. 2009) was applied to cluster the MLST data into genetically distinct groups and to estimate the level of admixture for each ST. The analyses were performed with the second-order Markov model and the standard MLST data input option, similarly as described in Hanage et al. (2009). The optimal clustering was obtained using 10 runs of the estimation algorithm with the prior upper bound of the number of clusters varying in the range (Corander et al. 2012) over the 10 replicates. Each of the estimation runs did yield a highly congruent partition of the ST data compared with the other runs, such that there were exactly 13 clusters, indicating a highly peaked posterior distribution in the neighborhood of these partitions (estimated posterior probability 1.000). The admixture analysis was subsequently performed using the 13 clusters in the estimated posterior mode partition with 100 Monte Carlo replicates for allele frequencies and by generating 100 reference genotypes to calculate *P* values. For reference cases, we used 10 iterations in estimation according

Table 1
List of *Escherichia coli* Genome Data Used in This Study

Strain	ST	BAPS Cluster	Pathotype	Accession Number
E. coli CE10	62	8	K1 ExPEC	NC_017646
E. coli DH1 –ECDH1	1,060	5	K12	CP001637.1
E. coli ME8659	1,060	5	K12	AP012030.1
E. coli DH10B	1,060	5	K12	NC_010473.1
E. coli W3110	10	5	K12	AC000091
E. coli MG1655	10	5	K12	U00096.2
E. coli BW2952	10	5	K12	CP001396.1
E. coli P12b	10	5	K12	NC_017663.1
E. coli H10407	48	5	ETEC	FN649414
E. coli UMNK88	100	5	K88	CP002729.1
E. coli REL606	93	5	B	CP000819.1
E. coli BL21 DE3	93	5	B	CP001509.3
E. coli ATCC9637	1,079	3	W	CP002185.1
E. coli SE11	156	3	Human commensal	AP009240.1
E. coli E23477A	1,132	3	ETEC	CP000800.1
E. coli IAI1	1,128	3	O:8	CU928160.2
E. coli 55989	678	1	EAEC	CU928145.2
E. coli C227_11	678	1	O104	AFRH00000000
E. coli 12009	17	1	O103 EHEC	AP010958.1
E. coli 11128	16	3	O111 EHEC	AP010960.1
E. coli 11368	21	3	O26 EHEC	AP010953.1
E. coli ATCC8739	1,120	5	K12	CP000946.1
E. coli HS	46	7	Human commensal	CP000802.1
E. coli CB9615	335	9	O55 EHEC	CP001846.1
E. coli EDL933	11	9	O157 EHEC	AE005174.2
E. coli Sakai	11	9	O157 EHEC	BA000007.2
E. coli TW14359	11	9	O157 EHEC	CP001368.1
E. coli EC4115	11	9	O157 EHEC	NC_011353.1
E. coli XuZhou21	11	9	O157 EHEC	NC_017906.1
E. coli RM12579	335	9	O55 EHEC	NC_017656.1
E. coli O42	414	6	EAEC	FN554766.1
E. coli UMNO26	597	6	O:7 ExPEC	CU928163.2
E. coli SMS35	354	8	Multidrug resistant	CP000970.1
E. coli E2348/69	15	4	O127 EPEC	FM180568.1
E. coli UTI18	131[a]	4	ExPEC	ERP001095
E. coli EC958	131[b]	4	ExPEC	CAFL01000001
E. coli NA114	131[c]	4	ExPEC	CP002797.1
E. coli P2U	131[d]	4	ExPEC	ERX159100
E. coli P5U	131[d]	4	ExPEC	ERX159106
E. coli P2B	131[d]	4	ExPEC	ERX159099
E. coli UTI24	131[a]	4	ExPEC	ERP001095
E. coli UTI32	131[a]	4	ExPEC	ERP001095
E. coli UTI62	131[a]	4	ExPEC	ERP001095
E. coli UTI188	131[a]	4	ExPEC	ERP001095
E. coli UTI226	131[a]	4	ExPEC	ERP001095
E. coli UTI306	131[a]	4	ExPEC	ERP001095
E. coli UTI423	131[a]	4	ExPEC	ERP001095
E. coli UTI587	131[a]	4	ExPEC	ERP001095
E. coli SE15	131[a]	4	Human commensal	AP009378.1
E. coli LF82	135	4	AIEC	NC_011993.1
E. coli IHE3034	95	4	ST95 ExPEC	CP001969.1
E. coli UTI89	95	4	ST95 ExPEC	CP000243.1
E. coli S88	95	4	O45 ExPEC	CU928161.2
E. coli APEC01	95	4	APEC	CP000468.1
E. coli UM146	643	4	AIEC	CP002167.1
E. coli 536	127	4	O6 ExPEC	CP000247.1
E. coli LF82	135	4	AIEC	CU651637.1
E. coli NRG857c	135	4	AIEC	CP001855.1
E. coli ED1a	452	4	O81	CU928162.2
ABU83972	73	4	Asymptomatic	CP001671
E. coli CFT073	73	4	ExPEC	AE014075.1
E. coli Di14	73	4	ExPEC	CP002212.1
E. coli Di12	73	4	ExPEC	CP002211.1

[a]2009 UK ST131 isolates.
[b]2004 UK ST131 Isolate.
[c]Indian ST131 isolate.
[d]2012 UK ST131 isolates.

to the guidelines of Corander and Marttinen (2006). STs were concluded to be significantly admixed when the *P* value did not exceed the threshold of 5%.

MLST-Based Phylogeny

The phylogenetic distribution of the BAPS clusters was determined using an ML tree estimated with FastTree (Price et al. 2009) using the default settings (1,000 bootstrap replicates with the general time-reversible model and Gamma model for rate heterogeneity) based on the concatenated MLST data over the seven loci for all identified MLST STs.

BratNextGen

Software package BratNextGen (Marttinen et al. 2012) was used to determine recombining regions in the whole-genome data comprising the 62 sequences. The estimation was performed with the default settings as in Marttinen et al. (2012) using 10 iterations of the estimation algorithm, which was assessed to be sufficient because changes in the hidden Markov model parameters were already negligible over the last couple of iterations. Significance of a recombining region was determined as in Marttinen et al. (2012) using a permutation test with 100 permutations executed in parallel on a cluster computer (threshold of 5% was used to conclude significance for each region).

Statistical Testing of Significance in Differences of Recombination Levels

To investigate whether the observed differences in the amount of recombination are accountable by random variation within and between lineages, we performed standard permutation tests. For a given labeling of strains into two groups, we calculated first the absolute value of the difference in the mean estimated amount of recombination between the two groups. This resulted in the observed statistic T_obs. Then, under the null hypothesis of no systematic difference in recombination between the groups, the group labels were randomly permuted among strains and the corresponding value of the statistic T_perm was calculated. To obtain a *P* value for T_obs under the null hypothesis, the permutation procedure was repeated 1,000,000 times, which yields an estimate of the probability $P(T\_perm > T\_obs)$.

