What can we learn from over 100,000 *Escherichia coli* genomes?

Kaleb Abram[1]*, Zulema Udaondo[1]*, Carissa Bleker[2,3], Visanu Wanchai[1], Trudy M. Wassenaar[4], Dave W. Ussery[1#]

[1]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

[2]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN, USA

[3]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA

[4]Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany

#Corresponding author: DWUssery@uams.edu

*These authors contributed equally

Running title: Insights from 100,000 *E. coli* genomes

Keywords: *Escherichia coli*, phylotype, pangenome, core genes, comparative genomics, population structure, taxonomy, Mash

ABSTRACT

The explosion of microbial genome sequences in public databases allows for large-scale population studies of model organisms, such as *Escherichia coli*. We have examined more than one hundred-thousand *E. coli* and *Shigella* genomes. After removing outliers, genomes were classified into two broad clusters based on a semi-automated Mash analysis, which distinguished 14 distinct phylotypes, graphically illustrated by Cytoscape. From a set of more than ten-thousand good quality *E. coli* and *Shigella* genomes from GenBank, we find roughly 2,700 gene families in the *E. coli* species core, and more than 135,000 gene families in the *E. coli* pan-genome. Based on a set of 2,613 single-copy core proteins taken from one representative genome per phylotype, we constructed a robust phylogenetic tree. This is the largest *E. coli genome* dataset analyzed to date, and provides valuable insight into the population structure of the species.

E. coli is a common inhabitant of the gastrointestinal tract of warm-blooded animals including humans, and also can be found in soil and freshwater (Jang *et al.*, 2017). The species is comprised of both commensal and pathogenic strains, and can cause disease in a wide variety of animal hosts. In humans, pathogenic *E. coli* strains are a leading cause of diarrhea-associated hospitalizations (Fischer Walker *et al.*, 2010). Some of the attributes that make *E. coli* an intensely studied microorganism include: rapid growth rate in the presence of oxygen, easy adaptation to environmental changes, and the relative ease with which it can be genetically manipulated (Dunne *et al.*, 2017). The extraordinary plasticity of *E. coli* genomes is demonstrated by differences in their size, which ranges from slightly less than 3 million basepairs (Mbp) to more than 7 Mb. Genomic diversity of the species, to which *Shigella* species should be included (Pettengill *et al.*, 2016; Chattaway *et al.*, 2017), is furthermore reflected by the existence of distinct phylogenetic groups (phylotypes) that have been identified using a variety of different methods (Clermont *et al.*, 2000; Gordon *et al.*, 2008; Tenaillon *et al.*, 2010).

Historically, four phylotypes have been recognized: A, B1, B2, and D (Clermont *et al.*, 2000; Tenaillon *et al.*, 2010) to which three more were added later: phylotypes C (closest relative to B1) (Clermont *et al.*, 2013), F (as a sister group of phylotype B2), and E to which many D members were reassigned (Clermont *et al.*, 2013). Some studies have subdivided these into more

groups, with D1 to D3, subdivisions of F, and separate phylotypes for *Shigella* species (Meier-Kolthoff *et al.*, 2014). These phylotypes are thought to be monophyletic (Tenaillon *et al.*, 2010; Meier-Kolthoff *et al.*, 2014) and partly coincide with different ecological niches and lifestyles, whose members differ in metabolic characteristics such as their ability to exploit different carbon sources, the presence of virulence genes and even antibiotic resistance profiles (Walk *et al.*, 2009; Carlos *et al.*, 2010; Tenaillon *et al.*, 2010; Vangchhia *et al.*, 2016).

Previously, population structure analysis has been performed using datasets of various size and composition that did not fully capture the diversity of the species. With the availability of a large number of genome sequences and high-performance computers, population genomics within the whole species can now be feasibly studied, although efficient programming is required for analysis of large amounts of data.

Here, we describe a comprehensive comparison of over 100,000 publicly available genome sequences, consisting of 12,602 assembled genomic sequences from GenBank, and over 102,000 unassembled raw genome sequences from the Sequence Read Archive (SRA). This study combined whole genome sequences (WGS) and SRA unassembled genomes using high-performance computing resources to cover the largest and most complete analysis to date of the population structure of *E. coli*. We have quantified the similarities and differences between phylotypes to identify genomic phylogroups that encompass the so-far recognized phylotypes and to characterize the genetic heterogeneity of these different phylogenetic lineages. We have identified 14 'medoid' genomes, one for each phylogroup, that can be used as a representation of the population groups within the species.

RESULTS

To conduct the analysis, 12,602 genome sequences labeled with *Escherichia* or *Shigella* were downloaded from Genbank on 26 June, 2018 (including plasmid sequences when applicable). This dataset (Supplementary Table 1) was cleaned using a variety of steps as described in the Methods section, to obtain an informative and diverse set of 10,667 *E. coli* and *Shigella* genomes that captures the actual diversity of the species as sequenced to date, with minimal risk of producing artifactual findings from low quality genomes. Underrepresented

genomes upon which previous research knowledge is biased against are also included. Our final genome dataset is heterogeneous and contains many different genomes within each of the phylotypes, thereby better reflecting the interests of the complete scientific community in the area of genomic research on this species. In addition to the GenBank genomes, a total of 102,091 read sets were downloaded from the SRA database that were labeled as either *E. coli* or *Shigella* (see Methods).

Mash analysis of E. coli genomic sequences reveals 14 phylotypes. After cleaning the dataset, we used Mash distances (Ondov *et al.*, 2016) to produce a matrix of the 10,667 genomes based on hierarchical clustering with Pearson's correlation coefficient. A clustered heatmap was used for visualization to illustrate the population structure of these genomes (**Fig. 1**). This methodology differentiated 14 different phylogroups, named: G, B2-1, B2-2, F, D1, D2, D3, E2(O157), E1, A, C, B1, Shig1 and Shig2 [ordered as in Fig. 1], which are all distinct based on their genetic sequence, according to the Mash distances. The phylogroups Shig1 and Shig2 exclusively contained *Shigella* species, but some *Shigella* sp. genomes were found in phylogroups A, B1, B2-2, D2, D3, E and F (Supplementary Fig. 1).

The heatmap shown in Fig. 1 reveals that some phylogroups share more genetic similarity to each other than to other phylogroups, such as B2-1 to B2-2, Shig1 to Shig2, and C to B1. On the other hand, B1 is quite distinct from the B2 groups. We have utilized Microreact (https://microreact.org/project/10667ecoli/b4431cf8) (Argimón *et al.*, 2016) to visualize the resultant Mash distance-based clustering. The assembly accession ID was used as the identifier for each genome. To this identifier we mapped the organism name, strain name, sequence size (Mb), Bioproject ID, Biosample ID, phylogroup, average genome quality score, and genome sequence quality score. To optimize the search function of Microreact, we downloaded (on 6/20/2019) all entries from PATRIC-labeled (Wattam *et al.*, 2017) *Escherichia coli* or *Shigella sp.* and mapped some of their data to each genome. This allows the exploration of subclusters within the dendrogram for a number of shared characteristics that is outside the scope of the current study and will be a topic for future exploration to increase our understanding of the *E. coli* species.

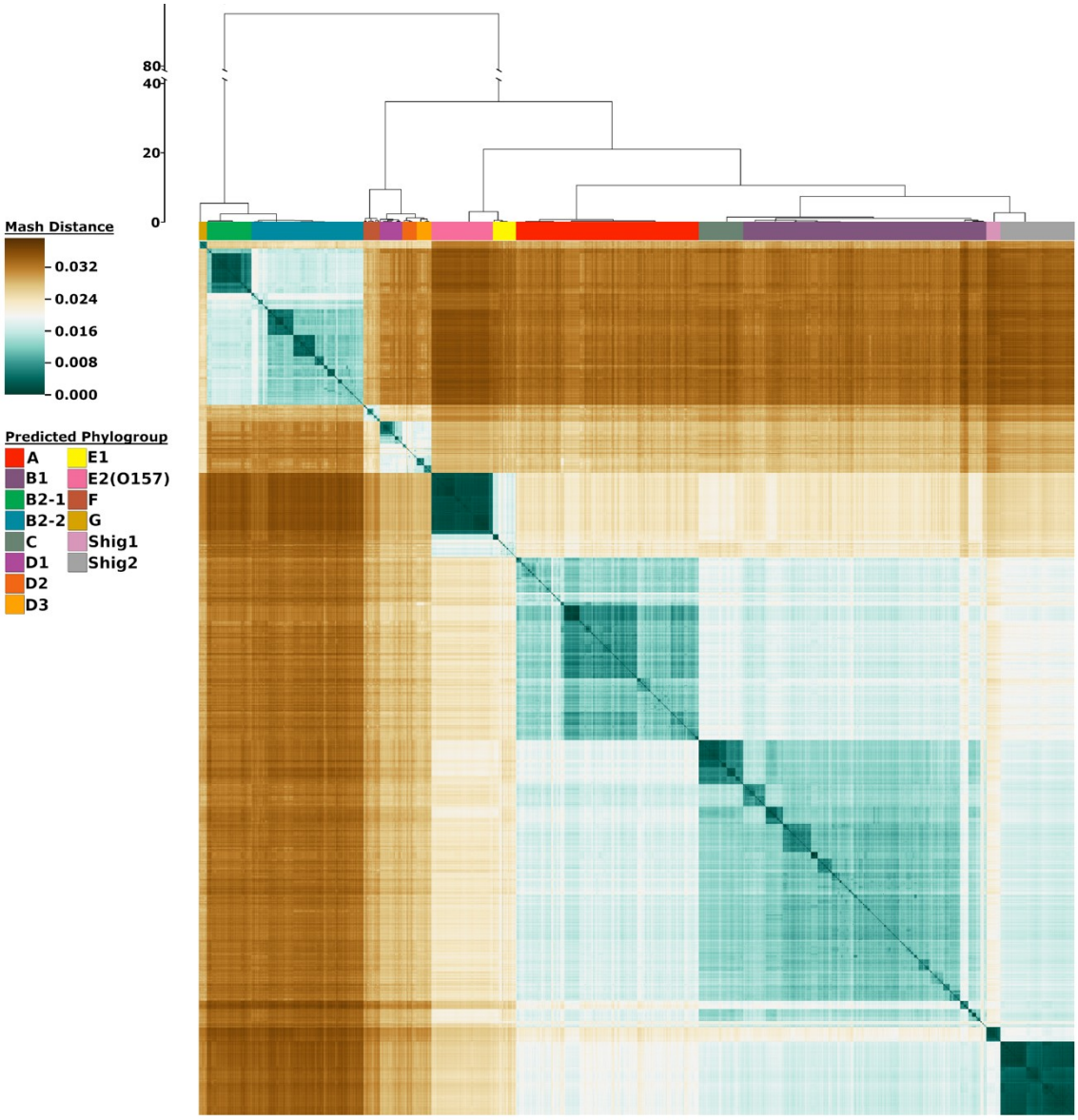


Fig. 1. Heatmap representation of 10,667 genomes using Mash distances. The color bars at the top of the heatmap identify the phylogroups as predicted from the analysis (see key). The scale to the left of the dendrogram corresponds to the resultant cluster height of the entire dataset obtained from hclust function in R (see Methods for more details). The colors in the heatmap are based on the pairwise Mash distance between the genomes. Blue-green colors represent similarity between genomes with the darkest blue-green corresponding to identical genomes reporting a Mash distance of 0. Brown colors represent low genetic similarity per Mash distance, with the darkest brown indicating a maximum distance of ~ 0.039. Genomes of relative median genetic similarity have the lightest color.

Most sequenced E. coli genomes belong to only 4 phylotypes. To increase the utility of our analysis, a minimal set of genomes was defined that represents the diversity of the 10,667 genomes without suppressing any of the predicted phylogroups. As 14 main phylogroups were predicted, we tested if one genome from each of these would be enough to accurately predict the phylotype of any given genome sequence claiming to be *Escherichia coli* or *Shigella*. Each of these 14 genomes represents the medoid (Struyf *et al.*, 1997) or the "genomic center" of each phylogroup based on the 10,667 analyzed genomes. In order to further increase the sequence dataset, we added a total of 102,091 read sets labeled as either *E. coli* or *Shigella sp.* from the SRA database. This dataset was first filtered by quality of the sequences (see Methods) which resulted in a set of 95,525 genomes to which a phylogroup could confidently be assigned. As a way to reduce computational load for classifying SRA reads, we compared these to each medoid of the 14 phylogroups in an asymmetric matrix. A heatmap plotting SRA reads that have a Mash distance equal to or less than 0.04 for at least three medoids is shown (Fig. 2) and a breakdown of the SRA results is summarized in Supplementary Table 2.

Two-thirds (67%) of the analyzed SRA reads were predicted to belong to one of four phylogroups: A (23%), C (15%), B1 (15%), and E2(O157) (14%). The most prominent predicted phylogroup in the SRA dataset was A, covering about 23% of the reads. This large disparity in phylogroup diversity in the SRA dataset is most likely explained by the interest of the scientific and medical communities. Strains belonging to phylogroups B1, C, and E2(O157) are often pathogenic and of interest to medical research, while phylogroup A includes strains frequently used in the laboratory (*e.g.*, strain K-12) or engineered strains (such as strain BL21 and REL606).

Similarly, approximately two-thirds of the 10,667 assembled genomes also belong to four phylogroups: B1 (28%), A (21%), B2-2 (13%) and Shig2 (8%). However, in the assembled genomes, phylogroup C is only about 5%, whilst E2(O157) is about 7%.

