Signatures of negative frequency dependent selection in colonisation factors and the evolution of a multi-drug resistant lineage of *Escherichia coli*

Alan McNally[1*+], Teemu Kallonen[2,3*], Christopher Connor[1*], Khalil Abudahab[2], David M. Aanensen[2], Carolyne Horner[4], Sharon J. Peacock[2,5,6], Julian Parkhill[2], Nicholas J. Croucher[7&], Jukka Corander[2,3,8 &+]

[1]Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK

[2]Infection Genomics, Wellcome Sanger Institute, Cambridge, UK

[3]Department of Biostatistics, University of Oslo, Oslo, Norway

[4]British Society of Antimicrobial Chemotherapy, Birmingham, UK

[5]Department of Medicine, University of Cambridge, Cambridge, UK

[6]London School of Hygiene and Tropical Medicine, London, UK

[7]Faculty of Medicine, School of Public Health, Imperial College, London, UK

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

*Equal contributions

[&]Equal contributions

+Corresponding authors

Abstract

Escherichia coli is a major cause of bloodstream and urinary tract infections globally. The wide dissemination of multi-drug resistant (MDR) strains of extra-intestinal pathogenic *E. coli* (ExPEC) pose a rapidly increasing public health burden due to narrowed treatment options and increased risk of failure to clear an infection. Here, we present a detailed population genomic analysis of the ExPEC ST131 clone, in which we seek explanations for its success as an emerging pathogenic strain beyond the acquisition of antimicrobial resistance (AMR) genes. We show evidence for a stepwise evolution towards separate ecological niches for the main clades of ST131 and differential evolution of anaerobic metabolism, key colonisation and virulence factors. We further demonstrate that negative frequency-dependent selection acting on these loci is a major mechanism that has shaped the population evolution of this pathogen.

Introduction

Escherichia coli is now the most common cause of blood stream infections in the developed world, outnumbering cases of *Staphylococcus aureus* bacteraemia by 2:1 [1]. *E. coli* is also the most common cause of urinary tract infections (UTI), which in turn are the most common bacterial infections in the world [2]. Bacteraemia and UTI are caused by a subset of *E. coli* termed extra-intestinal pathogenic *E. coli* (ExPEC). ExPEC are not a phylogenetically distinct group of *E. coli* but rather represent strains which have acquired virulence-associated genes that confer the ability to invade and cause disease in extra-intestinal sites [3]. Genes associated with virulence that confer the ability to adhere to extra-intestinal tissues, to sequester extracellular iron, to evade the non-specific immune response, and toxins resulting in localised tissue destruction have all been described as essential in the process of ExPEC pathogenesis [4].

The problem presented by the scale of ExPEC infections is exacerbated by the number of cases involving multi-drug resistant (MDR) strains [1,5,6]. Epidemiological surveys report as many as 60% of UTI ExPEC isolates as being resistant to three or more classes of antibiotics, and as many as 50% of bacteraemia isolates [5,6]. The increase in MDR ExPEC prevalence has been rapid and primarily attributable to a small number of ExPEC lineages [5]. The most common of these is the *E. coli* ST131 lineage, which has rapidly become a dominant cause of ExPEC UTI and bacteraemia globally [5–7]. *E. coli* ST131 is particularly associated with carriage of the CTX-M class of extended-spectrum β-lactamase (ESBL) which confers resistance to 3rd-generation cephalosporins [7], and there have been a small number of reports of *E. coli* ST131 isolates carrying metallo-β-lactamases conferring resistance to

57 carbapenems [8]. The carriage of these resistance genes is driven by acquisition and
58 stable maintenance of large MDR plasmids [9].

59 The phylogenetic structure of *E. coli* ST131 is well characterised [10–14] and shows the
60 emergence of a globally disseminated, MDR-associated clade C from primarily drug
61 susceptible clades A and B. The lack of phylogeographic signal and phylogenetic
62 structure based on host source suggests rapid global dispersal and frequent host
63 transitions within clade C [14]. Research has suggested that the acquisition of
64 fluoroquinolone resistance via point mutations in DNA gyrase and DNA
65 topoisomerase genes was the primary driver in the rapid emergence of clade C,
66 alongside the predated acquisition of well-defined ExPEC virulence factors [11,12].
67 Later work also suggested that clade C *E. coli* ST131 may dominate as a successful
68 MDR clade due to the ability to offset the fitness cost of MDR plasmid acquisition
69 and maintenance via compensatory mutations in gene regulatory regions [14].
70 Genome-wide association studies (GWAS) have been used to identify loci and
71 lineage specific alleles significantly associated with clade C *E.coli* ST131, which
72 suggested a secondary flagella locus encoding lateral flagella (Flag-2[15]), and a
73 number of hypothetical proteins and promoter regions as being clade C *E. coli*
74 ST131 associated loci [14].

75 Recent work on *E. coli* causing bacteraemia provided compelling evidence that
76 resistance to antimicrobials has not been the major driver of the success of ST131
77 [16]. Analysis of a large 11-year population survey across the UK showed that ST131
78 rapidly stabilised at a level of approximately 20% after its emergence around 2002 in
79 the UK. This was far in excess of already-resident MDR clones, such as ST88 or
80 ST405. Nevertheless, the overall prevalence of resistance phenotypes remained
81 approximately constant in the population. Furthermore, most currently known major

82 ExPEC clones (primarily ST12, ST73, ST95, and ST69, the last of which also rapidly
83 emerged in 2002) show a similar stable population frequency across the 10 years
84 following the introduction of ST131, despite exhibiting far less extensive resistance
85 profiles. These observations suggested the distribution of ExPEC strains was
86 shaped by negative frequency-dependent selection (NFDS) [16].

87 Clues as to the mechanistic basis of this selection may come from the recent
88 multilocus NFDS model of post-vaccination *Streptococcus pneumoniae* population
89 dynamics [17]. Unexpectedly, frequencies of accessory genes were found to be highly
90 conserved across multiple populations on different continents, despite these
91 populations themselves being composed of different strains, defined in terms of core
92 genome sequences. Detailed modelling and functional analysis indicated changes in
93 strain prevalence could be understood in terms of NFDS driving accessory loci
94 towards equilibrium frequencies through mechanisms involving interactions with
95 other bacteria, hosts, or mobile elements [17]. The level of the selective force was
96 estimated to be similar across the populations and manifested itself in the
97 maintenance of stable population frequencies of the accessory loci despite a
98 substantial perturbation of the population strain composition by the introduction of
99 the pneumococcal vaccine [17].

100 Here, we present the analysis of 862 genomes collated from previous large scale *E.*
101 *coli* ST131 phylogenomic studies [11–14,16,18] and newly sequenced isolates from the
102 BSAC bacteraemia resistance project from the UK and Ireland. This allowed us to
103 perform sufficiently powered population genetic analyses and identify the key steps
104 in the evolution from the largely drug susceptible clades A and B to the globally
105 dominant MDR clade C. By utilising pan-genome analyses we identify accumulation
106 of allelic diversity underpinning the formation of clade C. This diversity occurs in loci

involved in colonisation of the human host by ExPEC. We also provide evidence that the accumulation of allelic diversity in genes involved in anaerobic metabolism is a key trait in the emergence of clade C ST131. Together, our data present a picture of adaptation towards prolonged human colonisation as the primary evolutionary force in the emergence of the globally dominant MDR clade C *E. coli* ST131 and indicate that MDR evolution is likely a consequence of the prolonged exposure to the human host. Such a scenario would give more time for the host immune system to adapt to successful/common clones, presenting a possible mechanism for the population-level NFDS observed in the emergence of the ST131 lineage.

Methods

Genome data

We utilised a collection of 862 *E. coli* ST131 genomes (Table S1), of which 684 were previously sequenced as part of phylogenomic investigations of the ST131 lineage [10,11,13,14,16,19]. We added 184 previously unpublished ST131 isolates from the British society for antimicrobial chemotherapy (BSAC) bacteraemia resistance surveillance project which were selected from *E. coli* in the BSAC resistance surveillance bacteraemia collection from the UK and Ireland between 2001–2011.

In an attempt to avoid any issues arising from different assembly or annotation metrics employed in the previous projects, we downloaded only raw sequence data in fastq format using the previously published accession data. We then performed de novo assembly on all the genomes using Velvet [20] and annotation using Prokka [21] as previously described [16]. A pan-genome of the entire data set was constructed using Roary with 95% identity cut-off [22]. A concatenated core CDS alignment was made from the Roary output and a maximum likelihood phylogenetic tree was constructed

from the alignment using RaxML version 8.2.8 [23] and the GTR model with Gamma rate heterogeneity.

For comparative lineage analysis we utilised the 264 ST73 genomes, and162 ST95 genomes that were sequenced and fully characterised as part of the UK BSAC genome study [16].

Accessory genome analysis

The pan-genome matrix from Roary was utilised to investigate the presence of clade specific loci. The PANINI tool was used with the default setting to visualize the accessory gene sharing patterns in the population https://microreact.org/project/BJKoeBt2b [24]. PANINI has been demonstrated to provide efficient complementary visual means to phylogenetic trees to accurately extract both distinct lineages present in a population-wide genomic dataset, and to highlight clusters within lineages, that are explained by rapidly occurring, homoplasic alterations, such as phage infection. Roary was run on the entire data set using the default 95% sequence identity threshold to cluster genes, allowing us to separate genes based on allelic as well functional differences. Based on a frequency distribution histogram (Figure S1), we assigned a locus as being clade specific if it occurred at a frequency > 95% in one clade and at < 5% in the other two clades. Loci identified as clade specific were functionally annotated by performing a tBlastn analysis of the nucleotide sequence of the loci against the NCBI non-redundant database.

Functional categorisation of pangenomes

To assess the functional composition of the accessory pangenome we assigned Gene Ontology (GO) terms to gene sequences from the pangenome. Briefly, representative sequences from the pan genome of ST131 were mapped to orthologous groups in the bactNOG database using the eggNOG emapper utility [25] Mapping was performed using the diamond search algorithm. Output from eggNOG was filtered to remove Orthologous Groups with no GO terms, a score was assigned to each Orthologous Group based on gene mapping frequency.

Comparisons of lineage and clade specific loci

In order to compare lineage pan-genomes whilst accounting for differences in the number of genomes a sampling approach was utilised. Specifically, a subset with size equal either to the number of ST73 or ST95 genomes was selected at random from the ST131 Clade C. The functional enrichment of genes in the subset was quantified and statistically compared to the ST73 or ST95 pangenome using a Chi Squared test. This process was repeated 100 times to produce 100 p-values, from which the median p-value was calculated. Utilising the same subsampling approach, the pangenome composition of Clade C ST131 genomes was compared to both the Clade A and Clade B pangenomes.