Phylogeny of Recombining Regions and Core Genomes with Recombining Regions Removed

FastTree was used with the same settings as for the MLST to determine the phylogeny of significantly recombining regions and core genomes with significantly recombining regions removed. For each group of strains defined by the MLST BAPS clustering, the fractions of recombinations shared within and between groups were determined from the BratNextGen output. In addition, the *r/m* ratio was estimated for each

group from the number of single-nucleotide polymorphisms (SNPs) residing within and outside of the recombinant segments.

Results

Whole-Genome Phylogeny of *E. coli* and Determination of Population Structure from MLST Data

To determine the phylogeny of the 62 *E. coli* strains in our analysis (table 1), a whole-genome alignment was performed using Mugsy and the core genome extracted to infer phylogeny using RaxML. The core genome is composed of 2,336,639 bp and shows concordance with previous *E. coli* whole-genome phylogenies (Rasko et al. 2011) with hierarchical clustering based around Phylogroup (fig. 1). The *E. coli* ST131 strains also show very tight clustering in concordance with previous findings (Clark et al. 2012), though additionally this phylogeny places strains from the United Kingdom and India isolated between 2004 and 2010 (Avasthi et al. 2011; Totsika et al. 2011; Clark et al. 2012) in a monophyletic clade exhibiting very low diversity, furthering the suggestion that ST131 may be a globally disseminated clone. However, there is also the formation of a second cluster of *E. coli* ST131 strains comprising the reference genome strain SE15 and two additional strains isolated in the United Kingdom in 2012 none of which exhibit antimicrobial resistance.

In an attempt to provide a higher level of resolution to the population, we performed BAPS analysis using the data available on the entire *E. coli* MLST database (mlst.ucc.ie) as of 1 September 2012 (supplementary table S1, Supplementary Material online), which resulted in the identification of 13 BAPS clusters. The database contained 2,880 STs for which public and nonaberrant allele sequences were available. The phylogenetic distribution of the clusters was determined using an approximate ML tree estimated with FastTree (fig. 2). This shows that recombination has blurred the boundaries of lineages to a considerable degree but not uniformly over all the lineages. Notably, apart from a small subset of STs within BAPS cluster 4, this cluster forms a monophyletic clade. Mapping of BAPS clusters onto the whole-genome phylogeny (fig. 1 and supplementary fig. S1, Supplementary Material online) identified BAPS cluster 4 isolates as all belonging to phylogroup B2 and all being ExPEC strains with the exception of E234869, which is the reference O127 EPEC strain. All K12 strains are contained within BAPS cluster 5, except HS that belongs to BAPS cluster 7, and all O55 and O157 EHEC strains are within BAPS cluster 9. The phylogroup B1 clade contains two discrete BAPS clusters within it. The majority are within BAPS cluster 3 except the two EAEC strains 55989 and C227_11 (*E. coli* O104) and strain 12009 (O103 EHEC), which are in BAPS cluster 1. This population grouping confirms that pathotypes are not a robust way to differentiate *E. coli* and that phylogroups can also be distributed across the population. Our data provide a population framework to MLST supporting 13 distinct populations and in particular a clearly distinct BAPS cluster 4 containing only phylogroup B2 extraintestinal pathogenic STs. Determination of the levels of admixture across the BAPS clusters based on MLST data (fig. 3) supports the idea of discrete clusters but with significant recombination. A summary of the admixture results is given in supplementary table S2, Supplementary Material online. Notably, BAPS MLST cluster 4 is the sole cluster harboring STs from a single ancestral group only, B2 (supplementary table S4, Supplementary Material online). For a majority of the BAPS clusters, several ancestral groups are found within a single cluster. Among all clusters with at least 50 STs assigned to allow for more robust estimation of subpopulation characteristics, the frequency of admixed STs is smallest in cluster 4, and furthermore, the mean fraction of DNA atypical for the cluster is also smallest.

Quantification of Recombination with BratNextGen Shows Varied Recombination Frequencies across BAPS Clusters

To further examine the level of recombination across the BAPS clusters, we interrogated our data set comprising 62 aligned whole genomes using BratNextGen (fig. 4). The results clearly show an uneven level of recombination across the BAPS clusters with the *E. coli* ST131 clade within BAPS cluster 4 displaying very little recombination with an average value of just 0.39% of the core genome undergoing recombination. There is then an increase in recombination moving into the remaining BAPS cluster 4 ExPEC isolates and BAPS cluster 9 EHEC strains, and onward into the BAPS cluster 1, 3, 6, 7, and 8 strains associated with intestinal disease, culminating in the BAPS cluster 5 K12 strains exhibiting the highest level of recombination at an average 2.19% of the core genome. Quantitative analysis of differences between the recombination levels of the BAPS clusters was performed using permutation tests (fig. 5), which showed that the resistant ST131 strains within BAPS cluster 4 had significantly less recombination than the other strains with multiple isolates present ST73, ST95, and ST135 ($P = 0.000001$) and that the K12 strains of BAPS cluster 5 (ST10, ST1060) exhibited significantly higher levels of core genome recombination than strains in ST73, ST95, and ST135 ($P = 0.00001$). This is indicative of a sliding scale of core genome recombination, starting with high levels in the commensal K12 strains, reducing into the intestinal pathogenic strains, further still into the highly virulent intestinal pathogenic EHEC and the extraintestinal pathogenic strains, culminating in the pandemic, multidrug-resistant ST131 extraintestinal pathogenic strains. To ensure that these differences in core genome recombination cannot be explained away by reduced level of genome diversity within any particular MLST lineage included in the comparisons, we calculated from the nonrecombinant genome segments

