Within-host Diversity of Multidrug-Resistant *Escherichia coli* in the Gut and Bladder

Sofiya G. Shevchenko

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading committee:

Evgeni Sokurenko, Chair

John Mittler

Kevin Hybiske

Program Authorized to Offer Degree:

Microbiology

University of Washington

Abstract

Within-host Diversity of *Escherichia coli* in the Gut and Bladder

Sofiya G. Shevchenko

Chair of the Supervisory Committee:

Evgeni V. Sokurenko

Department of Microbiology

Uropathogenic *E. coli* are paradoxically able to both cause disease in the urinary tract, and reside there asymptomatically. The pandemic, multi-drug resistant *E. coli* subclone ST131-H30 (*H*30) is of special interest, as it has been found to persist in the gut and bladder of healthy people. In order to understand this persistence, we investigated whether *H*30 is competitive in these niches and thus able to persist by excluding other *E. coli*, as well as whether *H*30 may persist via within-host adaptation. In order to assess the *E. coli* clonal landscape, we developed a novel method based on deep sequencing of two loci, along with an algorithm for analysis of resulting data. Using this method, we assessed fecal and urinary samples from healthy women carrying *H*30, and found that even in the absence of antibiotic use, *H*30 could completely dominate the gut and, especially, urine of healthy carriers. In order to ascertain whether *H*30 adapts within-host, we employed population-level whole genome sequencing, and determined that *H*30 undergoes changes in genes affecting respiration in the gut, similar to commensal gut *E. coli*. Furthermore, we find that in the bladder, *H*30 undergoes changes potentially adaptive for growth in urine, including nonsynonymous mutations in iron and amino acid metabolism genes.

Contents

CHAPTER 1: INTRODUCTION

The human gut microbiome

Far from sterile, the human body is colonized with millions of microbes that are part of the basic functioning of the body. This includes bacteria, fungi, and other microorganisms that participate in digestion, resistance to disease, and even some emotional states(1–5). While various parts of the body are colonized, the gut is the most diverse and most often studied. The gut is first seeded by bacteria in utero, then undergoes changes throughout development as the human environment and diet change and diversify, until it takes on the adult form at about 3 years of age(6–8). The gut microbial community does not remain static however, as geographical location, stress, food supply and variety, and psychological state modulate it on a continual basis(1, 6, 9–11). Based on these factors, the gut microbiome may be more or less susceptible to disease, itself a cause of significant shifts in the makeup of intestinal microbial community.

This assembly of microorganisms that colonize the gut is immensely diverse. An estimated 1,000 bacterial species inhabit the gut, together with some archaea, phages, nematodes, and two protozoan genera(12, 13). The majority of these microbes live in a state of symbiosis or commensalism with the human host. However, when the community undergoes a disturbance, some of these may become pathogenic. Microbiome disturbance can also leave the community open to invasion by pathogenic species or strains, often belonging to the same genera as commensal or symbiotic residents. This makes the study of the gut microbiome both complex and important in the understanding of human health.

Benefits of the gut microbiome

While the function of the gut microbiome is not fully understood, its role in some aspects of human health has already begun to be described. For example, intestinal microbes are vital for uptake of nutrients and minerals, as well as biosynthesis of vitamins, amino acids, and short-chain fatty acids(1, 3, 14, 15). The gut microbiome provides intestinal epithelial cells with energy by producing acetate, propionate, and butyrate, and preventing the buildup of toxic byproducts in the intestine like D-lactate(1, 3). Some gut bacteria also degrade oxalate, itself a byproduct of intestinal bacterial metabolic activity, which helps prevent oxalate kidney stones(1, 16). Intestinal microbes also break down normally-inactive polyphenols into beneficial compounds(1, 17).

In addition to the digestive activity, a gut microbiome is required for formation and normal function of mucosal immunity in the intestine(18). This includes synthesis of antimicrobial proteins, activation of host immune cells, and shaping of host gut associated lymphoid tissues(18). Furthermore, the gut microbiota participate in maintenance of the structure of the intestine, by preventing cytokine-induced apoptosis of intestinal epithelial cells, as well as decreasing gut permeability by increasing levels of endocannabinoids(1, 19, 20). Overall, the gut microbiome is highly important for gut formation, function, and health.

Role of *Escherichia coli* in the gut microbiome

Escherichia coli is a bacterium found both in the environment and in nearly every mammalian host gut studied so far(21). While its role in human health has not been fully described, *E. coli* is known to be an important player in the microbiome. Commensal *E. coli* have been shown to be protective against pathogenic *E. coli* strains, preventing host colonization by outcompeting the invader(22, 23). Commensal *E. coli* have also been shown to be important during establishment of the infant gut microbiome by consuming oxygen and thus allowing anaerobes to colonize, as well as for host iron acquisition and host appetite regulation in general(24–26). *E. coli* is best

known, however, as a human pathogen, with several pathotypes including enterohemoragic, enteroinvasive, and enterotoxigenic *E. coli*. Various intestinal conditions like IBS and Crohn's disease are also associated with high numbers of *E. coli*; indeed IBS can be partially treated by administering the "probiotic" *E. coli* strain Nissle 1917(27, 28). Finally, *E. coli* is implicated in reproductive diseases, including infertility, miscarriage, SIDS, and others(29). Importantly, commensal and pathogenic *E. coli* alike are able to establish a population in the human gut.

E. coli gut colonization

The human gut is complex environment with various sugars, amino acids, proteins, lipids available for exploitation(21, 27). However, the rapid turnover of the intestinal mucin layer and the diversity of the microbiome necessitate metabolic flexibility to maintain colonization(21, 30). *E. coli* is well-known as a metabolically flexible gut resident, able to utilize various sugars like gluconate, arabinose, N-acetylgalactosamine, and hexuronate as carbon sources(21). It is also so known to use gluconeogenesis to maintain a glycogen reserve to survive carbon limitation(21, 31). *E. coli* are also known to grow in mixed biofilms in the intestine, relying on anaerobes to break down complex polysacchardies into mono- and di-saccharides that *E. coli* can metabolize(21). However, while characteristics like these are applicable to any *E. coli* strain colonizing the gut, a more in-depth look reveals interesting within-species differences.

Different strains of *E. coli* employ somewhat differing colonization strategies, as mouse studies have shown. For instance, *E. coli* K-12 strain MG1655 requires the glycolytic and Entner-Doudoroff pathways to colonize the mouse intestine, while the enterohemoragic strain EDL933 also requires the glyoxylate bypass(21). Moreover, MG1655 does not use gluconeogenesis during colonization, in contrast to the strain Nissle 1917 which does, and EDL933 which only uses it when competing with another *E. coli* strain(21). The latter is particularly interesting, as gluconeogenic conditions are also the trigger for EDL933 virulence factor expression and therefore pathogenesis(21). The particular nutrients used during

colonization also differ strain to strain. MG1655, for instance, does not use fucose or galactose while colonizing the gut, while Nissle 1917 does(21). Another commensal strain, *E. coli* HS, utilizes galactose but not fucose, as does EDL933(21). In fact, ability to use different sugars during colonization has been linked to improved gut colonization, which has been proposed as a contributing factor to the success of Nissle 1917 usage as a probiotic(21). This suggests that not all *E. coli* strains are the same in terms of metabolic function, and are not all equally capable in terms of gut colonization.

Adaptation of E. coli to the gut

While *E. coli* strains can differ in terms of gut colonization ability, their overall metabolic flexibility and the relative ease with which *E. coli* recombines and acquires new genetic elements (genes, operons, as well as plasmids) has allowed for within-host adaptation to the gut. This has been shown extensively in experiments involving germ-free mice, where mutations that allow for metabolism of galactitol, de-repress of the sorbitol operon, and inactivate transport of formate and fumarate among others, have been regularly observed(21, 32–35). Similar within-host adaptation of *E. coli* to the gut in humans has been inferred from a longitudinal study of locus variations in a dominant gut *E. coli* strain(36). Changes were observed in the form of mutations in drug efflux, formate metabolism, heat shock, cAMP metabolism, and ppGpp synthesis, among others. However, studies of *E. coli* adaptation specifically in humans are few. Thus, little is currently known about gut adaptation of specific strains within human hosts.

Antibiotic-resistance and E. coli

The current problem of spread of antibiotic resistance provides a compelling reason for study of *E. coli* (37). Along with the rise of drug-resistant infections, studies have shown a similar rise of resistant strains in the environment, in farm animals, and in healthy people(38–40). Since *E. coli* is acquired primarily through contact with the environment and via consumption of meat

products, this means that resistant strains may spread even in the absence of antibiotic pressure (41). Indeed, studies have shown that infants can be colonized by multidrug-resistant gut *E. coli* strains if their mother is also a carrier(42). To further compound the issue, *E. coli* can rapidly transmit antibiotic resistance to co-habitant bacteria provided the resistance genes are located on a plasmid(43). These data are of concern, as *E. coli* is the most frequent cause of bacterial infection, including bloodstream infections(44). Thus, study of colonization by and transmission of *E. coli* is vital to the understanding of antibiotic resistance spread.

Uropathogenic *E. coli*: pathogenicity and adaptation

UPEC is a subgroup of extrainestinal pathogenic *E. coli* that both resides in the gut and causes UTI. UPEC are usually distinguished from commensal *E. coli* by the presence of certain virulence factors on the genome. These include fimbiral adhesins, pili, toxins, capsule-encoding genes, which are required for colonization of the bladder and for host damage, as well as metabolic genes such as redundant iron acquisition systems(45).

While antibiotic resistance is a challenge with respect to *E. coli* infections in general, this is especially true of urinary tract infections (UTIs). Uropathogenic *E. coli* (UPEC) are one of the most common types of drug-resistant *E. coli* infections, with up to half of all UPEC resistant to at least one antibiotic, and UPEC causing approximately 11 million doctor visits yearly in the US alone(46). UPEC are also often isolated from fecal samples, including in healthy carriers(47–49). Since UPEC first establish a gut population prior to infecting the bladder, carriage of drug-resistant UPEC by healthy people may be involved in the incidence of drug-resistant UTI and spread of antibiotic resistance in general. Thus, study of UPEC may provide information on how antibiotic resistance spreads aside from antibiotic pressure.

UPEC pathogenicity and metabolism

UPEC are unique among *E. coli* in that they are able to colonize two significantly different niches. Since all UPEC establish a gut population prior to bladder colonization, with no known intestinal pathology associated with UPEC, study of UPEC in the gut may provide important insights into commensal *E. coli* gut colonization. Similarly, UPEC can colonize the bladder environment. While virulence factors like adhesin are required for bladder colonization, the UPEC metabolism is similarly vital for continued survival. This is due to the bladder being relatively poor in sugars, lipids, and certain amino acids, necessitating a switch in metabolic activity(45). Particularly important in this respect are biosynthesis of certain amino acids, redundant iron acquisition systems, catabolism of sialic acid and gluconate, and catabolism of D-serine (normally an inhibitor of bacterial growth)(45, 50). Thus, UPEC offer a way to study how *E. coli* adapt to a novel environment.

Importantly, long-term gut carriage of UPEC without bladder colonization is possible, and UPEC are not necessarily pathogenic in the bladder(48). Asymptomatic bacteriuria (ABU) is caused by UPEC strains which are likely very similar to those causing acute cystitis (symptomatic bladder colonization) in terms of ability to grow in urine and uropathogenic qualities. Since the true prevalence of ABU is difficult to ascertain, and prevalence and duration of UPEC carriage without UTI is currently unknown, it may be true that UPEC causes disease in a relative minority of hosts. Therefore, study of UPEC adaptation can tell us what avenues of *E. coli* adaptation can lead to pathogenesis and which lead to commensalism.

Preliminary evidence for UPEC adaptation to the host

There have already been studies indicating that UPEC adapt within-host, specifically to the bladder. For example, the Svanborg lab studied within-host evolution of the ABU strain 83972 by directly inoculating the bladders of three volunteers with the strain and resequencing(51). The study found that the majority of changes present in re-sequenced isolates are nonsynonymous SNPs, and that these targeted genes involved in oxidative stress response,

osmoregulation, and in *barA*, a regulator of carbon storage also associated with virulence. The Frimodt-Møller lab, sequencing fecal and bladder single-colony isolates from patients with UTI observed SNPs in urinary isolates affecting motility, sugar transport, stress response, and carbon storage (52). These studies indicate that UPEC may adapt to the host bladder, primarily via changes to metabolic and stress response genes.

ST131-H30: a pandemic UPEC clone

One of the most common and well-studied UPEC clones is ST131-*H*30, a pandemic subclone that has spread rapidly from its first appearance in the late 1990s, to its current position as a globally-distributed clone and the cause of up to 30% of all UTIs in the US(53, 54). It is also responsible for up to 75% of fluoroquinolone resistant infections, as well as strongly associated with drug-bug mismatches, clinical persistence, and adverse outcomes in elderly and immunocompromised individuals(55–58). Paradoxically, *H*30 is also known to colonize the bladder asymptomatically more frequently than other UPEC clones, and is known to persist in the gut of healthy carriers(48). Since this asymptomatic carriage may be part of *H*30's overall success as a pathogen, understanding whether this clonal group adapts to the host gut and/or bladder is important to unraveling *H*30 pathogenesis.

Our lab has begun study of *H*30 within-host adaptation. Research conducted in our lab has demonstrated that *H*30 changes when transmitted between hosts by acquisition of SNPs, primarily in transcriptional regulators(59). These were found to be inactivating and significantly changed gene expression and pathogen phenotype. Additionally, analysis of a fecal and urinary isolates from one person showed 3 SNPs differentiating the two. However, SNPs in individual isolates are insufficient to determine if *H*30 does indeed change and adapt within its host, as these may be spurious. Therefore, a population-level study is required. Furthermore, the persistence of *H*30 in the host gut may be due to its competitiveness compared to other *E. coli*

strains rather than adaptation to the host. Thus, a study of *H*30 as a part of the overall gut *E. coli* population is needed to determine if *H*30 is a competitive gut resident.

CHAPTER 2: DEVELOPMENT OF A NOVEL METHOD FOR ASSESSMENT OF *E. COLI* STRAIN CONTENT IN HUMAN SAMPLES

Introduction

Microbiomes, both in terms of function and diversity, have recently been a topic of considerable interest. The gut microbiome has gotten special attention due to its high complexity and importance to health(60). So far, studies have almost exclusively focused on species or higher-level diversity. However, this paints an incomplete picture, since strains within the same species can be of distinct clonal origin and have vastly different metabolic, pathogenic, and antibiotic resistance profiles(61, 62). Importantly, multidrug-resistant bacterial strains have been found competing with commensal strains in the gut, even without antibiotic pressure(63, 64). Thus, there is a pressing need to identify strains in the human microbiome for species of critical health importance.

Escherichia coli is one of the most common residents of the gut. While primarily a commensal colonizer, extra-intestinal pathogenic *E. coli* clones are implicated in a variety of diseases, including urinary tract infections (UTIs) - a leading cause of human antibiotic use(38, 53, 63, 65, 66). The spread of multi-drug resistant *E. coli* is now a major health concern, especially the pandemic *fimH*30 subclone of sequence type ST131 (*H*30). Though recently-emerged, *H*30 is now globally distributed and comprises up to half of all urinary and bloodstream isolates of *E. coli* that are fluoroquinolone-resistant and produce extended-spectrum beta-lactamases (ESBL)(48, 55, 56). Somewhat paradoxically, *H*30 is also a persistent gut colonizer of healthy people and frequently causes asymptomatic bacteriuria (ABU) in such carriers(35). Yet, the relative clonal predominance of *H*30 strains among *E. coli* colonizing the gut or bladder in healthy carriers remains unknown. Answering these questions could have a significant impact on understanding the spread of antibiotic resistance and its reservoirs.

Currently, microbiome diversity is studied by sequencing the 16S rRNA gene, but this cannot capture clonal diversity(67, 68). Conventional methods for assessing clonal diversity, such as metagenomic sequencing and single colony typing, are costly and labor intensive. For reliable clonal diversity analysis, metagenomic sequencing requires very high coverage per sample, while single colony typing requires handpicking large numbers of colonies for multi-locus sequence typing (MLST)(67–70). In *E. coli*, MLST requires assessment of 7 genes per isolate which is analytically complex, costly, labor intensive, and therefore difficult to implement. Previously, we reported an alternative clonotyping method that requires sequencing regions of only 2 genes – *fumC* which is part of the MLST scheme and fimH that encodes a rapidly-evolving fimbrial adhesin(69). The *fumC*/*fimH*-based (CH) typing of *E. coli* is widely accepted due to its simplicity and ability to not only identify specific STs but subdivide them into smaller subclones(69). Specifically, *H*30 is identified using the allele combination *fumC*40/*fimH*30, while other less resistant ST131 strains have the same *fumC* but different *fimH* alleles.

Here, we report a high-throughput method for clonal typing of *E. coli* strains by combining CH typing and deep amplicon sequencing. We developed a new algorithm - Population-Level Allele Profiler (PLAP) - for detecting alleles and predicting the relative prevalence of each allele in a sample. We were able to assess the prevalence of clonal groups (including *H*30) in multiple fecal and urine samples concurrently, with a limit of relative abundance detection at <1% of the total population.

Results

Deep amplicon sequencing of defined samples

To validate our approach and establish a limit of detection for strain presence, we first tested our deep amplicon sequencing procedure on a set of defined samples. To create the defined samples, we first selected a fecal sample from our lab collection known to contain *H*30 and

ST101. Next, we isolated a single colony from each and confirmed them to be strains of *H*30 (*fumC*40/*fimH*30) and ST101 (*fumC*41/*fimH*86) using CH typing. From these single colonies, we first created *H*30-only and ST101-only mixtures of *fumC* and *fimH* amplicons. We also created four ST101/*H*30 mixed samples by combining the *fumC* and *fimH* amplicons from ST101 and *H*30 in ST101:*H*30 ratios of 1:1, 1:4, 1:100, and 1:1000.

Analysis of raw sequencing data from *H*30-only and ST101-only samples showed the average coverage of erroneous bases was 0.08% ± 0.09% for both strains. Erroneous bases were observed in both genes across most nucleotide positions. The highest coverage for an erroneous base was 0.66% of aligned reads in *fumC* and 0.45% in *fimH* for *H*30, and 0.68% in *fumC* and 0.46% of reads in *fimH* for ST101. The frequency distribution for erroneous base coverage is presented in Figure 1.

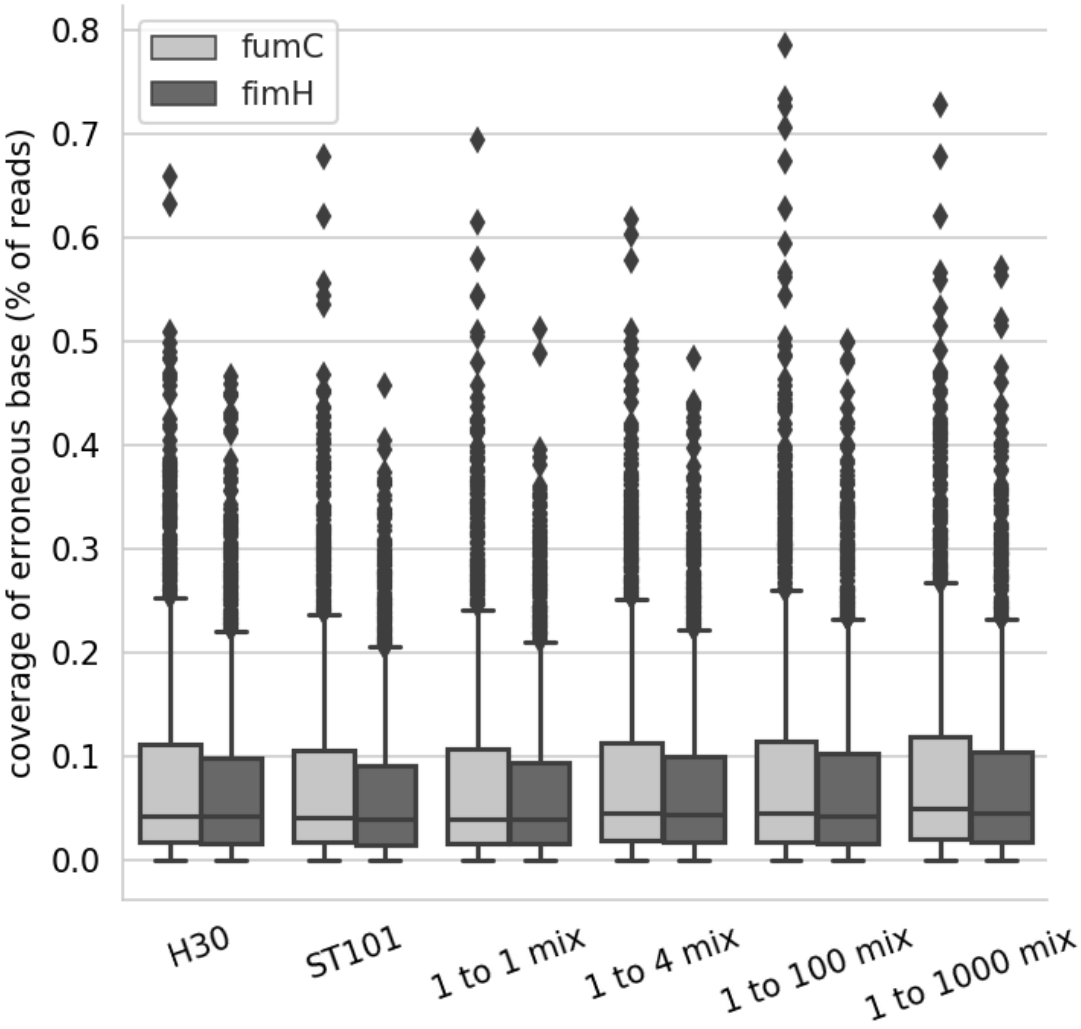


Figure 1. Coverage of erroneous bases in *H*30-only, ST101-only, and mix sample sequencing. Coverage is expressed in percentage of total reads aligned to each gene.

Analysis of raw sequencing data from ST101/*H*30 mixes showed that both *H*30 and ST101 alleles were detectable in the 1:1, 1:4, and 1:100 mixes. In the 1:1000 mix, only alleles of the dominant *H*30 strain were observed. In the 1:1, 1:4, and 1:100 mixes, input and observed allele prevalence was highly correlated for both *fumC* and *fimH* (R^2=0.996 and 0.997 respectively, Fig. 2). Erroneous bases were observed at 0.09% ± 0.1% and 0.08% ± 0.09% of aligned reads in *fumC* and *fimH*, respectively (Fig. 1). The highest coverage for erroneous bases among all mixes was 0.79% of aligned reads for *fumC* and 0.57% of aligned reads for *fimH*.

Since 0.79% of aligned reads was the highest coverage for an erroneous base, we established 0.8% as a cutoff for correct base calling in both genes. This cutoff was used for all further PLAP analysis.

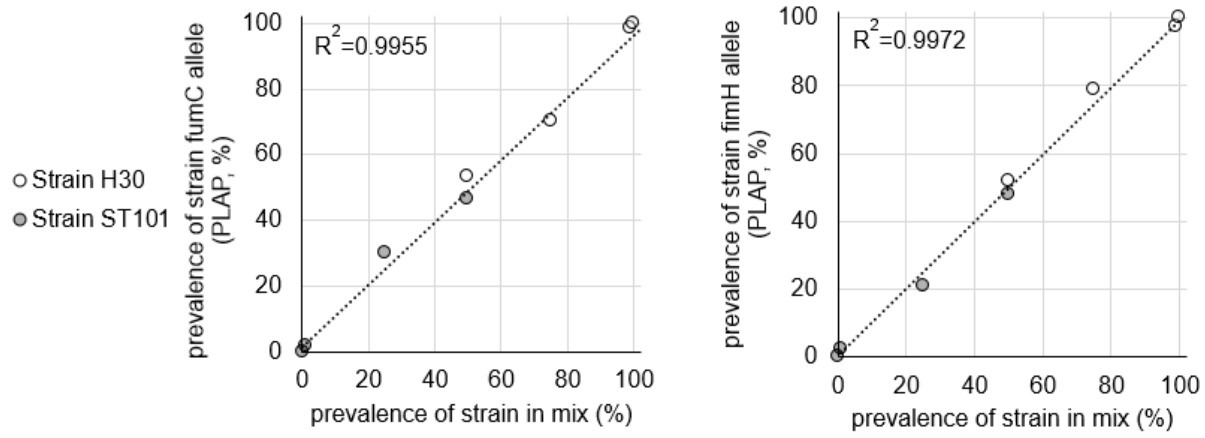


Figure 2. Correlation between input and PLAP-derived (deep seq) prevalences of *fumC* and *fimH* alleles of *H*30 and ST101 in 1:1, 1:4, and 1:100 mixes.

