# Distinct *Escherichia coli* transcriptional profiles in the guts of recurrent UTI sufferers revealed by pan-genome hybrid selection

Mark G. Young[1*], Timothy J. Straub[1*], Colin J. Worby[1], Hayden C. Metsky[1], Andreas Gnirke[1], Ryan A. Bronson[1], Lucas R. van Dijk[1,2], Christopher A. Desjardins[1], Christian Matranga[1], James Qu[1], Karen Dodson[3,4], Henry L. Schreiber IV[3,4], Abigail L. Manson[1], Scott J. Hultgren[3,4], Ashlee M. Earl[1**]

*These authors contributed equally to this work.

**Corresponding author

Affiliations: [1]Infectious Disease & Microbiome Program, Broad Institute, Cambridge, MA 02142, USA; [2]Delft Bioinformatics Lab, Delft University of Technology, Van Mourik Broekmanweg 6, Delft, 2628 XE, The Netherlands; [3]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA.; [4]Center for Women's Infectious Disease Research, Washington University School of Medicine, St. Louis, MO, USA

## Abstract

Low-abundance members of microbial communities are difficult to study in their native habitat. This includes *Escherichia coli*, a minor, but common inhabitant of the gastrointestinal tract and opportunistic pathogen, including of the urinary tract, where it causes most infections. While our understanding of the interactions between uropathogenic *Escherichia coli* (UPEC) and the bladder is increasing, comparatively little is known about UPEC in its pre-infection reservoir, partly due to its low abundance there (<1% relative abundance). In order to specifically and sensitively explore the genomes and transcriptomes of diverse *E. coli* from gastrointestinal communities, we developed *E. coli* PanSelect, a set of probes designed to enrich *E. coli*'s broad pangenome. First we demonstrated the ability of PanSelect to enrich diverse strains in an unbiased way using a mock community of known composition. Then we enriched *E. coli* DNA and RNA from human stool microbiomes by 158 and 30-fold, respectively. We also used *E. coli* PanSelect to explore the gene content and transcriptome of *E. coli* within the gut microbiomes of women with history of recurrent urinary tract infection (rUTI), finding differential regulation of pathways that suggests that the rUTI gut environment promotes respiratory vs fermentative metabolism. *E. coli* PanSelect technology holds promise for investigations of native *in vivo* biology of diverse *E. coli* in the gut and other environments, where it is a minor component of the microbial community, using unbiased, culture-free shotgun sequencing. This method could also be generally applied to other highly diverse, low abundance bacteria.

## Introduction

Urinary tract infections (UTIs) are among the most common bacterial infections worldwide, with significant individual and societal impacts, including highly recurrent UTIs (rUTIs), which affect millions, mostly women (Foxman 2002), and drain precious antibiotic resources (The Lancet 2018; Mediavilla et al. 2016; Ajiboye et al. 2009; Karlowsky et al. 2006; Manges et al. 2001). The vast majority of UTIs are caused by uropathogenic *Escherichia coli* (UPEC). In contrast to enterohemorrhagic *E. coli* (EHEC), the genetic definition of UPEC has remained elusive given the wide diversity of *E. coli* that can cause UTIs and rUTIs(Russo and Johnson 2000; Kaper, Nataro, and Mobley 2004; Schreiber et al. 2017). UPEC are common but minor (<1% relative abundance) members of the gastrointestinal tract (Nielsen et al. 2014; Jantunen et al. 2001; Yamamoto et al. 1997)from which they can be excreted in feces and go on to cause rUTIs. While culture- and metagenomic sequencing-based investigations have revealed much about this functionally and evolutionarily diverse group, these approaches have drawbacks, particularly for low abundance organisms, that leave open key questions about the interplay between UPEC in the gastrointestinal tract and infection of the bladder.

Advances in selective nucleic acid enrichment, including selective whole genome amplification (SWGA)(Leichty and Brisson 2014) and hybrid selection (HS)(Mamanova et al. 2010), have made higher resolution study of low abundance organisms in their native habitats possible. SWGA relies on PCR primers that leverage inherent differences in the frequencies of taxon-specific sequence motifs to selectively amplify a target organism above background(Leichty and Brisson 2014), but can result in considerable variation in sequencing coverage across a genome, and is not suited for expression-based investigations. HS, also called hybrid capture, relies on specific oligonucleotides, or probes, that selectively hybridize (*i.e.*, bind) to sequences from the target organism (**Figure 1a-d**)(Mamanova et al. 2010). This method can be used for both genomic and transcriptomic applications and has been used successfully to sequence human exomes(Fisher et al. 2011), malarial parasite (Bright et al. 2012) and viral (Matranga et al. 2014) genomes, as well as the transcriptome of a specific strain of *Bacteroides fragilis* from murine intestinal compartments(Donaldson et al. 2020).

One key challenge to HS implementation has been in the design of probes that recognize the full suite of desired targets in a sample. While the previous studies using HS focused on species with limited overall diversity, or the full genome of the targeted organism was known *a priori*, many bacterial species, like *E. coli*, have large pangenomes, defined as the entire set of genes found across a species. Having emerged >100Mya from a common ancestor of *Salmonella* (Ochman and Wilson 1987), *E. coli* has evolved considerably with more than eight recognized distinct phylogroups (Clermont et al. 2019) collectively accumulating *a* pangenome estimated at over 120,000 diverse gene families(Tantoso et al. 2022). This functional variation can have broad implications for a strain's physiology, interactions, and disease potential ("Structure, Function and Diversity of the Healthy Human Microbiome" 2012; Tenaillon et al. 2010).

To increase our ability to understand UPEC in their native gut habitat including in the context of rUTI, we developed an *E. coli* pangenome-based HS probe set using a combination of comparative genomics, taking into account the known sequence diversity of the entire *E. coli* species, and a recently developed algorithm (Metsky et al. 2019) that enables HS probe design for efficient tiling across diverse sequences. We demonstrated the utility of "*E. coli* PanSelect" to selectively enrich, by orders of magnitude, sequences of diverse, low abundance *E. coli* genomes from mock communities, and *E. coli* genomes and transcriptomes from stool communities from healthy women and rUTI sufferers. These enriched views revealed UPEC's

adaptive strategies for life in the rUTI dysbiotic gut, including a shift from fermentative to aerobic metabolism, which may have relevance for this recurrent disease. The pangenome-based HS approach described here holds promise for investigating other diverse taxa in a wide variety of community contexts where they exist in low abundance.

## Results

### Probe set design covers vast majority of *E. coli* pan-genome

The human gastrointestinal tract can be colonized by many diverse strains of *E. coli*, often at the same time ("Structure, Function and Diversity of the Healthy Human Microbiome" 2012; Tenaillon et al. 2010). In order to design a set of hybrid selection probes to cover the full scope of genetic diversity present in *E. coli*, we downloaded all available RefSeq complete *E. coli* reference assemblies, together with additional diverse, high quality *E. coli* reference assemblies from Genbank (**Figure 1e**, **Supplementary Figure 1, Supplementary Table 1**; Methods), for a total of 3,436 genomes representing all major *E. coli* phylogroups (**Supplementary Table 2**). After clustering to remove near-identical references, we identified shared orthogroups to use with the CATCH algorithm to design 60-75 bp probes (Metsky et al. 2019). To reduce computational runtime during probe design, we first clustered orthogroup genes with >80% identity. After filtering steps to remove probes that would likely also hybridize with sequences from other taxa including common gut bacteria (e.g., *Bacteroidetes*) (**Methods**), we were left with a final set of 892,415 short (60-75 bp) oligonucleotide probes designed to specifically hybridize to and cover the *E. coli* pan-genome, including the vast majority of the known diversity of *E. coli* (**Figure 1e; Supplementary Figure 1**).

In order to determine the expected coverage of our final probe set against the *E. coli* pan-genome, we used blastn to query our probe sequences against all protein-coding gene sequences in our final set of references. We classified genes as 'selected' by the probe set if they had a blast hit of at least 65 bp and no more than 8 mismatches (**Methods**). We successfully selected 99.95% of the genes (8,131,231 genes) and 99.54% of the orthogroups (52,688 orthogroups) targeted by our probe design, even after removing thousands of potentially off-target probes. In addition, the probe set selected 29,232 of the *E. coli* genes that had not been included in the probe design (likely due to sequence homology), bringing the total coverage to 97.97% (8,160,463) of genes found in the total pangenome, and suggesting that the probe set may be generalizable to additional sequence diversity not captured by our design. Most missing genes were annotated as hypothetical proteins (67%) or mobile genetic elements (33%).

### *E. coli* PanSelect enriched four strains of *E. coli* in a mock metagenomic community without bias in strain composition

We first assessed the ability of our probe set to enrich *E. coli* sequences from a previously described DNA-based mock community (van Dijk et al. 2021) including an uneven mixture of DNA from four distinct, previously sequenced isolates of *E. coli* totaling approximately 1% overall relative abundance (RA), set against a background of human DNA. We performed HS on Illumina libraries generated from this mock community and then sequenced both HS-enriched and unenriched libraries (**Figure 2, Supplementary Results A, Supplementary Table 3, Methods**). Using StrainGE (van Dijk et al. 2021) output to compute and compare the representation and relative proportions of *E. coli* strains pre- and post-HS, we observed an overall 40-fold

increase in *E. coli* RA post-HS (one sample t-test, $p < 0.001$) with, strikingly, little change to the strain RA ratios (paired t-test, $p > 0.1$, **Figure 2a**).

As a result, the overall depth of *E. coli* genome coverage was increased from 6x to 200x for the most abundant strain, and from 0.6x to 27x for the least abundant strain (**Figure 2b**), with 72% (from 0.8%) of the least abundant strain's reference covered with 5 or more reads (a standard cutoff for minimum depth) (**Figure 2c).** As expected, we were able to produce substantially more complete assemblies using the HS-enriched data (**Table S4-S5, Supplementary Results B**).

Given that our probes were designed to select for only relatively common coding sequences from the *E. coli* pangenome, we sought to understand the degree to which non-coding regions, representing ~5% of a typical *E. coli* genome, or rarer coding regions not included in our probe design, were also found in our post-HS enriched sample data due to proximity to selected regions. We first estimated the fraction of each reference that would be covered by our probe set and then empirically determined the presence of regions not explicitly covered and their distances to probe binding sites (**Figure 2d**). Overall, we observed significant enrichment in regions up to ~300 bp away from predicted probe hybridization sites, similar to the average length of Illumina library fragments (**Figure 2d**). Despite only 83 - 86% of each reference being covered by probes, between 97% and 99.6% of the sequence from each strain's genome was enriched post-HS (**Table 1**), suggesting that *E. coli* PanSelect provides significant information even of regions not explicitly covered by probes.

## *E. coli* PanSelect provides high-level, unbiased enrichment of *E. coli* DNA and RNA within human stool samples

To determine how well *E. coli* PanSelect performed in a real world scenario, and to determine whether it would provide sufficient resolution to examine *E. coli* gene content and expression directly from stool, we applied the probes to DNA-based Illumina libraries from 191 stool samples, and a corresponding set of 130 RNA-based (RNA-Seq Illumina libraries (**Supplementary Table 6, Supplementary Table** 7). These samples were collected as part of a previously published clinical cohort study investigating the gastrointestinal tract as a uropathogen reservoir in recurrent urinary tract infection (rUTI)(Worby et al. 2022). Samples were previously shown to contain complex mixtures of low-abundance *E. coli* strains (average of 0.07% *E. coli* RA), within the diverse background of the full bacterial gut microbiomes.

As a measure of overall *E. coli* RA, we enumerated the fraction of reads from each sample, pre- and post-HS, mapping to a single *E. coli* reference genome (UTI89). Compared to pre-HS DNA libraries, post-HS libraries showed a median 158-fold increase in *E. coli* RA (range of 5 - 2,232-fold fold-increase), leading to the detection of one additional strain for every 2.8 samples. Overall, hybrid selection enriched coding sequences uniformly; the greatest fluctuation in pre- and post-HS comparisons occurred for samples with lower pre-enrichment *E. coli* abundance (**Figure 3a**). Enrichment did not change the relative ratios of strains within samples predicted to harbor multiple strains (**Figure 3b**; Wilcoxon signed-rank test, $p=0.29$ for subset of samples with two strains before and after enrichment). As expected, the increased coverage of *E. coli* allowed us to produce far more accurate and complete *E. coli* assemblies. For a sample from which we were able to obtain a near-complete reference assembly for an isolate, we assembled <1% of the genome from metagenomic data pre-HS, and > 99% post-HS (**Supplementary Results B**, **Supplementary Tables S8-S**9).

For post-HS RNA-Seq libraries, the median increase in *E. coli* content was 30-fold (range of 1.2 - 5,084 fold-increase), from a median starting abundance of 0.12%. Among *E. coli*-enriched metatranscriptomes, we observed greater *E. coli* transcript diversity; after uniform downsampling and removing three outliers where enrichment failed (Methods), we were able to observe 20-fold more unique transcripts after enrichment. On average, 693 new transcripts per sample were observed post-HS that were not detected pre-HS, out of a total of 728 transcripts observed post-HS (95%). Only 41 transcripts detected pre-HS were not detected post-HS (<1/sample), 25 of which came from a single sample (UMB12_02) with an anomalously high pre-HS RA of *E. coli* (**Figure 3c**).