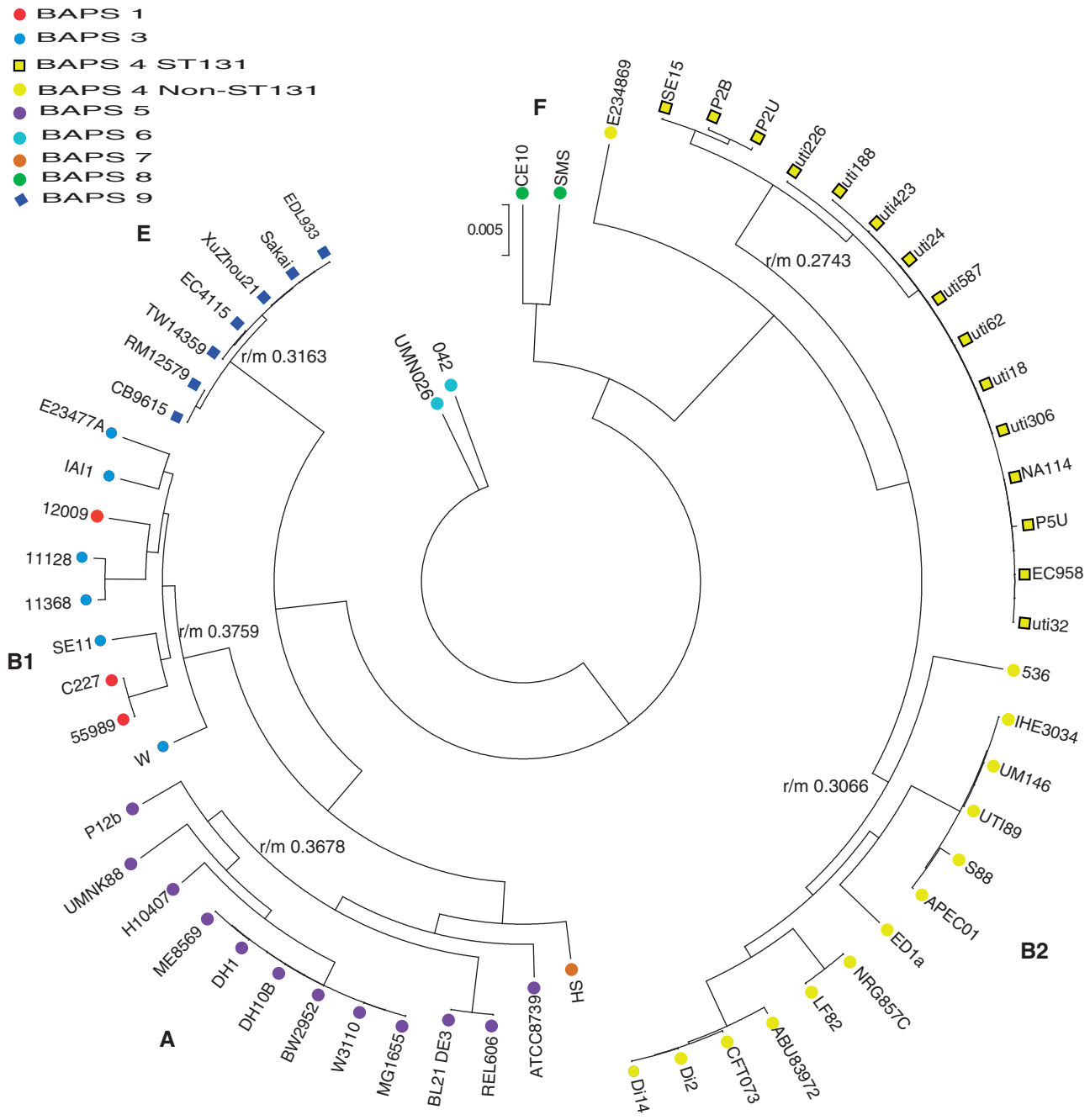

Fig. 1.—Whole-genome-based phylogeny of *Escherichia coli*. The phylogeny is based on approximately 2.3-Mbp core genome aligned using Mugsy, with an ML phylogeny determined using RAxML. Classical phylogroup is indicated at each clade of the tree with the appropriate capital letter. Strains are further color coded according to allocated BAPS cluster (yellow = BAPS 4, ST131 isolates with black border; brown = BAPS 7; purple = BAPS 5; light blue = BAPS 3; red = BAPS 1; dark blue = BAPS 9; cyan = BAPS 6; green = BAPS 8). The calculated *r/m* value for each major clade is also presented.

relative SNP distances between all pairs of isolates within each lineage (ST73, ST95, ST135, and resistant ST131). Two-sample Kolmogorov–Smirnov (K-S) test was used to examine whether the distribution of SNP distances was markedly different between any pair of these lineages. Notably, mean relative SNP distance between two isolates was highest within the resistant ST131 among these four MLST lineages, thus indicating that the identified reduction in core genome recombination cannot be plausibly explained by the lineage being "younger" and less diverse than the other lineages. K-S test yielded a nonsignificant result when the distributions of relative SNP distances were compared between resistant ST131 and

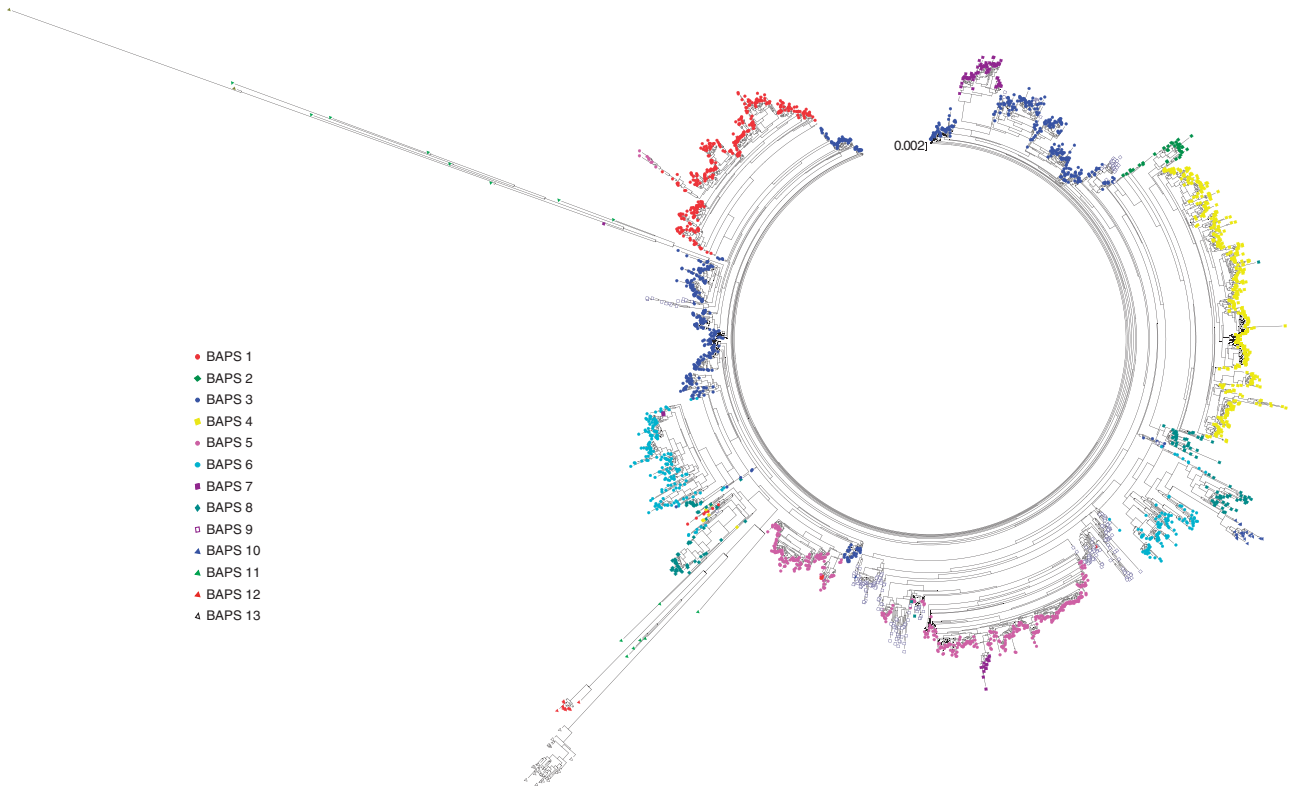

Fig. 2.—ML phylogeny of the *Escherichia coli* population based on concatenated MLST data. Each taxa in the tree is a different ST present in the *E. coli* MLST database and is color coded by its allocated BAPS cluster.