Deep sequencing of study samples and allele prediction

Next, we applied PLAP to 67 participant samples (43 fecal and 24 urine) collected from a previous study[35]. A total of 128 *fumC* and 129 *fimH* alleles were predicted across all samples, of which 123 (96.1%) and 125 (96.9%) were previously known *fumC* and *fimH* alleles, respectively. 5 novel *fumC* and 4 novel *fimH* alleles were potentially detected. All novel *fumC* and *fimH* alleles were phylogenetically distant from other alleles predicted in the sample, indicating that these alleles are not artifacts of sequencing (Fig. 3, 4). These novel alleles nonetheless clustered with other *E. coli fumC* and *fimH* alleles, indicating that these are novel *E. coli* alleles rather than alleles belonging to other species.

The average number of alleles predicted per sample was 1.91 ± 0.96 for *fumC* and 1.93 ± 1.01 for *fimH*. 43 samples had same numbers of predicted *fumC* and *fimH* alleles; 24 samples had

different numbers of predicted *fumC* and *fimH* alleles (Fig. 5). Overall, the number of predicted *fumC* alleles correlated to the number of predicted *fimH* alleles with an R^2 of 0.88 (Fig. 5).

To assess the performance of PLAP for predicting alleles, we used samples containing criterion clones - strains previously identified by single colony typing. PLAP detected criterion *fimH* and *fumC* alleles in 52 of these samples (90%). In the 6 samples where criterion allele(s) were not found, the criterion clones were ciprofloxacin-resistant, but their isolation from the sample required ≥2 plating attempts. This leads us to believe that these alleles were not detected because they were absent in the MacConkey-plated population prior to deep sequencing.

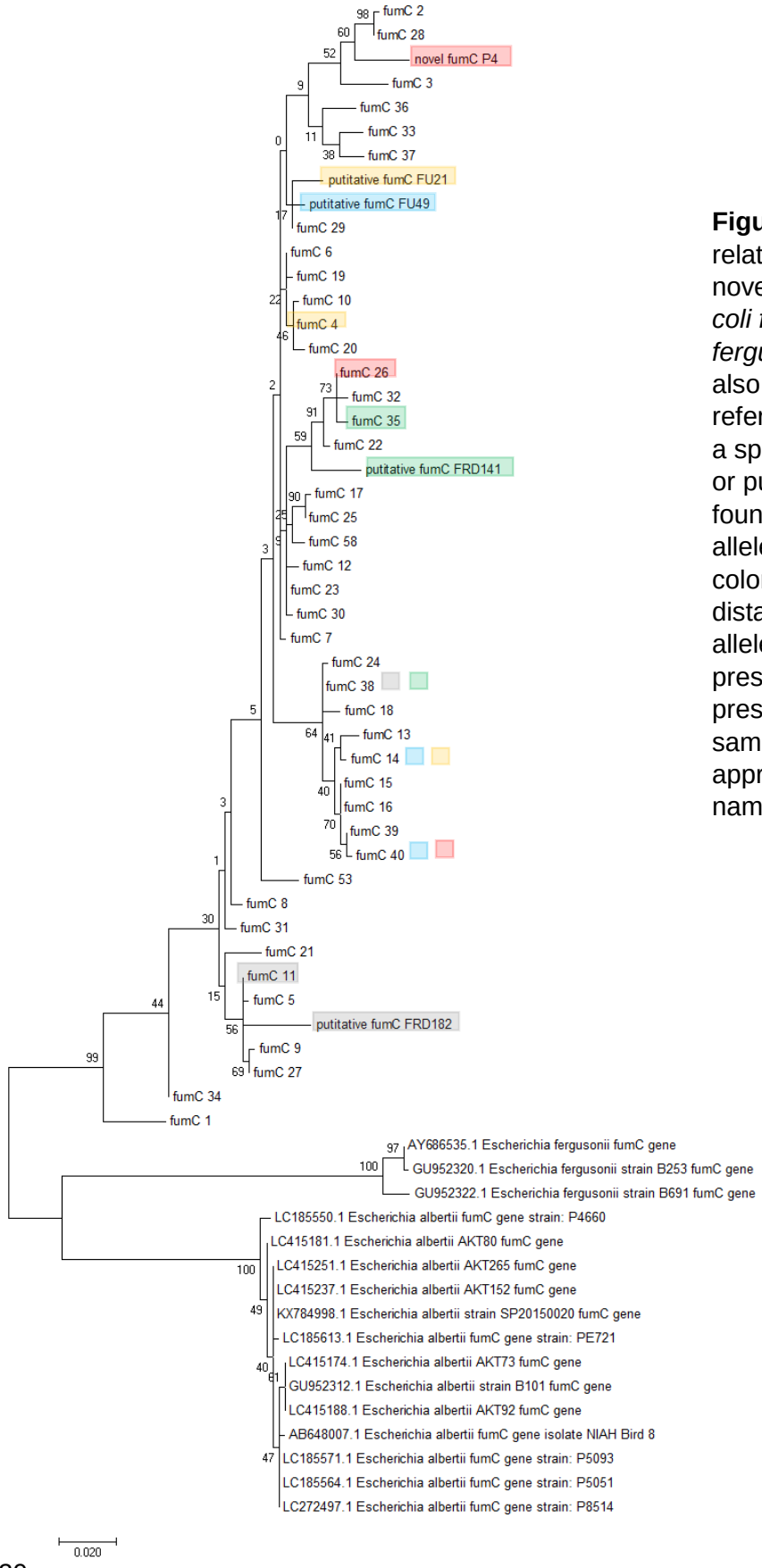


Figure 3. Phylogenetic relationships between predicted novel *fumC* alleles and known *E. coli fumC* alleles. *Escherichia fergusonii* and *albertii fumC* alleles also presented for outgroup reference. Alleles not labelled with a species are known *E. coli* alleles or putative novel alleles. Alleles found in the sample as the novel allele are highlighted in the same color as the novel allele to show distance between predicted novel alleles and other *fumC* alleles present in the sample. Alleles present in multiple different samples are marked with the appropriate colors next to the allele name.

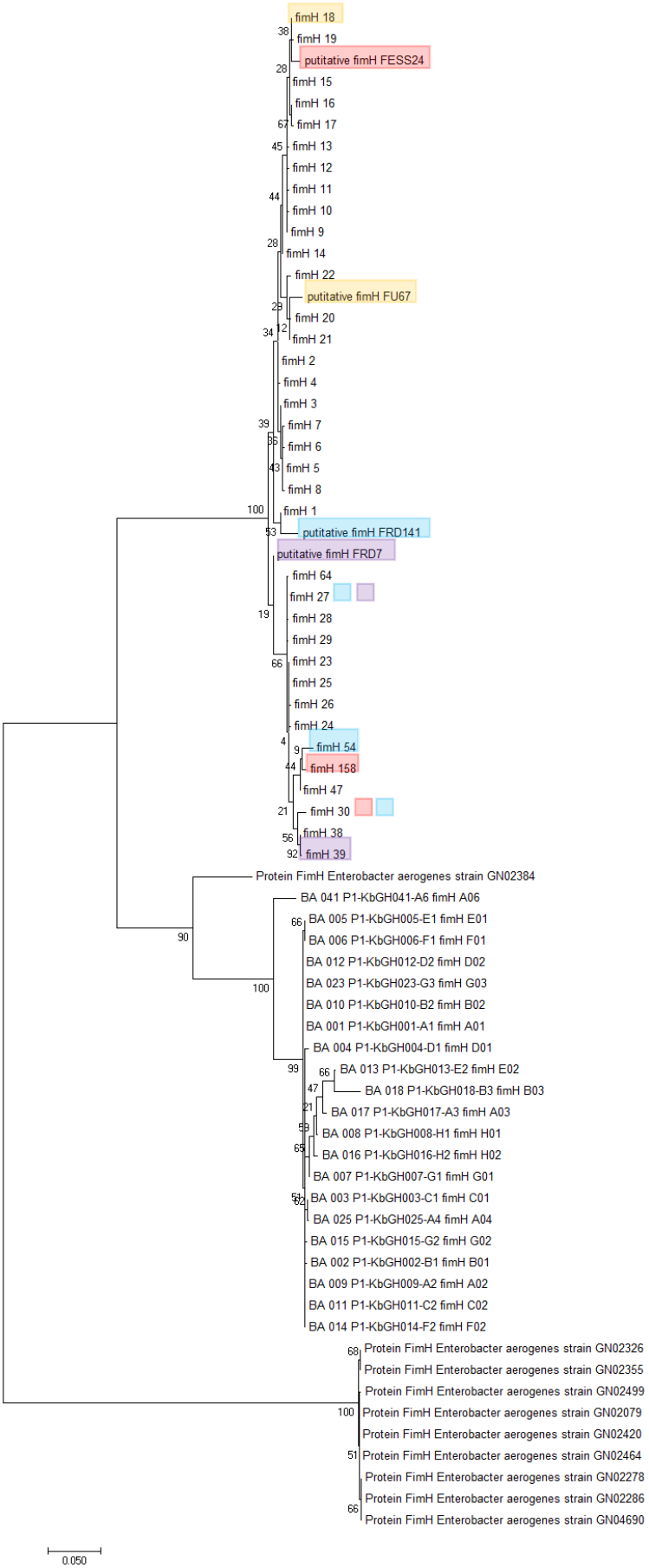


Figure 4. Phylogenetic relationships between predicted novel *fimH* alleles and known *E. coli fimH* alleles. *Klebsiella pneumoniae* and *Enterobacter aerogenes fimH* alleles also presented for outgroup reference. Alleles not labelled with a species are known *E. coli* alleles or putative novel alleles. Alleles found in the sample as the novel allele are highlighted in the same color as the novel allele to show distance between predicted novel alleles and other *fimH* alleles present in the sample. Alleles present in multiple different samples are marked with the appropriate colors next to the allele name.

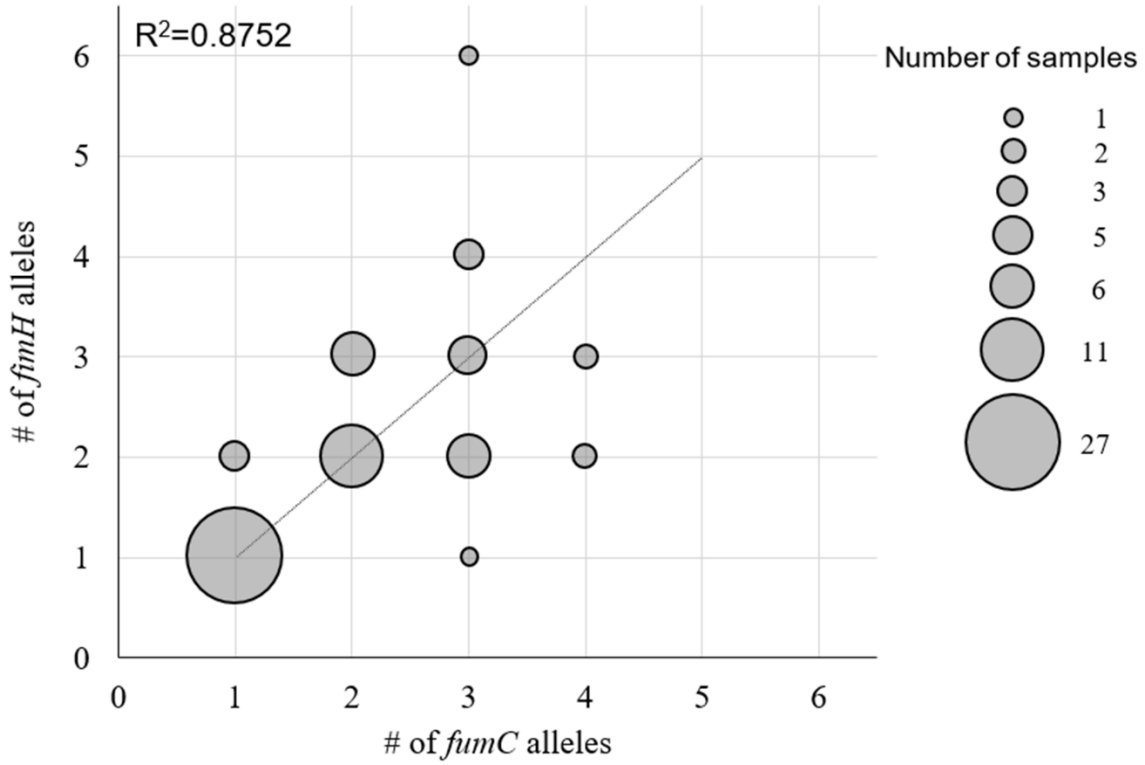


Figure 5. Congruency of *fumC* and *fimH* allele counts in fecal and urine samples. Size of bubbles corresponds to number of samples with designated *fumC/fimH* allele counts (i.e. 1 sample with one *fumC* allele and three *fimH* alleles). Linear fit with Pearson square correlation index shown.

A total of 72 non-criterion (previously unidentified) *fumC* and 71 non-criterion *fimH* alleles were predicted by PLAP across all 67 samples. To assess the performance of PLAP on non-criterion alleles, we analyzed 14 samples (10 fecal, 4 urine) predicted to contain 22 non-criterion *fumC* and 22 non-criterion *fimH* alleles. 12 of these samples had at least one non-criterion allele alongside criterion alleles; the remaining 2 had multiple non-criterion alleles in each gene only. For each sample ≥40 single colonies were isolated and CH type determined using 7-SNP qPCR, with each CH type verified by sequencing. With these data, we confirmed 19 (86%) predicted non-criterion alleles for each gene. This included one predicted novel *fumC* allele. Of the unconfirmed alleles, one was not distinguishable by 7-SNP qPCR and had a predicted prevalence of 1%; therefore, we did not attempt to locate it. The remaining unconfirmed alleles

had predicted prevalences of <3% and therefore may have been missed due to insufficient sampling. Additionally, all criterion alleles in these samples, 12 per gene, were predicted by PLAP.

Prediction of allele prevalence in multi-allele samples

We have also designed PLAP to predict the within-sample prevalence of each allele. The average allele prevalence in fecal samples was 47.3% ± 4.3% SEM (range 0.88 – 100%) for *fumC* and 48.4% ± 4.22% SEM (range 1 – 100%) in *fimH*. The average allele prevalence in urine samples was 64.8% ± 6.91% SEM (range 1.4 – 100%) for *fumC* and 58.3% ± 7.18% SEM (range 1 – 100%) in *fimH*.

In order to verify that the prevalences predicted by PLAP were accurate, we compared predictions to actual in-sample prevalence using two different methods.

In the first method, we used *H*30 since ascertaining its prevalence is relatively simple. By plating the sample on MacConkey agar then patching onto LB-ciprofloxacin, it is possible to compare the number of ciprofloxacin-resistant (*H*30) colonies to the total number of *E. coli* colonies. The ratio of these two numbers provides the *H*30 load in a sample. We compared the predicted prevalences of *fumC*40 and *fimH*30 to the *H*30 load in 17 fecal samples containing ciprofloxacin-resistant *H*30.

Correlations between the *H*30 load and the predicted prevalence of *fumC*40 and *fimH*30 were 0.86 and 0.84 respectively (Fig. 6), indicating that prevalences given by PLAP were representative of actual allele prevalences. To determine whether outliers were present, we calculated the 99% CI range for every sample (see Methods). Three outlier samples were identified (open circles, Fig. 6). Since it is possible that these outliers contain ciprofloxacin-sensitive non-*H*30 *fimH*30-containing clones, *fumC*-null or *fimH*-null clones, and/or ciprofloxacin-sensitive *H*30, we decided to employ screening of a large number of single colonies.

In this second method, we used single colony typing for the in-depth characterization of 14 multi-allele samples described above, alongside 4 additional single-allele samples (2 fecal, 2 urine) for which only one allele per gene was predicted. This set of 18 samples included 11 of the 17 fecal samples used for the *H*30-based analysis above, including one of the outlier samples. For all 18 samples, we used CH typing of ≥40 single colonies per sample to determine the prevalence of each *fumC* and *fimH* allele. Correlation between the PLAP-predicted prevalence and the experimental allele prevalence was 0.98 for both *fumC* and *fimH* alleles (Fig. 7). As in the *H*30 analysis above, we determined whether outliers were present using the 99% CI range for every sample. Only one outlier was detected, corresponding to the only sample that contained colonies from which *fimH* could not be amplified (*fimH*-null colonies). Furthermore, the sample that was an outlier in the *H*30-based analysis was found to contain a relatively rare ciprofloxacin-sensitive *H*30.

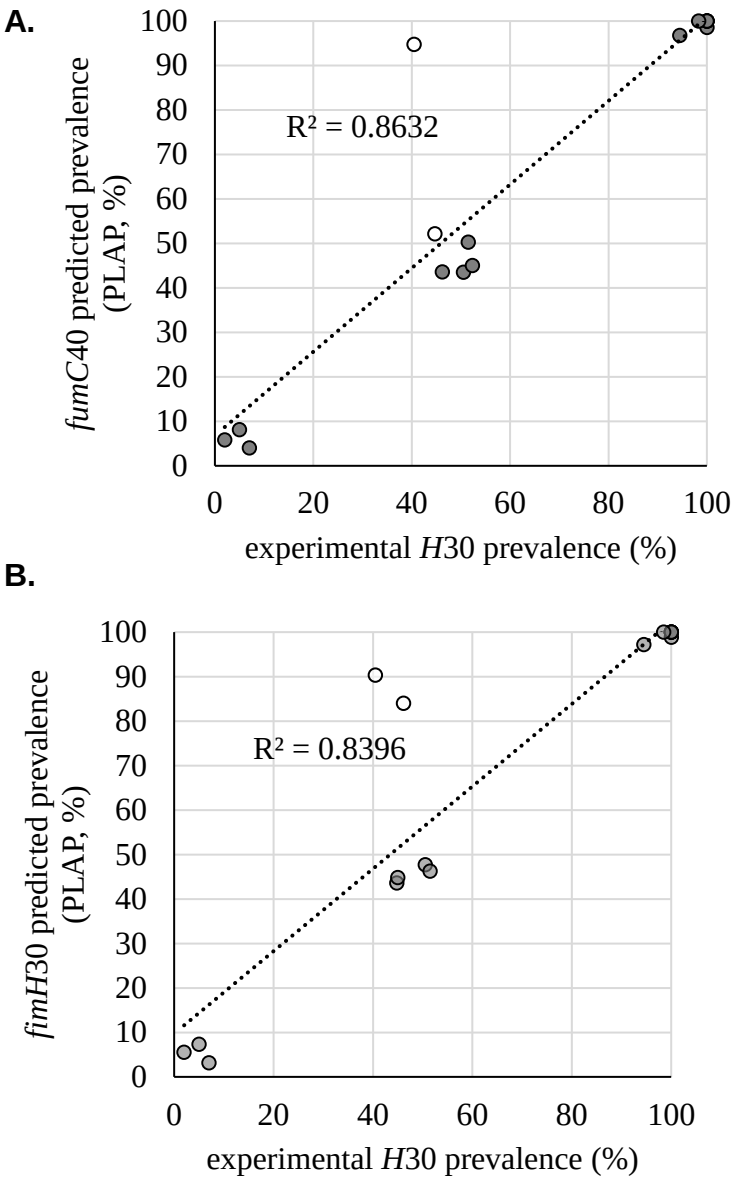


Figure 6. **Validation of predicted *H*30 allele prevalence.** PLAP-predicted prevalence of *H*30 alleles vs actual *H*30 load in *H*30-containing fecal samples. Prevalence of predicted *fumC*40 (**A**) and predicted *fimH*30 (**B**). Predicted prevalence of *fumC*40 and *fimH*30 is expressed as percentage of all *E. coli* in each sample. Experimentally confirmed *H*30 load is expressed as percent of *H*30 (ciprofloxacin-resistant) single colonies to all plated *E. coli* single colonies in percent. At least 130 colonies were tested per sample. Outliers, marked in open circles, were outside the 99% confidence interval of the number of colonies tested.

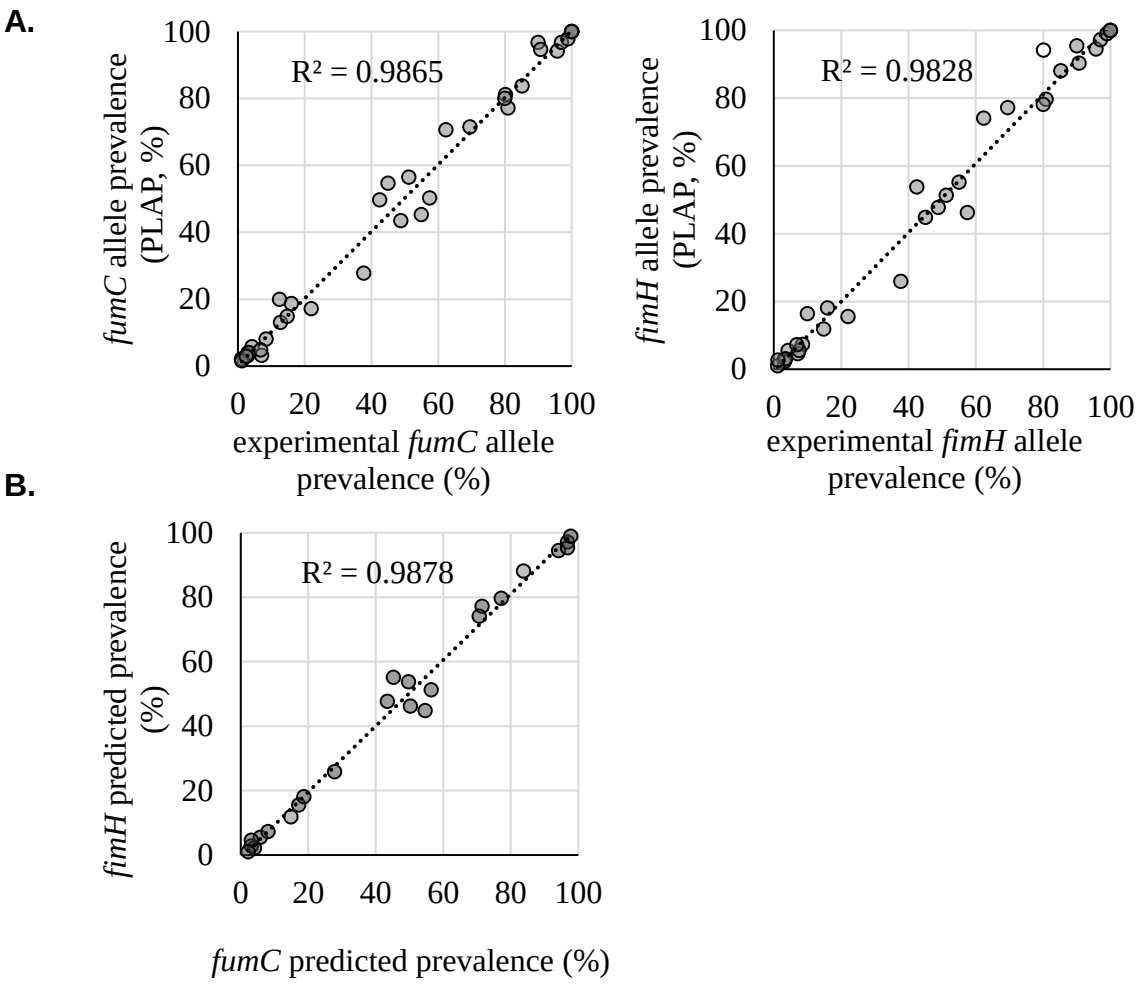


Figure 7. Validation of predicted *fumC/fimH* allele prevalence. A. PLAP-predicted vs experimental within-sample *fumC/fimH* allele prevalence in 18 samples. Experimental allele prevalence was determined by CH typing of at least 40 single bacterial colonies per sample. Outliers (open circles) were outside the 99% confidence interval of the number of colonies sampled. **B.** Predicted prevalence of *fumC* vs *fimH* alleles from the same CH type in 11 samples where no sharing of alleles between strains was present.

