Case Report

Not peer-reviewed version

Phylogroup Homeostasis of *Escherichia coli* in the Human Gut Reflects the Physiological State of the Host

Maria S. Frolova [*] , Sergey S. Kiselev , Valery V. Panyukov , Olga N. Ozoline [*]

Posted Date: 16 May 2025

doi: 10.20944/preprints202505.1278.v1

Keywords: *E. coli* phylogroups; *k*-mer-based phylotyping; intraspecific phylotyping; colorectal adenoma; carcinoma; antibiotic treatment; probiotic therapy; obesity; Mediterranean diet; interspecific relationships

Case Report

Phylogroup Homeostasis of *Escherichia coli* in the Human Gut Reflects the Physiological State of the Host

Maria S. Frolova [1,*], Sergey S. Kiselev [1], Valery V. Panyukov [2] and Olga N. Ozoline [1,*]

[1] Department of Functional Genomics of Prokaryotes, Institute of Cell Biophysics of the Russian Academy of Sciences, Federal Research Center Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences, 142290 Pushchino, Russia

[2] Department of Bioinformatics, Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Science, 142290 Pushchino, Russia

* Correspondence: mosmasha@mail.ru (M.S.F.); Tel.: +7-496-739-156 (M.S.F.); ozoline@rambler.ru (O.N.O.); Tel.: +7-496-739-140 (O.N.O.)

Abstract: The development of alignment-free approaches using short *k*-mers as barcodes has revolutionized taxonomic analysis, enabling bacterial identification at the phylogroup level even within natural communities. We applied this approach to characterize *Escherichia coli* intraspecific diversity in human gut microbiomes using publicly available datasets. By estimating the relative abundance of eight *E. coli* phylogroups defined by their 18-mer markers in 558 fecal samples, we employed correlation analysis and machine learning techniques to assess the discriminatory power of their patterns under different host physiological states. We compared *E. coli* phylogroup distribution between microbiomes of healthy individuals and patients with chronic colorectal disorders, tracked their dynamic alterations following prolonged antibiotic exposure with subsequent probiotic-assisted recovery, and evaluated changes induced by a physiologically beneficial dietary intervention for obesity. As a result, binary classification methods distinguished most compared datasets, while correlation analysis revealed shifts in phylogroups connectivity networks and their bimodal interactions with *Prevotella*, suggesting intra- and inerspecific epistatic relationships. Finally, a technique based on nonlinear dimensionality reduction visualized all distinct clusters of all sets, providing a chance to test the feasibility of using intraspecific phylotyping of *E. coli* for medical applications.

Keywords: *E. coli* phylogroups; *k*-mer-based phylotyping; intraspecific phylotyping; colorectal adenoma; carcinoma; antibiotic treatment; probiotic therapy; obesity; Mediterranean diet; interspecific relationships

1. Introduction

Escherichia coli (*E. coli*) is a facultative anaerobic bacterium commonly found in the intestinal microbiota [1] and involved in maintaining gut microbiome homeostasis [2]. Advances in genomic sequencing and phylotyping have led to the progressive classification of *E. coli* into increasingly refined phylogenetic groups, initially categorized into four [3], then five [4], seven [5,6], eight [7–10], and most recently, twelve phylogroups [11]. For intraspecific classification, two main approaches have been widely used: multilocus sequence typing (MLST) [5,7] and the PCR-based method developed by Clermont and colleagues [3,4,6,8], which targets virulence-associated genes. These studies have demonstrated that pathogenic *E. coli* strains are not randomly distributed across phylogroups [12], leading to the general assumption that pathogenicity emerging in some phylotypes due to horizontal transfer of virulent genes [13] requires a specific genetic background for their expression [14]. Consequently, phylotyping has gained significant importance in epidemiological surveillance and pathogenicity studies.

In 2018 in silico PCR-based classification and the MASH clustering using marker *k*-mers [15] were combined in a new tool, ClermonTyping [6]. By analyzing a dataset of over 300 *E. coli* genomes, the authors observed only a few discrepancies between the two approaches: one relying on virulence genes detection and the other assessing genomic background. In our previous study [10], we compared MLST-based typing of 124 genomes with *k*-mer-based clustering and found that both methods generated topologically congruent trees, consistently identifying the eight established *E. coli* phylogroups (A, B1, B2, D, E, F, C and G). However, a subsequent MASH-based analysis of a much larger dataset, containing 10,667 *E. coli* and *Shigella* genomes, divided them into 12 clusters [11]. While phylogroups A, B1, F, C and G remained distinct, B2 was split into clusters B2-1 and B2-2; D diversified into subgroups D1, D2 and D3, while E separated into E1 and E2. Thus, despite the significantly expanded dataset, no entirely new phylogroups were discovered, suggesting that the eight major phylogroups represent the core genetic structure of *E. coli*.

While all phylogroups include strains prone to pathogenicity, epidemiological evi-dence indicates that ancient groups F and D harbor more pathogenic bacteria compared to groups A and B1 [9,16], whereas recently diverged lineages are frequently associated with severe infections [8,12,14]. A striking example is the 2011 German outbreak of *E. coli* infections, caused by the enteroaggregative hemorrhagic strain O104:H4 C227-11 [17–19], belonging to group B1. This strain exhibited exceptional virulence, attributed to the accu-mulation of horizontally acquired genetic determinants of pathogenicity [17,19–21]. These virulence factors included enhanced adhesion mediated by AAF plasmid-encoded fim-briae [19], potent cytotoxicity due to prophage-derived Shiga toxin (Stx-2) [17], and multi-drug resistance mainly conferred by extended-spectrum β-lactamases.

Large phylogroup B1, which comprises about 30% of known *E. coli* genomes, includes only a small portion (2.4%) of potentially virulent strains, such as those of serotypes O104:H4 and O121:H19 [11]. In contrast, group E, currently comprising just 10.9% of known genomes, harbors a disproportionally high fraction (10.2%) of highly pathogenic *E. coli* strains of serotype O157:H7 [11]. The remarkable virulence of O157:H7 stems from extensive horizontal gene transfer, having incorporated genetic material from 53 species [22] and over 460 prophages [23] (compared to just 29 in non-pathogenic *E. coli* K-12 [24]). Major chromosomal pathogenicity factors include Shiga toxin-encoding prophages (similar to those in O104:H4) [25] and 5 LEE (locus of enterocyte effacement) pathogenicity islands, encoding a type III secretion system and adhesins [26]. Over a thousand genes are absent from non-pathogenic *E. coli* K-12 [27,28], forming about 180 O157:H7-specific O-islands [29], and including 131 potentially virulent genes [28]. Given these distinctive genomic features, phylogroup E strains are readily identifiable through both PCR-based typing and *k*-mer-based clustering [6,10,11].

The pathogenicity of B2 strains is strongly associated with 54-kb *pks* genomic island, which encodes enzymes for colibactin biosynthesis [30,31]. However, the presence of such islands is not exclusive to pathogenic bacteria. Even the probiotic strain *E. coli* Nissle 1917, which is widely used to treat various intestinal disorders [32], has *pks* island in the genome [33,34]. Nevertheless, neither live bacteria of this strain nor the spent culture supernatant had a genotoxic effects [33], exemplifying epistatic suppression of virulence genes. While B2 is often considered as the most pathogenic phylogroup, its clinical prevalence may reflect enhanced ecological fitness rather than intrinsic virulence [35]. This is exemplified by CTX-M-15-producing B2 strains (serotypes O16:H5 and O25b:H4) that caused major outbreaks of infectious diseases in the early 2000s through their extended-spectrum β-lactamase-mediated antibiotic resistance [36,37]. Additional B2 virulence factors include: cytolethal distending toxins (CDTs) causing DNA damage in infected cells [30] and cytotoxic necrotizing factor (CNF) [38]. Both types of genes are frequently plasmid-encoded but many are located within genomic islands [30,38,39], contributing to the genomic background traced by virtual *k*-mer-based screening of chromosomes.

The effectiveness of PCR-based phylotyping indicates that phylogroups have a specific combination of marker genes. This was recently validated in a comprehensive study of 844 uropathogenic *E. coli* strains, which revealed certain associations between phylogroups and specific virulence factors, including genes of antimicrobial resistance, motility and biofilm formation [40]. According to the efficiency of PCR typing, 111 strains of phylogroup G, showed no isolates carrying genes for the adhesin Air, toxin Sat and the transcription factor EilA [9], of which Air and EilA are specific to groups D and F, while Sat is produced by groups D and B2 bacteria [41]. Phylogroup A virulence is most often associated with the adhesive fimbriae FimA and YfcV, as well as receptors for yersiniabactin (FyuA) and ferric aerobactin (IutA) [42]. Isolates from the less studied group C, encode the enzyme HlyF [43], which triggers eukaryotic autophagy due to toxin release via outer membrane vesicles [44]. In combination with Shiga toxin of *E. coli* O80:H2, this caused bacteremia in Europe [45]. Therefore, not all but many virulence genes are distributed across most phylogroups (Shiga toxins Stx-1/2 so far have not found only in group G genomes [46,47]). However, the mere presence of virulence genes does not guarantee pathogenicity. As proposed two decades ago, their functional integration requires a compatible genetic background [14].

Current understanding suggests that this genetic background emerges through complex interactions between cellular regulatory networks and horizontally acquired genes via epistatic relations. The enhanced recombination efficiency within phylogroups likely promotes the preferentially maintenance of beneficial and/or virulence genes among phylogenetically related strains. Yet, even phylogroups harboring multiple virulence determinants show unpredictable expansion patterns, implying the existence of more complex, higher-order epistatic interactions within microbial communities. To investigate this phenomenon at the level of *E. coli* phylogroup homeostasis, we employed *k*-mer-based profiling of natural microbiomes. The study specifically examined how endogenous and exogenous factors triggering adaptive responses in the human gut microbiome influence intraspecific equilibrium within *E. coli* populations.

2. Materials and Methods

2.1. Datasets Used for Intraspecific Taxonomic Analysis

The metagenomics data taken for this study were obtained from the European Nucleotide Archive (ENA), containing whole-genome sequencing (WGS) reads derived from human gut microbiomes (**Table 1**). The PRJEB7774 dataset, which includes samples from healthy volunteers and patients with colon adenoma and carcinoma [48], was chosen as representative of "stable" microbial communities that have adapted to distinct host physiological states. In contrast, the PRJEB28097 dataset comprised stool samples collected from healthy volunteers before and after the acute treatment with ciprofloxacin and metronidazole [49,50]. Previous study demonstrated that this combination of antimicrobials significantly reduced the abundance of all *L. paracasei* phylogroups [51], leading us to anticipate a similar response from other bacteria, including *E. coli*. The PRJEB33500 dataset contained fecal samples from overweight individuals, collected before and after adherence to Mediterranean diet, which modulated gut microbiome composition [52]. This dataset was particularly valuable as it enable direct comparison of intraspecific bacterial homeostasis caused by a relatively comfortable transition to healthier dietary patterns, representing a more physiologically gradual perturbation compared to antibiotic intervention or disease states.

Table 1. Datasets used for intraspecific taxonomic analysis.

Type of dataset	Donor types and number of samples		Bioproject
	Types	Number	
Inflammatory bowel diseases and cancer	Healthy individuals	65	PRJEB7774 [48]
	Patients with adenoma	49	
	Patients with carcinoma	46	
	Fifteen healthy donors before antibiotic treatment	92	

Antibiotic treatment with or without probiotic recovery	Twelve donors of the same group during antibiotic treatment	49	PRJEB28097 [49,50]
	Seven donors of the same group during self-recovery	42	
	Eight donors of the same goup during recovery with probiotics	43	
Overweight donors before and after diet	Samples from 43 overweight or obese individuals	86	PRJEB33500 [52]
	Samples from the same 43 persons after Mediterranean diet	86	

Fastq files underwent pre-processing steps, such as removal of adapter sequences and quality control using Trim Galore v.0.6.10 [53]. Reads shorter than 20 nucleotides and low-quality reads with Phred scores below 20 were excluded from the analysis.

2.2. Genome Collections

Two genome sets were used in this study. Set 1 contained 124 genomes previously aplied for phylotyping in our earlier study [10]. It consisted of 59 complete genomes whose phylogroups were previously identified in the original publications and 65 randomly selected chromosomes not assigned to phylogroups (listed in Supplementary Table S1 in [10]). Set 2 was assembled to validate phylotyping consistency. It was composed from 6,300 completed bacterial genomes deposited to NCBI GenBank. After removing identical sequences and the 124 genomes from Set 1, the remaining genomes were randomly split into 10 subsets. Each subset was then back-supplemented with the 124 genomes from Set 1. Using the sets of non-redundant 18-mers from each genome and *E. albertii* KF1 chromosome as an outgroup, we constructed 10 draft phylogenetic trees via the neighbor-joining method based on pairwise Sørensen's distance matrices [54]. Clusters corresponding to all eight phylogroups were identified in each subset, and genomes distinct from Set 1 and each other were selected and combined. After iterative phylogenetic refinement to minimize duplicates and overlaps with Set 1, 154 genomes were finalized for Set 2 (listed in **Supplementary Table 1**).

