Genome evolution and pathoadaptation of *Shigella*

1.1 Author names

Zaira Seferbekova[1,2], Alexey Zabelkin[3,4,5], Yulia Yakovleva[5,6], Robert Afasizhev[2], Nikita Alexeev[3], Mikhail S. Gelfand[2,7,#], Olga O. Bochkareva[2,8,#]

1.2 Affiliation

[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia
[2] Institute for Information Transmission Problems (the Kharkevich Institute, RAS), Moscow, Russia
[3] Computer Technologies Laboratory, ITMO University, Saint-Petersburg, Russia
[4] JetBrains Research, Saint-Petersburg, Russia
[5] Bioinformatics Institute, Saint-Petersburg, Russia
[6] Department of Cytology and Histology, Saint-Petersburg State University, Russia
[7] Skolkovo Institute of Science and Technology, Moscow, Russia
[8] Institute of Science and Technology (IST Austria), Klosterneuburg, Austria

1.3 Corresponding authors

[#] OOB: olga.bochkaryova@gmail.com, MSG: mikhail.gelfand@iitp.ru

1.4 Keyword

Shigella, *Escherichia coli,* genomic rearrangements, pathogens, recombination, IS, *ipaH*

2. Abstract

The genus *Shigella* comprises a polyphyletic group of facultative intracellular pathogens that evolved from *Escherichia coli*. *Shigella* genomes have accumulated mobile elements, which may have been caused by decreased effective population size and concomitant reduction of purifying selection that allowed their proliferation. Here, we investigated the interplay of the accumulation of genomic repeats with genomic rearrangements and their impact on adaptation in bacterial evolution. We studied 414 genomes of *E. coli* and *Shigella* strains to assess the contribution of genomic rearrangements to *Shigella* pathoadaptation. We show that *Shigella* accumulated a variety of insertion sequences (ISs), experienced exceptionally high rates of intragenomic rearrangements and had a decreased rate of homologous recombination. IS families differ in the expansion rates in *Shigella* lineages, as expected given their independent origin. In contrast, the number of IS elements and, consequently, the rate of genome rearrangements in the enteroinvasive *E. coli* strain (EIEC) strain are comparable to those in other *E. coli*. We found two chromosomal E3 ubiquitin-protein ligases (putative IpaH family proteins) that are functional in all *Shigella* strains, while only one pseudogenised copy is found in the EIEC strain and none in other *E. coli.* Taken together, our data indicate that ISs played an important role in the adaptation of *Shigella* strains to a intracellular lifestyle and that the composition of functional types of ubiquitin-protein ligases may explain the differences in the infectious dose and disease severity between *Shigella* and EIEC pathotypes.

3. Impact statement

Pathogenic *Escherichia coli* frequently cause infections in humans. Many *E. coli* exist in nature and their ability to cause disease is fueled by their ability to incorporate novel genetic information by extensive horizontal gene transfer of plasmids and pathogenicity islands. The emergence of antibiotic-resistant *Shigella*, which is a pathogenic form of *E. coli*, coupled with the absence of an effective vaccine against them, highlights the importance of continued study of these pathogenic bacteria. Our study contributes to the understanding of genomic properties associated with molecular mechanisms underpinning the pathogenic nature of *Shigella*. We show the contribution

of insertion sequences in adaptation of these intracellular pathogens and indicate a role of chromosomal *ipaH* genes in *Shigella* pathogenesis. The approaches developed in our study are broadly applicable to investigation of genotype-phenotype correlation in historically young bacterial pathogens.

4. Data summary

The datasets supporting the conclusions of this article and used scripts are available at https://github.com/zseferbekova/ShigellaProject. The authors confirm all supporting data have been provided in the article or as supplementary data files.

5. Introduction

Escherichia coli is likely the best studied organism, at least on the molecular level. It is widely used to study fundamental aspects of bacterial genomics and is the subject of extensive research as an important component of the normal gut microbiota of vertebrates, including humans. While most *E. coli* strains are harmless, a non-negligible fraction are pathogenic, causing such diseases as diarrhoea, urinary tract infection, or meningitis (1,2). Another group of pathogens, *Shigella*, which causes a severe form of bacillary dysentery, evolved from *E. coli* (3,4). *Shigella* spp. are polyphyletic relative to *E. coli*, but the genus name is maintained in part due to the medical tradition (5–7). Nevertheless, from an evolutionary perspective, *Shigella* is just a set of strains causing a specific disease within the broader *E. coli* phylogenetic group.

Shigella strains are distinguished from *E. coli* by their nonmotility with the associated absence of decarboxylated lysine, and by various biochemical characteristics, such as inability to ferment lactose and mucate (8). One more *E. coli* pathovar, enteroinvasive *E. coli* (EIEC), generally exhibits the same pathogenic and biochemical features as *Shigella*, including invasiveness that depends on the virulence plasmid (9–14). Such phenotypic similarity may be attributed to adaptation to similar environmental conditions as *Shigella* and EIEC spend most of their lifecycle inside eukaryotic cells, while most *E. coli* strains inhabit extracellular space. Thus, EIEC could represent either a non-toxin-producing *Shigella* 'prototype', which could be a precursor for a typical *Shigella*, or a distinct group of pathogenic *E. coli* that have adapted to an intracellular lifestyle but, unlike *Shigella*, have not lost the ability to live outside eukaryotic cells (15–17).

The acquisition of the virulence plasmid enabling intracellular lifestyle was likely a key event of *Shigella* evolution that facilitated further pathoadaptation. This pathoadaptation may have incorporated a variety of events, such as point mutations, suppression of certain genes, deletion of anti-virulence genes, or acquisition of insertion sequences. On the other hand, the intracellular niche may have provided a more relaxed selective pressure due to abundant resources and a lack of competitors (18,19) that, in conjunction with lower effective population size, would have decreased the negative selection rate (20) and caused substantial changes in its genome arrangement and composition.

The chromosome and plasmids of *Shigella* species contain many insertion sequences (ISs) — small mobile DNA fragments that translocate throughout the genome. Such events often lead to genome rearrangements and changes in the expression of adjacent genes (21–24). The type of rearrangement depends on the mutual arrangement of the repetitive elements that have been involved in the recombination event. Recombination between inverted repeats leads to inversions, recombination between direct repeats leads to deletions, and recombination between direct repeats during replication leads to duplication (25). Since large deletions, insertions, and duplications are often under negative selection and are rare, inversions are the main drivers of structural changes in bacterial chromosomes (26). The frequency of rearrangements varies and may correlate with the presence of mobile genetic elements such as ISs (27,28).

Genome rearrangements may exert a profound effect on the phenotype by changing gene expression and function (29,30). Gene deletions may also contribute to the specialization of bacteria, for example, by eliminating genes that have been essential in the original environment but are not necessary after the transition to an intracellular lifestyle. For instance, *Burkholderia mallei*, a young obligate intracellular pathogen, has lost numerous clusters of genes through IS-mediated elimination as demonstrated by the comparison of its genome to strains of the ancestral species, *Burkholderia pseudomallei (31,32)*. *Shigella* genomes also feature loss or inactivation of many genes, which has been attributed to the relaxation of selection acting on those genes (33). Changes in gene content enable rapid adaptation to different conditions in the host cell (34,35).

Here, we provide a comprehensive analysis of the genomes of *E. coli* and *Shigella* strains and assess the contribution of genomic rearrangements to *Shigella* pathoadaptation. We show that *Shigella* genomes accumulated a large variety of ISs and experienced exceptionally high rates of intragenomic rearrangements in comparison to other, in particular, pathogenic *E. coli*. We also focused on rates of expansion of different ISs families and described rearrangements that have occurred independently in separate lineages of *Shigella*. Finally, we compared the rates of homologous recombination in *E. coli* and *Shigella* strains.

6. Methods

1. Genomes

We used all complete and annotated genomes of *Shigella* and *E. coli* available in GenBank as of April 2019 (36). We constructed a phylogenetic tree for *E. coli* strains only (see Methods below) and excluded clones and closely related *E. coli* strains that formed plateaus in the tree from further analysis (Supplementary Figure 1, the list of excluded genomes is available on GitHub: https://github.com/zseferbekova/ShigellaProject/1Tree/Data/excluded_strains.csv). In total we analysed 414 complete genomes, including 35 *Shigella* spp., 41 STEC, 31 ExPEC, 8 APEC, 7 ETEC, 3 EPEC, 3 AIEC, 2 EAEC, and 1 EIEC genome. More details are available on GitHub: https://github.com/zseferbekova/ShigellaProject/1Tree/Data/total_stats.csv.

2. Orthologs

To construct orthologous groups, we used ProteinOrtho V5.13 (37) with parameters cov=67 (minimal coverage of the best BLAST alignments in percent) and identity=50 (minimal percent identity of the best BLAST alignments).

3. Phylogenetic tree

For the construction of the strains' phylogenetic tree, we used 238 universal, single-copy orthologous groups found in all 414 genomes. We constructed a nucleotide multiple sequence alignment of genes in each orthologous group using Mafft (38) in the linsi mode. We then used RAxML (39) with the GTR+Gamma model and 100 bootstrap replicates to construct a phylogenetic tree based on the concatenated alignment of these genes. Finally, we used the GGRaSP (40) R package to divide all strains in the tree into seven clusters corresponding to the standard phylogroups (41). All trees were visualised using online iTOL (42).