Matching fumC and fimH alleles to predict sample strain content

In CH typing, unique combinations of *fumC* and *fimH* alleles are used to determine the identities of strains in a sample. Since a strain contains one copy of *fumC* and *fimH*, the prevalences of alleles of these two genes in the sequencing data should be identical. For example, in a sample

containing 30% *H*30 (*fumC*40/*fimH*30) and 70% ST101 (*fumC*41/*fimH*86), we expect to see 30% of *fumC* reads to be *fumC*40 and 30% of *fimH* reads to be *fumH*30. In reality, however, the prevalences will be slightly different due to PCR and sequencing errors. To establish an acceptable difference between the prevalences of same-strain *fumC* and *fimH* alleles, we looked at 11 samples containing unique CH types (i.e. without allele sharing). In these 11 samples, the predicted prevalences of *fumC* and *fimH* were highly correlated (0.99, Fig. 7). First, we calculated the absolute difference between the predicted *fumC* and *fimH* prevalence for each matched pair of alleles. Next, each absolute difference was divided by the predicted *fumC* or *fimH* prevalence to obtain a relative deviation (Fig. 8). Finally, we used the relative deviations to derive an equation for the maximum acceptable difference between matching *fumC* and *fimH* alleles (Fig. 8).

While some samples, like those discussed above, contain only unique CH types, others contain CH types with shared alleles. For example, in a sample containing 30% *H*30 and 70% ST131, which share *fumC*40, the prevalence of *fumC*40 is not representative of either *H*30 or ST131 prevalence. For such samples, the minority rule was applied to resolve the strain content. Thus, under the minority rule, the percentage of *H*30 in the example above would be determined by *fimH*30, rather than *fumC*40, since the *fimH*30 prevalence is smaller. We tested this approach on both the *H*30 and the 18-sample analysis described above to see if this resolved outliers. In both cases, using the minority rule removed outliers and improved the correlation between predicted and experimental prevalence (Fig. 9). Thus, we were able to assign strain content and strain prevalence in all samples, including samples with allele sharing.

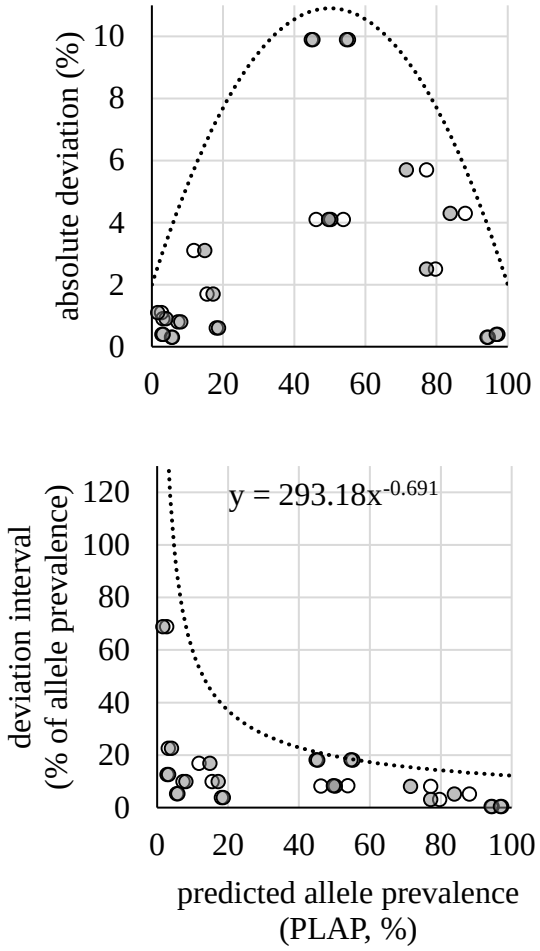


Figure 8. Difference in predicted prevalence between *fumC* and *fimH* alleles from the same *E. coli* strain. Deviation in absolute numbers is shown on the top. Deviation as a percentage of the prevalence of the allele is shown on the bottom. Open circles indicate *fimH* data points. Shaded circles indicate *fumC* data points. Trend lines and equations were used to determine intervals for matching (i.e. belonging to the same CH type) *fumC* and *fimH* alleles.

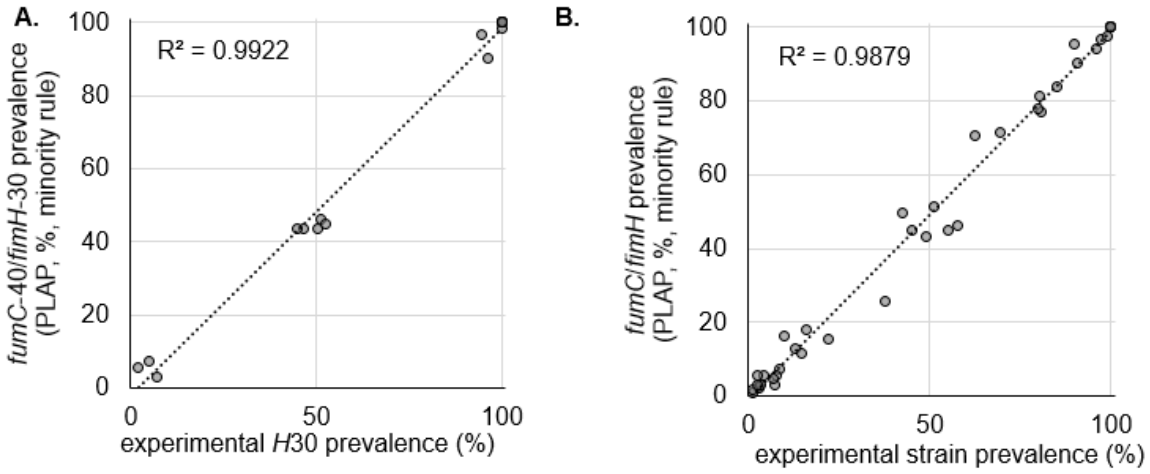


Figure 9. A. Comparison of actual *H*30 load in *H*30-containing fecal samples to PLAP-predicted *fumC*-40/*fimH*-30 prevalences with minority rule correction (i.e. the smaller prevalence of the two was used). Prevalence of *fumC*-40/*fimH*-30 is expressed as percentage of all *E. coli* in each sample. *H*30 load is expressed as ratio of *H*30 (ciprofloxacin-resistant) single colonies to all plated *E. coli* single colonies in percent. **B.** PLAP-predicted allele prevalence (with minority rule correction) compared to experimental allele prevalence as determined by surveying at least 40 single colonies per sample.

Predicted strain diversity of fecal and urine samples

Using the equation described above, we were able to classify all samples in our study into 4 categories (see Fig. 10): samples with only one CH type (uniclonal); samples with multiple unique CH types (unambiguous); samples with one dominant unique CH type and multiple minor non-unique CH types (ambiguous-simple), and samples where the dominant CH type was not unique (ambiguous-complex). Fecal samples were 33% uniclonal, 23% unambiguous, 21% ambiguous-simple, and 23% ambiguous-complex. Urine samples were 54% uniclonal, 8% unambiguous, 25% ambiguous-simple, and 12.5% ambiguous-complex.

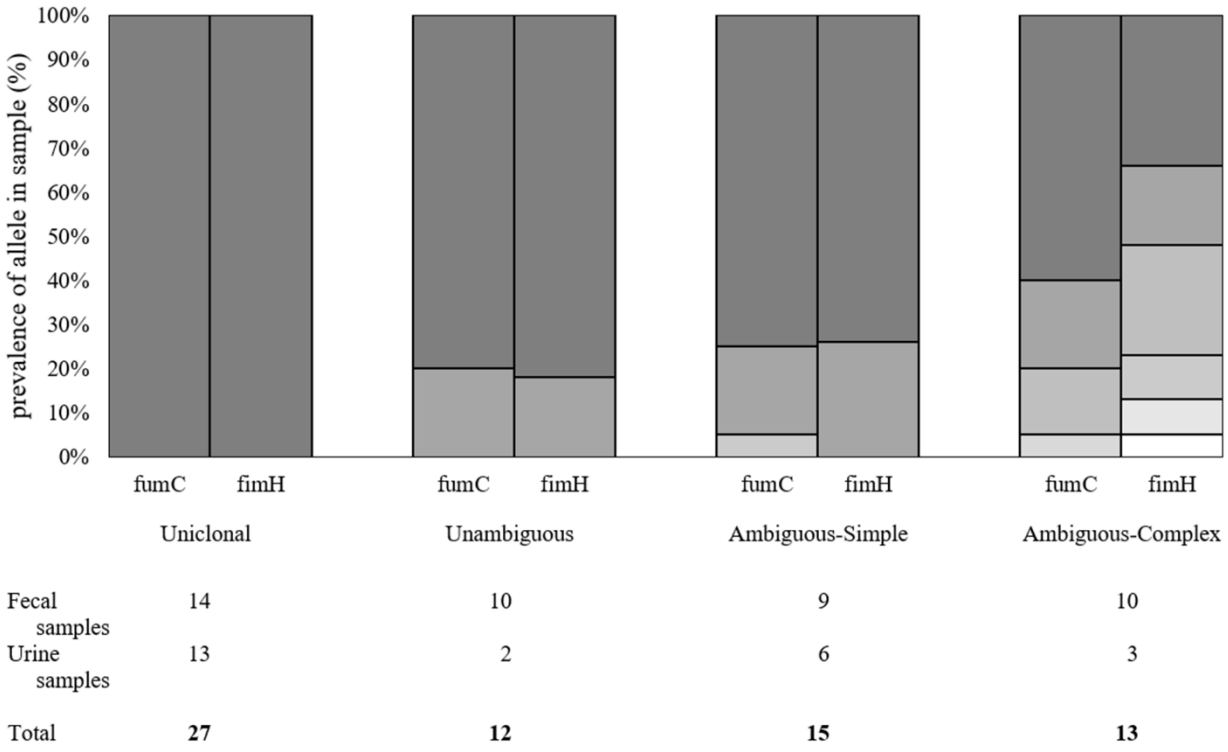


Figure 10. Representative examples of each sample category defined by within-sample breakdown of prevalence for *fumC* and *fimH* alleles. Number of fecal and urine samples belonging in each category is listed below.

Overall, 107 fecal and 48 urine strains were predicted, corresponding to 68 clones in fecal samples and 33 clones in urine samples. Of these clones, 50 (73.5%) and 24 (73%) were found in Enterobase, respectively. A sample of the clones detected can be found in Table 1. A full list can be found in Supplemental Table 1.

Out of the 155 total strains predicted, 6 were *fumC*-null (3.9%) and 2 were *fimH*-null (1.3%). This is congruent with the occurrence of null alleles in our 18-sample subset, where 1 (3%) out of 35 total strains predicted was a null-allele strain.

The average number of strains per sample was 2.47 ± 1.32 for fecal samples and 1.96 ± 1.40 for urine samples. Based on Enterobase's ST-phylogroup data, we determined that B2 was the most common (14 out of 47, 30%) among non-criterion fecal strains. Other phylogroups

included A (26%), B1 (19%), C (8.5%), D (11%), E (2%), and F (4%). Non-criterion strains in urine samples included strains from phylogroups B2 (8 out of 16, 50%), B1 (19%), D (19%), A and F (6% each).

Participant	Sample	*fumC* allele	*fumC* allele prevalence (%)	*fimH* allele	*fimH* allele prevalence (%)	Predicted CH type (ST type)	Phylogroup
P2	First fecal	**4**	**81.2%**	**34**	**94.2%**	**4-34 (ST399) 80.2%**	C
		214	**13.1%**	***30***	**3.1%**	**214-0 (ST1316) 12.8%**	A
		40	**4%**	**9**	**2.7%**	***40-30 (H30) 3.5%***	B2
		24	**1.6%**			**24-9 (Novel) 1.2%**	
	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100%*	B2
	Second fecal	**40**	**100%**	**30**	**100%**	**40-30 (*H*30) 100%**	B2
P11	First fecal	40	75.60%	5	49.90%	40-5 () 49.9%	B2
		4	12.90%	*21*	23.00%	*40-21 (ST357) 23%*	B2
		11	11.50%	35	13.50%	11-34 () 11.5%	A
				34	12.80%	4-35 () 13.5%	C
	Urine	***40***	**80%**	***21***	**78.10%**	***40-21 (ST357) 78.1%***	B2
		38	**20%**	**5**	**16.40%**	**38-5 (ST569) 16.4%**	B2
				35	**5.50%**	**38-35 () 5.5%**	B2
	Second fecal	23	98.50%	38	98.20%	23-38 () 98.2%	B1
		40	1.60%	*21*	1.80%	*40-21 (ST357) 1.6%*	B2

Table 1. Sample of predicted CH types. CH types and alleles marked in italics are criterion CH types/alleles. CH types and alleles marked in bold are confirmed to be present by single colony isolation and sequencing. Enterobase was used to translate CH type into strain (ST) type, and to determine phylogroup.

Novel clones

17 fecal samples (40%) and 8 urine samples (33%) in our study were found to contain at least one novel CH type. This included 19 fecal and 9 urine CH types not found in Enterobase. Of these, 5 fecal and 3 urine CH types included at least one novel allele, and 14 fecal and 6 urine CH types were combinations of *fumC* and *fimH* that were not previously observed (novel CH combinations). Both CH types involving novel alleles and novel CH combinations were observed to be primarily low-frequency clones. The average predicted prevalence for novel CH combinations was 8.7% ± 3.5% SEM (range 1-64.2%), and 13 out of 20 novel CH combinations had predicted prevalences of <5%. One such combination was confirmed in our 14 characterized sample set, consisting of *fumC*24 and *fimH*9, with a predicted prevalence of 1.6% and experimental prevalence of 1.2%.

Similarly, 7 out of 8 novel allele-containing CH types had predicted prevalences of <2%. The remaining CH type had a predicted prevalence of 70.7% and was detected using single colony typing. The novel *fumC* allele was paired with *fimH*47 and was verified to be 8 SNPs away from the closest known allele. The remaining MLST gene alleles for this strain were *adk*46, *icd*260, *mdh*160, *gyrB*266, *purA*1, and *recA*221.

Clones below error threshold

To ascertain if we could identify alleles at prevalences below our defined error threshold of 0.8%, we ran PLAP on the set of 14 multi-allele samples using an error threshold of 0.5%. In 8 and 6 samples, respectively, prevalence of *fumC* and *fimH* alleles was <0.8%. None of the alleles corresponded to known *fumC* or *fimH* alleles. These apparent novel alleles clustered alongside known alleles identified in the sample (Fig. 11, 12), leading us to conclude that these arose due to sequencing or amplification error rather than belonging to clonally different strains.

Predicted strain diversity in H30-containing urine and fecal samples

Strain diversity in first fecal samples was comparable with diversity in second fecal samples (paired t-test, $p>0.1$). Distinguishing between *H*30-containing and non-*H*30 samples showed that there was no statistical difference in strain diversity between *H*30-containing and non-*H*30 fecal samples of either kind (unpaired t-test, $p>0.1$), and that there was no difference in diversity between first and second fecal samples in either non-*H*30 or *H*30-containing samples (Fig. 13, paired t-test, $p>0.1$). Both *H*30 and non-*H*30 urine samples were less diverse than corresponding fecal samples (paired t-test, $p<0.01$ and 0.02, respectively). However, *H*30 urine samples were less diverse than non-*H*30 urine samples (t-test, $p=0.04$).

It is also noteworthy that in 6 out of 23 *H*30-containing fecal samples, *H*30 was the only strain predicted, indicating that it may be fully dominant in the gut niche in these participants.

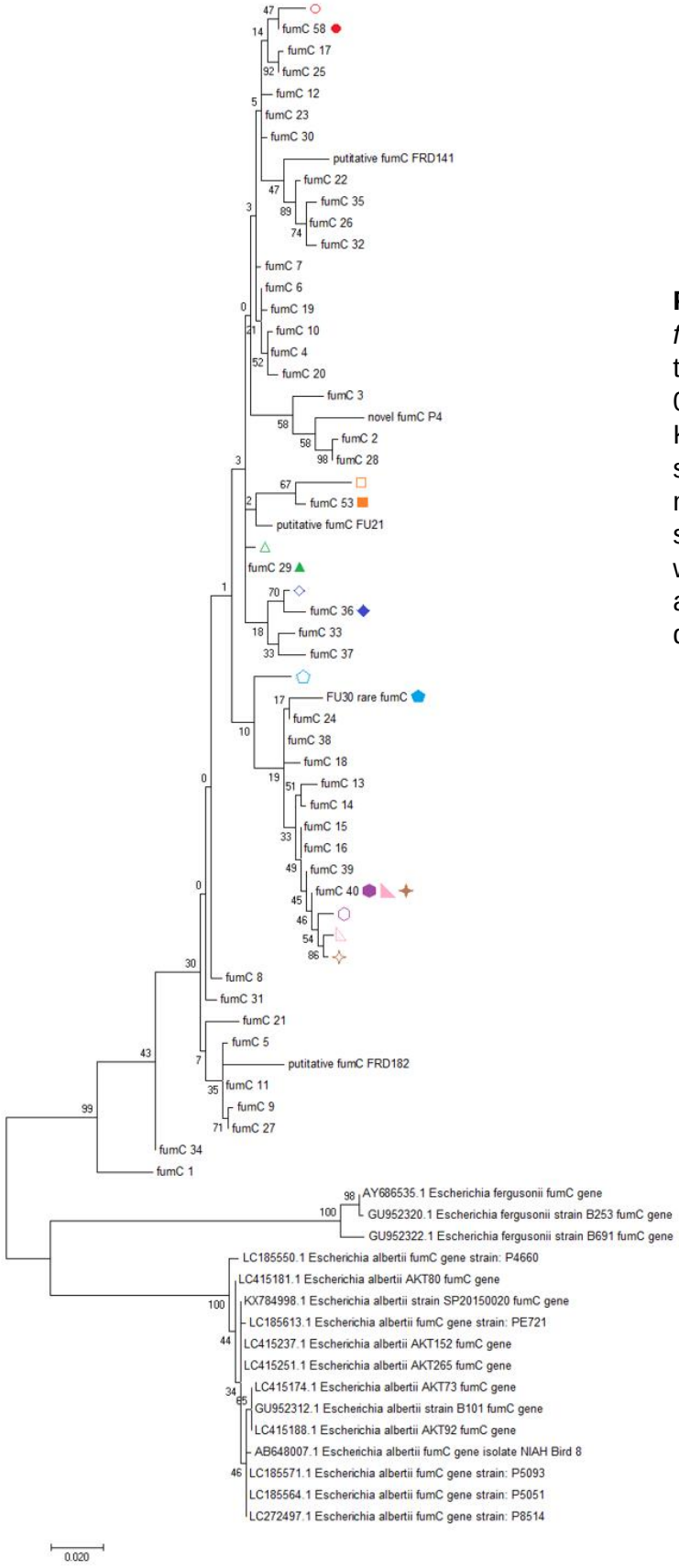


Figure 11. Putative rare novel *fumC* alleles identified by lowering the error threshold from 0.8% to 0.5%, marked in open shapes. Known alleles from the same sample as the rare novel allele are marked in filled-in shapes of the same type and color. FumC-40 was present in 3 different samples and therefore is marked by 3 different shapes.

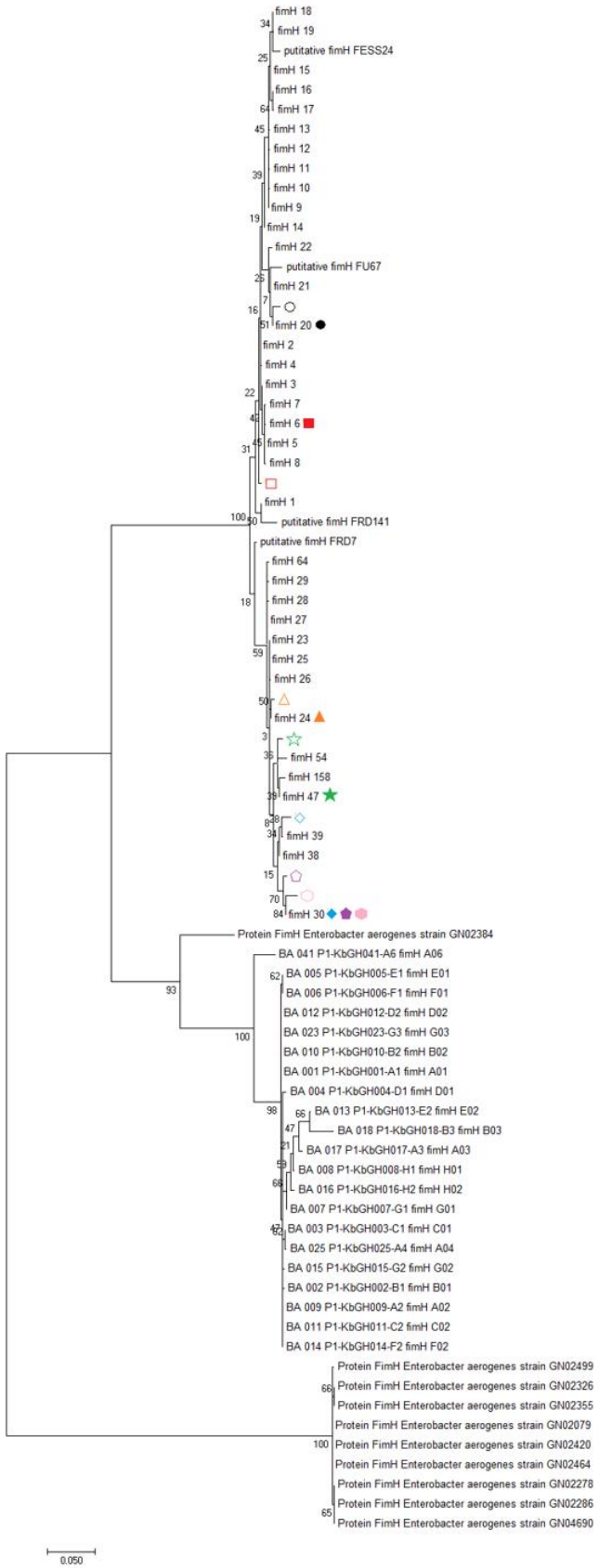


Figure 12. Putative rare novel *fimH* alleles identified by lowering the error threshold from 0.8% to 0.5%, marked in open shapes. Known alleles from the same sample as the rare novel allele are marked in filled-in shapes of the same type and color. FimH-30 was present in 3 different samples and therefore is marked by 3 different shapes.

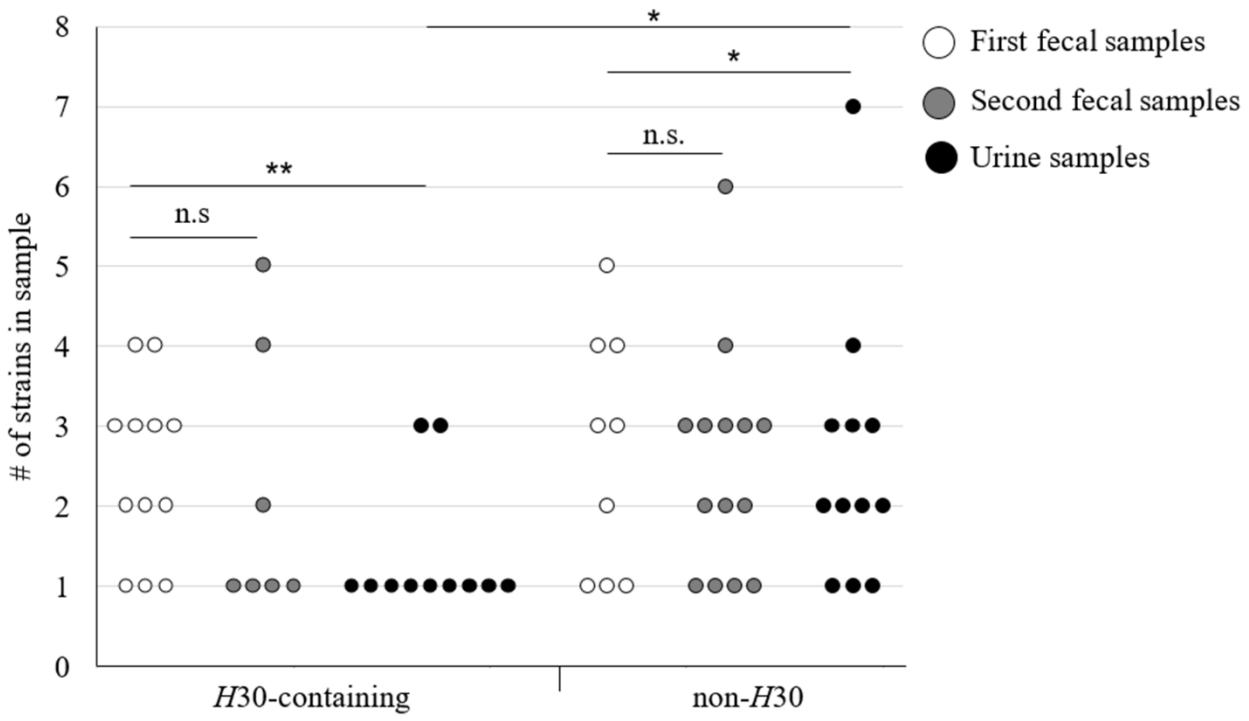


Figure 13. **Diversity of *E. coli* in individual fecal/urine samples.** *H*30 content was determined by PLAP and/or culturing.