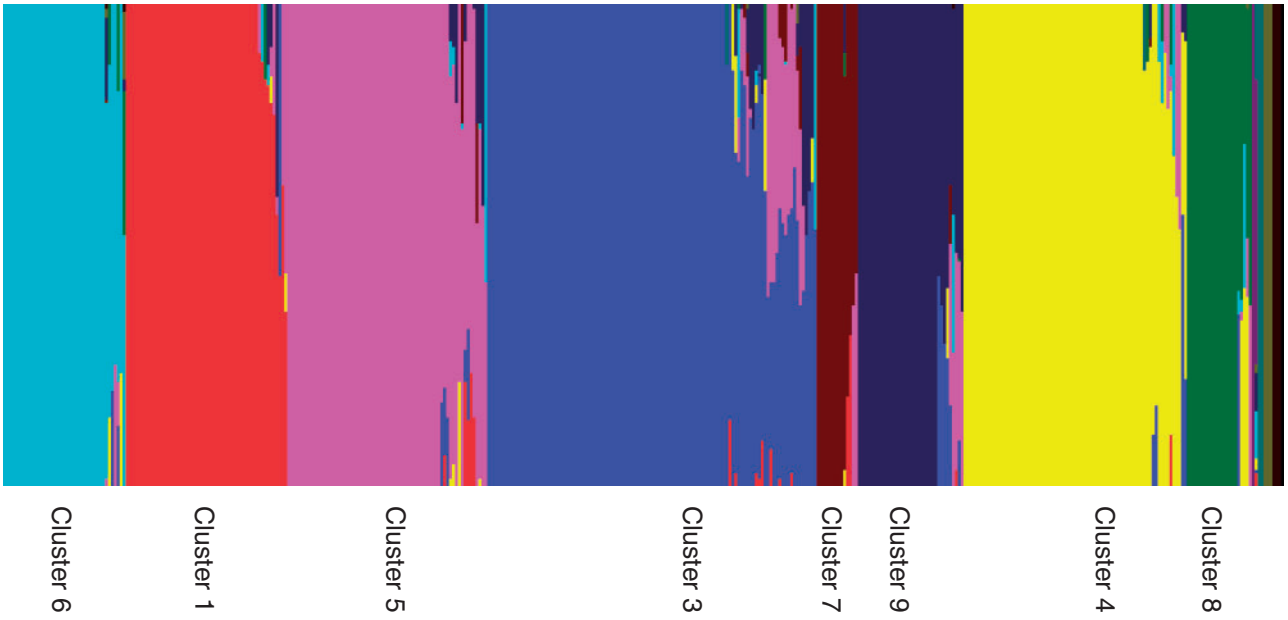

Fig. 3.—Graphical representation of genetic admixture between the *Escherichia coli* BAPS population clusters as determined by BAPS analysis of MLST data. The colors on the fringes of each cluster denote introgression of DNA from that source cluster into the recipient cluster.

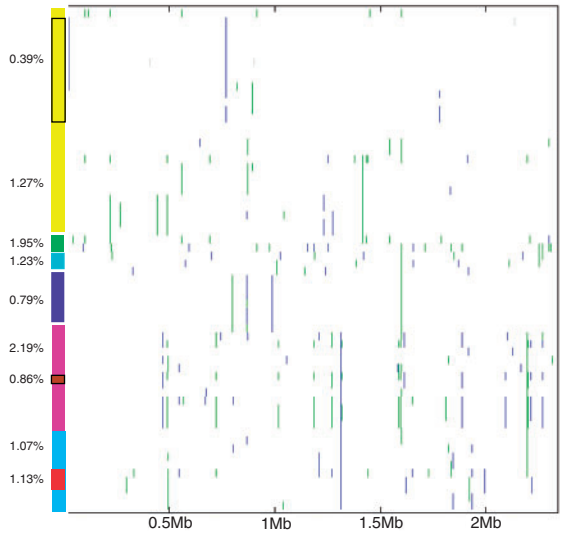

Fig. 4.—Graphical representation of recombination across 62 *Escherichia coli* genome sequences as determined by BRATNextGen analysis. Colored bars on the left indicate the BAPS cluster to which each strains in the analysis belongs to (yellow = BAPS 4; brown = BAPS 7; purple = BAPS 5; light blue = BAPS 3; red = BAPS 1; dark blue = BAPS 9; cyan = BAPS 6; green = BAPS 8). Each strain in the analysis is a dash on the *y* axis of the diagram. The *x* axis is marked by base pair position relative to the core genome pseudomolecule formed from the whole-genome alignment. Bars in the diagram represent regions of recombination detected within the core genome of each strain, with the color coding of the bars allocated in an arbitrary manner. The figure to the left of BAPS color indicator denotes the average percentage of recombination in the core genome for that lineage.

ST73, ST95 ($P = 1.000$), whereas significant difference was observed between ST135 and the other lineages ($P = 0.0001$).

Our BratNextGen analysis also appears to suggest that the recombining regions in each BAPS cluster are largely specific to that particular cluster and that there is a degree of sexual isolation between the clusters. This has been suggested before with respect to phylogroups A + B1, B2, and E (Didelot et al. 2012), but our analysis indicates this could be occurring at a level beyond phylogroups. To determine whether core genome recombination is sexually limited across BAPS clusters, we extracted the recombining regions from our core genome data set and inferred phylogeny from them using FastTree (fig. 6). The resulting phylogeny mirrors that of the entire core genome supporting the suggestion that there is no significant recombination between BAPS clusters at the core genome level. To confirm the phylogeny, we calculated the proportion of recombinant segments in each BAPS cluster (including a separate calculation for ST131), which were intracluster specific, and the proportion of recombinant segments, which were shared across clusters, or intercluster recombination events. The resulting data clearly support the theory of recombination being favored to occur between strains from the same cluster within BAPS cluster 4 and 9, with ST131 showing a large bias toward intra-ST131 recombination. Together our data provide extra insights into the sexual isolation of the ST131 group, which phylogenetically is very tightly clustered and at a further distance from all other strains than at a whole-genome level.