4. IS elements

Online tool ISsaga (43) was used for annotation of IS elements in the chromosome sequences. For general statistical calculations, we used the number of predicted ORFs associated with ISs. For more precise annotation of ISs from different families, we used the number of predicted ISs that could include several ORFs.

5. Synteny blocks

Synteny blocks for each cluster were constructed using the Sibelia software (44,45). We used common single-copy synteny blocks with the length threshold of 5 kb to infer the number of inversions. In each phylogenetic cluster, we used distance matrices where each element is a number of synteny blocks between given strains. We constructed trees based on the obtained matrices using PHYLIP (46,47) and the neighbour-joining algorithm. We used synteny blocks with the length threshold of 1 kb for the detection of parallel events across the phylogenetic tree. Chromosome maps were visualised using Circos (48).

6. Inversions scenario

We reconstructed the history of inversion events using MGRA (49). This tool takes as an input a phylogenetic tree and genomes represented as sets of synteny blocks. This analysis included chromosomes of all *Shigella*, related *E. coli* and several representative *E. coli* from each cluster not containing *Shigella* (51 genomes in total).

7. Parallel rearrangements

To analyse inversions, we considered common single-copy blocks. For these 377 blocks, we constructed the breakpoint graph (50) as follows. The graph is built on 377×2 vertices. For each block *B,* we introduce two vertices B_H and B_T, its head and tail, respectively. If two blocks *B* and *C* are adjacent in genome *g*, the vertices B_H and C_H are linked by an edge of colour *g*. We note that since we consider only common blocks and all genomes are circular, the edges of each colour form a perfect matching on the graph vertices. Since some adjacency edges of different colours may link the same pair of vertices, we introduce multi-edges — a multi-edge is a set of parallel edges of different colours. The breakpoint graph for our data contains 754 vertices and 656 multi-edges. If two strains differ from each other by one inversion, this corresponds to a 4-cycle in the breakpoint graph (51). For each multi-edge, we split the set of strains into patterns depending on the presence of the corresponding colour edge in the multi-edge. Thus we associated each breakpoint (and each inversion) with a partition of the set of strains into patterns.

To analyse indels and duplications, we considered all blocks which were present in different copy numbers in some strains. For each block, we split the set of strains into patterns so that the copy number of this block in each pattern was the same (i.e. the pattern with strains containing 0 copies of the block, the pattern with strains containing 1 copy of the block, etc). Thus, we associated each copy number variation with a partition of the set of strains into patterns.

8. Rates of homologous recombination

For this analysis we considered genomes from phylogroup B1 as it contains most of the *Shigella* strains. Pairwise full-genome alignments were performed using MAUVE (52). Then for fragments without gaps, we calculated the number of non-identical columns in each 1 kb segment. Thus, for each pair of genomes, we constructed the distributions of the number of mutations across the genomes.

In the case of strictly vertical inheritance, this distribution would be Poisson with the parameter λ reflecting the time of strain divergence. Fragments transferred horizontally from distant strains would contain more mutations yielding deviation from the Poisson distribution in the form of a heavy tail (53). The latter, being a mixture of the Poisson distributions with unknown parameters may be fitted by the Erlang distribution.

We have used the Python SciPy module (54) to fit all pairwise distributions by the function $F_{\lambda,k,\mu,W}(x) = W\times\mathbf{P}_\lambda(x)+(1-W)\times\mathbf{E}_{k,\mu}(x)$, where $\mathbf{P}_\lambda(x) = e^{-\lambda}\lambda^x/x!$ is the Poisson distribution with parameter λ, $\mathbf{E}_{k,\mu}(x) = (x/\mu)^{k-1}e^{-x/\mu}/(\mu(k-1)!)$ is the Erlang distribution with the shape *k* and scale μ (mode=$(k-1)\mu$, mean=$k\mu$, variance=$k\mu^2$), and the weight *W* in the range [0,1] measures the vertically inherited fraction of genome while $(1-W)$ corresponds to horizontally transferred fraction.

7. Results

1. Structure of the phylogenetic tree of *E. coli* and *Shigella* spp.

We found 238 universal single-copy orthologous groups in 414 genomes and used them to construct the phylogenetic tree shown in Fig. 1. The tree is unrooted as we were interested only in the relationships between strains within *E. coli* phylogroups.

The structure of the obtained phylogenetic tree recapitulates seven known *E. coli* phylogroups — A, B1, B2, C, D, E, F (41,55). The phylogeny supports the hypothesis that *Shigella* spp. evolved several times independently from *E. coli.* Most *Shigella* strains form clusters in phylogroup B1, with a notable exception of two *S. dysenteriae* strains from phylogroup E, as reported previously (56). In contrast, other pathogenic *E. coli* strains did not form a separate cluster. The only complete and annotated EIEC strain did not cluster with any *Shigella (57)*. One of the *Shigella* genomes (GenBankID: GCA_001596115.1) was unclassified and had an atypical location in the phylogenetic tree. Moreover, the source of the sample was lichen, which is highly unusual and unlikely for *Shigella*. Thus, we assumed that in this case the taxonomic annotation was wrong and did not consider this genome as *Shigella*.

Shigella have a significantly smaller genome size than all pathogenic and non-pathogenic *E. coli* (Fig. 2a). Genome reduction is a well-known feature of many pathogenic and symbiotic bacteria (58). By contrast, the genome size of the two most abundant *E. coli* pathovars (ExPEC and STEC) is larger than that of non-pathogenic *E. coli* (Fig. 2b, Supplementary Figure 2). These data may indicate that for *Shigella* spp. the process of gene loss has already started, whereas

ExPEC and STEC have only recently acquired various genetic elements that made them pathogenic. *E. coli* BLK9 and *E. coli* BLK16 derived from a strain routinely used for high level plasmid-based recombinant protein production have unusually short genomes. Indeed, to improve the genetic stability of bacterial hosts resident prophages and active IS elements were removed from their genome resulting in a downsized and clean genome (59).

2. Accumulation of IS elements

Shigella genomes generally encode more IS elements than non-pathogenic *E. coli* (60). To estimate the density of IS elements, we calculated their number in the chromosomes of all 414 strains (Fig. 1). The number of IS elements in *Shigella* genomes was significantly higher than in the genomes of pathogenic and non-pathogenic *E. coli* strains (Fig. 3a; the Wilcoxon-Mann-Whitney test, *p*-value = 2.4×10^{-21} and 7.8×10^{-18}, respectively). Interestingly, the two most sequenced pathovars, STEC and ExPEC, had, respectively, significantly higher ($p = 1.9\times10^{-4}$) and lower ($p = 3.5\times10^{-4}$) number of IS elements than the average in non-pathogenic *E. coli* (Fig. 3b).

To test whether the distribution of IS families differed in two clusters of *Shigella* we merged IS elements into larger families and mapped the results on the phylogenetic tree (Fig. 4, 5). We found four IS families (IS1, IS3, IS4, IS91) which were enriched in all *Shigella* in comparison to *E. coli*. Moreover, we detected IS families that were specific for some *Shigella* lineages (IS21, IS110, IS630, IS66). However, as our dataset included only three *S. boydii* and two *S. dysenteriae* genomes, the results for these lineages should be considered as preliminary. By contrast, the *S. sonnei* group was represented by 11 strains and showed a significantly higher number of IS21, IS110 and IS630 elements (the Wilcoxon-Mann-Whitney test, $p = 1.3\times10^{-8}$, 5.1×10^{-7}, 1.9×10^{-9}, respectively).

3. Rates of rearrangements

Shigella genomes are thought to be dynamic due to numerous IS elements that promote non-homologous recombination (12,21,24). Thus, the rate of rearrangement in *Shigella* genomes may be higher in proportion to the accumulation of single nucleotide substitutions. To test this possibility we compared the number of syntenic blocks with the number of single nucleotide substitutions in the universal genes for pairs of genomes in each phylogroup. The number of common single-copy synteny blocks between two genomes was inversely proportional to collinearity of the genomes, i.e. more collinear genomes had fewer blocks while more blocks corresponded to genomes with a large number of inversions. While the number of syntenic blocks may not accurately reflect the number of rearrangement events, especially when the number of such events is substantial, our approach provides a lower bound estimate of the number of inversion events. We then plotted the number of syntenic blocks relative to the number of single nucleotide substitutions for each pair of genomes (Supplementary Fig. 3, 4).

Indeed, within the same interval of of evolutionary distances between strains, in pairs of *Shigella* and *E. coli* strains ratio of number of syntenic blocks to single nucleotide substitutions was substantially higher, in comparison to pairs of *E. coli* strains (the Wilcoxon-Mann-Whitney test, $p=2.22\times10^{-16}$) (Fig. 6, Supplementary Fig. 3). Thus, genome rearrangements occurred more frequently in *Shigella* history compared to *E. coli*. Interestingly, four *Shigella* lineages showed different rates of rearrangements (Supplementary Fig. 5). In other pathogenic *E. coli* ratio of number of syntenic blocks to single nucleotide substitutions did not differ from the average in non-pathogenic strains (Supplementary Fig. 4). The unclassified *Shigella* strain from lichen did not differ from *E. coli* strains (Supplementary Fig. 4b), further supporting our assumption that this strain has been misclassified.

We reconstructed the history of inversions in 51 strains. This dataset includes all *Shigella* spp., closely related *E. coli* and several representative *E. coli* from each cluster without *Shigella* (Fig. 7). Of 114 reconstructed inversion events, 103 were mapped to *Shigella* branches, while only 11 corresponded to *E. coli* branches. Of these inversions, 33 were mapped to the branch separating two *S. dysenteriae*; three other *Shigella* species demonstrated rearrangements at internal and terminal branches. These results are consistent with the estimation of rearrangement rates using the number of syntenic regions as the indicator.