Strain turnover in fecal samples

There was no correlation between number of strains in the first and second fecal sample, as well as no correlation between number of strains in the urine sample and either fecal sample (Fig. 14). When comparing the strain content of first and second fecal samples, we found that 92% of non-criterion strains appeared to be transient i.e. were detected in one of the fecal samples only. Transient non-criterion strains were also skewed towards lower-frequency strains (t-test, $p<0.001$, Fig. 15B). It is possible that these strains are present in both fecal samples but are below our limit of detection in one. However, we find that in one participant (P2, Suppl. Table 1) the first fecal sample contains 3 ciprofloxacin-sensitive non-criterion strains while the second fecal sample contains only ciprofloxacin-resistant *H*30 as verified by single colony

testing. This leads us to believe that there may be significant strain turnover in our fecal samples overall.

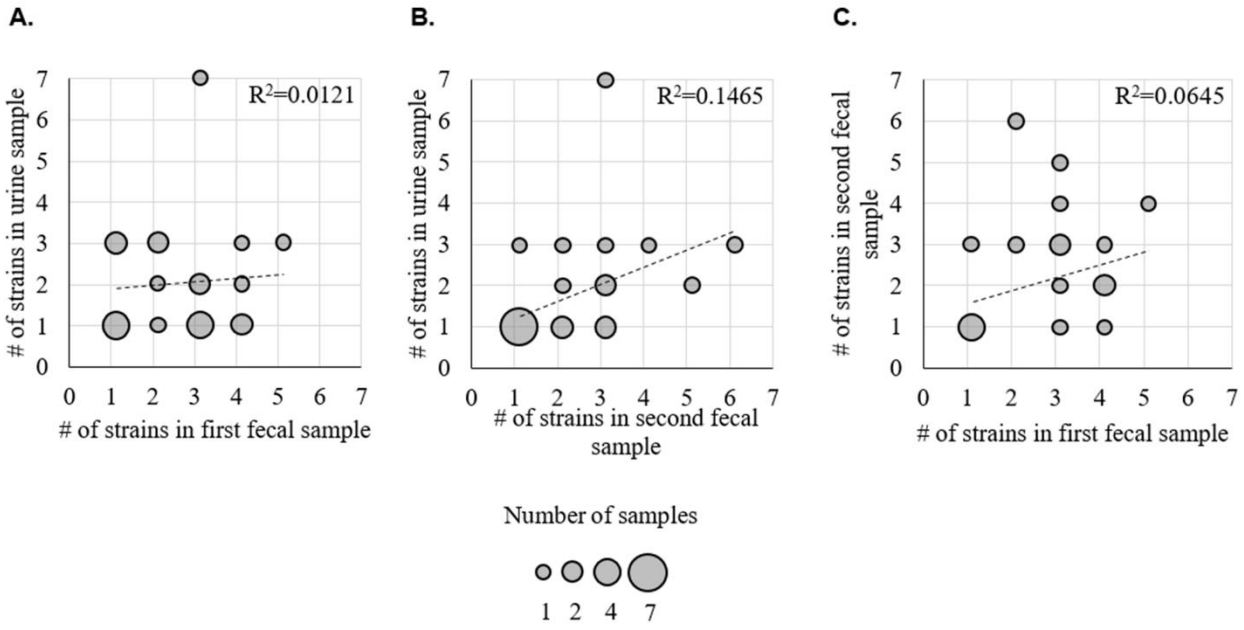


Figure 14. Counts of *E. coli* strains in fecal and urine samples. Number of strains detected by PLAP in (**A**) first fecal vs urine, (**B**) second fecal vs urine, and (**C**) first fecal vs second fecal samples. Each bubble indicates participants with the corresponding number of *E. coli* strains in the designated sample. The bubble size indicates number of participants with the determined number of strains. Linear fit with Pearson square correlation index shown.

Discussion

We combined conventional *fumC/fimH* typing with deep amplicon sequencing to assess *E. coli* clonal diversity in a high-throughput manner. Our method has several advantages over existing protocols. Firstly, our method has high sequencing resolution for target species. Since we only sequence *E. coli fumC* and *fimH*, we can generate ≥0.5 million reads per sample, yielding ≥5,000 reads per base. In contrast, metagenomic sequencing, which is nonspecific to target species, yields only 20 reads per base per genome (assuming a 5Mb genome). Secondly, our method assessed up to 46 samples per sequencing run. In contrast, MLST requires typing ≥100 single colonies per sample to capture the low-prevalence strains that PLAP detects. Finally,

while we developed PLAP for *E. coli*'s CH typing, PLAP is not limited to *E. coli* clonotyping and may be generalized to other MLST schemes. For those attempting to use or adapt our approach, we have provided guidelines for both the experimental and algorithm portions on PLAP's web page: github.com/marade/PLAP.

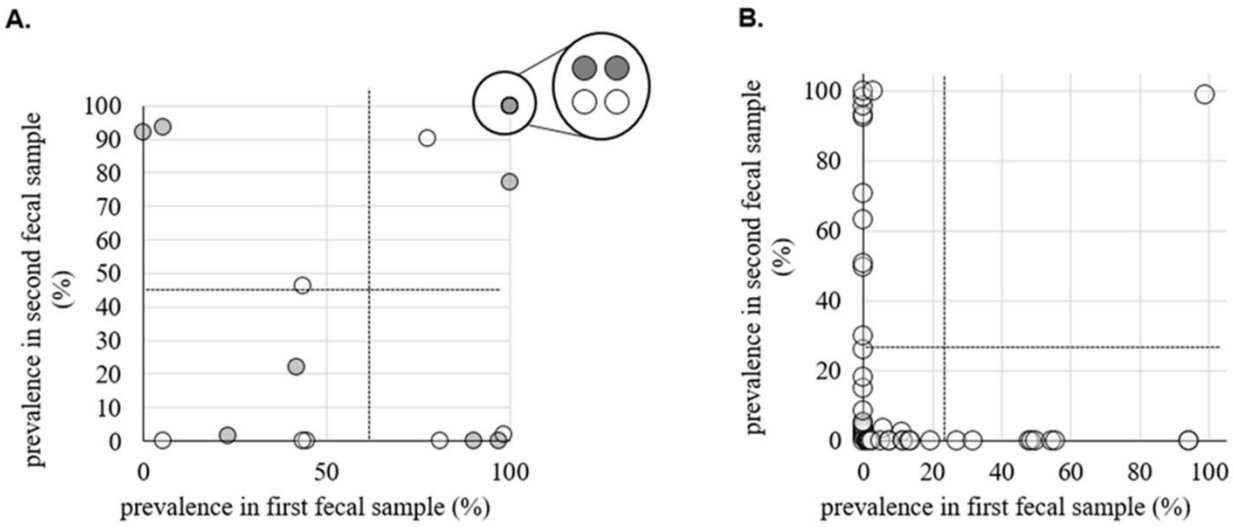


Figure 15. Persistence of *E. coli* strains in fecal samples. (**A**) Prevalence of criterion fecal strains in first vs second fecal samples. White data points indicate *H*30 strains while shaded data points indicate non-*H*30 strains. Circled cluster represents 4 strains present at 100% prevalence in both samples. Dotted lines indicate the mean prevalence for strains in first and second fecal samples. Distribution of prevalences in both first and second fecal samples is not significantly different from random (t-test, $p>0.05$). (**B**) Prevalence of non-criterion fecal strains in first vs second fecal samples. Dotted lines indicate the mean prevalence for transient strains in first and second fecal samples. Transient strains are defined as strains that are present in only one of the two fecal samples from the same participant. Distribution of prevalences in both first and second fecal samples is significantly skewed towards lower prevalences (t-test, $p<0.01$).

Despite studies showing that the healthy gut *E. coli* population typically includes multiple clones, we show that the pandemic multidrug-resistant subclone *H*30 can dominate the gut in healthy women, sometimes as the only detectable clone(39, 71–73). This builds upon previous research which has found multidrug-resistant bacteria in healthy people, and healthy people who appear to harbor only one gut clone(71). While we selected participants *H*30 in at least two out of three

samples, meaning that our study looked at “successful” *H*30 strains, total dominance is nonetheless concerning. Since antibiotic pressure was absent, these results indicates that *H*30 is potentially outcompeting other clones by alternative means. Whether these mechanisms are metabolic, or whether certain virulence factors give *H*30 an advantage is unclear, though previous studies have speculated that some virulence factors may be beneficial for *E. coli* gut survival(72, 74). Additionally, our study involved a small number of participants in which *H*30 was present in the gut and bladder. Therefore, it is possible that host differences play a significant role. Another novel observation was that *H*30 was the sole detected urinary strain more frequently than other clones, regardless of *H*30 gut dominance/non-dominance. This may indicate that *H*30 might be an especially well-adapted uropathogen, potentially explaining its association with UTI. Since it is unknown how ABU converts to UTI, further study into *H*30 dominance in both ABU and UTI are needed.

We also uncovered substantial diversity in our samples. This includes significant *E. coli* diversity in non-*H*30 urine samples from healthy women. Reports of multi-strain bacteriuria are rare, likely due to the convention of selecting one isolate per urine sample(52, 75). Therefore, it is unknown how common multi-strain bacteriuria may truly be. Remarkably, we also detected low-prevalence strains in the gut, some of which were novel clones, with up to 6 clones in a single sample. Gut *E. coli* diversity of this magnitude is supported by studies typing >200 single colonies per sample(71). Studies using smaller counts usually report fewer clones, indicating that there may be undescribed *E. coli* diversity when manageable numbers of colonies are used(39). Therefore, we believe that microbiome-like approaches to *E. coli* diversity are necessary to fully understand intra-species dynamics in both the gut and bladder.

Our approach does have limitations. Firstly, our lowest detectable strain prevalence is 0.8% of the *E. coli* population. This limit may be addressed in several ways including use of a high fidelity polymerase and preferential selection of *E. coli* colonies. However, we also recognize

that detection of rare strains may still prove difficult and that methods like ours may not fully replace current techniques. Secondly, our method relies on sub-culturing *E. coli*. We are aware that, theoretically, some strains could be suppressed during growth on selective media, forming no/smaller colonies and skewing prevalence results. However, we did not encounter this during our study. While amplification of *fumC* and *fimH* may be applied to urine samples without culturing, attempts at doing this directly from fecal samples were unsuccessful, possibly due to *E. coli* comprising <1% of the gut microbiome, making *E. coli* DNA too rare to effectively amplify. Therefore, we used culturing for all samples and believe that evaluating target species abundance using 16S sequencing is warranted in such cases. Lastly, we used antibiotic resistance for validation, which is not possible with clones/species where antibiotic resistance is absent or not strongly clonal. In these cases, validation using single-colony typing should be considered. These issues lower the reliability of our approach, but we believe that it remains an important step towards development of comprehensive clonal diversity (clonobiome) assessment tools for any species of interest.

CHAPTER 3: ASSESSMENT OF WITHIN-HOST *H*30 MICROEVOLUTION

Introduction

Uropathogenic *E. coli* (UPEC) are the primary cause of urinary tract infections (UTIs) – one of the most common types on infections, with approximately 11 million cases occurring per year in the US alone(46). While many UTIs are sporadic, ~25% will become recurring UTIs or UTIs that return approximately every 6 months(76, 77). These infections can be damaging to the bladder leading to increased bladder permeability and inflammation, pain during urination, and blood in the urine(78). UTIs can also travel up the urinary system into the kidneys (pyelonephritis) and from there, the bloodstream (urosepsis). The former can cause significant renal damage and the latter is fatal in >15% of cases(79, 80). Furthermore, many UPEC are drug- resistant – at least 20% but up to half in some regions – complicating treatment(81). The most common drug-resistant UPEC clone is the pandemic ST131-*H*30 (*H*30), which is resistant to multiple antibiotics, including fluoroquinolones(56). Additionally, *H*30 is associated with drug-bug mismatches, UTI in long-term care facilities, and negative clinical outcomes in the elderly and immunocompromised(57, 58). Despite this association with disease, recent findings show that *H*30 is also associated with asymptomatic bladder carriage in healthy people at higher frequency than other UPEC, and with higher persistence in the gut(48). Thus, there is a possibility that within- host adaptation plays a key role in both asymptomatic carriage and symptomatic disease involving this clone.

Preliminary evidence for *H*30 adaptation, and UPEC adaptation in general, has already been found. Studies of symptomatic UTI have shown that urinary UPEC sustain mutations in various genes, including those encoding flagellar components, iron permeases, and carbon storage regulators, as well as genes involved in various stress responses(52). Both gain and loss of plasmids was observed and presumed to occur in the fecal environment(52). Studies of symptomatic *H*30 infection have yielded similar results, with cystitis-causing *H*30 sustaining

mutations in protein secretion and sugar transport genes as it moved from the fecal to the bladder environment(59). Furthermore, cystitis-causing *H*30 acquired from a previous host with pyelonephritis underwent changes including mutations in transcriptional regulators affecting response to nitrogen limitation, ribose catabolism, and respiration, suggesting that microevolution between- or within-host could affect the severity of infection(59). Studies of UPEC isolated from carriers colonized asymptomatically have also shown adaptive changes occurring. The *E. coli* strain 83972, the most well-studied asymptomatic bladder-colonizing strain, accumulated mutations in genes affecting oxidative stress response, carbon storage, and iron uptake, when re-sequenced from inoculated hosts.(51) These data suggest that within-host adaptation may occur in UPEC, including in *H*30.

The question of whether true within-host adaptation occurs still stands however. Previous studies have used single-colony isolates, which are not necessarily representative of the population that they are isolated from. Furthermore, metabolic activity during colonization is known to differ between strains, meaning that potentially the nature of within-host adaptation may differ depending on the strain or clone studied. Thus, to determine if and how *H*30 in particular adapts in the host gut and/or bladder, studying *H*30 using a population-centered approach is required.

In order to investigate *H*30 within-host adaptation, we utilized a set of fecal and urinary samples from healthy *H*30-colonized female participants. Using population-level genomic sequencing, we were able to characterize the differences between fecal and urinary *H*30 populations, as well as the genetic heterogeneity within fecal and urinary *H*30 populations in these samples. Our findings showed that *H*30 primarily undergoes reductive changes in both gut and bladder and that respiration genes are disproportionally affected by functional changes in fecal *H*30 populations.

Results

Within-host relatedness of gut and bladder H30 isolates

In order to assess the similarities and differences between fecal and urinary *H*30, we must first establish their relatedness within our set of carriers. To do this, we sequenced a single-colony *H*30 isolate from each sample and used the algorithm kSNP3.0 to construct an allele-based phylogeny. All isolates clustered on the phylogeny by host, meaning that isolates from the same host were more closely related to each other than to isolates from other hosts (Fig.16). This confirms that *H*30 isolates from urinary samples are related to same-host fecal *H*30 isolates.

Genomic changes in fecal and bladder H30 isolates

In order to gain a preliminary assessment of differences between fecal and urinary *H*30 in our samples, we compared same-host isolate genomes to each other while considering the first fecal isolate genomes ancestral. We determined that genomic changes, in the form of SNPs, plasmids, and/or mobile regions present in at least one isolate but not all isolates from the same host, were observed in all sets of same-host fecal and urinary samples. The most common type of change was mutation, with a total of 107 SNPs observed in across all isolates. On average, isolates had 7.5 ± 3.4 SEM (range 0 – 52) SNPs versus isolates from the same host. Such SNPs were observed in 8 out of 11 sets of isolates, with 7 sets having SNPs unique to second fecal isolates only, 4 sets having SNPs unique to urinary isolates only, and 4 sets having SNPs shared by urinary and second fecal isolates only. 3 sets of isolates were isogenic. A visual representation of number of SNPs presents in same-host isolates can be found in Figures 17 and 18.

Loss of mobile regions or plasmids was observed as well. These included 9,952 bp genomic integrase-associated region and a 7,961 bp genomic fragment in the P9 urinary isolate, a 15,776 bp transposon-associated region present in both P1 urinary and second fecal isolates,

and a 3,489 bp portion of a previously-described plasmid (RefSeq ID NC_022651) present in both P7 urinary and second fecal isolates. No single-copy, virulence, or antibiotic resistance genes were lost in any of these regions. Gain of mobile regions or plasmids were observed in 2 sample sets. This included gain of 4,199 bp and 3,502 bp fragments of two previously-described plasmids (Genbank IDs CP012629 and MK416183) present in both P1 urinary and second fecal isolates, and 8,006 bp and 4,375 bp fragments of two different plasmids present in the P8 urinary isolate. The latter are undescribed but showed 97% and 89% identity to previously-described plasmids (Genbank ID LT985255 and LT985254, respectively). As with gene losses, it appears that no virulence or antibiotic resistance genes were gained in either case.

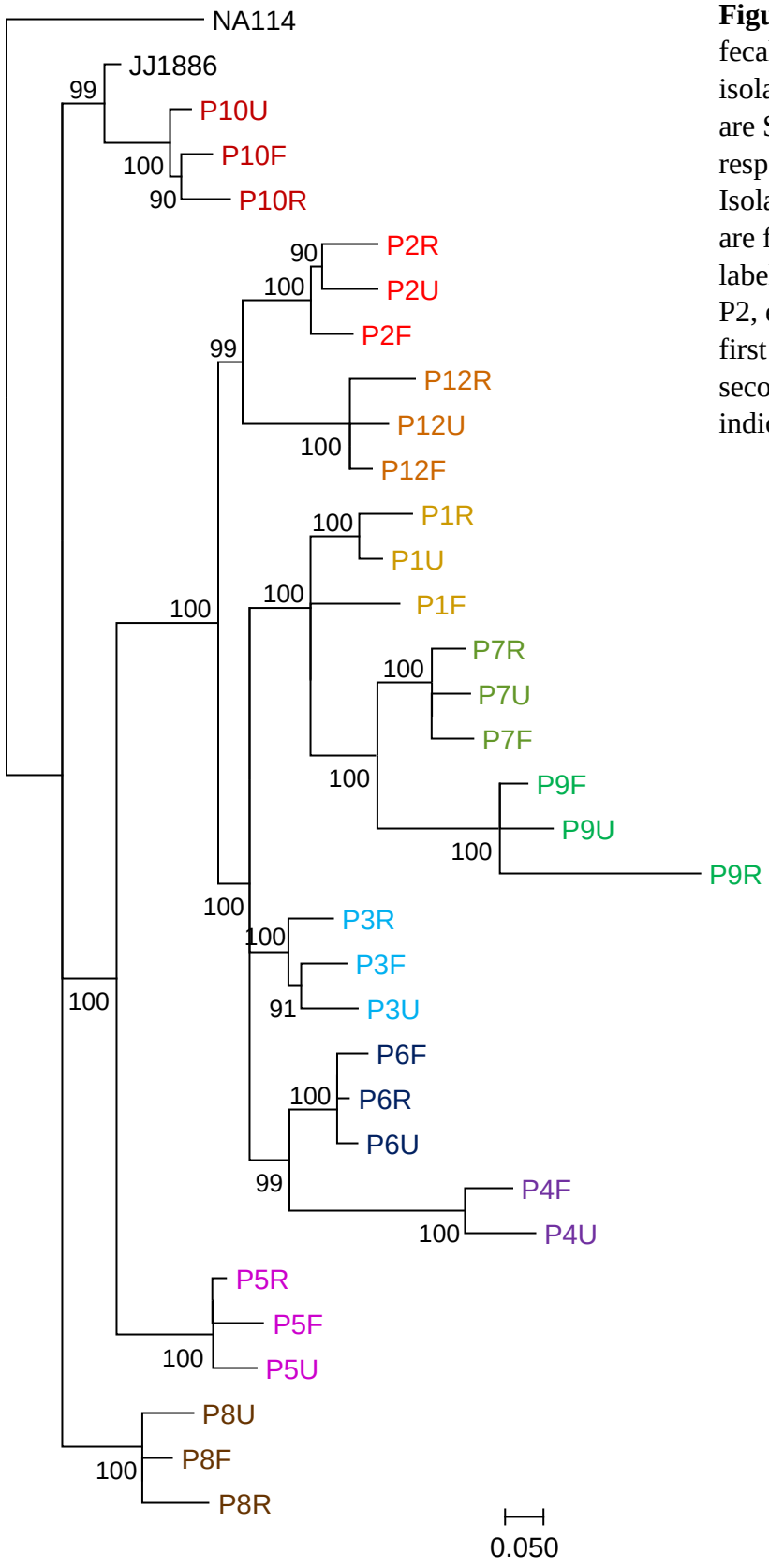
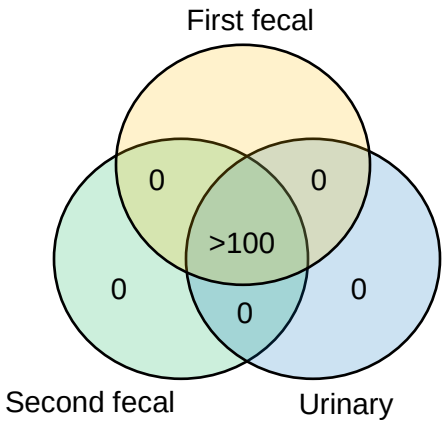
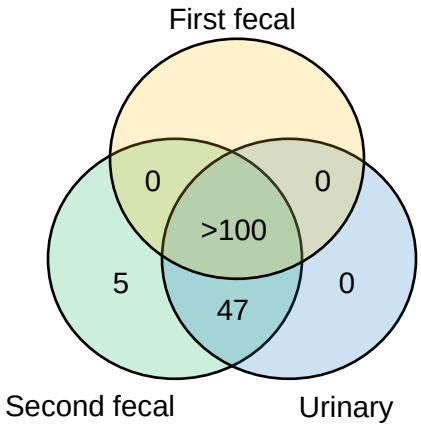


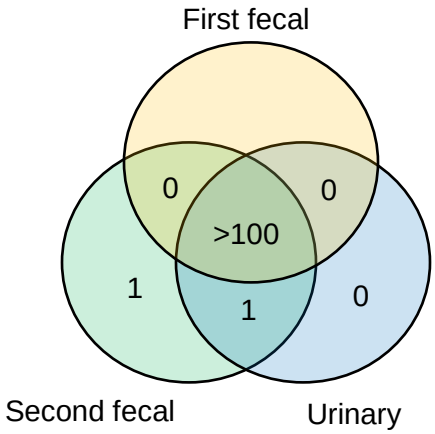
Figure 16. Phylogenetic analysis of fecal and urinary single-colony isolates of *H*30. NA114 and JJ1886 are ST131 and *H*30 cystitis genomes respectively, used here for reference. Isolates with the same color name are from the same host. Isolates are labeled by participant number (P1, P2, etc) and sample of origin (F for first fecal, U for urinary, and R for second fecal). Bootstrap values are indicated at nodes.