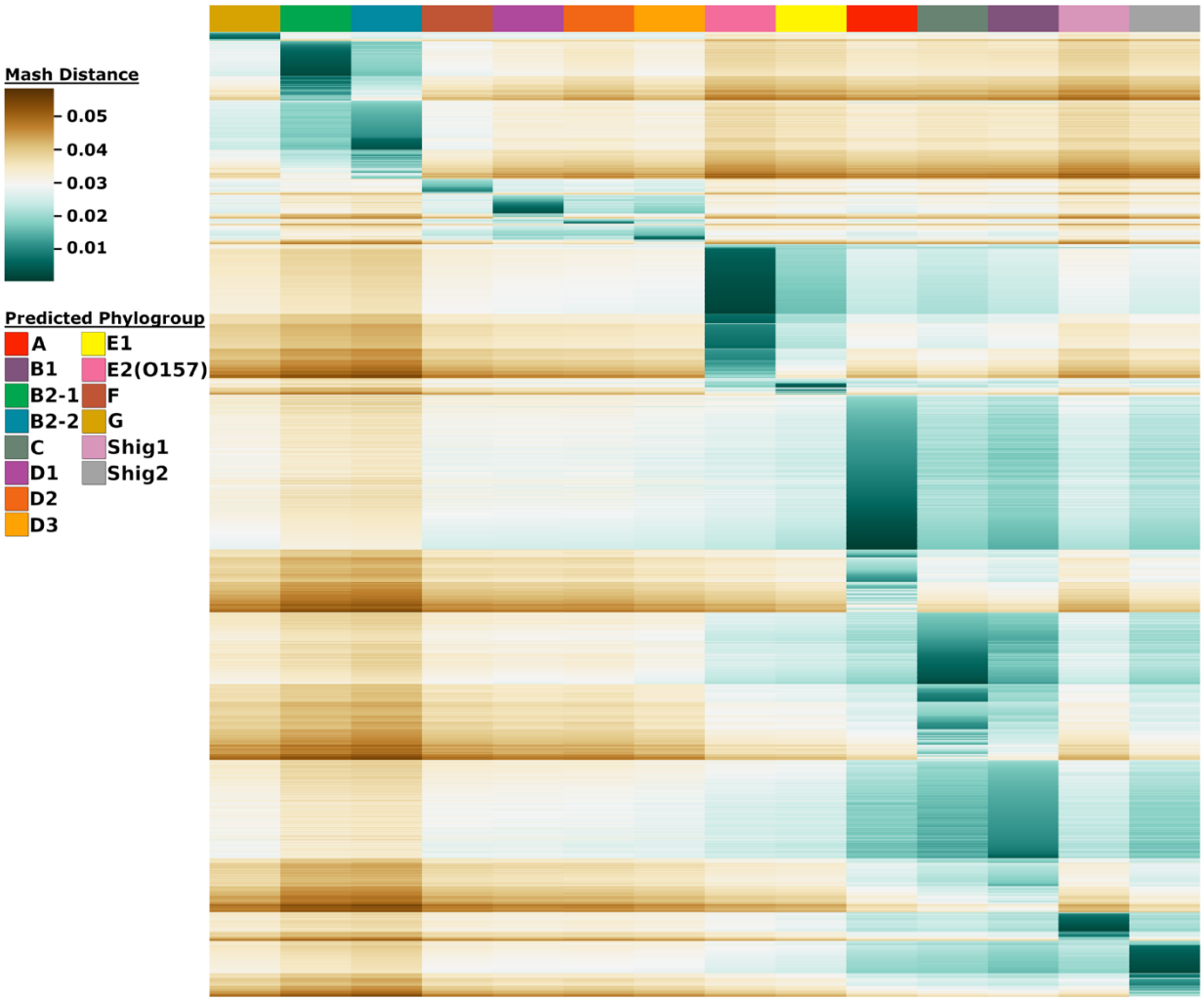


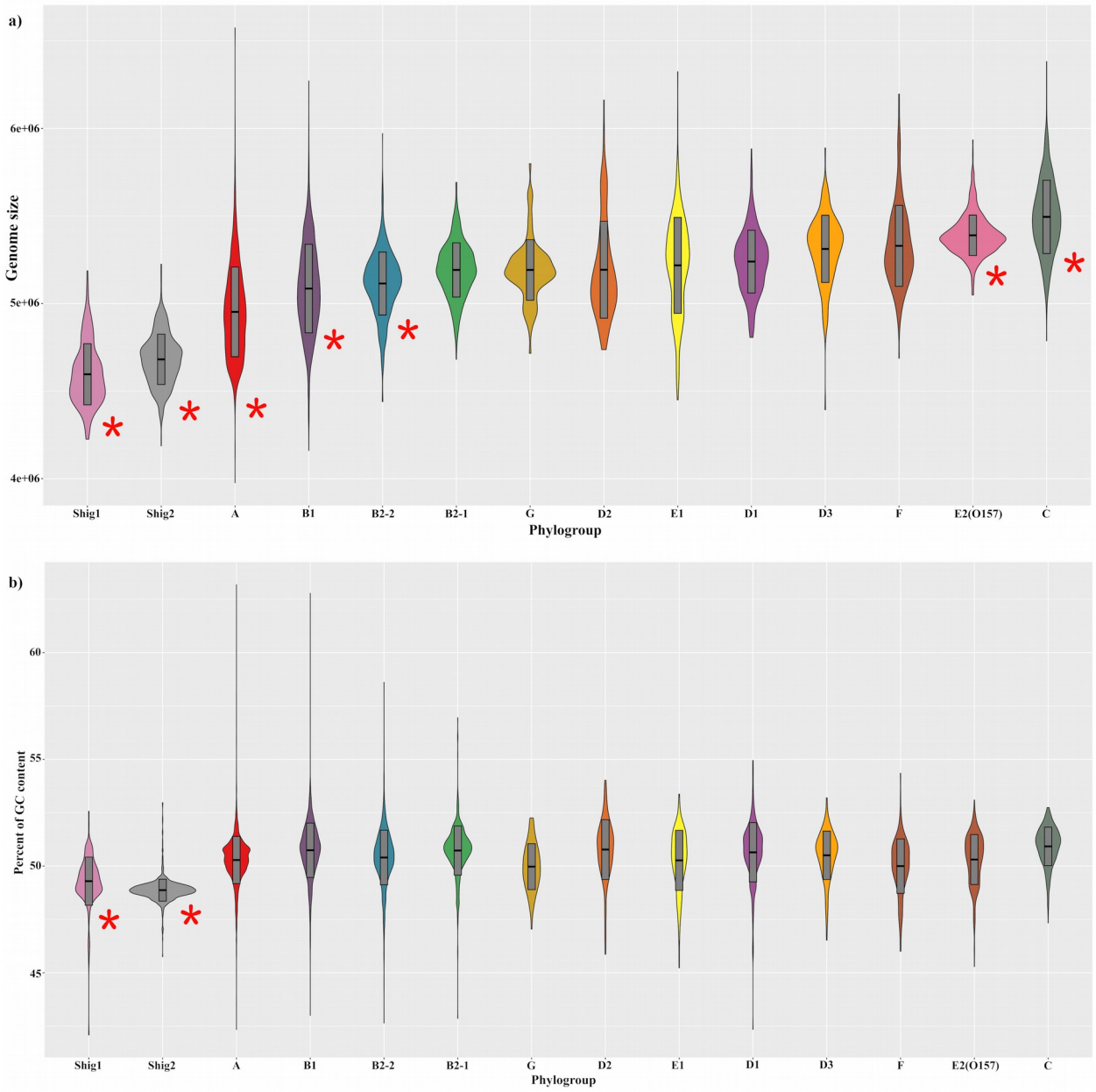
Fig. 2. Heatmap representation of 91,261 sequence reads from the SRA database. The heatmap colors are based on the pairwise Mash distance between the SRA read sets and the 14 medoid genomes of each phylogroup, which are presented in the same order as in Fig. 1. To be included, SRA reads sets had to have 3 or more medoid comparisons producing a Mash distance equal to or less than 0.04. This removed 4,264 SRA read sets from the dataset. The number of SRA reads mapped to each medoids is given below the heatmap. Supplementary Fig. 2 contains additional cut-offs ranging from one to 14 phylogroups.

The currently sequenced E. coli and Shigella species can be represented by 14 medoid genomes. To investigate whether our clustering results were due to the data itself and not due to bias in hierarchical clustering methods, we utilized Cytoscape (Shannon *et al.*, 2003) to represent the raw Mash distance outputs. During this clustering, the medoids were used as anchors to

evaluate how the rest of the genomes distributed around them. Using the graph visualization abilities of Cytoscape, we verified the concept of our medoids being representatives of entire phylogroups. This demonstrated the medoids were suitable to decrease visual complexity without sacrificing accuracy. The resulting Cytoscape graphs visualize the relative genetic diversity of the genomes as calculated by Mash genomic distance. A video is available as Supplementary information and a collection of stills is available upon request. This analysis shows that the two B2 phylogroups are the most genetically distinct from the remainder of the species, in terms of sequence content, as they start to split off first. The next set of genomes to separate represent the D/F/G phylogroups, with G splitting off from the B2 complex. Group F then splits off from the D/F complex. Next, the E complex separates from the bulk of the species, which then still contains A/B1/C/Shig. Of the *Shigella* groups, Shig1 splits off before Shig2 does, showing that Shig1 (containing predominantly *S. flexneri* strains) is more genetically distant from the A/B1/C complex than Shig2 is (the latter is mainly composed of *S. sonnei* strains (Supplementary Fig. 1). At the final Mash value cutoff of 0.0095, the C and B1 phylogroups become the last two groups to separate. This last split is indicative of the relatively large shared genetic content by these two phylogroups. Between the initial Cytoscape frame and the final frame, the number of genomes represented decreased by 43% while the edges (connections between genomes and medoids) decreased by 96%. These results show that the medoids represent the species accurately. As expected, the overall interconnectivity of the different phylogenetic groups drops significantly with the cutoff, but intraconnectivity within the phylotypes does not.

Members of Mash phylogroups possess different genomic characteristics. Since Mash values provide a distance measure of similarity between a pair of genomes, the phylogroups of Fig. 1 are the consequence of differences/similarities in the genetic content of each genome with respect to the rest of the genomes included in the analysis. Differences in genome size and percentage of GC content between these phylogenetic groups are observed (Fig. 3). Statistical tests were performed (ANOVA and Turkey test, see methods) to identify significant differences between the average genome size and GC content per phylogroup. Significant differences in genome size of members belonging to different phylogroups are observed for phylogroups A and B1 (significantly smaller genomes ($P<0.01$)) and C and E2(O157) (significantly larger genomes

(P<0.01)). The two *Shigella* phylogroups also contain genomes with significantly smaller genomes (P<0.01), on average, indicative of reductive genome evolution of these organisms as was noted before (Weinert and Welch, 2017). However, reduced genome size is not associated with pathogenicity *per se*, as the large genomes of E2(O157) illustrate. Larger genome sizes associated with virulence may result from the accumulations of virulence genes in prophages, pathogenicity islands and plasmids (Bhunia, 2018). We also compared overall genomic GC content, which is less variable and only differs significantly for the two *Shigella* phylogroups (P<0.01). These characteristics might reflect the different evolutionary strategies and opposite selection pressures as a consequence of adaptation to diverse niches in which the different phylotypes have evolved (Balbi *et al.*, 2009).



188 **Fig. 3. Violin-plots of the distribution of genome size (A) and genomic GC content (B) by phylogroup.** Bar-
189 plots inside the violins represent values for mean and mean plus one standard deviation per phylogroup.
190 Phylogroups that have values significantly different to all other phylogroups (according to F statistics test) are
191 marked with a red asterisk**.**

192

Level of preservation of homologous genes varies between phylotypes. To evaluate the existence of functional traits associated with each of the phylogroups, we performed pan- and core genomic analyses using the proteomes of the set of 10,667 assembled genomes. In addition, separate core genomes were calculated for the 14 individual phylogroups. For this, all protein-coding genes were newly annotated using standardized criteria.

The overall pangenome of all assembled genomes is comprised of 135,983 clusters of homologous proteins. A core genome of this total dataset representing homologs found in 100% of the strains ($^{TOT}core_{100}$) only contains one gene, which happens to be a hypothetical protein with a functional domain of a peptidase superfamily. This vanishingly small core is likely due to individual genes that are missed in genome sequencing, assembly, or gene calling; with very large numbers of genomes, a 100% core can be quite small, regardless of high sequence quality scores. From testing the cutoffs for conservation in 99% to 90% of the genomes (Supplementary Fig. 4) we concluded that, while the traditional cutoff for core calculation of 95% of genomes would suffice, a cutoff of 97% can minimize erroneous core genes due to over-representation of genomes, in turn providing a more stringent core. Therefore, we defined the core genome as homologous genes shared by at least 97% of the genomes, which produces a $^{TOT}core_{97}$ of 2,663 clusters (1.96% of the total pangenome's clusters). These core genes comprise on average a bit more than 50% of the protein content per strain, as illustrated in Fig. 4. The $^{TOT}core_{97}$ contains the well-preserved genes that define the species, and for the shortest genomes that have been sequenced, approximately 74% of their protein content belong to this core (GCA_000350185.1; *Escherichia coli* str. K-12 substr. MDS42); in contrast, for the largest genomes this fraction is only about 32% (GCA_000937575.1; *E. coli* Ec138B_L1).

By defining phylotype-specific core genomes it becomes apparent that large differences exist between the level of gene preservation for each of the phylotypes. Table 1 summarizes the phylogroup-specific sizes of core genomes, accessory gene clusters and singletons. Predictably, the phylogroup with the largest number of genes in their phylogroup-specific core genome is E2(O157). Not only do its members have large genomes, but this phylogroup is also very homogeneous and mostly contains *E. coli* O157:H7 strains that have a clonal relationship (Sharma et al., 2019). Relatively large phylogroup-specific core genomes are also observed for

phylogroup C, harboring strains of clinically relevant non-O157 enterohemorrhagic (EHEC) serotypes such as O111, O26 and O103. A third phylogroup with a large core genome is Shig2, whose members have relatively short genomes, on average, suggesting this phylogroup is relatively homogeneous, which increases the size of the core genome. However, phylogroups with fewer members produce larger core genome fractions with respect to their pangenome due to sample size bias, as illustrated by G. The phylotype with the smallest core genome is Shig1 followed by B1 and A (Table 1). The small core genome of Shig1 is related to its small genome size, while phylotypes A and B1 contain more diverse members, resulting in a larger fraction of accessory genes and a smaller phylogroup-specific core.

Table 1. Summary of pangenome analysis results. Values obtained from the different pangenome analysis using the 14 phylogroups separately and the entire set of assembled genomes (10,667 genomes) using UCLUST (Edgard, 2010). Same parameters were used to all the analysis.