4. Shigella-specific rearrangements

To analyse the impact of genome rearrangements on *Shigella* speciation, we focused on inversions, insertions, deletions and duplications that occurred independently in the *Shigella* lineages. Events that occur multiple times independently on a phylogeny, called homoplasies,

could indicate selection pressure acting on populations adapting to an intracellular lifestyle.

In all studied *Shigella* and *E. coli* genomes we found 377 universal synteny blocks, such that each block was found in all of the considered genomes. We also found 3642 synteny blocks that were not classified as universal because the block was not found, or found more than once, in at least one genome. The focus on presence or absence of synteny blocks allowed us to distinguish between genome rearrangements not affecting copy numbers (such as inversions) and those leading to copy number variations (such as insertions, deletions, and duplications). We then separately constructed the breakpoint graph for universal blocks and analysed phyletic patterns of non-universal blocks (see Methods). Thus we identified genetic features that could not be parsimoniously explained by common ancestry.

The universal blocks spanned only 25-29% of the chromosome length, moreover, the distribution of these blocks was not uniform with long sections of the chromosome not harboring any universal blocks (Fig. 8a). For the universal synteny blocks, we found 31 parallel inversions and 30 recombination hotspots where a region was involved in various inversions several times independently. We focused on the regions that were syntenic in *E. coli* but involved in recombination events in *Shigella* (Supplementary Fig. 6a). These inversions were flanked by different transposases (IS1, IS3, IS4) that were independently integrated into the same locus in different *Shigella* lineages (Fig. 8b).

We identified parallel insertions, deletions and duplications in 2256 out of 3642 non-universal synteny blocks. *Shigella* adaptation had been accompanied by massive pseudogenization (61), therefore, we anticipated to see an increased rate of loss of non-universal synteny blocks in *Shigella*. However, we did not find any strictly *Shigella*-specific large-scale deletion events.

Interestingly, we found one non-universal synteny block that was present in *Shigella* but absent in *E. coli*, likely indicating acquisition of this block in *Shigella* rather than multiple independent losses in *E. coli*. This region contained a gene encoding an E3 ubiquitin-protein ligase, one of *Shigella* invasion-plasmid antigens, with a highly conserved 250 bp upstream region (Supplementary Fig. 6b). Moreover, in *S. flexneri* chromosomes we identified two copies of this block, located at a substantial distance from each other. In several strains, one of the copies contained a pseudogenized gene (*ipaH0887/ipaH2610)*. One more E3 ubiquitin-protein ligase gene (*IpaH1880*) with conserved upstream region was shared by all *Shigella* and EIEC chromosomes (Supplementary Fig. 6b), but in the EIEC strain this gene had been disrupted by a recent insertion. The evolutionary history of this syntenic block seems relevant to the functional specificity of *Shigella* and EIEC as the products of *IpaH* genes are secreted by intracellular bacteria via the type III secretion system (T3SS) (62) and, therefore, may confer *Shigella*- or EIEC-specific functionality.

5. Rates of homologous recombination

We hypothesized that disruption of syntenic regions should decrease the rate of homologous recombination. To check this, we calculated fractions of horizontally transferred fragments in strains from the B1 phylogroup using pairwise genome alignments of *E. coli*, *S. flexneri*, *S. boydii,* and *S. sonnei* (see Methods). Indeed, within the same interval of evolutionary distances between strains (with the Poisson λ in the vertically inherited fraction ranging from 0 to 1.45), pairs of *Shigella* strains had a significantly lower fraction of horizontally transferred fragments, $(1–W)=0.094±0.017$, in comparison to pairs of *E. coli* strains, $(1–W)=0.612±0.002$ ($p=4.17×10^{-124}$, the Wilcoxon test) (Fig. 9).

8. Discussion

Compared to other pathogenic *E.coli*, *Shigella* (i) accumulated a large variety of ISs, (ii) acquired new chromosomal genes, (iii) experienced exceptionally high rates of intragenomic rearrangements and (iv) had a decreased rate of homologous recombination.

The diversity of *E. coli* pathotypes is explained by the high plasticity of its genome, as genes responsible for pathogenic traits are mostly acquired by extensive horizontal gene transfer and are often conveyed by mobile genetic elements (63–65). Both *Shigella* and EIEC spend much of their life cycle within eukaryotic cells and share many invasion-related functional systems. The adaptation to an intracellular lifestyle was conferred by the acquisition of the pINV plasmid encoding a type III secretion system (T3SS) (62). Phylogenetic analysis supports the hypothesis that four existing *Shigella* lineages and the EIEC strain have arisen from different ancestral *E. coli* strains on multiple independent occasions (13,14,55,66).

Two chromosomal E3 ubiquitin-protein ligases (putative IpaH family proteins) appeared functional in all *Shigella* strains, while only one pseudogenised copy was found in the EIEC strain. The differences in the composition of functional types of ubiquitin-protein ligases may explain the differences in the infectious dose and disease severity between *Shigella* and EIEC pathotypes (67) Indeed, the delivery of bacterial virulence proteins, called “effectors,” into host cells via T3SS plays a crucial role in the infection strategies of *Shigella*. Such effectors are involved in the reorganization of the host cell cytoskeleton and in the modulation of cellular signalling pathways that allow the bacteria to evade the host immune response (68). Each IpaH family protein likely has a specific host target protein due to the substrate recognition domain, and thus makes a distinct contribution to bacterial pathogenesis (62). Most T3SS effectors are encoded by plasmid *pINV* genes, while the biological role of chromosomally encoded *ipaH* genes remains obscure. Based on presented analysis we hypothesize that the absence and/or disruption of chromosomal *ipaH* in the EIEC strain might be a factor that distinguishes *Shigella* and EIEC pathogenesis.

High numbers and the variety of mobile elements in evolutionarily young intracellular pathogens is usually explained by weaker selection against repetitive elements due to decreased effective population size (20,69,70). IS elements may drive the evolution of chromosome organisation by causing intragenomic recombination (71). Genome rearrangements change the chromosomal architecture, which can alter gene composition and gene expression. These events are largely detrimental for free-living bacteria and are expected to be selected against (25). On the other hand, in a new environment intragenomic recombinations and the associated functional changes may provide a base for adaptation. For example, the accumulation of IS elements is likely responsible for frequent genome rearrangements leading to substantial genome reduction in *B. mallei* (31). *Shigella* strains have smaller genomes than other *E. coli*, which cannot be explained with a small number of large-scale deletions in *Shigella*. The boundaries of inversions in *Shigella* are usually flanked by transposases, thus it is likely that the high rate of inversions is associated with intensive integration of ISs.

IS families differ in the expansion rate in *Shigella* lineages, which is expected given the independent origin of these lineages. In contrast to *Shigella*, the number of IS elements and, consequently, the rate of genome rearrangements in the EIEC strain were comparable with those in other *E. coli*. These differences in the number of IS genomic elements may have influenced different stages of pathoadaptation of EIEC and *Shigella* spp.

An expected consequence of frequent genome rearrangement is a decrease in the rate of homologous recombination. Indeed, *Shigella* genomes contain fewer horizontally transferred DNA segments in comparison to *E. coli*. However, this also could be explained by a smaller population size and isolated lifestyle of *Shigella* strains.

Due to high genomic plasticity, pathogenic *E. coli* are among the most frequent causes of bacterial infections in humans (67,72). In particular, the *E. coli* O104:H4 outbreak in Germany in 2011 was caused by a strain that had acquired characteristics of two previously described pathotypes (73). The emergence of antibiotic-resistant *Shigella* with the absence of an effective vaccine highlights the importance of a detailed investigation of this pathogen (74). Our results contribute to the understanding of genomic properties associated with pathoadaptation of *Shigella* and EIEC and the general approaches we applied here should be broadly applicable to other young bacterial pathogens.

9. Author statements

9.1 Authors and contributors

MSG and OOB conceived the study, ZS, OOB, and MSG designed the study; ZS, RA, AZ, and NA developed the methods, ZS, YY, RA, and OOB analyzed the data; ZS, NA, OOB, and MSG wrote the manuscript. All authors read and approved the final version of the manuscript.

9.2 Conflicts of interest

The authors declare that they have no competing interests.

9.3 Funding information

This study was supported by the Russian Science Foundation, grant 18-14-00358. The work of OOB is supported by the European Union’s Horizon 2020 research and innovation programme

under the Marie Skłodowska-Curie grant agreement No 754411. The work of NA is supported by the Government of the Russian Federation through the ITMO Fellowship and Professorship Program. The funding bodies had no role in the design of the study, collection, analysis, and interpretation of data and in writing the manuscript.

9.4 Ethical approval

Not applicable

9.5 Consent for publication

Not applicable

9.6 Acknowledgements

We thank Fyodor Kondrashov for valuable advice and manuscript proofreading. We also thank Alla Mikheenko for assistance with Circos.