A large fraction of data (median of 82%) from each post-HS metagenomic sample continued to represent organisms other than *E. coli*, consistent with our prior experience using a less complex set of HS probes on a less microbe-rich sample type (Donaldson et al. 2020). We thus sought to determine whether the non-targeted fraction of the microbiome was minimally perturbed post-HS, as we had seen before, for potential use in downstream analyses. We observed very little change in overall metagenome composition post-HS compared to pre-HS ($p > 0.05$; PERMANOVA), suggesting that *E. coli*-enriched datasets could be used for both analysis of the target organism and the full microbiome (**Supplementary Results C, Supplementary Figure** 2).

## *E. coli* gene content did not differ significantly between healthy women and those with rUTI

Our *E.coli* PanSelect method allowed us to gain insight into long-standing questions about *E. coli* populations in the GIT. One outstanding question is whether women with rUTI history harbor gut *E. coli* which are genetically distinct from that of healthy women, predisposing them to recurrent disease. Most attempts to answer this question have relied on individual cultured isolates(Schreiber et al. 2017; Bunduki et al. 2021; Terlizzi, Gribaudo, and Maffei 2017), which miss much of the diversity now recognized to exist in the gastrointestinal tract ("Structure, Function and Diversity of the Healthy Human Microbiome" 2012). Our prior work using primary stool metagenomes suggested that *E. coli* in healthy and rUTI guts are similar in their RAs and phylogroup membership (Worby et al. 2022) though the amount of *E. coli* data going into these analyses was relatively scant.

Thus, we used the enriched metagenomic data to look directly at *E. coli* gene content present in healthy participants versus those carried in the rUTI gut. Of the 13 participants in the rUTI cohort with sufficient *E. coli* content, we opted to exclude four who did not report UTI symptoms over the course of a full year. We reasoned that focusing on those who had a UTI ("recurrers") during the study window would improve sensitivity to find rUTI-associated effects (**Supplementary Table 7)**. We first performed an inter-cohort (healthy vs recurrers) comparison of overall *E. coli* gene family presence/absence profiles using output from PanPhlAn 3 (Beghini et al. 2021). An average of 4,613 *E coli* UniRef90 clusters (each roughly equivalent to a gene) was detected per sample, which was similar to the expected number of genes in the typical *E. coli* genome (**Supplementary Results D, Supplementary Figure 3, Methods**). A Jaccard index-based comparison of total gene content revealed no clear distinction between cohorts (marginal PERMANOVA; $p=1$) (**Supplementary Figure 3c),** but was significantly associated with the phylogroup of the most abundant *E. coli* strain detected within each sample (marginal PERMANOVA; $p < 0.05$) (**Supplementary Figure 3d**)**.**

To determine whether any specific gene typified *E. coli* in the GIT of women with or without a history of rUTI, we considered the distribution of all 16,729 gene families detected across samples. We tested for

differential abundance using linear mixed effects models on the 4,114 gene families that were neither core nor rare in our sample set (detected in 10-90% of samples) and found in at least one sample per cohort. We found no difference in gene carriage between cohorts (**Methods**, **Supplementary Table 10, Supplementary Results** D). We used Fisher's exact tests to examine genes that we could not fit logistic regression models for, concluding that none differed significantly between cohorts (**Methods**).

## Type 1 pili are not differentially expressed in the feces of healthy women vs. those with rUTI

The lack of gene content differences between cohorts extended to some of the best known urovirulence factors, including type 1 pili (T1P), encoded by the *fim* operon, which bind to and aid invasion of bladder epithelium during UTI (Mulvey et al. 1998; Rahdar et al. 2015), and were predicted in all but one sample across the enriched metagenomes. More recently, T1P were shown to have a role in colonization of the gastrointestinal tracts of mice, which could explain its ubiquity in our samples (Spaulding et al. 2017). However, *fim*'s role in naturally occurring *E. coli* strains in their native human gut habitats is unclear. For instance, the degree to which *fim* is expressed in the gut, whether *fim* expression levels in the gut track with rUTI susceptibility, or whether key components of *fim* regulation, studied in detail in only a few strains, are universal, are all unknown.

In order to examine the regulation and expression of *fim* in the natural gut ecology, we selected enriched samples with sufficient *E. coli* DNA content and at least five reads aligning to *fimS* (**Supplementary Table 7**, **Methods**), an invertible, non-coding promoter element directly upstream of the *fim* operon that tightly controls *fim* transcription (Abraham et al. 1985). Though the non-coding *fimS* sequence was not included in our probe design, 140 of 188 (74%) of post-HS data sets had ≥5 sequencing reads extending into the *fimS* region that we could use to determine whether *fimS* was in the "on" (presumed *fim* operon expressed) or "off" (presumed *fim* operon not expressed) orientation **(Supplementary Figure S4-S5)** .

We found that, in feces, the vast majority of *fimS* was in the "off" orientation, (**Figure 4a**), which is in contrast to reports from urine, where as much as 90% of *E. coli* cells expressed T1P, with *fimS* in the "on" orientation (Schwan and Ding 2017). Only 7.7% ± 1.9% of *fimS* in each sample was "on", regardless of rUTI history. However, there was considerable inter-sample variation (**Supplementary Figure S5B**), with some samples having ≥10% of *fim* in the "on" orientation, and some having none. Using the set of samples that also had sufficient *E. coli* RNA content (**Supplementary Table 7, Methods**), we confirmed that changes in *fimS* regulatory status were proportional to changes in expression of the downstream *fim* operon (p=0.0004; **Methods**), as expected (Abraham et al. 1985; Schwan and Ding 2017; Subashchandrabose et al. 2014).

## Broad shift towards expression of aerobic metabolic genes in the rUTI gut

Due to recently identified differences in GIT microbiota composition between rUTI and healthy subjects (Worby 2022) and the importance of transcriptional state of UPEC in the urinary tract (Schreiber), we explored expression-level differences between healthy and recurrer gut *E. coli* more broadly. To do this, we chose samples with sufficient *E. coli* coverage in both PanSelected RNA and DNA datasets for uniform comparisons (**Supplementary Figure 6**, **Supplementary Table 7**, **Methods**), which enabled expression comparisons for a total of 2,182 genes (**Supplementary Table 11**; **Methods**). Using linear mixed effects models (Y. Zhang et al. 2021), with depth-based weights to control for inter-sample variation in *E. coli* coverage (**Supplementary Figure 7**) and false discovery rate correction, we identified four genes that were significantly differentially

expressed (**Supplementary Table 11, Figure 4b**): *ydeE*, encoding an arabinose exporter, was significantly over-expressed in rUTI *E. coli* (p=0.048); and *fucI*, encoding L-fucose isomerase; *aspA*, encoding aspartate ammonia lyase; and *frdA*, encoding fumarate reductase subunit, were all significantly under-expressed in rUTI *E. coli* (all p=0.035). Just missing our threshold was *dcuA,* encoding a C4-dicarboxylate transporter capable of importing fumarate in exchange for succinate, which was under-expressed in rUTI E. coli (p=0.054). Taken together, reduced expression of enzymes for conversion of aspartate to fumarate (*aspA*), import of fumarate (*dcuA*) and reduction of fumarate (*frdA*) indicate a shift away from utilization of fumarate as a terminal electron acceptor for anaerobic respiration, while decreased isomerization of fucose (*fucI*) and increased efflux of arabinose (*ydeE*) suggest differences in metabolism of secondary sugars in recurrent vs healthy gut *E. coli*.

Given that multiple genes from the same pathways were significant in our gene-based analysis, we next sought to explore whether other functionally related groups of genes were differentially expressed between cohorts. To group genes into related sets that take into account their known functional and regulatory relationships, we used i) transcription factor (TF) networks as per RegulonDB (Huerta et al. 1998; Gama-Castro et al. 2008), which classify TF activity as either activators (noted by "+") or repressors (noted by "-"), and ii ) KEGG Pathways and Modules (Kanehisa and Goto 2000). Roughly two-thirds (1,417 of 2,182) of genes included in DE testing belonged to at least one of 284 related gene sets. Using Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005), we determined that 22 of these related gene sets were over-represented among genes found to be more highly expressed (11 sets) or more lowly expressed (11 sets) in rUTI gut *E. coli* compared to those in healthy guts (p<0.05) (**Supplementary Table 12**, **Figure 4b-c**).

Most (233 of 332) genes within these related sets were controlled by two master regulators, ArcA and FNR, both involved in coordinating functions related to anaerobic metabolism. Compared to healthy gut *E. coli*, genes activated by FNR+ were less expressed and genes repressed by ArcA- were more highly expressed in *E. coli* residing in the rUTI gut. Less expressed genes under FNR+ control included several that were also identified in the individual gene-based analysis (*aspA*, *frdA* and *dcuA*); their expression is also controlled by NarP and NarL, which together regulate adjustments to nitrite and nitrate availability. Together, the differential expression of genes in these four regulons suggested that the rUTI gut ecology is richer in oxygen and/or nitrate, requiring rUTI gut *E. coli* to upregulate functions for respiring nitrate and/or oxygen and downregulate functions for respiring fumarate and nitrite.

Nearly all of the other 17 enriched sets were part of regulatory networks that were at least partly nested within these main regulons, and many of these had nested relationships with each other (**Figure 4d**). Overall, these gene sets pointed to differences in energy metabolism, amino acid biosynthesis, carbon catabolism and stress responses for *E. coli* living in these two different gut habitats. rUTI gut *E. coli* had higher expression of: (i) aspects of the TCA cycle (KEGG modules for the dicarboxylate-hydroxybutyrate- and reductive citrate-cycles); (ii) amino acid biosynthetic machinery, including the KEGG pathways for C5-branched dibasic acid metabolism and 2-oxocarboxylic acid metabolism, as well as genes under the control of the TyrR- repressor, encoding aromatic amino acid biosynthetic machinery; (iii) Rob- repressed genes involved in $\beta$-oxidation fatty acid catabolism; (iv) reactive oxygen and nitrogen detoxification via the SoxR+ regulon; (v) ribosomes; and (vi) genes in the the only regulon not obviously connected into the three main regulons - the YhaJ+ regulon, which has unclear, strain-specific roles related to virulence (Connolly et al. 2019). As compared to healthy gut *E. coli*, rUTI *E. coli* had lower expression of: (i) additional regulons involved in fumarate respiration (DcuR+, FhlA+, FlhDC+); (ii) AraC+ controlled genes that are also part of the pentose/glucuronate interconversion pathway for catabolism of arabinose and other sugars; and (iii) machinery for redox balancing during fermentation as part

of the YdeO+ and AppY+ regulons. All together, these results suggest a broad shift of *E. coli* towards aerobic respiration in the rUTI gut (**Figure 5**).

Conditions that favor oxygen/nitrogen-based respiration during growth in the gut are a known driver of Proteobacteria blooms, which have been shown to occur during infection and at times of intestinal inflammation (Winter et al. 2013; Winter, Lopez, and Bäumler 2013). Given the broad evidence for a switch from fermentative to a more oxygen and/or nitrate-based respiratory metabolism in rUTI gut *E. coli*, including the increased expression of ribosomal proteins, we hypothesized that *E. coli* in the rUTI gut may be growing at a faster rate than in the healthy gut. Using SMEG (Emiola, Zhou, and Oh 2020), which determines the difference in coverage between the origin and terminus of replication as a proxy for growth rate (Emiola, Zhou, and Oh 2020), we observed no significant difference in estimated growth rate between the cohorts (p=0.54), with a numerically greater growth rate in the healthy cohort (**Figure 4e**).

# Discussion

Here, we provide a framework to generate a set of probes to enrich a low abundance, diverse bacterial species, where the diversity and genetic content within a sample are not known a priori, but are expected to be diverse. We have shown the efficacy of *E. coli* PanSelect, a set of ~900,000 probes which cover the large pangenome of *E. coli*, to enrich for *E. coli* DNA and RNA in both a controlled mock community and for biological samples containing complex mixtures of diverse representatives. Applied to short fragment libraries from DNA and RNA extracted from human stool where *E. coli* are at very low relative abundance, *E. coli* PanSelect increased the RA of *E. coli* DNA and RNA 158-fold and 30-fold, respectively, without evidence of skewing of the *E. coli* or non-*E. coli* data fractions.

In benchmarking using metagenomic samples from human stool samples, *E. coli* PanSelect enabled detection of novel strains that could not be observed without HS, as well as a dramatically increased ability to assemble low-abundance *E. coli* strains. The striking enrichment of *E. coli* with HS allowed for a wide range of analyses that were previously cost-prohibitive due to the requirement for very deep sequencing. To obtain the same sequencing depth of *E. coli* without HS, we would need to generate enough sequencing reads to cover the human genome 150 times over. We assembled large contigs of *E. coli* sequence from post-HS metagenomic data, demonstrating the ability to obtain near-complete *E. coli* references from clinical samples without the need for culture. Even in mixed-strain samples, we were able to assemble large fractions of species-specific genes on contigs long enough to provide genomic context. This ability to see the genomic context surrounding strain-specific genes, even in low-abundance strains, is a powerful capability enabled by *E. coli* PanSelect. Tools that specifically attempt to disentangle strain genomes in metagenomic assemblies, such as DESMAN (Quince et al. 2017) and STRONG (Quince et al. 2020), may provide an even greater increase in our ability to assemble multiple genomes from this data.