We examined the areas of the core genome in which recombination was detected in each BAPS cluster by mapping the regions onto an annotated pseudomolecule of the core *E. coli* genome, highlighting the presence of the majority of recombination events in CDS, though with some intergenic regions also recombining. There was no physical clustering of recombination in distinct regions of the chromosome, which may be suggestive of hotspots or multiple insertions via a single recombination event in any of the 62 strains. Similarly, there was no association with recombination in any functional category of gene in any BAPS cluster nor was there any association with a particular gene in any cluster, which might infer some obvious biological relevance regarding niche or pathogenesis. This is in contrast to a recent article suggesting recombination hotspots in the *rfb* operon, *fimA*, and the *aroC* locus (Didelot et al. 2012); however, we make no comment on the validity of these findings. Indeed, further analysis of all the recombinant regions across all 62 genomes is currently the focus of a significant body of work, in the hope it may add to our hypothesis on the role of ecology in defining the recombination patterns described here.

Discussion

Escherichia coli is a highly diverse organism, which ranges from the intestinal commensal K12 strains, through to intestinal pathogenic variants such as ETEC, EAEC, and EPEC, to severe intestinal pathogens such as *E. coli* O157, and then extraintestinal pathogenic variants causing UTIs and bacteremia. This enormous diversity has made *E. coli* the subject of countless comparative and evolutionary studies attempting to determine the mechanisms by which each of the subgroups has diversified and specialized (Dykhuizen and Green 1991; Touchon et al. 2009). The recent emergence of the multidrug-resistant *E. coli* O:25b:H4 ST131 as the globally dominant strain type isolated from extraintestinal infections (Peirano and Pitout 2010) provides a new and important dimension to the study of how *E. coli* evolves and diversifies. ST131 strains first came to prominence in the early 21st century as the strain type responsible for drug-resistant outbreaks of community acquired bacteremia (Lau et al. 2008; Nicolas-Chanoine et al. 2008; Ender et al. 2009; Pitout et al. 2009; Rooney et al. 2009) since when it has become the dominant strain type associated with UTIs (Johnson et al. 2010; Croxall et al. 2011; Platell et al. 2011; Zong and Yu 2011). Genome sequence data are suggestive of global dissemination of a highly successful clone of ST131 (Clark et al. 2012), unlike other recently emerged globally successful *E. coli* such as

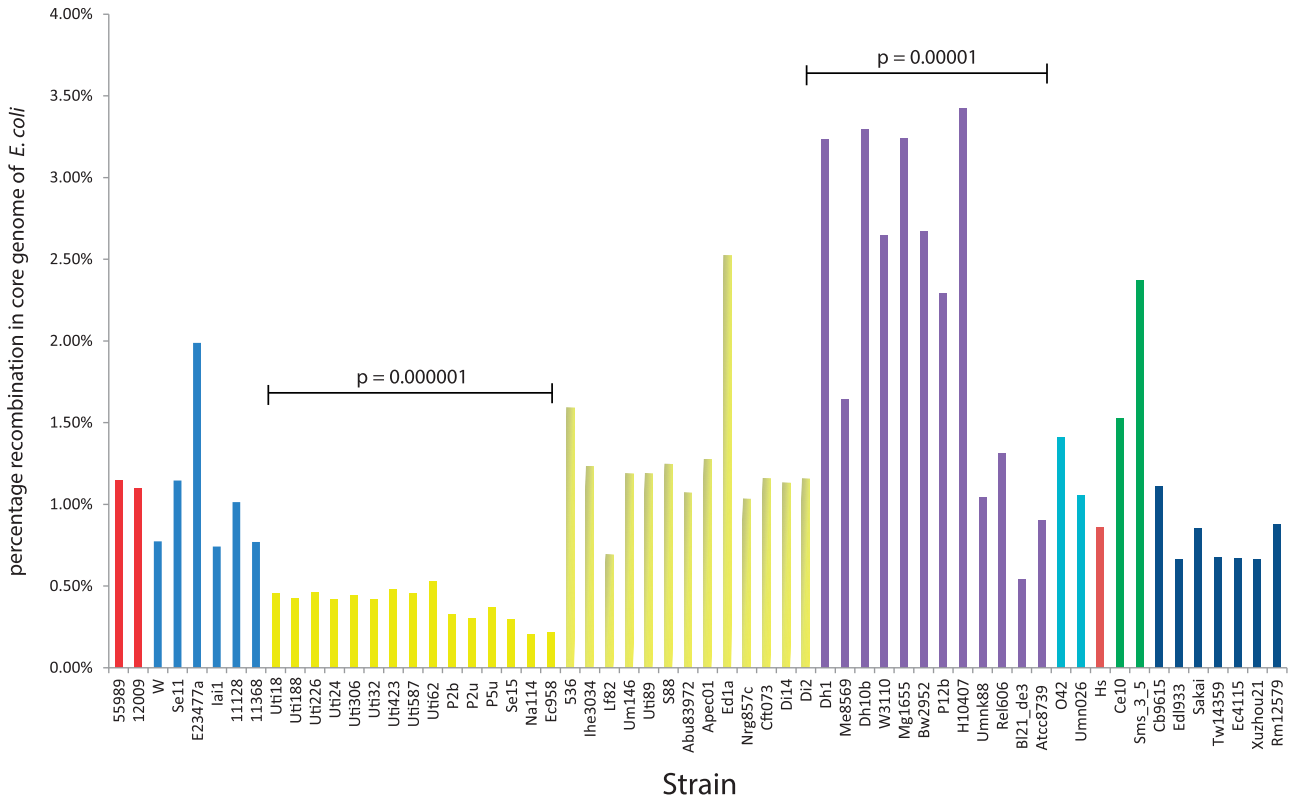

Fig. 5.—Graph showing the percentage of core genome undergoing recombination in each of the 62 genomes in our analysis. Bars are color coded according to the BAPS cluster that strain is allocated to (yellow = BAPS 4, light yellow ST131 strains within BAPS 4; brown = BAPS 7; purple = BAPS 5; light blue = BAPS 3; red = BAPS 1; dark blue = BAPS 9; cyan = BAPS 6; green = BAPS 8). Bars with values above the ST131 and K12 strains indicate *P* values for significant difference between that group of strains and all others as determined by standard permutation tests.

E. coli O157, which is a diverse population of organisms (Bono et al. 2012). Also unlike O157, the dominant emergence of ST131 does not seem to be linked to any increased virulence phenotype (Johnson et al. 2012) or any set of specific or unique genetic loci (Clark et al. 2012) other than dissemination of CTX-M 15.