2.3. Barcoding of the E. coli Phylogroups

Barcoding and intraspecific taxonomic analysis were performed as previously described [51]. In brief, we used a local copy of the NCBI GenBank database (February 28, 2023 release), containing 45,566 fully assembled bacterial genomes, 51,914 plasmid sequences and the human reference genome (GRCh38). Hash indices were generated for all non-redundant 18-mers across these sequences. For *E. coli* genomes (both sets), we excluded 18-mers present in any bacterial genome or plasmid in the local copy of NCBI GenBank database (including *E. coli* plasmids), except for 3,616 *E. coli* genomes and 233 genomes of closely related *Shigella*. 18-mers from *Homo sapiens* chromosomes were also filtered out. Unique 18-mer barcodes for each phylogroup were obtained by stepwise removal of shared 18-mers across phylogroups. The distribution of model genomes among the eight phylogroups and their barcode sizes are summarized in **Table 2**.

Table 2. Number of genomes in phylogroups and size of sets with unique 18-mers (barcodes).

Phylogroups	**Number of genomes**		**Number of 18-mers**	
	Set 1	**Set 2**	**Set 1**	**Set 2**
A	17	21	415335	354997
B1	25	25	710784	524927
B2	23	29	1014716	783899
C	14	17	242272	170224
D	11	15	673338	524936
E	13	19	680604	802163
F	11	15	445835	313383
G	10	13	254624	171176

The 18-mer barcode sets for all phylogroups are available in our GitHub repository (https://github.com/marsfro/ecoli_counter). Hashing and barcoding were conducted using the 64-bit UniSeq software on a high-performance server (configuration: 2 Xeon Gold 5218, 64 GB RAM).

2.4. Alternative Intraspecific Phylotyping of E. coli

Alternative UniSeq-based phylotyping of Set 1 genomes was performed as described above, but with a modified *k*-mer selection approach. Instead of using unique marker *k*-mers absent in other bacterial genomes (one of the two UniSeq program options [10]), we used all non-redundant 18-mers present in each genome for clustering. This approach is functionally similar to MASH [15], differing from phylotyping based on unique barcodes [10, 51]. Next, pairwise Sørensen similarity indices [54] were computed for the 124 sets of 18-mers. A phylogenetic tree was then reconstructed from the resulting distance matrix using the neighbor-joining method [55] in MEGA X [56].

2.5. Analysis of Human Metagenomes Based on Phylogroup-Specific Taxonomy

Phylogroup-specific *k*-mers were used to assess the relative abundance of each *E. coli* phylogroup in target metagenomes. Reads containing at least one phylogroup-specific *k*-mer were counted and assigned to their respective phylogroups. To enable comparison across phylogroups with differing barcode size (**Table 2**), read counts were normalized using the average barcode size (554,688 *k*-mers). A custom Python script for identifying and counting phylogroup-specific reads in WGS metagenomes is available at https://github.com/marsfro/ecoli_counter/kmers_ecoli_counter.py. The resulting relative abundances are provided in **Supplementary Table S2**. To evaluate potential links between *E. coli* interspecific diversity and interspecific relationships, we estimated the abundance of three dominant human gut genera (*Bacteroides*, *Prevotella*, and *Ruminococcus*) characteristic of different enterotypes. Following standard pre-processing, taxonomic profiling was performed using Centrifuge v.1.0.4 software [57].

2.6. Machine Learning Methods for Binary Clustering

The normalized percentages of the eight *E. coli* phylogroups from control and pathological/treatment-derived samples served as features (variable Y) in machine learning models. Clustering was performed in Python (v. 3.10.9) using the Scikit-learn libraries (sklearn, v. 1.3.0) [58]. Datasets of the PRJEB77747774 [48] and PRJEB33500 [52] with large number of independent variables were partitioned into training (80%) and test sets (20%) via train_test_split. Three binary classification models were used: logistic regression (LR), random forests (RF), and gradient boosting (GB). To optimize the hyperparameters of these models, we performed a grid search using the GridSearchCV function from sklearn. For LR, the regularization parameter C was tuned with values of 0.1, 1, and 10. The RF classifier was tuned with the number of estimators (n_estimators) set to 50, 100, and 200, and the maximum tree depth (max_depth) set to 3, 4, and 5. For GB classifier, n_estimators were set to 50, 100 and 200, and learning_rate to 0.1, 0.05 or 0.01. GridSearchCV systematically evaluated all parameter combinations for each model. The best parameters were selected based on 5-fold cross-validation and the area under the Receiver Operating Characteristic curve (ROC-AUC) as the evaluation metric. The optimal values of hyperparameters are given in figure captions.

2.7. Datasets Visualization

Virtual visualization of clusters was performed using the Uniform Manifold Approximation and Projection (UMAP) technique for dimension reduction [59]. The relative abundances of the eight phylogroups were converted to a standard input format using the StandardScaler option from the sklearn.preprocessing module in Python. The fit_transform option was then applied to center the features around their mean and scale them to unit variance. Next, the UMAP technique was used to reduce dimensionality and visualize the datasets in a space with lower dimensionality. Finally, the image was created using UMAP CLASS option from the UMAP library (version 0.5.2) with the hyperparameters: n_neighbors = 35 (number of neighboring data points to consider), min_dist = 0.1 (the effective minimum distance between embedded points).

2.8. Statistical Analysis

For statistical analysis, numerical values corresponding to duplicated probes in the PRJEB33500 project and multiple samples from 7-15 donors in PRJEB28097 were averaged. The normality of data distribution across datasets was assessed using One-sample *t*-test in SigmaPlot (v.11). As most datasets did not meet normality assumption, group comparisons were performed using the nonparametric Mann-Whitney-Wilcoxon test via SigmaPlot Compare Two Groups option based on median differences [60]. Due to non-normality, variability was also assessed using mean absolute deviations (MAD) instead of two-way ANOVA. Sample-specific deviations from the mean were compared between dataset using the median based test. Datasets exhibiting a statistically significant increase in variance ($p < 0.05$) and a greater interquartile range (IQR) than baseline (see Figure Legends for cutoffs) were classified as significantly diversified. We used the Pearson correlation coefficient (R) to evaluate intraspecific associations with *Bacteroides*, *Prevotella* or *Ruminococcus* in all datasets. The statistical significance of R was estimated using a VassarStats online tool (http://vassarstats.net/tabs_r.html) [61]. To check the robustness of statistical assessments, we performed jackknife resampling by iteratively excluding single sample at a time from the original dataset [62]. This analysis identified five outliers of 568 samples that disproportionately influenced the outcomes; these were excluded from subsequent analysis (see **Supplementary Tables S2** and **S3**). For machine learning-derived data, statistical significance was assessed using Student's *t*-test in Python.

3. Results

In our previous study [10], we performed *E. coli* phylotyping using *k*-mer-based approach and MLST. For MLST, a combined set of 27 marker genes proposed in [4,63–65] was implemented. The individually aligned sequences of these genes were concatenated and a phylogenetic tree was constructed using the IQ-TREE [66]. For the *k*-mer-based approach, phylogenetic trees were inferred from a pairwise distance matrix of *Escherichia coli*/*Shigella*-specific 18- and 22-mers (124 genomes total), identified using the UniSeq algorithm [10]. Both methods produced topologically identical trees. Their clustering precisely matched ClermonTyping [6], though 14 strains showed discrepancies compared to the MASH-based clustering used in [11]. These included 12 strains classified as C, which in [11] were assigned to group B1, and 2 strains, classified in our study as B1 bacteria were assigned to group C. Since the MASH algorithm, originally developed to assess intraspecific polymorphism in bacteria, estimates genomic distances based on “mutation rates” using representative 21-mers [15], the observed discrepancies between our classification and MASH-based typing might be due to difference in the marker *k*-mer sets, used for clustering. Given that the accuracy of taxonomic analysis critically depends on the quality of *k*-mer barcodes, we re-evaluated the phylotyping of our 124 genomes using the option of UniSeq software analogous to MASH.

3.1. Validation of Intraspecific E. coli Phylotyping Using MASH-Like Option of UniSeq Pipeline

Unlike our previous analysis, which used k-mers unique for E. coli/Shigella, this implementation considered all non-redundant 18-mers present in each genome. Then, pairwise Sørensen similarity indices [54] were computed for all 124 18-mer sets and the phylogenetic tree was constructed from the pairwise distance matrix using neighbor-joining method [55]. As a result, the constructed tree (**Figure 1**) showed identical topology to our earlier phylogeny based on unique 18-mers (**Figure 3** in [10]).

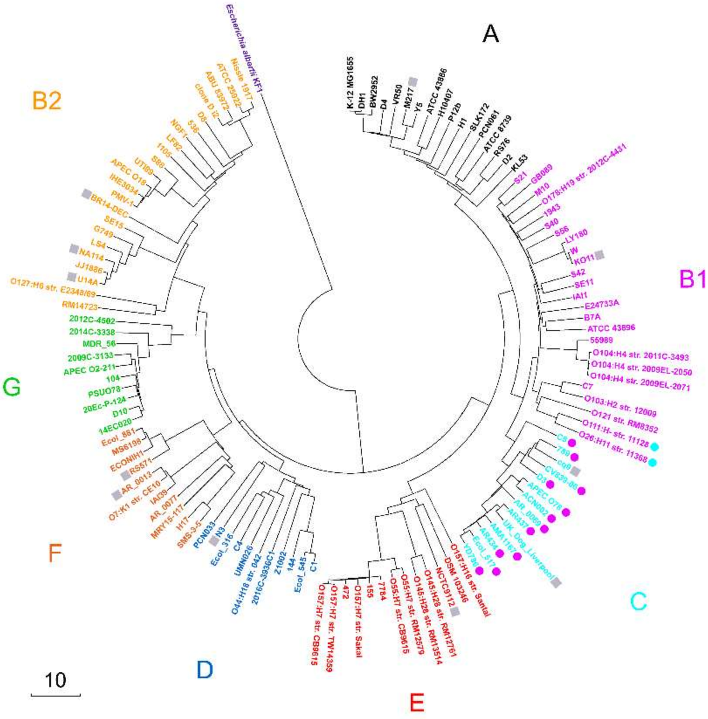


Figure 1. Phylogenetic tree of 124 *E. coli* strains constructed using the neighbor-joining method [55] in MEGA X [56] based on a pairwise distance matrix derived from sets of represetative 18-mers. The scale bar shows the Sørensen distance (percentage). The eight *E. coli* phylogroups are color-coded. A set of *Escherichia albertii* KF1 representative 18-mers was used as an outgroup. Strains with discordant phylotyping (B1 vs. C) relative to [11] are highlighted with colored circles, while strains not analyzed in [11] are denoted as gray squares.

Thus, five distinct approaches, including MLST analysis using 27 marker genes [10], in silico PCR and MASH via ClermonTyping resource [6], and two UniSeq-based techniques ([10] and **Figure 1**) consistently yielded congruent phylotyping results. Thus, the same 124 genomes as in our previous study [10] were used for phylotyping with their distribution across phylogroups detailed in **Table 2** (Set 1). Strains O111:H- str. 11128 and O26:H11 str. 11368 were assigned to phylogroup B1, while strains C8, 789, cq9, CV839-06, D3, APEC O78, ACN002, AR_0069, AR437, UK_dog_Liverpool, AM1167, AR434, Ecol_517 and YD786 were considered as members of phylogroup C.

3.2. Assessment of Barcoding Specificity

While both UniSeq- and MASH-based typing methods are equally effective for phylogenetic analysis, only UniSeq, which relies on unique marker *k*-mers absent in other bacterial genomes, enables intraspecific taxonomic analysis of natural microbial communities. To estimate the abundance of bacterial groups in metagenomes, the algorithm quantifies reads containing their specific 18-mer markers. The accuracy of these estimates depends on both the number and specificity of the 18-mer markers used. Increasing the number of known genomes within each phylogroup improves barcode specificity. However, expanding the reference database for filtering reduces the cumulative barcode size. Since natural microbiomes contain many bacteria not represented in reference database, uncharacterized microorganisms will inevitably generate reads misassigned to known groups. To account for this bias, two complementary approaches were applied.

To assess the species specificity of marker k-mers, we estimated their average number in the 124 genomes used for barcoding. These values were then compared with the *k*-mer content of other *Escherichia* species (*E. albertii*, *E. fergusonii*, and *E. marmotae*) (**Figure 2a**). Genome selection was based on two criteria: (1) absence from the reference dataset of their close homologs to prevent 18-mers filtering during barcode construction, and (2) mutual evolutionary dissimilarity as determined by the phylogenetic tree topology (constructed similarly to **Figure 1**).