In addition to using *E. coli* PanSelect to analyze the *E. coli* content, it can be used to analyze non-*E. coli* content, which was only minimally biased after HS. This suggests that a single *E. coli* PanSelect sequencing run can be used to thoroughly investigate both the *E. coli* and its metagenomic background, saving time, money, and precious sample material. *E. coli* PanSelect of metatranscriptomic libraries also allowed for a new range of analyses of the *E. coli* transcriptome. In benchmarking, we were able to observe 20-fold more

unique transcripts after enrichment. We demonstrated that *E. coli* PanSelect can provide the sensitivity required to analyze and compare otherwise cryptic gene expression patterns of low-abundance *E. coli* in the human gut.

We applied *E. coli* PanSelect to enrich the *E. coli* fraction in a large set of human stool samples from women with and without history of rUTI. The gut microbiome is an important pre-infection reservoir for UTI-causing uropathogenic *E. coli* (UPEC). In a genome-wide analysis of HS-enriched transcriptomes from rUTI and healthy stool samples, we observed differential expression of *E. coli* genes and key TF regulons, (**Figure 5**) including the ArcA and FNR regulons, indicating a shift from fermentative to aerobic metabolism in rUTI gut *E. coli* in response to an overall change in their environment. A facultative anaerobe, *E. coli* carries cellular machinery for respiration on multiple electron acceptors, but will shift its metabolism towards respiration using nitrate and oxygen when possible(Ingledew and Poole 1984). The gut environment is normally deficient in nitrate and oxygen, and colonized by fermentative, obligate anaerobes. However, during periods of inflammation, the intestinal epithelium produces antimicrobial reactive oxygen and nitrogen species (RONS), resulting in changes to microbiome composition based on RONS tolerance. Facultative anaerobes, like *E. coli*, are more resistant to oxidative stress, and are even able to co-opt secondary products of RONS, such as S-oxides, N-oxides, and nitrate, as terminal electron acceptors for respiration (Winter, Lopez, and Bäumler 2013). Our data suggest that *E. coli* in the rUTI gut is likely responding to the presence of RONS and availability of exogenous electron acceptors in the environment, resulting in decreased expression of genes repressed by the nitrogen response regulators NarL and NarP, while increasing expression of nitric oxide and and superoxide response regulons and cellular machinery for aerobic respiration (**Figure 6**). Partially due to similar shifts in the gut environment, chronic inflammatory diseases, such as inflammatory bowel disease, are often characterized by an outgrowth of facultative anaerobes (Winter, Lopez, and Bäumler 2013). Interestingly, we saw neither an inflammation-driven outgrowth of *E. coli* in our prior analysis of unenriched stool metagenomes, nor an increase in growth rate within the HS-enriched dataset. Possibly, the increased growth potential provided by RONS is counterbalanced by antimicrobial activity, or small enough that it can only be detected at the transcriptional level.

Despite evidence for potential environmental similarities between the rUTI and IBD gut, there are key differences in *E. coli's* response. In addition to differences in growth rate, IBD-associated *E. coli* are characterized by expression of pro-inflammatory pilins (including type I pili), which were notably unchanged in our analysis (Y. Zhang et al. 2022). Furthermore, where growth on mucin-derived metabolites, including fucose, has been described for Crohn's disease-linked Adherent Invasive *E. coli* (AIEC), expression of a gene for growth on fucose and other secondary sugars was one of the most downregulated in rUTI vs healthy gut E. coli in our analysis (S. Zhang et al. 2022). Interestingly, Hadjifrangiskou et al. found that UPEC requires aerobic respiration to cause UTI, with oxygen scavenging by the conserved respiratory quinol oxidase cytochrome *bd* within host cells promoting a shift towards aerobic glycolysis(Beebout et al. 2022), and Mobley et al. found that nitrate reductase is a competitive virulence factor in UTI(Martín-Rodríguez et al. 2020).

There were several limitations in the design of our probe set. First, we excluded a substantial fraction (63%) of orthogroups from our probe design in order to i) exclude orthogroups that hybridized to other common gut bacteria, ii) reduce computational runtime for the CATCH algorithm, and iii) obtain a manageable final number of probes to synthesize; however, most excluded orthogroups were rare and contained few genes. Cross-hybridization with other community bacteria is an important factor that must be taken into account in probe design, considering an organism's ecology and community context. Despite the exclusion of these orthogroups from the probe design process, the vast majority of individual *E. coli* genes (98%) were included in

the probe design. Importantly, the exclusion of a gene from probe design did not necessarily mean that the probe set would not capture the gene, as the excluded gene could still share sequence homology or be found in close proximity to genes that were included in the probe design. For example, although the fimS promoter region was not included in probe design, we obtained sufficient enrichment of regions surrounding the *fimS* invertible promoter to be able to confidently identify its orientation (**Supplementary Figure 4**).

Second, our probe design was limited by the set of assemblies in the NCBI database at the time of probe design (2017), which may not represent an accurate view of *E. coli* in the target sample. However, our methodology could be used to update the probe design with an expanded database. Furthermore, sequenced strains may be biased towards *E. coli* found in certain environments, such as clinical ones, and may not be an accurate representation of overall *E. coli* diversity. However, our data showed that probes did cover some genes and orthogroups that were not specifically part of our design set, so we may be able to successfully enrich genes from unsampled genomes.

Finally, these probe sets were designed with short-read sequencing libraries in mind; however, long read technologies are becoming increasingly used for assessing and assembling metagenomic samples(Bertrand et al. 2019). Whether these probes could be directly applicable to long, native strands of metagenomic DNA is currently unknown. Nevertheless, a combination of hybrid-captured short-read sequencing and native long-read sequencing may be a fruitful avenue for metagenomic assembly of low-abundant organisms.

Further experiments are also needed to optimize sample pooling protocols, as the lower level of enrichment in the metatranscriptomic data (30-fold) as compared to the metagenomic data (158-fold) could be due to the fact that our RNA data was processed in larger pools (24 samples per pool) as compared to the DNA samples (8 samples per pool). Furthermore, in assessing the effectiveness of *E. coli* PanSelect, the interpretability of our fold-change estimates were impacted by our use of relative, rather than absolute abundance. Because of this, samples with lower starting quantities of *E. coli* had systematically greater apparent fold changes. Likely due to this effect, the increase in performance of E. coli PanSelect on human stool samples, relative to enrichment of the mock community, was attributable to the reduced initial quantities of *E. coli* in the stool samples.

Although we showcased the use of the *E. coli* PanSelect probes on gut metagenomes, broader applications include detection of low-abundance *E. coli* in other specimens and environments, including the study of other body sites such as the skin (Ranjan et al. 2017), catheter, or bladder (Rosen et al. 2007; Flores-Mireles et al. 2015; Pietrucha-Dilanchian and Hooton 2016), or applications such as pathogen surveillance at hospitals (Hilty et al. 2012) and in the food industry (Allard et al. 2018). In addition to applications to *E. coli,* our pan-genome-based approach for probe design should also represent a cost-effective alternative to ultra-deep metagenomic sequencing to examine other low-abundance bacteria, particularly those that are well-studied with numerous representative reference genomes; and are genetically complex and diverse.

## Conclusions

Computationally designed *E. coli* PanSelect probes targeting the entire pan-genome provide a promising method to enrich sequences derived from low-abundance, highly diverse species in complex

metagenomic and metatranscriptomic communities. Using this method, we successfully enriched *E. coli* sequence from both a mock mixture of four diverse strains of *E. coli* in 99% human background, as well as in human stool samples. We found that *E. coli* was enriched, on average, 158-fold (for DNA) and 30-fold (for RNA) post-HS. With our HS-enriched dataset, we specifically investigated the *in-vivo* regulation and expression of type1 pili, and determined differences in *E. coli* transcriptional state between *E. coli* resident within healthy individuals and rUTI sufferers, suggesting increased RONS signaling in the guts of rUTI sufferers. We believe that our approach to probe design will provide a valuable tool to push forward the study and understanding of *E. coli* in metagenomic contexts, including giving us a better *in vivo* understanding of how *E. coli* strains differ in their gene content and expression, respond to antibiotics, and dynamically change through time and disease. This tool will accelerate new biological discoveries and may aid in developing alternative therapeutic options for diseases caused by *E. coli* and other bacteria.

## Methods

### Probe design and synthesis

E. coli *reference genome selection, clustering, and annotation.* All *E. coli* and *Shigella* (hereafter collectively referred to as *E. coli*) complete genomes were downloaded from NCBI RefSeq as of June 2017 (total of 295 genomes). To supplement the Refseq collection with additional diverse genomes, 3,141 publicly available, high quality (L50 < 20) genomes of *E. coli* that were listed in the NCBI Pathogen Detection database were downloaded from GenBank from July to August 2017 (**Supplementary Table 1**). In order to remove nearly identical Genbank genomes, we performed k-mer based clustering. All 3,141 Genbank genomes were k-merized (using 23mers) with the StrainGST “kmerize” tool from StrainGE(van Dijk et al. 2021), then their pairwise all-vs-all Jaccard similarities were calculated. Single-linkage clustering was performed on these similarities at a 95% threshold to construct 1,485 genome clusters, of which 1,124 contained a single genome and 361 contained two or more Genbank genomes. Of the 361 multi-genome clusters, 67 included a genome identical to one of the RefSeq references, and were discarded. For the remaining 294, the largest reference was chosen as a representative for the cluster. Our final set included these 294 representatives, the 1,124 singletons, and the 295 RefSeq genomes. This final set of 1,713 genomes represented a large, diverse collection, with references from all eight major clades of *E. coli* (**Supplementary Table 2; Figure 2; Supplementary Figure 1**), as determined by the tool ClermonTyping (Beghain et al. 2018), and 515 distinct multi-locus sequence types, as determined by the tool mlst (https://github.com/tseemann/mlst).

The 1,713 genomes were then uniformly re-annotated with the Broad Institute prokaryotic genome pipeline(Valentino et al. 2014; Schreiber et al. 2017). Genes were clustered into orthogroups using SynerClust(Georgescu et al. 2018), which resulted in a total of 174,584 orthogroups containing 8,334,026 total genes. As the computational time to analyze all these orthogroups using CATCH was prohibitive, we filtered out rare orthogroups found in fewer than three genomes, leaving 64,146 orthogroups (containing a total of 8,165,358 genes). In order to ensure that our set contained all potential instances of key genes important in clinically relevant *E. coli*, we retained all instances of orthogroups containing instances of 59 Pfam domains of interest, obtained from a curated list (**Supplementary Table 13**). Using this list, we added back a total of 2,434 orthogroups (3,479 genes) that were found in fewer than three genomes. Our final set contained 64,580 orthogroups comprising 8,168,837 genes. In order to reduce design constraints in CATCH, thereby decreasing computational cost and time, we further clustered each orthogroup using UCLUST(Edgar 2010), with an 80% identity threshold. This generated one or more clusters of genes within each orthogroup, in which all cluster

members had ≥80% identity to one other. This generated 87,218 gene clusters from the 64,580 orthogroups. These gene clusters were the input for CATCH probe design.

*Probe design and filtering.* CATCH(Metsky et al. 2019) was run to generate probes for each gene cluster using the following parameters: 2 bp mismatch allowed; 25 bp cover extension; expand "N" to ACGT; 30 bp island of exact match; 60-75 bp length; no more than 8 total mismatches. In addition, three non-*E. coli Enterobacteriaceae* assemblies were used as part of the CATCH algorithm to "blacklist" probes that matched off-target sequences: a *Citrobacter ,* a *Salmonella*, and a *Klebsiella* genome (Genbank accessions GCA_000648515.1, GCF_000195995.1, and GCA_000240185.2, respectively). Representatives from these three genera were chosen as they represent well-characterized *Enterobacteriaceae*, closely related to *E. coli*; blacklisting them was thought to help to improve the specificity of the probe set to *E. coli* vs. other similar organisms. Duplicate probes were removed, resulting in a total of 911,618 unique probe sequences.

We also used an additional set of filters to remove remaining probes that might capture off-target sequences. We used the blastn tool from BLAST+(Altschul et al. 1990; Camacho et al. 2009) to search probe sequences for homology against the NCBI prokaryote reference genome database (downloaded in October 2017), using the following parameters: max_target_seqs 30; evalue 1e-5; qcov_hsp_perc 80; perc_identity 80. Using these results, we removed probes that had matches of 65bp or more to: 1) ≥100 references in the database (1,798 probes removed); 2) *Bacteroidetes* references (2,470 probes removed); 3) *Firmicutes* references (14,935 probes removed). We were left with a final total of 892,415 probes that were unlikely to hit other commonly found bacterial species in the human gut.

*In silico probe set validation.* In order to verify that our probe set would actually be able to capture the vast majority of genes in the *E. coli* pangenome, we used blastn from BLAST+ to query our probe sequences against the entire pan-genome from our set of 1,713 references, which included genes that had previously been filtered out at the probe design stage. We used probe sequences as queries for blastn with the following parameters: max_target_seqs 30; e-value 1e-5; qcov_hsp_perc 80; perc_identity 80. We retained alignments with >65 bp length and no more than 8 mis-matches in the entire alignment. The probe set was considered to capture a gene if one or more probes met these criteria for a given gene.