10. References

1. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of Shigella evolution, adaptation and geographical spread. Nat Rev Microbiol [Internet]. 2016 Apr;14(4):235–50. Available from: http://dx.doi.org/10.1038/nrmicro.2016.10

2. Zhang Y, Lin K. A phylogenomic analysis of Escherichia coli / Shigella group: implications of genomic features associated with pathogenicity and ecological adaptation. BMC Evol Biol [Internet]. 2012 Sep 7;12:174. Available from: http://dx.doi.org/10.1186/1471-2148-12-174

3. Pupo GM, Lan R, Reeves PR. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics [Internet]. Vol. 97, Proceedings of the National Academy of Sciences. 2000. p. 10567–72. Available from: http://dx.doi.org/10.1073/pnas.180094797

4. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of Shigella evolution, adaptation and geographical spread. Nat Rev Microbiol [Internet]. 2016 Apr;14(4):235–50. Available from: http://dx.doi.org/10.1038/nrmicro.2016.10

5. Pupo GM, Lan R, Reeves PR. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics [Internet]. Vol. 97, Proceedings of the National Academy of Sciences. 2000. p. 10567–72. Available from: http://dx.doi.org/10.1073/pnas.180094797

6. Chaudhuri RR, Henderson IR. The evolution of the Escherichia coli phylogeny. Infect Genet Evol [Internet]. 2012 Mar;12(2):214–26. Available from: http://dx.doi.org/10.1016/j.meegid.2012.01.005

7. Lan R, Reeves PR. Escherichia coli in disguise: molecular origins of Shigella. Microbes Infect [Internet]. 2002 Sep;4(11):1125–32. Available from: http://dx.doi.org/10.1016/s1286-4579(02)01637-4

8. Pupo GM, Lan R, Reeves PR. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics [Internet]. Vol. 97, Proceedings of the National Academy of Sciences. 2000. p. 10567–72. Available from: http://dx.doi.org/10.1073/pnas.180094797

9. Pupo GM, Lan R, Reeves PR. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics [Internet]. Vol. 97, Proceedings of the National Academy of Sciences. 2000. p. 10567–72. Available from: http://dx.doi.org/10.1073/pnas.180094797

10. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of

Shigella evolution, adaptation and geographical spread. Nat Rev Microbiol [Internet]. 2016 Apr;14(4):235–50. Available from: http://dx.doi.org/10.1038/nrmicro.2016.10

11. Lan R, Reeves PR. Escherichia coli in disguise: molecular origins of Shigella. Microbes Infect [Internet]. 2002 Sep;4(11):1125–32. Available from: http://dx.doi.org/10.1016/s1286-4579(02)01637-4

12. Beld MJC, Reubsaet FAG. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli [Internet]. Vol. 31, European Journal of Clinical Microbiology & Infectious Diseases. 2012. p. 899–904. Available from: http://dx.doi.org/10.1007/s10096-011-1395-7

13. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR. Molecular Evolutionary Relationships of Enteroinvasive Escherichia coli and Shigella spp [Internet]. Vol. 72, Infection and Immunity. 2004. p. 5080–8. Available from: http://dx.doi.org/10.1128/iai.72.9.5080-5088.2004

14. Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, et al. Revisiting the Molecular Evolutionary History of Shigella spp [Internet]. Vol. 64, Journal of Molecular Evolution. 2007. p. 71–9. Available from: http://dx.doi.org/10.1007/s00239-006-0052-8

15. Lan R, Reeves PR. Escherichia coli in disguise: molecular origins of Shigella. Microbes Infect [Internet]. 2002 Sep;4(11):1125–32. Available from: http://dx.doi.org/10.1016/s1286-4579(02)01637-4

16. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR. Molecular Evolutionary Relationships of Enteroinvasive Escherichia coli and Shigella spp [Internet]. Vol. 72, Infection and Immunity. 2004. p. 5080–8. Available from: http://dx.doi.org/10.1128/iai.72.9.5080-5088.2004

17. Beld MJC, Reubsaet FAG. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli [Internet]. Vol. 31, European Journal of Clinical Microbiology & Infectious Diseases. 2012. p. 899–904. Available from: http://dx.doi.org/10.1007/s10096-011-1395-7

18. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of Shigella evolution, adaptation and geographical spread. Nat Rev Microbiol [Internet]. 2016 Apr;14(4):235–50. Available from: http://dx.doi.org/10.1038/nrmicro.2016.10

19. Beld MJC, Reubsaet FAG. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli [Internet]. Vol. 31, European Journal of Clinical Microbiology & Infectious Diseases. 2012. p. 899–904. Available from: http://dx.doi.org/10.1007/s10096-011-1395-7

20. Mamirova L, Popadin K, Gelfand MS. Purifying selection in mitochondria, free-living and obligate intracellular proteobacteria. BMC Evol Biol [Internet]. 2007 Feb 12 [cited 2020 Jun 8];7(1):1–12. Available from: https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-7-17

21. Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'Hauteville H, Kunst F, et al. The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of Shigella flexneri. Mol Microbiol [Internet]. 2000 Nov;38(4):760–71. Available from: http://dx.doi.org/10.1046/j.1365-2958.2000.02179.x

22. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res [Internet]. 2005 Nov 7;33(19):6445–58. Available from: http://dx.doi.org/10.1093/nar/gki954

23. Beld MJC, Reubsaet FAG. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli [Internet]. Vol. 31, European Journal of Clinical Microbiology & Infectious Diseases. 2012. p. 899–904. Available from: http://dx.doi.org/10.1007/s10096-011-1395-7

24. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev [Internet]. 2014 Sep;38(5):865–91. Available from: http://dx.doi.org/10.1111/1574-6976.12067

25. Darling AE, Miklós I, Ragan MA. Dynamics of genome rearrangement in bacterial populations. PLoS Genet [Internet]. 2008 Jul 18;4(7):e1000128. Available from: http://dx.doi.org/10.1371/journal.pgen.1000128

26. Achaz G, Coissac E, Netter P, Rocha EPC. Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics [Internet]. 2003 Aug;164(4):1279–89. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12930739

27. Darling AE, Miklós I, Ragan MA. Dynamics of genome rearrangement in bacterial populations. PLoS Genet [Internet]. 2008 Jul 18;4(7):e1000128. Available from: http://dx.doi.org/10.1371/journal.pgen.1000128

28. Achaz G, Coissac E, Netter P, Rocha EPC. Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics [Internet]. 2003 Aug;164(4):1279–89. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12930739

29. Darling AE, Miklós I, Ragan MA. Dynamics of genome rearrangement in bacterial populations. PLoS Genet [Internet]. 2008 Jul 18;4(7):e1000128. Available from: http://dx.doi.org/10.1371/journal.pgen.1000128

30. Couturier E, Rocha EPC. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol [Internet]. 2006 Mar;59(5):1506–18. Available from: http://dx.doi.org/10.1111/j.1365-2958.2006.05046.x

31. Bochkareva OO, Moroz EV, Davydov II, Gelfand MS. Genome rearrangements and selection in multi-chromosome bacteria Burkholderia spp. BMC Genomics [Internet]. 2018 Dec 27;19(1):965. Available from: http://dx.doi.org/10.1186/s12864-018-5245-1

32. Losada L, Ronning CM, DeShazer D, Woods D, Fedorova N, Kim HS, et al. Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. Genome Biol Evol [Internet]. 2010 Jan 22;2:102–16. Available from: http://dx.doi.org/10.1093/gbe/evq003

33. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of Shigella evolution, adaptation and geographical spread. Nat Rev Microbiol [Internet]. 2016 Apr;14(4):235–50. Available from: http://dx.doi.org/10.1038/nrmicro.2016.10

34. Beld MJC, Reubsaet FAG. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli [Internet]. Vol. 31, European Journal of Clinical Microbiology & Infectious Diseases. 2012. p. 899–904. Available from: http://dx.doi.org/10.1007/s10096-011-1395-7

35. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res [Internet]. 2005 Nov 7;33(19):6445–58. Available from: http://dx.doi.org/10.1093/nar/gki954

36. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res [Internet]. 2018 Jan 4;46(D1):D8–13. Available from: http://dx.doi.org/10.1093/nar/gkx1095

37. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics [Internet]. 2011 Apr 28;12:124. Available from: http://dx.doi.org/10.1186/1471-2105-12-124

38. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability [Internet]. Vol. 30, Molecular Biology and Evolution. 2013. p. 772–80. Available from: http://dx.doi.org/10.1093/molbev/mst010

39. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies [Internet]. Vol. 30, Bioinformatics. 2014. p. 1312–3. Available from: http://dx.doi.org/10.1093/bioinformatics/btu033

40. Clarke TH, Brinkac LM, Sutton G, Fouts DE. GGRaSP: a R-package for selecting representative genomes using Gaussian mixture models. Bioinformatics [Internet]. 2018 Sep 1;34(17):3032–4. Available from: http://dx.doi.org/10.1093/bioinformatics/bty300

41. Dusek N, Hewitt AJ, Schmidt KN, Bergholz PW. Landscape-Scale Factors Affecting the Prevalence of Escherichia coli in Surface Soil Include Land Cover Type, Edge Interactions, and Soil pH. Appl Environ Microbiol [Internet]. 2018 May 15;84(10). Available from: http://dx.doi.org/10.1128/AEM.02714-17

42. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees [Internet]. Vol. 44, Nucleic Acids Research. 2016. p. W242–5. Available from: http://dx.doi.org/10.1093/nar/gkw290

43. Varani AM, Siguier P, Gourbeyre E, Charneau V, Chandler M. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. Genome Biol [Internet]. 2011 Mar 28;12(3):R30. Available from: http://dx.doi.org/10.1186/gb-2011-12-3-r30

44. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees [Internet]. Vol. 44, Nucleic Acids Research. 2016. p. W242–5. Available from: http://dx.doi.org/10.1093/nar/gkw290