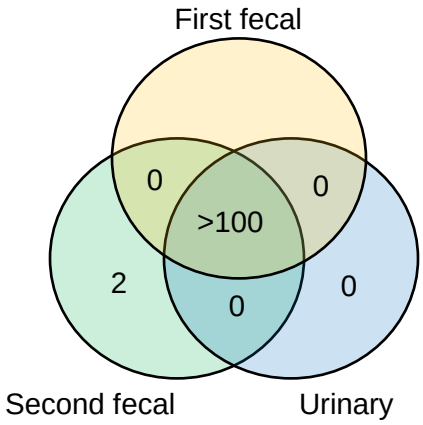
P2, P3, P12



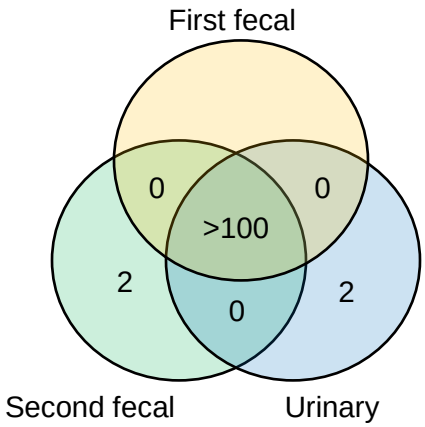
P1



P5

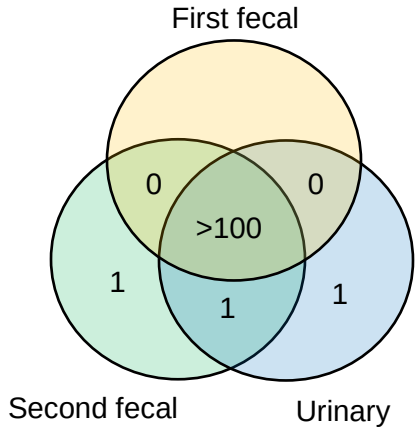


P6

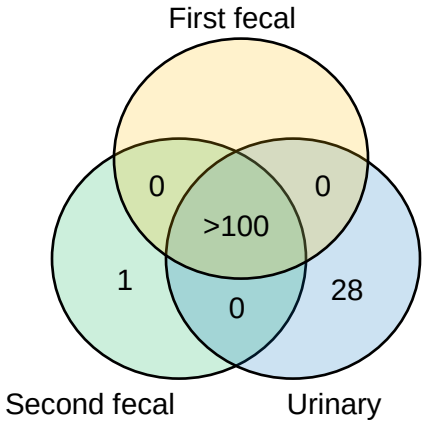


P7

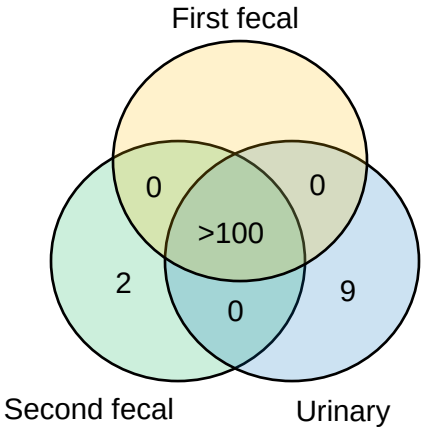
Figure 17. Number of SNPs differentiating same-host fecal and urinary *H*30 isolates. All first fecal positions were considered ancestral and therefore all SNPs are listed as present in urinary and second fecal isolates. Number of SNPs shared by all three isolates, indicated in the middle as >100, refers to number of SNPs differentiating same-host *H*30 isolates from the *H*30 reference genome, JJ1886. Isolates from participants P1-3, P5-7, and P12 are indicated here. See next figure for the remaining sets of isolates.



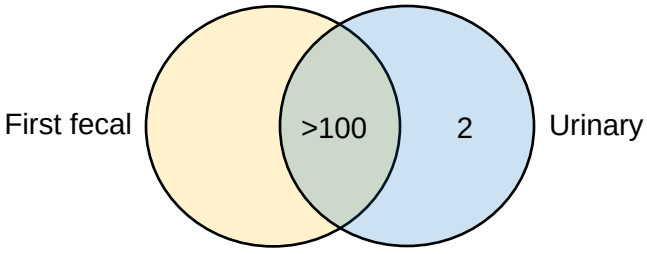
P8



P9



P10



P4

Figure 18. Number of SNPs differentiating same-host fecal and urinary *H*30 isolates. All first fecal positions were considered ancestral and therefore all SNPs are listed as present in urinary and second fecal isolates. Number of SNPs shared by all three isolates, indicated in the middle as >100, refers to number of SNPs differentiating same-host *H*30 isolates to the *H*30 reference genome, JJ1886. Isolates from participants P4 and P8-10 are indicated here. See above figure for the remaining sets of isolates. Participant P4 did not have *H*30 in the second fecal sample. Therefore, only first fecal and urinary isolates are depicted.

Validation of H30 fecal and urinary changes

To confirm that the SNPs observed in *H*30 isolates were truly present, we employed two different approaches.

Firstly, we employed PCR and sequencing of genes wherein the urinary isolate of participant P9 had a unique SNP when compared to the first fecal isolate. These included *barA, fliH, araH, frdD, psuG*, *waaZ*, *aceA*, and *fhlA*. All SNPs were confirmed to be present in the urine isolate and absent in the first fecal isolate.

Secondly, we used deep amplicon sequencing to both confirm the presence of isolate SNPs and to determine their within-sample frequency. For this approach we chose 5 genes that had SNPs in either one or two isolates from the same host: *cysZ* (P8 urine isolate only), *fldB* (P8 urine isolate only), *frdB* (P10 urine isolate only), *rpiA* (P10 urine isolate only), and *rssB* (P1 urine and second fecal isolates). For each gene, we amplified the gene from the urinary and second fecal samples, and sequenced the amplicons to >5,000X coverage. This sequencing confirmed the within-sample presence of all SNPs and their absence from samples whose isolates did not contain the SNP. Furthermore, the SNPs in *cysZ* and *fldB* were found to be at 90% and 95% within-sample frequency respectively, suggesting that these SNPs reached near-fixation in the P8 urinary population in which they were originally detected (Table 2). The SNPs in *rssB* was similarly at >95% frequency in both the P1 urinary and second fecal samples. The SNPs in *frdB* and *rpiA*, however, were found to be at 18% frequency in the P10 urinary sample, suggesting that these arose in the urinary population but did not reach fixation. This also suggests that other SNPs may exist in *H*30 populations that were not detected by single-colony sequencing.

Participant	Gene	SNP	SNP-containing isolates	Frequency in urinary population (deep seq %)	Frequency in second population (deep seq %)
P8	*cysZ*	P168L	Urinary only	88.9%	0%
	fldB	I133V	Urinary only	94.7%	0%
P10	*frdB*	F236Y	Urinary only	18.5%	0%
	rpiA	D74 (syn)	Urinary only	18.2%	0%
P1	*rssB*	A174S	Urinary and second fecal	98.5%	99.5%

Table 2. **Population frequencies of urinary and second fecal isolate mutations as determined by deep amplicon sequencing.** Frequencies are provided in percent of total reads aligned to the position. Coverage was at least 5,000X per position per gene. Synonymous SNPs are indicated as '(syn)'. None of the first fecal isolates from P8, P10, or P1 had the polymorphisms listed.

Population frequency of within-host H30 changes

Since not all mutations detected in fecal and urinary *H*30 isolates were found to be dominant, we decided to use a population-centered sequencing approach to determine the potential population frequency of mutations detected these isolates. By sequencing a pool of ≥200 *H*30 colonies from each fecal and urinary sample to a depth of ≥100X coverage across the genome, we generated population genomes for each sample. Thus, we were able to assess relatedness of *H*30 fecal and urinary populations, determine the population frequency of changes observed in isolates, as well as detect additional population-level differences. As with the isolate genomes, population genomes clustered together by host of origin, suggesting that *H*30

populations from the same host are more closely related to each other than to *H*30 populations from a different host (Fig. 19).

Overall, we determined that 87 out of 107 (81%) SNPs detected in fecal and urinary *H*30 isolates were dominant (>90% frequency) in corresponding population genomes. Of these, 32 (37%) were dominant in urinary population genomes but absent in other same-host population genomes. An additional 48 (55%) were dominant in both urinary and second fecal population genomes but were absent from corresponding first fecal population genomes, and another 7 (8%) were dominant in second fecal population genomes only. A total of 7 SNPs that were detected in *H*30 isolates were absent from corresponding population genomes. A further 7 were non-dominant in corresponding population genomes and had frequencies <60%. Additionally, 2 SNPs which were detected in urinary and second fecal isolates from participants P8 and P10 were found to be dominant in the corresponding population genomes, but were dominant and non-dominant in first fecal population genomes from these participants, respectively. These data confirm that a majority of mutations observed in *H*30 isolates were dominant in their respective populations, and that *H*30 populations undergo changes via genomic mutations in both gut and bladder.

Similarly to isolate SNPs, some of the gene loss and gain events detected in urinary and fecal isolates above, were also detected on a population level. The loss of the ~15k bp transposon-associated region in both P1 urinary and second fecal isolates was also observed in the corresponding population genomes, but was further expanded to a broader region of 76,781 bp. The gene losses detected in P9 and P7 isolates were not found in population genomes. However, the P8 second fecal population genome was found to have lost a 38,114 bp transposase-associated region and a 1,771 bp fragment of a plasmid (RefSeq ID NC_022662). Additionally, the P5 second fecal population genome was found to have lost a 9,088 bp fragment of a phage-associated region. The gain of genetic material observed in P1 urinary and

second fecal isolates was confirmed and expanded to include 9,417 bp, 5,014 bp, and 5,362 bp fragments of three described plasmids (GenBank IDs MK416183, CP035333, and LR134252, respectively). The gain of genetic material in the P8 urinary isolate was found to exist in all three P8 population genomes. These data show that while gene loss occurs in out sample set, gene gain is rare.

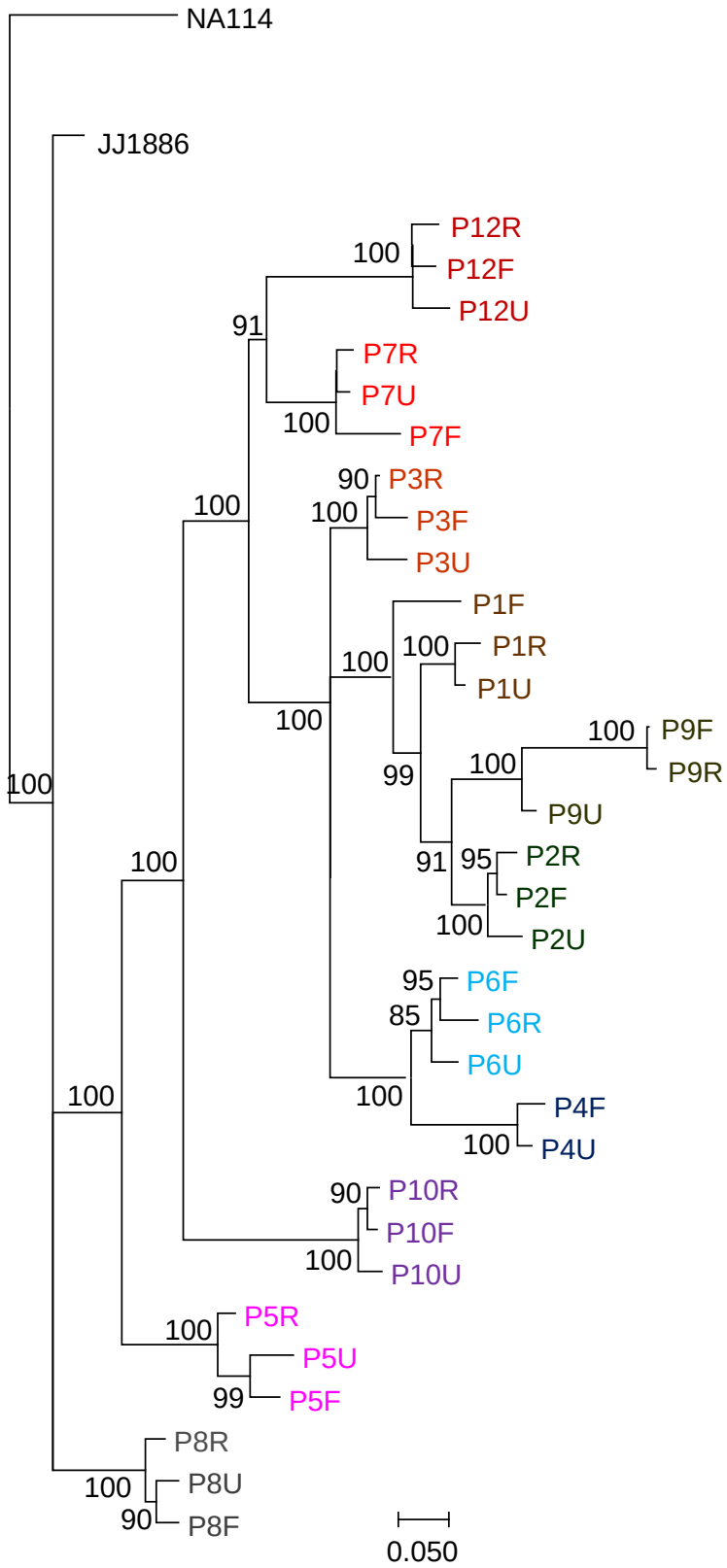


Figure 19. Phylogenetic analysis of fecal and urinary *H*30 population genomes (≥100X coverage). NA114 and JJ1886 are ST131 and *H*30 cystitis genomes respectively, used here for reference. Isolates with the same color name are from the same host. Isolates are labeled by participant number (P1, P2, etc) and sample of origin (F for first fecal, U for urinary, and R for second fecal). Bootstrap values are indicated at nodes.

Genomic heterogeneity in H30 fecal and urinary populations

Since some of the mutations detected in the *H*30 isolate sequencing proved to be non-dominant in the population genome sequencing, we were interested in the within-sample heterogeneity present in fecal and urinary *H*30. Thus, we investigated the population genomes for polymorphic positions i.e. positions at which at least two bases appear to be present in the population genome. We found that such heterogeneity was present in all fecal and urine samples with on average 29.8 ± 4.2 SEM (range 7 to 141) polymorphic positions per population genome. We found no significant difference between number of polymorphic positions in fecal and urine population genomes (paired t-test, p=0.4). Similarly, there was no difference in number of polymorphic positions between first and second fecal population genomes (paired t-test, p=0.2). We did not find any positions with more than two bases. A sample of polymorphic position can be found in Table 3.

The overall frequency of minority polymorphisms (i.e. bases at polymorphic positions that had a frequency of <50%) were on average 5.0% ± 0.2% SEM (range 1 – 50%). Of the 954 minority polymorphisms total were observed across all *H*30 population genomes, 750 (79%) had frequencies of <10%. Such polymorphisms occurred in all population genomes (see Fig. 20 for example). Minority polymorphisms at >10% frequency were similarly observed in all samples.

Of the 954 polymorphic positions observed across all *H*30 population genomes, 326 (34%) were positions were the polymorphism was synonymous, 450 (47%) were positions with nonsynonymous polymorphism, and 178 (19%) occurred in intergenic regions. Notably, while synonymous polymorphisms were present in most *H*30 population genomes (30 out of 32), all *H*30 population genomes contained positions with nonsynonymous polymorphism. This suggests that changes in *H*30 gut and bladder populations may be functional.

Gene	Polymorphism	Gene product	Frequency in P2 samples		
			First fecal	Urinary	Second fecal
mipA	S94N	MltA-interacting protein	0%	18.4%	0%
murQ	G247E	N-acetylmuramic acid 6-phosphate etherase murQ	0%	16.9%	96.1%
purL	G623 (syn)	phosphoribosylformylglycinamidine synthase	0%	5.3%	100%
uacT	V211I	Urate/xanthine permease	0%	1.7%	100%
fbaA	D215	ketose-bisphosphate aldolase	0%	64.7%	0%
uvrD	G669D	DNA-dependent helicase	11%	0%	97.3%
recQ	G372C	DNA helicase RecQ	0%	0%	97.3%
			Frequency in P12 samples		
			First fecal	Urinary	Second fecal
mdoH	P98S	glucan biosynthesis protein H mdoH	0.7%	0%	7.7%
entE	V307F	salicylate synthase	96%	92.1%	97.3%
nac	Upstream	LysR family transcriptional regulator nac	0.7%	0%	6.8%
cysZ	P168L	sulfate transporter CysZ	0.7%	24.2%	0.6%
tdcD	Upstream	propionate kinase	21.1%	31.3%	7.5%
ulaD	A93T	3-dehydro-L-gulonate-6-phosphate decarboxylase UlaD	0%	0%	6.9%
bsmA	L28R	lipoprotein BsmA	0.7%	0.9%	30.8%

Table 3. Polymorphic positions and their frequency in population genomes from select participants. All frequencies are reported as percentage of total aligned reads. Synonymous SNPs are indicated as ‘(syn)’.

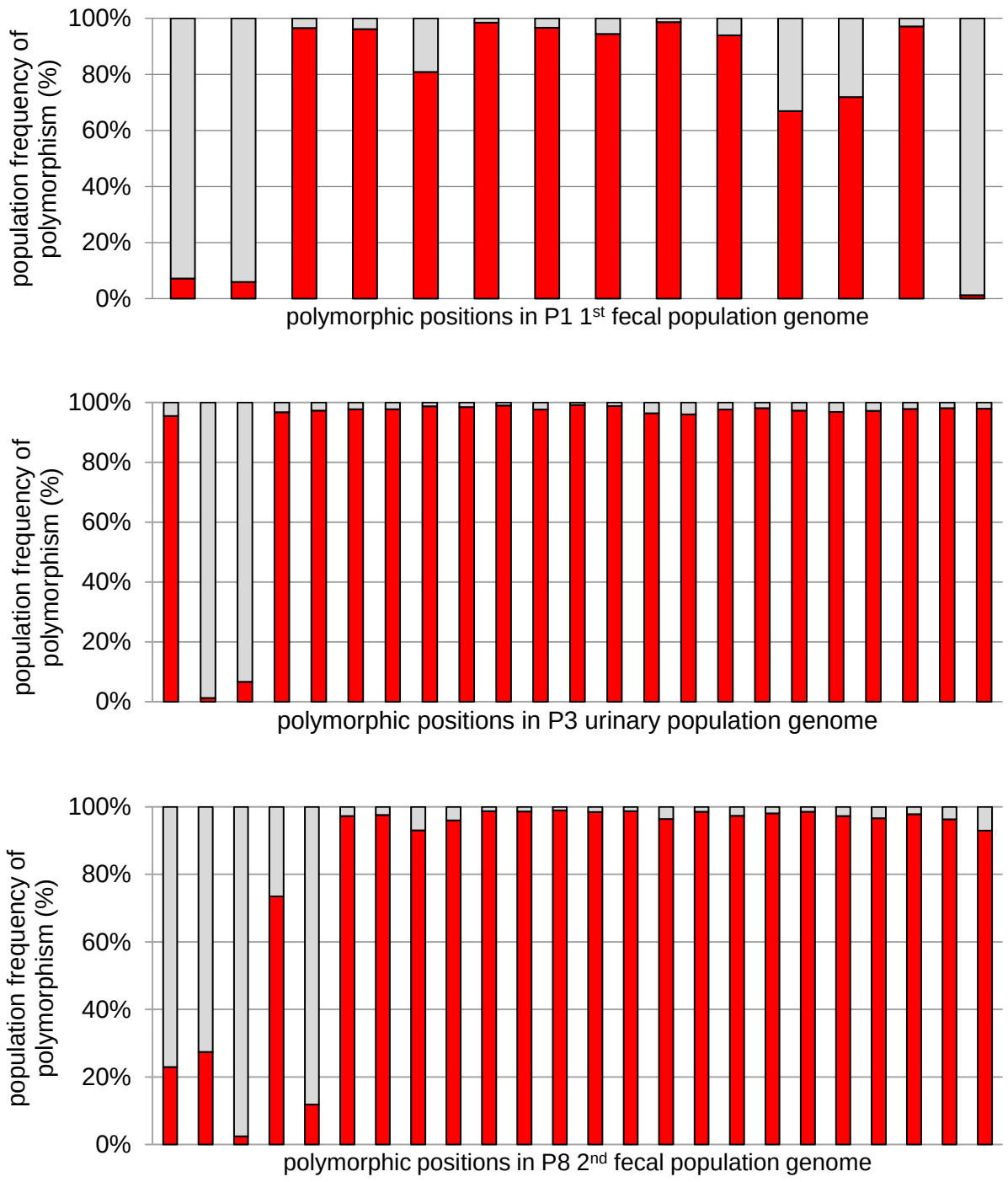


Figure 20. Polymorphic positions in select population genomes. Polymorphic positions were defined as those positions on the genome at which at least 2 bases were observed in the population genome, provided these were not located in multi-copy or mobile-element-associated genes. For each position, the frequency of the non-ancestral base is shown here in red. All positions are shown irrespective of their location on the genome.

Validation of H30 population genomic heterogeneity

Since the number of polymorphic positions detected in *H*30 fecal and urinary populations was substantial, we decided to validate it using two different methods.

First, we wanted to determine whether the polymorphisms detected were an artifact of sequencing the genome to a relatively high coverage. To test this, we took the single-colony *H*30 isolates sequenced from P1, P5, and P6 first fecal and urine samples above, and sequenced these to an average genomic coverage of ≥100X, similar to sequencing coverage of the population genomes. We found that the sequencing output for these deeper sequenced isolates showed polymorphism in phage-associated, transposon-associated, tRNA, rRNA, and multi-copy genes. However, we had already excluded these genes when assessing polymorphism in population genomes. Thus, we cannot explain our observations using depth of sequencing.

Secondly, we sequenced an additional 11 isolates from the P9 first fecal and urinary samples each. Thus, including the initial isolate we sequenced for each of these samples, we had a total of 12 isolate genomes per sample. Analysis of the 12 first fecal isolate genomes showed 3 polymorphic positions, all of which were also detected in the first fecal population genome. Minority polymorphisms at these positions had population genome frequencies of 5.7%, 7.8%, and 40%. Additionally, analysis of the 12 urinary isolate genomes showed 20 polymorphic positions, 19 of which were also detected in urinary population genome. Minority polymorphisms at these positions included 18 minority polymorphisms with population genome frequencies of 1.2% to 9%, and 1 polymorphism at 50% frequency in the urinary population genome. Thus, using these isolates, we were able to confirm the presence of polymorphic positions in both the first fecal and urinary population genomes.

Substructure in urinary and fecal H30 populations

Since the first fecal and urinary population genomes from participant P9 have been shown to be heterogeneous, we wanted to determine whether population substructure is the origin of this heterogeneity. In other words, we wanted to determine if subgroups existed in the first fecal and urinary *H*30 populations. In order to determine this, we compared all sequenced 24 *H*30 genomes to each other. We found that of the 12 first fecal *H*30 genomes, 6 were isogenic and possessed majority positions only i.e. did not have any bases that had population genome frequencies of <50%. 2 additional genomes diverged from this group by only one SNP each. 4 first fecal genomes differed from the majority group by one shared SNP with a 40% frequency in the first fecal population genome. 3 of these had no other changes, while the remaining genome had an additional deletion. Similarly, analysis of the 12 isolates sequenced from the urinary sample showed that there was a core set of 8 genomes which had majority positions only, and that this core set was split into two groups of 4 genomes each, with one group of genomes having a SNP at one position. The population genome frequency of this SNP was 50%. The remaining 4 urinary isolate genomes differed from the core set at 18 positions. These 4 genomes had minority positions at all 18 of these positions; additionally, these minority positions were ancestral. Further analysis showed that 3 of these 4 genomes were in fact isogenic with the majority group from the first fecal sample, while the remaining genome was isogenic with another first fecal subgroup. A tree based on the SNPs differentiating these 24 fecal and urinary isolates can be found in Figure 21. With these data, we show that there are at least 5 subgroups in the first fecal population and at least 4 subgroups in the urinary population. We also show that remnants of the first fecal population can be found in the urinary population.

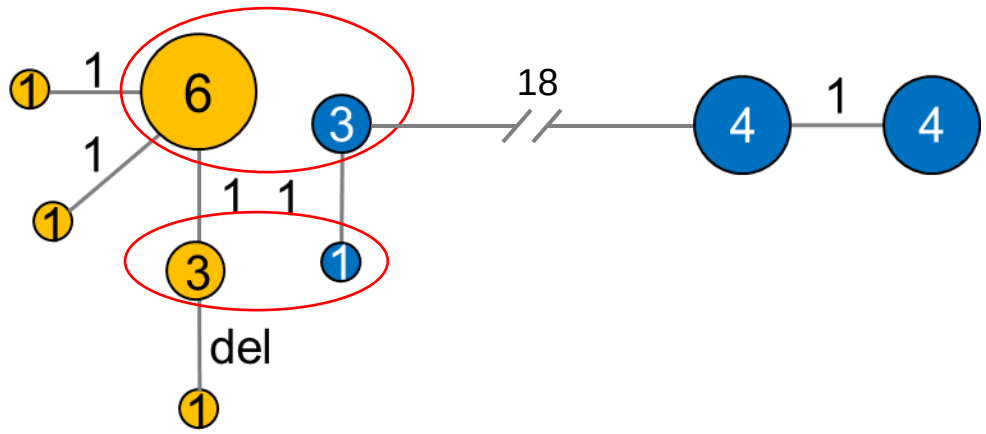


Figure 21. Phylogeny of sequenced P9 fecal and urinary isolates based on genomic SNPs. Number of SNPs differentiating isolates is indicated near branches. Yellow circles indicate first fecal *H*30 population subgroups. Blue circles indicate *H*30 urinary population subgroups. Number of isolates belonging to each subgroup is indicated in each circle. 'del' indicates a deletion was sustained by the subgroup. Isogenic fecal and urinary subgroups are indicated by circling the subgroups in red.