*Probe synthesis.* For this study, the probes were synthesized by Roche, though the probe set was not specifically tailored to their technology and could be synthesized by other manufacturers. All probes could be synthesized, although 330,387 (37%) probes had one or more bases truncated from the 3' end. The average number of bases trimmed per probe was 1.27 ± 2.16. Only 5,423 probes had 10 or more bases trimmed. All of the most highly truncated probes had low nucleotide complexity, primarily due to long stretches of homopolymers. As these changes were unlikely to affect the performance of the probe set as a whole, we used this slightly modified probe set in our experiments. The average probe length after synthesis was 73.7 ± 2.2 bp.

## Analysis of four-strain *E. coli* mock community

*Library construction and sequencing.* We used a previously reported mock community (van Dijk et al. 2021), which included 99% human DNA and 1% *E. coli* DNA. The *E. coli* DNA was composed of: i) H10407 (clade A; 80%), ii) E24337A (clade B1; 15%), UTI89 (clade B2; 4.9%), and Sakai (clade E; 0.1%). The Nextera XT library construction kit (Illumina) was used to generate sequencing libraries following the manufacturer's recommended protocol. To enrich *E. coli* sequences in the mock library (~100 ng into the HS

reaction), we performed HS using a Roche SeqCap EZ Hypercap kit with our designed custom capture probe set. Hybridization and target capture followed the SeqCap kit instructions except that we diluted the probe pool 1:2 before use, and substituted custom Nextera adapter blocking oligonucleotides(Metsky et al. 2019) for the SeqCap HE Universal adapter and index blocking oligonucleotides. After hybridization (18 h), bead capture and washes, we performed 10 cycles of PCR with generic universal Illumina P7 and P5 primers. The final libraries were quantified by Qubit fluorometry (Thermo Fisher Scientific), and the size distribution was analyzed by TapeStation electrophoresis (Agilent) prior to Illumina sequencing. Then, pre- and post-HS libraries were sequenced on an Illumina HiSeqX, generating 21,460,598 and 75,576,717 paired-end 151 bp reads for the pre- and post-HS libraries, respectively.

*Downsampling and quality control.* Pre-HS and HS libraries were downsampled to equal depth (approximately 20,000,000 paired-end reads) with Picard-Tools prior to analysis (https://broadinstitute.github.io/picard/). Quality control was performed with FastQC(Andrews and Others 2010) and MultiQC(Ewels et al. 2016). Due to observed heightened rates of PCR duplications in the HS libraries, both HS and pre-HS sequencing datasets were *de novo* deduplicated with FastUniq(Xu et al. 2012).

*Calculation of enrichment, bias, and coverage levels*. We assessed the enrichment of total *E. coli* within the metagenome with a one-sample t-test of the log2 fold-change for all four strains. In order to examine enrichment bias between different strains within a sample, we compared the ratios of the RAs of individual strains in HS metagenomes to the ratios in pre-HS metagenomes with paired t-tests. RAs and depth of coverage for each of the four strains were estimated using StrainGST(van Dijk et al. 2021). We first built a StrainGST database consisting of just the reference genomes of the four strains of *E. coli* in the mock mixture–H10407, UTI89, Sakai, and E24377A–all downloaded from NCBI Genbank. Then, we k-merized both pre- and post-HS data and ran StrainGST (without k-mer fingerprinting) against the database to determine the RAs and depth of coverages for all four strains.

Coverage levels for each of the four strains were obtained by aligning downsampled and deduplicated data with Bowtie2(Langmead and Salzberg 2012) with default parameters to a concatenation of all four strains' reference genomes. Only properly paired aligned reads with a minimum mapping quality (MQ) of 5 were retained with samtools (http://www.htslib.org/). This filtering was done to exclude reads and regions of the genomes where reads aligned equally well to different strains, with the goal of reducing bias in less abundant strains due to sequence homology to sequences deriving from the more abundant strains. Then, coverages of MQ≥5 reads were assessed using Bedtools (https://bedtools.readthedocs.io/en/latest/).

*Assembly of* E. coli *mock community.* Downsampled and deduplicated data were used to generate metagenomic assemblies. First, pre- and post-HS data were digitally normalized with the program khmer(Crusoe et al. 2015). Then, downsampled data were processed with Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove leftover adapter content. Then, as the background was 99% human DNA, the normalized data were cleaned of human data using Bowtie2 (with the "very sensitive" flag) against the hg38 reference. Reads that did not align to the human genome were then assembled with MetaSPAdes (Nurk et al. 2017) with default parameters. Contigs and scaffolds <1 kb were removed. Then, GAEMR (http://software.broadinstitute.org/software/gaemr/) was used to assess assembly metrics and determine the taxonomy of each contig/scaffold.

To calculate strain genome coverage in metagenomic assemblies, we used BLAST+ to search for reference genome sequence in assembled *E. coli* contigs >1kb. Hits >1kb and with >90% identity were included in the final coverage calculation. To identify strain-specific genes in each strain in the mock community, we used an all vs. all BLAST+ approach to look for homologous genes between all reference

genomes. Any gene that did not match a gene in any other strain (E-value <1e-10) was considered strain-specific.

*Comparison of actual to expected probe coverages.* To determine probe hybridization sites on each strain's reference genome, all probe sequences were aligned using Bowtie2 to each of the four reference genomes, individually. The intervals where probes aligned were designated as putative probe hybridization sites. Bedtools was used to calculate the probe coverage of the four reference genomes, as well as coverage of the reference genomes by the pre-HS and post-HS metagenomes. A custom R script was used to determine the correlation between probe coverage of regions of the references and enrichment in read coverage with HS.

## Sequencing of clinical stool samples

Total nucleic acid was extracted from stool metagenomes collected during the Urinary Microbiome (UMB) project, as previously described (Bioproject PRJNA400628) (Worby et al. 2022). The total (~100 µL) nucleic acid per sample was divided into equal aliquots for DNA and RNA sequencing. 130 samples were used for metagenomic HS, as well pre and post-HS metatranscriptomics, while 61 were used for just metagenomic HS sequencing, for a total of 191 samples (**Supplementary Table 7**). The RNA aliquots were treated with DNase and Agencourt AMPure beads for a SPRI clean-up. The 130 non-enriched RNA libraries were sequenced with an Illumina NovaSeq, generating an average of 56.0 ± 56.7 million paired-end 151 bp reads. The 191 non-enriched DNA libraries were sequenced with a combination of Illumina HiSeq 2500 and HiSeq X10 technologies, as previously reported(Worby et al. 2022).

*Enrichment and sequencing of* E. coli PanSelect *libraries.* We enriched *E. coli* sequences from both DNA and RNA samples with multiplex solution HS using a Roche SeqCap EZ Hypercap kit with our designed custom capture probe set. DNA-Seq libraries were processed in pools of 8 libraries (~200 ng each). RNA-Seq libraries were prepared as pools of multiplex RNAtag-Seq libraries from 24 RNA samples(Shishkin et al. 2015; Bhattacharyya et al. 2019) and amplified by 14 cycles of PCR to generate at least 100 (mean 140) ng of each library pool for HS (one 24-plex pool per reaction). Hybridization and target capture were performed as for the mock community above. All post-HS libraries were run on an Illumina NovaSeq, generating an average of 9.6 ± 3.9 million paired-end 151 bp reads for post-HS DNA libraries, and 10.2 ± 10.7 million paired-end 151 bp reads for post-HS RNA libraries. Three samples for which HS DNA sequencing failed (low post-QC read depth) were removed from analysis (UMB13_09, UMB08_04, UMB24_08).

Quality assessment. Quality of sequencing files was assessed with FastQC and MultiQC. We did not *de novo* deduplicate DNA reads, as we observed far lower PCR duplication in the HS DNA libraries than for the mock community. We processed DNA and RNA data with KneadData (https://huttenhower.sph.harvard.edu/kneaddata/) to remove low quality sequence, adapter content, and human contamination.

## Benchmarking *E. coli* PanSelect enrichment using stool samples

*Enrichment estimation.* We used the same kneaddata QC workflow as we used for analysis of the mock community. For both DNA and RNA, per-sample *E. coli* enrichment was estimated as the difference in the proportion of reads aligned using Bowtie2 v. 2.3.4.3(Langmead and Salzberg 2012) to the UTI89 reference genome between pre-HS and post-HS sample pairs. Strain-level composition was estimated with StrainGST. In order to match up strains across pre- and post-HS sample pairs, we matched pre- and post-HS strains

assigned to the same reference using StrainGST (n=184 strain pairs). Of the remaining strains, we matched pre- and post-HS strains assigned to the same phylogroup using StrainGST (n=14 strain pairs). If a strain in one sample matches to two strains from the same phylogroup in its paired sample, these were removed from analysis (n=16 strain trios). The remaining strains had no match in its paired sample (n=55).

For tabulating and visualizing (**Figure 3b**) enrichment of transcripts, post-HS/pre-HS sample pairs with >1 million post-QC RNA read pairs were selected and downsampled to 1 million reads. Transcript relative abundance was estimated in counts per million (CPM), with 10 CPM used as the detection threshold for both pre-HS and post-HS. Outlier samples were identified by performing t-tests on studentized residuals from an ordinary least squares regression between log10-transformed pre-HS and post-HS total *E. coli* relative abundance (benjamini-hochberg FDR correction, α=0.1). Using this procedure, three RNA samples were identified as enrichment outliers and removed from transcript counts.

*Background metagenome analyses.* Taxonomic profiles for all metagenomes (pre- and post-HS) were calculated using MetaPhlAn3(Beghini et al. 2021), run on metagenomes downsampled to a maximum depth of 3.5 Gb. *E. coli* and *Shigella* RAs were removed from the taxonomic table and RAs of the remaining taxa were renormalized to sum to 100%. Bray-Curtis Dissimilarity values were calculated with the Python package *scipy* v.1.7.1 ( https://scipy.org/). PERMANOVA was implemented by the *adonis2* function from the R package *vegan* (https://cran.r-project.org/package=vegan).

Sequencing and assembly of *E. coli* isolate for benchmarking of metagenomic assembly. We sequenced a single *E. coli* isolate from Participant 8, timepoint 2, using a combination of Illumina and Oxford Nanopore Sequencing. Sequencing and hybrid assembly was performed as previously published (Bronson et al. 2021). This assembly has been submitted to NCBI Genbank with accession GCF_011751425.1.

*Metagenomic assembly.* Metagenomic assemblies were generated for ten samples using all available sequencing data. Digital normalization, adapter trimming, assembly, and calculation of assembly metrics were performed as for the mock community data. Metrics were calculated based on only contigs assigned to *E. coli* using blastn. After assemblies were produced, a binning program, MetaBat2(Kang et al. 2019), was used to produce metagenome-assembled genomes (MAGs). MAGs were analyzed with CheckM (Parks et al. 2015) to determine taxonomy and assembly completeness for MAGs that were classified as *Enterobacteriaceae* by CheckM.

## Analyses of full set of enriched stool samples

*Comparison of gene content between cohorts*. We calculated gene family presence/absence profiles for enriched metagenomes with PanPhlAn3 (v. 3.1), run with the UniRef90 *Escherichia coli* pangenome generated on Nov 2, 2020 by the Segata Lab (Beghini et al. 2021) (**Supplementary Table 14**). Samples were filtered based on evenness of *E. coli* coverage with the profiling workflow contained within PanPhlAn (*panphlan_profiling.py)*, run with the 'very sensitive' parameters (params here) (**Supplementary Table 7**).

We used omnibus and per-gene tests to test for differences in gene content between the recurrer and healthy cohorts. First, we quantified differences in overall gene content profiles between samples with the Jaccard Index and tested for differences between cohorts with PERMANOVA. We used the *adonis2* implementation of PERMANOVA (R package *vegan*) to assess the marginal effect of cohort, controlling for subject. In order to test for differential abundance of individual gene families, we used logistic regression, followed by Fisher's exact tests for genes where we had issues with model fitting. For logistic regression, we selected the genes that were found in between 10% and 90% of all samples, and detected in at least one

sample from both the recurrer and healthy cohorts (4,114 gene families). We fit models with the function *pglmm* from R package *phyr* v.1.1.2(D. Li et al. 2020), using random intercepts for subjects and a phylogenetic covariance structure. For the phylogenetic covariance structure, we used the most abundant strain identified within each sample by StrainGST. No gene families were significant after benjamini-hochberg false discovery rate correction. To evaluate the remainder of gene families for which we had issues with model fitting, we calculated Fisher's exact tests to find associations between presence and cohort. Because Fisher's exact test does not account for distribution of samples across different subjects, we tested for association at both the sample and subject carriages levels. For the subject level, we called a gene present in a subject if it was identified in at least one sample in a subject series. For both sets of models, the most significantly differentially abundant genes were those that were already shown to be insignificant by logistic regression, so we concluded that the remaining genes were unlikely to be significantly differentially abundant.

*fimS structural variation profiling.* We estimated the fraction of fimS in the "on" orientation within enriched metagenomes by aligning reads to a reference containing both a copy of the UTI89 genome with fimS in the 'on' orientation and a copy with *fimS* in the 'off' orientation. We required at least 5 reads to align specifically to one of these two copies of the *fimS* region (**Supplementary Table 7**), in order to be able to determine fimS orientation. We used Bowtie2 (version 2.3.4.3, default parameters, MQ >5) for alignment of metagenomes.