45. Minkin I, Patel A, Kolmogorov M, Vyahhi N, Pham S. Sibelia: A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes [Internet]. Lecture Notes in Computer Science. 2013. p. 215–29. Available from: http://dx.doi.org/10.1007/978-3-642-40453-5_17

46. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees [Internet]. Vol. 44, Nucleic Acids Research. 2016. p. W242–5. Available from: http://dx.doi.org/10.1093/nar/gkw290

47. Minkin I, Patel A, Kolmogorov M, Vyahhi N, Pham S. Sibelia: A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes [Internet]. Lecture Notes in Computer Science. 2013. p. 215–29. Available from: http://dx.doi.org/10.1007/978-3-642-40453-5_17

48. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Sep;19(9):1639–45. Available from: http://dx.doi.org/10.1101/gr.092759.109

49. Avdeyev P, Jiang S, Aganezov S, Hu F, Alekseyev MA. Reconstruction of Ancestral Genomes in Presence of Gene Gain and Loss. J Comput Biol [Internet]. 2016 Mar;23(3):150–64. Available from: http://dx.doi.org/10.1089/cmb.2015.0160

50. Alekseyev MA, Pevzner PA. Breakpoint graphs and ancestral genome reconstructions [Internet]. Vol. 19, Genome Research. 2009. p. 943–57. Available from: http://dx.doi.org/10.1101/gr.082784.108

51. Bafna V, Pevzner PA. Genome rearrangements and sorting by reversals. In: Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science [Internet]. IEEE; p. 148–57. Available from: http://ieeexplore.ieee.org/document/366872/

52. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res [Internet]. 2004 Jul;14(7):1394–403. Available from: http://dx.doi.org/10.1101/gr.2289704

53. Dixit PD, Pang TY, Studier FW, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proc Natl Acad Sci U S A [Internet]. 2015 Jul 21;112(29):9070–5. Available from: http://dx.doi.org/10.1073/pnas.1510839112

54. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods [Internet]. 2020 Mar;17(3):261–72. Available from: http://dx.doi.org/10.1038/s41592-019-0686-2

55. Chaudhuri RR, Henderson IR. The evolution of the Escherichia coli phylogeny. Infect Genet Evol [Internet]. 2012 Mar;12(2):214–26. Available from: http://dx.doi.org/10.1016/j.meegid.2012.01.005

56. Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA. Investigating the Relatedness of Enteroinvasive Escherichia coli to Other E. coli and Shigella Isolates by Using Comparative Genomics. Infect Immun [Internet]. 2016 Aug;84(8):2362–71. Available from: http://dx.doi.org/10.1128/IAI.00350-16

57. Bordenstein SR, Reznikoff WS. Mobile DNA in obligate intracellular bacteria [Internet]. Vol. 3, Nature Reviews Microbiology. 2005. p. 688–99. Available from: http://dx.doi.org/10.1038/nrmicro1233

58. Ochman H, Moran NA. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science [Internet]. 2001 May 11;292(5519):1096–9. Available from: http://dx.doi.org/10.1126/science.1058543

59. Umenhoffer K, Draskovits G, Nyerges Á, Karcagi I, Bogos B, Tímár E, et al. Genome-Wide Abolishment of Mobile Genetic Elements Using Genome Shuffling and CRISPR/Cas-Assisted MAGE Allows the Efficient Stabilization of a Bacterial Chassis. ACS Synth Biol [Internet]. 2017 Aug 18;6(8):1471–83. Available from: http://dx.doi.org/10.1021/acssynbio.6b00378

60. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res [Internet]. 2005 Nov 7;33(19):6445–58. Available from: http://dx.doi.org/10.1093/nar/gki954

61. Feng Y, Chen Z, Liu S-L. Gene decay in Shigella as an incipient stage of host-adaptation. PLoS One [Internet]. 2011 Nov 16;6(11):e27754. Available from: http://dx.doi.org/10.1371/journal.pone.0027754

62. Ashida H, Toyotome T, Nagai T, Sasakawa C. Shigella chromosomal IpaH proteins are secreted via the type III secretion system and act as effectors. Mol Microbiol [Internet]. 2007 Feb;63(3):680–93. Available from: http://dx.doi.org/10.1111/j.1365-2958.2006.05547.x

63. van Elsas JD, Semenov AV, Costa R, Trevors JT. Survival of Escherichia coli in the environment: fundamental and public health aspects. ISME J [Internet]. 2010 Jun 24 [cited 2020 Jun 8];5(2):173–83. Available from: https://www.nature.com/articles/ismej201080

64. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet [Internet]. 2009 Jan;5(1):e1000344. Available from: http://dx.doi.org/10.1371/journal.pgen.1000344

65. Dobrindt U, Chowdary MG, Krumbholz G, Hacker J. Genome dynamics and its impact on evolution of Escherichia coli. Med Microbiol Immunol [Internet]. 2010 Aug;199(3):145–54. Available from: http://dx.doi.org/10.1007/s00430-010-0161-2

66. Pupo GM, Lan R, Reeves PR. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics [Internet]. Vol. 97, Proceedings of the National Academy of Sciences. 2000. p. 10567–72. Available from: http://dx.doi.org/10.1073/pnas.180094797

67. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, et al. The Intriguing Evolutionary Journey of Enteroinvasive (EIEC) toward Pathogenicity. Front

Microbiol [Internet]. 2017 Dec 5;8:2390. Available from: http://dx.doi.org/10.3389/fmicb.2017.02390

68. Mattock E, Blocker AJ. How Do the Virulence Factors of Work Together to Cause Disease? Front Cell Infect Microbiol [Internet]. 2017 Mar 24;7:64. Available from: http://dx.doi.org/10.3389/fcimb.2017.00064

69. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev [Internet]. 2014 Sep;38(5):865–91. Available from: http://dx.doi.org/10.1111/1574-6976.12067

70. Bordenstein SR, Reznikoff WS. Mobile DNA in obligate intracellular bacteria [Internet]. Vol. 3, Nature Reviews Microbiology. 2005. p. 688–99. Available from: http://dx.doi.org/10.1038/nrmicro1233

71. Raeside C, Gaffé J, Deatherage DE, Tenaillon O, Briska AM, Ptashkin RN, et al. Large chromosomal rearrangements during a long-term evolution experiment with Escherichia coli. MBio [Internet]. 2014 Sep 9;5(5):e01377–14. Available from: http://dx.doi.org/10.1128/mBio.01377-14

72. Allocati N, Masulli M, Alexeyev MF, Di Ilio C. Escherichia coli in Europe: an overview. Int J Environ Res Public Health [Internet]. 2013 Nov 25;10(12):6235–54. Available from: http://dx.doi.org/10.3390/ijerph10126235

73. Burger R. EHEC O104:H4 IN GERMANY 2011: LARGE OUTBREAK OF BLOODY DIARRHEA AND HAEMOLYTIC URAEMIC SYNDROME BY SHIGA TOXIN–PRODUCING E. COLI VIA CONTAMINATED FOOD. In: Improving Food Safety Through a One Health Approach: Workshop Summary [Internet]. National Academies Press (US); 2012 [cited 2020 Jun 8]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK114499/

74. Hassan R, Tantawy M, Gouda NA, Elzayat MG, Gabra S, Nabih A, et al. Genotypic characterization of multiple drug resistant Escherichia coli isolates from a pediatric cancer hospital in Egypt. Sci Rep [Internet]. 2020 Mar 5;10(1):4165. Available from: http://dx.doi.org/10.1038/s41598-020-61159-z