Phylogroup	Core genome (97% strains)		Accessory genome		Unique		Total (Pan genome)		Core/pan (%)	No. of strains
	clusters	proteins	clusters	proteins	clusters	proteins	clusters	proteins	clusters	
All	2,663	28,566,052	82,821	22,783,754	50,499	51,099	135,983	51,400,905	1.96	10,667
A	3,184	7,142,893	41,769	3,246,591	24,501	24,828	69,454	10,414,312	4.58	2,232
B1	3,141	9,365,646	44,019	4,887,086	24,590	24,844	71,750	14,277,576	4.38	2,960
B2-1	3,708	2,016,812	10,990	619,867	7,048	7,180	21,746	2,643,859	17.05	541
B2-2	3,425	4,709,983	22,762	1,819,538	12,566	12,763	38,753	6,542,284	8.84	1,367
C	3,899	2,132,258	10,413	738,879	5,242	5,290	19,554	2,876,427	19.94	540
D1	3,666	1,006,271	10,012	318,372	7,659	7,770	21,337	1,332,413	17.18	273
D2	3,524	626,693	11,703	221,033	6,765	7,181	21,992	854,907	16.02	177
D3	3,754	668,359	7,252	201,292	4,814	4,936	15,820	874,587	23.73	177
E1	3,151	885,018	14,883	471,354	7,969	8,088	26,003	1,364,460	12.12	279
E2(O157)	4,060	3,080,073	6,128	743,413	4,442	4,535	14,630	3,828,021	27.75	750
F	3,486	698,031	9,465	288,420	5,381	5,480	18,332	991,931	19.02	199
G	3,783	365,756	5,716	98,269	4,016	4,066	13,515	468,091	27.99	96
Shig1	3,128	564,868	4,903	256,426	2,815	2,883	10,846	824,177	28.84	177
Shig2	3,732	3,383,814	6,870	719,247	4,751	4,799	15,353	4,107,860	24.31	899

Fig. 4. Pangenome representation of *E. coli* and *Shigella* species. In this circular barplot, each bar length represents the total number of proteins of a single genome, grouped by phylogroup. The proteins belonging to the overall $core_{97}$ genome are shown in green. Additional proteins shared in each phylogroup-specific $core_{97}$ genome are shown in blue, while purple is reserved for accessory proteins.

The comparison of the core genomes per each phylogroup allowed us to establish the existence of exclusive genetic signatures that would confer exclusive characteristics to each of the phylogroups found in *E. coli* species using Mash. Unique phylogroup-specific core genes are those genes present in the core genome of each phylogroup, but not found in the core genome of the other phylogroups. Such genes represent genetic signatures for all members of a phylogroup where they are conserved, but not members of the other phylogroups. The existence of unique phylogroup-specific core genes is shown in Fig. 5.

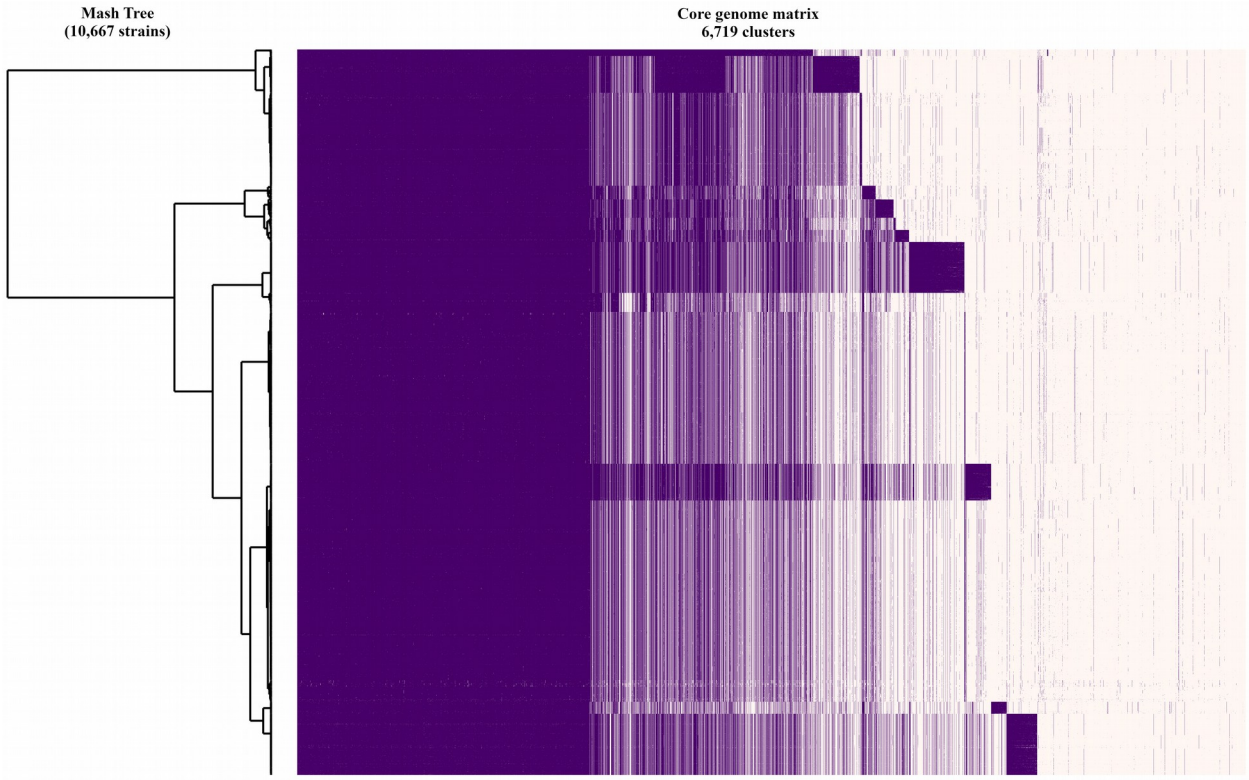


Fig. 5. Graphical representation of presence/absence of gene clusters using phylogroups core genomes sorted by Mash dendrogram phylogroup order. The plot shows 6,719 core gene clusters. The large purple region at the beginning of the plot represents the gene clusters present in the core genome of all phylogroups. Subsequent blocks of purple represent core gene clusters with a large representation in the members of a phylogroup.

Phylogroups evolve with different rates of gain/loss of protein families. Taking advantage of the existence of the medoids as representative entities of the phylogenetic groups and the $^{TOT}core_{97}$ genome we identified, a very robust phylogenic analysis was performed based on the concatenated alignment of 2,613 core genome clusters for the entire species (without paralogs) and a maximum likelihood approach (see methods) using IQ-TREE software (Nguyen *et al.*, 2015). From a total of 9,293 gene families built with these core genes (defined using UCLUST for the 14 medoid strains, see Methods), CAFE-based analysis of gene gain and loss patterns (Han *et al.*, 2013) was used to identify the evolution of protein family sizes across the species, using a random birth/death process along each lineage of the phylogenetic tree. The resulting tree is shown in Fig. 6.

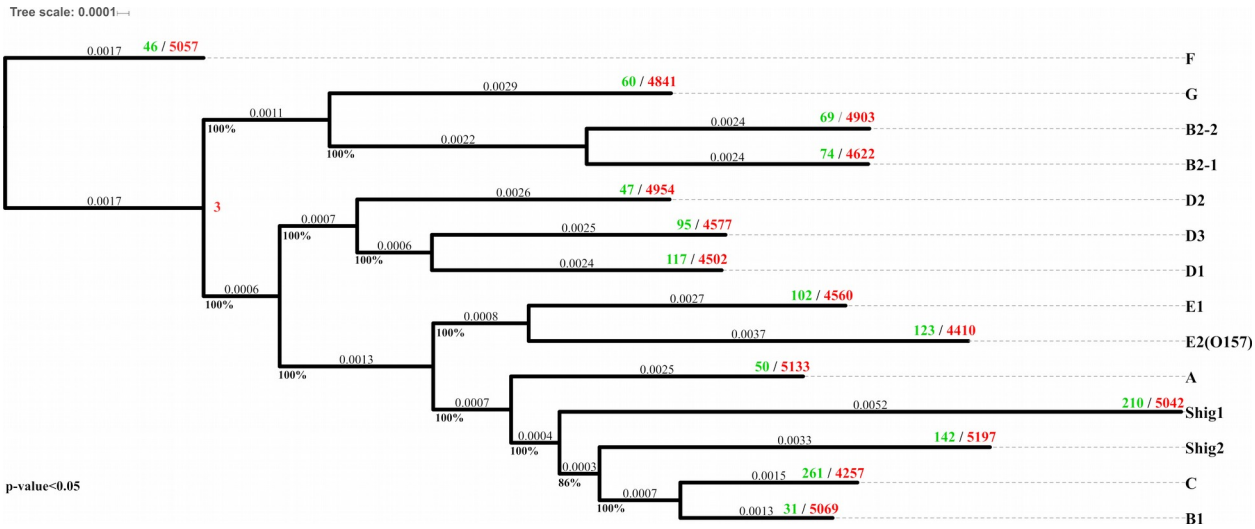


Fig. 6. Phylogenetic representation of *E. coli* species using a set of 2,613 core clusters. Branch support is shown at the beginning of each node. Maximum likelihood distances are showed on the middle of each branch. The number of gained (green color) and lost (red color) protein families was estimated for each branch using an ultrametric tree and the pangenomic matrix for the 14 phylotypes with 9,293 protein families.

This analysis led to the observation that the different phylogroups have evolved with different rates of gain/loss of protein families (Supplementary Table 3). Branches with higher ratios of gene expansion correspond to phylogroups C and Shig1 and these differed largely from the other

phylogroups. At the other end of the spectrum, phylogroups D2, F and B1 represent the lowest ratios, indicating limited gene expansion. Note, that these observations are not related to genome size (*cf*. Fig. 3) or phylogroup-specific core size (Fig. 4).

Discussion

Mash-based analyses provides a fast and highly scalable K-mer based approach that can be used on very large sets of genomes. Based on more than a hundred thousand genomes, the population structure of *E. coli* species appears to be more diverse than currently thought. The methodology applied here detected 14 phylogroups with a remarkably unequal distribution of the number of genomes across the 14 phylogroups. The current bias in the sequencing data decreases the probability of finding the genetic signatures that captures the relative homogeneity of all members of the phylogroups. As a consequence, less numerously represented phylogroups may actually contain additional, as yet unidentified phylogroups within them and at presence conclusions about their open or closed nature cannot be drawn.

Findings based on Mash analyses were supported by differences found in the analysis of the core genomes of the 14 phylogroups. These differences can be broadly defined into two categories: size of the core genome and genetic content. Differences in size could be reflective of the possible clonal nature of some of the phylogroups. Nevertheless, phylogroups that harbor commensal and environmental strains such as B1 and A possess smaller core genomes as a result of a wide variety of environmental pressures. Genomes belonging to Shig1 phylogroup have one of the smallest sets of core genes; however, this number represents almost 29% of the clusters found in this phylogroup, which is the highest ratio of core gene clusters to total pangenome clusters. Therefore, the ratio of core cluster/total cluster is an indication of the intragroup diversity (Table 1). The presence of different clusters of genes belonging to the core genome of each phylogroup support the existence of multiple non-overlapping phylogroups within the species. The comparison of shared and unique core gene clusters for each phylogroup relative to other phylogroups provides the intergroup diversity of the species.

Findings based on Mash analyses were supported by differences found in the analysis of

the core genomes of the 14 phylogroups. These differences can be broadly defined into two categories: size of the core genome and genetic content. Differences in core size expose the possible clonal nature of some of the phylogroups, in particular those containing mainly clinically relevant strains. Phylogroups that harbor commensal and environmental strains such as B1 and A generally possess smaller core genomes while their members have adapted to a wide variety of environmental pressures. Genomes belonging to the Shig1 phylogroup have the smallest number of phylogroup-specific core genes, but due to their smaller genomes this number represents almost 29% of the total clusters found in this phylogroup, which is the highest ratio of core gene clusters per phylogroup-specific pangenome. The ratio of core clusters/total clusters can be used as an indication of the intragroup diversity (Table 1). The presence of different clusters of genes belonging to the core genome of each phylogroup support the existence of multiple non-overlapping phylogroups within the species. The comparison of shared and unique core gene clusters for each phylogroup relative to other phylogroups provides the intergroup diversity of the phylogroup within the species.

The dataset of 10,667 WGS genomes was used to calculate the size of the total core genome for *E. coli* that contained 2,663 gene clusters. Such analyses have been reported multiple times in the literature, using different cut off values and criteria, as summarized in Fig. 7 and Supplementary Table 4. The data sets and analytical parameters varied widely between these studies, resulting in a variation in core genome size between 867 and 3,472, ignoring the comparison of the first two *E. coli* genomes that were published, and analyses with subsets of *E. coli*. Compared to our results, most determined core genomes were too low, in part because the parameters of inclusion were too strict for the quality of genomes that were analyzed.

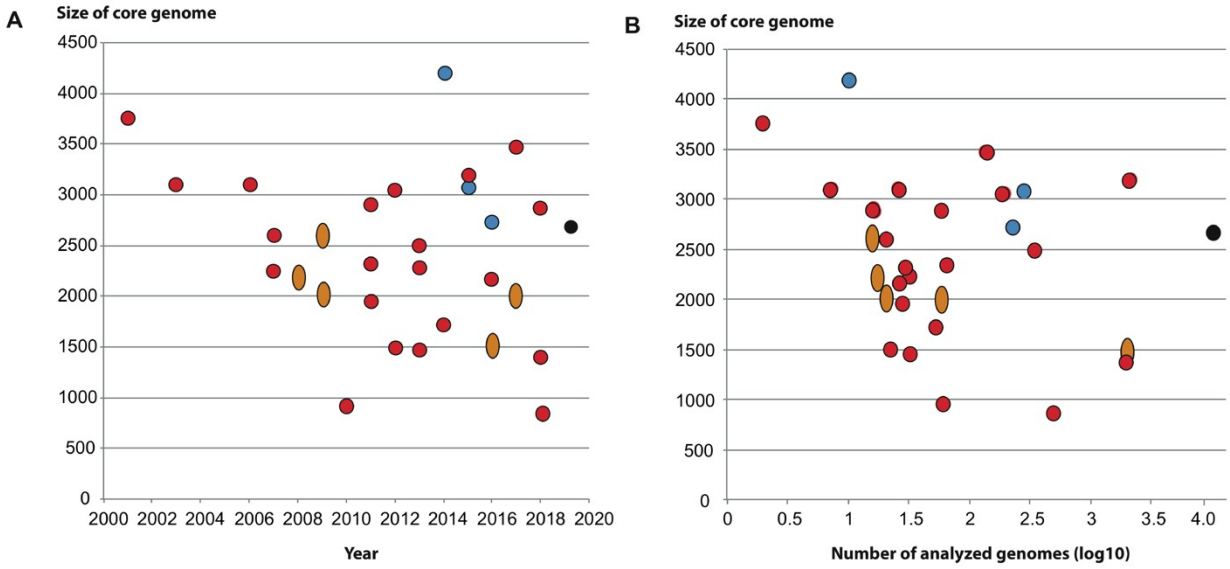


Figure 7. Core genome analyses of *E. coli* in the literature. (A) Core genome size related to the year of publication. (B) Core genome size is plotted to the log 10 number of genomes included in each study. Black symbols represent the data from this study. Oval symbols represent reported approximate sizes only. Blue symbols represent studies in which a subset of *E. coli* genomes was analyzed (EHEC only in the 2014 study and ExPEC only in the 2015 study). More information, including the sources of these data, are provided in Supplementary Table 4.