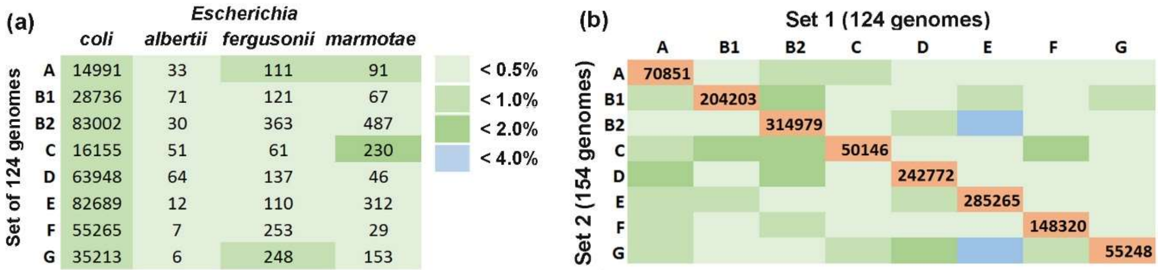


Figure 2. Intraspecific barcoding of *E. coli* using 124 genomes with verified phylogroup identity provided marker 18-mer sets, demonstrated **(a)** limited cross-species similarity and **(b)** acceptable for taxonomic analysis specificity. (**a**) Average percentage of marker 18-mers detected in 124 *E. coli* genomes (**Figure 1** and **Table 1**) compared to chromosomes of *E. albertii* (CP130156.1, CP141901.1, CP157789.1), *E. fergusonii* (CP099328.1, CP125351.1, CP137855.1) and *E. marmotae* (CP099344.1, CP099351.1, CP173213.1). **(b)** Overlap of barcodes derived from 124- and 154-genome sets, illustrating shared markers within the same phylogroup (magenta) and unintended cross-phylogroup overlaps (shades of green and blue). The percentage of matches was assessed relative to the size of the smaller barcode in the pair. A unified color scheme denotes overlaps in both panels.

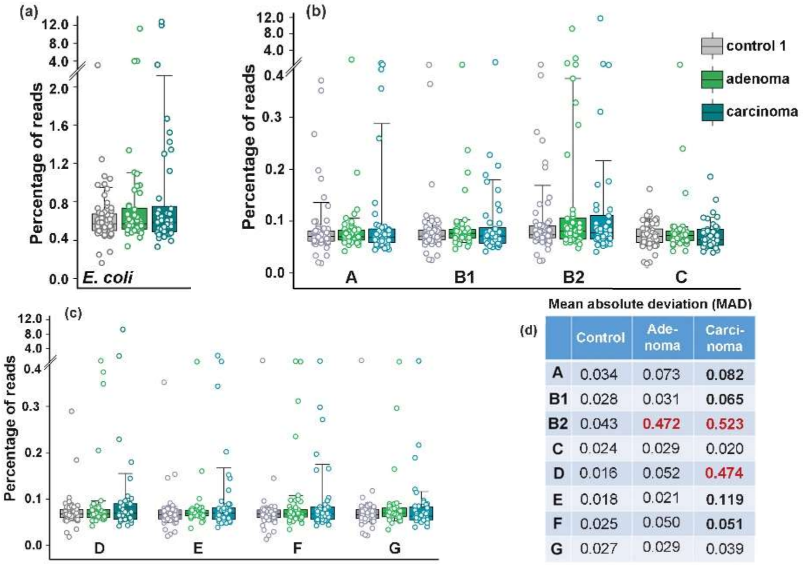
Three genomes of each species appeared to be sufficiently representative and the highest overlap was observed between *E. coli* group C barcodes and 18-mers of *E. marmotae* RHB35-E2-C08 (1.4%). The average overlap across other combinations was significantly lower (0.34 ± 0.07%). These estimates suggest that the presence of uncharacterized *Escherichia* is unlikely to substantially affect the accuracy of *E. coli* phylogroup quantification in natural microbiomes.

For the second trial, we constructed another training set containing 154 genomes with phylogroup identities assigned via ClermonTyping and UniSeq approach (Set 2). These genomes were selected from NCBI GenBank using the same criteria as stated above with additional requirement of no close homologs present in Set 1. Several iterations of phylogenetic analysis were required to choose suitable genomes from 6300 completed genomes of *E. coli* deposited in the GenBank database to create a phylogenetically balanced subset. Set 2 (**Table 2**) was independently barcoded using the same reference database as for Set 1. Phylogenetic reconstruction confirmed clear separation of genomes into eight phylogroups with no discordance against expected identities (**Supplementary Figure 1**), although all groups except E exhibited a reduction in the number of marker 18-mers (**Table 2**). This is at least partly due to the more stringent interphylogroup filtering in Set 2, which was difficult to avoid during genome selection.

We next quantified the overlap between the two sets for each phylogroup and revealed that 20-52% of 18-mers from the smaller set (usually Set 2) were shared with Set 1 barcodes. Overlaps between different groups ranged from 0.049% to 3.950% (**Figure 2b**) with the highest cross-phylogroup similarity observed between group E of Set 1 with groups B1 (2.89%) and G (3.95%) from Set 2. The average overlap across other phylogroup combinations was substantially lower (0.55 ± 0.08%), only marginally exceeding typical interspecific overlap levels (0.34 ± 0.07%) (**Figures 2b and 2a**). Therefore, we proceeded with Set 1 barcodes (experimentally verified by multiple methods) for all subsequent intraspecific taxonomic analyses.

3.3. Colorectal Adenoma and Carcinoma Had Different Impact on the Distribution of E. coli Phylogroups in the Intestinal Microflora

Consistent with previous reports [35], we anticipated that at least carcinoma would either promote greater persistence of phylogroup B2 in the gut or otherwise affect E. coli phylogroup homeostasis. The average abundance of this phylogroup was indeed higher in the samples obtained from patients with adenoma and carcinoma (0.349% and 0.362, respectively) than in the control group (0.097%) (**Figure 3**).



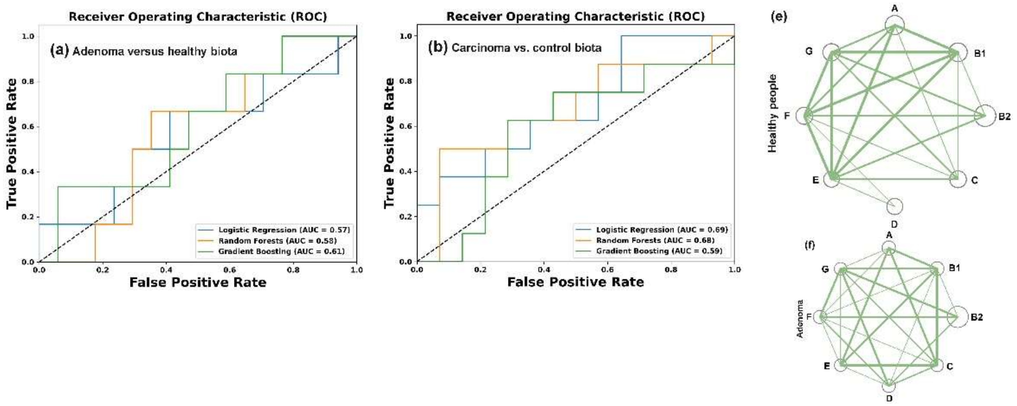
Mean absolute deviation (MAD)

	Control	Ade-noma	Carci-noma
A	0.034	0.073	**0.082**
B1	0.028	0.031	**0.065**
B2	0.043	**0.472**	**0.523**
C	0.024	0.029	0.020
D	0.016	0.052	**0.474**
E	0.018	0.021	**0.119**
F	0.025	0.050	**0.051**
G	0.027	0.029	0.039

Figure 3. The diversity of E. coli (**a**) and its phylogroups (**b**,**c**) increased in the microbiomes associated with colon diseases (**d**). The box plots show the percentages of reads with phylogroup-specific 18-mers (**b**,**c**) or E. coli/Shigella-specific 18-mers (**a**). Significance of MAD alterations was assessed using Mann-Whitney-Wilcoxon test [60] and Interquartile Interval Range (IQR). MADs with statistically significant changes (p < 0.001) and IQR increased for at least 1.5 times are shown in bold. An increase in IQR by more than 2 times is indicated in red.

Only group D bacteria known to produce Air adhesin and Sat toxin [41] increased their presence to almost the same level (from 0.173% to 0,319%) in microbiomes associated with carcinoma, but all changes, along with those in total *E. coli* abundance were not statistically significant (**Figure 3a–c**). Nevertheless, we observed a priori unexpected increase in the variability of *E. coli* phylogroups abundance across different biological samples. In particular, adenoma increased the mean absolute deviation (MAD) in the presence of B2 bacteria, while carcinoma promoted significant variability for six phylogroups, with the greatest impact on B2 and D group bacteria (**Figure 3d**).

Assuming that the difference in *E. coli* homeostasis between the usually harmless adenoma and the dangerous carcinoma reflects variable ways in which *E. coli* adapts to the specific environment created in the gut, we used machine-learning algorithms to assess whether the frequency profile of phylogroups can distinguish between healthy and pathological state. Normalized percentages of the eight phylogroups in control and pathologi-cal/treatment conditions were used as features defining the target variable Y in machine learning models for each type of samples. Binary classification was performed using Lo-gistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) models (**Figure 4a–c**). The best hyperparameters for each model were obtained as described in Materials and Methods section. Their performance was evaluated using the Area Under the Curve (AUC) as a distance-based metric for the Receiver Operating Characteristics (ROC) curves.



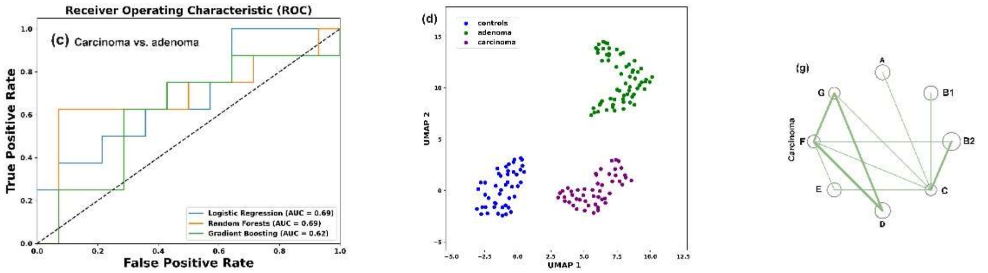


Figure 4. While binary clustering failed to differentiate *E. coli* phylogroup composition between control samples and adenoma-associated microbiomes (**a**), RF and LR models (**b,c**), UMAP dimensionality reduction algorithm [59] (**d**) and intraspecific correlations (**e-g**) effectively detected changes in *E. coli* homeostasis. The C parameters used for binary classification with LR were 0.1 (**a**) and 1.0 (**b,c**). The parameters for clustering with RF were: max_depth = 5, n_estimator 200 (**a**,**b**) and 50 (**c**). The best learning_rate/n_estimator combinations for GB were: 0.1/100 (**a**), 0.05/200 (**b**) and 0.1/200 (**c**). UMAP clustering (**d**) was performed using parameters n_neighbors = 35, min_dist = 0.7. (**e–g**) Network visualization of intraspecific correlations among *E. coli* phylogroups based on Pearson's correlation coefficient (R). Node size reflects mean phylogroup abundance in samples. Statistically significant correlations are indicated by lines with thickness reflecting their strength: $p < 0.00001$ (thick), $p < 0.001$ (medium), $p < 0.05$ (thin).

The resulting ROC-AUC scores of 0.57 (LR), 0.58 (RF) and 0.61 (GB) were close to 0.5, indicating a random distribution of the control samples and samples from patients with colorectal adenoma (**Figure 4a**). However, when the control set was compared with the carcinoma patient samples (**Figure 4b**), the AUC scores were higher and, at least for the LR model, separated the two sets with 73% accuracy. The discriminatory power of this model also distinguished between E. coli populations in the microbiomes of people with colorectal carcinoma and adenoma with approximately the same accuracy (**Figure 4c**). An algorithm based on nonlinear dimensionality reduction (Uniform Manifold Approxima-tion and Projection, UMAP [59]) clearly separated all three sets of biological samples (**Fig-ure 4d**), reflecting distinct phylogroup compositions.

Based on this observation, we next evaluated intraspecific correlations in the abundance of E. coli phylogroups within the microbiomes (**Supplementary Table S3**). In the control set (**Figure 4e**) and in the microbiota associated with adenoma (**Figure 4f**) all phylogroups exhibited positive correlation with each other. Although not all correlations were statistically significant, this likely suggests that bacteria from different groups do not specialize in entirely independent functions within E. coli populations. The number of correlated groups in samples obtained from the microbiomes of patients with carcinoma, on the contrary, was much smaller, and B2 group showed non relationships with A, B1, D and E phylogroups (**Figure 4g**). Thus, while the abundance of E. coli phylogroups did not differ significantly among microbiomes adapted to different host physiological states (**Figure 3b,c**), their interaction patterns appeared sensitive to the chronic alterations induced by carcinoma.

3.4. Bimodal Response of E. coli to Antibiotics and Recovery with Probiotics

Despite being inherently susceptible to most clinical antimicrobials, E. coli has a remarkable capacity to acquire resistance genes through horizontal gene transfer [40,67]. They are distributed across all phylogroups and may function collectively to counteract antimicrobial drags. Using data from the PRJEB28097 project [49,50] we aimed to characterize the individual responses of E. coli groups to ciprofloxacin, resistance to which was observed among isolates of all phylotypes [43,68,69]. The selected dataset contained stool samples from healthy volunteers receiving ciprofloxacin (500 mkg, twice daily) and metronidazole (500 mkg, three times daily) [49,50]. As metronidazole exhibits activity against E. coli only in the presence of other susceptible bacteria [70], this antibiotic combination enabled detection of community-dependent response. While the abundance of E. coli in these microbiomes was comparable to that observed in the PRJEB7774 project (**Figures 2a**, and **5a**), this set allowed for longitudinal tracking of antibiotic-induced microbiome alterations (**Figure 5b**).