Chi squared statistical tests were performed to assess the significance of the observed differences in functional enrichment. Briefly, with each iteration of the sampling procedure a Chi squared test was performed using the functional proportion of the subsampled pangenome as the observed values and the proportions for ST73 or ST95 as the expected value. This generated 100 p-values from which one can use the average, maximum, or median to assess significance of the observed differences. In addition, proportional Z statistic tests were also

performed to assess the significance of the observed difference. The measurements from the 100 replicates of the subsampling procedure were used to generate an average for the proportions as well as to estimate the variance. The tests were conducted using the proportional measurements from ST73 and ST95 as the 'true' means and quantifying how distinct the ST131 subsamples were from these reference values.

The sequences of 64 anaerobic metabolic genes in which allelic diversity was observed were extracted from individual genomes. The nucleotide sequences were then clustered at 80% identity and 80% length using CD-HIT which was run using the accurate flag and 'word size' of 5 [26]. An additional CD-HIT script was used to extract gene sequences for clusters with more than 3 genes, the minimum required by MEGA-CC for analysis. The sequences were then aligned using Muscle with default settings [27]. Resulting alignment files were analysed in MEGA-CC to produce measurements of Tajima's D [28].

ST131 clade specific SNPs

To visualise the ST131 clades A, B, C, C1 and C2 within the ML tree and the PANINI clustering we identified clade specific SNPs (Table S1) as previously described [16].

NFDS modelling

NFDS modelling used genomic data from the previous publication analysing the population dynamic of blood stream infection *E. coli* isolates in the UK [16]. The analysis effectively assumes these isolates were random, opportunistic infections corresponding to a representative sample of the evolving *E. coli* population. Isolates were assigned to genotypes based on a hierBAPS analysis of the core genome [29].

200 The previously-defined sequence types were used to divide any diverse clusters to
201 the appropriate level of resolution. Therefore the clusters used corresponded to the
202 largest hierBAPS cluster that either contained only a single sequence type, or the
203 nearest-neighbour sequence types within the cluster were single or double locus
204 variants; if neither condition could be satisfied, the third level of clustering was used.
205 This identified 62 sequence clusters across the population. The sets of orthologous
206 sequences were those defined by a previous Roary analysis [16] those present at
207 between 5% and 95% frequency in the first sample, from 2001, were modelled as
208 evolving under NFDS, and tending towards an equilibrium frequency, e_l,
209 corresponding to that in the 2001 sample.

210 Seven resistance phenotypes, present within this frequency range in 2001, were also
211 modelled as evolving under NFDS: amoxicillin, clavulanic acid, ciprofloxaxin,
212 cefuroxime, gentamicin, piperacillin-tazobactam, and trimethoprim. The first six of
213 these were directly inferred from the previously published analysis. Trimethoprim
214 was instead inferred from the *sul* and *dfrA* alleles identified by Roary; data from the
215 Cambridge University Hospitals collection [16] was used to train a model constructed
216 with the randomForest R library (https://cran.r-
217 project.org/web/packages/randomForest/) which had 93% accuracy when applied
218 back to the training dataset. This was used to infer resistance phenotypes for the
219 BSAC collection.

220 Analysis used the heterogeneous multilocus NFDS model described previously [17],
221 modified to treat a vaccine cost, v, as a fitness advantage, r. All individuals, i, of the
222 sequence clusters corresponding to ST131 and ST69 were assigned the same
223 fitness advantage, $r_i = r$; $r_i = 0$ for all other i. Hence the function defining the number
224 of progeny, $X_{i,t}$, produced by i at time t was:

$$X_{i,t} \sim Pois\left(\left(\frac{\kappa}{N_t}\right)(1+r_i)(1-m)\left(\left(1+\sigma_f\right)^{\pi_{i,t}} + (1+\sigma_w)^{\omega_{i,t}}\right)\right)$$

In this formula, density-dependent competition is parameterised by the carrying capacity κ, set at 50,000 to represent a large population that is still computationally feasible, and the total number of cells in the simulated population at t, N_t. The strength of NFDS was determined by the parameters p_f, σ_f and σ_w. As previously, the accessory loci and resistance phenotypes were ordered according to the statistic Δ_l:

$$\Delta_l = \frac{(f_{l,t>0} - e_l)^2}{(1 - e_l(1 - e_l))}$$

Where $f_{l,t>0}$ is the mean post-2001 locus frequency. If the L loci and phenotypes considered to be under NFDS were ordered by ascending values of Δ_l, then l_f was the highest ranking locus meeting the criterion $\frac{l_f}{L} \leq p_f$. This determined the strength of NFDS acting on each locus, and therefore the reproductive fitness of individual i, based on which loci were encoded in its genome, as represented by the binary variable $g_{i,l}$, and the deviation of their simulated locus frequency at time t, $f_{l,t}$, from their corresponding equilibrium frequencies:

$$\pi_{i,t} = \sum_{l=1}^{l_f} g_{i,l}\left(e_l - f_{l,t}\right)$$

And:

$$\omega_{i,t} = \sum_{l=l_f+1}^{L} g_{i,l}\left(e_l - f_{l,t}\right)$$

These summed deviations served as the exponents for the NFDS terms of the reproductive fitness, with $\pi_{i,t}$ and σ_f corresponding to those loci under stronger NFDS, and $\omega_{i,t}$ and σ_w corresponding to those loci under weaker NFDS.

The simulations were initialised with a random selection of κ genotypes from the genomic data, which were biased such that those isolates observed in 2001 were represented at one thousand fold greater frequency than genotypes collected in later years. This was necessary to 'seed' the initial population with ST131 and ST69, to facilitate their expansion in a realistic manner in subsequent years. The parameter *m* represented the rate at which all isolates entered the population through migration; this was biased to import all sequence clusters at the same rate, to avoid any fits in which high rates of migration would artefactually replicate the population observed in the later years of the collection [17].

Model fitting to genomic data

As in Corander et al. [17] the simulation model was fitted through Approximate Bayesian Computation (ABC) using the BOLFI algorithm, which has been shown to accelerate ABC inference 1000-10000 times without loss of accuracy [30]. The prior constraints placed on the parameter values were as follows: the lower bound on all parameters was set to 0.0009 and the upper bounds were r_i – 0.99, m – 0.2, p_f – 0.99, σ_f – 0.03, σ_w. – 0.005. We used 600 iterations of the BOLFI algorithm to minimise the Jensen-Shannon divergence of the sequence cluster frequencies in the genomic data and in the simulations, as ascertained through randomly sampling discrete sets of isolates in accordance with the size and timings of the genomes selected for sequencing from the original collection. Convergence of BOLFI was monitored each 100 iterations and the approximate likelihood estimate was

assessed to have been stabilized by the end of the 600 iterations [30]. The 95% posterior credible intervals for the parameters were obtained using three generations of sequential Monte Carlo sampling with the same default settings as used in Corander et al [17].

Results

NFDS on accessory loci can explain ExPEC population dynamics

Previous work on this population suggested it was subject to balancing selection based on the persistent diversity of strains, and stable prevalence of resistance phenotypes, despite the invasion of genotypes ST69 and ST131, the latter of which has an MDR phenotype [16]. It is possible this could represent strains being adapted to distinct niches through unique gene content. However, using the previous analysis of gene content with Roary, the 18 strains with at least ten representatives in the population had a mean of only 16.7 private genes (range: 1-49), defined as those loci present at >95% in one strain, and <5% in all others. This is consistent with strains being defined by a characteristic combination of common accessory loci, rather than distinctive sequence [14,31].

Such distribution of gene content is similar to that observed in *S. pneumoniae*, in which NFDS acting on variable phenotypes encoded by genomic islands was suggested to shape the population [17]. The Roary analysis had identified 6,824 intermediate-frequency genes, present in between 5% and 95% of the overall population. Comparisons between the pre-ST131 2001 samples, and subsequent data from up to 2011, found strong, linear correlations between the prevalence of their intermediate-frequency genes (Fig 1A, Fig S2). This is consistent with these loci existing at ‘equilibrium’ frequencies, determined by their costs and frequency-

286 dependent benefits. Furthermore, these correlations with the first sample, in 2001,
287 did not successively weaken year-on-year, as might be expected with neutral drift
288 (Fig 1B). Instead, deviation from the first sample increased until 2008, as the
289 sequence clusters (SCs) primarily associated with ST131 and ST69 became more
290 prevalent (Fig 1C). The rise of ST131 was primarily driven by a dramatic rise in the
291 prevalence of MDR clade C isolates, with clade B persisting at a lower, but stable,
292 level. This was followed by a reversion back towards the equilibrium gene
293 frequencies up to 2010, which does not correspond to major changes in the
294 frequency of either ST131 or ST69, suggesting a reconfiguration of other lineages in
295 the population.

296 In order to obtain a population-wide view of these dynamics, the previously-
297 described multilocus NFDS model was therefore applied to this dataset to test
298 whether these strain dynamics were consistent with selection at the accessory locus
299 level. The model was initialised with the 2001 population, which was seeded with
300 genotypes observed in later years at a low level, representing the possibility they
301 were present in the population but unsampled. Subsequent simulation with a Wright-
302 Fisher framework included these post-2001 genotypes migrating into the population
303 at a rate m, while the BAPS clusters corresponding to ST131 and ST69 expanded at
304 a rate determined by their increased reproductive fitness relative to the rest of the
305 population, r. The equilibrium frequencies of 7,211 intermediate frequency loci,
306 corresponding to genes identified by Roary that were between 5% and 95% in the
307 2001 sample plus ten antibiotic resistance phenotypes, were assumed to be those
308 observed in 2001 sample of genomes. These were then simulated as evolving under
309 NFDS; a fraction p_f evolved under strong NFDS, determined by the parameter σ_f,
310 while the rest evolved under weak NFDS, according to parameter σ_w (see Methods).

311 Fitting this model using BOLFI estimated the parameters listed in Table S2, which
312 identified significant evidence for NFDS (σ_f and p_f greater than zero), providing a
313 gene-level mechanistic basis for NFDS underlying the previous strain-level
314 observations of Kallonen *et al* [16].

315 These simulations successfully reproduced several aspects of the observed data
316 (Fig 2, Fig S3). Both ST131 and ST69 rapidly spread through the population, and
317 stabilise at an equilibrium frequency. This does not occur at the expense of the
318 established, common clones, such as ST73 and ST95. Instead, in accordance with
319 the genomic data, the displaced sequence clusters include ST10, ST14, ST144 and
320 ST405. Although their expansions were expected to be the same, as both were
321 driven by the fitness advantage *r*, NFDS at the gene level constrains the expansion
322 of ST69 in this population, whereas ST131 reaches the higher prevalence observed
323 in the genomic sample, with both antibiotic-sensitive and resistant isolates rising.
324 This seems unlikely to reflect ST69 being confined to narrow niche, as the diversity
325 in the intermediate frequency loci within ST69 is greater than that within ST131 (Fig
326 S4). Notably, in these pairwise comparisons of gene content, the maximal pairwise
327 distances between representatives of clade C were similar to those between random
328 representatives selected from the same sequence cluster, suggesting some
329 members of this recently-emerged clade had undergone substantial changes in their
330 genome content. Instead, the rapid invasion of ST131 into the population appears to
331 represent a highly-fit genotype emerging in both antibiotic-sensitive (clade B) and
332 resistant (clade C) forms. The latter partially displaced existing MDR clones, none of
333 which were as successful as SC1. Therefore a comprehensive genomic dataset
334 encompassing all known ST131 genome sequences was created to understand the

unique characteristics of the ST131 lineage, with particular focus on the successful clades B and C.