METHODS

Data Acquisition and Cleaning

A set of 12,602 genomes labeled either *Escherichia* or *Shigella* were downloaded from NCBI's Genbank on June 26, 2018. To evaluate the quality of the data set, various sequence quality scores were calculated as described elsewhere (Land *et al.*, 2014). Following the recommended cutoff value, the dataset was filtered to include only genomes with a Total Quality Score of 0.8 or higher. Applying the same cutoff value to the Sequence Quality Score alone resulted in an extremely restricted dataset that no longer addressed the goals of this study. Genome size was restricted to greater than 3 Mb and less than 6.77 Mb to trim questionably sized genomes. After applying these two steps, 10,855 genomes remained in the dataset for analysis.

To further clean the dataset, we filtered genomes that were outside the statistical distribution of

Mash distances within the dataset. Assuming that *Shigella* species are all members of *E. coli*, we chose to use type strains for the *Escherichia* and *Shigella* genera (accession numbers GCA_000613265.1 and GCA_002949675.1, respectively) to quickly filter the set of 10,855 genomes for erroneous or low-quality genomes that may have slipped through the previous cleaning steps. The Mash values of the 10,855 genomes compared to each type strain were broken into percentiles ranging from 10% to 99.995%. A cutoff percentile of 98.5% was determined to provide sufficient cleaning without risking a large loss of data (data not shown) and was applied to each type strain Mash value set. Genomes that were found in both sets after filtering were retained to produce the final dataset of 10,667 genomes.

Mash and Clustering Analysis

Genetic distances between all 10,667 genomes were calculated using 'Mash dist' with a k-mer size of 21 and a sampling size of 10,000. The resulting output was converted into a distance matrix with matching columns and rows. To improve the clustering results and to provide a standard metric that allows comparison of different analytical methods, we converted the Mash distance value into a similarity measure via the Pearson correlation coefficient (Kirch, 2008). This returns values ranging from -1 (total negative linear correlation) to 1 (total positive linear correlation), where 0 is no linear correlation. Since clustering-based methods require a distance measure, the values were subtracted from 1 to convert them into a distance measure. These distance measures were then clustered using hclust and the "ward.D2" method. A heatmap was generated using the hclust dendrogram to reorder the heatmap, while values from the raw distance matrix of Mash distances were mapped to color. To determine the height to cut the hclust dendrogram and to accurately predict phylogroups that optimally overlapped with existing phylotypes, we compared multiple different cutoff values and methods to obtain cutoff values. Taking the maximum height present in the hclust dendrogram and multiplying it by 1.25^{-02} was found to provide both accurate predictions and a standard method that scales with the data supplied. Sufficient accuracy was defined by the cutoff at which the last accepted phylotype was visible, in this case representing the C phylotype splitting off from B1. Some detailed results of both the cutoff percentile and hclust height testing are included for 10,667 genomes in

Supplementary Table 1.

Medoid Selection for Species Representation

Using the Mash values for the entire species, a medoid was defined for each phylogroup. This was defined as the member of a phylogroup to which the average dissimilarity for all other phylogroup member is the smallest. This was done by using the aggregate function of R to find the mean across each phylotype. The alternative approach, isolating each phylogroup and then reclustering and calculating the medoid, did not yield as accurate results as calculating the medoid per phylogroup with respect to the entire 10,667 genome dataset.

Scaling up to over 100,000 Escherichia coli by addition of SRA read.

The keyword "*Escherichia coli*" filtered with "DNA" for biomolecule and "genome" for type was used to retrieve SRA ids from the NCBI SRA website on March 22, 2019. For large scale data transfer, these SRA genomes were downloaded using the high throughput file transfer application Aspera (http://asperasoft.com). According to the variety of sequencing technologies used to generate genome data, the obtained read sets of 102,091 genomes in SRA format were divided into five subsets to ease computational and organizational load as follows: 3 Illumina paired read sets, 1 mixed technology with paired reads, and 1 mixed technology with single reads. The 5 sets of reads were then converted from fastq to fasta format to be processed by Mash using a python script.

The sequence reads were sketched using Mash (v2.1). This version change was due to the addition of read pooling in the read mode which automatically joins paired reads, eliminating the need to concatenate or otherwise process paired read sets. All read sets were sketched individually so that read sets that caused an error when sketching were dropped from the analysis before sketching. The -m setting was set to 2 to decrease noise in the sketches of the reads. After sketching the reads within the subsets, all sketches were concatenated into a sketch for that subset using the paste command of Mash. The concatenated sketch of each subset was then compared to the 14 medoids using Mash dist. As all five subsets had the same reference, the dist

output from each subset was concatenated to one file. This single SRA dist output file was then analyzed to evaluate the quality of the SRA dataset. Due to how Mash distances are calculated, the k-mer size and sketches sampled settings can consistently flag genomes of very low quality, since major basis of a Mash value is how many hits are present out of sketches sampled. The top 5 most numerous distances of the SRA read sets corresponded to 0 to 4 hits of the possible 10,000 sketches per genome. This indicates presence of extremely low-quality genomes. A histogram of the SRA Mash distance results was created to visualize the distribution of Mash distances of the entire 102,091 SRA reads dataset (results not shown), initially using a cutoff of 0.1 Mash distance. However, a final Mash distance cutoff of 0.04 was chosen based on the maximum Mash value in the 10,667 whole set that was 0.393524. Although this higher cutoff might potentially eliminate useful information, it insured quality of the SRA dataset. This retained 95,525 reads that had at least one Mash distance to a phylogroup medoid. The distances were transferred into a matrix with reads as columns and rows containing a phylogroup medoid. For each read the smallest Mash distance to a medoid was identified, and the corresponding medoid noted (Supplementary Table 2). We then created a distance matrix from the Mash distance output of the 95,525 reads that met the above cutoff with reads as rows and medoids as columns. Due to computational pressure this distance matrix was loaded into Python 3 instead of R. A clustered heatmap was made using Seaborn, Matplotlib, and Scipy with the clustermap function. Instead of clustering both rows and columns, columns (phylogroups) were ordered the same as Figure 1 and rows were sorted as follows: number of hits to phylogroups (ascending = True) and Mash distance (ascending = False). This provided a quick visualization method for the SRA dataset with a consistent sorting criterion to make comparison between Figure 2 and the Supplemental heatmaps much easier.

Cyoscape visualization of MASH analysis

The Mash distance matrix of the 10,667 genomes was transmuted into a new 3 column matrix where the first two columns contains two genomes to be compared and the third column contains the Mash value for that pairwise comparison. A sliding cutoff ranging from 0.04 to 0.0095 with increments of 0.005 was applied to the Mash value column. After each cutoff filter was applied a data table was compiled with the cutoff identified in the name. These data tables were imported

into Cytoscape, the Prefuse Force Directed layout was applied, and phylotype membership mapped with a metadata table. For each cutoff the resultant graph was output as an SVG. All SVGs were then compiled into an animated transition.

Statistical analysis of genome sizes and percent GC content

Genome sizes and percent of GC content was calculated using infoseq package from EMBOSS suite v6.6.0.0. A dataframe with sequence ID, percentage of GC content, genome size and phylogroup ID was made. Library ggplot2 from R was used to plot genome sizes and GC content. Library dplyr from R was used to perform analysis of Variance ANOVA test and Turkey HSD tests. The homogeneity of variances was tested using Levene's test and the normality assumption of the data was checked using Shapiro-Wilk test. As some of the groups didn't meet the criteria of the assumption of normality, Kruskal-Wallis test was performed as well as non-parametric alternative to one-way ANOVA. Kruskal-Wallis test rejected both null hypothesis (means of genome size or percent of GC content are similar between the different phylogroups), with p-value < 2.2e-16 in both cases. Raw results from these test are available in Supplementary Table 5.

Pangenome analyses and clustering

All 10,667 genomes were reannotated using Prokka v1.13 (Seemann, 2014), with parameters --rnammer --kingdom Bacteria --genus *Escherichia* –species *coli* --gcode 11. All protein-coding sequences (n=51,400,905) were clustered using UCLUST from USEARCH v.10.0.240 (Edgar, 2010), into protein families using cut-off values of 80% of protein sequence similarity, 80% of query sequence coverage, e-value equal or lower than 0.0001 (parameters -evalue 0.0001 -id 0.8 -query_cov 0.8). The total pangenome was recorded. For the core genome various inclusion percentages were compared, since we included draft genomes existing in multiple contigs. The optimum was defined that allowed 3% omissions, giving a core genome was defined as those genes present in 97% of the genome collection. Therefore, protein families with presence in at least 97% of the total set strains, were considered as part of the core genome of *E. coli* species.

The pan- and core genome of each of the 14 phylogroups were then separately clustered using

the same cut-off parameters as for the entire set at species level.

Core genome matrix

Core genome clusters for the 14 phylotypes obtained using UCLUST v.10.0.240 in the previous analysis were used again with UCLUST v.10.0.240 using the same parameters as input to find the intersection of core genes between the core clusters of the 14 phylotypes. A binary matrix with cluster ID as column labels, genomes as row names, and the number of genes belonging to that cluster as the cell value was constructed using the main output from UCLUST. This matrix was then supplied to an "in house" python script that sorts the pangenome matrix such that the gene clusters found in all phylogroups are placed first (species' core genome). Then groups are sorted by abundance per phylogroup to isolate phylogroup core genes. All leftover gene groups are sorted by phylogroup and abundance and added to the end of the sort list. The Mash tree obtained earlier for the 10,667 dataset was then loaded and used to sort the order of the organisms. Finally, Matplotlib was used to visualize the sorted matrix.

Phylogenetic analysis of core gene families

The set of core gene clusters of the 14 medoids was extracted from the core genome clusters of the entire species and from them single copy ortholog groups were identified to construct a phylogenomic tree. In total a set of 2,613 single gene (clusters without paralogs paralogs) ortholog groups were aligned using MAFFT v.7.110 (Katoh and Standley, 2013). The model of evolution per each of the 2,613 protein clusters was calculated using IQ-TREE v.1.6.10 (Nguyen *et al.*, 2015) with parameters -m TESTONLY nt AUTO. Once the best model of evolution was obtained for each of the core protein families, those clusters that shared model of evolution were sent together to IQ-TREE for a better estimation of the substitution model parameters using -m MF+MERGE, -nt AUTO and selecting the final model of evolution with mset. In a last step, all partitions obtained with their corresponding model of evolution were sent again to IQ-TREE for final estimation of the phylogenetic tree for the 14 medoids using ultrafast bootstraping approach (-bb 1000).

For estimation of protein family gain and loss events, the Maximum Likelihood tree was used as

482 an ultrametric tree using ace library from ape (Paradis *et al.*, 2004) in R v.3.6.0 (R Core Team,
483 2013). To obtain the pangenome matrix needed as input for CAFE program v.4.2.1, the
484 pangenome of the 14 medoids was constructed using UCLUST (with same parameters as in
485 previous analyses). A pivot table was built using the main output from UCLUST and pandas
486 library in a python3 script using the function pivot_table with agglomeration function=sum.
487 CAFE program was used for gene family expansion/contraction analysis, using option -s for an
488 optimization algorithm to find the value(s) of λ that maximize the log likelihood of the data for
489 all families. Families showing significant size variance were identified based on 1,000 random
490 samples and a p-value cutoff of 0.05. Deviated branches were further identified based on the
491 Viterbi algorithm in CAFE with a p-value cutoff of 0.05.

492

493 **REFERENCES**

Jang, J., Hur, H.-G., Sadowsky, M.J., Byappanahalli, M.N., Yan, T., and Ishii, S. (2017) Environmental *Escherichia coli*: ecology and public health implications-a review. *Journal of Applied Microbiology* **123**: 570–581.

Fischer Walker, C.L., Sack, D., and Black, R.E. (Beriotto, I., Browning, D.F., Squire, D., et al. (2017) Sequencing a piece of history: complete genome sequence of the original *Escherichia coli* strain. *Microbial Genomics* **3**(3):mgen000106.

Pettengill, E.A., Pettengill, J.B., and Binet, R. (2016) Phylogenetic Analyses of *Shigella* and Enteroinvasive *Escherichia coli* for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. *Front Microbiol* **6**:1573.

Chattaway, M.A., Schaefer, U., Tewolde, R., Dallman, T.J., and Jenkins, C. (2017) Identification of *Escherichia coli* and *Shigella* Species from Whole-Genome Sequences. *Journal of Clinical Microbiology* **55**: 616–623.

Clermont, O., Bonacorsi, S., and Bingen, E. (2000) Rapid and Simple Determination of the *Escherichia coli* Phylogenetic Group. *Applied and Environmental Microbiology* **66**: 4555–4558.

Gordon, D.M., Clermont, O., Tolley, H., and Denamur, E. (2008) Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method: MLST versus Clermont method. *Environmental Microbiology* **10**: 2484–2496.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010) The population genetics of commensal *Escherichia coli. Nature Reviews Microbiology* **8**: 207–217.