Since the number of subgroups detected in first fecal and urinary populations was based on sequencing of 12 single colonies per population, while the population genome which shows more heterogeneity involved sequencing of at least 200 single colonies, we wanted to determine how many subgroups may exist in the first fecal and urinary population genomes overall. In order to do this, we considered each unique genotype its own subgroup and used Monte Carlo simulation to project the number of subgroups from up to 12 single colonies to 200 colonies. For both first fecal and urinary *H*30 populations, the number of subgroups at 200 colonies i.e. in the population genome, was projected to be approximately 9 (Fig. 22). Thus, we determined that in first fecal and urinary *H*30 populations from participant P9, at least 9 subgroups exist each.

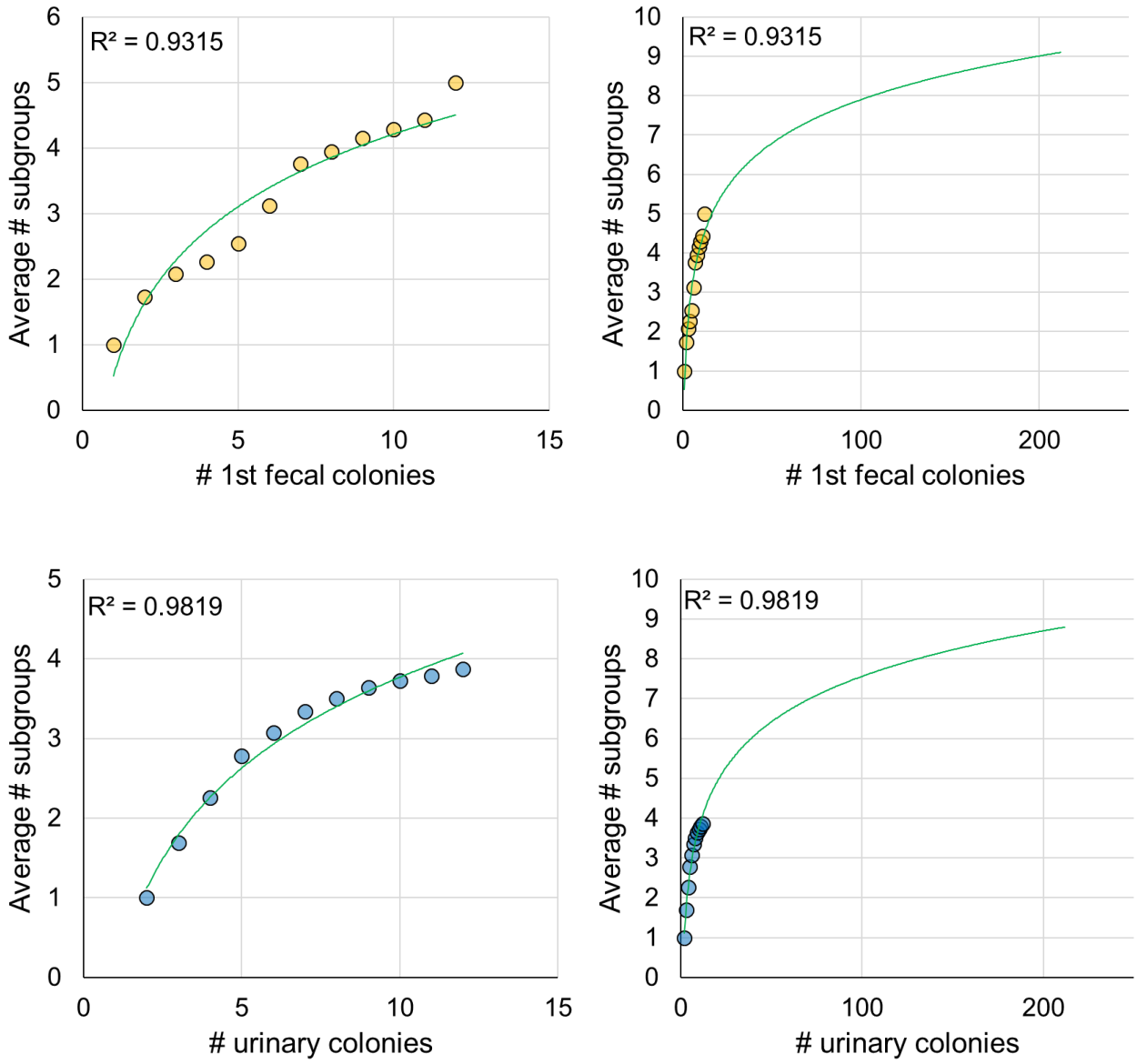


Figure 22. Projections of number of subgroups in first fecal and urinary *H*30 population genomes. Monte Carlo simulation was used to create a logarithmic trend line for number of subgroups in 1-12 colonies, then project number of subgroups in 200 colonies (number of single colonies used for population genome sequencing).

Patterns of within-host change in H30

Since we were able to observe heterogeneity within *H*30 in each fecal and urinary sample, we were interested in whether this heterogeneity is similar across samples from the same host. Firstly, were interested in whether heterogeneity in fecal *H*30 changes over time i.e. was different between the first and second fecal samples. To determine this, we plotted the frequency of each non-ancestral position (SNP) in the first fecal sample versus its frequency in the second fecal sample. We found that the frequencies of SNPs in same-host fecal samples were highly correlated ($R^2 > 0.7$) in 7 out of 10 participants with *H*30 in both fecal samples. In the remaining 2 participants (P1, P2, P10), the correlation between SNP frequencies between fecal samples was under 0.5 due to the accumulation of SNPs in the second fecal sample. However, some participants (P5, P6, P8) with high correlations between fecal samples also showed accumulation of singular SNPs in second fecal samples. These data indicate that while overall the fecal *H*30 population remains genetically stable in most participants, change via accumulation of SNPs may occur.

Secondly, we were interested in whether heterogeneity in urinary *H*30 was similar to heterogeneity in fecal *H*30 i.e. did the fecal *H*30 population undergo changes in the bladder. Using the same method as above, we found that SNP frequencies in urinary *H*30 were highly correlated ($R^2 > 0.7$) to SNP frequencies in same-host first fecal *H*30 in 7 out of 11 participants. As with fecal samples, the remaining urinary H30 populations (P1, P4, P8, P9) accumulated SNPs not present in same-host first fecal *H*30. Singular SNPs accumulating in urinary *H*30 from participants with high correlation between urinary and first fecal *H*30 were also observed (P2, P6). Furthermore, while the urinary H30 population in participant P10 did not accumulate novel SNPs, SNPs that were subdominant in first fecal *H*30 rose to near-dominance in urine. These data indicated that urinary *H*30 populations may undergo change via accumulation of SNPs.

Taking fecal and urinary *H*30 population comparisons together, we were able to observe 4 different patterns of genetic stability: 1) significant differences between first fecal and the other two *H*30 populations (time-ordered; P1), 2) significant difference between the urinary *H*30 population and both fecal *H*30 populations (urinary divergence; P9), 3) change or flux in both urinary and fecal H30 populations (instability; P2, P4, P6, P8, P10), and 4) overall *H*30 stability (stable; P3, P5, P7, P12). Representative examples of each pattern can be found in Figure 23.

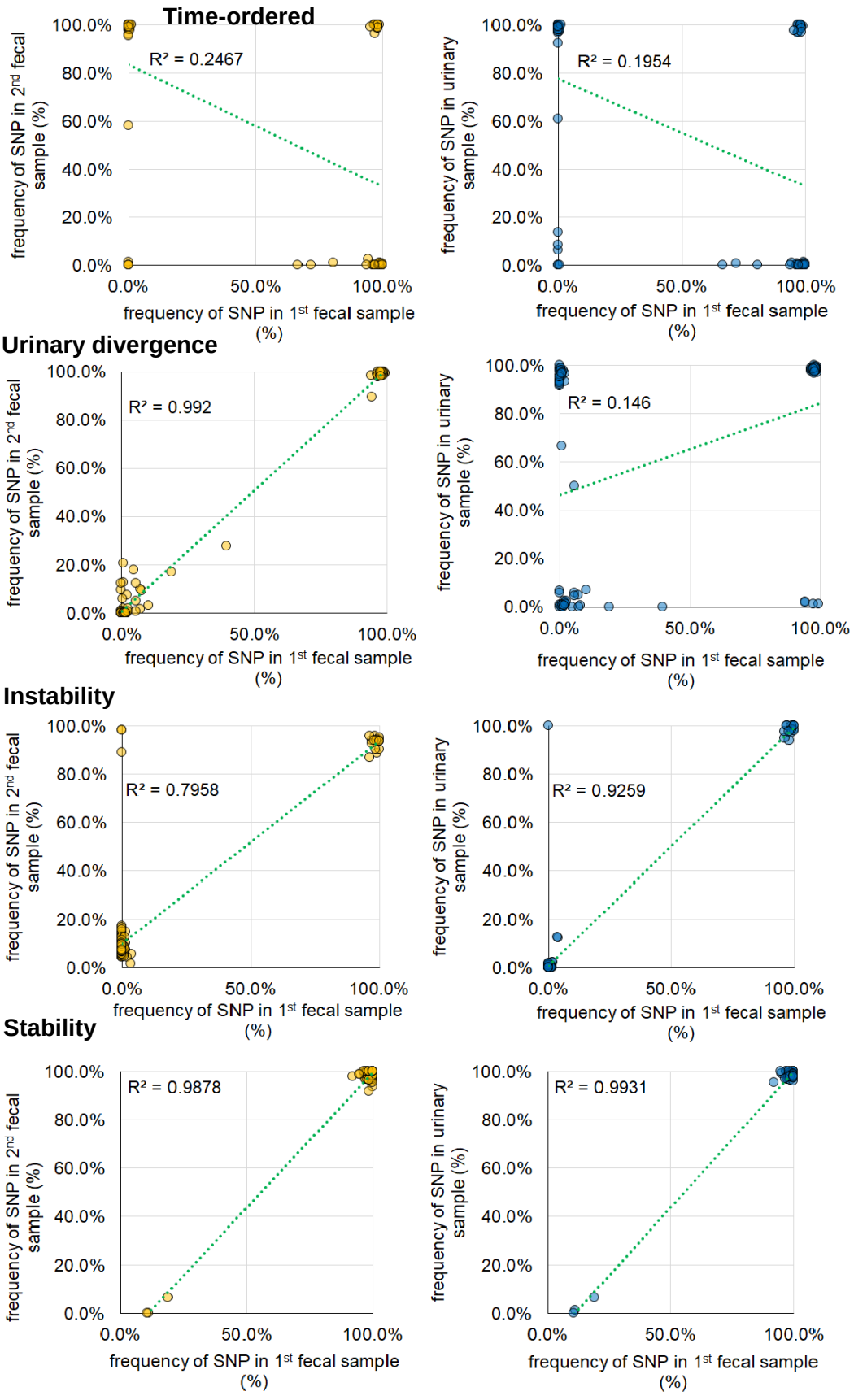


Figure 23. Comparison of SNP frequency between same-host samples. The frequency of each SNP in first fecal versus other samples was analyzed.

Location of changes in H30 fecal and urinary populations

Processes targeted by polymorphism in fecal samples included carbohydrate metabolism, DNA metabolism, stress response, iron acquisition, and amino acid metabolism, among others (Fig. 24). Assessment of synonymous versus nonsynonymous polymorphisms showed that respiration genes were affected by nonsynonymous polymorphisms more often than expected (Fisher's exact, p=0.012) while DNA metabolism genes were affected less often than expected (Fisher's exact, p=0.042). Processes targeted by polymorphism in urine samples were largely the same as those targeted in fecal samples (Fig. 25). However, compared to fecal samples, urinary samples had more nonsynonymous polymorphisms than synonymous polymorphisms, suggesting adaptation (Fisher's exact, p= 0.018). Furthermore, the number of nonsynonymous polymorphisms in carbohydrate metabolism genes was significantly fewer than synonymous polymorphisms (Fisher's exact, p=0.0309), and nonsynonymous polymorphisms occurred in processes not targeted by synonymous polymorphisms, including cofactor biosynthesis, virulence and defense, sulfur metabolism, and motility.

Between-sample changes, i.e. positions that were present in either urinary or second fecal population genomes but were absent in first fecal population genomes, included SNPs unique to urinary populations in genes affecting cell wall biosynthesis (*mipA, waaZ, lptA, pbp2*), central metabolism (*barA, rpiA, fbaA*), and anaerobic respiration (*frdD, torZ*), among others (Table 4). Changes unique to second fecal populations were detected in genes affecting cell wall biosynthesis (*lysM*, *plsC*), biofilm formation (*bsmA*, *hipB*), and growth under limiting conditions (*hexR*, *glnG*) (Table 5). Importantly, the carbon storage transcriptional regulator *barA* was targeted by SNPs in two urinary samples from different participants. This regulator, together with *uvrY*, serves as a switch that activates metabolic transition from glycolytic to gluconeogenic carbon sources (Fig. 26). A list of all polymorphic positions and all between-sample changes can be found in Supplemental Table 3.

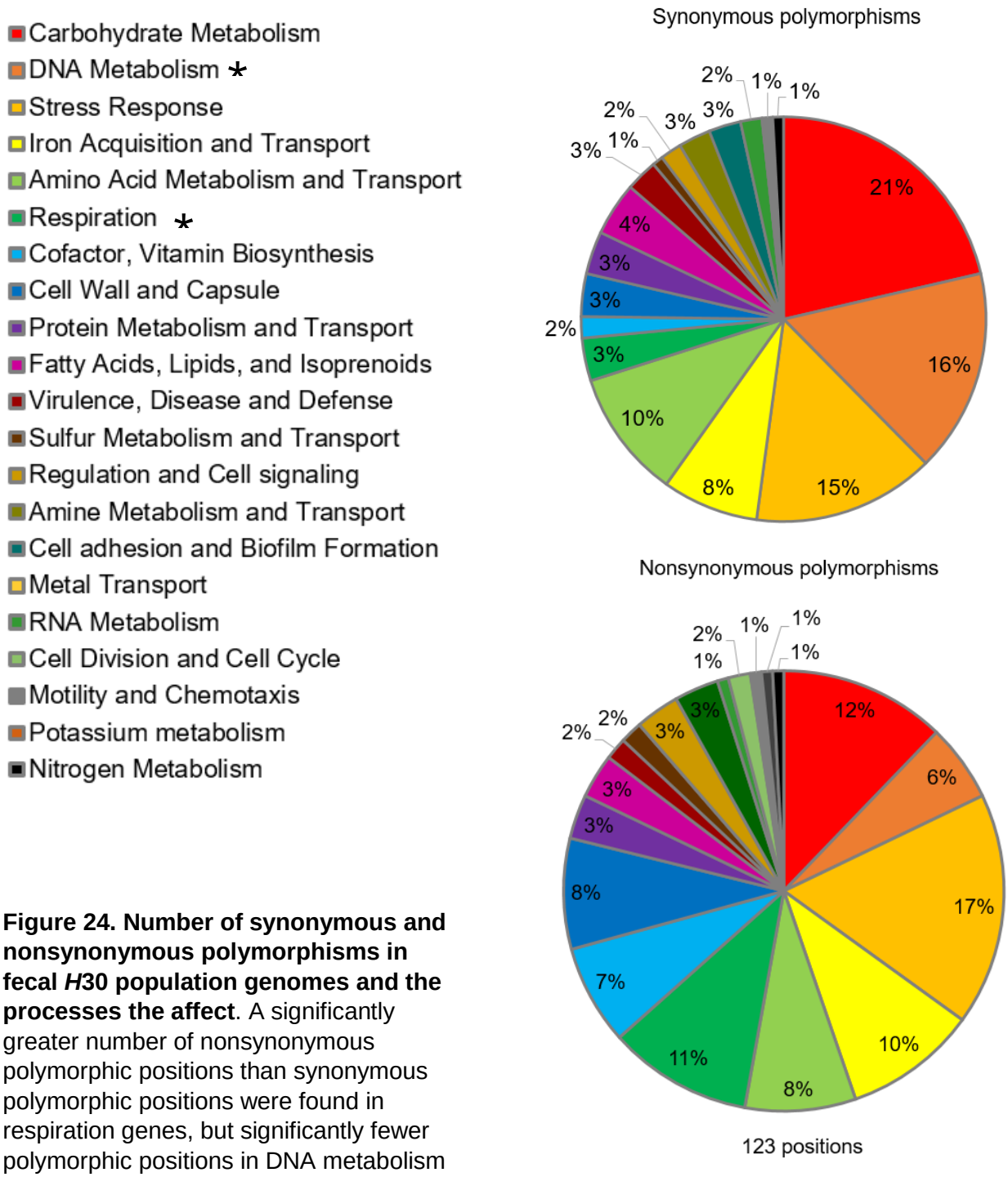
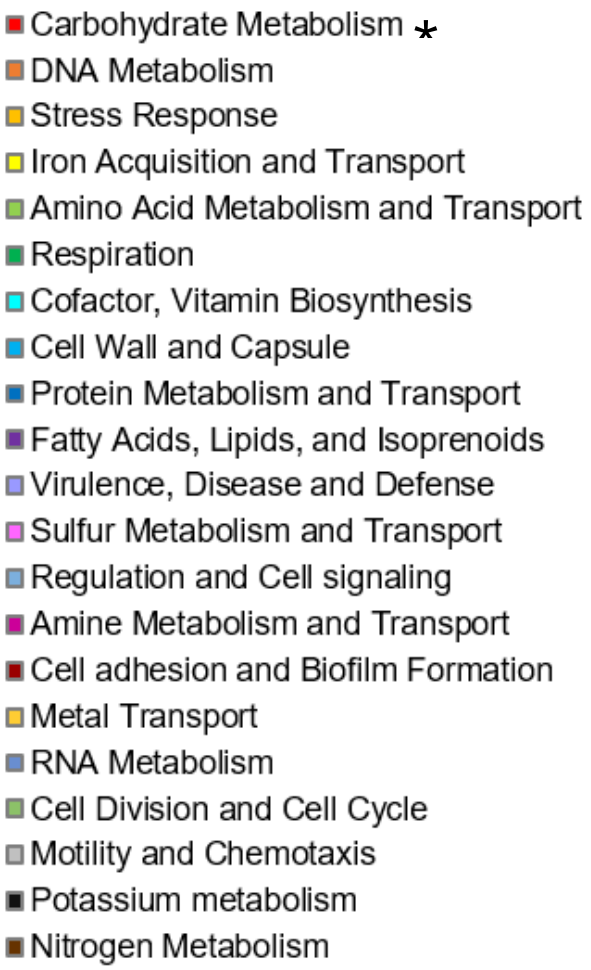
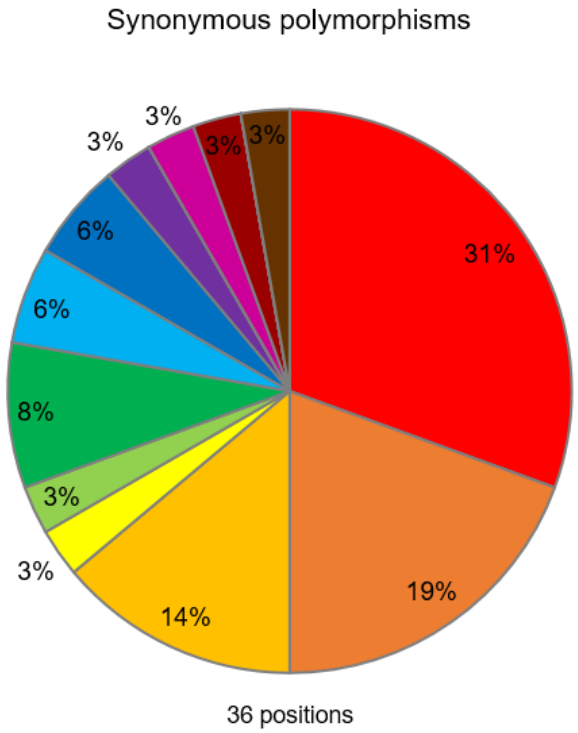


Figure 24. Number of synonymous and nonsynonymous polymorphisms in fecal *H*30 population genomes and the processes the affect. A significantly greater number of nonsynonymous polymorphic positions than synonymous polymorphic positions were found in respiration genes, but significantly fewer polymorphic positions in DNA metabolism genes.





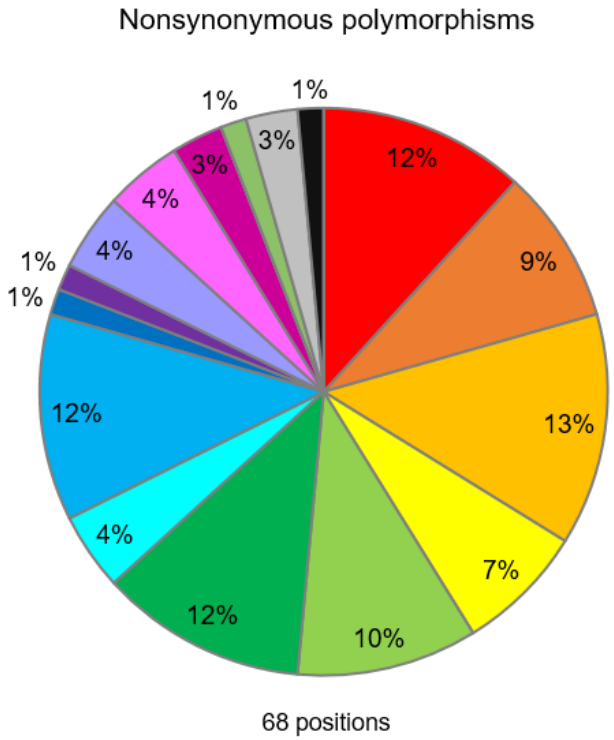


Figure 25. Number of synonymous and nonsynonymous polymorphic positions in urinary *H*30 population genomes and the processes they affect. A significantly fewer number of nonsynonymous polymorphic positions that synonymous polymorphic positions were found in carbohydrate metabolism genes.

Participant	Gene	Polymprphism	Gene product	Second fecal population frequency
P1	*mglB*	L326 (syn)	Galactose ABC transporter substrate-binding protein mglB	49%
	tldD	A29P	Metalloprotease TldD	96%
P2	*murQ*	G247E	N-acetylmuramic acid 6-phosphate etherase murQ	100%
	purL	G623 (syn)	phosphoribosylformylglycinamidine synthase	100%
	uacT	V211I	Urate/xanthine transporter	100%
	recQ	G372C	DNA helicase RecQ	97%
P5	*hexR*	L124 (syn)	MurR/RpiR family transcriptional regulator hexR	31%
	yfcV	upstream	Fimbrial yfcV	21%
P6	*wcaD*	L128V	Putative colanic acid polymerase WcaD	89%
	plsC	A46 (syn)	1-acylglycerol-3-phosphate O-acyltransferase plsC	98%
	caiA	A129 (syn)	Crotonobetainyl-CoA dehydrogenase caiA	15%
	hpt	A19 (syn)	Hypoxanthine phosphoribosyltransferase hpt	17%
	thpR	A111 (syn)	RNA 2',2C3'-cyclic phosphodiesterase thpR	11%
	prpB	T25 (syn)	Methylisocitrate lyase	13%
	acnA	T850M	Aconitate hydratase	11%
	rhtC	V8 (syn)	threonine export protein RhtC	16%
P8	*apaG*	A75 (syn)	Co2+/Mg2+ efflux protein ApaG	23%
	cysH	A280G	Phosphoadenosine phosphosulfate reductase CysH	99%
P9	*cadC*	G464R	Transcriptional regulator CadC	21%
P10	*hipB*	W26L	Antitoxin HipB	92%
	glnG	D61N	Nitrogen regulation protein NR(I)	97%
P12	*bsmA*	L28R	Lipoprotein BsmA	31%

Table 4. Polymorphic positions that were observed in second fecal *H*30 population genomes that were absent in same-host first fecal and urinary *H*30 population genomes. Genes with unknown function and polymorphisms with <10% frequency are not included. Synonymous polymorphisms are marked with '(syn)'.