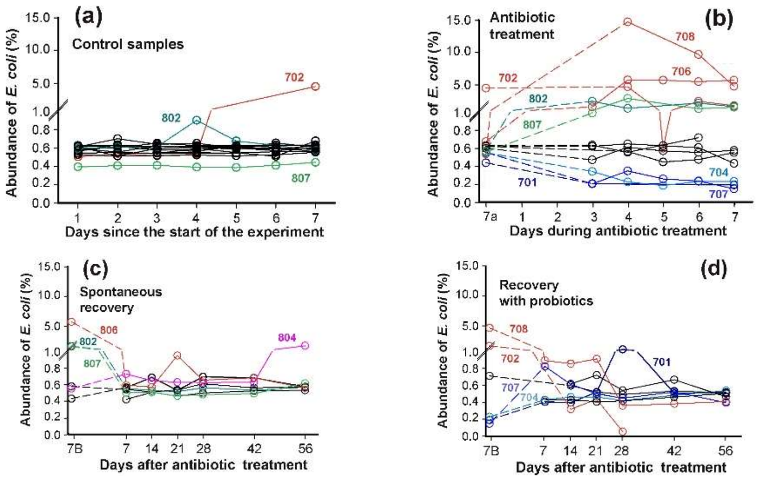


Figure 5. Dynamics of E. coli abundance (sum of all phylogroups) in healthy human microbiomes. (**a**): Stable baseline levels during 7-day pretreatment period. (**b**): Differential response patterns following antibiotic administration. (**c**,**d**): Recovery patterns in the absence (**c**) or presence (**d**) of probiotic supplementation. Numerals indicate sample IDs (**Supplementary Table S2**). Left time points in **b–d** represent endpoints of preceding stages. Whenever possible, dashed lines connect longitudinal measurements from the same microbiome. Colored samples are discussed in the text.

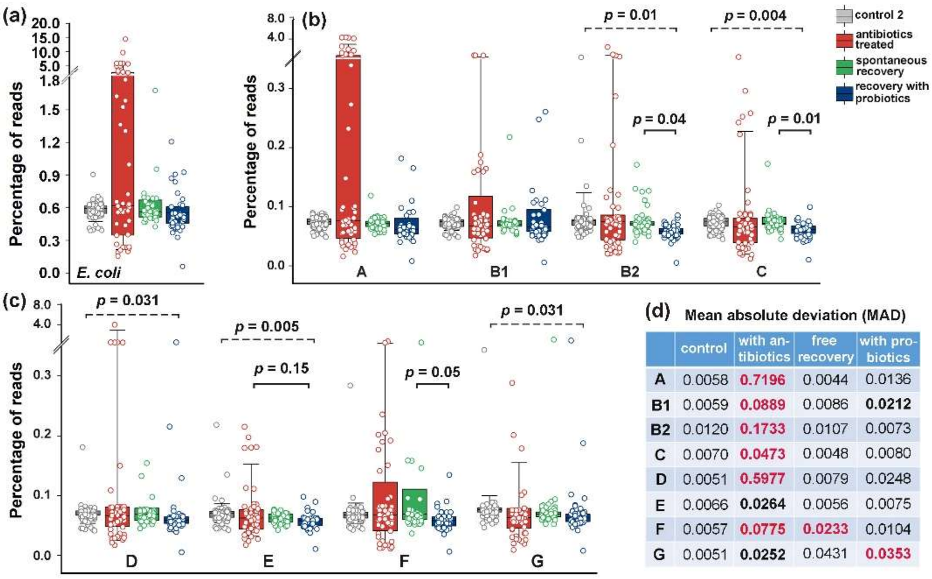
Prior to antibiotic administration, *E. coli* abundance was stable in 13 out of 15 microbiomes (**Figure 5a**). Two exceptions were observed: sample 802 showed only a transient increase, while in sample 702 *E. coli* exhibited significant spontaneous expansion beginning on day 5 of the pre-testing phase. This elevated level persisted until day 5 of antibiotic treatment, when a temporarily decline occurred (**Figure 5b**). As expected, antibiotics exposure caused divergent effects on *E. coli* abundance across individuals. Among the twelve samples analyzed dynamically, four (703, 802-804) demonstrated complete antibiotic insensitivity. Conversely, five microbiomes showed significant *E. coli* proliferation, including sample 807, which had minimal baseline abundance (**Figure 5a**). Samples 701, 704 and 707 displayed the opposite pattern, with notable *E. coli* reduction.

Following seven days of spontaneous recovery, most microbiomes returned to near-baseline *E. coli* levels. Only sample 804, previously identified as antibiotic-insensitive, displayed unexpected *E. coli* expansion at the end of experiment (**Figure 5c**). Spontaneous recovery of *E. coli* in microbiomes took less than 7 days (**Figure 5c**). However, supplementation with an 11-strain probiotic mixture [49,50], containing four strains of the genus *Bifidobacterium*, this period was longer (**Figure 5d**). This lag is not surprising, as administration of only *B. longum* and *L. paracasei* to laboratory rats significantly suppressed *Escherichia* in at least one enterotype of their gut microbiomes [71]. Thus, the abundance of *E. coli* in fecal microbiomes is sensitive to the presence of both antibiotics and probiotics.

3.5. The Response of E. coli to Antibiotics Was Not Uniform Among Phylogroups

The lack of strong time dependence in samples from the same donor across particular experimental stages (days 1-7 pre-treatment, days 3-7 during treatment and days 7-56 post-treatment) (**Figure 5**) allowed us to use the mean phylogroup abundance values in each microbiome as independent variables for statistical assessment (**Figure 6**).

doi:10.20944/preprints202505.1278.v1



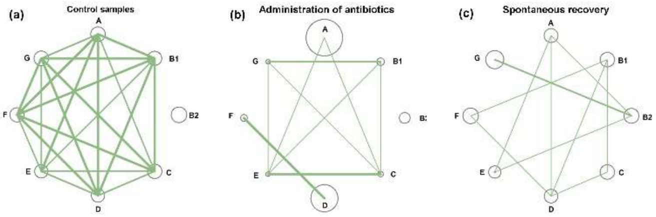
	control	with antibiotics	free recovery	with probiotics
A	0.0058	**0.7196**	0.0044	0.0136
B1	0.0059	**0.0889**	0.0086	**0.0212**
B2	0.0120	**0.1733**	0.0107	0.0073
C	0.0070	**0.0473**	0.0048	0.0080
D	0.0051	**0.5977**	0.0079	0.0248
E	0.0066	**0.0264**	0.0056	0.0075
F	0.0057	**0.0775**	**0.0233**	0.0104
G	0.0051	**0.0252**	0.0431	**0.0353**

Figure 6. Phylogroup-dependent response of *E. coli* to antibiotics and probiotic supplementation. Scatter plots show the percentages of reads containing (**a**) *E. coli*-specific or (**b**,**c**) phylogroup-specific 18-mers. (**b**–**c**) Symbols show percentage of phylogroups in microbiomes, while box plots display their mean values averaged across samples from individual donors. (**d**) Statistical significance of MAD alterations was assessed using Mann-Whitney-Wilcoxon test [60] and Interquartile Interval Range (IQR). MADs with IQR increase of more than 3-fold and statistically significant changes ($p < 0.05$ vs. controls) are bolded. Changes with $p \leq 0.001$ are highlighted in red.

The bidirectional response of *E. coli* to antibiotics (**Figures 5b and 6a**) significantly increased the MAD of all phylogroups without statistically significant changes of their abundance (**Figure 6b–d**). The main contribution to the variability of *E. coli* was made by group A (**Figure 6a,b**), the variance of which was less pronounced in microbiomes associated with colon diseases (**Figure 3d**). Consistent with the dynamic plots (**Figure 5c**) one week of spontaneous recovery following antibiotic treatment significantly reduced *E. coli* variance and abundance to near-baseline levels (**Figure 6d**). However, this effect was not similar for all groups. Phylogroup F, in particular, retained significantly higher variability compared to controls (**Figure 6c,d**). The influence of probiotics tested on independent group of volunteers was also apparent: most groups exhibited reduced persistence in microbiomes, though group G remained variable throughout the extended recovery period (**Figure 6c,d**). Therefore, the response of *E. coli* to antibiotics and probiotics was not uniform for all of its phylogroups.

3.6. Post-Treatment Recovery Partially Restored the Intra- and Interspecies Balance Disrupted by Antibiotics, But Not the Original Correlation Between Phylogroups and UMAP Cluster

In control microbiomes, all *E. coli* phylogroups were present at nearly equal proportions, and their percentages were highly correlated (**Figure 7a**). Only group B2 showed no statistically significant intraspecific associations (**Supplementary Table S3**). By inducing divergent changes in all groups, antibiotics disrupted this equilibrium. Consequently, more than half of the intraspecific correlations were lost (**Figure 7b**).



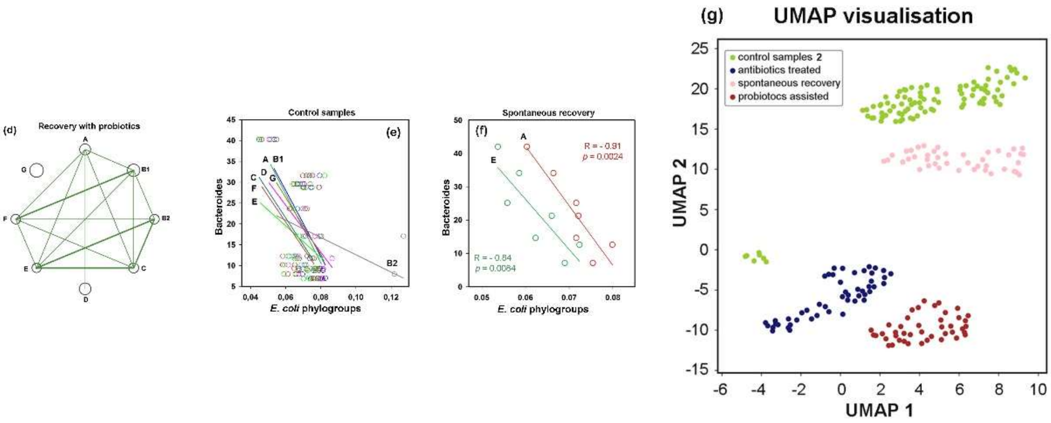


Figure 7. Networks of intraspecific *E. coli* correlations and their connections with other taxa **of** (**a**–**d**) Network visualization of intraspecific correlations among *E. coli* phylogroups, which were disrupted by antibiotics (**b**) and partly restored in a probiotic-dependent manner (**c**,**d**). Node size represents mean phylogroup abundance across samples. Lines depict statistically significant correlations, with thickness reflecting their strength: $p \leq 0.00001$ (thick), $p < 0.001$ (medium), $p < 0.05$ (thin). (**e**,**f**) A trend toward negative interspecific correlations with *Bacteroides* in control samples (**e**) became statistically significant for two phylogroups during spontaneous recovery (**f**). (**g**) UMAP clustering (parameters: n_neighbors = 35, min_dist = 1.0) visualize differences in *E. coli* homeostasis among four gut microbiota states.

Within eight weeks after antibiotic exposure, the relative abundance of all phylogroups had nearly returned to baseline levels (circles in **Figure 7a,c**). This recovery was particularly evident for groups A and D, which had expanded during antibiotic treatment, as well as for the antibiotic-suppressed groups B1, C, E, F and G, which rebounded close to their original abundances. However, the number of intraspecific links remained reduced compared to control samples, and the network structure differed from both baseline and antibiotic-perturbed states. For instance, group B2 bacteria, which showed no significant correlations with other groups in control and antibiotic-treated samples, developed significant connections with groups A, E and G (**Figure 7c**). Probiotic-assisted restoration, which suppressed the expansion of groups A and D established statistically significant associations between B2 bacteria and groups B1, C, E and F, but disrupted all intraspecific links of group G, which "survived" antibiotic exposure.