99

Clermont, O., Christenson, J.K., Denamur, E., and Gordon, D.M. (2013) The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups: A new *E. coli* phylo-typing method. *Environmental Microbiology Reports* **5**: 58–65.

Meier-Kolthoff, J.P., Hahnke, R.L., Petersen, J., Scheuner, C., Michael, V., Fiebig, A., et al. (2014) Complete genome sequence of DSM 30083T, the type strain (U5/41T) of *Escherichia coli,* and a proposal for delineating subspecies in microbial taxonomy. *Standards in Genomic Sciences* **9**: 2.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.A., Tiedje, J.M., and Whittam, T.S. (2009) Cryptic Lineages of the Genus *Escherichia*. *Applied and Environmental Microbiology* **75**: 6534–6544.

Carlos, C., Pires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I., Gomes, T.A., et al. (2010) *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiology* **10**: 161.

Vangchhia, B., Abraham, S., Bell, J.M., Collignon, P., Gibson, J.S., Ingram, P.R., et al. (2016) Phylogenetic diversity, antimicrobial susceptibility and virulence characteristics of phylogroup F *Escherichia coli* in Australia. *Microbiology* **162**: 1904–1912.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**(1):132.

Struyf, A., Hubert, M., and Rousseeuw, P. (1997) Clustering in an Object-Oriented Environment. *J Stat Soft* **1**:1-30.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13**: 2498–2504.

Weinert, L.A. and Welch, J.J. (2017) Why Might Bacterial Pathogens Have Small Genomes? *Trends in Ecology & Evolution* **32**: 936–947.

Bhunia, A.K. (2018) *Escherichia coli*. In, *Foodborne Microbial Pathogens*. New York, NY: Springer New York, pp. 249–269.

Balbi, K.J., Rocha, E.P.C., and Feil, E.J. (2009) The Temporal Dynamics of Slightly Deleterious Mutations in *Escherichia coli* and *Shigella* spp. *Mol Biol Evol* **26**: 345–355.

Sharma, V.K., Akavaram, S., Schaut, R.G., and Bayles, D.O. Comparative genomics reveals structural and functional features specific to the genome of a foodborne *Escherichia coli* O157:H7. *BMC Genomics*. 2019;20(1):196.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**: 268–274.

Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W. (2013) Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Mol Biol Evol* **30**: 1987–1997.

Land, M.L., Hyatt, D., Jun, S.R., Kora, G.H., Hauser, L.J., Lukjancenko, O., Ussery, D.W. Quality scores for 32,000 genomes (2010). Stand Genomic Sci. 2014;9:20.

Kirch, W. ed. (2008) Pearson's Correlation Coefficient. In, *Encyclopedia of Public Health*. Dordrecht: Springer Netherlands, pp. 1090–1091.

Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Katoh, K. and Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**: 772–780.

Paradis, E., Claude, J., and Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

ACKNOWLEDGMENTS

This work was supported by NIH/NIGMS grant 1P20GM121293 and from the Helen Adams & Arkansas Research Alliance Endowment in the Department of Biomedical Infomatics, College of Medicine.

Conflict of interest

None declared

Legends of Tables

Table 1. Summary of pangenome analysis results. Values obtained from the different pangenome analysis using the 14 phylogroups separately and the entire set of assembled genomes (10,667 genomes) using UCLUST (Edgard, 2010). Same parameters were used to all the analysis

Legends of Figures

Fig. 1. Heatmap representation of 10,667 genomes using Mash distances. The color bars at the top of the heatmap identify the phylogroups as predicted from the analysis (see key). The scale to the left of the dendrogram corresponds to the resultant cluster height of the entire dataset obtained from hclust function in R (details in Methods). The colors in the heatmap are based on the pairwise Mash distance between the genomes. Blue-green colors represent similarity between genomes with the darkest blue-green corresponding to identical genomes reporting a Mash distance of 0. Brown colors represent low genetic similarity per Mash distance, with the darkest brown indicating a maximum distance of ~ 0.039. Genomes of relative median genetic similarity have the lightest color.

Fig. 2. Heatmap representation of 91,261 sequence reads from the SRA database. The heatmap colors are based on the pairwise Mash distance between the SRA read sets and the 14 medoid genomes of each phylogroup, which are presented in the same order as in Fig. 1. To be included, SRA reads sets had to have 3 or more medoid comparisons producing a Mash distance equal to or less than 0.04. This removed 4,264 SRA read sets from the dataset. The number of SRA reads mapped to each medoids is given below the heatmap. Supplementary Fig. 2 contains additional cut-offs ranging from one to 14 phylogroups.

Fig. 3. Violin-plots of the distribution of genome size (A) and genomic GC content (B) by phylogroup. Bar-plots inside the violins represent values for mean and mean plus one standard deviation per phylogroup. Phylogroups that have values significantly different to all other phylogroups (according to F statistics test) are marked with a red asterisk.

Fig. 4. Pangenome representation of *E. coli* and *Shigella* species. In this circular barplot, each bar length represents the total number of proteins of a single genome, grouped by phylogroup. The proteins belonging to the overall core$_{97}$ genome are shown in green. Additional proteins shared in each phylogroup-specific core$_{97}$ genome are shown in blue, while purple is reserved for accessory proteins.

Fig. 5. Graphical representation of presence/absence of gene clusters using phylogroups core genomes sorted by Mash dendrogram phylogroup order. The plot shows 6,719 core gene clusters. The large purple region at the beginning of the plot represents the gene clusters present in the cores genome of all phylogroups. Subsequent blocks of purple represent core gene clusters with a large representation in the members of a phylogroup.

Fig. 6. Phylogenetic representation of *E. coli* species using a set of 2,613 core clusters. Branch support is shown at the beginning of each node. Maximum likelihood distances are showed on the middle of each branch. The number of gained (green color) and lost (red color) protein families was estimated for each branch using an ultrametric tree and the pangenomic matrix for the 14 phylotypes with 9,293 protein families.

Figure 7. Core genome analyses of *E. coli* in the literature. (A) Core genome size related to the year of publication. (B) Core genome size is plotted to the log 10 number of genomes included in each study. Black symbols represent the data from this study. Oval symbols represent reported approximate sizes only. Blue symbols represent studies in which a subset of *E. coli* genomes was analyzed (EHEC only in the 2014 study and ExPEC only in the 2015 study). More information, including the sources of these data, are provided in Supplementary Table 3.

Supplementary information

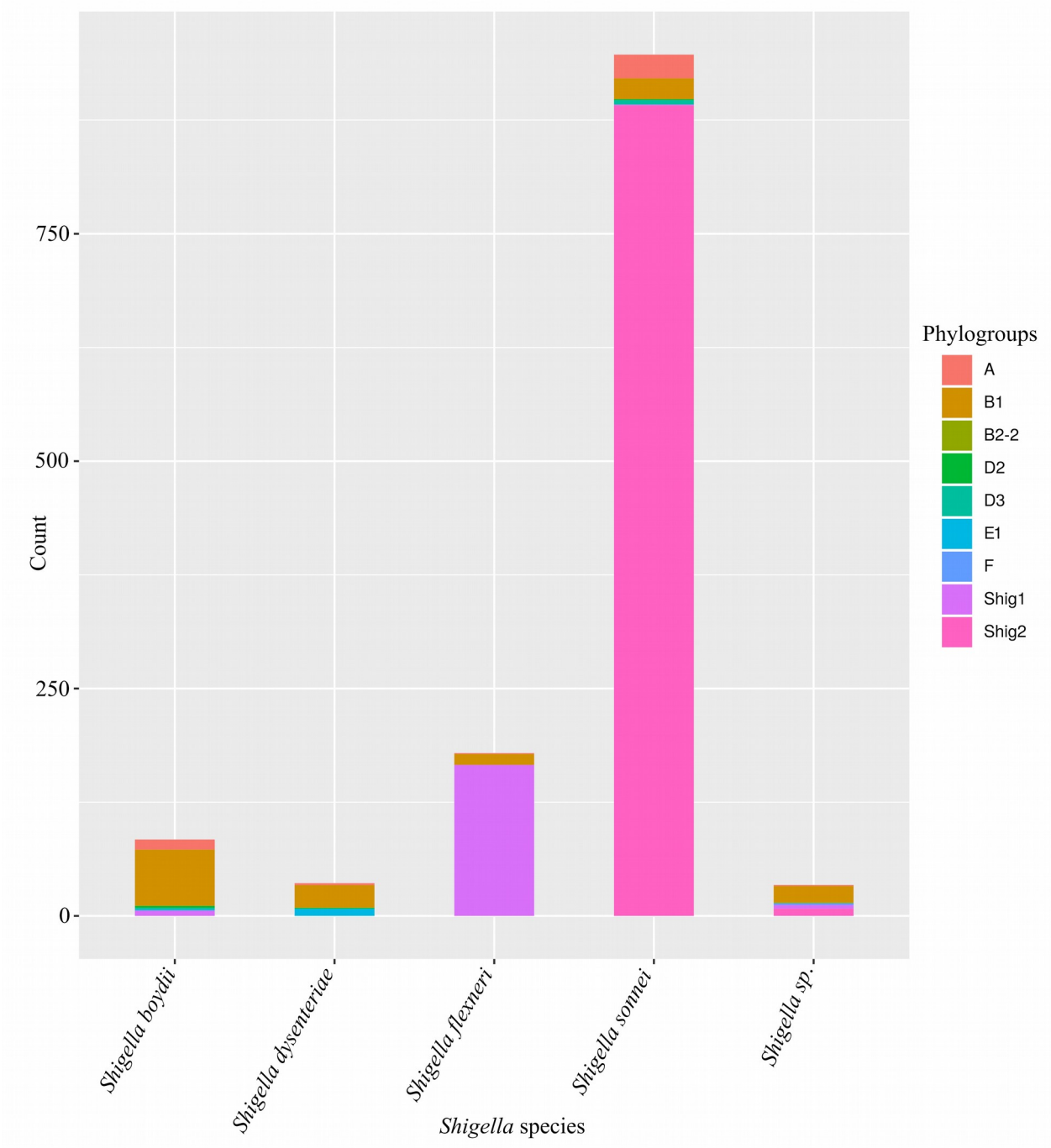
Supplementary Table 1. 10,667 WGS annotation numbers and strain names used in this study, their metadata and quality scores. This file also includes a list of the medoid genomes used in this study.

Supplementary Table 2. SRA metadata including read name, the predicted phylogroup, the number of hits a read has to phylogroup medoids that is above a cutoff of 0.04.

Supplementary Table 3. Gene gain and gene loss analysis using CAFE v3 software.

559 **Supplementary Table 4.** *E. coli* pangenome timescale obtained from literature.