Core and accessory genomic structure of the ST131 population.

To facilitate the comprehensive pan-genome analysis that can provide a high-resolution view into a lineage's evolution [14], the ST131 isolates from the UK collection were combined with collections sampled internationally. A maximum likelihood phylogeny made from an alignment of concatenated core CDS from all 862 genomes confirmed the earlier consensus three clade structure of the lineage (Fig 3a). The collation of all available genomes into a single phylogeny had no impact on the previous observed population structure, with no phylogeographic signal or host source clustering evident in the phylogeny https://microreact.org/project/BJKoeBt2b. To confirm that the collation of the 862 genomes was consistent with previous descriptions of the accessory genome distribution in ST131, isolate relatedness based on shared accessory gene content was visualized as a two-dimensional projection using PANINI (Fig 3b) [24]. Clades A and B largely resided in dense clusters at the periphery of the projection. In contrast, clade C isolates were more diffuse, overlapping with some clade B isolates, forming a cloud with discernible sub-structuring into distinct groups. This reflects the very high diversity of the gene content across the C clade, and the previous finding of multiple accessory genome sub-clusters in clade C [14].

Low frequency accessory genes suggest differential ecology of clade A and clade B/C *E. coli* ST131

Given that the vast majority of accessory genes occur at very low frequency, we sought to determine if these represented mobile genetic elements circulating

transiently in the population. We extracted genes occurring in less than 20% of the entire population (based on the distribution of the gene frequencies in Fig S1) that were confined to a single clade, and then functionally categorised them. In both clade A and clade B/C (Dataset S1-S3) the overwhelming majority of low frequency accessory genes encode hypothetical proteins (64.4% clade A, 58% clade B/C). Excluding the hypothetical proteins from the analysis showed unexpected bias in functional gene categories differentially observed in the lineages (Fig 4). The most common gene types were functional phage, plasmid and other mobile genetic element (MGE) genes, with more private phage genes present in clade B/C than in clade A. Conversely, there were more private plasmid genes in clade A than clade B/C, despite the presence of a diverse number of MDR plasmids within clade C.[14] Together this suggests that clade A strains of *E. coli* ST131 and clade B/C strains of *E. coli* ST131 are exposed to different plasmid and phage pools, an observation which is most parsimoniously explained by them having different ecological habitats.

Clade-specific and intermediate frequency genes in the population.

To identify which aspects of the accessory genome differed between the clades of ST131, the distributions of the 32,631 sets of orthologous genes identified by Roary were analysed (Dataset S3). Characterising the full set of loci present at intermediate frequencies was not feasible, as even focussing on the 3,354 present at between 5% and 95% frequency found the majority of these were present at a frequency below 20% (Fig S1). Therefore, the search was refined to clade specific genes, occurring at a frequency > 95% in one clade but at <5% in the other two clades (Dataset S1).

Clade A contained the highest number of loci exclusive to a lineage (54) despite constituting the least sampled clade. Clade B had only 2 exclusive loci and clade C

had 18. When clades B and C were combined against clade A, there were 60 loci exclusively present in the B/C combination. The majority of clade A private genes encode hypothetical proteins whilst those private to clade C encode DNA modification proteins and metabolic functions. The genes private to clade B/C combined also encode hypothetical proteins and metabolic functions, notably five dehydrogenase enzymes involved in anaerobic metabolism labelled *yihV*, *garR_3, fadJ, fdhD,* and *gnd* in our dataset (Dataset S2). Blast analysis against the NCBI non-redundant database suggested that the dehydrogenase enzyme gene annotated as *pdxA*, in our Roary dataset was confined to clade C ST131 strains. These dehydrogenase enzyme genes were found to be present across phylogroup B2 *E. coli* strains via blastn against the NCBI non-redundant database. Therefore these loci are not unique to clade C ST131, and were either acquired by an ancestral clade B/C strain, or have been lost by clade A.

High diversity in core anaerobic metabolism genes unique to clade B/C

Analysis of accessory loci private to clade B/C (present in >95% of that population) identified two separate loci encoding 3-hydroxyisobutyrate dehydrogenase enzymes, and loci encoding 3-hydroxyacyl-CoA dehydrogenase, 6-phosphogluconate dehydrogenase, and formate dehydrogenase. Analysis of clade B/C loci circulating at low frequency of <20% also identified a significant over-representation of genes encoding dehydrogenase enzymes involved in anaerobic metabolism (a total of 64 loci), including seven variants of formate dehydrogenase. There were also seven variants of the *eutA* gene found in the ethanolamine utilisation pathway (the *eut* operon) and a distinct version the *cobW* gene which encodes the sensor kinase for activation of the cobalamin biosynthesis operon. Closer investigation of the sequences of these loci suggested that these were not genes private to clade B/C

408 per se, but rather represented multiple unique alleles of genes that are core to the
409 ST131 population which differ at nucleotide sequence level by more than 5%. This
410 infers a unique selection pressure is acting on these core genes in clade B/C
411 compared to clade A.

412 Further scrutiny of low frequency loci in clade B/C also identified alternative alleles of
413 a large number of well characterised extra-intestinal pathogenic *E. coli* virulence-
414 associated genes, including: antigen 43 (7 alternative alleles); heavy metal
415 resistance such as arsenic (5 loci), copper (4 loci), and mercury (5 loci); capsule
416 biosynthesis (20 loci); cell division and septation (14 loci); antibiotic resistance to
417 chloramphenicol (3 loci), macrolides (2 loci), rifampicin (1 locus), and MDR efflux
418 pumps (21 loci); iron acquisition (39 loci); curli and type I fimbriae and P pili (42 loci);
419 lateral and classical flagella (26 loci); and LPS synthesis (9 loci). These loci
420 represent alternative alleles of genes found widely across the *E. coli* phylogeny
421 indicating there are multiple allelic variants of important genes that are confined to
422 clade B/C of the *E. coli* ST131 lineage.

423 We sought to determine the distribution of this allelic diversity across the *E. coli*
424 ST131 phylogeny by annotating the tips of the phylogenetic tree with the
425 presence/absence of each of the anaerobic metabolism (Figure 5), and capsule, cell
426 division, MDR efflux, iron acquisition, pili, and flagella divergent loci (Figure 6). Our
427 analysis shows that the alternative alleles occur at very low frequency but are
428 randomly distributed throughout the phylogeny of the C clade, and are exclusive to
429 clade C. Given that these alleles differ from the normal conserved versions of genes
430 by >5% at nucleotide level, it is implausible that these alleles would be arising
431 repeatedly and independently via mutation. Instead, the most parsimonious
432 explanation is that the minor frequency alternative alleles are being distributed

through the population via recombination. This conclusion is supported by the fact that every one of the allele variants identified in our analysis has 100% nucleotide identity matches with genes present in other *E. coli* in the NCBI non-redundant database.

Given that our data set is biased towards clade C genomes, we performed comparative analyses of the frequency with which allelic diversity occurs in anaerobic metabolism genes. We randomly subsampled clade C 100 times and compared an equal number of clade A, B, and C genomes for allelic diversity. Our data shows that even when randomly subsampling clade C, the levels of diversity observed in anaerobic metabolism genes is significantly higher than in clade A, providing evidence that the accumulation of sequence diversity is specific to the MDR clade C (Figure 5).

Finally, we sought to exclude the possibility that the presence of these allelic variants was skewed by some form of geographically localised expansion of variants. To do this we compared the relative frequency of all accessory genes, highlighting the allele variants in anaerobic metabolism, capsule, cell division, MDR efflux, iron acquisition, fimbriae, and flagella present in UK versus non-UK isolate genomes (Figure S5). Our data showed a strong linear relationship between the frequency of genes in the two populations, indicating that the data was not biased by expansion of alleles in a given geographical location, and that this accumulated diversity was equally as likely to happen in any given strain independent of its geographical origin.

Allelic diversity of anaerobic metabolism genes in Clade C ST131 is not observed in other dominant ExPEC lineages

456 The possibility exists that the above observations made for the clade C of *E. coli*
457 ST131 simply reflect the general evolutionary path of a successful extra-intestinal
458 pathogen. To test this we performed an identical analysis on the pangenome of 261
459 ST73 isolates and of 160 ST95 isolates from the UK BSAC population survey [16]. *E.*
460 *coli* ST73 and ST95 represent two of the most dominant lineages associated with
461 clinical extra-intestinal disease alongside ST131 [5,16], but are predominantly non-
462 MDR lineages and rarely associated with MDR plasmids [16]. As with our inter-clade
463 comparisons, we randomly subsampled clade C ST131 100 times to allow equal
464 numbers of genomes per lineage to be compared. Our analysis showed a similar
465 ratio of plasmid, phage and hypothetical proteins in the accessory genome as in
466 ST131 (Fig 7). ST73 and ST95 displayed similar ratios of alternative alleles in P and
467 Type 1 fimbriae, cell division and septation genes, and multiple iron acquisition
468 genes as observed in ST131. However, enrichment in allelic variation in anaerobic
469 metabolism genes was significantly higher in any given subsampled set of clade C
470 ST131 genomes compared to both lineages. This supports the hypothesis that the
471 observation of increased diversity accumulating in anaerobic metabolism genes is
472 not a more general extra-intestinal pathogenic *E. coli* trait but is particularly enriched
473 in the ST131 lineage.

474 The accumulation of nucleotide diversity in a given set of loci can often be
475 interpreted as a signature of some form of selection occurring on those genes.
476 However the low levels of frequency of any given allele across clade C strains
477 contradicts a hypothesis for positive selection, where one would expect successful or
478 beneficial alleles to sweep to a high frequency or fixation. Indeed comparison of the
479 sequences of each of the 64 anaerobic metabolism loci in which diversity was

observed identified just three loci which showed signatures of positive selection as indicated by a Tajima's D score above two.

However, these results can be reconciled with a lineage evolving under NFDS. Different resource use strategies can facilitate co-existence between competing strains, such those co-colonising a host, resulting in frequency-dependent selection [32,33]. This would explain the sustained intermediate frequencies of genes encoding dehydrogenases over multiple years (Fig S6). Should the equilibrium frequency of any of these loci become limiting on the expansion of ST131, it would be expected that successful subclades might appear to avoid being constrained by the potentially limiting equilibrium frequencies of some accessory loci. Hence this diversification of metabolic loci could represent the adaptive radiation of a successful genetic background able to efficiently compete with the resident *E. coli* population through a diverse panel of metabolic capacities suited to exploiting resources under anaerobic conditions.