11. Figures and tables

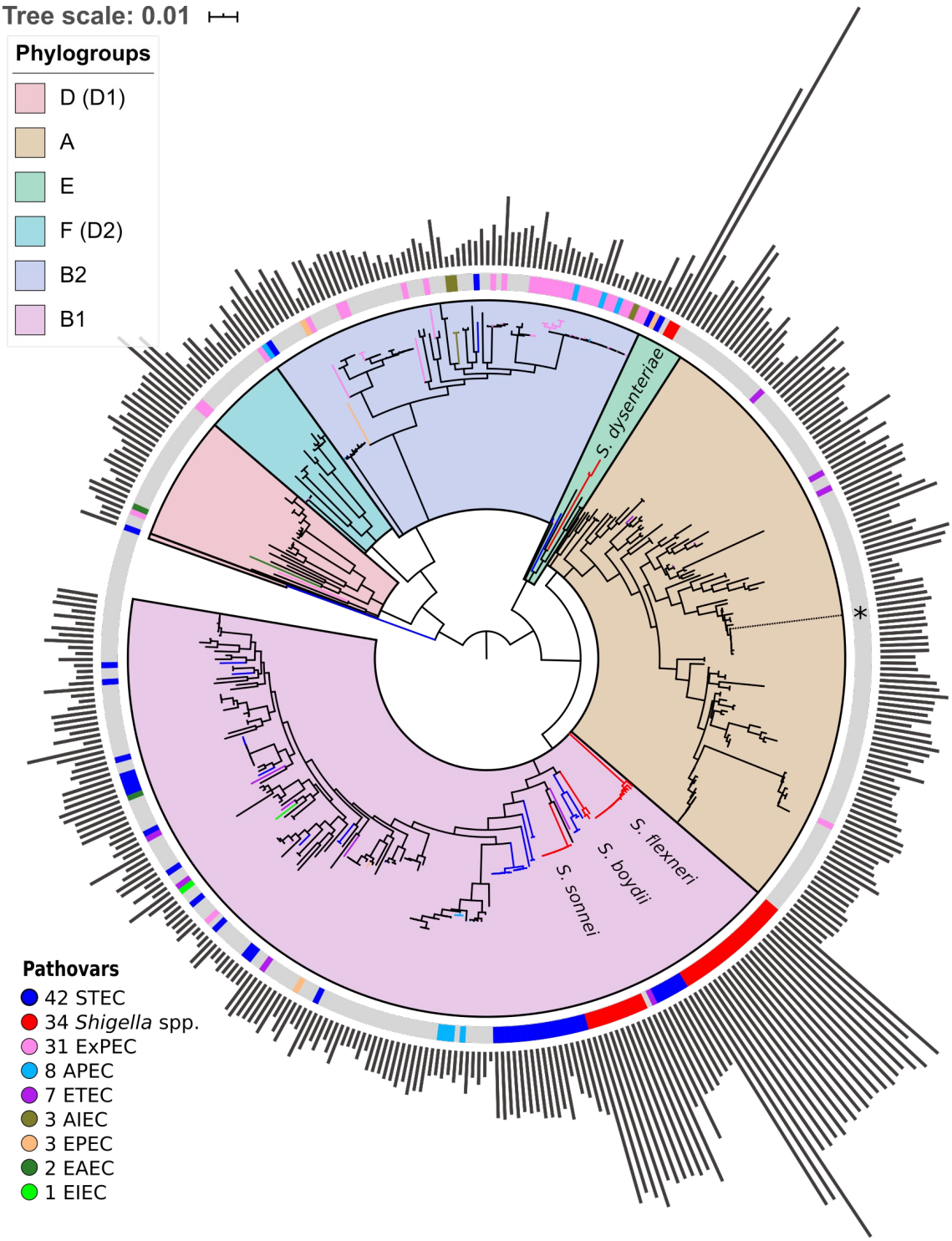


Figure 1. Phylogenetic tree of *E. coli* and *Shigella* spp. The tree is based on the nucleotide alignment of universal single-copy orthologs. Clusters and corresponding phylogroups are shown in different colours; the number of genomes of each pathovar is indicated to the right of the pathovar's name. Pathogenic *E. coli* strains are marked with coloured bands on the outer circle. The location of the unclassified *Shigella* strain is shown with an asterisk. The bars indicate the number of ISs found by ISsaga in bacterial chromosomes. The tree is unrooted. The tree is also available online (iTOL): https://itol.embl.de/tree/8589127100415001556065531.

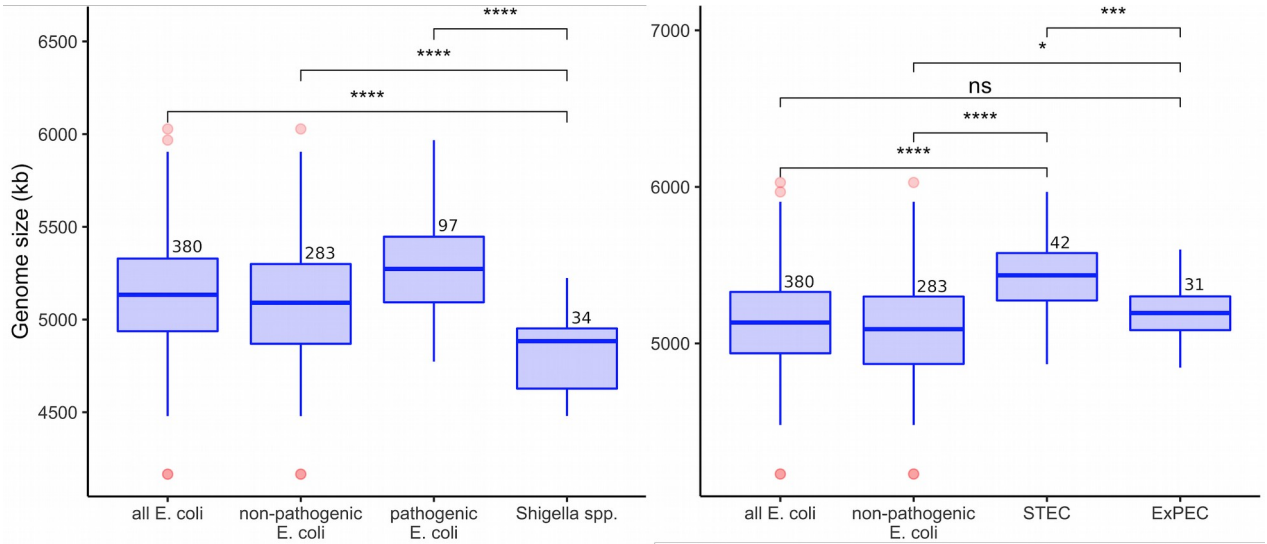


Figure 2. Comparison of genome sizes in **(a)** *Shigella*, non-pathogenic *E. coli*, other pathogenic *E. coli* and all *E. coli* (excluding *Shigella* spp); **(b)** two most abundant pathovars — ExPEC and STEC, non-pathogenic *E. coli* and all *E. coli* (excluding *Shigella* spp). The numbers above boxplots indicate the number of genomes in each group. The Wilcoxon-Mann-Whitney test with the Bonferroni correction: (*) $p \leq 0.05$, (**) $p \leq 0.01$; (***) $p \leq 0.001$; (****) $p \leq 0.0001$.

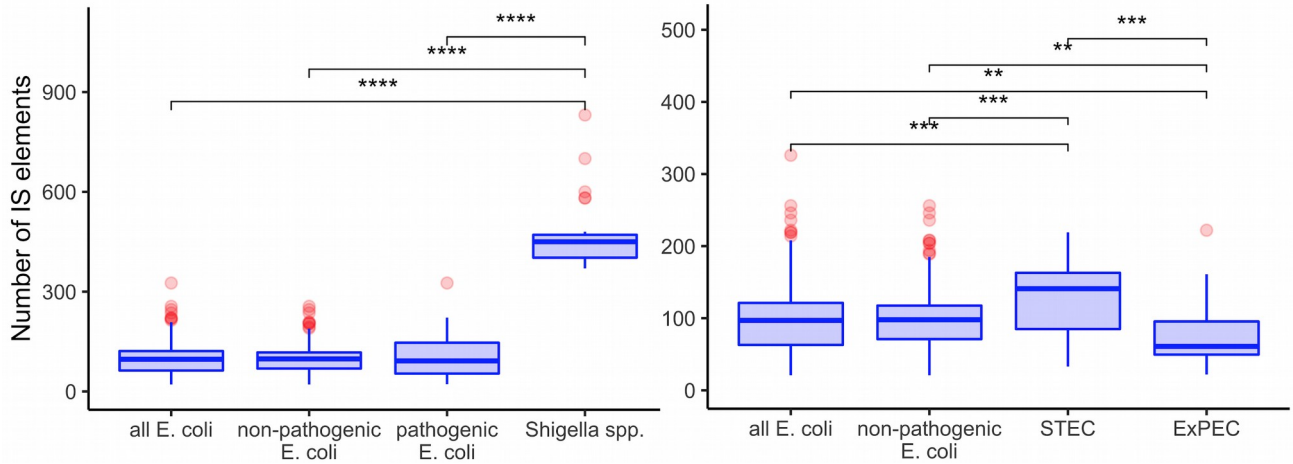


Figure 3. Comparison of IS numbers in **(a)** *Shigella* spp., non-pathogenic *E. coli*, other pathogenic *E. coli*, and all *E. coli* (excluding *Shigella* spp); **(b)** two most abundant pathovars — ExPEC and STEC, non-pathogenic *E. coli*, and all *E. coli* (excluding *Shigella* spp). The Wilcoxon-Mann-Whitney test with the Bonferroni correction: (*) $p \leq 0.05$, (**) $p \leq 0.01$; (***) $p \leq 0.001$; (****) $p \leq 0.0001$.