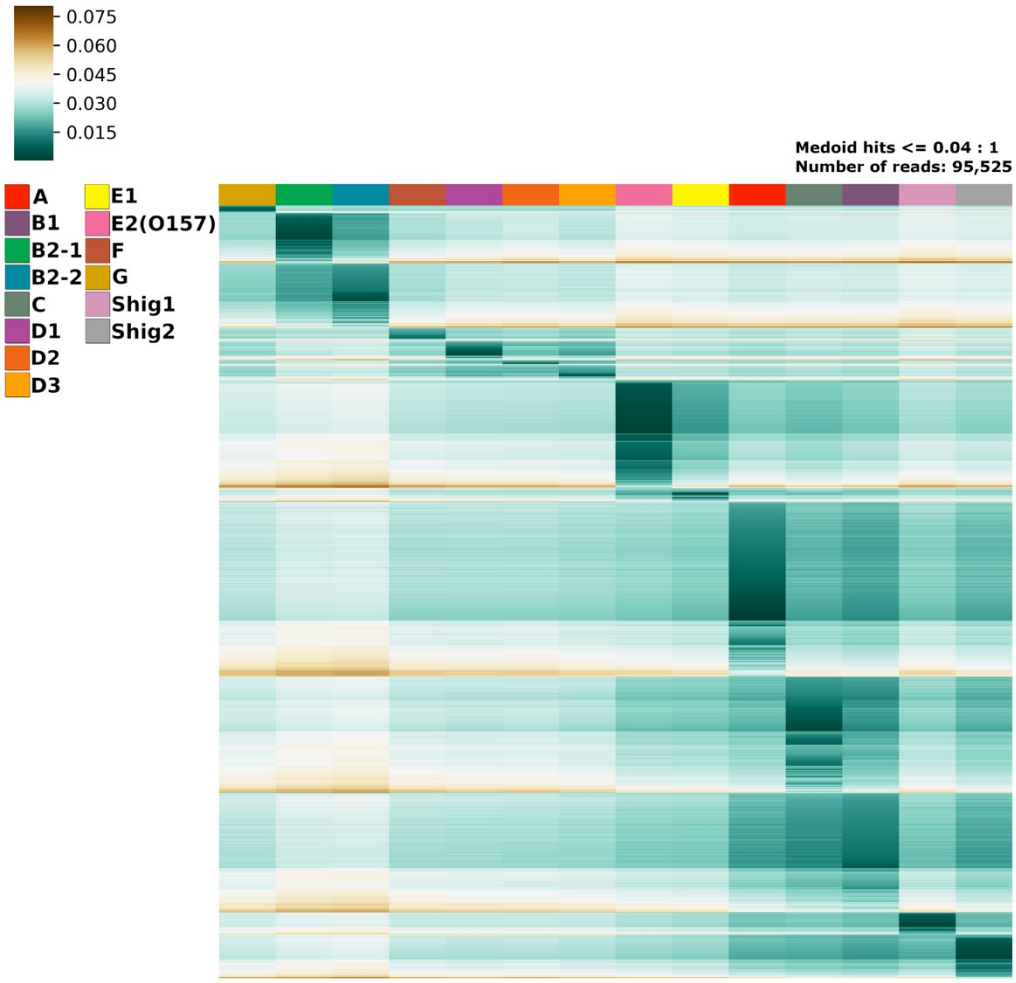
560 **Supplementary figures**



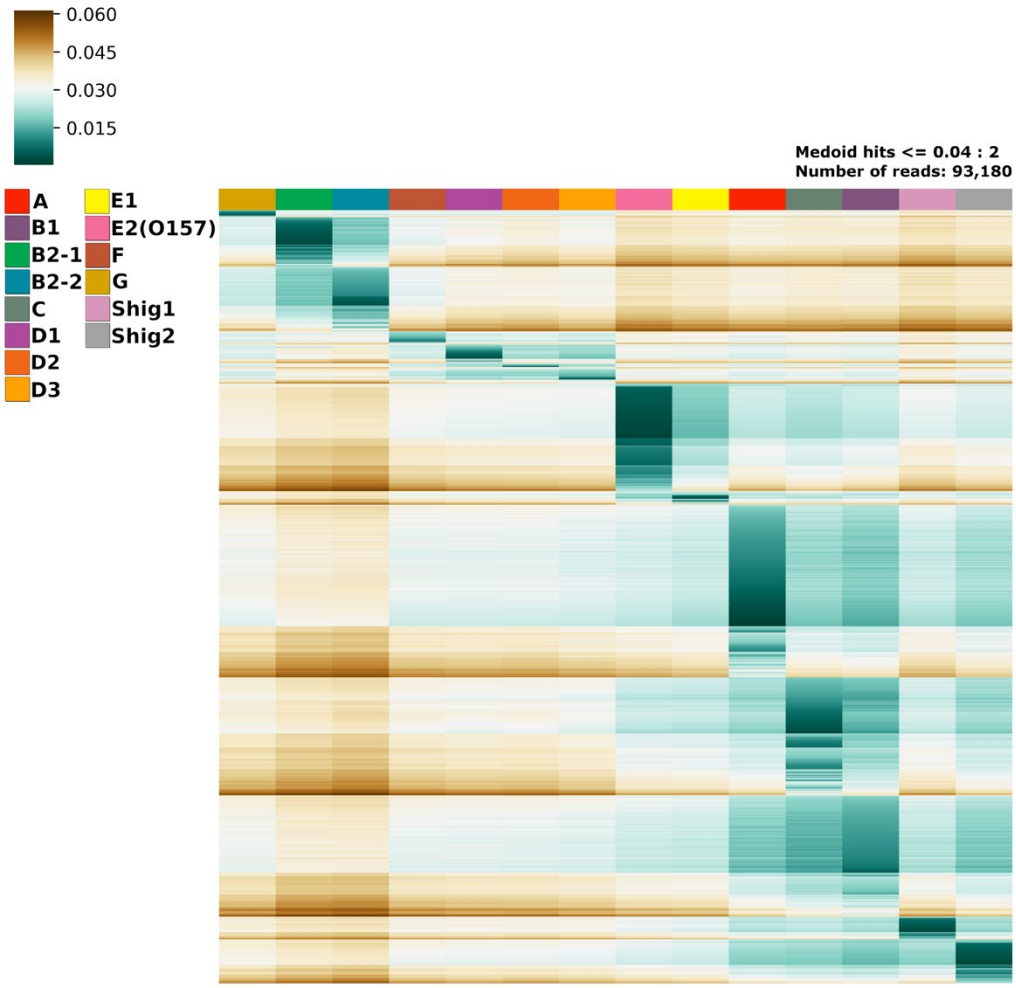
562 Supplementary Figure 1. Distribution of *Shigella* genomes over phylogroups.

Supplementary Figure 2. Heatmap of all SRA reads that had a Mash score of at least 0.04 to one medoid.

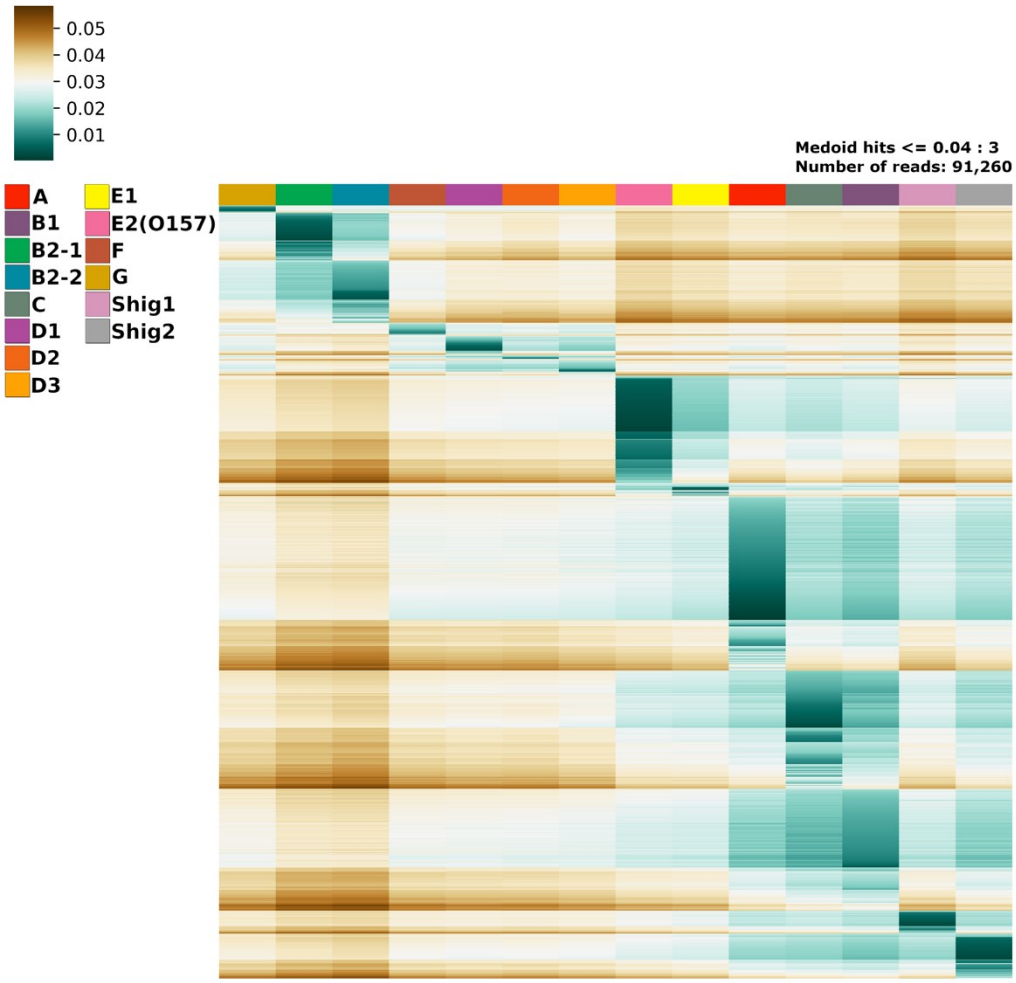
a)



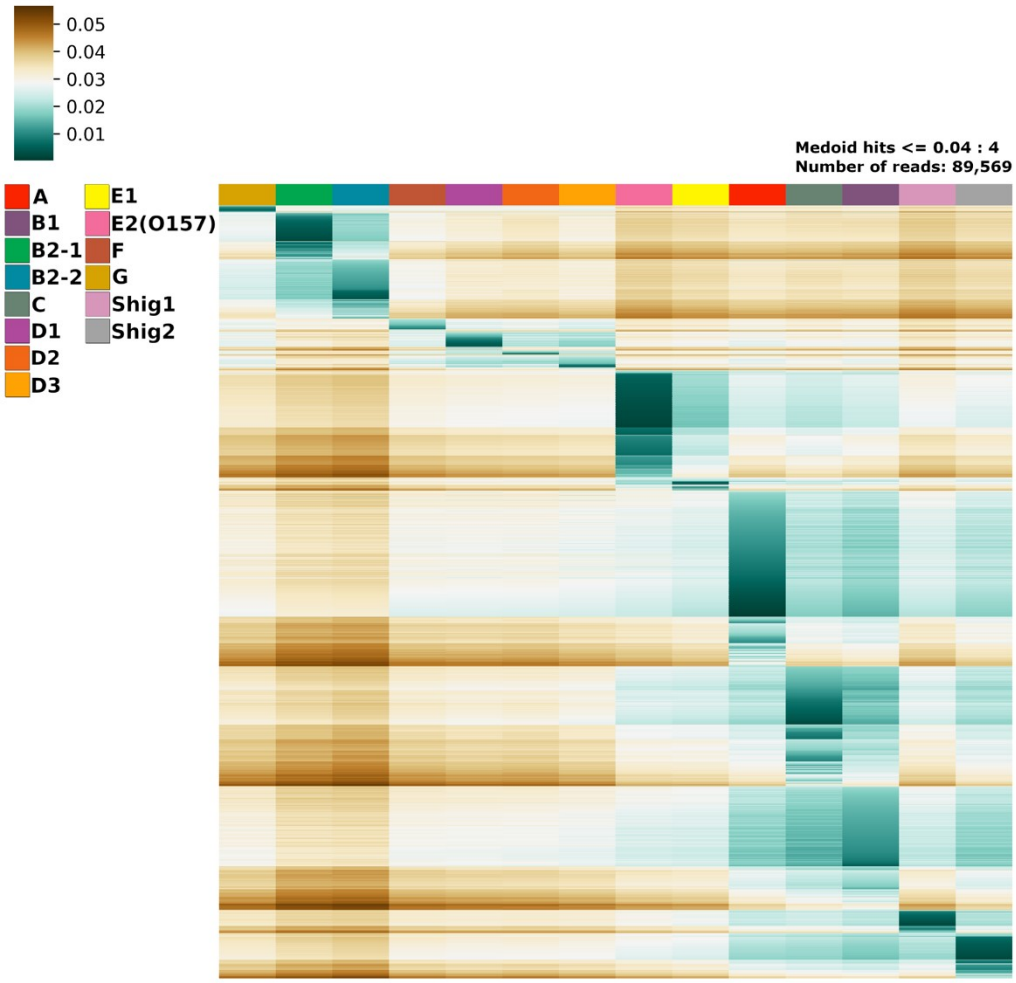
b)



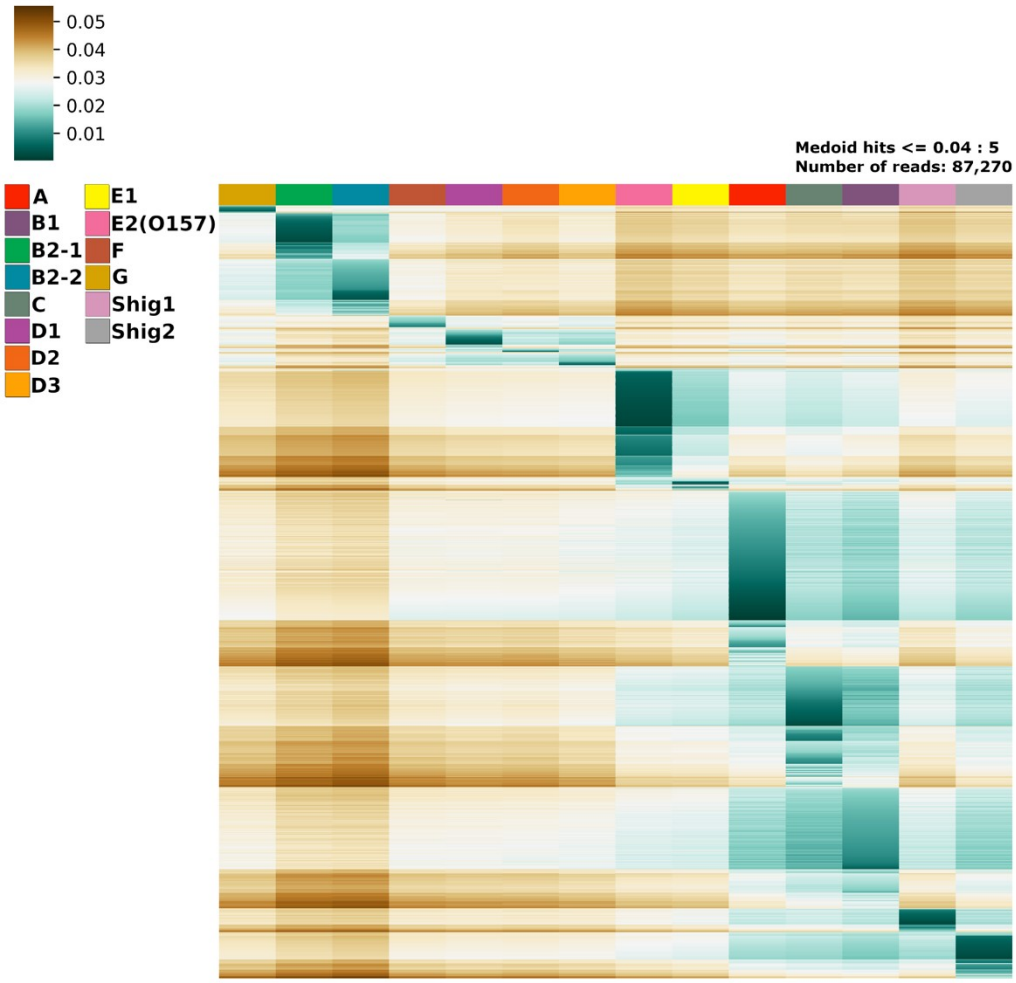
c)