Our reconstruction of the whole-genome-informed phylogeny of *E. coli* is in good agreement with previously published phylogenies (Rasko et al. 2008, 2011; Lukjancenko et al. 2010; Didelot et al. 2012) but is the first to contain all the ST131 genomes sequenced to date (Avasthi et al. 2011; Totsika et al. 2011; Clark et al. 2012), as well as three new genome sequences isolated in the United Kingdom in 2012 (P5U, P2U, and P2B) and the strain SE15 that is published as an O150 strain isolated as a human commensal (Toh et al. 2010) but which STs as an ST131 using the MLST scheme. The phylogenetic tree shows ST131 are clustered within the phylogroup B2 ExPEC strains as expected and form two distinct groups, the first containing SE15 and two of the new UK 2012 isolates. The second contains all the O:25b and multidrug-resistant isolates spanning an 8-year period and from the United Kingdom and India and shows very little diversity, in concordance with earlier data from our group (Clark et al. 2012).

Phylogeny alone provides very little detail as to how distinct lineages of *E. coli* arose or indeed what distinct lineages are beyond the classical phylogrouping. Our BAPS analysis based on all publicly available MLST data suggests the presence of 13 distinct population clusters within the *E. coli* population. These clusters are remarkably well in agreement with the lineages detected in the core genome phylogeny (fig. 1), and when the concatenated MLST phylogeny and resulting BAPS groupings for the whole-genome sequence strains is compared with the whole-genome phylogeny, there is direct concordance (supplementary fig. S1, Supplementary Material online). The major exception to this agreement is found within the phylogroup B1 where the genome-wide information intermixes strains from the BAPS clusters 1 and 3, possibly due to that fact that both these clusters are human intestinal pathogens, which also colonize the intestinal tracts of livestock, therefore sharing ecological niche. ST131 is found in BAPS cluster 4, which contains all the phylogroup B2 ExPEC strains in our phylogeny and also displays the lowest levels of admixture across all the *E. coli* populations. All other classical phylogroups are disseminated across multiple BAPS clusters. This would suggest BAPS cluster 4 is a population of more "clonal" strains of *E. coli* that are linked by their association with

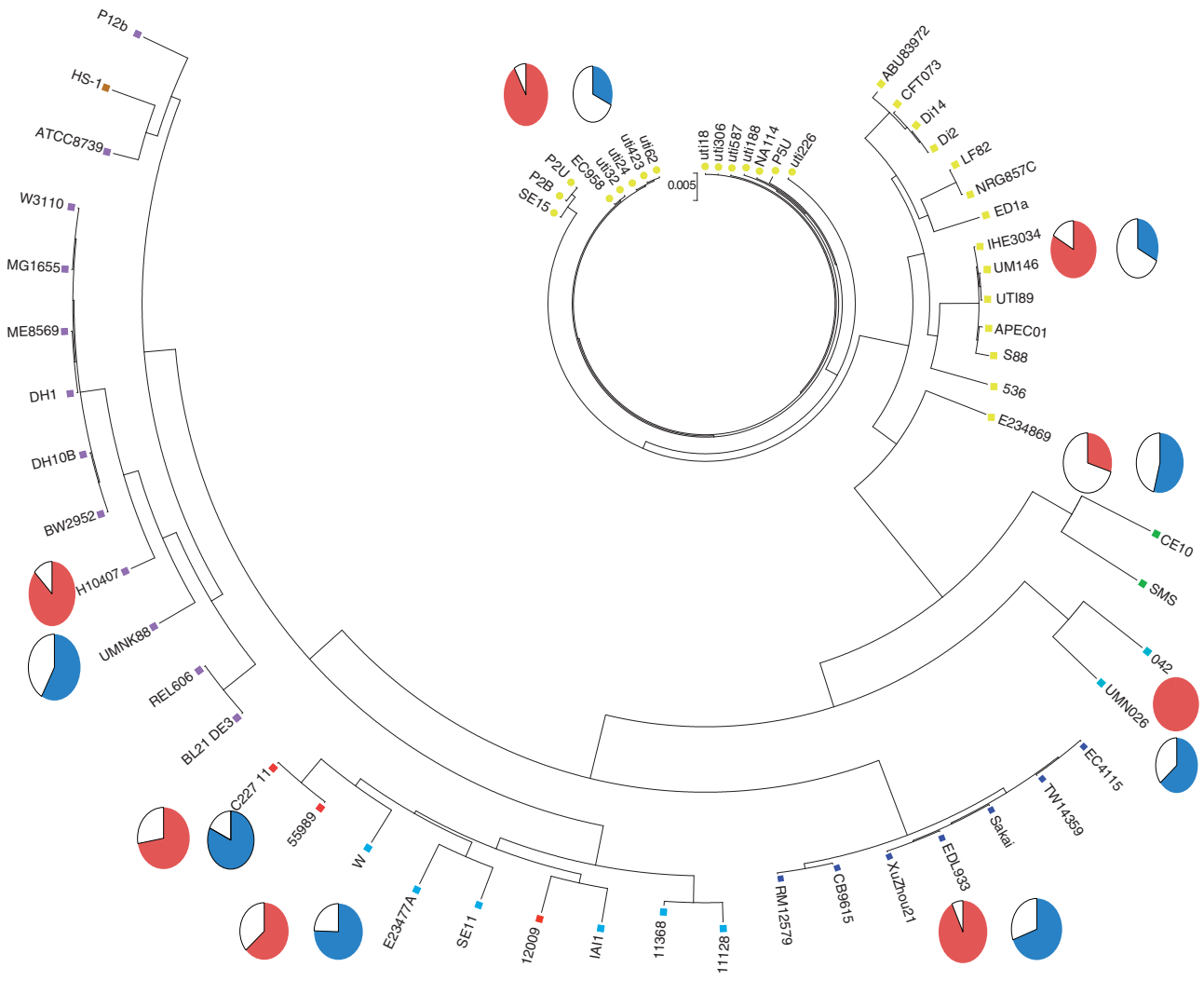

Fig. 6.—ML phylogeny of the detected recombining regions in each genome. The recombinant regions as determined by BRATNextGen were extracted and concatenated for each genome before being aligned and interrogated using RAxML. Strains are color coded according to their allocated BAPS cluster (yellow = BAPS 4; brown = BAPS 7; purple = BAPS 5; light blue = BAPS 3; red = BAPS 1; dark blue = BAPS 9; cyan = BAPS 6; green = BAPS 8). Pie charts next to clades indicate the proportion of recombining regions, which are intra-BAPS cluster specific (red in white) and inter-BAPS cluster recombination events (blue in white).