Discussion

The evolutionary events that led to the emergence of *E. coli* ST131 have been an intense focus of research, with consensus opinion suggesting that, following acquisition of key ExPEC virulence factors, acquisition of fluoroquinolone resistance in the 1980's by the clade C sub-lineage of ST131 was a key event in that emergence [11,12]. However, a recent nationwide UK population survey rejected this hypothesis and suggested that success of the major ExPEC clones is not dictated by resistance traits [16]. Here, we identify the conserved frequencies of accessory genes in the *E. coli* population which strongly suggest this species' population structure and dynamics are shaped by NFDS acting on genomic islands. Such multilocus NFDS is

504 able to account for how an otherwise stable population was disrupted by the invasion
505 of ST131 and ST69, displacing some lineages while leaving other, largely antibiotic-
506 susceptible, genotypes at almost untouched prevalences. However, the model could
507 not explain the selective advantage driving ST131 into the UK *E. coli* population, an
508 emergence that has been replicated worldwide. One possibility is that *E. coli*
509 populations were not at equilibrium prior to the emergence of ST131. Instead,
510 selection by antibiotic use could have sustained a set of low fitness MDR genotypes,
511 the presence of which disturbed the population composition. Yet this seems unlikely
512 to have been a situation that would have persisted for years in the many locations in
513 which ST131 has been successful. Alternatively, ST131's prominence could
514 represent an increased propensity to cause disease, from which isolates were
515 sampled, rather than a genuine rise in the carried *E. coli* population. This also seems
516 unlikely given the publications which fail to uncover any specific virulence-associated
517 phenotypic traits unique to the ST131 lineage [34,35]. Understanding such a change
518 requires detailed analysis of the ST131 lineage. Compiling all of the available
519 genome sequence data for the lineage to date permitted new events associated with
520 the emergence of clade C to be identified.

521 Previous work has suggested that clade C strains of *E. coli* ST131 undergo reduced
522 levels of detectable core genome recombination compared to other phylogroup B2 *E.*
523 *coli* [36] or ST131 clade A strains [14]. We have previously postulated that this may be a
524 result of ecological separation between clade C strains and other common ExPEC
525 [14,36]. Our analysis of nearly 900 genomes has allowed us to interrogate accessory
526 gene movement to a far greater resolution than previously possible, and with
527 sufficient numbers of genomes in each of the three ST131 clades to provide
528 statistical rigour. From the analysis of the accessory genome we identified thousands

529 of plasmid, phage and other mobile genetic element genes which are private to clade
530 A and the combined clade B/C, respectively. Such an observation is a classic
531 signature of ecological separation of the two populations [37,38], particularly given that
532 the genetic distance between clade A and clade B/C is much smaller than it is to
533 other lineages and species from which the circulating genes are also found in the
534 NCBI non-redundant database.

535 Our analysis also identified a significantly increased level of sequence diversity in
536 genes involved in key host colonisation processes in clade C. This diversity was
537 uncovered through our pan-genome analysis as allelic variants of core genes.
538 Primary amongst these is a large number of genes involved in anaerobic
539 metabolism, including seven allelic variants of the formate dehydrogenase gene, as
540 well as allelic variants of genes involved in ethanolamine utilisation and cobalamin
541 biosynthesis. The pivotal role of ethanolamine production and cobalamin
542 biosynthesis in the ability of Gram negative pathogens to outcompete bacteria in the
543 human intestine is well documented [39,40], and this phenomenon only occurs when
544 supported by an increased ability to perform anaerobic respiration in the presence of
545 inflammation [39]. It has been shown that MDR *E. coli* ST131 is able to colonise the
546 gastro-intestinal tract of humans for months or years in the absence of antibiotic
547 selection [41,42], and that this colonisation results in a displacement of the *E. coli*
548 colonising the host prior to exposure to the MDR strain [41].

549 Whilst this diversity in anaerobic metabolism genes was unique to clade C ST131,
550 the allelic variation observed in other human colonisation and virulence factors such
551 as iron acquisition, fimbriae, and cell division was also observed in two of the other
552 most commonly isolated lineages of *E. coli* from extra-intestinal infections, ST73 and
553 ST95. This diversity likely reflects selection occurring on genes important for ExPEC

554 pathogenesis. Iron acquisition is well characterised as a key virulence determinant in
555 ExPEC, with the ability to initiate a successful UTI completely abrogated in the
556 absence of functional iron acquisition systems [43]. Recent experimental vaccine work
557 exploiting siderophore production by ExPEC has shown to be highly effective in
558 rodent models on ExPEC UTI [44]. The importance of iron acquisition can also explain
559 many of the MDR efflux allele variants seen in this data set, with half occurring in the
560 *acrD* gene which has been experimentally shown to play a role in iron acquisition in
561 *E. coli* [45]. We identified multiple alleles of genes in the type 1 fimbriae operon and in
562 genes in the P pilus operon which are classical virulence determinants in UTI [46], and
563 multiple genes involved in capsule biosynthesis, which we have previously reported
564 as being a hotspot for recombination in *E. coli* ST131 [13,35]. We also identified
565 multiple alleles of genes involved in controlling incomplete septation and filamentous
566 growth, which is a crucial process in the formation of the filamentous intracellular
567 bacterial communities (IBCs) which are thought to be fundamental in the ability of
568 ExPEC to survive inside bladder epithelial cells and cause UTI [47]. There are a small
569 number of allelic variants in anaerobic metabolism genes also present in ST73 and
570 ST95, possibly reflecting recent experimental studies suggesting a crucial role for the
571 cytochrome-bd oxidase system in the ability to cause urinary tract infection [48]. Also
572 previous studies using saturated mutagenesis techniques and studying global
573 transcriptional patterns during urinary tract infection of ExPEC strains have
574 suggested a key role for dehydrogenase enzymes involved in anaerobic metabolism
575 in the ability to cause pathology in the mammalian urinary tract [49–51].

576 Recent modelling data on why drug resistant and drug susceptible populations of
577 bacteria co-exist highlighted that any factors which increase the duration of
578 colonisation in a human host will also increase the selective pressure for it to evolve

579 antibiotic resistance [52]. Hence both the success of ST131 in invading the population,
580 and the association of many isolates in this lineage with an MDR phenotype, would
581 be consistent with its distinctive anaerobic metabolism loci facilitating enhanced
582 persistence within its host, perhaps through an improved ability to outcompete
583 resident commensal *E. coli* strains. The fact that this selection is only seen in clade
584 C of ST131 suggests that this occurred around the time of the emergence of the
585 lineage as a human clinical threat [13] alongside the development of fluoroquinolone
586 resistance. Subsequent acquisition of MDR plasmids, and the consequent selection
587 for an ability to offset the fitness costs of long term MDR plasmid maintenance [14], is
588 likely to have occurred as a result of prolonged exposure to selective antibiotic
589 environments during colonisation of humans. Nevertheless, neither anaerobic
590 metabolism genes nor antibiotic resistance loci have swept to fixation in ST131,
591 reflecting their fluctuating but stable prevalence in the broader *E. coli* population (Fig
592 S6). This diversification can instead be explained by NFDS, under which these
593 genes are beneficial when rare, because they provide an advantage over co-
594 colonising strains which will typically lack the same metabolic capacities. However,
595 as these traits become more common as ST131 expands, representatives of this
596 lineage will more commonly encounter one another, therefore necessitating further
597 diversification for different clade C representatives to sustain their advantage over
598 competitors. Similarly, the capsule locus diversification previously observed within
599 clade C, resulting in the capsule synthesis locus corresponding to a 'hotspot' of
600 recombination [35], could result from NFDS of variable antigens [54], with the host
601 immune system selecting for a diversity of capsule structures as the dominant type
602 becomes more common following ST131's emergence[16].

Our data present a novel and important hypothesis for the mechanisms by which dominant lineages of multi-drug resistant *E. coli* evolve and emerge from their background populations. It is essential that this knowledge is taken forward to study other emerging and emergent lineages of MDR *E. coli* and of other MDR Gram-negative pathogens.

Data accession

Accession numbers for the reads used in this study are listed in Table S1 with information of year and place of isolation and the results of the *in silico* PCR for clade specific SNPs.

Acknowledgements

This publication presents independent research supported by the Health Innovation Challenge Fund (HICF-T5-342 and WT098600), a parallel funding partnership between the UK Department of Health and Wellcome. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health, Public Health England or Wellcome. T.K. was funded by the Norwegian Research Council JPIAMR grant no. 144501. J.C. was funded by the ERC grant no. 742158. NJC is funded by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and Royal Society (Grant Number 104169/Z/14/Z). CC is funded by the Wellcome MIDAS doctoral training program at UoB.

References

1. de Kraker, M. E. A. *et al.* The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System. *Clin. Microbiol. Infect.* **19,** 860–868 (2013).

626 2. Foxman, B. The epidemiology of urinary tract infection. *Nat. Rev. Urol.* **7,** 653–
627 660 (2010).

628 3. Kohler, C.-D. & Dobrindt, U. What defines extraintestinal pathogenic
629 *Escherichia coli*? *Int. J. Med. Microbiol.* **301,** 642–647 (2011).

630 4. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. J. Genomic islands in
631 pathogenic and environmental microorganisms. *Nat. Rev.* **2,** 414–424 (2004).

632 5. Alhashash, F., Weston, V., Diggle, M. & McNally, A. Multidrug-Resistant
633 *Escherichia coli* Bacteremia. *Emerg.Infect.Dis* **19,** 1699–1701 (2013).

634 6. Croxall, G. *et al.* Molecular epidemiology of extraintestinal pathogenic
635 *Escherichia coli* isolates from a regional cohort of elderly patients highlights the
636 prevalence of ST131 strains with increased antimicrobial resistance in both
637 community and hospital care settings. *J. Antimicrob. Chemother.* **66,** (2011).

638 7. Banerjee, R. & Johnson, J. R. A new clone sweeps clean: the enigmatic
639 emergence of *Escherichia coli* sequence type 131. *Antimicrob. Agents*
640 *Chemother.* **58,** 4997–5004 (2014).

641 8. Peirano, G., Schreckenberger, P. C. & Pitout, J. D. D. Characteristics of NDM-
642 1-producing *Escherichia coli* isolates that belong to the successful and virulent
643 clone ST131. *Antimicrob. Agents Chemother.* **55,** 2986–2988 (2011).

644 9. Mathers, A. J., Peirano, G. & Pitout, J. D. D. The role of epidemic resistance
645 plasmids and international high-risk clones in the spread of multidrug-resistant
646 Enterobacteriaceae. *Clin. Microbiol. Rev.* **28,** 565–591 (2015).