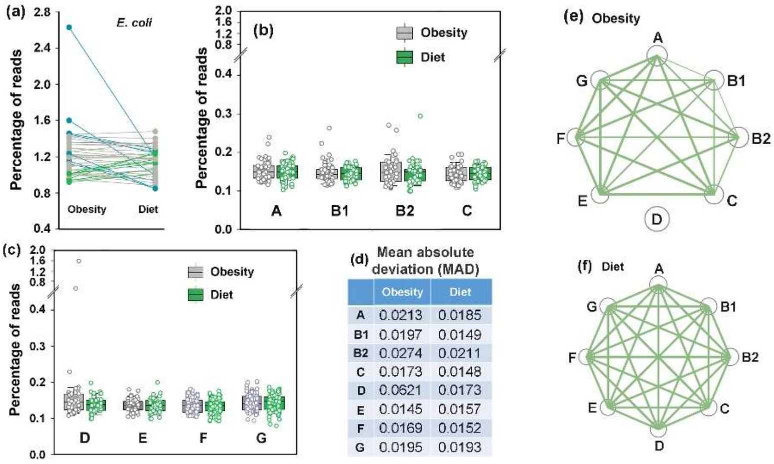
Probiotic-induced rearrangements in *E. coli* intraspecific networks suggested that phylogroup homeostasis is dependent on interspecific relationships. To investigate this, we estimated the abundance of dominant enterotype-associated genera (*Bacteroides*, *Prevotella* and *Ruminococcus*) [1] in all samples using the metagenomic classifier Centrifuge [57]. In control samples, we observed consistent negative correlations between all *E. coli* phylogroups and *Bacteroides*, with the strongest association for phylogroup D (R = –0.5) and weakest for group B1 (R = –0.21) (exemplified in **Figure 7e**). *Prevotella* showed positive correlations with all *E. coli* phylogroups (statistically significant associations for groups A, B1 and D in the range 0.52 < R < 0.61). *Ruminococcus*, on the contrary, exhibited weak negative correlations with all groups, except B2. Antibiotics preserved the negative correlations between *Bacteroides* and three *E. coli* phylogroups but inverted the relationship with group B2, resulting in a stable positive correlation with R = 0.40 ± 0.04 in jackknife analysis. During spontaneous recovery, *Bacteroides* re-established their negative link with B2 and strengthened negative associations with groups A and E (**Figure 7f**). In the presence of probiotics, only group G maintained its original connections with *Bacteroides*, while all other groups showed an inverse trend (**Supplementary Table 3**).

Antibiotics converted the positive correlations between *E. coli* groups and *Prevotella* into weak negative connections, while simultaneously strengthening positive links with *Ruminococcus*. During spontaneous recovery, *Prevotella* re-established positive links with groups A, B2, E and G; whereas probiotics restored positive associations only with groups B2 and C. Interactions with *Ruminococcus* spontaneously reverted to negative correlations for all *E. coli* phylogroups. However, probiotic intervention not only re-established their positive links with group B2 but also created new positive association with group C.

Therefore, although interspecific connections were weaker than intraspecific ones**,** our analysis indicated that phylogroups tend to explore similar connections with dominant taxa (**Supplementary Table 3**). However, we also found several examples, revealing individual interspecific links. These distinctive connections may play a crucial role in shaping the networks of intraspecific connections (**Figure 7a-d**), which difference was confirmed by UMAP clustering (**Figure 7g**). It has become clear that therapeutic antibiotic doses induce significant adaptive changes, leading to intraspecific reorganization and affecting interspecific interactions of *E. coli* phylogroups.

3.7. Idealizing E. coli Intraspecific Balance, the Mediterranean Diet Intensified Its Negative Link to Bacteroides and Unlocks Bidirectional Connections with Prevotella

The PRJEB33500 project dataset [52] comprises duplicate fecal samples from 43 overweight/obese volunteers collected before and after an 8-week of Mediterranean diet (MD) restriction. Both intra- and interspecific *E. coli* relationships were assessed using the mean values of duplicate samples (**Figure 8**). The average abundance of *E. coli* in the microbiomes of overweight individuals (**Figure 8a**) was higher than in healthy donors from the other two analyzed projects (**Figures 3a** and **6a**). Although the adaptive response of microbiomes to dietary restriction was also bidirectional (**Figure 8a–c**), the divergence was much less pronounced than that observed under antibiotic exposure. Only six microbiomes exhibited a 10 – 30% increase in *E. coli* persistence, while in the biota of seven donors, its abundance decreased by 10 – 55%, including samples, whose elevated *E. coli* levels were primarily driven by phylogroup D bacteria (two outliers in **Figure 8c**). Notably that dietary restriction led to a reduction in mean absolute deviation (**Figure 8d**), possibly reflecting the stabilizing effect of a balanced diet. While no significant changes were observed in the average abundance of *E. coli* phylogroups (**Figure 8b,c**), intraspecific balance improved markedly. The equilibration abundance of all phylogroups (nodes in Figure 8f) and the reduced persistence of phylogroup D bacteria in microbiomes with their initial overabundance resulted in strong correlations among all groups (**Figure 8f**). This observation suggest that the Mediterranean diet may be used to restore *E. coli* intraspecific homeostasis.



(d) Mean absolute deviation (MAD)

	Obesity	Diet
A	0.0213	0.0185
B1	0.0197	0.0149
B2	0.0274	0.0211
C	0.0173	0.0148
D	0.0621	0.0173
E	0.0145	0.0157
F	0.0169	0.0152
G	0.0195	0.0193

Figure 8. The Mediterranean diet did not significantly alter the overall abundance of *E. coli* or its phylogroups in the gut microbiomes, but improved intraspecific balance without induced diversification. (**a–d**) Box plots display the percentages of reads containing either all *E. coli*-specific 18-mers (**a**) or phylogroup-specific 18-mers (**b,c**) in the metagenomes. (**d**) Mean absolute deviations estimated from the mean values of 43 paired samples. (**e,f**) Network visualization of intraspecific correlations among *E. coli* phylogroups before and after MD. Node sizes correspond to mean phylogroup abundances, while the connecting lines represent statistically significant correlations, with thickness reflecting their strength: $p \leq 0.00001$ (thick), $p < 0.001$ (medium), $p < 0.05$ (thin).

Based on the previous studies, demonstrated significant underrepresentation of *Bacteroides* in the gut microbiomes of obese individuals [72], we anticipated to detect a difference in their presence in response to MD or resulting changes in their relationships with *E. coli*. The expected increase of at least 10% in *Bacteroides* abundance relative to baseline was indeed observed in 18 out of 43 model microbiomes, but in 14 samples, the percentage of *Bacteroides* decreased (**Figure 9a**).

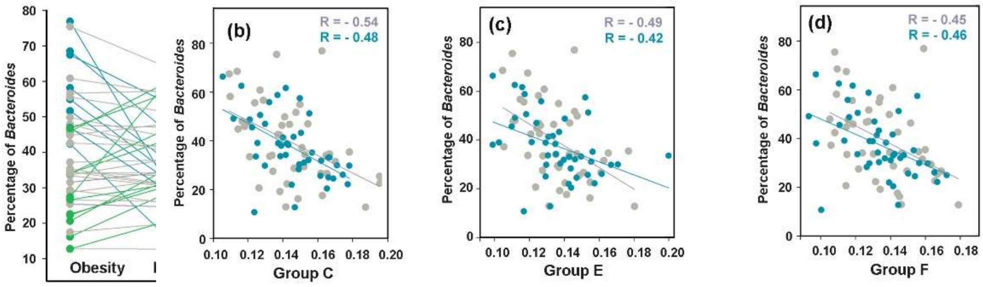
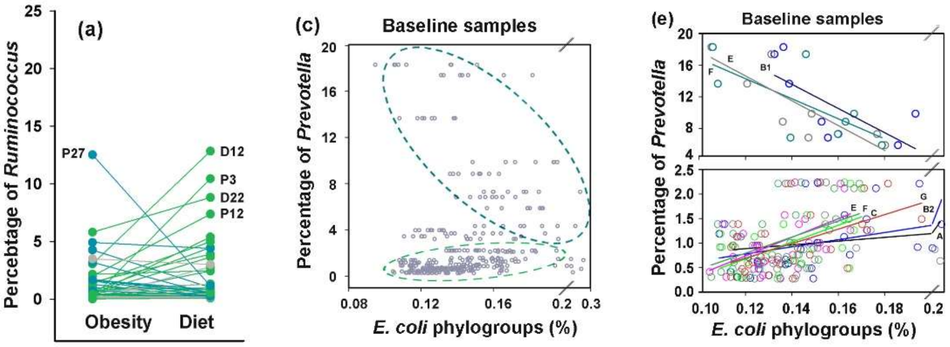


Figure 9. Bidirectional changes in *Bacteroides* abundance induced by Mediterranean diet adherence retained negative correlations with *E. coli* phylogroups. (**a**) Changes in *Bacteroides* prevalence were calculated from mean percentages in replicate fecal samples collected from 43 overweight individuals before and after 8 weeks of a diet. Plots are color-coded to indicate ≥10% increase (green) or decrease (blue) from baseline. Changes within 10% of baseline are shown in gray. (**b**-**d**): Representative negative correlation patterns between *Bacteroides* and *E. coli* phylogroups C (**b**), E (**c**), and F (**d**). Baseline and post-intervention samples are shown in gray and blue, respectively.

Therefore, even dominant bacterial genera employ bimodal adaptive response to dietary restriction. Prior to dietary intervention, only *E. coli* groups B1, C, E and F (**Figure 9b–d**) exhibited statistically significant negative links with *Bacteroides* ($0.00044 \leq p \leq 0.036$), but following 8 weeks of diet, these antagonistic relationships became significant for all *E. coli* phylogroups (**Supplementary Table S3**). This connection aligns with observations from the PRJEB28097 project datasets (**Figure 7e,f**), suggesting that the antagonistic interaction with *Bacteroides* may represent a fundamental property of *E. coli*.

While the plant-based diet increased *Ruminococcus* abundance in 25 of 43 samples, it had little effect on the average percentage of bacteria from this genus (**Figure 10a**). Furthermore, we found no robust evidence of interspecific associations between *Ruminococcus* and *E. coli* (**Supplementary Table S3**).



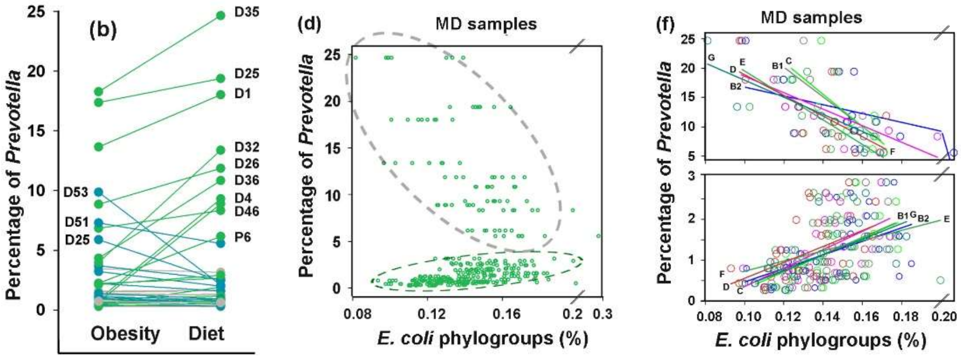


Figure 10. While inducing bidirectional changes in *Ruminococcus* and *Prevotella* abundance, MD did not significantly alter their presence in the gut, but revealed a bimodal relationship between the abundance of *E. coli* and the persistence of *Prevotella*. (**a**,**b**) Changes in *Ruminococcus* (**a**) and *Prevatella* prevalence were calculated from mean percentages in replicate fecal samples collected from 43 overweight individuals before and after 8 weeks of a diet. Plots are color-coded to indicate ≥10% increase (green) or decrease (blue) from baseline. (**c–f**) Scatter plots showing either the entire set of samples (**c**,**d**), or samples divided into two categories based on the percentage of *Prevotella* in the microbiomes (**e**,**f**). Ovals outline symbols with different correlation modes.

However, correlation with *Prevotella*, anticipated based on its symbiosis with *E. coli* [1] and the relationships found in the PRJEB28097 project datasets (**Supplementary Table S3**) emerged in unexpected manner (**Figure 10c–f**). As a dominant genus in the second human enterotype [1], *Prevotella*, was detected in all microbiomes, with a mean abundance approximately twice as high as that of *Ruminococcus* (**Figures 10a,b**). The high fruit/vegetable, low meat Mediterranean diet is beneficial for *Prevotella* consuming complex carbohydrates and resulted in bidirectional adaptive changes in the persistence of this genus (**Figure 10b**). As a result, its weak positive correlations across the entire baseline dataset with groups C (R = 0.31, p = 0.021) and D (R = 0.46, p = 0.0009) disappeared (**Supplementary Table 3**). However, the scatter plots between abundances of *Prevotella* and *E. coli* phylogroups (**Figure 10c,d**) were far from both correlative and random, suggesting a potential bimodal relationship governing *E. coli* phylogroup abundance as a function of *Prevotella* levels. When the control set containing samples from overweight individuals was divided into two subsets (**Figure 10e**), the estimated R-values for correlation in the eight *Prevotella*-rich microbiomes (>5%) ranged from -0.53 to -0.84, with significant negative relationships observed for groups B1, D, E and F (0.009 < P ≤ 0.050). In contrast, significant positive correlations were detected in the 30 samples with low *Prevotella* abundance (<3%) for groups A, B2, C, E, F and G (0.42 < R < 0.54, 0.0019 < P ≤ 0.022). Thus, at least groups E and F may employ distinct communication modes with *Prevotella*. Following dietary intervention, when intraspecific homeostasis became balanced (**Figure 8f**), the 11 *Prevotella*-rich microbiomes exhibited R-values ranging from -0.48 to -0.77 with statistically significant correlations in all groups except A (0.003 < p ≤ 0.019, **Figure 10f**), and all groups in 31 *Prevotella*-low samples demonstrated significant positive correlations (**Supplementary Table 3**). These findings are among the most important in our study, as they highlighted adaptive capacity of *E. coli* to modulate interactions with *Prevotella* based on its abundance and phylogroup-specific connections with dominant taxa, suggesting divergent ecological strategies among *E. coli* lineages.