extraintestinal infection and that is marked by a reduction in admixture events from outwith the subpopulation. To confirm this further, we conducted a fuller investigation of recombination at the whole-genome level using BRATNextGen. The whole-genome recombination analysis clearly shows the ST131 strains within BAPS cluster 4 having a marked decrease in the level of detectable core genome recombination when compared with the other *E. coli* in our analysis, though the levels in the non-ST131 BAPS cluster 4 strains are similar to those observed in the intestinal pathogens with the exception of BAPS cluster 9 containing *E. coli* O157 and its direct ancestral relative *E. coli* O55 (Leopold et al. 2009), which display the next lowest levels of detectable core recombination after ST131. The highest levels of core genome recombination are found in K12 strains located in BAPS cluster 5, which is perhaps unsurprising given many of these are derivatized laboratory strains and intestinal commensals. The finding of reduction in recombination associated with increased pathogenesis in *E. coli* is in direct contrast to the findings of Wirth et al. (2006), which utilized MLST data and population genetic analysis on the data set to infer that increased virulence in *E. coli* was a result of increased recombination. The discrepancy between our study and theirs may be due to level of resolution afforded by the analysis of dozens of complete and draft genome sequences, as well as the Bayesian analysis programs utilized in our study, providing a detailed analysis of recombination across entire core genomes rather than a very small subset of selected genes as in MLST. This argument is

strengthened by the fact that our data presented here are in agreement with that of Didelot et al. (2012) in that recombination patterns indicate a trend toward sexual isolation in phylogroup B2.

Our finding of reduction in recombination of the core genome in the globally disseminated, multidrug-resistant ST131 clone is in direct agreement with recent findings in other pathogens. The most striking analogy is with hospital-associated strains of *Ent. faecium* displaying increased virulence and antimicrobial drug resistance, which have arisen through limited recombination events, followed by a marked decrease in detectable core genome recombination leading to clonal expansion of a successful variant (Willems et al. 2012). The successful spread of such a clonal lineage could be due to a selective advantage for strains, which avoid the loss of advantageous phenotypes via recombination. Another observation from our data is how the recombination that is detected is primarily limited to within BAPS clusters in *E. coli*, which is seen most clearly when a phylogeny is derived from the recombining regions alone, mirroring the core genome phylogeny, and the levels of intracluster recombination are determined. This would support a similar pattern of recombination in evolution of *E. coli* as that described in the hospital-associated *Ent. faecium*, with recombination being restricted as a pathogenic clone becomes successful and more niche adapted. As such in *E. coli,* we observe high levels of recombination in the intestinal commensal K12, reducing through the intestinal pathogens, and then restriction and sexual isolation as the dominant ST131 clone emerges as an extraintestinal pathogen with high levels of multidrug resistance. Of course such restriction is limited to the core genome as it is well known that accessory genome elements such as plasmids and phage easily transfer across our determined BAPS clusters as is seen with the CTXM-15 plasmid and the shiga-toxin-like phage crossing into the *E. coli* O104 strain (Rasko et al. 2011). However, transfer of pathogenicity islands may actually also be restricted if one considers that the islands associated with uropathogenicity are only found in phylogroup B2 *E. coli* (Lloyd et al. 2007, 2009). Given that ST131 is a classical phylogroup B2 strain containing classical ExPEC-specific virulence factors and that the main difference between ST131 and its near BAPS cluster 4 neighbors is extended dissemination of the CTX-M-15 ESBL, it seems most likely that ST131 has specialized via horizontal gene transfer to become a multidrug-resistant ExPEC and then recombined less due to ecological or genetic factors. The emergence of this strain as a highly successful and dominant lineage would then most probably be as a result of selection through drug resistance.

When one considers the reasons for sexual restriction of admixture in *E. coli* ST131, then there are two main plausible discussion points. The first would be mechanistic barriers to cross BAPS cluster admixture via physical prevention or a gradual increase in genetic incompatibility as there is a selective advantage to reduce recombination and limit loss of advantageous genes. The second would be ecological isolation by reduced opportunity to meet and recombine with strains of other BAPS clusters. Fully addressing this question requires further focused research, particularly in furthering our inadequate understanding of the ecology of human pathogenic *E. coli* and especially ExPEC. It has been reported that ST131 is found in companion animals (Ewers et al. 2010; Platell, Johnson, et al. 2011; Platell et al. 2011) and in poultry food products (Vincent et al. 2010); however, the actual route of dissemination and transmission is still not known for certain with some studies suggesting UTI caused by *E. coli* may sometimes be transmitted as a sexually transmitted infection (Foxman 2010) as opposed to the classical fecal–urethral route considered clinical dogma. A favoring for the role of ecological separation leading to the detected recombination pattern in *E. coli* may be taken when one considers the cluster with the next lowest level of detectable core recombination to ST131, that of BAPS cluster 9. This cluster contains *E. coli* O157 strains and their ancestral O55 relatives. *Escherichia coli* O157 is a pathogenic variant, which has become globally successful. Moreover, *E. coli* O157 is ecologically distinct in that it causes acute infections in humans leading to a transient and brief colonization of the intestinal tract and is found to only colonize the recto-anal junction of livestock as opposed to the more microbially rich intestinal lumen (Naylor et al. 2003). This would suggest that the lineage has less opportunity to recombine with distant *E. coli* lineages due to reduced opportunity to interact in the mammalian intestinal tract. Similarly, the recent observations on the phylogeography of *Shigella* also fit with our postulated model of HGT-driven divergence and then ecological separation-driven reduction in recombination (Holt et al. 2012). *Shigella* are a subset of *E. coli,* which have become niche restricted to the human intestinal tract where they are highly pathogenic, and display extreme monomorphism across a large population set, with little to no recombination. Only by better understanding the ecology of ST131 would we be able to draw fuller inferences and comparisons.

In conclusion, we present novel data on the population structure of *E. coli*. Recombination analysis on this scale has not been previously undertaken for *E. coli* due to its computational complexity, and advanced model-based tools are the key to unraveling the mysteries of complex evolutionary processes. Our data suggest that the emergence of new pathogenic and drug-resistant lineages of *E. coli* are marked by reduction in detectable recombination within the core genome, which is typified by analysis of the *E. coli* ST131 clone. Our data raise new questions on the evolution and emergence of new pathogenic *E. coli* variants and opens up new avenues of research to further our understanding of the ecology and interactions of extraintestinal pathogenic *E. coli*

Supplementary Material

Supplementary figure S1, tables S1–S6, and file S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org).

Acknowledgments

The authors thank Prof Nicholas Thomson for helpful comments on the manuscript. This work was supported by ERC grant no. 239784, Academy of Finland grant no. 251170, and a grant from Sigrid Juselius Foundation to J.C.; by population genetics graduate school and a grant from Sigrid Juselius Foundation to L.C.; and by grant no CRD-2-1 from EMDA iNET sponsored by EU to A.M.

Literature Cited

Avasthi TS, et al. 2011. Genome of multidrug-resistant uropathogenic *Escherichia coli* strain NA114 from India. J Bacteriol. 193:4272–4273.

Bono JL, et al. 2012. Phylogeny of shiga toxin-producing *Escherichia coli* O157 isolated from cattle and clinically ill humans. Mol Biol Evol. 29: 2047–2062.