647 10. Price, L. B. *et al.* The epidemic of extended-spectrum-β-lactamase-producing

648 *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-
649 Rx. *MBio* **4,** e00377-13 (2013).

650 11. Stoesser, N. *et al.* Evolutionary History of the Global Emergence of the
651 *Escherichia coli* Epidemic Clone ST131. *MBio* **7,** e02162 (2016).

652 12. Ben Zakour, N. L. *et al.* Sequential acquisition of virulence and fluoroquinolone
653 resistance has shaped the evolution of *Escherichia coli* ST131. *MBio* **7,**
654 e00347-16- (2016).

655 13. Petty, N. K. *et al.* Global dissemination of a multidrug resistant *Escherichia coli*
656 clone. *Proc Natl Acad Sci.***111,** 5964-9 (2014).

657 14. McNally, A. *et al.* Combined Analysis of Variation in Core, Accessory and
658 Regulatory Genome Regions Provides a Super-Resolution View into the
659 Evolution of Bacterial Populations. *PLoS Genet.* **12,** e1006280 (2016).

660 15. Ren, C.-P., Beatson, S. A., Parkhill, J. & Pallen, M. J. The Flag-2 Locus, an
661 Ancestral Gene Cluster, Is Potentially Associated with a Novel Flagellar
662 System from *Escherichia coli*. *J. Bacteriol.* **187,** 1430–1440 (2005).

663 16. Kallonen, T. *et al.* Systematic longitudinal survey of invasive *Escherichia coli* in
664 England demonstrates a stable population structure only transiently disturbed
665 by the emergence of ST131. *Genome Res.* **27,** 1437-49 (2017).

666 17. Corander, J. *et al.* Frequency-dependent selection in vaccine-associated
667 pneumococcal population dynamics. *Nat. Ecol. Evol.* **1,** 195-60 (2017).

668 18. Clark, G. *et al.* Genomic analysis uncovers a phenotypically diverse but
669 genetically homogeneous *Escherichia coli* ST131 clone circulating in unrelated

670 urinary tract infections. *J. Antimicrob. Chemother.* **67,** 868-77 (2012).

671 19. Salipante, S. J. *et al.* Large-scale genomic sequencing of extraintestinal
672 pathogenic *Escherichia coli* strains. *Genome Res.* **25,** 119–128 (2015).

673 20. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly
674 using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).

675 21. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,**
676 2068–9 (2014).

677 22. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis.
678 *Bioinformatics* **31,** 3691–3693 (2015).

679 23. Stamatakis, A., Ludwig, T., Maier, H. RAxML-III: a fast program for maximum
680 likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21,** 456
681 (2005).

682 24. Abudahab, K. *et al.* PANINI: Pangenome Neighbor Identification for Bacterial
683 Populations. *bioRxiv* DOI: 10.1101/174409 (2017).

684 25. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through
685 Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34,** 2115–2122
686 (2017).

687 26. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large
688 sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

689 27. Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high
690 throughput. *Nucleic Acids Res* **32,** 1792–7 (2004).

691 28. Kumar, S., Stecher, G., Peterson, D. & Tamura, K. MEGA-CC: computing core

692 of molecular evolutionary genetics analysis program for automated and
693 iterative data analysis. *Bioinformatics* **28,** 2685–2686 (2012).

694 29. Corander, J., Marttinen, P., Sirén, J. & Tang, J. Enhanced Bayesian modelling
695 in BAPS software for learning genetic structures of populations. *BMC*
696 *Bioinformatics* **16,** 539 (2008).

697 30. Gutmann, M. U. & Corander, J. Bayesian Optimization for Likelihood-Free
698 Inference of Simulator-Based Statistical Models. *J. Mach. Learn. Res.* **17,** 1–
699 47 (2016).

700 31. Croucher, N. J. *et al.* Diversification of bacterial genome content through
701 distinct mechanisms over different timescales. *Nat. Commun.* **5,** 5471 (2014).

702 32. Stewart, F. M. & Levin, B. R. Partitioning of Resources and the Outcome of
703 Interspecific Competition: A Model and Some General Considerations. *Am.*
704 *Nat.* **107,** 171–198 (1973).

705 33. Friesen, M., Saxer., Travisano, M., & Doebeli, M. Experimental evidence for
706 sympatric ecological diversification due to frequency-dependent competitipon
707 in *E. coli*. *Evolution* **58,** 245–260 (2004).

708 34. Alqasim, A. *et al.* Phenotypic microarrays suggest *Escherichia coli* ST131 is
709 not a metabolically distinct lineage of extra-intestinal pathogenic *E. coli*. *PLoS*
710 *One* **9,** e88374 (2014).

711 35. Alqasim, A., Scheutz, F., Zong, Z. & McNally, A. Comparative genome
712 analysis identifies few traits unique to the *Escherichia coli* ST131 H30Rx clade
713 and extensive mosaicism at the capsule locus. *BMC Genomics* **15,** 830 (2014).

714 36. McNally, A. *et al*. The evolutionary path to extra intestinal pathogenic, drug
715 resistant *Escherichia coli* is marked by drastic reduction in detectable
716 recombination within the core genome. *Genome Biol.Evol.* **5,** 699–710 (2013).

717 37. Shapiro, B. J. *et al.* Population genomics of early events in the ecological
718 differentiation of bacteria. *Science* **336,** 48–51 (2012).

719 38. Reuter, S., *et al*. Directional gene flow and ecological separation in *Yersinia*
720 *enterocolitica*. *Microb. Genomics* **1,** e000030 (2015).

721 39. Winter, S. E. *et al.* Gut inflammation provides a respiratory electron acceptor
722 for Salmonella. *Nature* **467,** 426–429 (2010).

723 40. McNally, A., Thomson, N. R., Reuter, S. & Wren, B. W. 'Add, stir and reduce':
724 *Yersinia spp*. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.*
725 **14,** (2016).

726 41. Arcilla, M. S. *et al.* Import and spread of extended-spectrum beta-lactamase-
727 producing Enterobacteriaceae by international travellers (COMBAT study): a
728 prospective, multicentre cohort study. *Lancet. Infect. Dis.* **17,** 78–85 (2017).

729 42. Overdevest, I. *et al.* Prolonged colonisation with *Escherichia coli* O25:ST131
730 versus other extended-spectrum beta-lactamase-producing *E. coli* in a long-
731 term care facility with high endemic level of rectal colonisation, the
732 Netherlands, 2013 to 2014. *Euro Surveill.* **21,** 1560 (2016).

733 43. Reigstad, C. S., Hultgren, S. J. & Gordon, J. I. Functional genomic studies of
734 uropathogenic *Escherichia coli* and host urothelial cells when intracellular
735 bacterial communities are assembled. *J. Biol. Chem.* **282,** 21259–21267
736 (2007).

44. Mike, L. A., Smith, S. N., Sumner, C. A., Eaton, K. A. & Mobley, H. L. T. Siderophore vaccine conjugates protect against uropathogenic *Escherichia coli* urinary tract infection. *Proc. Natl. Acad. Sci.* **113,** 13468-73 (2016).

45. Horiyama, T. & Nishino, K. AcrB, AcrD, and MdtABC multidrug efflux systems are involved in enterobactin export in *Escherichia coli*. *PLoS One* **9,** e108642 (2014).

46. Wright, K. J., Seed, P. C. & Hultgren, S. J. Development of intracellular bacterial communities of uropathogenic *Escherichia coli* depends on type 1 pili. *Cell. Microbiol.* **9,** 2230–2241 (2007).

47. Anderson, G. G. *et al.* Intracellular bacterial biofilm-like pods in urinary tract infections. *Science (80).* **301,** 105–107 (2003).

48. Shepherd, M. *et al.* The cytochrome bd-I respiratory oxidase augments survival of multidrug-resistant *Escherichia coli* during infection. *Sci. Rep.* **6,** 35285 (2016).

49. Wiles, T. J. *et al.* Combining quantitative genetic footprinting and trait enrichment analysis to identify fitness determinants of a bacterial pathogen. *PLoS Genet.* **9,** e1003716 (2013).

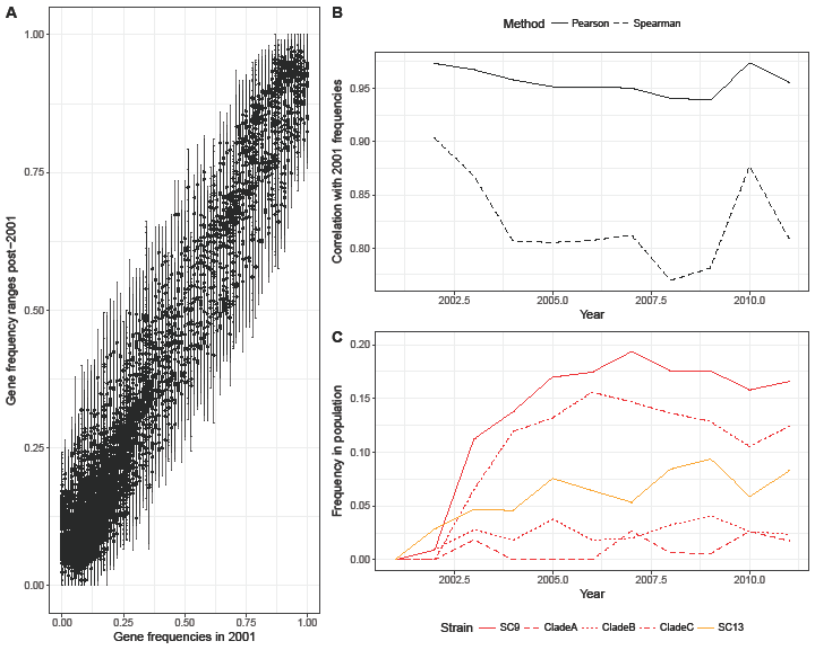
50. Subashchandrabose, S., Smith, S. N., Spurbeck, R. R., Kole, M. M. & Mobley, H. L. T. Genome-wide detection of fitness genes in uropathogenic *Escherichia coli* during systemic infection. *PLoS Pathog.* **9,** e1003788 (2013).

51. Subashchandrabose, S. *et al.* Host-specific induction of *Escherichia coli* fitness genes during human urinary tract infection. *Proc. Natl. Acad. Sci.* **111,** 18327–18332 (2014).

52. Lehtinen, S. *et al.* Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc. Natl. Acad. Sci.* **114,** 1075–1080 (2017).

53. Gupta, S., Ferguson, N. & Anderson, R. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* **280,** 912–915 (1998).