Participant	Gene	Polymorphism	Gene product	Urinary population frequency
P1	*fruK*	A226 (syn)	1-phosphofructokinase	14%
	metH	Upstream	Cobalamin-dependent methionine synthase	61%
P2	*mipA*	S94N	MltA-interacting protein	18%
	fbaA	D215 (syn)	Ketose-bisphosphate aldolase	65%
P4	P423_RS09705	A103D	CDP-alcohol phosphatidyltransferase family protein	83%
P5	*prfS*	A17 (syn)	Peptide chain release factor 3	61%
P7	*barA*	R477H	Two-component sensor histidine kinase BarA	32%
P8	*cysZ*	P168L	Sulfate transporter CysZ	90%
	tas	D287A	NADP(H)-dependent aldo-keto reductase	19%
	fldB	I133V	Flavodoxin-2	95%
P10	*rpiA*	D74 (syn)	Ribose-5-phosphate isomerase	13%
	atpD	L389 (syn)	ATP synthase subunit beta	21%
P12	*cysZ*	P168L	Sulfate transporter CysZ	24%
P9	*caiA*	R352	Crotonobetainyl-CoA dehydrogenase caiA	97%
	pbp2	D312G	Penicillin-binding protein 2	91%
	rnfD	T187 (syn)	Electron transport complex subunit D	98%
	fliH	Q152*	Flagellar assembly protein FliH	92%
	psuG	P202S	Pseudouridine-5'-phosphate glycosidase psuG	94%
	barA	S106*	Two-component sensor histidine kinase BarA	98%
	uxaA	Upstream	Altronate dehydratase	93%
	garL	P81 (syn)	5-keto-4-deoxy-D-glucarate aldolase garL	96%
	mtr	T356S	Tryptophan permease	98%
	lptA	L144 (syn)	Lipopolysaccharide ABC transporter substrate-binding protein LptA	99%
	torZ	G406S	Molybdopterin guanine dinucleotide-containing S/N-oxide reductase	93%
	waaZ	L116S	3-deoxy-D-manno-oct-2-ulosonate III transferase WaaZ	98%
	frdD	M87I	Fumarate reductase subunit D	95%

Table 5. Polymorphic positions that were observed in urinary *H*30 population genomes while absent in either same-host fecal *H*30 population genomes. Genes with unknown function and polymorphisms with <10% frequency are not included. Synonymous polymorphisms are indicated with '(syn)'. Stop codons are indicated with an asterisk.

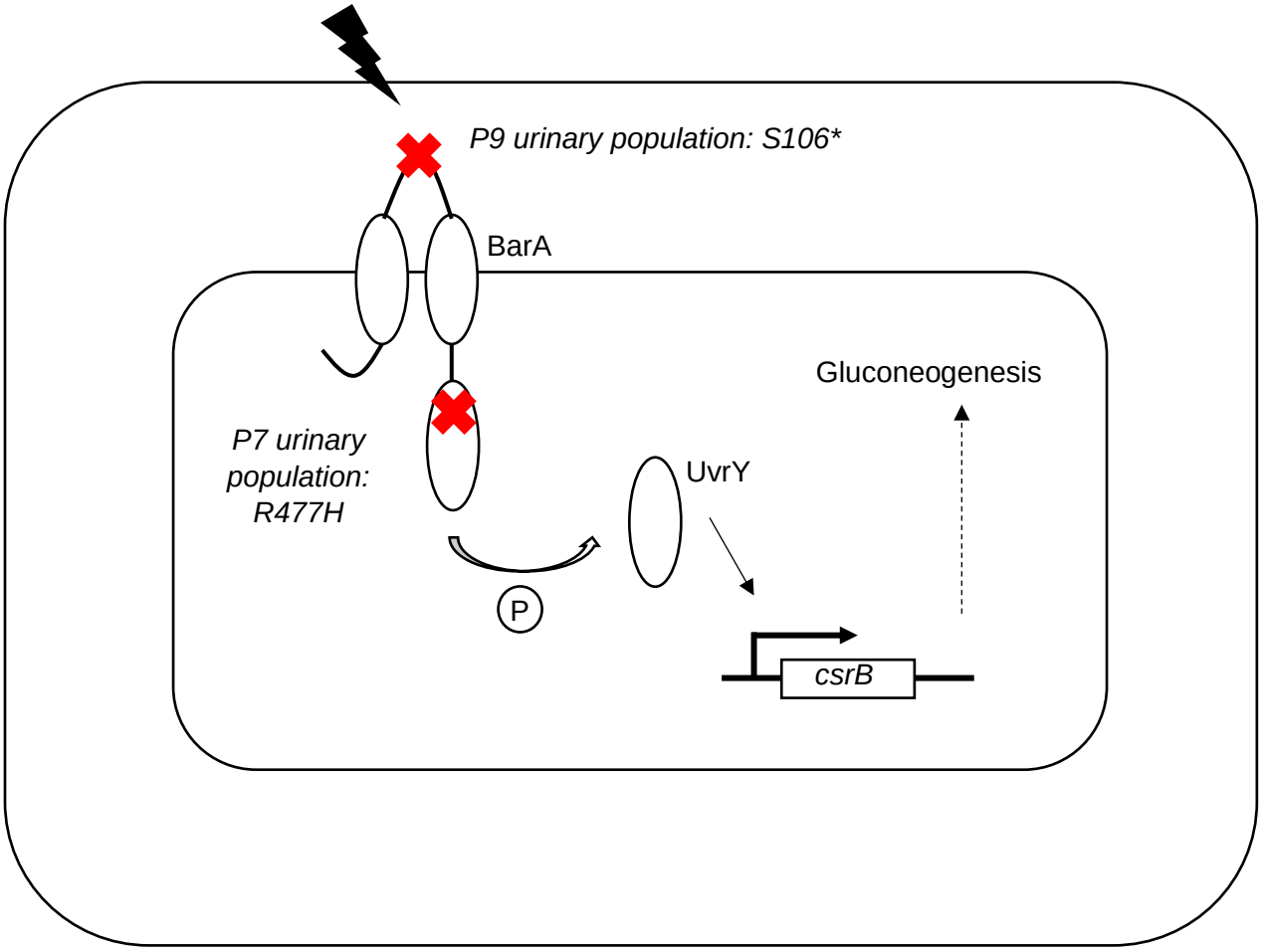


Figure 26. Activity of BarA, a global regulator targeted by SNPs in urinary *H*30 from two different participants. Description of SNPs, including amino acid change, participant, and sample are indicated in italics. Locations of SNPs on the protein are shown in red. The SNP in the participant P9 urinary population results in a premature stop codon and thus a truncated protein. The SNP in the participant P7 urinary population is located in the histidine kinase domain potentially resulting in an inability to phosphorylate UvrY. The premature stop codon was detected in both the P9 urinary isolate, and in the population. The P7 SNP was detected in the population genome only.

Discussion

This study is the first thus far to apply a population-centered approach to the study of *H*30 within-host heterogeneity. While studies comparing same-host urinary and fecal isolates of UPEC have determined that within-host changes may occur, concerns still lingered that the differences between isolates were spurious false positives or were not representative of the overall populations in these carriers. Our study, by creating population genomes for both fecal and urinary *H*30 is able to address these concerns and distinguish between low-frequency or isolate-specific differences between fecal and urinary UPEC, and true population-scale differences.

To this end, our analysis of urinary *H*30 has found that urinary populations undergo genomic changes in many processes, with potentially functional changes in carbohydrate metabolism genes occurring less often than expected. These findings are similar to previous research of adaptation of the asymptomatic bladder-colonizing *E. coli* strain 83972 to inoculated hosts, which detected genetic changes primarily in stress response, motility, and antibiotic resistance genes, but not in metabolic genes(82). Furthermore, these findings are in contrast to research comparing fecal and urinary UPEC isolates from symptomatic UTI, which primarily sustain nonsynonymous mutations in metabolic genes, carbohydrate metabolism in genes in particular(52). This suggests that *H*30 asymptomatically colonizing the bladder may adapt in a manner distinct from symptomatic UPEC. However, whether *H*30 adaption during symptomatic bladder infection is different from the adaptation described in this study requires further study.

Interestingly, similar to studies of both symptomatic UTI and asymptomatic bladder colonization, we found that the carbon storage regulator *barA* was targeted twice by nonsynonymous changes(52, 82). This is somewhat in conflict with other studies that show that gluconeogenesis is necessary for bladder colonization(83). Furthermore, *barA* – has been found to be necessary

for UPEC virulence in one study(84). This discrepancy may be due to several factors. Firstly, research that has found the necessity of gluconeogenesis and *barA* involved administering a high dosage of *E. coli* into catheterized mice, which may not be representative of colonization in humans in terms of the size of the pioneer population. Secondly, the genetic background of the strain(s) studied may significantly impact which characteristics are necessary for colonization, as metabolic abilities may differ between strains. It may be that the UPEC with sustain carbon storage gene mutations have a metabolic strategy distinct from other UPEC. Thus, our study, alongside others, suggests that the need for further study of UPEC metabolism may be necessary.

Unlike previous studies, we found that sulfur metabolism genes are affected in urinary populations. Interestingly, urine is not sulfur-limited as it can contain up to 20 mM sulfate, 1mM cysteine, and 0.8 mM taurine(85). It is therefore possible that sulfur/cysteine is abundant enough in urine that loss of some transport and/or metabolic genes does not sufficiently reduce fitness to remove these mutants from the urinary population. Overall, these findings require further study, both in an *H*30 background and in other UPEC clones.

Our analysis of fecal *H*30 samples found that changes that arise in these populations affect many processes, but that respiration genes were affected by potentially functional changes more often than expected while DNA metabolism genes were affected by such changes less often. This partially supports previous research of *E. coli* adaptation to the mouse gut, which was observed to include deactivation of *dcuB* and *focA*, genes important for anaerobic respiration system(32). Furthermore, previous studies of a commensal gut *E. coli* strain as it changed over a year reported changes in genes affecting many of the same processes we report occurring in fecal *H*30(86). This suggests that *H*30 within-host gut adaptation may be similar to that of commensal *E. coli*. Since these changes are deactivating, which may impair long-term colonization as it limits metabolic flexibility, further research is needed to determine

whether these changes may ultimately lead to *H*30 eventually being outcompeted by a different strain. Furthermore, study of *H*30 in carriers with short-term carriage is needed to determine whether any of the changes we report in this study can be associated with persistence in the gut.

Our study does have notable weaknesses. Our study included a small number of participants and only one urinary sample per participant. It is therefore unknown how adaptation occurs in the bladder over time and what the true range of within-host diversity in *H*30 may be. Furthermore, more sequenced isolates are needed to verify the accuracy of the population genome frequencies and to determine whether the potential substructure detected in both fecal and urinary *H*30 is truly present. While the number of isolates currently sequenced do suggest that the within-sample heterogeneity we detect is real, the true frequency of each minority position may deviate from the population genome prediction due to PCR and sequencing errors. However, if the linkage between low-frequency positions and higher ~50% frequency positions that are implied by our existing sequenced isolates is true, the existence of *H*30 subpopulations in the gut and bladder suggests that there are different niches that *H*30 occupies in both compartments of the body. Since the characteristics that would differentiate these niches are currently unknown, further study into *H*30 population structure, especially in the bladder, is needed to determine what causes this differentiation and how it impacts overall *H*30 microevolution.

CHAPTER 4: EXPERIMENTAL EVOLUTION OF FECAL *H*30 ISOLATES IN HUMAN URINE

Introduction

Uropathogenic *E. coli* (UPEC) are the main cause of urinary tract infections (UTI), which cost the US alone an estimated $ 1.6 billion every year(87). Interestingly, UPEC are also able to colonize the bladder asymptomatically, sometimes for long periods of time(88). Recently, the pandemic multidrug-resistant UPEC clone ST131-H30 (*H*30), the most common cause of drug-resistant UTI, was found to asymptomatically colonize the bladder more frequently that other UPEC clones(48). Such disparity in clinical outcome has given rise to the question of whether *H*30 (and other UPEC) can become more or less pathogenic via within-host adaptation to the bladder. Research into this question has shown that during UTI-causing UPEC may adapt via acquisition of nonsynonymous mutations, particularly in genes involved in stress response, metabolism, and fimbria(89). Our laboratory's study (see above) has shown that during asymptomatic colonization, *H*30 accumulates nonsynonymous mutations on a population level, in genes affecting sulfur metabolism, carbon storage, cell wall biogenesis, and others. These results were similar to previous studies of the asymptomatic bladder colonizer *E. coli* strain 83972 in inoculated hosts, in which carbon storage genes are also targeted by nonsynonymous mutation(82).

However, bacteria may adapt to many different factors, including the medium as an environment for growth, the physical aspects of the environment, and/or the host immune response. For instance, deactivation of the carbon storage regulator *barA* could be adaptive for growth in urine if the urine has glycolytic products to serve as a carbon source or if lysed bacteria are present(90). *barA* deactivation could also be useful in slowing down growth which may help evade immune response. In order to determine what mutations, if any, are adaptive to growth in urine in particular, an experimental evolution approach is necessary. Thus, we wanted to

determine which changes, if any, that we observed in natural urine samples from *H*30 carriers were adaptive to growth in urine specifically.

In order to address this question we decided to take single-colony first fecal isolates from participants P1 and P4 (described above) and passage them in sterilized human urine for 30 days, with samples saved every 3 days (Fig. 27). This experiment allowed us to determine that mutations in iron acquisition, amino acid biosynthesis, and amino sugar biosynthesis may be adaptive to growth in urine.

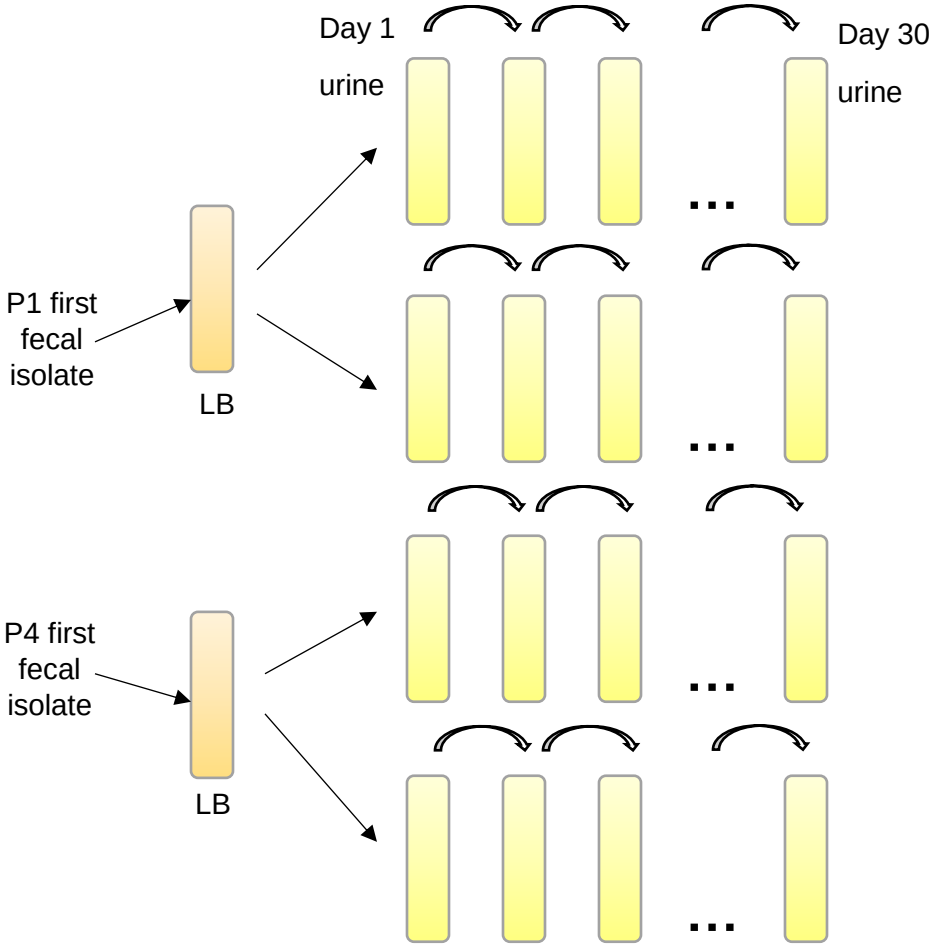


Figure 27. Experimental procedure for *in vitro* evolution of first fecal isolates in human urine. All passages (LB, urine) were done overnight with 10ul as the inoculum.

Results

To assess the preliminary differences between day 1 and day 30 H30, we performed whole-genome sequencing on single-colony isolates from each day 1 and day 30 sample. Comparison between day 1 and day 30 isolates showed that after 30 days, mutations were observed in every day 30 isolate.

A total of 37 mutations were observed. 16 of these mutations were observed in both passaging runs of the P4 isolate (Table 6). Interestingly, these included mutations in *fhuD,* and *ilvY*. *fhuD* and *ilvY* were genes in which the P6 second fecal population contained low-frequency polymorphic positions. Furthermore, the mutation in *nagC* which was observed in both P4 evolved isolates was also detected in all three P10 *H*30 populations at 50% (first fecal), 87% (urinary), and 100% (second fecal) frequencies.

Aside from the repeated mutations described above, sequencing of the day 1 and day 30 isolates produced 3, 3, 4, and 7 unique mutations in the P1 passage 1, P1 passage 2, P4 passage 1, and P4 passage 2 day 30 isolates respectively (Table 7). These include genes involved in sulfur metabolism (*cysP*, *cysW*), motility (*pdeH*, *qseC*), and cell wall biosynthesis (*lptD*) which were affected by within-sample polymorphism and between-sample changes in the natural fecal and *H*30 populations described above. Furthermore, as with the *nagC* mutation above, the mutation in *xanQ* which was observed in P4 passage 1 was also observed in all three P10 *H*30 populations at 43% (first fecal), 92% (urine), and 100% (second fecal) frequencies.

Gene	Mutation	Gene product	P1 passage 2 day 30 population frequency	P4 passage 1 day 30 population frequency	P4 passage 2 day 30 population frequency
fhuD	L144F	Iron-hydroxamate transporter substrate-binding subunit	98%	97%	100%
degQ	Q84L	Serine endoprotease DegQ	99%	96%	100%
nagC	A341S	N-acetylglucosamine repressor	97%	100%	100%
ybjL	Upstream	Transporter	99%	98%	98%
dapA	N156I	4-hydroxy-tetrahydrodipicolinate synthase	100%	99%	100%
yqeB	I247L	Hypothetical protein	99%	98%	100%
P423_RS17910	L398Q	Fimbrial biogenesis outer membrane usher protein	98%	95%	100%
P423_RS17930	I27N	Hypothetical protein	99%	100%	100%
gspA	W56*	General secretion pathway protein GspA	100%	100%	100%
slp	G116 (syn)	Outer membrane protein slp	96%	96%	100%
hmuV	L198I	Hemin import ATP-binding protein HmuV	99%	99%	100%
waaY	T45I	Lipopolysaccharide core heptose(II) kinase RfaY	100%	100%	100%
gmk	E61D	Guanylate kinase	98%	98%	100%
ilvY	E248K	HTH-type transcriptional activator IlvY	98%	100%	100%
pldB	E243D	Lysophospholipase	99%	99%	100%
yiiM	Upstream	6-N-hydroxylaminopurine resistance protein	99%	N/A	N/A

Table 6. Mutations first observed in both P4 passage 1 and P4 passage 2 day 30 single-colony isolates as compared to corresponding day 1 isolates. Frequencies in P4 passage 1, P4 passage 2, and P1 passage 2 day 30 population genomes (100X genome coverage) is indicated where mutations were present. Synonymous mutations are indicated with '(syn)'. Stop codons are indicated with an asterisk.

Evolved Isolate	Gene	Mutation	Gene product	Frequency in day 30 population
P1 passage 1	*lptD*	Y63S	LPS-assembly protein LptD	N/A
	ydcS	Upstream	spermidine/putrescine ABC transporter substrate-binding protein	N/A
	yphD	A204S	ABC transporter permease	N/A
P1 passage 2	P423_RS14855	Upstream	hypothetical protein	87%
	nrdI	S41P	class Ib ribonucleoside-diphosphate reductase assembly flavoprotein NrdI	99%
	P423_RS17495	A137	NAD dependent epimerase/dehydratase	100%
P4 passage 1	P423_RS02925	L166 (syn)	sugar-binding transcriptional regulator	96%
	xanQ	F94I	xanthine permease XanQ	95%
	yiiR	L41I	DUF805 domain-containing protein	97%
	yjbB	L455I	Na/Pi cotransporter family protein	97%
P4 passage 2	*cysP*	Upstream	sulfate ABC transporter substrate-binding protein	N/A
	tqsA	Upstream	AI-2 transporter TqsA	98%
	P423_RS09205	Y212 (syn)	Bcr/CflA family multidrug efflux MFS transporter	99%
	uvrC	Q232K	excinuclease ABC subunit C	100%
	P423_RS13965	N211Y	sulfate ABC transporter permease	100%
	pdeH	P64S	cyclic-guanylate-specific phosphodiesterase	98%
	sgaT	Upstream	transporter SgaT	N/A

Table 7. Unique mutations observed day 30 isolates from P1 and P4 passages as compared to corresponding day 1 isolates. Frequency in corresponding day 30 population genomes (100X genome coverage) is indicated where the mutation was present. Synonymous mutations are indicated with '(syn)'.

In order to determine if the mutations detected in day 30 isolates are representative of the overall evolved populations, we applied the same population-level sequencing approach as utilized with participant fecal and urinary *H*30 populations. Sequencing population genomes of all four day 1 and day 30 samples to ≥100X coverage showed that 12 of the 17 unique mutations detected in P1 and P4 day 30 isolates were dominant (>95%) in their respective day 30 population genomes. The remaining 5 unique mutations were not detected at any frequency in respective population genomes. This included all unique mutations detected in the day 30 P1 passage 1 isolate and 2 mutations (*cysP*, *sgaT*) detected in the day 30 P4 passage 2 isolate.

Out of the 16 repeat mutations detected in day 30 isolates from both P4 passages, 15 were observed in both P4 passage 1 and 2 population genomes at >95% frequency (Yable 6). The remaining repeat mutation was not observed in either population genome at any frequency. Additionally, all 16 repeat mutations were also observed in the P1 passage 2 day 30 population genomes at >95% frequency.

Further analysis showed that polymorphic positions were observed in all population genomes, with all day 30 population genomes having >500 low-frequency (<10%) polymorphisms each, when compared to day 1 population genomes. A total of 8 >10% frequency polymorphisms were observed in day 30 population genomes. These included P1 passage 1 polymorphisms in sugar transporter, *acrR*, glycosyl transferase, and allantoin permease genes, P4 passage 1 polymorphisms in hypothetical protein and L-serine ammonia-lyase genes, and a P1 passage 2 polymorphism in a hypothetical protein gene. All polymorphisms had frequencies of <20% in the corresponding day 30 population genome, indicating the kind of substructure seen in participant population genomes is potentially absent. Aside from the mutations described above, no additional positions were found to be both absent in day 1 population genomes and dominant in day 30 population genomes.

Discussion

Our study is the first to utilize experimental evolution in urine to understand adaptation of UPEC over time. Using this approach, we were able to determine that up to 15 mutations arose and reached fixation in 3 of our 4 day 30 *H*30 populations. These included nonsynonymous mutations in iron acquisition (*fhuD*, *hmuV*), proteolysis (*degQ*), protein secretion (*gspA*), amino sugar and amino acid biosynthesis (*nagC*, *ilvY*, *dapA*), DNA metabolism (*gmk*), and fimbria biogenesis (P423_RS17910). This can be explained by our experimental design: since we grew up fecal isolates in LB overnight, thus creating up to 10^6 mutants in this initial culture, the initial inoculum may have contained up to 10^4 mutants. Due to the selective pressure of urine as a significantly different growth environment than LB, the emergence of these mutations suggests that they may be truly adaptive. Their presence in one of the P1 day 30 populations further reinforces this, as the P1 and P4 fecal isolates are differentiated by 47 genomic SNPs and by plasmid content. Importantly, we detected similar mutations in natural *H*30 populations, arising both in the processes these genes are involved in, and in some of the genes themselves. *ilvY*, *hmuV*, and *fhuD*, for example, also sustained mutations in the P6 second fecal population. Furthermore, the *nagC* mutation occurred first at a 50% frequency in the P10 first fecal population, then reaching dominance in the P10 urinary population. Thus, our data suggest that some adaptive mutations may in fact arise in the gut prior to bladder colonization. Further study into fecal and urinary *H*30, including *H*30 from UTI, is needed to determine the prevalence of this phenomenon and whether it applies to asymptomatic carriage only or to all *H*30 bladder colonization.

We also observed mutations unique to a single passage. These included mutations in genes involved in sugar transport (*xanQ*), sulfur metabolism (P423_RS13965), and motility (*pdeH*). These were also processes targeted by changes in urinary *H*30 populations, including *xanQ*, which sustained a mutation in P10 *H*30 populations in a manner similar to *nagC*. This suggests

that these changes may potentially improve growth but may not be necessary for adaptation. However, further research is needed to determine whether such mutations truly improve growth or if these are neutral changes.