For many samples, we observed a mixture of alignments to both the 'on' and 'off' references, indicating subpopulations of *fim*-expressing (piliated) and non-expressing (smooth) *E. coli* within the same samples. We used the proportion of reads aligning uniquely to the 'on' orientation as an estimate for the size of the piliated population. Because there was significant inter-sample variation in *E. coli* RA (and thus *fimS* coverage), we were able to estimate the RA of the piliated population more accurately in some samples than others. In order to reflect this variation in our estimates of average *fimS* activation, we used weighted averages and regressions with weights proportional to per-sample *fimS* coverage ('on' + 'off').

*Selection of* E. coli *coverage threshold for* fim *and differential expression analysis.* Post-HS RNA and DNA samples were aligned to the UTI89 reference genome with bwa-mem v.0.7.17-r1188 (H. Li and Durbin 2009), and alignments were counted with FADU v.1.8 (Chung et al. 2021) (run with parameters: *-M -p*). We then filtered samples based upon total post-HS DNA and RNA *E. coli* content, using a threshold of 60,000 UTI89 coding sequence-aligning reads. We determined this threshold by examining the relationship between RNA *E. coli* content and observed RNA transcript diversity. We used 'relative abundance-weighted transcript diversity' as a metric, defined as the sum of average relative abundances of all unique transcripts detected in a sample. We selected our threshold of 60,000 *E. coli* reads, because this value allowed us to observe unique transcripts representative of 80% of the average *E. coli* transcriptome (**Supplementary Figure 6B**).

*Relationship between fimS orientation and fim operon expression.* For testing the relationship between *fimS* orientation and *fim* operon expression, we used the subset of samples for which we could both estimate *fimS* orientation (5 *fimS*-aligning reads) and measure gene coverage and expression (60,000 *E. coli* DNA and RNA reads) (**Supplementary Table 7**) We quantified *fim* operon coverage and expression as the sum of coverage of all genes in the *fim* transcriptional unit (*fimAICDFGH*). We used mixed effects models for metagenome-controlled differential expression testing, as described in(Y. Zhang et al. 2021). Models were fit with the form: $fim\ RNA(log\ cpm)\ \sim fim\ DNA(log\ cpm)\ +\ fimS(log\ \%)\ +\ (1|subject)$ with the function *lme* from the R package *nlme*. To control for variation in *E. coli* content between samples, we used weights proportional to sample *E. coli* content: *weights=varFixed(~ (1/DNA_ecoli + 1/RNA_ecoli)).*

*Selection of samples and genes for differential expression testing.* For differential expression testing between cohorts, we filtered samples based upon total coverage of *E. coli* (60,000 DNA and RNA reads, see above) and coverage evenness, as determined by the profiling workflow contained within PanPhlAn3, run with 'very sensitive' parameters on alignments to the UTI89 reference (**Supplementary Table 7)**. We used genes that were classified as 'single copy' within metagenomes by PanPhlAn. For differential expression testing, we selected a subset of genes present in at least 50% of all metagenomes from both the recurrer and healthy cohorts (15 samples/cohort) and expressed in at least 15 metagenome/metatranscriptome pairs across the full study. Gene presence was defined as coverage by at least 20 reads in both metagenomes and metatranscriptomes, as 20 reads in the sample with the 15th-greatest RNA *E. coli* content corresponds to a non-zero expression rate (>2 reads) at our threshold for sample inclusion based on *E. coli* content (60,000 reads).

*Between cohort differential expression testing.* We used mixed effects models for per-gene metagenome-controlled differential expression tests(Y. Zhang et al. 2021). We fit models of the form: $RNA(log\ cpm)\ \sim DNA(log\ cpm)\ +\ cohort\ +\ E.\ coli\ RA\ (log\ \%)\ +\ (1|subject)$ in with the function *lme* from the R package *nlme*. To control for variation in *E. coli* content between samples, we used weights proportional to sample *E. coli* content: *weights=varFixed(~ (1/DNA_ecoli + 1/RNA_ecoli)).* We included *E. coli* RA as a fixed effect based on model comparisons using Akaike's information criterion (AIC). We quantified the coverage of genes within metagenomes and metatranscriptomes as copies per million scaled by *E. coli* content (CPM), and used a natural log transformation for variance stabilization of RNA and DNA CPM values, as well as *E. coli* RAs. RNA zeroes were replaced before transformation with a gene-specific pseudocount equal to ½ the lowest non-zero RPM value measured for each gene, as done in Zhang et al. (2021). There were no zero values for DNA (because we used single copy genes) or *E. coli* RA.

*Gene set enrichment analysis.* We used Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) to test for overrepresentation of gene sets among genes over and under-expressed in the recurrer cohort. We grouped genes into TF regulons, using gene:TF interactions reported in RegulonDB(Huerta et al. 1998), as well as KEGG Modules and Pathways. For TFs with dual activity, we grouped genes into separate regulons consisting of genes activated and repressed by the TF. Because RegulonDB reports regulatory information for *E. coli K12* and we used a *E. coli UTI89* reference, the TF:gene interactions were not immediately transferable to our dataset. We used SynerClust (Georgescu et al. 2018) to pair orthologs between the two reference genomes, and transferred annotations from the *K12* reference to UTI89 orthologs. For annotation of KEGG Pathway and Module membership, we used KEGG gene name annotations from RegulonDB and the KEGG *E. coli K12* (*eco*) Pathway and Module maps. For GSEA, we used gene sets that were five genes in size or larger. We used the t-scores from the differential expression tests (reported above) as input for GSEA. We reported GSEA results at an FDR-corrected significance threshold of 0.05.

*Metagenomic growth rate estimation.* The growth rate of *E. coli* strains within metagenomes was estimated as the difference in coverage between the origin and terminus of replication, as implemented in SMEG(Emiola, Zhou, and Oh 2020). We constructed a SMEG species database using the strain reference genomes reported by StrainGST, and ran SMEG using the 'reference based' mode with the strain RA estimated by StrainGST. Per sample, we calculated the average *E. coli* growth rate as the average of strain growth rates, weighted by strain RA. We used linear mixed effects models to compare strain growth rate between cohorts, of the form $growth\ rate \sim cohort\ +\ (1|subject)$.

## Statistical analysis and graphical plotting

All statistical analysis and plotting was performed in R v3.6(Team and Others 2012) ggplot2(Wilkinson 2011), data.table (https://rdatatable.gitlab.io/data.table/), Rmisc (https://cran.r-project.org/package=Rmisc). Python libraries used include pandas v.1.3.2(McKinney, n.d.), numpy v.1.21.2(Harris et al. 2020), scipy v.1.7.1(Virtanen et al. 2020), statsmodels v.0.12.1(Seabold and Perktold, n.d.), biopython v.1.7.9(Cock et al. 2009), pysam v.0.19.1(*Pysam: Pysam Is a Python Module for Reading and Manipulating SAM/BAM/VCF/BCF Files. It's a Lightweight Wrapper of the Htslib C-API, the Same One That Powers Samtools, Bcftools, and Tabix* n.d.), matplotlib v.3.4.3(Hunter 2007), and seaborn v.0.11.2(Waskom 2021).

## Data availability

Post-HS sequencing data has been submitted to SRA to NCBI's Sequence Read Archive under Bioprojects PRJNA685748 (mock community) and PRJNA400628 (UMB stool samples). Pre-HS sequencing data was previously submitted under these same Bioprojects (pre-HS DNA data for the mock community and stool samples from the UMB project). The assembly of the *E. coli* isolate UMB08_02 has been submitted under accession GCF_011751425.1.

## Acknowledgements

We would like to thank Curtis Huttenhower and members of his lab as well as members of the Bacterial Genomics group for helpful discussions. We would also like to thank Broad's Genomic Platform and Microbial 'Omics Core for their assistance with data generation.

This project has been funded with Broad NextGen funds to AME and Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Grant Number U19AI110818 to the Broad Institute and R01DK121822 to Washington University and the Broad Institute.

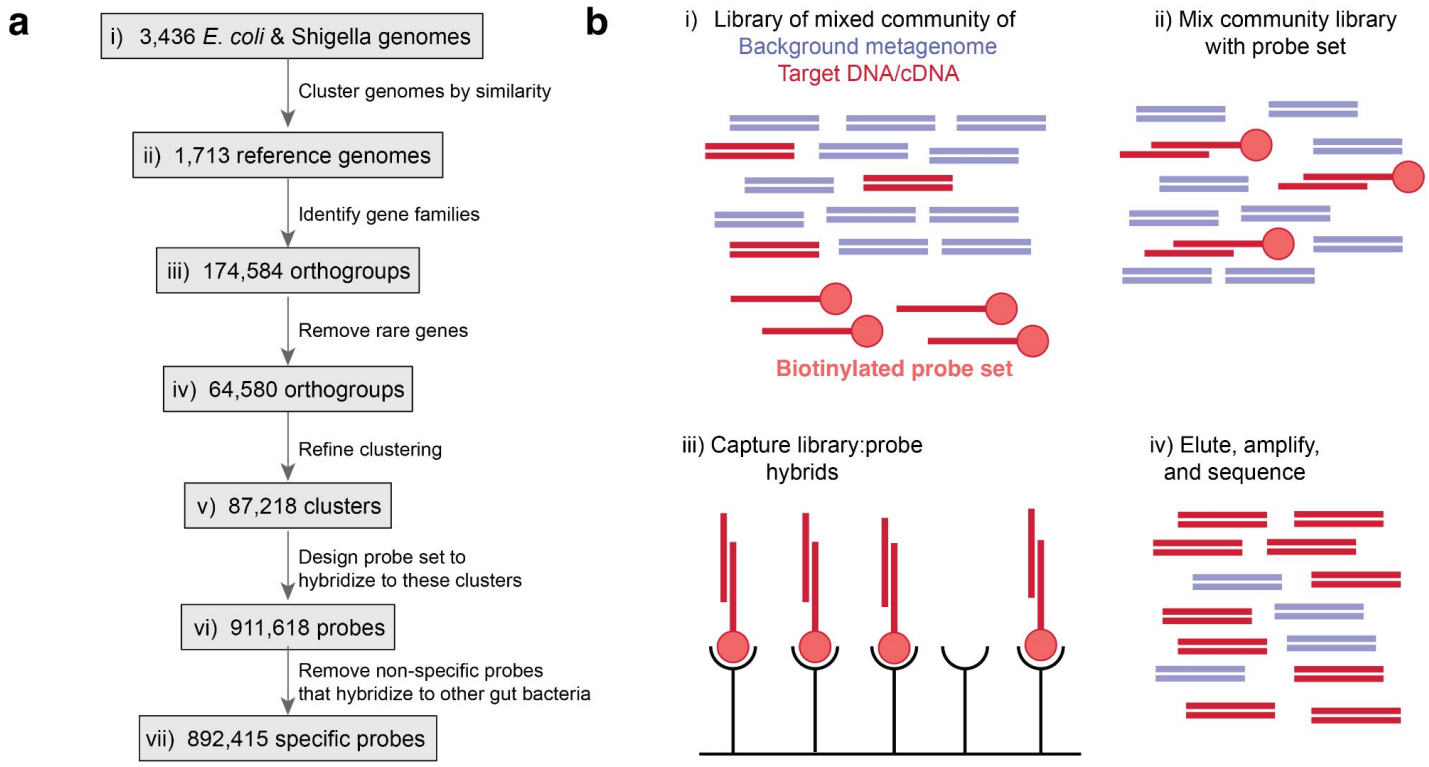


**Figure 1. *E. coli* PanSelect probe design and hybrid capture process.** a-d) Schematic illustration of hybrid capture process for enriching nucleic acids from minor members of a complex community. b) Initial sequencing libraries containing a small number of reads from a low-abundance target organism (red) in a metagenomic background (blue). c) Short oligonucleotide probes representing a bait set attached to biotin (red circles), hybridized with complementary library molecules. d) Sequencing library, biotinylated probes, and capture beads with streptavidin (open semi circles) incubated overnight. e) Bound library molecules eluted off the beads, which are further amplified by PCR and sequenced. This process results in a final sequencing library highly enriched for the species of interest, ready for sequencing. e) *E. coli* PanSelect probe design: i) Complete genomes were downloaded from RefSeq and the NCBI Pathogens database, then ii) clustered using a k-mer based algorithm. iii) Orthogroups were constructed with SynerClust (Georgescu et al. 2018), iv) filtered to remove rare orthogroups, and v) further clustered at 80% identity with UCLUST (Edgar 2010). vi) These clusters were each used as input into CATCH (Metsky et al. 2019) to generate 60-75bp probe sequences. vii) Probes were filtered to remove those that hit many bacteria, as well as probes that hit genomes of major members of the human gut microbiome from the *Bacteroidetes* and *Firmicutes* phyla, retaining a total of 892,415 probes for synthesis.