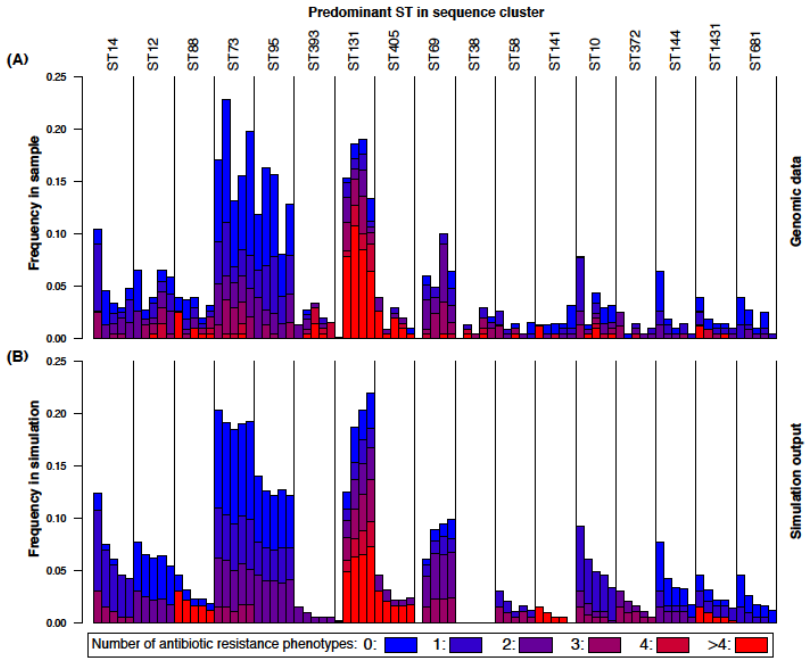
767 Figure 1: Summarising the population dynamics of the British Society for
768 Antimicrobial Chemotherapy extraintestinal pathogenic *E. coli* collection. These
769 isolates were collected from bacteraemia cases around the UK between 2001 and
770 2011. (A) Conservation of gene frequencies. Each point corresponds to one of the
771 6,824 genes identified by ROARY in the BSAC collection with a mean frequency
772 between 0.05 and 0.95. The horizontal axis position indicates the starting frequency
773 in 2001, and the vertical axis indicates the mean frequency over all years, with the
774 error bars indicating the full range observed across annual samples. (B) Correlation
775 of gene frequencies with those observed in 2001. This shows the changing
776 correlation of gene frequencies, calculated by both the Pearson and Spearman
777 methods, in each year relative to those observed in 2001. Both measures indicate a
778 divergence in gene frequencies as ST69 and ST131 emerge, until 2010, at which
779 point there is a reversion to the frequencies seen in the original population. (C)
780 Emergence of ST69, in orange, and ST131, in red. The frequencies of the subclades
781 of ST131 are shown by the red dashed lines.



782

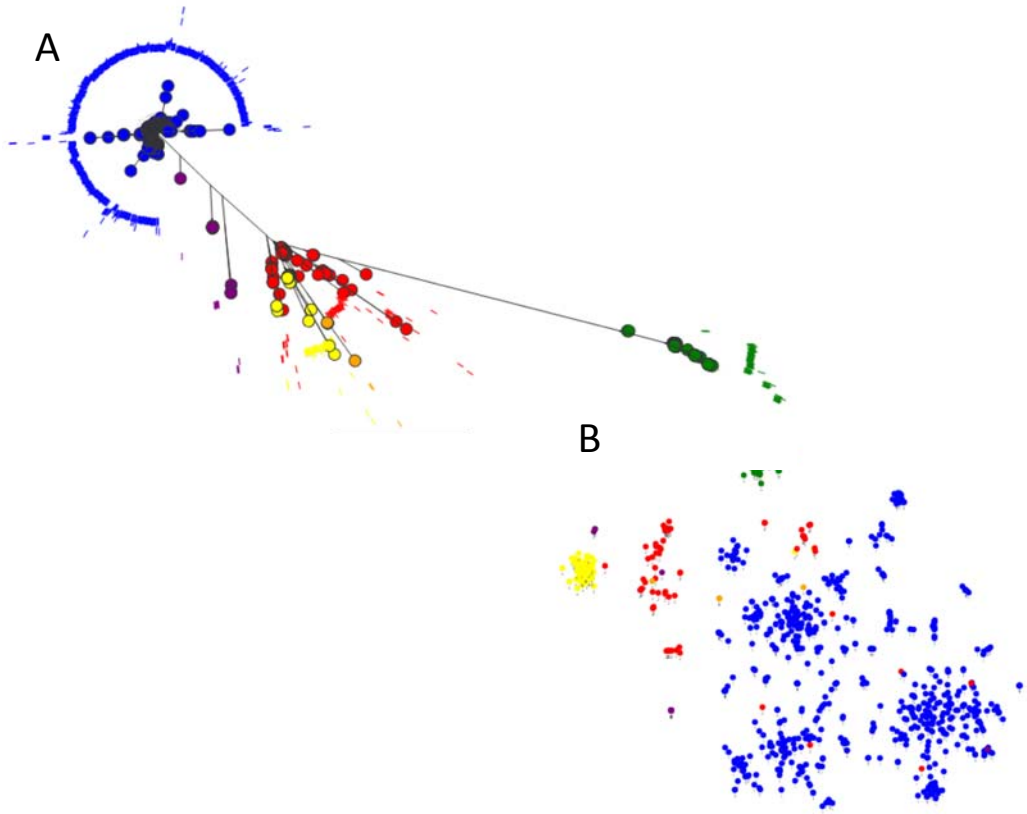
783 Figure 2: Simulations of changes in the BSAC extra-intestinal pathogenic *E. coli*
784 population evolving under multilocus NFDS. The top panel shows the genomic data,
785 and the bottom panel show the median frequencies observed from 100 simulations
786 run with the best-matching parameter set identified by fitting the model with BOLFI.
787 This corresponded to σ_f = 0.029, r = 0.179, m = 0.001, p_f = 0.425 and σ_w = 0.0048.
788 Each column corresponds to a sequence cluster identified by hierBAPS (see
789 Methods), and is annotated with the predominant sequence type with which it is
790 associated. Each bar indicates the frequency of the sequence cluster in subsequent
791 time periods, from left to right. The bars are coloured according to the number of
792 antibiotic resistance phenotypes associated with the isolates within the sequence
793 cluster at different timepoints. Only sequence clusters reaching a frequency of at
794 least 2.5% at one timepoint in the genomic sample are shown; the full results of the
795 simulation, including measures of between-simulation variation, are shown in Fig S3.

796

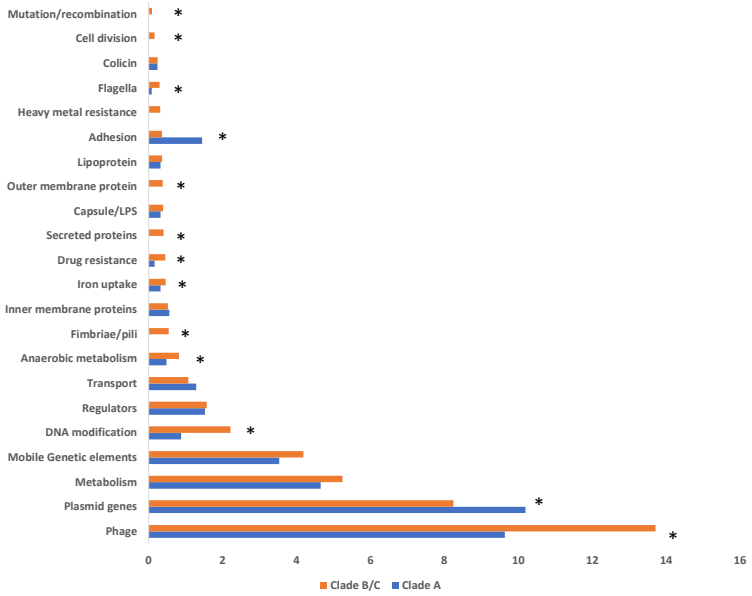


797

Figure 3: (A) Maximum likelihood phylogeny of 862 *E. coli* ST131 strains. The phylogeny was inferred using RaxML with a GTR GAMMA model of substitution, on an alignment of concatenated core CDS as determined by Roary. (B) PANINI plot of the accessory genome content of all 862 strains based on a tSNE plot . The plot is a diagrammatical representation of the relatedness of each strain based on the presence/absence of accessory genes, and is presented as a two dimensional representation. The taxa are colour coded by BAPS grouping (Table S1) and show clade A (Green, BAPS-3), clade B (red, yellow and purple – BAPS 2, 4, and 5) and clade C (blue, BAPS-1).



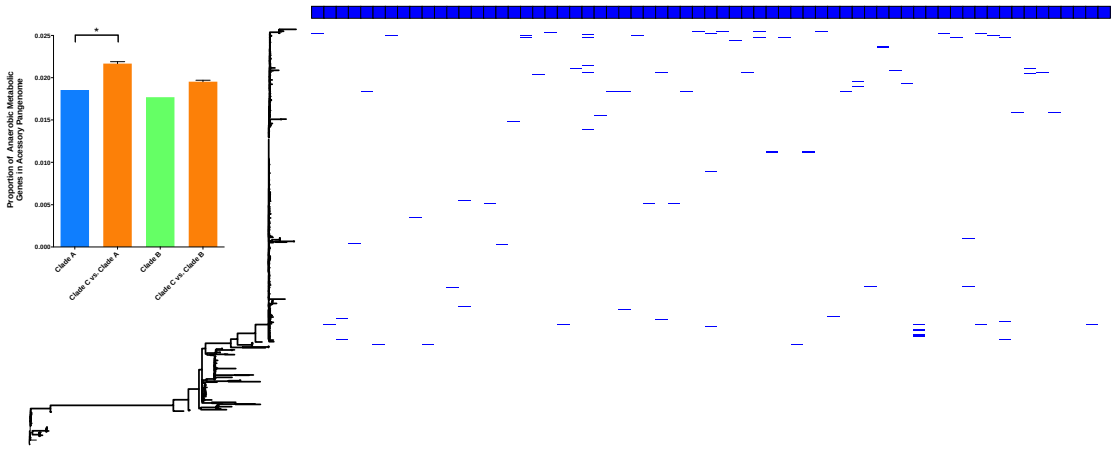
810 Figure 4: Bar chart depicting functional classes of accessory genes differentially
811 present in clade A (blue bars) and clade B/C (orange bars) *E. coli* ST131. Functional
812 classes are based on GO classes as described in methods. Bars marked with *
813 indicate where a significant difference exists between clade A and clade C as
814 determined by t-test.