3.8. Assessing Difference Between Samples, Machine-Learning Approaches May Also Be Implemented to Reveal Individual Similarity

UMAP clustering clearly segregated pre- and post-MD microbiome samples into two distinct groups (**Figure 11a**) and binary classification using at least two models (RF and GB) reliable distinguished between them (**Figure 11b**).

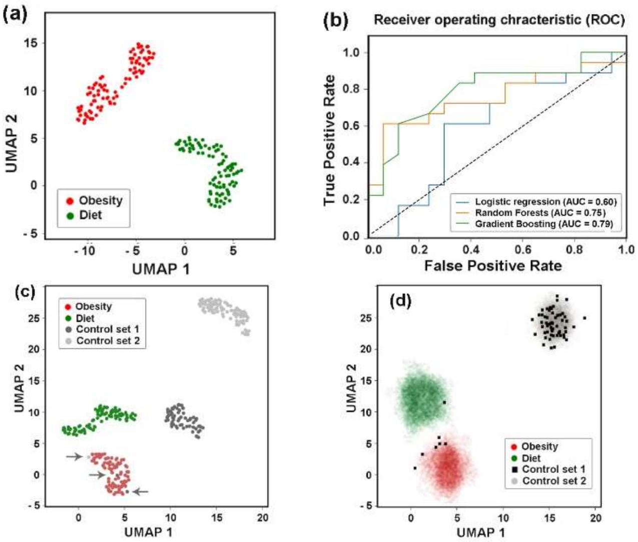


Figure 11. Visualization and statistical assessment of differences between *E. coli* populations (**a**,**b**) and optimization of UMAP for individual identity testing (**c**,**d**). (**a**) UMAP clustering of all 172 samples from the PRJEB33500 project (parameters: n_neighbors = 35, min_dist = 0.7). (**b**) ROC curves and AUC values for the same dataset, calculated using mean values of paired samples. (**c**) Joint UMAP clustering of Control set 1 and 2 (baseline samples from PRJEB7774 and PRJEB28097) with PRJEB33500 project samples. Arrows highlight samples with altered clustering in the combined classification. (**d**) Superimposed clusters from 65 images obtained in a “virtual diagnostics” experiment, when each sample from Control set 1 (black squares) was individually added to a combined set of 264 samples from three other datasets. Clustering parameters for (**c**,**d**): n_neighbors=20, min_dist=1.

However, when these two sets were jointly classified with control samples from the PRJEB7774 and PRJEB28097 projects, we observed three samples, in which the *E. coli* populations closely resembled those of overweight individuals (indicated by arrows in **Figure 11c**). It was not surprising that both sets of the last project differed from the control sets of PRJEB7774 and PRJEB28097, as the PRJEB33500 samples were obtained from apparently healthy individuals with a clear physiological peculiarity. The divergence of the two sets with control samples obtained from healthy donors (**Figure 11c**) likely reflects a batch effect – a phenomenon when non-biological experimental factors introduce artefactual changes. Known contributors to batch effects include differences in laboratory conditions, reagents, and sequencing instrumentation [73,74]. As the PRJEB7774 and PRJEB28097 samples were originated from different sources (Beijing Genome Institute and Weizmann Institute of Science, respectively), their national/geographical characteristics may have driven their separation more strongly than shared *E. coli* phylogroup distributions could combine them. The most important observation made in this part of the study, was the presence of three control samples located outside their primary clusters (**Figure 11c**), suggesting their similarity to the samples of overweight people.

To evaluate, how samples from Control set 1 would be distributed across the other three clusters when added individually, we performed a “virtual diagnostic” experiment (**Figure 11d**). This ignored the influence of Control set 1’s specific features, while emphasizing the intraspecific similarities inherent to “healthy microbiomes”. The joint classification was performed 65 times with 265 samples, of which 92 belonged to Control set 2, 172 were from dietary experiment sets, and 1 test sample was from Control set 1 in each iteration. Because each substitution altered the UMAP projection, we overlaid only clusters rather than entire images (**Figure 11d**). Due to variations in cluster configurations in the individual UMAP images, the overlapping regions appeared as diffuse clouds. Black squares mark the locations of all Control set 1 samples in the UMAP images. The majority (89.2%) of them clustered with samples obtained from healthy donors who had not received antibiotics or probiotics prior to sampling (Control set 2). Six samples grouped with microbiomes from overweight individuals, while one joined to samples from Mediterranean diet followers. Even if six healthy donors from Control set 1 were not clinically obese and one did not prefer a plant-rich diet, this non-random distribution of Control set 1 samples among health-associated categories demonstrated the predictive power of intraspecific *E. coli* characteristics. Therefore, the distribution patterns of *E. coli* phylogroups or the intraspecific homeostasis of some other gut bacteria may reflect host physiological state, potentially serving as a basis for diagnostic applications.

4. Discussion

Based on the assumption that the genetic background establishing epistatic interactions with horizontally acquired genes also shapes phylogroup homeostasis of *E. coli*, which in turn governs both intraspecific and interspecific relationships within microbiomes, we investigated how *E. coli* phylogroups distribution in the human gut correlates with host physiological state. Our study addressed two key questions: (1) Are there any phylogroup-specific responses of *E. coli* to environmental changes caused by either chronic intestinal disorders or acute temporary disturbances? (2) How sensitive is intraspecific balance of *E. coli* to such disruptions? While seemingly straightforward, these questions led us to discover unexpectedly profound rearrangements in both intraspecific relationships and in interspecific connections.

When assessing differences in the abundance of *E. coli* phylogroups between the control samples and microbiomes from individuals formed under different physiological conditions, we observed a statistically significant decrease only in phylogroups B2, C and F following probiotic-mediated recovery after antibiotic treatment (**Figure 6b,c**). In most other cases, changes in mean abundance were not statistically significant. However, a phylogroup-specific response was evident in their variability. Even in microbiomes adapted to chronical colon diseases, phylogroups B2 and D exhibited a significant increase in mean absolute deviation (**Figure 3d)**. More pronounced individual changes were observed in response to antibiotic treatment, with groups A, B2, D and F contributing most to adaptive variability of *E. coli* (**Figure 6b,d**).

Bidirectional changes in isogenic bacterial populations have long been recognized as "bistability" [75] or "bimodality" [76]. This phenomenon enables bacteria to adopt alternative survival strategies in adverse environments [75,76] or enhance virulence [77]. A classic example of such diversification is the emergence of persisters, i.e. subpopulations of cells that develop resistance to toxic agents or achieve antimicrobial tolerance by entering a dormant, slow-growing state [78]. Several mechanisms drive population bifurcation, including structural rearrangements and mutations in the genome [79], epigenetic modifications [77] and "transcriptional noise" [80,81], which, due to stochastic gene expression and feedback regulation [76] can cause segregation into two or more sub-populations. Some of the most compelling evidences for behavioral variability comes from single-cell studies [82,83]. Particularly relevant to our work are findings on species variability in natural microbiomes. For instance, using a mouse model of chronic colonization, W. Elhenawy and coauthors showed that Crohn's disease-associated adherent-invasive *E. coli* (AIEC) isolates undergo host-specific adaptive diversification [84]. The authors identified two lineages that outcompeted the ancestral strain by enhancing invasion or improving acetate utilization in the gut. Although AIEC bacteria are distributed across all *E. coli* phylogroups, they are predominantly associated with group B2 [85], which displayed significantly increased variability in the gut microbiota of patients with chronic colorectal disorders and following antibiotic exposure (**Figures 3d** and **6d**). Consequently, phylogroups B2 and D formed the fewest intraspecific connections in the analyzed datasets (**Figures 4e–g**, **7b–e**, **8e,f**).

Phylogroup E, on the contrary, demonstrated the highest stability in baseline samples (0.008 ≤ MAD ≤ 0.0145) and along with the group F formed the most extensive network of intraspecific correlations. By incorporating over a thousand alien genes [27,28] into the chromosomes of *E. coli* serotype O157:H7 alone [22] and domesticating more than 460 prophages [23], group E bacteria are forced to control expression of a larger number of genes than bacteria with smaller genomes of other groups. Thus, a weaker correlation between them was intuitively expected. Yet our analysis revealed the opposite pattern. It is therefore possible that the genetic background of group E bacteria, evolutionarily tuned to integrate alien genes, was also tuned to maintain the balance of *E. coli* phylogroups.

In evaluating interspecific relationships between *E. coli* phylogroups and dominant enterotype taxa, we observed predominantly negative correlations with *Bacteroides* (**Figures 7,9** and **Supplementary Table S3**). This is in line with several publications describing competitive relationships between these genera [86–88]. Following complete disruption of these links by antibiotics, the negative correlations with *Bacteroides* were spontaneously restored by group A bacteria and increased from an insignificant level for group E bacteria. By the end of the experiment, the interspecies connectivity network with *Bacteroides* had nearly returned to baseline levels. However, probiotic-mediated restoration converted phylogroups A and D links with *Bacteroides* from significantly negative to weakly positive (R = 0.46, p = 0.12). That means that the type of interspecific relations between *E. coli* phylogroups and dominant gut taxa may change depending on the presence of probiotic bacteria, which are only minor components in the human gut biota.

Interspecific interactions with *Ruminococcus* were mostly non-significant. Surprisingly, however, significant positive correlations were observed with phylogroups B1, C, E and G under antibiotic exposure, i.e. a condition when most microbial connections were disrupted (**Supplementary Table S3**). Interestingly, phylogroups B1, C, and G formed similar associations in the microbiomes of overweight individuals with high *Prevotella* levels. After adherence to the Mediterranean diet, positive associations with *Ruminococcus* expanded to all groups, though statistically significant links shifted to groups D and E.

An even more unexpected observation emerged from our analysis of interspecific relationships between *E. coli* phylogroups and *Prevotella*. While the MetaHIT consortium's human gut microbiome analysis predicted negative correlations between these taxa [1], our evaluation of the PRJEB28097 project control dataset revealed positive correlations between *Prevotella* and three *E. coli* groups (A, B1 and D). The same links persisted in samples from overweight individuals with low *Prevotella* abundance (positive correlation with all *E. coli* phylogroups except B1). Only when *Prevotella* abundance exceeded 5%, we observed the predicted negative correlations with groups B1, E, and F (**Figure 10e**). Following dietary restriction, antagonistic relationships in *Prevotella*-rich biota displayed all phylogroups except A, while positive correlations in *Prevotella*-depleted microbiomes were enhanced and maitained by all phylogroups except E. Thus, both interaction types became significantly stronger and cannot be ignored. To our knowledge, this represent the first documented evidence of *E. coli* phylogroups switching between interspecific correlation types based on dominant taxa abundance. Unfortunately, we were unable to validate this phenomenon using the other analyzed datasets. In the colorectal disease project (PRJEB774), *Prevotella* abundance in all samples was very low (< 1.2%) and no significant correlations with *E. coli* were observed, while the small number of independent variables in PRJEB28097 project prevented their meaningful stratification into two categories.

5. Conclusions

Our data indicate that intraspecific homeostasis of *E. coli* relies on positive correlations between all phylogroups, which are typically persisted at roughly equal levels in microbiomes. Intraspecific balance is highly sensitive to host physiological sate. External perturbations, compositional shift in biota, or chronic diseases induced adaptive diversification in the abundance of individual phylogroups, which form intraspecific connections but never establish antagonistic relations. Correlation analysis revealed phylogroup-specific differences in interspecific connectivity networks with dominant taxa across all model datasets, consistent with the proposal that intraspecific homeostasis depends on epistatic relationships between *E. coli* phylogroups and evolutionary tuned regulatory networks formed with other genera. The amazing sensitivity of intraspecific homeostasis to both chronic abnormalities and artificial interventions allowed us to identify unexpected patterns and formulate a number of questions for further research. In particular, it has become important to understand to what extent *E. coli* homeostasis depends on enterotype, how widespread is the bimodal correlation of *E. coli* with *Prevotella* and to what extent are phylogroups of other species sensitive to physiological state of the host? Using profiles of *E. coli* phylogroups, we observed a high predictive power of UMAP clustering in assessing the physiological status of at least healthy donors. The creation of a database containing sets of reference samples of various intestinal pathologies and sets of phylogroup-specific *k*-mers for potentially virulent bacteria will pave the way for the implementation of intraspecific phylotyping in clinical medicine and diagnostics.

Supplementary Materials: The following supporting information can be downloaded at: Preprints.org, **Figure S1**: Phylogenetic tree of 154 *E. coli* strains constructed with the Set 2 barcodes using neighbor-joining method. **Table S1:** *E. coli* strains whose genomes were used for intraspecific phylotyping (Set 2). **Table S2:** Statistics and normalized percentage of reads with *E. coli* phylogroup-specific 18-mers in human gut microbiomes; **Table S3:** Intraspecific and interspecific relationships of *E. coli* phylogroups in the human gut microbiomes.