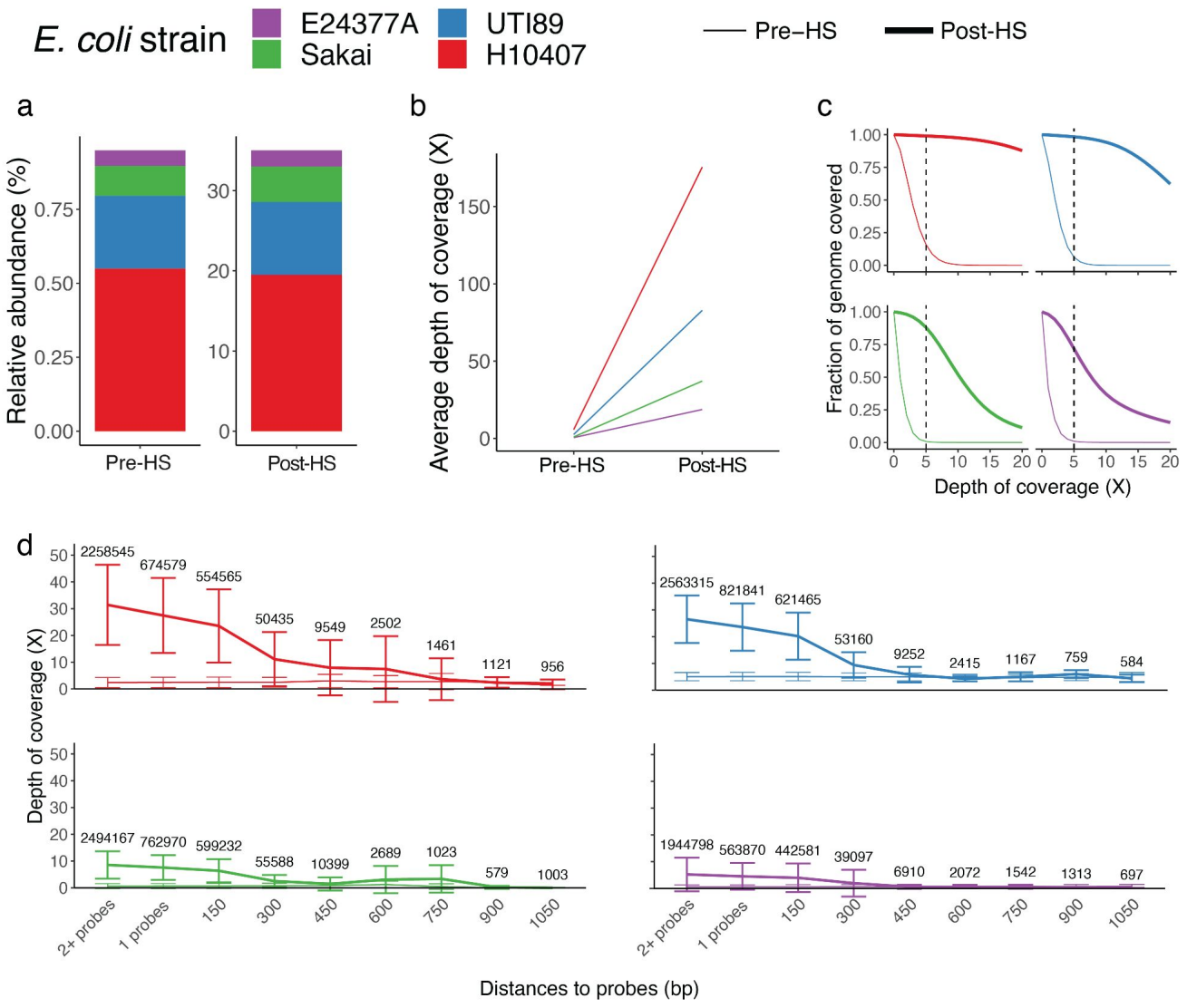


**Figure 2. *E. coli* PanSelect enriches *E. coli* DNA without bias from a 4-strain mock community.** The four strains included in the mock community are represented by color in each panel. a) The predicted RAs of the four *E. coli* strains are shown pre- and post-HS, calculated using StrainGE. b) Average depth of coverage pre- and post-HS for each strain. c) Genome coverage pre-HS (thin lines) and post-HS (thick lines) for each strain. The dashed vertical line represents 5x coverage. d) Average depth of coverage in relation to the closest predicted probe binding site(s), for each of the four strains. “1 probe”and “2+ probes” indicate regions where probes are predicted to bind. Error bars denote standard deviations. Numbers above error bars indicate the number of positions across the genome in each category. Thin lines represent pre-HS data; thick lines post-HS data.

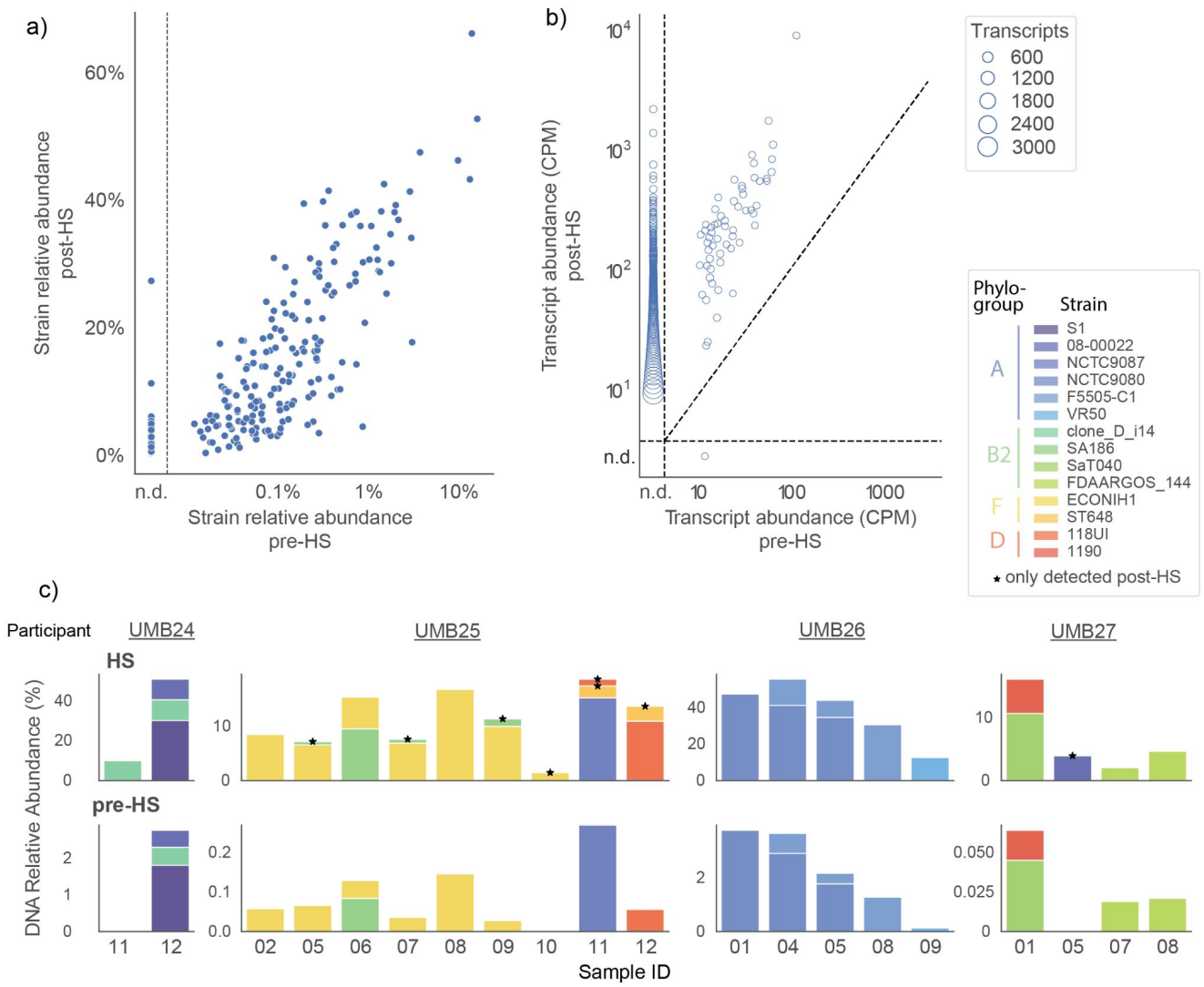


**Figure 3. *E. coli* PanSelect enriched *E. coli* from human stool samples, revealing previously missed strains and transcripts.** a) Pre- and post-enrichment relative abundances of *E. coli* strains detected within 191 human stool metagenomes. Points to the left of the dashed vertical line represent strains that were exclusively detected within enriched metagenomes. Strains were identified by StrainGST. b) Strain composition of two-strain samples, before and after enrichment, for samples from four randomly chosen patients. Strain relative abundances were estimated with StrainGST. Stars indicate strains that could not be detected before HS.c) Pre- and post-enrichment relative abundances of individual *E. coli* transcripts, in copies per million (CPM).Transcripts below and to the left of the dashed lines were not detected (n.d.) before or after enrichment, respectively. Transcripts expressed above 10 CPM were classified as detected. Transcripts with similar expression values were clustered with hierarchical clustering. Circles indicate cluster centers and are sized by the number of transcripts that they represent.

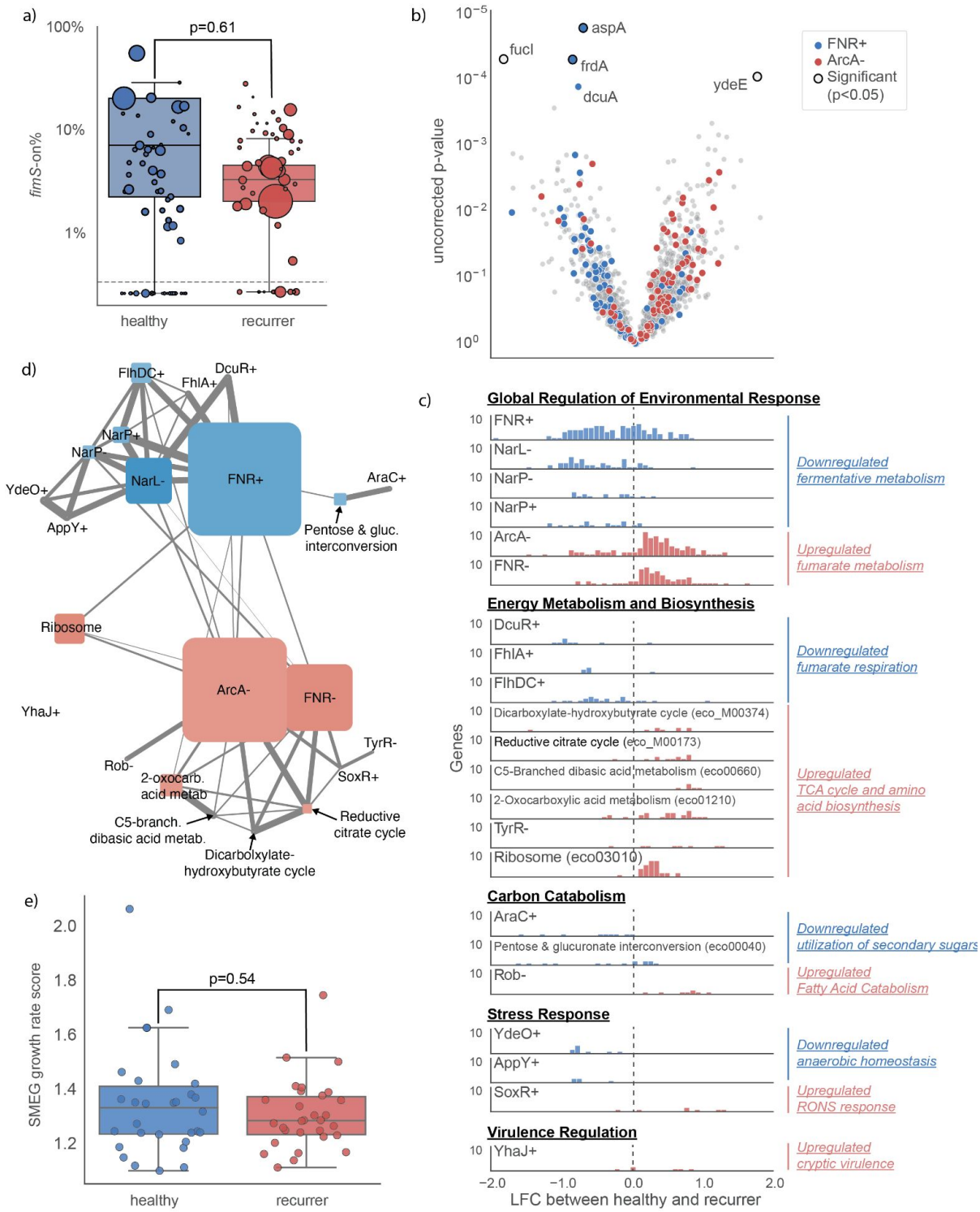


**Figure 4. Shift towards aerobic metabolism in the rUTI gut in the rUTI gut.** a) Percent of *fimS* in the "on" orientation for samples in the healthy (blue) and recurrer (red) cohorts. b) Differential expression (DE) of 2,182 *E. coli* genes between stool from healthy and rUTI women. The log fold change is plotted against the uncorrected p-value. The 4 individual genes that were significantly DE after false discovery rate correction are indicated with black outlines and labeled. *dcuA*, which was near the significance threshold, is also labeled. c) Distribution of fold-changes for genes in each of the 22 under- or over-expressed gene sets (KEGG pathways/modules or regulons) identified using Gene set enrichment analysis (GSEA) (Supplementary Table S14). Datapoints from gene sets are colored by whether the gene set was upregulated (red) or downregulated (blue). d) Network diagram showing the interconnectedness of the DE gene sets, colored according to up- or down-regulation as in c). e) *E. coli* growth rate is shown for the healthy (blue) and recurrer (red) cohorts. Growth rate was estimated from stool metagenomes with SMEG (Methods), which measures the difference in metagenomic coverage between the origin and terminus of replication.