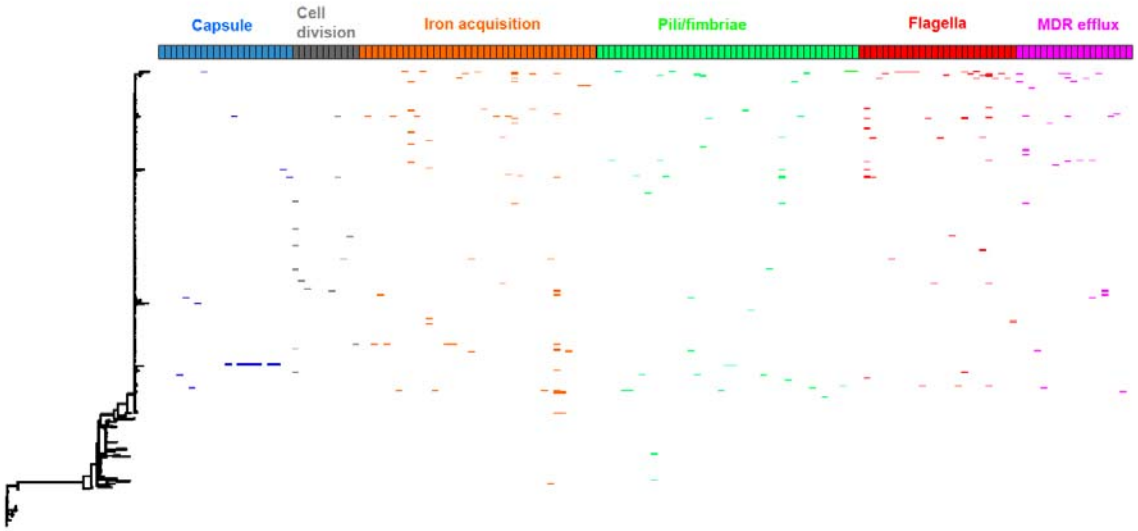
815

816

Figure 5: Annotation of a maximum likelihood phylogeny of *E. coli* ST131, based on concatenated core CDS, with the presence of alternative alleles of 64 loci involved in anaerobic metabolism. Each blue box along the top of the tree annotation represents an individual anaerobic metabolism gene, and its presence in the ST131 population is indicated by a blue line. The inset is a bar chart displaying the proportion of the accessory pangenome that is occupied by genes involved in anaerobic metabolism for ST131 Clade A (light blue), Clade B (light green), subsampled Clade C vs. Clade A (orange) and subsampled Clade C vs. Clade B (orange). P = 0.042 for Clade C vs. Clade A and P = 0.086 for Clade C vs. Clade B. Error bars represent standard error of the mean. Significance was determined using the median value p-value from Chi squared tests performed on random subsamples of the C clade.



830 Figure 6: Annotation of a maximum likelihood phylogeny of *E. coli* ST131, based on
831 concatenated core CDS, with the presence of alternative alleles of loci involved in
832 capsule production (blue boxes), cell division (grey boxes), iron acquisition (orange
833 boxes), pili/fimbriae production (green boxes), flagella (red boxes), and MDR efflux
834 pumps (pink boxes). Each box represents an individual gene, and its presence in the
835 ST131 population is indicated by an appropriately coloured line.

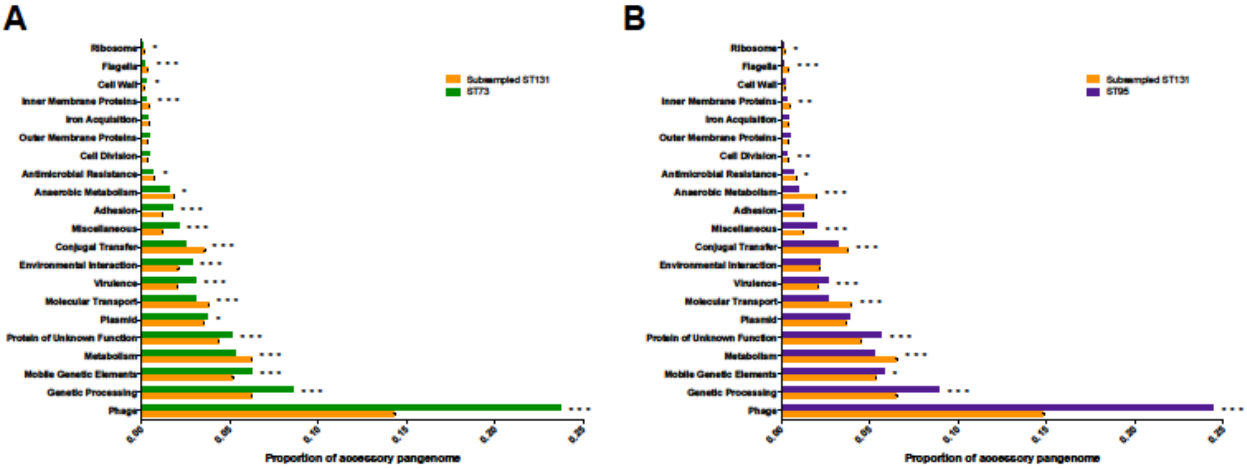


836

837

838 Figure 7: Bar charts depicting the composition of the accessory genome of ST73
839 (green) and ST95 (purple) compared to a repetitively sampled Clade C ST131
840 (orange). The proportion of the accessory genome is plotted against manually
841 assigned functional categories. Hypothetical proteins are responsible for the majority
842 of the accessory pan genome and are omitted from the graphs. Error bars are
843 standard error of the mean. Iterative Chi squared tests were performed to assess
844 significance, as described in methods, p<0.05 (*), p<0.01 (**) and p<0.001 (***).

845



846
847

Figure S1: Histogram of the relative frequency of genes within the accessory genome of *E. coli* ST131. The x-axis indicates the relative frequency with which a gene appears, whilst the y-axis indicates the number of accessory genes which appear at that given frequency.

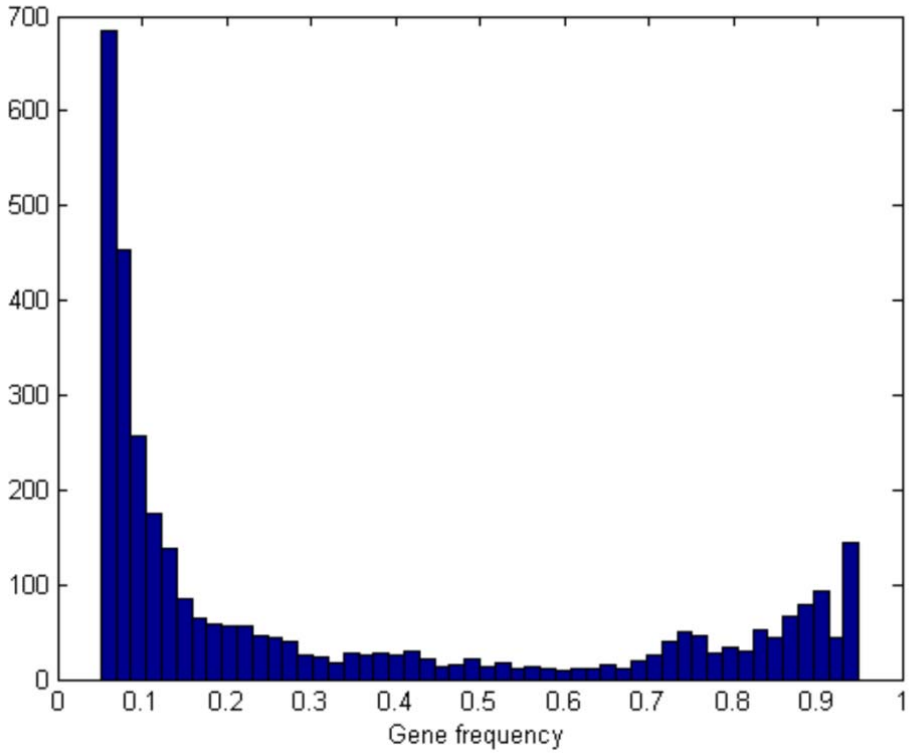


Figure S2: Correlations of gene frequencies in the BSAC collection over time. Each plot shows the frequencies of those genes, identified by ROARY, that were found to be present at a mean frequency between 0.05 and 0.95 across the entire collection. In each panel, the horizontal axis shows the frequency in 2001, and the vertical axis shows the frequency in a subsequent year. These graphs show how the correlation between the starting frequencies, in 2001, and later years weakened until 2008, at which point the correlation strengthen considerably in 2010 and 2011.