Author Contributions: Conceptualization, M.F. and O.); methodology, O.O.; software, V.P.; validation, M.F., S.K. and O.O.; formal analysis, S.K.; investigation, M.F. and S.K.; resources, O.O.; data curation, S.K.; writing—original draft preparation, M.F.; writing—review and editing, O.O.; visualization, M.F.; supervision, O.O.; project administration, O.O.; funding acquisition, O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Russian Science Foundation, grant number 24-14-00433 (M.F., S.K. and O.O.).

Institutional Review Board Statement: All human datasets used in this study were obtained from the publicly available database (NCBI SRA). BioProject ID numbers are indicated in Table 2.

Informed Consent Statement: Not applicable. All human datasets used in this study were obtained from the publicly available database (NCBI SRA).

Data Availability Statement: Sets with unique 18-mers of eight *E. coli* phylogroups are available in the GitHub repository located at https://github.com/marsfro/ecoli_counter. A Python script developed for counting reads containing phylogroup-specific *k*-mers in WGS metagenomes is available at https://github.com/marsfro/ecoli_counter/kmers_ecoli_counter.py.

Acknowledgments: The authors acknowledge the use of DeepSeek Chat for language editing assistance during the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

References

1. Arumugam, M.; Raes, J.; Pelletier, E.; Le Paslier, D.; Yamada, T.; Mende, D.R.; Fernandes, G.R.; Tap, J.; Bruls, T.; Batto, J.-M.; et al. Enterotypes of the human gut microbiome. *Nature* **2011**, *473*, 174–180. doi: 10.1038/nature09944.
2. Martinson, J.N.V.; Walk, S.T. *Escherichia coli* residency in the gut of healthy human adults. *EcoSal Plus* **2020**, *9*, ESP-0003-2020. doi: 10.1128/ecosalplus.ESP-0003-2020.
3. Clermont, O.; Bonacorsi, S.; Bingen, E. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl. Environ. Microbiol.* **2000**, *66*, 4555–4558. doi: 10.1128/AEM.66.10.4555-4558.2000.
4. Jaureguy, F.; Landraud, L.; Passet, V.; Diancourt, L.; Frapy, E.; Guigon, G.; Carbonnelle, E.; Lortholary, O.; Clermont, O.; Denamur, E.; et al. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* **2008**, *9*, 560. doi: 10.1186/1471-2164-9-560.
5. Clermont, O.; Olier, M.; Hoede, C.; Diancourt, L.; Brisse, S.; Keroudean, M.; Glodt, J.; Picard, B.; Oswald, E.; Denamur, E. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infect. Genet. Evol.* **2011**, *11*, 654–662. doi: 10.1016/j.meegid.2011.02.005.
6. Beghain, J.; Bridier-Nahmias, A.; Le Nagard, H.; Denamur, E.; Clermont, O. ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb. Genom.* **2018**, *4*, 000192. doi: 10.1099/mgen.0.000192.
7. Clermont, O.; Christenson, J.K.; Denamur, E.; Gordon, D.M. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* **2013**, *5*, 58–65. doi: 10.1111/1758-2229.12019.
8. Lu, S.; Jin, D.; Wu, S.; Yang, J.; Lan, R.; Bai, X.; Liu, S.; Meng, Q.; Yuan, X.; Zhou,J.; et al. Insights into the evolution of pathogenicity of *Escherichia coli* from genomic analysis of intestinal *E. coli* of *Marmota himalayana* in Qinghai-Tibet plateau of China. *Emerg. Microbes Infect.* **2016**, *5*, e122. doi: 10.1038/emi.2016.122.

9. Clermont, O.; Dixit, O.V.A.; Vangchhia, B.; Condamine, B.; Dion, S.; Bridier-Nahmias, A.; Denamur, E.; Gordon, D. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ. Microbiol.* **2019**, *21*, 3107–3117. doi: 10.1111/1462-2920.14713.
10. Panyukov, V.V.; Kiselev, S.S.; Ozoline, O.N. Unique *k*-mers as strain-specific barcodes for phylogenetic analysis and natural microbiome profiling. *Int. J. Mol. Sci.* **2020**, *21*, 944. doi: 10.3390/ijms21030944.
11. Abram, K.; Udaondo, Z.; Bleker, C.; Wanchai, V.; Wassenaar, T.M.; Robeson, M.S.; Ussery, D.W. Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups. *Commun. Biol.* **2021**, *4*, 117. doi: 10.1038/s42003-020-01626-5.
12. Picard, B.; Garcia, J.S.; Gouriou, S.; Duriez, P.; Brahimi, N.; Bingen, E.; Elion, J.; Denamur, E. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect. Immun.* **1999**, *67*, 546–553. doi: 10.1128/IAI.67.2.546-553.1999.
13. Ochman, H.; Lawrence, J.G.; Groisman, E.A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **2000**, *405*, 299–304. doi: 10.1038/35012500.
14. Escobar-Paramo, P.; Clermont, O.; Blanc-Potard, A.B.; Bui, H.; Le Bouguenec, C.; Denamur, E. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol. Biol. Evol.* **2004**, *21*, 1085–1094. doi: 10.1093/molbev/msh118.
15. Ondov, B.D.; Treangen, T.J.; Melsted, P.; Mallonee, A.B.; Bergman, N.H.; Koren, S.; Phillippy, A.M. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **2016**, *17*, 132. doi: 10.1186/s13059-016-0997-x.
16. Chakraborty, A.; Saralaya, V.; Adhikari, P.; Shenoy, S.; Baliga, S.; Hegde, A. Characterization of *Escherichia coli* phylogenetic groups associated with extraintestinal infections in South Indian population. *Ann. Med. Health Sci. Res.* **2015**, *5*, 241–246. doi: 10.4103/2141-9248.160192.
17. Rasko, D.A.; Webster, D.R.; Sahl, J.W.; Bashir, A.; Boisen, N.; Scheutz, F.; Paxinos, E.E.; Sebra, R.; Chin, C.-S.; Iliopoulos, D.; et al. Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. *N. Engl. J. Med.* **2011**, *365*, 709–717. doi: 10.1056/NEJMoa1106920.
18. Bielaszewska, M.; Mellmann, A.; Zhang, W.; Kock, R.; Fruth, A.; Bauwens, A.; Peters, G.; Karch, H. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect. Dis.* **2011**, *11*, 671–676. doi: 10.1016/S1473-3099(11)70165-7.
19. Brzuszkiewicz, E.; Thurmer, A.; Schuldes, J.; Leimbach, A.; Liesegang, H.; Meyer, F.D.; Boelter, J.; Petersen, H.; Gottschalk, G.; Daniel, R. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch. Microbiol.* **2011**, *193*, 883–891. doi: 10.1007/s00203-011-0725-6.
20. Schiller, P.; Knodler, M.; Berger, P.; Greune, L.; Fruth, A.; Mellmann, A.; Dersch, P.; Berger, M.; Dobrindt, U. The superior adherence phenotype of *E. coli* O104:H4 is directly mediated by the aggregative adherence fimbriae type I. *Virulence* **2021**, *12*, 346–359. doi: 10.1080/21505594.2020.1868841.
21. 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*; Choffnes E.R., Relman D.A., Olsen L., Hutton R., Mack A.; National Academies Press: Washington (DC), USA, 2012; pp. 115–129.
22. Putonti, C.; Luo, Y.; Katili, C.; Chumakov, S.; Fox, G.E.; Graur, D.; Fofanov, Y. A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences. *Mol. Biol. Evol.* **2006**, *23*, 1863–1868. doi: 10.1093/molbev/msl053.
23. Lim, J.Y.; Yoon, J.; Hovde, C.J. A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. *J. Microbiol. Biotechnol.* **2010**, *20*, 5–14. doi: 10.4014/jmb.0908.08007.
24. Wick, L.M.; Qi, W.; Lacher, D.W.; Whittam, T.S. Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. *J. Bacteriol.* **2005**, *187*, 1783–1791. doi: 10.1128/JB.187.5.1783-1791.2005.
25. Rahal, E.A.; Kazzi, N.; Nassar, F.J.; Matar, G.M. *Escherichia coli* O157:H7 - clinical aspects and novel treatment approaches. *Front. Cell. Infect. Microbiol.* **2012**, *2*, 138. doi: 10.3389/fcimb.2012.00138.
26. Franzin, F.M.; Sircili, M.P. Locus of enterocyte effacement: a pathogenicity island involved in the virulence of enteropathogenic and enterohemorrhagic *Escherichia coli* subjected to a complex network of gene regulation. *Biomed. Res. Int.* **2015**, *2015*, 534738. doi: 10.1155/2015/534738.
27. Perna, N.T.; Plunkett, G.; Burland, V.; Mau, B.; Glasner, J.D.; Rose, D.J.; Mayhew, G.F.; Evans, P.S.; Gregor, J.; Kirkpatrick, H.A.; et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **2001**, *409*, 529–533. doi: 10.1038/35054089.
28. Hayashi, T.; Makino, K.; Ohnishi, M.; Kurokawa, K.; Ishii, K.; Yokoyama, K.; Han, C.G.; Ohtsubo, E.; Nakayama, K.; Murata, T.; et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **2001**, *8*, 11–22. doi: 10.1093/dnares/8.1.11.
29. Jiang, L.; Yang, W.; Jiang, X.; Yao, T.; Wang, L.; Yang, B. Virulence-related O islands in enterohemorrhagic *Escherichia coli* O157:H7. *Gut Microbes* **2021**, *13*, e1992237. doi: 10.1080/19490976.2021.1992237.

30. Taieb, F.; Petit, C.; Nougayrede, J.P.; Oswald, E. The enterobacterial genotoxins: cytolethal distending toxin and colibactin. *EcoSal Plus* **2016**, *7*, ESP-0008-2016. doi: 10.1128/ecosalplus.ESP-0008-2016.
31. Nougayrede, J.P.; Homburg, S.; Taieb, F.; Boury, M.; Brzuszkiewicz, E.; Gottschalk, G.; Buchrieser, C.; Hacker, J.; Dobrindt, U.; Oswald, E. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* **2006**, *313*, 848–851. doi: 10.1126/science.1127059.
32. Kruis, W.; Fric, P.; Pokrotnieks, J.; Lukas, M.; Fixa, B.; Kascak, M.; Kamm, M.A.; Weismueller, J.; Beglinger, C.; Stolte, M.; et al. Maintaining remission of ulcerative colitis with the probiotic *Escherichia coli* Nissle 1917 is as effective as with standard mesalazine. *Gut* **2004**, *53*, 1617–1623. doi: 10.1136/gut.2003.037747.
33. Dubbert, S.; Klinkert, B.; Schimiczek, M.; Wassenaar, T.M.; Bunau, R.V. No genotoxicity is detectable for *Escherichia coli* strain Nissle 1917 by standard in vitro and in vivo tests. *Eur. J. Microbiol. Immunol.* **2020**, *10*, 11–19. doi: 10.1556/1886.2019.00025.
34. Homburg, S.; Oswald, E.; Hacker, J.; Dobrindt, U. Expression analysis of the colibactin gene cluster coding for a novel polyketide in *Escherichia coli*. *FEMS Microbiol. Lett.* **2007**, *275*, 255–262. doi: 10.1111/j.1574-6968.2007.00889.x.
35. Wassenaar, T.M. *E. coli* and colorectal cancer: a complex relationship that deserves a critical mindset. *Crit. Rev. Microbiol.* **2018**, *44*, 619–632. doi: 10.1080/1040841X.2018.1481013.
36. Coque, T.M.; Novais, A.; Carattoli, A.; Poirel, L.; Pitout, J.; Peixe, L.; Baquero, F.; Canton, R.; Nordmann, P. Dissemination of clonally related *Escherichia coli* strains expressing extended-spectrum β-lactamase CTX-M-15. *Emerg. Infect. Dis.* **2008**, *14*, 195–200. doi: 10.3201/eid1402.070350.
37. Boyd, D.A.; Tyler, S.; Christianson, S.; McGeer, A.; Muller, M.P.; Willey, B.M.; Bryce, E.; Gardam, M.; Nordmann, P.; Mulvey, M.R. Complete nucleotide sequence of a 92-kilobase plasmid harboring the CTX-M-15 extended-spectrum beta-lactamase involved in an outbreak in long-term-care facilities in Toronto, Canada. *Antimicrob. Agents Chemother.* **2004**, *48*, 3758–3764. doi: 10.1128/AAC.48.10.3758-3764.2004.
38. Feng, Y.; Mannion, A.; Madden, C.M.; Swennes, A.G.; Townes, C; Byrd, C; Marini, R.P.; Fox, J.G. Cytotoxic *Escherichia coli* strains encoding colibactin and cytotoxic necrotizing factor (CNF) colonize laboratory macaques. *Gut Pathog.* **2017**, *9*, 71. doi: 10.1186/s13099-017-0220-y.
39. Chaoprasid, P.; Dersch, P. The cytotoxic necrotizing factors (CNFs)—a family of Rho GTPase-activating bacterial exotoxins. *Toxins* **2021**, *13*, 901. doi: 10.3390/toxins13120901.
40. Wang, M.C.; Fan, Y.H.; Zhang, Y.Z.; Bregente, C.J.B.; Lin, W.H.; Chen, C.A.; Lin, T.P.; Kao, C.Y. Characterization of uropathogenic *Escherichia coli* phylogroups associated with antimicrobial resistance, virulence factor distribution, and virulence-related phenotypes. *Infect. Genet. Evol.* **2023**, *114*, 105493. doi: 10.1016/j.meegid.2023.105493.
41. Paramita, R.I.; Nelwan, E.J.; Fadilah, F.; Renesteen, E.; Puspandari, N.; Erlina, L. Genome-based characterization of *Escherichia coli* causing bloodstream infection through next-generation sequencing. *PLoS One* **2020**, *15*, e0244358. doi: 10.1371/journal.pone.0244358.
42. Deku, D.G.; Duedu, K.O.; Kinanyok, S.; Kpene, G.E.; Feglo, P.K. Phylogenicity and virulence profiles of clinical *Escherichia coli* isolates in the Ho Teaching Hospital of Ghana. *Biomed. Res. Int.* **2022**, *2022*, 1347033. doi: 10.1155/2022/1347033.
43. Flament-Simon, S.C.; Nicolas-Chanoine, M.H.; Garcia, V.; Duprilot, M.; Mayer, N.; Alonso, M.P.; Garcia-Menino, I.; Blanco, J.E.; Blanco, M.; Blanco, J. Clonal structure, virulence factor-encoding genes and antibiotic resistance of *Escherichia coli*, causing urinary tract infections and other extraintestinal infections in humans in Spain and France during 2016. *Antibiotics (Basel)* **2020**, *9*, 161. doi: 10.3390/antibiotics9040161.
44. Murase, K.; Martin, P.; Porcheron, G.; Houle, S.; Helloin, E.; Penary, M.; Nougayrede, J.P.; Dozois, C.M.; Hayashi, T.; Oswald, E. HlyF produced by extraintestinal pathogenic *Escherichia coli* is a virulence factor that regulates outer membrane vesicle biogenesis. *J. Infect. Dis.* **2016**, *213*, 856–865. doi: 10.1093/infdis/jiv506.
45. Gigliucci, F.; van Hoek, A.H.A.M.; Chiani, P.; Knijn, A.; Minelli, F.; Scavia, G.; Franz, E.; Morabito, S.; Michelacci, V. Genomic characterization of *hlyF*-positive shiga toxin-producing *Escherichia coli*, Italy and the Netherlands, 2000-2019. *Emerg. Infect. Dis.* **2021**, *27*, 853–861. doi: 10.3201/eid2703.203110.
46. Ferreira, M.R.A.; Silva, T.S.; Stella, A.E.; Conceicao, F.R.; dos Reis, E.F.; Moreira, C.N. Detection of virulence factors and antimicrobial resistance patterns in shiga toxin-producing *Escherichia coli* isolates from sheep. *Pesq. Vet. Bras.* **2015**, *35*, 775–780. doi: 10.1590/S0100-736X2015000900002.
47. Karakaya, E.; Aydin, F.; Kayman, T.; Abay, S. *Escherichia coli* in different animal feces: phylotypes and virulence genes. *World J. Microbiol. Biotechnol.* **2022**, *39*, 14. doi: 10.1007/s11274-022-03451-w.
48. Ma, Y.; Zhang, Y.; Xiang, J.; Xiang, S.; Zhao, Y.; Xiao, M.; Du, F.; Ji, H.; Kaboli, P.J.; Wu, X.; et al. Metagenome analysis of intestinal bacteria in healthy people, patients with inflammatory bowel disease and colorectal cancer. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 599734. doi: 10.3389/fcimb.2021.599734.
49. Suez, J.; Zmora, N.; Zilberman-Schapira, G.; Mor, U.; Dori-Bachash, M.; Bashiardes, S.; Zur, M.; Regev-Lehavi, D.; Ben-Zeev Brik, R.; Federici, S.; et al. Post-antibiotic gut mucosal microbiome reconstitution is impaired by probiotics and improved by autologous FMT. *Cell* **2018**, *174*, 1406–1423. doi: 10.1016/j.cell.2018.08.047.