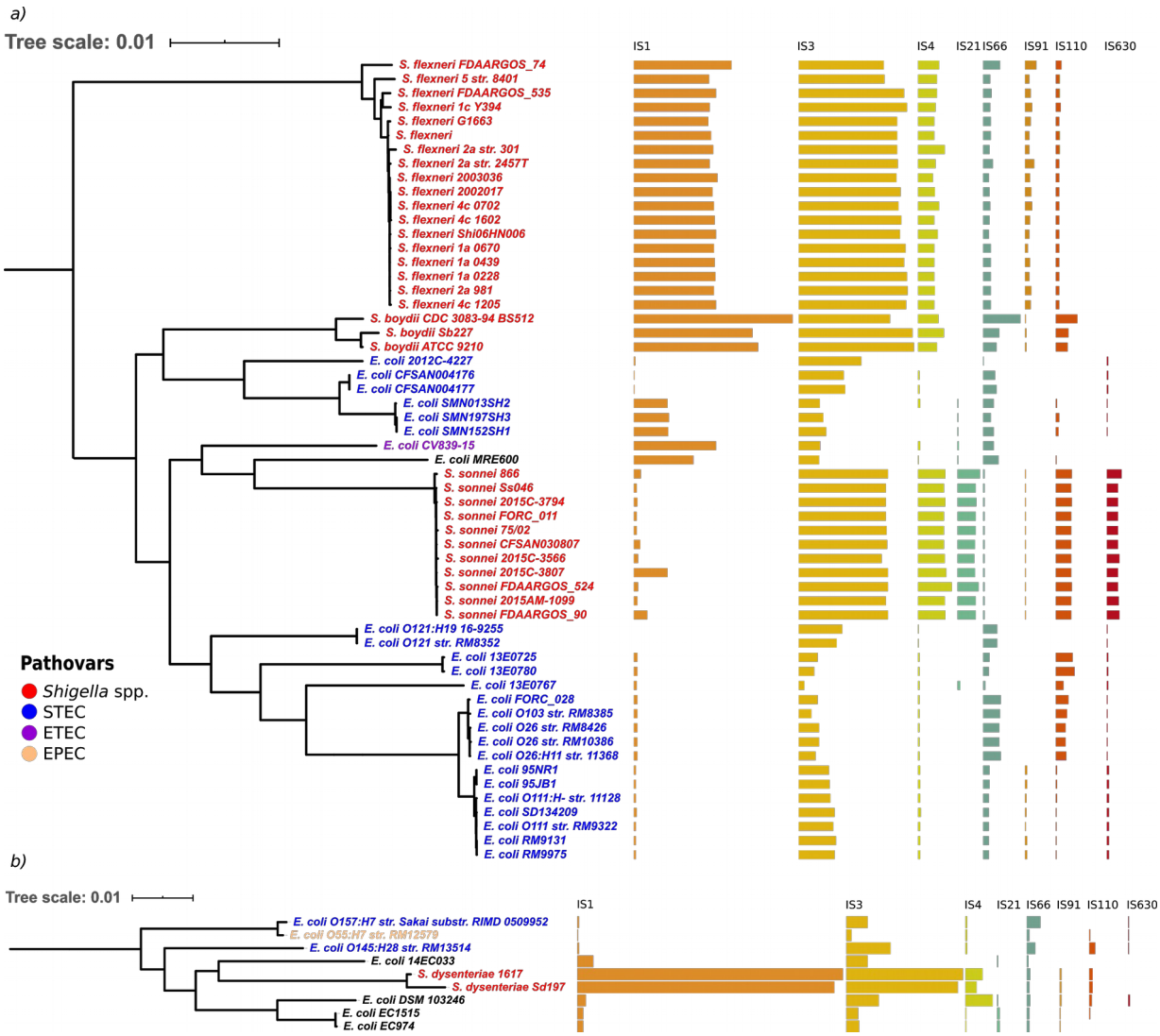


Figure 4. Distribution of IS families in **(a)** a part of phylogroup B1 and **(b)** phylogroup E. The bars indicate the number of ISs in each family. The trees are also available online (iTOL):
a) https://itol.embl.de/tree/8589127100479001583084903#,
b) https://itol.embl.de/tree/8589127100208261560504457#.

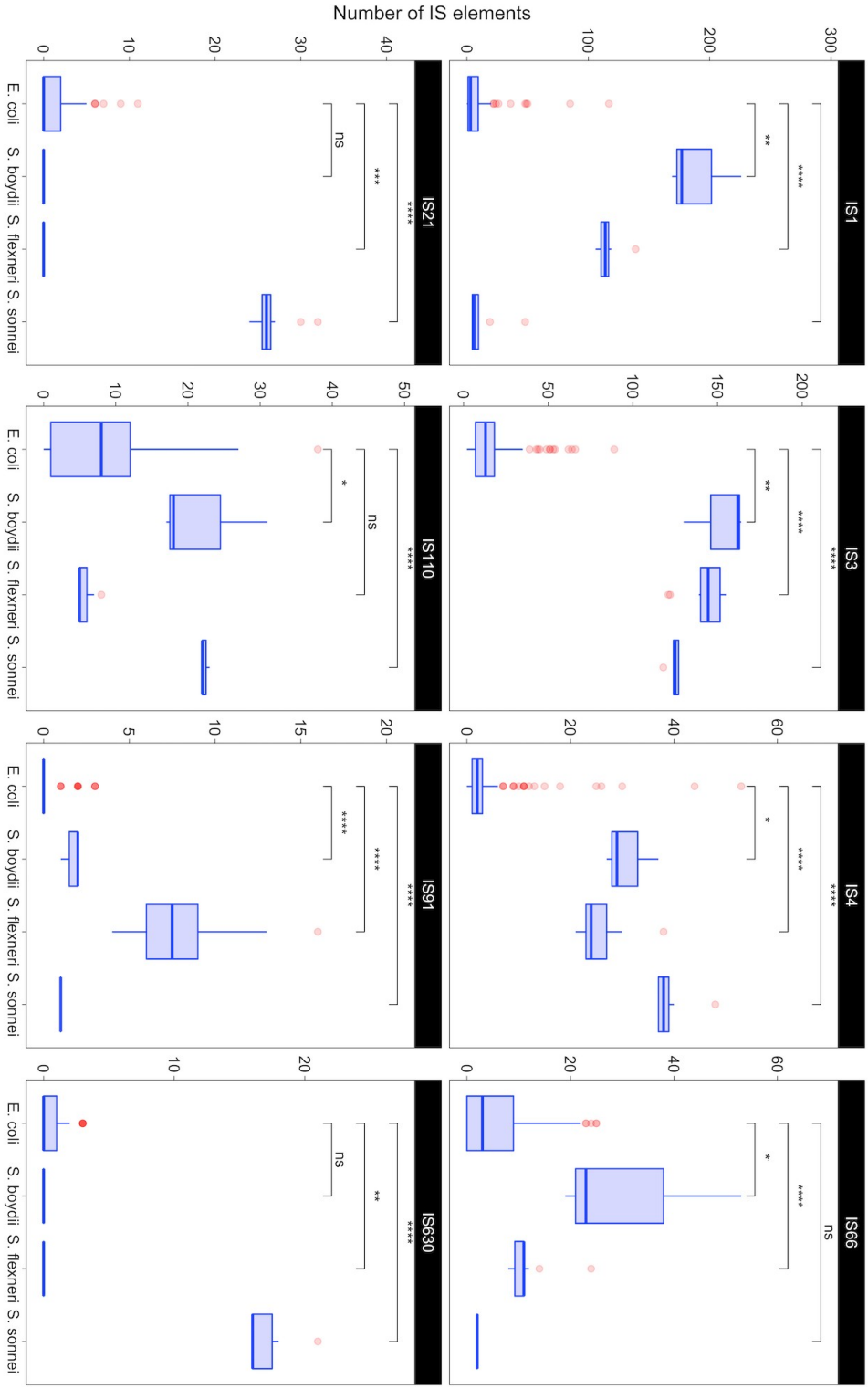


Figure 5. The comparison of selected IS families in *Shigella* and *E. coli* strains from phylogroup B1. The Wilcoxon-Mann-Whitney test with the Bonferroni correction: (*) $p \leq 0.05$, (**) $p \leq 0.01$; (***) $p \leq 0.001$; (****) $p \leq 0.0001$.

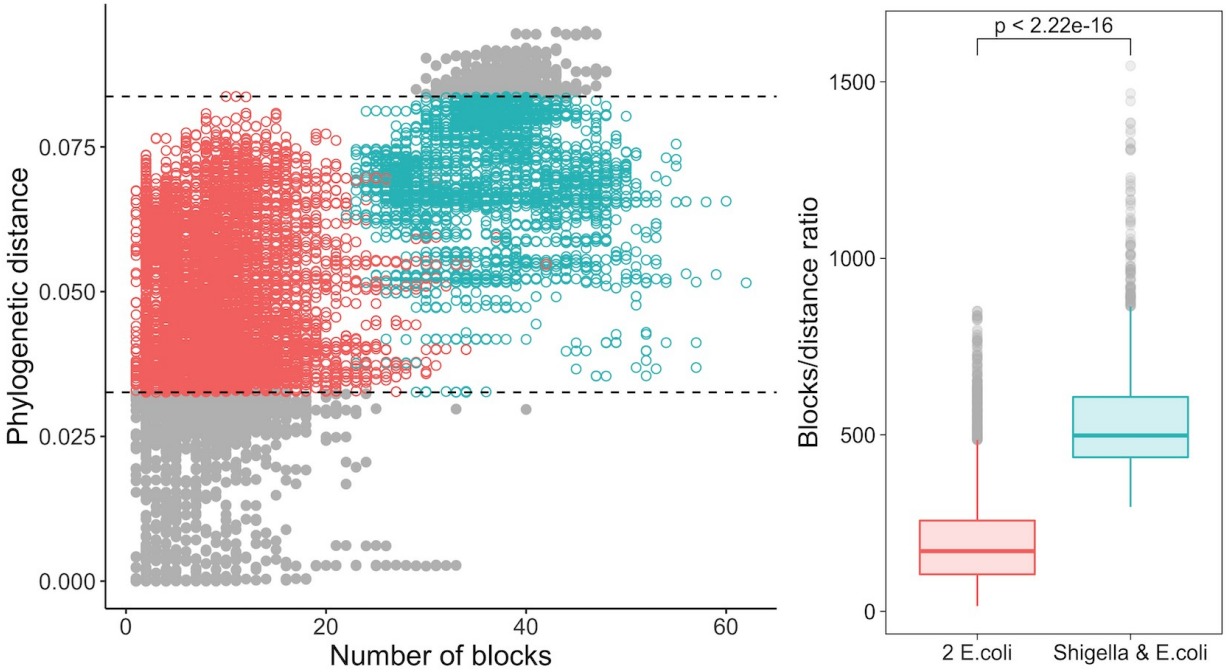


Figure 6. Comparison of the ratio of phylogenetic distance to the number of synteny blocks for two clusters from phylogroup B1. Only genome pairs in the intersection were compared. The Wilcoxon-Mann-Whitney test *p*-value is given.