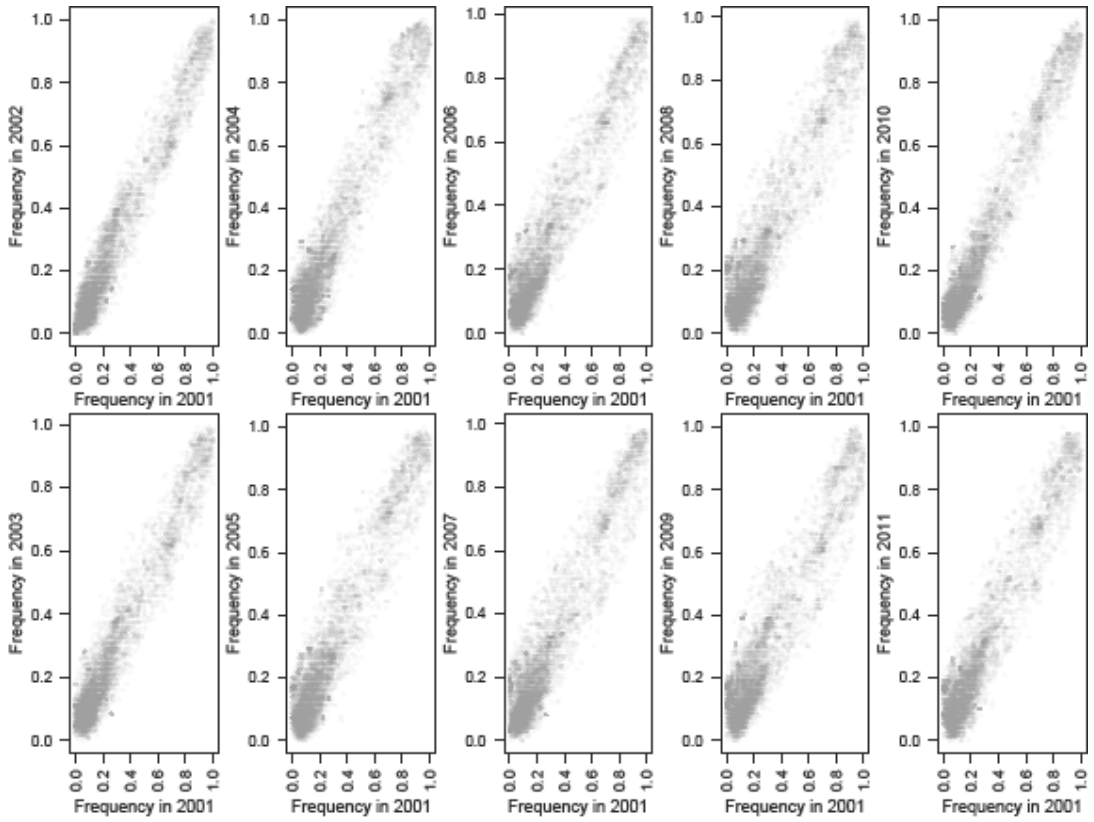
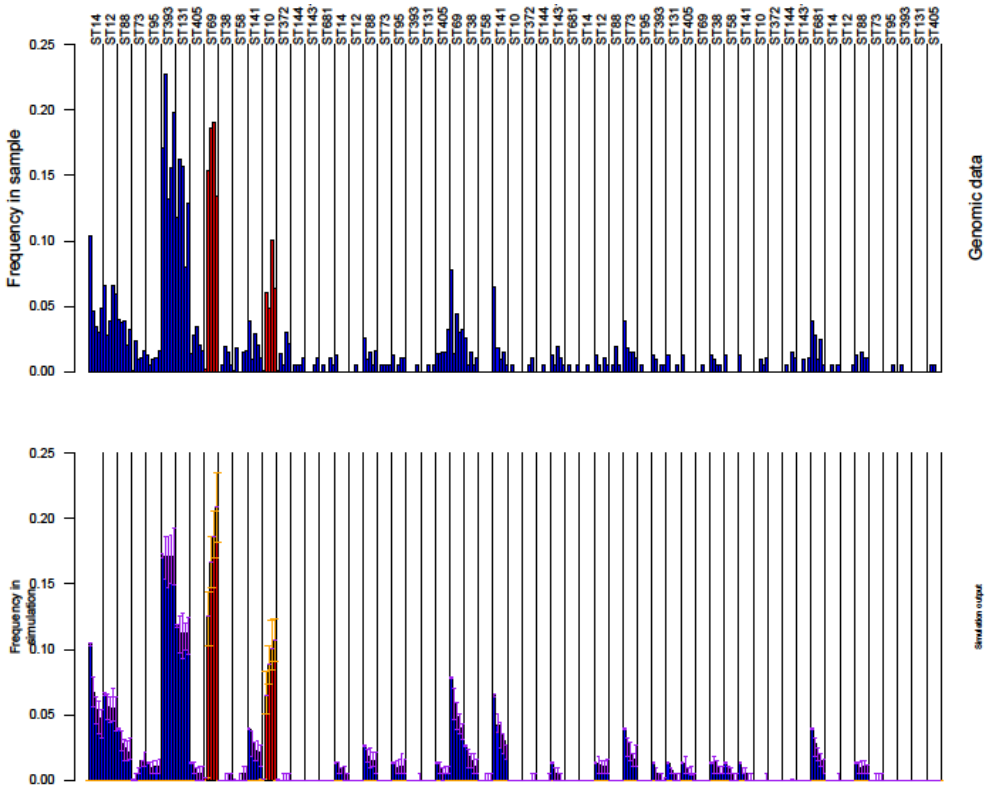
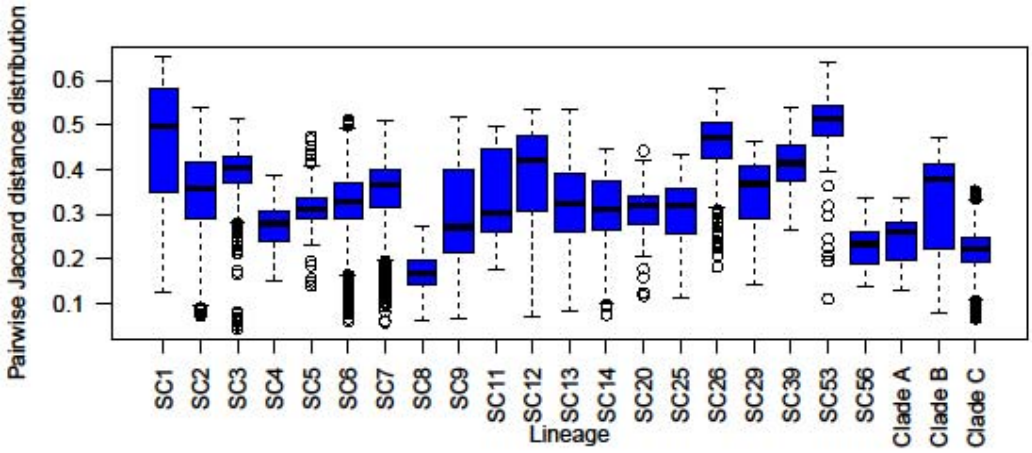


Figure S3: Full results of the NFDS simulations. These barcharts show the frequencies for all lineages from the one hundred simulations performed using the optimal parameters identified within the BOLFI model fitting, which are summarised in Fig 2. Each column again corresponds to a sequence cluster, and is annotated according to the predominant sequence type. The five bars within each column represent the frequency of the sequence cluster over subsequent time intervals: either that observed in the genomic samples for the top panel, or the median frequency in simulations in the bottom panel. The error bars on the bottom panel indicate the interquartile range for each bar from the 100 simulations. The red bars correspond to the ST69 and ST131 sequence clusters that had a reproductive fitness benefit, *r*, over the rest of the population.



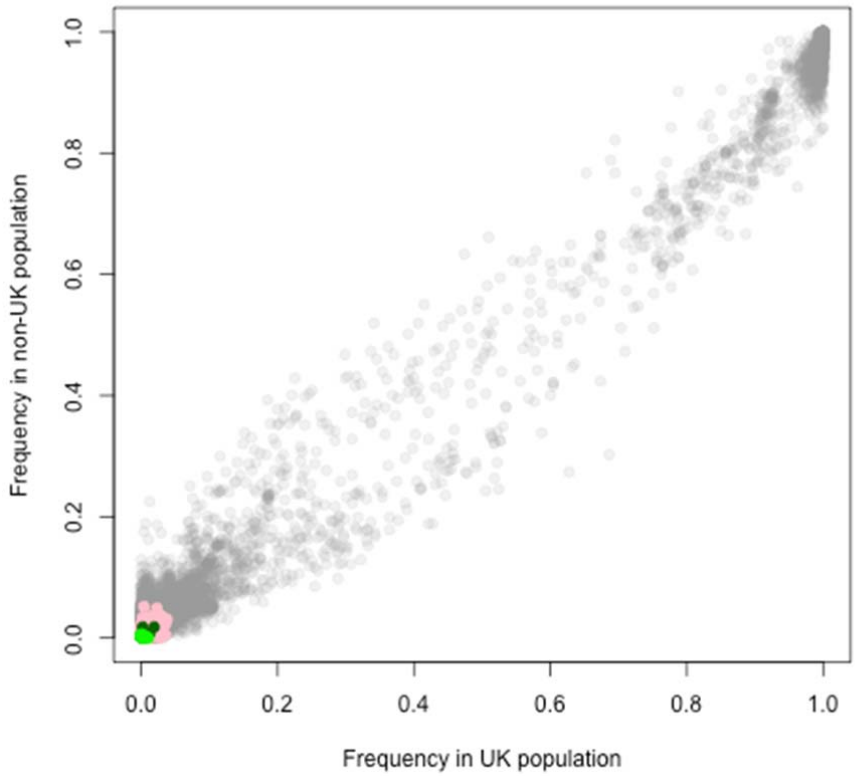
875 Figure S4: Diversity of intermediate frequency loci within *E. coli* lineages. The
876 dissimilarity between pairs of isolates was measured as the binary Jaccard distance
877 between them, based on the presence or absence of the intermediate frequency loci
878 simulated in the multilocus NFDS model. The genetic diversity of each sequence
879 cluster represented by at least ten isolates in the BSAC collection, and the three
880 clades of the ST131 *E. coli*, are represented by a boxplot that shows the distribution
881 of all such pairwise comparisons within the sequence cluster. This demonstrates the
882 success of ST131 cannot be attributed to it exhibiting a greater diversity of loci under
883 selection in the model relative to other lineages.

884



885

886 Figure S5: Frequency dependence plot showing the frequency at which all *E. coli*
887 ST131 accessory genes occur in strains isolated from the UK versus strains isolated
888 from outside the UK. The allele variants identified colour coded as in the previous
889 figures: anaerobic metabolism (blue boxes), capsule production (pale blue boxes),
890 cell division (black boxes), iron acquisition (orange boxes), pili/fimbriae production
891 (green boxes), flagella (red boxes), and MDR efflux pumps (pink boxes)



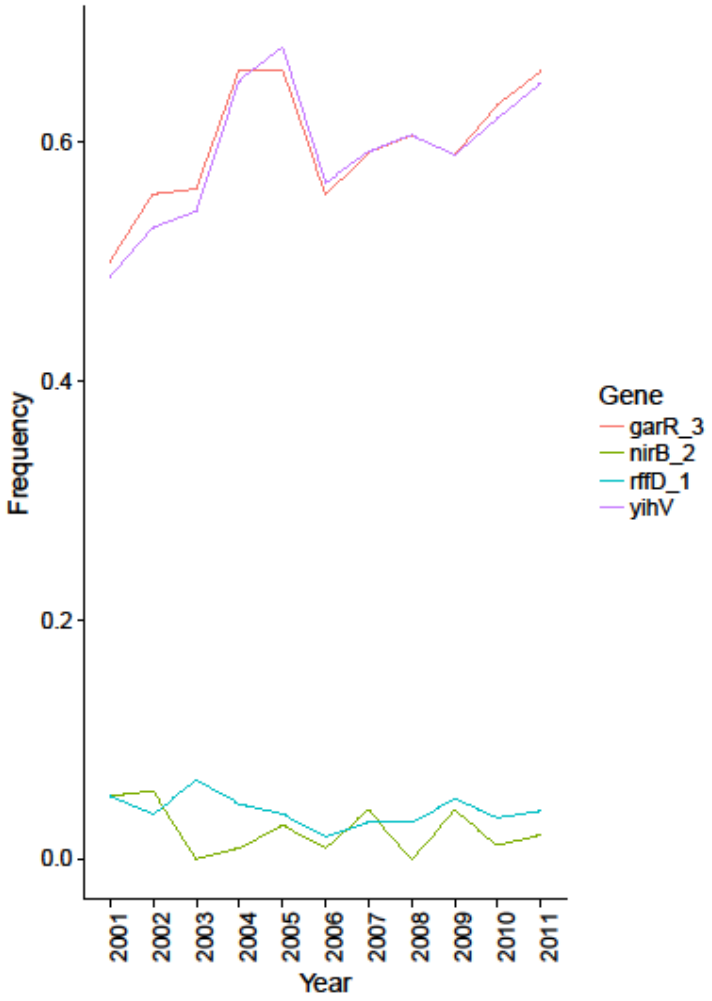
892

893

894 Figure S6: Stable intermediate frequencies of anaerobic metabolism loci. Four genes
895 involved in anaerobic metabolism were found to be present at intermediate
896 frequencies in the BSAC collection. All were absent from the ST131 lineage, except
897 nirB_2, which was found in a subset of the lineage. Nevertheless, plotting their
898 annual frequencies reveals distinct, stable frequencies over the period, despite the
899 rise to prominence of ST131.

900

901



902

903