Clark G, et al. 2012. Genomic and molecular epidemiology analysis of clinical *Escherichia coli* ST131 isolates suggests circulation of a genetically monomorphic but phenotypically heterogeneous ExPEC clone. J Antimicrob Chemother. 67:868–877.

Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl Environ Microbiol. 66: 4555–4558.

Corander J, Connor TR, O'Dwyer CA, Kroll JS, Hanage WP. 2012. Population structure in the *Neisseria*, and the biological significance of fuzzy species. J R Soc Interface. 9:1208–1215.

Corander J, Marttinen P. 2006. Bayesian identification of admixture events using multi-locus molecular markers. Mol Ecol. 15:2833–2843.

Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 16:539.

Croxall G, et al. 2011. Molecular epidemiology of extraintestinal pathogenic *Escherichia coli* isolates from a regional cohort of elderly patients highlights the prevalence of ST131 strains with increased antimicrobial resistance in both community and hospital care settings. J Antimicrob Chemother. 66:2501–2508.

Croxen M, Finlay B. 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. Nat Rev Microbiol. 8:26–38.

Didelot X, Meric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13:256.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2: 414–424.

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

Ender PT, et al. 2009. Transmission of an extended-spectrum-beta-lactamase-producing *Escherichia coli* (sequence type ST131) strain between a father and daughter resulting in septic shock and emphysematous pyelonephritis. J Clin Microbiol. 47:3780–3782.

Ewers C, et al. 2010. Emergence of human pandemic O25:H4-ST131 CTX-M-15 extended-spectrum-beta-lactamase-producing *Escherichia coli* among companion animals. J Antimicrob Chemother. 65:651–660.

Foxman B. 2010. The epidemiology of urinary tract infection. Nat Rev Urol. 7:653–660.

Hanage WP, Fraser C, Tang J, Connor T, Corander J. 2009. Hyper-recombination, diversity and antibiotic resistance in the pneumococcus. Science 324:1454–1457.

Holt KE, et al. 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet. 44:1056–1059.

Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M. 2010. *Escherichia coli* type ST131 as the major cause of serious multidrug-resistant *Escherichia coli* infections in the United States. Clin Infect Dis. 51:286–294.

Johnson JR, Porter SB, Zhanel G, Kuskowski MA, Denamur E. 2012. Virulence of *Escherichia coli* clinical isolates in a murine sepsis model in relation to sequence type ST131 status, fluoroquinolone resistance, and virulence genotype. Infect Immun. 80:1554–1562.

Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic *Escherichia coli*. Nat Rev Microbiol. 2:123–140.

Lau SH, et al. 2008. UK epidemic *Escherichia coli* strains A–E, with CTX-M-15 b-lactamase, all belong to the international O25:H4-ST131 clone. J Antimicrob Chemother. 62:1241–1244.

Leopold SR, et al. 2009. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. Proc Natl Acad Sci U S A. 106: 8713–8718.

Lloyd AL, Henderson TA, Vigil PD, Mobley HLT. 2009. Genomic islands of uropathogenic *Escherichia coli* contribute to virulence. J Bacteriol. 191: 3469–3481.

Lloyd AL, Rasko DA, Mobley HLT. 2007. Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*. J Bacteriol. 189:3532–3546.

Lukjancenko O, Wassenaar TM, Ussery DW. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol. 60:708–720.

Marttinen P, et al. 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acid Res. 40:e6.

Naylor SW, et al. 2003. Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic *Escherichia coli* O157:H7 in the bovine host. Infect Immun. 71: 1505–1512.

Nicolas-Chanoine M, et al. 2008. Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. J Antimicrob Chemother. 61:273–281.

Olesen B, et al. 2012. Enteroaggregative *Escherichia coli* O78:H10—the cause of an outbreak of urinary tract infection. J Clin Microbiol. 50: 3703–3711.

Peirano G, Pitout JDD. 2010. Molecular epidemiology of *Escherichia coli* producing CTX-M β-lactamases: the worldwide emergence of clone ST131 O25:H4. Int J Antimicrob Agents. 35:316–321.

Pitout JDD, Campbell L, Church DL, Gregson DB, Laupland KB. 2009. Molecular characteristics of travel-related extended-spectrum-β-lactamase-producing *Escherichia coli* isolates from the Calgary health region. Antimicrob Agents Chemother. 53:2539–2543.

Platell JL, et al. 2011. Commonality among fluoroquinolone-resistant sequence type ST131 extraintestinal *Escherichia coli* isolates from humans and companion animals in Australia. Antimicrob Agents Chemother. 8:3782–3787.

Platell JL, Johnson JR, Cobbold RN, Trott DJ. 2011. Multidrug-resistant extraintestinal pathogenic *Escherichia coli* of sequence type ST131 in animals and foods. Vet Microbiol. 153:99–108.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26:1641–1650.

Rasko DA, et al. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *Escherichia coli* commensal and pathogenic isolates. J Bacteriol. 190:6881–6893.

Rasko DA, et al. 2011. Origins of the *Escherichia coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. New Engl J Med. 365:709–717.
Rogers BA, Sidjabat HE, Paterson DL. 2011. *Escherichia coli* O25b-ST131: a pandemic, multiresistant, community-associated strain. J Antimicrob Chemother. 66:1–14.
Rooney PJ, et al. 2009. Nursing homes as a reservoir of extended-spectrum β-lactamase (ESBL)-producing ciprofloxacin-resistant *Escherichia coli*. J Antimicrob Chemother. 64:635–641.
Sahl JW, et al. 2011. Genomic comparison of multi-drug resistant invasive and colonizing *Acinetobacter baumannii* isolated from diverse human body sites reveals genomic plasticity. BMC Genomics 12: 291.
Schloss PD, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 75:7537–7541.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456.
Tang J, Hanage WP, Fraser C, Corander J. 2009. Identifying currents in the gene pool for bacterial populations using an integrative approach. PLoS Comp Biol. 5:e1000455.
Toh H, et al. 2010. Complete genome sequence of the wild-type commensal *Escherichia coli* Strain SE15, belonging to phylogenetic group B2. J Bacteriol. 192:1165–1166.
Totsika M, et al. 2011. Insights into a multidrug-resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. PLoS One 6:e26578.
Touchon M, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5: e1000344.
Vincent C, et al. 2010. Food reservoir for *Escherichia coli* causing urinary tract infections. Emerg Infect Dis. 16:88–95.
Willems RJ, et al. 2012. Restricted gene flow among hospital subpopulations of *Enterococcus faecium*. MBio 3:e00151–12.
Wirth T, et al. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 60:1136–1151.
Zong Z, Yu R. 2011. blaCTX-M-carrying *Escherichia coli* of the O25b ST131 clonal group have emerged in China. Diagn Microbiol Infect Dis. 69: 228–231.

Associate editor: Josefa Gonzalez