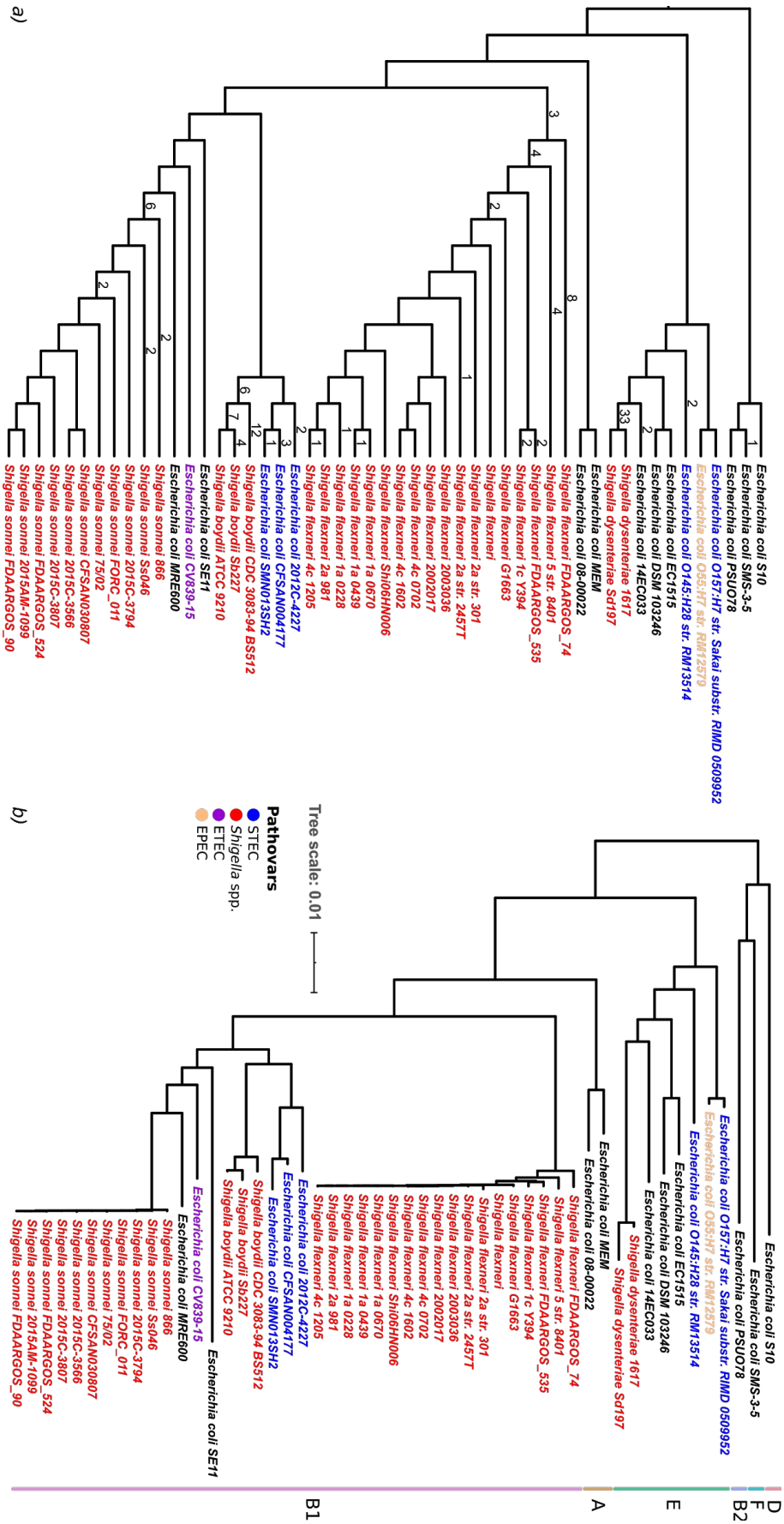
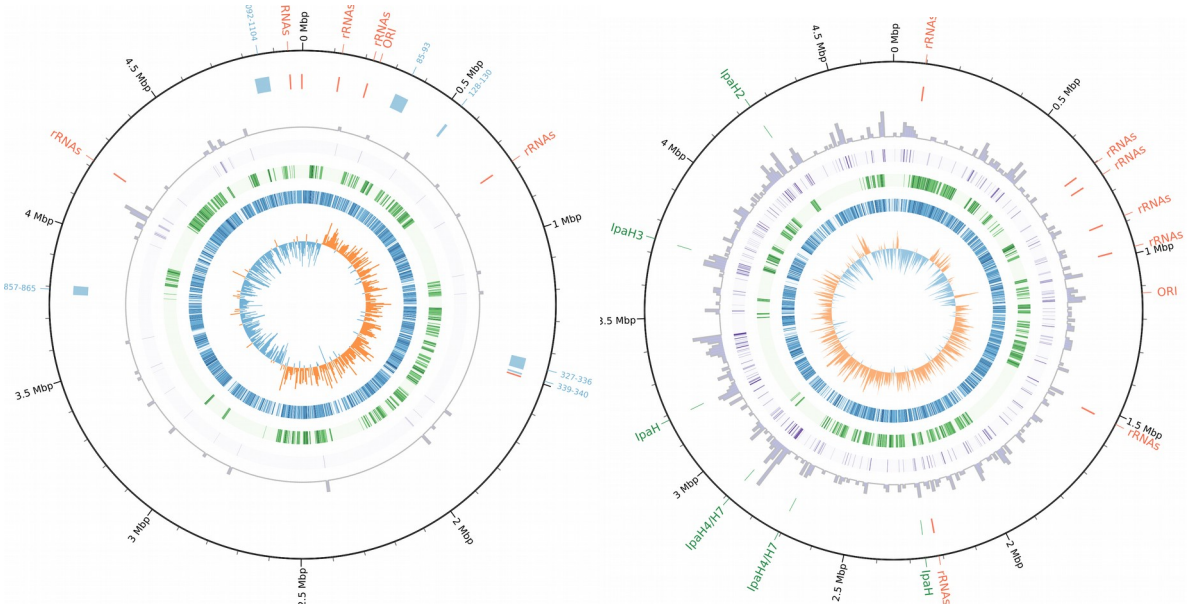


Figure 7. Inversion events reconstructed by MGRA software. (a) A cladogram with numbers of inversions shown for each branch. **(b)** The corresponding phylogenetic tree. Phylogroups are marked with coloured strips, pathogenic strains are shown in different colours. Both trees are unrooted. The trees are also available online (iTOL): https://itol.embl.de/tree/9318063252480721583313649.



Escherichia coli C4 *Shigella flexneri* 1a 228

Figure 8. Maps of *E.coli* (left) and *Shigella* (right) chromosomes. The inner circle — GC-skew, the second (blue) circle — synteny blocks, the third (green) circle — universal synteny blocks, the fourth circle — ISs, the fifth circle — density of ISs. The light blue boxes — regions of recombination hotspots in *Shigella*.

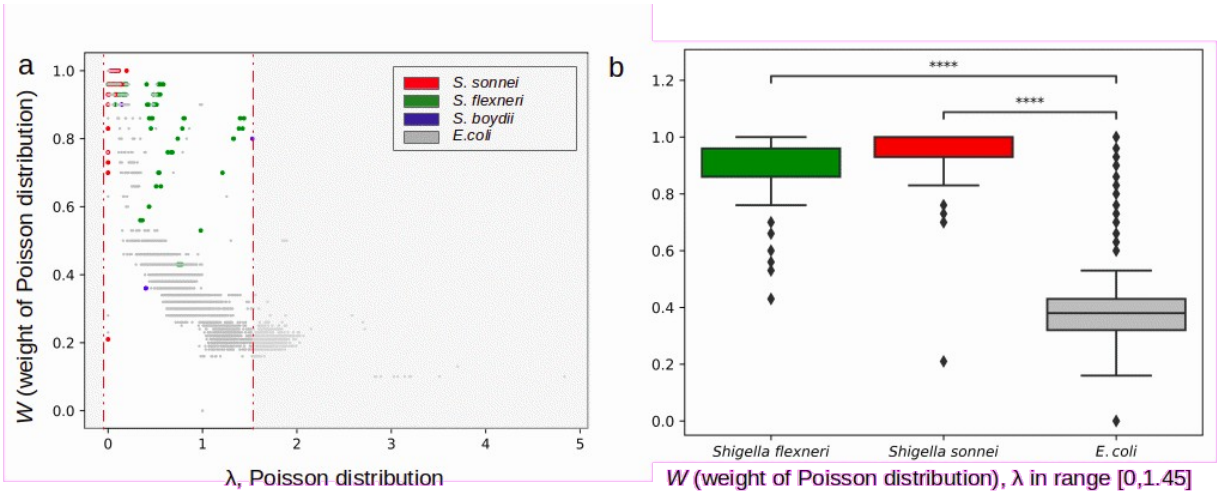


Figure 9. Rates of homologous recombination. (a) Fractions of horizontally transferred fragments in strains from the B1 phylogroup calculated as the weight of the Poisson distribution (see Methods). **(b)** The weight of Poisson distribution in *Shigella* and *E.coli* in the same interval of evolutionary distances between strains. The Wilcoxon-Mann-Whitney test: (*) $p \leq 0.05$, (**) $p \leq 0.01$; (***) $p \leq 0.001$; (****) $p \leq 0.0001$.