50. Montassier, E.; Valdes-Mas, R.; Batard, E.; Zmora, N.; Dori-Bachash, M.; Suez, J.; Elinav, E. Probiotics impact the antibiotic resistance gene reservoir along the human GI tract in a person-specific and antibiotic-dependent manner. *Nat. Microbiol.* **2021**, *6*, 1043–1054. doi: 10.1038/s41564-021-00920-0.
51. Frolova, M.; Yudin, S.; Makarov, V.; Glazunova, O.; Alikina, O; Markelova, N.; Kolzhetsov, N.; Dzhelyadin, T.; Shcherbakova, V.; Trubitsyn, V.; et al. *Lacticaseibacillus paracasei*: occurrence in the human gut microbiota and *k*-mer-based assessment of intraspecies diversity. *Life (Basel)* **2021**, *11*, 1246. doi: 10.3390/life11111246.
52. Meslier, V.; Laiola, M.; Roager, H.; De Filippis, F.; Roume, H.; Quinquis, B.; Giacco, R.; Mennella, I.; Ferracane, R.; Pons, N.; et al. Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. *Gut* **2020**, *69*, 1258–1268. doi: 10.1136/gutjnl-2019-320438.
53. Kruger, F. Babraham Bioinformatics - Trim Galore! Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 7 April 2025).
54. Sorensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab. Biol. Skrifter* **1948**, *4*, 1–34.
55. Saitou, N.; Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425. doi: 10.1093/oxfordjournals.molbev.a040454.
56. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. doi: 10.1093/molbev/msy096.
57. Kim, D.; Song, L.; Breitwieser, F.P.; Salzberg, S.L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **2016**, *26*, 1721–1729. doi: 10.1101/gr.210641.116.
58. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
59. McInnes, L.; Healy, J.; Melville, J. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv* **2018,** *preprint: nor peer reviewed*. doi: 10.48550/arXiv.1802.03426.
60. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. doi: 10.1214/aoms/1177730491.
61. VassarStats: Calculators for statistical table entries r to P. Available online: http://vassarstats.net/tabs_r.html (accessed on 18 April 2025).
62. Quenouille, M.H. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B* **1949**, *11*, 68–84. doi: 10.1111/j.2517-6161.1949.tb00023.x.
63. Qi, W.; Lacher; D.W.; Bumbaugh, A.C.; Hyma, K.E.; Quellette, L.M.; Large, T.M.; Tarr, C.L.; Whittam, T.S. *Ec*MLST: An online database for multi locus sequence typing of pathogenic *Escherichia coli*. *Comput. Syst. Bioinformatics Conf.* **2004**, 520–521. doi: 10.1109/CSB.2004.1332482.
64. Wirth, T.; Falush, D.; Lan, R.; Colles, F.; Mensa, P.; Wieler, L.H.; Karch, H.; Reeves, P.R.; Maiden, M.C.J.; Ochman, H.; et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol. Microbiol.* **2006**, *60*, 1136–1151. doi: 10.1111/j.1365-2958.2006.05172.x.
65. Lescat, M.; Hoede, C.; Clermont, O.; Garry, L.; Darlu, P.; Tuffery, P.; Denamur, E.; Picard, B. *aes*, the gene encoding the esterase B in *Escherichia coli*, is a powerful phylogenetic marker of the species. *BMC Microbiol.* **2009**, *9*, 273. doi: 10.1186/1471-2180-9-273.
66. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. doi: 10.1093/molbev/msu300.
67. Poirel, L.; Madec, J.-Y.; Lupo, A.; Schink, A.-K.; Kieffer, N.; Nordmann, P.; Schwarz, S. Antimicrobial resistance in *Escherichia coli*. *Microbiol. Spectr.* **2018**, *6*, ARBA-0026-2017. doi: 10.1128/microbiolspec.ARBA-0026-2017.
68. Mohsin, A.S.; Alsakini, A.H.; Ali, M.R. Outbreak of drug resistance *Escherichia coli* phylogenetic F group associated urinary tract infection. *Iran. J. Microbiol.* **2022**, *14*, 341–350. doi: 10.18502/ijm.v14i3.9771.
69. Neyestani, Z.; Khademi, F.; Teimourpour, R.; Amani, M.; Arzanlou, M. Prevalence and mechanisms of ciprofloxacin resistance in *Escherichia coli* isolated from hospitalized patients, healthy carriers, and wastewaters in Iran. *BMC Microbiol.* **2023**, *23*, 191. doi: 10.1186/s12866-023-02940-8.
70. Onderdonk, A.B.; Louie, T.J.; Tally, F.P.; Bartlett, J.G. Activity of metronidazole against *Escherichia coli* in experimental infra-abdominal sepsis. *J. Antimicrob. Chemother.* **1979**, *5*, 201–210. doi: 10.1093/jac/5.2.201.
71. Kolzhetsov, N.; Markelova, N.; Frolova, M.; Alikina, O.; Glazunova, O.; Safonova, L.; Kalashnikova, I.; Yudin, V.; Makarov, V.; Keskinov, A.; et al. Enterotype-dependent probiotic-mediated changes in the male rat intestinal microbiome in vivo and in vitro. *Int. J. Mol. Sci.* **2024**, *25*, 4558. doi: 10.3390/ijms25084558.
72. Andoh, A.; Nishida, A.; Takahashi, K.; Inatomi, O.; Imaeda, H.; Bamba, S.; Kito, K.; Sugimoto, M.; Kobayashi, T. Comparison of the gut microbial community between obese and lean peoples using 16S gene sequencing in a Japanese population. *J. Clin. Biochem. Nutr.* **2016**, *59*, 65. doi: 10.3164/jcbn.15-152.
73. Leek, J.T.; Scharpf, R.B.; Bravo, H.C.; Simcha, D.; Langmead, B.; Johnson, W.E.; Geman, D.; Baggerly, K.; Irizarry, R.A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **2010**, *11*, 733–739. doi: 10.1038/nrg2825.

74. Cuklina, J.; Pedrioli, P.G.A.; Aebersold, R. Review of batch effects prevention, diagnostics, and correction approaches. *Methods Mol. Biol.* **2020**, *2051*, 373–387. doi: 10.1007/978-1-4939-9744-2_16.
75. Dubnau, D.; Losick, R. Bistability in bacteria. *Mol. Microbiol.* **2006**, *61*, 564–572. doi: 10.1111/j.1365-2958.2006.05249.x.
76. Smits, W.K.; Kuipers, O.P.; Veening, J.W. Phenotypic variation in bacteria: the role of feedback regulation. *Nat. Rev. Microbiol.* **2006**, *4*, 259–271. doi: 10.1038/nrmicro1381.
77. Ronin, I.; Katsowich, N.; Rosenshine, I.; Balaban, N.Q. A long-term epigenetic memory switch controls bacterial virulence bimodality. *eLife* **2017**, *6*, e19599. doi: 10.7554/eLife.19599.
78. Xu, Y.; Liu, S.; Zhang, Y.; Zhang, W. DNA adenine methylation is involved in persister formation in *E. coli*. *Microbiol. Res.* **2021**, *246*, 126709. doi: 10.1016/j.micres.2021.126709.
79. Foster, P.L. Stress-induced mutagenesis in bacteria. *Crit. Rev. Biochem. Mol. Biol.* **2007**, *42*, 373–397. doi: 10.1080/10409230701648494.
80. Raj, A.; van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **2008**, *135*, 216–226. doi: 10.1016/j.cell.2008.09.050.
81. Raser, J.M.; O'Shea, E.K. Noise in gene expression: origins, consequences, and control. *Science* **2005**, *309*, 2010–2013. doi: 10.1126/science.1105891.
82. Korobkova, E.; Emonet, T.; Vilar, J.M.G.; Shimizu, T.S.; Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* **2004**, *428*, 574–578. doi: 10.1038/nature02404.
83. Elowitz, M.B.; Levine, A.J.; Siggia, E.D.; Swain, P.S. Stochastic gene expression in a single cell. *Science* **2002**, *297*, 1183–1186. doi: 10.1126/science.1070919.
84. Elhenawy, W.; Tsai, N.T.; Coombes, B.K. Host-specific adaptive diversification of Crohn's disease-associated adherent-invasive *Escherichia coli*. *Cell Host Microbe* **2019**, *25*, 301–312. doi: 10.1016/j.chom.2018.12.010.
85. Denamur, E.; Clermont, O.; Bonacorsi, S.; Gordon, D. The population genetics of pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* **2021**, *19*, 37–54. doi: 10.1038/s41579-020-0416-x.
86. Mirsepasi-Lauridsen, H.C.; Vallance B.A.; Krogfelt K.A.; Petersen A.M. *Escherichia coli* pathobionts associated with inflammatory bowel disease. *Clin Microbiol Rev*. **2019**, *30*, e00060-18. doi: 10.1128/CMR.00060-18.
87. Maier, L.; Pruteanu, M.; Kuhn, M.; Zeller, G.; Telzerow, A.; Anderson E.E.; Brochado, A.R.; Fernandez, K.C.; Dose, H.; Mori, H.; Patil, K.R.; Bork, P.; Typas. A. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*. **2018,** 29, 623-628. doi: 10.1038/nature25979 .
88. Wexler, A.G.; Goodman, A.L. An insider's perspective: Bacteroides as a window into the microbiome. *Nat Microbiol*. **2017**, 25;2:17026. doi: 10.1038/nmicrobiol.2017.26 .

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.