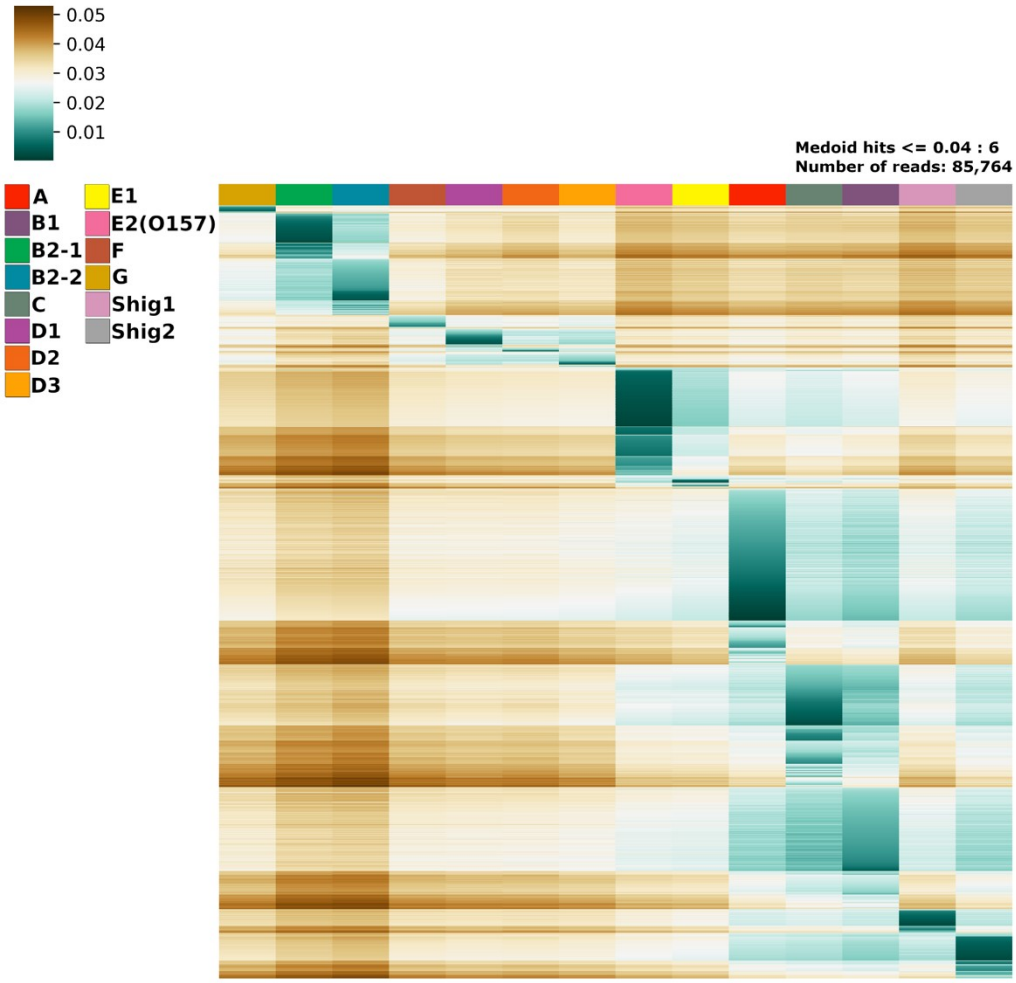
d)



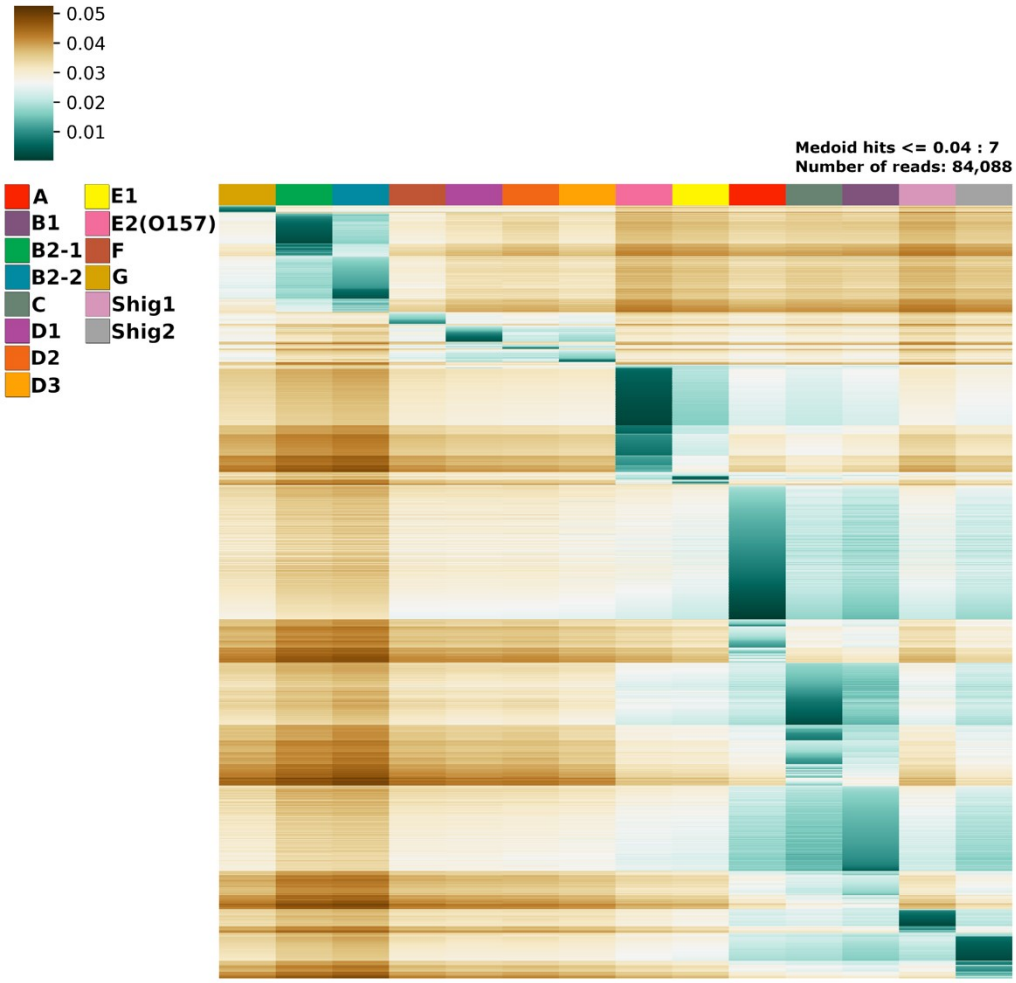
e)



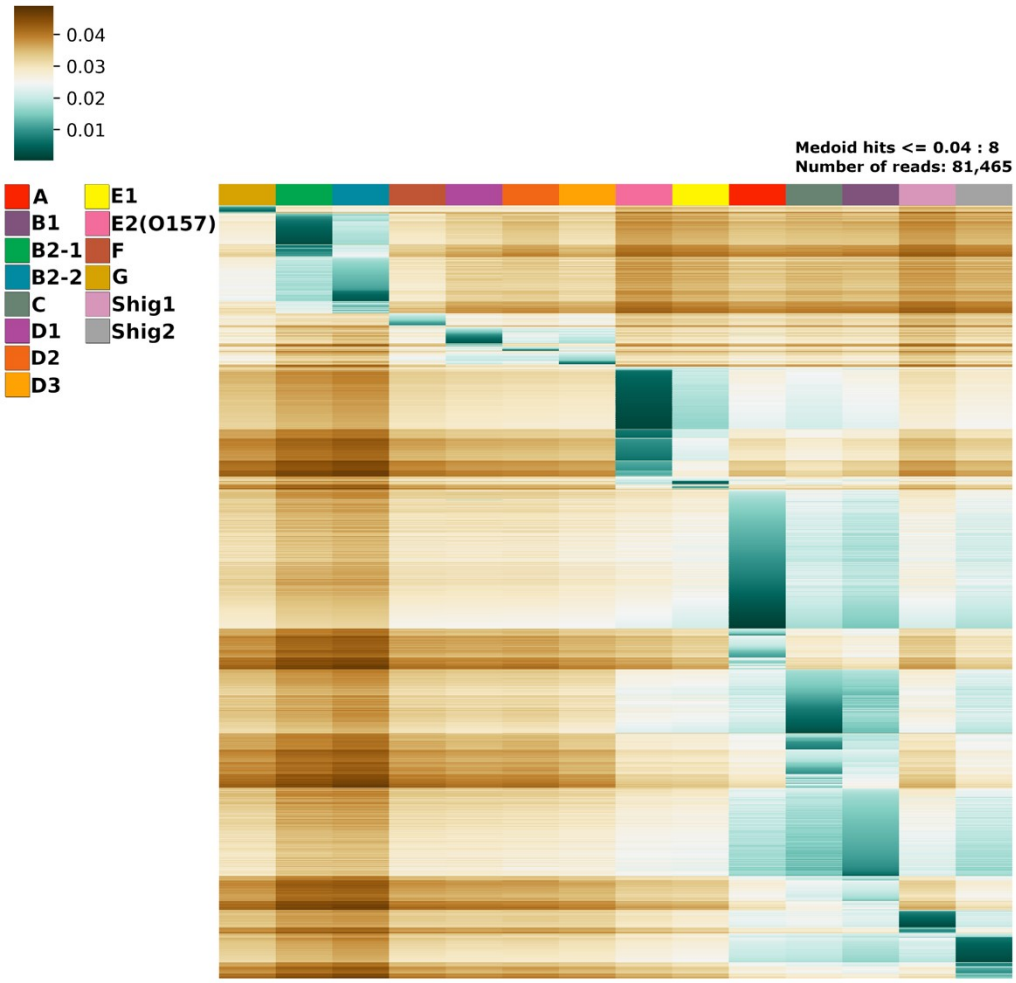
f)



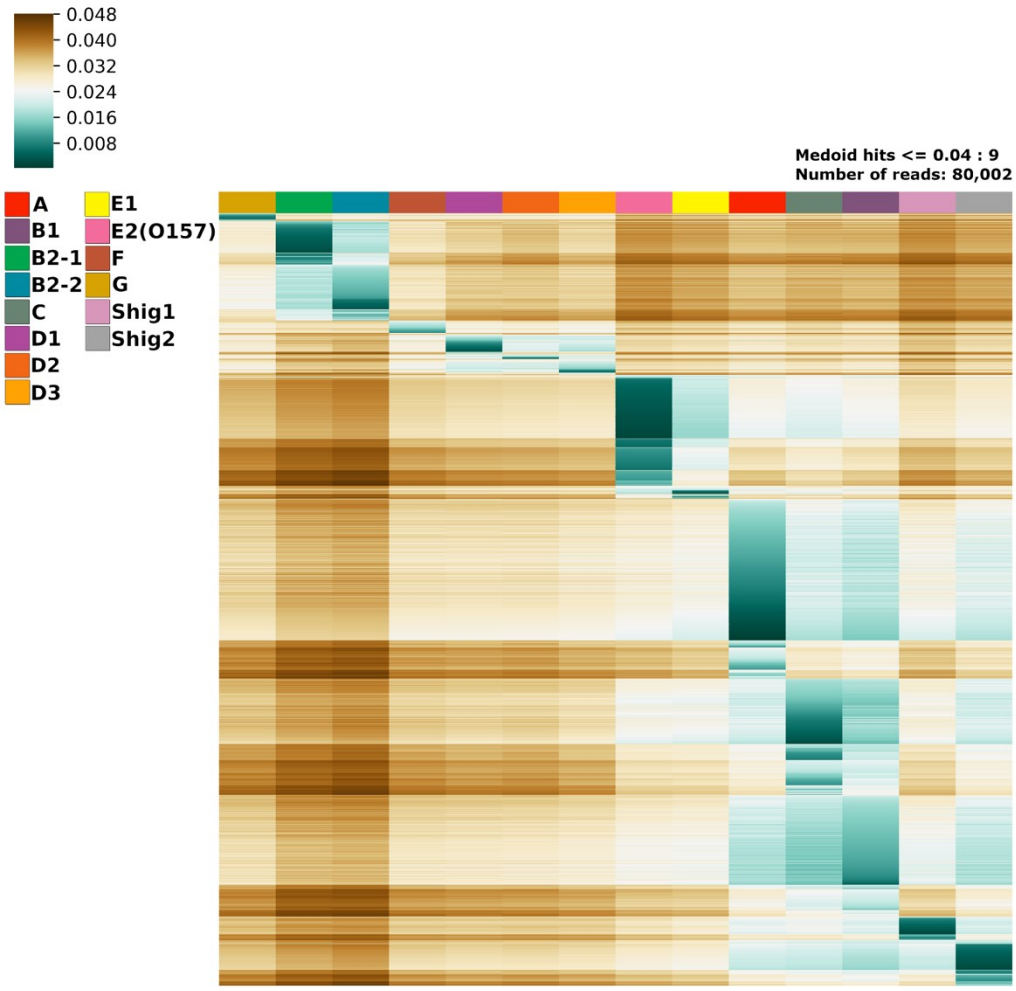
g)



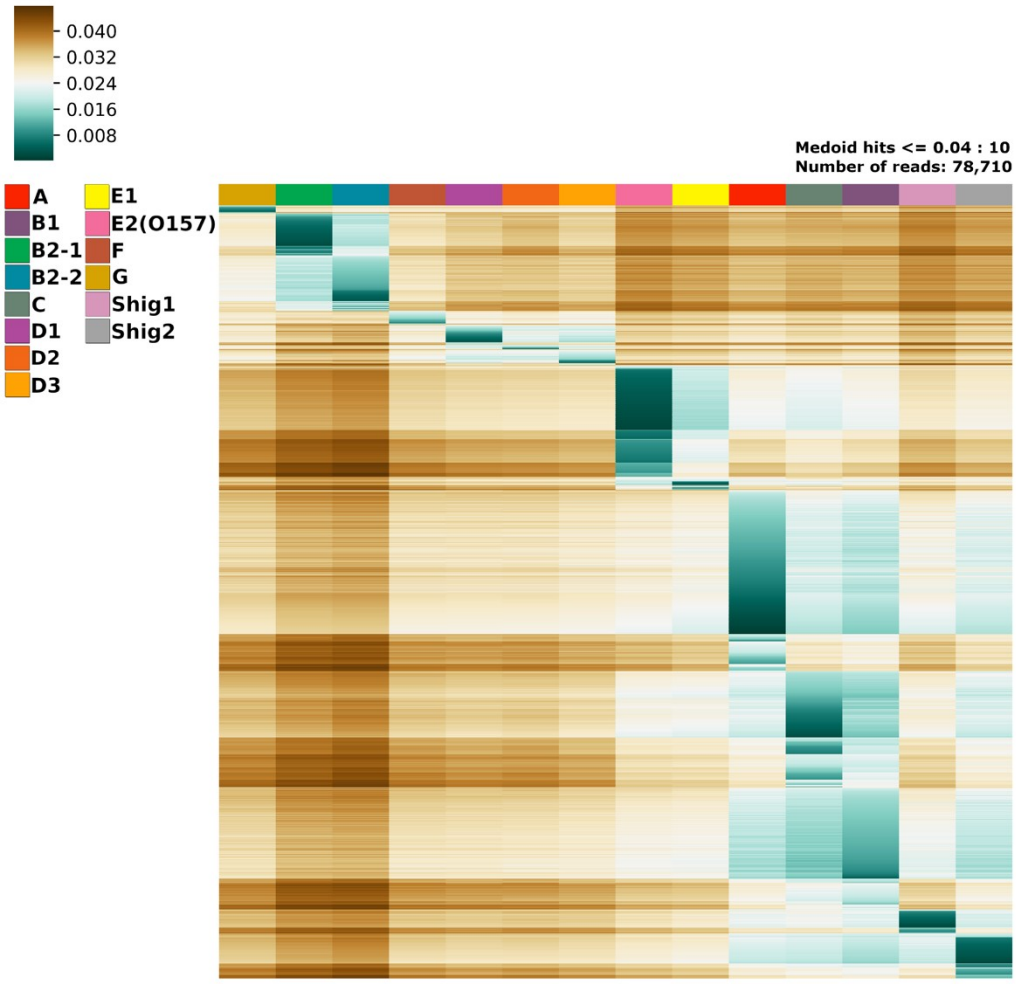
h)