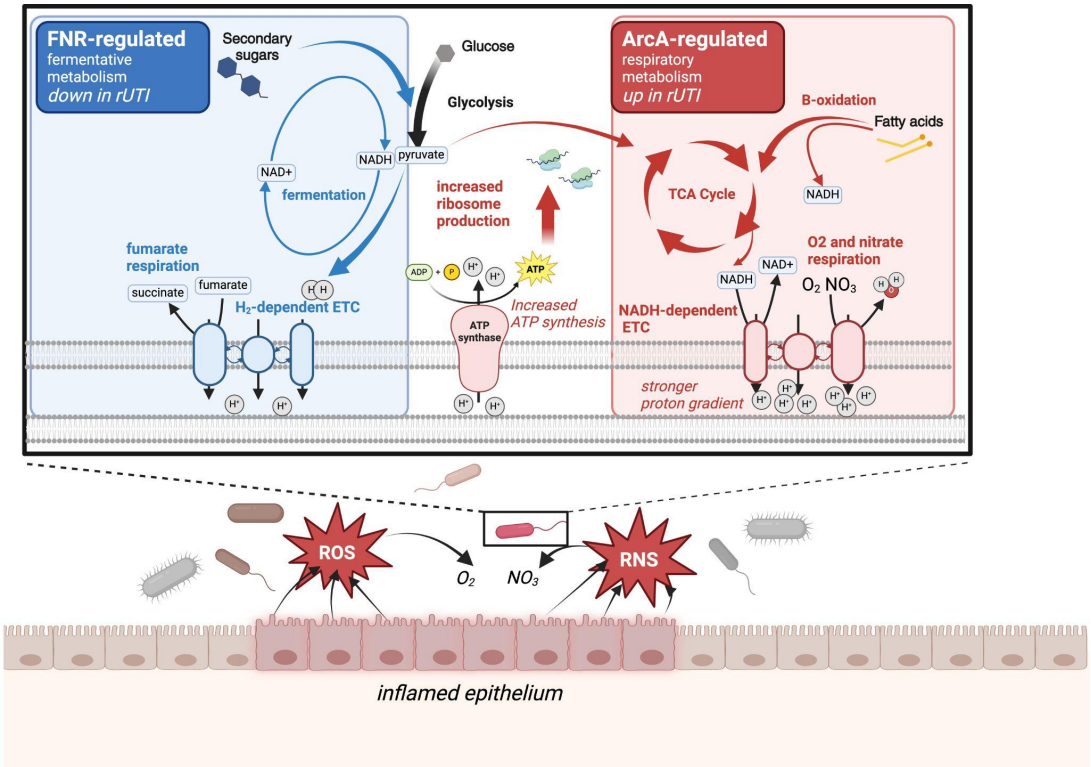


**Figure 5.** Model of transcriptional changes to *E. coli* in the rUTI gut. *E. coli* over-expressed ArcA-regulated respiratory metabolism while under-expressing components of FNR-regulated fermentative metabolism. ATP synthesis and ribosome production also showed evidence for increased activity. Inflammation-associated reactive oxygen and nitrogen species (RONS) may be responsible for driving changes in expression. Created with BioRender.com.

## Tables

**Table 1. Enrichment of strain genomes for the mock community: Post-HS depth of coverage for the mock community was increased for >96% of each strains' genome.**

| | Depth of coverage | | |
|---|---|---|---|
| | Unenriched | | Enriched |
| **Strain** | **Missing in both pre-HS and HS*** | **HS <= pre-HS** | **HS > pre-HS** |
| H10407 | 0.02% | 0.58% | 99.40% |
| UTI89 | 0.05% | 0.70% | 99.25% |
| Sakai | 0.37% | 1.18% | 98.44% |
| E24377A | 1.51% | 1.62% | 96.86% |

*Missing indicates neither pre-HS nor HS data had coverage. Other columns indicate HS data compared to pre-HS data depth of coverage.

# References


Abraham, J. M., C. S. Freitag, J. R. Clements, and B. I. Eisenstein. 1985. “An Invertible Element of DNA Controls Phase Variation of Type 1 Fimbriae of Escherichia Coli.” *Proceedings of the National Academy of Sciences of the United States of America* 82 (17): 5724–27.

Ajiboye, Remi M., Owen D. Solberg, Bryan M. Lee, Eva Raphael, Chitrita Debroy, and Lee W. Riley. 2009. “Global Spread of Mobile Antimicrobial Drug Resistance Determinants in Human and Animal Escherichia Coli and Salmonella Strains Causing Community-Acquired Infections.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 49 (3): 365–71.

Allard, Marc W., Rebecca Bell, Christina M. Ferreira, Narjol Gonzalez-Escalona, Maria Hoffmann, Tim Muruvanda, Andrea Ottesen, et al. 2018. “Genomics of Foodborne Pathogens for Microbial Food Safety.” *Current Opinion in Biotechnology* 49 (February): 224–29.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Andrews, Simon, and Others. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.” Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Beebout, Connor J., Gabriella L. Robertson, Bradley I. Reinfeld, Alexandra M. Blee, Grace H. Morales, John R. Brannon, Walter J. Chazin, et al. 2022. “Uropathogenic Escherichia Coli Subverts Mitochondrial Metabolism to Enable Intracellular Bacterial Pathogenesis in Urinary Tract Infection.” *Nature Microbiology* 7 (9): 1348–60.

Beghain, Johann, Antoine Bridier-Nahmias, Hervé Le Nagard, Erick Denamur, and Olivier Clermont. 2018. “ClermonTyping: An Easy-to-Use and Accurate in Silico Method for Escherichia Genus Strain Phylotyping.” *Microbial Genomics* 4 (7). https://doi.org/10.1099/mgen.0.000192.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. “Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3.” *eLife* 10 (May). https://doi.org/10.7554/eLife.65088.

Bertrand, Denis, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, et al. 2019. “Hybrid Metagenomic Assembly Enables High-Resolution Analysis of Resistance Determinants and Mobile Elements in Human Microbiomes.” *Nature Biotechnology* 37 (8): 937–44.

Bhattacharyya, Roby P., Nirmalya Bandyopadhyay, Peijun Ma, Sophie S. Son, Jamin Liu, Lorrie L. He, Lidan Wu, et al. 2019. “Simultaneous Detection of Genotype and Phenotype Enables Rapid and Accurate Antibiotic Susceptibility Determination.” *Nature Medicine* 25 (12): 1858–64.

Bright, A. Taylor, Ryan Tewhey, Shira Abeles, Raul Chuquiyauri, Alejandro Llanos-Cuentas, Marcelo U. Ferreira, Nicholas J. Schork, Joseph M. Vinetz, and Elizabeth A. Winzeler. 2012. “Whole Genome Sequencing Analysis of Plasmodium Vivax Using Whole Genome Capture.” *BMC Genomics* 13 (June): 262.

Bronson, Ryan A., Chhavi Gupta, Abigail L. Manson, Jan A. Nguyen, Asli Bahadirli-Talbott, Nicole M. Parrish, Ashlee M. Earl, and Keira A. Cohen. 2021. “Global Phylogenomic Analyses of Mycobacterium Abscessus Provide Context for Non Cystic Fibrosis Infections and the Evolution of Antibiotic Resistance.” *Nature Communications* 12 (1): 5145.

Bunduki, Gabriel Kambale, Eva Heinz, Vincent Samuel Phiri, Patrick Noah, Nicholas Feasey, and Janelisa Musaya. 2021. “Virulence Factors and Antimicrobial Resistance of Uropathogenic Escherichia Coli (UPEC) Isolated from Urinary Tract Infections: A Systematic Review and Meta-Analysis.” *BMC Infectious Diseases* 21 (1): 753.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (December): 421.

Chung, Matthew, Ricky S. Adkins, John S. A. Mattick, Katie R. Bradwell, Amol C. Shetty, Lisa Sadzewicz, Luke J. Tallon, et al. 2021. “FADU: A Quantification Tool for Prokaryotic Transcriptomic Analyses.” *mSystems* 6 (1). https://doi.org/10.1128/mSystems.00917-20.

Clermont, Olivier, Ojas V. A. Dixit, Belinda Vangchhia, Bénédicte Condamine, Sara Dion, Antoine Bridier-Nahmias, Erick Denamur, and David Gordon. 2019. “Characterization and Rapid Identification of

Phylogroup G in Escherichia Coli, a Lineage with High Virulence and Antibiotic Resistance Potential." *Environmental Microbiology* 21 (8): 3107–17.
Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics." *Bioinformatics* 25 (11): 1422–23.
Connolly, James P. R., Nicky O'Boyle, Natasha C. A. Turner, Douglas F. Browning, and Andrew J. Roe. 2019. "Distinct Intraspecies Virulence Mechanisms Regulated by a Conserved Transcription Factor." *Proceedings of the National Academy of Sciences of the United States of America* 116 (39): 19695–704.
Crusoe, Michael R., Hussien F. Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, Reed Cartwright, Amanda Charbonneau, et al. 2015. "The Khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis." *F1000Research* 4 (September): 900.
Dijk, Lucas R. van, Bruce J. Walker, Timothy J. Straub, Colin J. Worby, Alexandra Grote, Henry L. Schreiber, Christine Anyansi, et al. 2021. "StrainGE: A Toolkit to Track and Characterize Low-Abundance Strains in Complex Microbial Communities." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2021.02.14.431013.
Donaldson, Gregory P., Wen-Chi Chou, Abigail L. Manson, Peter Rogov, Thomas Abeel, James Bochicchio, Dawn Ciulla, et al. 2020. "Spatially Distinct Physiology of Bacteroides Fragilis within the Proximal Colon of Gnotobiotic Mice." *Nature Microbiology* 5 (5): 746–56.
Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster than BLAST." *Bioinformatics* 26 (19): 2460–61.
Emiola, Akintunde, Wei Zhou, and Julia Oh. 2020. "Metagenomic Growth Rate Inferences of Strains in Situ." *Science Advances* 6 (17): eaaz2299.
Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32 (19): 3047–48.
Fisher, Sheila, Andrew Barry, Justin Abreu, Brian Minie, Jillian Nolan, Toni M. Delorey, Geneva Young, et al. 2011. "A Scalable, Fully Automated Process for Construction of Sequence-Ready Human Exome Targeted Capture Libraries." *Genome Biology* 12 (1): R1.
Flores-Mireles, Ana L., Jennifer N. Walker, Michael Caparon, and Scott J. Hultgren. 2015. "Urinary Tract Infections: Epidemiology, Mechanisms of Infection and Treatment Options." *Nature Reviews. Microbiology* 13 (5): 269–84.
Foxman, Betsy. 2002. "Epidemiology of Urinary Tract Infections: Incidence, Morbidity, and Economic Costs." *The American Journal of Medicine* 113 Suppl 1A (July): 5S – 13S.
Gama-Castro, Socorro, Verónica Jiménez-Jacinto, Martín Peralta-Gil, Alberto Santos-Zavaleta, Mónica I. Peñaloza-Spinola, Bruno Contreras-Moreira, Juan Segura-Salazar, et al. 2008. "RegulonDB (version 6.0): Gene Regulation Model of Escherichia Coli K-12 beyond Transcription, Active (experimental) Annotated Promoters and Textpresso Navigation." *Nucleic Acids Research* 36 (Database issue): D120–24.
Georgescu, Christophe H., Abigail L. Manson, Alexander D. Griggs, Christopher A. Desjardins, Alejandro Pironti, Ilan Wapinski, Thomas Abeel, Brian J. Haas, and Ashlee M. Earl. 2018. "SynerClust: A Highly Scalable, Synteny-Aware Orthologue Clustering Tool." *Microbial Genomics*. https://doi.org/10.1099/mgen.0.000231.
Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62.
Hilty, Markus, Belinda Y. Betsch, Katja Bögli-Stuber, Nadja Heiniger, Markus Stadler, Marianne Küffer, Andreas Kronenberg, et al. 2012. "Transmission Dynamics of Extended-Spectrum β-Lactamase-Producing Enterobacteriaceae in the Tertiary Care Hospital and the Household Setting." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 55 (7): 967–75.
Huerta, A. M., H. Salgado, D. Thieffry, and J. Collado-Vides. 1998. "RegulonDB: A Database on Transcriptional Regulation in Escherichia Coli." *Nucleic Acids Research* 26 (1): 55–59.
Hunter. 2007. "Matplotlib: A 2D Graphics Environment" 9 (May): 90–95.
Ingledew, W. J., and R. K. Poole. 1984. "The Respiratory Chains of Escherichia Coli." *Microbiological Reviews* 48 (3): 222–71.
Jantunen, Maria E., H. Saxén, Susanna Lukinmaa, Marja Ala-Houhala, and Anja Siitonen. 2001. "Genomic Identity of Pyelonephritogenic Escherichia Coli Isolated from Blood, Urine and Faeces of Children with