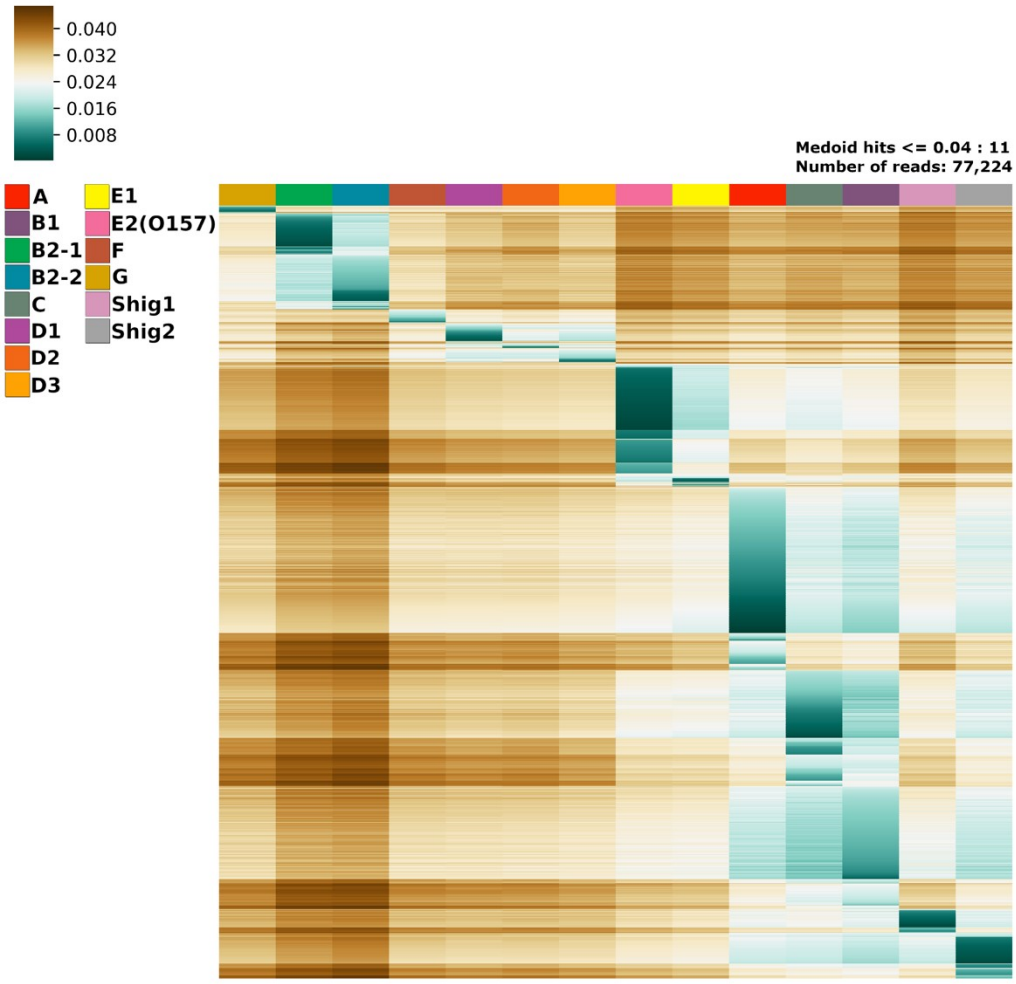
i)



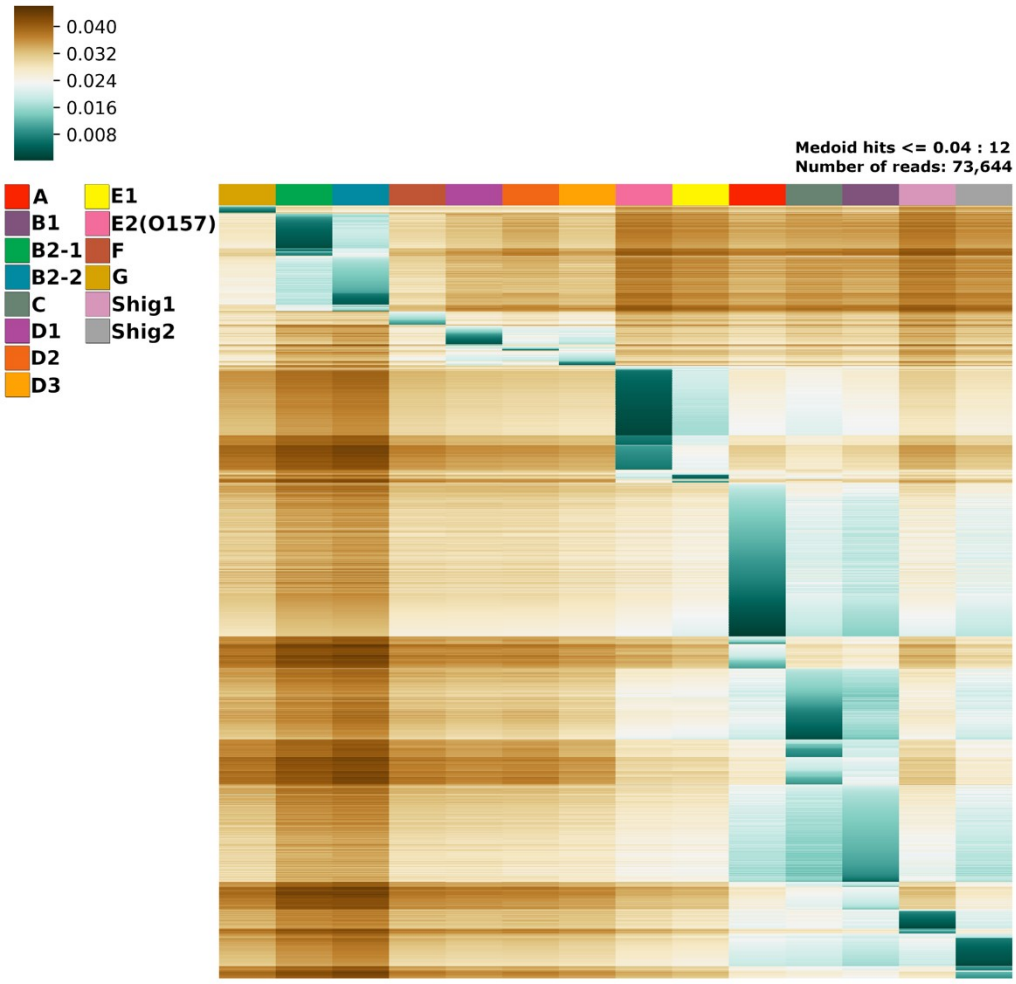
j)



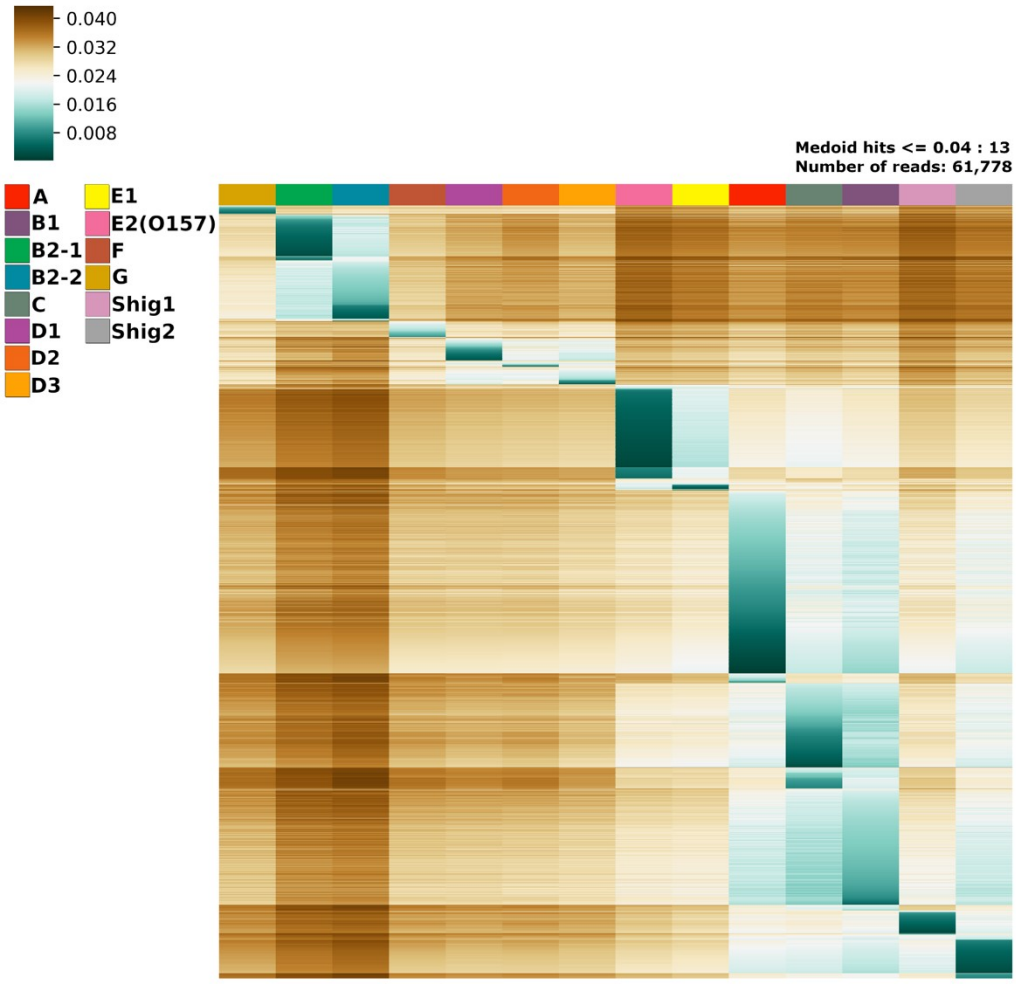
k)



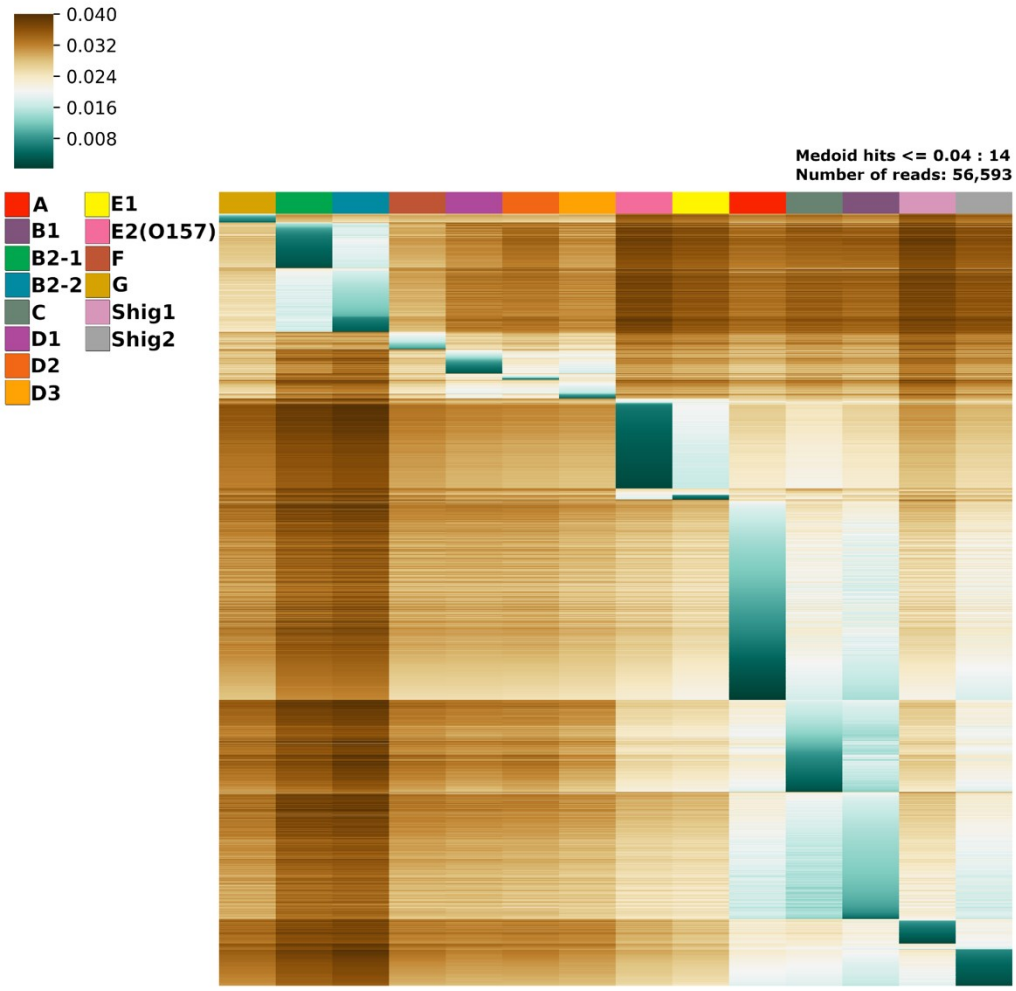
l)



m)



n)



Supplementary Figure 3. Core genomes established at a cutoff of 90% to 100% per phylogroup. Last section represents the rate of cluster drop-off between percentages (90% to 99%)