Urosepsis." *Journal of Medical Microbiology* 50 (7): 650–52.
Kanehisa, M., and S. Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1): 27–30.
Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. "MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies." *PeerJ* 7 (July): e7359.
Kaper, James B., James P. Nataro, and Harry L. Mobley. 2004. "Pathogenic Escherichia Coli." *Nature Reviews. Microbiology* 2 (2): 123–40.
Karlowsky, James A., Daryl J. Hoban, Melanie R. Decorby, Nancy M. Laing, and George G. Zhanel. 2006. "Fluoroquinolone-Resistant Urinary Isolates of Escherichia Coli from Outpatients Are Frequently Multidrug Resistant: Results from the North American Urinary Tract Infection Collaborative Alliance-Quinolone Resistance Study." *Antimicrobial Agents and Chemotherapy* 50 (6): 2251–54.
Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.
Leichty, Aaron R., and Dustin Brisson. 2014. "Selective Whole Genome Amplification for Resequencing Target Microbial Species from Complex Natural Samples." *Genetics* 198 (2): 473–81.
Li, Daijiang, Russell Dinnage, Lucas A. Nell, Matthew R. Helmus, and Anthony R. Ives. 2020. "Phyr: An R Package for Phylogenetic Species‐distribution Modelling in Ecological Communities." *Methods in Ecology and Evolution / British Ecological Society* 11 (11): 1455–63.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
Mamanova, Lira, Alison J. Coffey, Carol E. Scott, Iwanka Kozarewa, Emily H. Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J. Turner. 2010. "Target-Enrichment Strategies for next-Generation Sequencing." *Nature Methods* 7 (2): 111–18.
Manges, A. R., J. R. Johnson, B. Foxman, T. T. O'Bryan, K. E. Fullerton, and L. W. Riley. 2001. "Widespread Distribution of Urinary Tract Infections Caused by a Multidrug-Resistant Escherichia Coli Clonal Group." *The New England Journal of Medicine* 345 (14): 1007–13.
Martín-Rodríguez, Alberto J., Mikael Rhen, Keira Melican, and Agneta Richter-Dahlfors. 2020. "Nitrate Metabolism Modulates Biosynthesis of Biofilm Components in Uropathogenic Escherichia Coli and Acts as a Fitness Factor During Experimental Urinary Tract Infection." *Frontiers in Microbiology* 11 (January): 26.
Matranga, Christian B., Kristian G. Andersen, Sarah Winnicki, Michele Busby, Adrianne D. Gladden, Ryan Tewhey, Matthew Stremlau, et al. 2014. "Enhanced Methods for Unbiased Deep Sequencing of Lassa and Ebola RNA Viruses from Clinical and Biological Samples." *Genome Biology* 15 (11): 519.
McKinney. n.d. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science*. https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf.
Mediavilla, José R., Amee Patrawalla, Liang Chen, Kalyan D. Chavda, Barun Mathema, Christopher Vinnard, Lisa L. Dever, and Barry N. Kreiswirth. 2016. "Colistin- and Carbapenem-Resistant Escherichia Coli Harboring Mcr-1 and blaNDM-5, Causing a Complicated Urinary Tract Infection in a Patient from the United States." *mBio* 7 (4). https://doi.org/10.1128/mBio.01191-16.
Metsky, Hayden C., Viral Hemorrhagic Fever Consortium, Katherine J. Siddle, Adrianne Gladden-Young, James Qu, David K. Yang, Patrick Brehio, et al. 2019. "Capturing Sequence Diversity in Metagenomes with Comprehensive and Scalable Probe Design." *Nature Biotechnology*. https://doi.org/10.1038/s41587-018-0006-x.
Mulvey, M. A., Y. S. Lopez-Boado, C. L. Wilson, R. Roth, W. C. Parks, J. Heuser, and S. J. Hultgren. 1998. "Induction and Evasion of Host Defenses by Type 1-Piliated Uropathogenic Escherichia Coli." *Science* 282 (5393): 1494–97.
Nielsen, Karen L., Pia Dynesen, Preben Larsen, and Niels Frimodt-Møller. 2014. "Faecal Escherichia Coli from Patients with E. Coli Urinary Tract Infection and Healthy Controls Who Have Never Had a Urinary Tract Infection." *Journal of Medical Microbiology* 63 (Pt 4): 582–89.
Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. "metaSPAdes: A New Versatile Metagenomic Assembler." *Genome Research* 27 (5): 824–34.
Ochman, H., and A. C. Wilson. 1987. "Evolution in Bacteria: Evidence for a Universal Substitution Rate in Cellular Genomes." *Journal of Molecular Evolution* 26 (1-2): 74–86.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25 (7): 1043–55.

Pietrucha-Dilanchian, Paula, and Thomas M. Hooton. 2016. “Diagnosis, Treatment, and Prevention of Urinary Tract Infection.” *Microbiology Spectrum* 4 (6). https://doi.org/10.1128/microbiolspec.UTI-0021-2015.

*Pysam: Pysam Is a Python Module for Reading and Manipulating SAM/BAM/VCF/BCF Files. It’s a Lightweight Wrapper of the Htslib C-API, the Same One That Powers Samtools, Bcftools, and Tabix*. n.d. Github. Accessed September 9, 2022. https://github.com/pysam-developers/pysam.

Quince, Christopher, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. 2017. “DESMAN: A New Tool for de Novo Extraction of Strains from Metagenomes.” *Genome Biology* 18 (1): 181.

Quince, Christopher, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S. Soyer, J. Kimberly Summers, Antoine Limasset, A. Murat Eren, Rayan Chikhi, and Aaron E. Darling. 2020. “Metagenomics Strain Resolution on Assembly Graphs.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.09.06.284828.

Rahdar, Masoud, Ahmad Rashki, Hamid Reza Miri, and Mehdi Rashki Ghalehnoo. 2015. “Detection of Pap, Sfa, Afa, Foc, and Fim Adhesin-Encoding Operons in Uropathogenic Escherichia Coli Isolates Collected From Patients With Urinary Tract Infection.” *Jundishapur Journal of Microbiology* 8 (8): e22647.

Ranjan, Amit, Sabiha Shaik, Nishant Nandanwar, Arif Hussain, Sumeet K. Tiwari, Torsten Semmler, Savita Jadhav, et al. 2017. “Comparative Genomics of Escherichia Coli Isolated from Skin and Soft Tissue and Other Extraintestinal Infections.” *mBio* 8 (4). https://doi.org/10.1128/mBio.01070-17.

Rosen, David A., Thomas M. Hooton, Walter E. Stamm, Peter A. Humphrey, and Scott J. Hultgren. 2007. “Detection of Intracellular Bacterial Communities in Human Urinary Tract Infection.” *PLoS Medicine* 4 (12): e329.

Russo, T. A., and J. R. Johnson. 2000. “Proposal for a New Inclusive Designation for Extraintestinal Pathogenic Isolates of Escherichia Coli: ExPEC.” *The Journal of Infectious Diseases* 181 (5): 1753–54.

Schreiber, Henry L., 4th, Matt S. Conover, Wen-Chi Chou, Michael E. Hibbing, Abigail L. Manson, Karen W. Dodson, Thomas J. Hannan, et al. 2017. “Bacterial Virulence Phenotypes of Escherichia Coli and Host Susceptibility Determine Risk for Urinary Tract Infections.” *Science Translational Medicine* 9 (382). https://doi.org/10.1126/scitranslmed.aaf1283.

Schwan, William R., and Hua Ding. 2017. “Temporal Regulation of Fim Genes in Uropathogenic Escherichia Coli during Infection of the Murine Urinary Tract.” *Journal of Pathogens* 2017 (December): 8694356.

Seabold, and Perktold. n.d. “Statsmodels: Econometric and Statistical Modeling with Python.” *Proceedings of the 9th Python in*. https://pdfs.semanticscholar.org/3a27/6417e5350e29cb6bf04ea5a4785601d5a215.pdf.

Shishkin, Alexander A., Georgia Giannoukos, Alper Kucukural, Dawn Ciulla, Michele Busby, Christine Surka, Jenny Chen, et al. 2015. “Simultaneous Generation of Many RNA-Seq Libraries in a Single Reaction.” *Nature Methods* 12 (4): 323–25.

Spaulding, Caitlin N., Roger D. Klein, Ségolène Ruer, Andrew L. Kau, Henry L. Schreiber, Zachary T. Cusumano, Karen W. Dodson, et al. 2017. “Selective Depletion of Uropathogenic E. Coli from the Gut by a FimH Antagonist.” *Nature* 546 (7659): 528–32.

“Structure, Function and Diversity of the Healthy Human Microbiome.” 2012. *Nature* 486 (7402): 207–14.

Subashchandrabose, Sargurunathan, Tracy H. Hazen, Ariel R. Brumbaugh, Stephanie D. Himpsl, Sara N. Smith, Robert D. Ernst, David A. Rasko, and Harry L. T. Mobley. 2014. “Host-Specific Induction of Escherichia Coli Fitness Genes during Human Urinary Tract Infection.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (51): 18327–32.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Tantoso, Erwin, Birgit Eisenhaber, Miles Kirsch, Vladimir Shitov, Zhiya Zhao, and Frank Eisenhaber. 2022. “To Kill or to Be Killed: Pangenome Analysis of Escherichia Coli Strains Reveals a Tailocin Specific for Pandemic ST131.” *BMC Biology* 20 (1): 146.

Team, R. Core, and Others. 2012. “R: A Language and Environment for Statistical Computing.”

https://repo.bppt.go.id/cran/web/packages/dplR/vignettes/intro-dplR.pdf.
Tenaillon, Olivier, David Skurnik, Bertrand Picard, and Erick Denamur. 2010. “The Population Genetics of Commensal Escherichia Coli.” *Nature Reviews. Microbiology* 8 (3): 207–17.
Terlizzi, Maria E., Giorgio Gribaudo, and Massimo E. Maffei. 2017. “UroPathogenic Escherichia Coli (UPEC) Infections: Virulence Factors, Bladder Responses, Antibiotic, and Non-Antibiotic Antimicrobial Strategies.” *Frontiers in Microbiology* 8 (August): 1566.
The Lancet. 2018. “Balancing Treatment with Resistance in UTIs.” *The Lancet* 391 (10134): 1966.
Valentino, Michael D., Abigail Manson McGuire, Jason W. Rosch, Paulo J. M. Bispo, Corinna Burnham, Christine M. Sanfilippo, Robert A. Carter, et al. 2014. “Unencapsulated Streptococcus Pneumoniae from Conjunctivitis Encode Variant Traits and Belong to a Distinct Phylogenetic Cluster.” *Nature Communications* 5 (November): 5411.
Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72.
Waskom, Michael. 2021. “Seaborn: Statistical Data Visualization.” *Journal of Open Source Software* 6 (60): 3021.
Wilkinson, Leland. 2011. “ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H.” *Biometrics*. https://doi.org/10.1111/j.1541-0420.2011.01616.x.
Winter, Sebastian E., Christopher A. Lopez, and Andreas J. Bäumler. 2013. “The Dynamics of Gut-Associated Microbial Communities during Inflammation.” *EMBO Reports* 14 (4): 319–27.
Winter, Sebastian E., Maria G. Winter, Mariana N. Xavier, Parameth Thiennimitr, Victor Poon, A. Marijke Keestra, Richard C. Laughlin, et al. 2013. “Host-Derived Nitrate Boosts Growth of E. Coli in the Inflamed Gut.” *Science* 339 (6120): 708–11.
Worby, Colin J., Henry L. Schreiber 4th, Timothy J. Straub, Lucas R. van Dijk, Ryan A. Bronson, Benjamin S. Olson, Jerome S. Pinkner, et al. 2022. “Longitudinal Multi-Omics Analyses Link Gut Microbiome Dysbiosis with Recurrent Urinary Tract Infections in Women.” *Nature Microbiology* 7 (5): 630–39.
Xu, Haibin, Xiang Luo, Jun Qian, Xiaohui Pang, Jingyuan Song, Guangrui Qian, Jinhui Chen, and Shilin Chen. 2012. “FastUniq: A Fast de Novo Duplicates Removal Tool for Paired Short Reads.” *PloS One* 7 (12): e52249.
Yamamoto, S., T. Tsukamoto, A. Terai, H. Kurazono, Y. Takeda, and O. Yoshida. 1997. “Genetic Evidence Supporting the Fecal-Perineal-Urethral Hypothesis in Cystitis Caused by Escherichia Coli.” *The Journal of Urology* 157 (3): 1127–29.
Zhang, Shiying, Xochitl Morgan, Belgin Dogan, Francois-Pierre Martin, Suzy Strickler, Akihiko Oka, Jeremy Herzog, et al. 2022. “Mucosal Metabolites Fuel the Growth and Virulence of E. Coli Linked to Crohn’s Disease.” *JCI Insight* 7 (10). https://doi.org/10.1172/jci.insight.157013.
Zhang, Yancong, Amrisha Bhosle, Sena Bae, Lauren J. McIver, Gleb Pishchany, Emma K. Accorsi, Kelsey N. Thompson, et al. 2022. “Discovery of Bioactive Microbial Gene Products in Inflammatory Bowel Disease.” *Nature* 606 (7915): 754–60.
Zhang, Yancong, Kelsey N. Thompson, Curtis Huttenhower, and Eric A. Franzosa. 2021. “Statistical Approaches for Differential Expression Analysis in Metatranscriptomics.” *Bioinformatics* 37 (Suppl_1): i34–41.