Our study has a number of important caveats. Firstly, our experiment involved *H*30 adaptation to a singular batch of human urine. Since urine composition can vary, some of the mutations we report may be adaptive to the particular urine we utilized. For example, mutations in sulfur metabolism genes may not occur in urine that is relatively poor in cysteine; mutations in sugar transport genes may occur in urine that has a higher sugar content. Experimental evolution in concentrated and dilute urine may be able to clarify this matter. Importantly, this may explain why, unlike natural *H*30 urinary populations, evolved *H*30 populations did not sustain mutations in carbon storage regulation genes, as such mutations may be adaptive to a particular urine composition. An analysis of metabolites before and after growth of *H*30 in urine would be similarly useful in determining which components are utilized by *H*30 for growth. Secondly, one passage in our study did not produce any mutations that reached fixation in the day 30 population. It may be that mutations arose between day 1 and day 30, but that these were ultimately detrimental to long-term survival, and the subpopulations containing those mutations were eventually outcompeted. Study of the time points between day 1 and 30 would be needed to determine if this is the case. Indeed, our experimental protocol involved saving the sample not used for the inoculum every 3 days, making such a study possible. If such mutations did arise in this passage, comparison of these to mutations from the other 3 passages would determine whether any of the mutations we report are beneficial in the short-term only. Overall, this study shows that *H*30 adaptation to the bladder is complex and extensive study is needed to disentangle its drivers and mechanisms.

CHAPTER 5: MATERIALS AND METHODS

Study design and sample processing

We selected a subset of participants from a previous study carried out by Kaiser Permanente Washington and University of Washington (Seattle, WA)(48). That study identified healthy gut carriers of ciprofloxacin-resistant *E. coli*, including *E. coli H*30. These *E. coli* were found in initial fecal samples by soaking the swab in water and plating 0.5 mL of the suspension on LB-ciprofloxacin. CH typing of 1 to 8 single colonies was performed on any resulting *E. coli* colonies to identify the resistant clone. After the initial fecal sample was analyzed, *H*30 carriers as well as carriers of some other strains were asked to provide urine samples. These were received on average 152 ± 55.9 days after the initial sample (85% responded). The respondents were then asked to provide follow-up fecal samples, which were received on average 82 ± 41.1 days after the urine sample (84% responded). To identify any ciprofloxacin-resistant clones, 0.5mL of each urine sample was plated on LB-ciprofloxacin and any resulting colonies typed as with initial samples. Follow-up fecal sample swabs were analyzed as described above with initial samples to similarly identify ciprofloxacin-resistant clones. For this study, we chose 28 individuals who supplied all three samples. In 11 participants, *H*30 was identified in all three samples; in 4 additional participants *H*30 was isolated in two samples. In 8 participants ciprofloxacin-resistant ST1193 was found in at least two samples. In 5 participants the same ciprofloxacin-susceptible clone was found in at least two samples. The sample types, strains clonal identity, and sampling times for all participants are shown in Figure 28. Average age of participants was 66.7 ± 15.7 years.

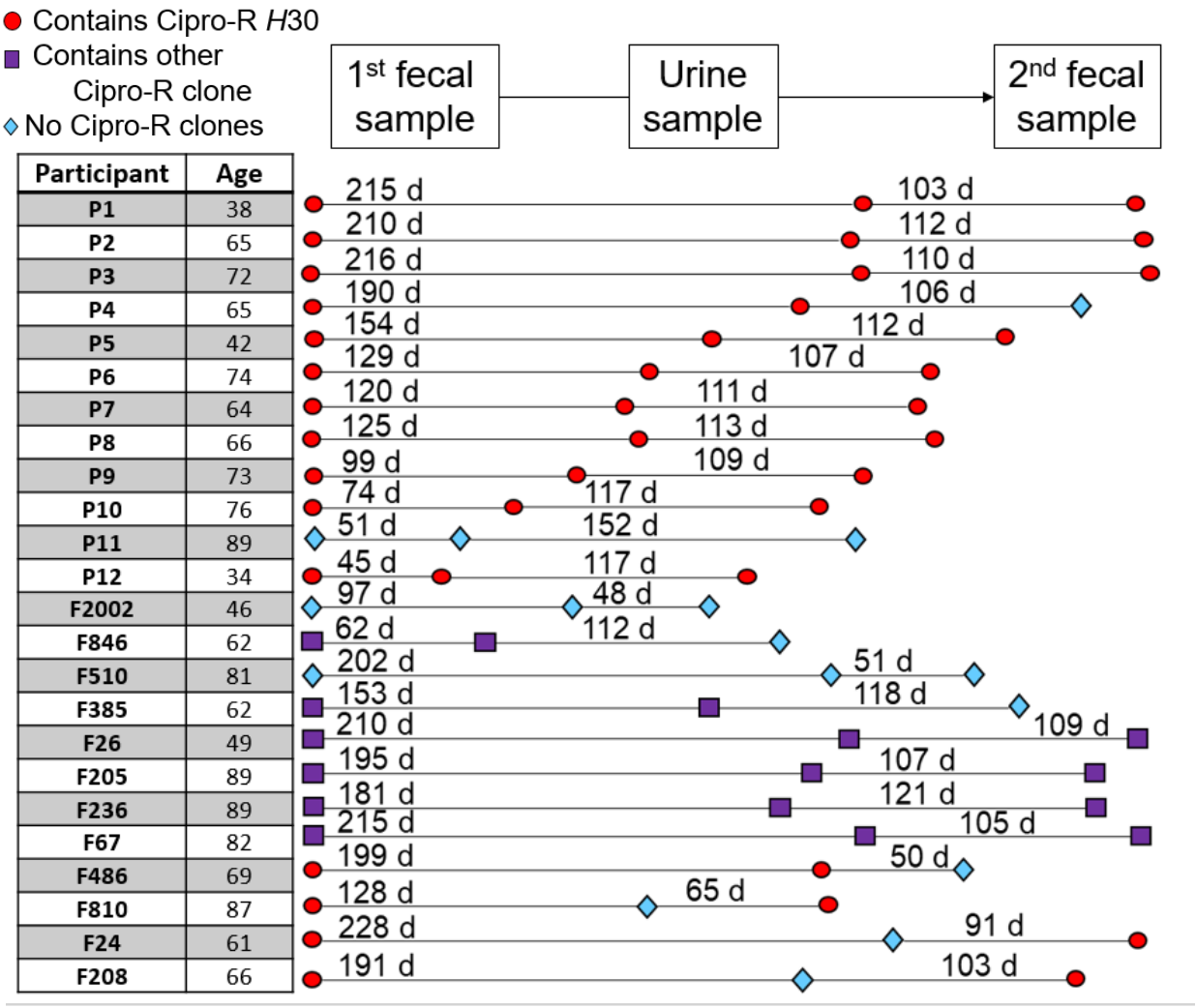


Figure 28. Sampling of volunteer sample sets. Length of segments is proportional to number of days between samples. Number of days is indicated above each segment. 'Cipro-R' refers to ciprofloxacin-resistance.

Deep amplicon sequencing for characterization of *E. coli* diversity

Preparation of predefined control samples

For control experiments, two predefined strains were chosen - *H*30 (*E. coli* FESS614.ds6) and clonal group ST101 (*E. coli* FESS614.ds4). DNA from these strains was extracted and *fumC* and *fimH* was amplified by PCR using the following conditions: 3min denaturation (95°C), 35 cycles of annealing (95°C for 45sec, 57°C for 45sec, 72°C for 45sec), 5min extension (72°C),

4°C hold. The primers (10 uM) used were as follows: 5'-TCACAGGTCGCCAGCGCTTC-3' (*fumC* forward), 5'-GTACGCAGCGAAAAAGATTC3' (*fumC* reverse), 5'-TCAGGGAACCATTCAGGCA-3' (*fimH* forward), 5-ACAAAGGGCTAACGTGCAG-3' (*fimH* reverse). Amount of PCR product was measured by Qbit. To create *H*30-only and ST101-only samples, the corresponding *fumC* and *fimH* PCR products were pooled together at a 1:1 ratio. To create mixes, *H*30 and ST101 amplicons of *fumC* were mixed together in ST101:*H*30 ratios of 1:1, 1:4, 1:10, 1:100, and 1:1000. The same was performed with *fimH* amplicons. The *fumC* and *fimH* mixes were then pooled together by ratio type to create mixes that had equal concentrations of total *fumC* and *fimH*. The DNA mixes were prepared for sequencing using Nextera XT DNA library prep kit using standard protocol. The resulting library was sequenced on the Illumina MiSeq (v3 kit). All mixes, except 1:10, reached coverage of ≥9,000X and were analyzed.

Deep sequencing and allele analysis of fecal and urine samples

Each fecal and urine sample was plated on MacConkey agar to reach ~1,000 *E. coli* single colonies per plate. All colonies were swabbed from the agar and DNA extracted using the Qiagen Blood & Tissue Kit. From this pooled DNA *fumC* and *fimH* genes were amplified by PCR by using the same primers and conditions as described above for control samples. Amplicons were then purified and pooled by sample using the Qiagen Gel Extraction kit, then prepared for sequencing using Nextera XT DNA library prep kit using standard protocol except for usage of 52.5ul of resuspension buffer in the final magnetic bead cleanup step. The resulting library was sequenced on the Illumina MiSeq (v3 kit). Sequencing data was analyzed using a Python program of our construction, Population-Level Allele Profiler (PLAP), and has been made available for public use on GitHub: github.com/marade/PLAP. The process is described below (see also Fig. 29).

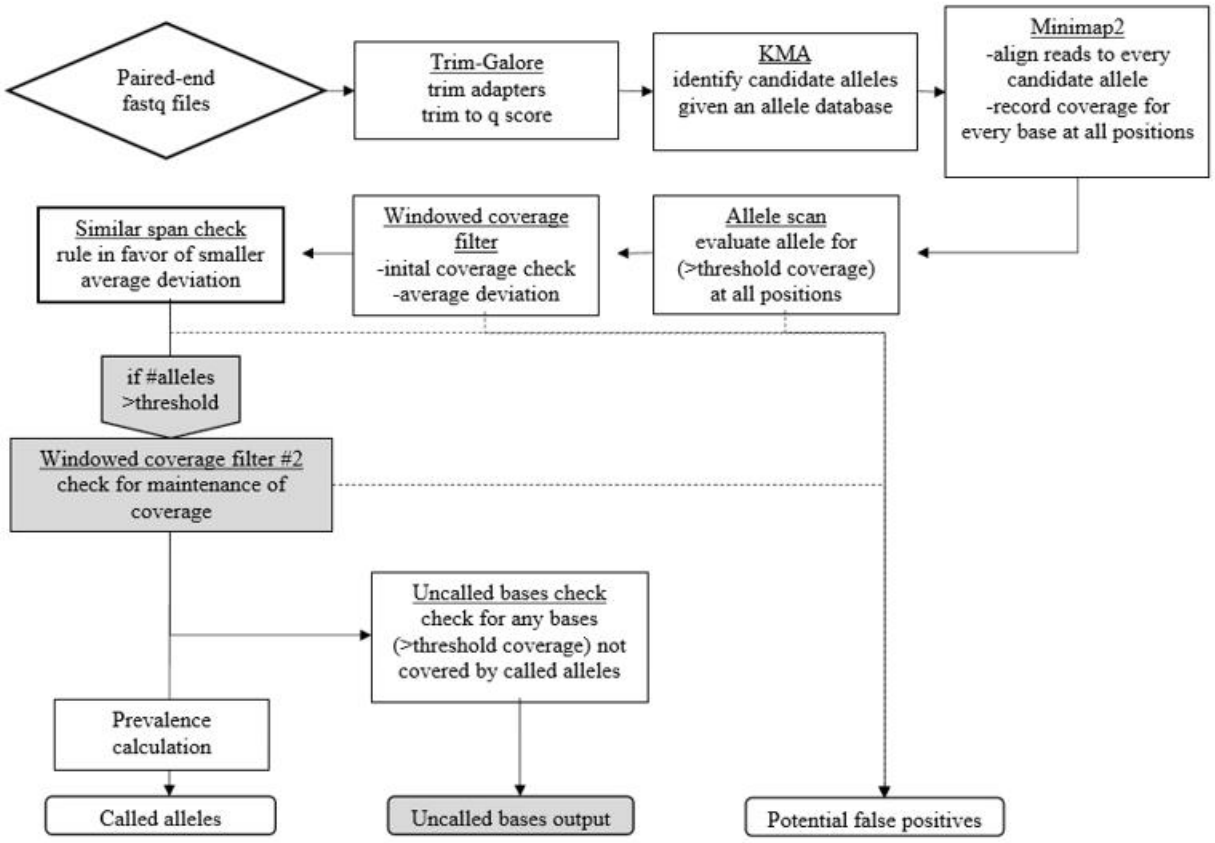


Figure 29. PLAP algorithm workflow. Algorithms previously developed by other groups include Trim-Galore, KMA, Minimap2. Not pictured but used during windowed coverage checks is SAMtools.

For each sample, adapter sequences were removed using Trim-Galore, and resulting trimmed reads were aligned to a list of all known *fumC* and *fimH* alleles using KMA with strict 99.99% identity matching(91, 92). For each KMA-detected allele per sample, trimmed reads were again aligned to the sequence using Minimap2 and SAMtools(93, 94). Any candidate allele which had at least 1 base supported by <0.8% of reads was removed from consideration. False positives were filtered using a moving 10bp window for each allele as follows. Reads of ≥100bp with 100% identity within the window were counted. Alleles with low initial coverage, unstable coverage (high average deviation from the mean), and high similarity in coverage pattern to an allele with more stable coverage were removed from consideration. If >3 alleles were left for

consideration for a gene, 10bp moving window analysis was repeated with ≥200bp reads. If for any interval in this second analysis, >60% of coverage was lost compared to the first moving window coverage, the allele was discarded. Heterogeneity at any positions that remained undescribed by surviving alleles was recorded. Relative abundance of all alleles was determined using the minimum coverage found during first moving window analysis. In samples found by PLAP to be ≥50% made up of <100bp reads (overtagmented samples), allele prevalence was calculated manually by ascertaining base(s) unique to each allele and using the coverage of these base(s) to calculate prevalence.

Out of the 28 total sets of fecal and urine samples chosen for this study, at least one sample failed PCR amplification or sequencing library prep in 4 sets and therefore all samples from these sets were dropped. From the remaining 24 sets we were able to sequence *fumC* and *fimH* in all three samples. Out of those, 67 (89%) samples – 22 first fecal, 24 urine, and 21 second fecal – reached ≥9,000X coverage per gene and were included in the analysis.

Determining within-sample clonal group breakdown

Identity of strains present in a sample was determined by combining *fumC* and *fimH* allele numbers and determining the ST type using Enterobase. In uniclonal and unambiguous samples, every allele had one match supported by the equation for maximum acceptable difference between same-strain *fumC* and *fimH*. Therefore, these alleles formed a CH type based on which ST type was determined.

For ambiguous-simple samples, the most prevalent *fumC* and *fimH* alleles formed an equation-supported CH type. Any alleles that also had a single equation-supported match were assigned to form a CH type. For all other alleles, Enterobase was consulted to determine which allele combinations have been observed. If the CH type(s) produced was between alleles that had different prevalences according to the equation, the “remaining” prevalence was calculated for

the allele with the greater prevalence. This allele was then paired with allele(s) for which an Enterobase-logged CH type was not available and/or any novel alleles until the “remaining” prevalence was consumed. If there were any allele(s) that remained after this step, they were paired with the major allele of the opposite gene. For ambiguous-complex samples, the most prevalent *fumC* and most prevalent *fimH* allele were assigned to the same CH type. The “remaining” prevalence was calculated for the allele with the greater prevalence and treated as an unmatched allele. From this step, we proceeded as with ambiguous-simple samples.

Determining prevalence of clonal groups by culturing

Prevalence of ciprofloxacin-resistant clones in each sample was determined by diluting ~1ul of sample with ≥300ul of H2O, plating 40ul of this dilution on MacConkey agar, picking >130 single *E. coli* colonies, patching on Hardy-Chrom UTI agar to verify *E. coli* identity, then patching colonies on LB-ciprofloxacin. Prevalence of other clonal groups was validated by plating on MacConkey agar and subsequent patching of single colonies onto Hardy-Chrom UTI agar to distinguish *E. coli*. *fumC* and *fimH* alleles of these colonies were then determined by 7SNP clonotyping and Sanger sequencing(95).

Statistical and phylogenetic analysis

To determine the 99% confidence interval (CI) for the prevalence of ciprofloxacin-resistant strains, the number of resistant colonies was treated as number of successes and the total number of picked colonies was treated as the total. To determine the 99% CI for the prevalence of ciprofloxacin-sensitive strains, the number of colonies of that strain was treated as number of successes and the total number of picked colonies was treated as the total. Confidence intervals were calculated using Stata(96). All t-tests were run using GraphPad (http://www.graphpad.com/quickcalcs/ConfInterval1.cfm).

Phylogenetic trees were constructed using MEGA7(97). Erroneous base coverage graph was generated using seaborn(98). *Escherichia coli fumC* alleles were downloaded from Enterobase MLST allele data. *Escherichia coli fimH* alleles used are publicly available(99). *Escherichia fergusonii* and *albertii fumC* alleles were downloaded from NCBI. *Klebsiella pneumonia* and *Enterobacter aerogenes* alleles of *fimH* were downloaded from the PATRIC database (www.patricbrc.org).

Analysis of diversity within fecal and urinary *H*30 populations

Sequencing of single-colony isolates and H30 population genomes

Single-colony isolates of *H*30 were taken from each sample by plating on LB with ciprofloxacin and verifying *H*30 identity using 7-SNP CH qPCR. DNA from these was then extracted using the Qiagen DNeasy Blood & Tissue kit and then prepared for sequencing using the Illumina Nextera XT kit using standard procedure. Genomes were sequenced using Illumina Miseq with a v3 kit. Average coverage across each sequenced genome was ≥20X.

*H*30 population genomes were generated by plating each sample on LB with ciprofloxacin for at least 200 single colonies. All colonies were then swabbed off the plate as a pool, and DNA extracted, prepared, and sequenced as described above for single colonies. Average coverage across population genomes reached ≥100X for each sample.

Phylogeny construction and analysis of sequencing data

Isolate and population genome phylogenies were constructed using kSNP3.0, with the JJ1186 and NA114 *H*30 genomes as outgroups(100). All sequencing data was analyzed with JJ1886 as a reference genome using BWA-MEM and VarScan to determine presence of polymorphisms as well as loss of plasmids and mobile regions (101, 102). Identity of genes containing polymorphisms was verified using BLAST against *E. coli* K-12, and then assigned to categories

of cellular processes using GO terms listed in the BioCyc Database Collection (https://biocyc.org/). The presence of gene gain was determined by assembling isolate and population genomes using the SPAdes assembler and using blastn to compare same-host genomes to each other(103, 104). Kmer sizes used for SPAdes assembly were 21, 33, 55, and 77. Monte Carlo simulation was run using Microsoft Excel in the following manner: random genomes were chosen from the pool of sequenced P9 first fecal or urinary *H*30 genomes in numbers ranging from 1 (a single genome) to 12 (all genomes), then the number of resulting subgroups or polymorphic positions was calculated. This random choice was done 100 times per number of genomes selected (ex. 100 rounds of choosing 2 random genomes, 100 rounds of choosing 3 random genomes, etc). Then average number of subgroups or average number of polymorphic positions was calculated for each number of genomes selected. These averages were graphed versus the corresponding number of genomes selected and a logarithmic trend line was constructed based. This trend line was then allowed to project forward to 200 total "number of genomes" i.e. 200 total single colonies.

Experimental evolution of fecal H30 isolates

Single-colony isolates of *H*30 from P1 and P4 first fecal samples were taken, inoculated into LB medium, and left to grow overnight at 37°C. 10ul of this culture were used to inoculate 2ml of sterile human urine, which was then left to grow at 37°C overnight. Each subsequent passage involved inoculating 2ml of sterile human urine in a fresh tube with 10ul of the previous culture. This continued for 30 days. Two independent 30-days passages per first fecal isolate were performed, resulting in 4 separate evolution experiments. Glycerol stocks of every passage were created every 3 days starting with day 1 and included a stock of day 30 culture. Single-colony isolates and population genomes were prepared and sequenced from all day 1 and day 30 samples the same way as natural (participant) fecal/urine samples above.

CHAPTER 6: FUTURE DIRECTIONS

Dominance of *H*30 in the gut and bladder

As determined by our first study, *H*30 may be highly competitive in the bladder and may outcompete some gut strains as well. While the dominance of *H*30 in the bladder suggests that it is an effective colonizer, comparison to other UPEC clones is needed to confirm this. Notably, in our study, some non-*H*30 strains also rose to dominance in the bladder, including ST69, ST297, and ST1193. The latter is an emerging multidrug-resistant UPEC clone, that like *H*30 is able to both asymptomatically colonize the bladder and cause serious UTI, including fluoroquinolone-resistant infections, and appears to be associated with older host age(48, 105, 106). Unlike *H*30, however, ST1193 comes from an ST14-complex background rather than an ST131 background, meaning it is potentially different from *H*30 genetically and/or metabolically(106). This may make ST1193 an interesting comparison to *H*30 in terms of dominance in the bladder. If ST1193 is similarly dominant, similarities between *H*30 and ST1193 can be drawn to determine what characteristics may lead to their success. If ST1193 is not dominant, however, then their similarities (antibiotic resistance, UPEC virulence factors) may not be causative to bladder dominance. Indeed, our study finds that out of 4 ST1193-containing urine samples, only one was found to contain ST1193 only, suggesting that the latter case may be closer to the truth. However, the number of participants in our study, both *H*30- and ST1193-colonized, is too few to draw conclusions. A larger cohort is needed to fully address this question. Additionally, further studies should include samples from patients with symptomatic UTI to determine whether *H*30 (and ST1193) bladder dominance occurs in cystitis.

We also noted *H*30 dominance in the gut in some carriers. Since, according to our previous studies, *H*30 persists in the gut longer than other UPEC clones, it may be that *H*30 is highly competitive in the gut of some people. A larger cohort study may be able to determine if *H*30 is

more likely to outcompete some *E. coli* clones but not others. However, host and microbiome factors may also play a significant role. Since the interactions between *E. coli* and other species in the microbiome are currently unknown, it is possible that other species contribute to the competitiveness of *E. coli* strains in some hosts. For example, the anaerobic contingent of the microbiome in some hosts is made up of species/strains that provide a di-saccharide that *H*30 prefers to utilize, while in other hosts, a different carbohydrate is produced, lowering *H*30 competitiveness. Host nutrition may play a role here as well. Overall, *H*30 competitiveness requires further study, potentially alongside broader surveys of the microbiome and intestinal metabolome.

Perspectives in *H*30 within-host adaptation

Our second study has found that *H*30 undergoes potential adaptation both in the gut and in the bladder. Overall, the changes were primarily reductive: nonsynonymous mutations were common, and gene loss was observed in some cases as well. Such reductions may be detrimental to long-term carriage of *H*30, since gene inactivation may impair metabolic flexibility. Thus, a cohort with longer-term sampling may be needed to determine if such inactivation eventually leads to loss of the strain (in our case, *H*30). However, some changes observed in fecal *H*30 were similar to changes observed in colonizing commensal gut *E. coli*, suggesting that at least some changes are truly adaptive. It may be that *H*30 changes similarly to commensal gut *E. coli* within-host, suggesting that its activity in the intestine is commensal-like. A study of *H*30 together with other *E. coli* strains from the same host would be useful in determining if the changes fecal *H*30 undergoes are similar to commensal adaptation to the host.

We also detected changes in urinary *H*30, with nonsynonymous mutations more common in urinary than in fecal *H*30. The nature of the changes appeared to be similar to those observed in

the ABU strain 83972, which raises the question of ABU-specific and UTI-specific adaptation. Indeed, the changes previously observed in UPEC isolated from UTI patients were different from changes reported in 83972, but since these studies involved strains from different phylogroups, making direct comparisons is difficult. Since we describe asymptomatic *H*30 carriage specifically, a study of *H*30 UTI which may allow us to compare and contrast ABU and UTI adaptation (if such exist), could be relatively easily conducted. Additionally, studies with multiple urine samples would be useful in determining which changes, if any, accumulate or diminish over time in urine. It may be that some changes that we observe are adaptive in the short-term but are outcompeted in the relative long-term in the bladder. The mutations we observed in *barA* are of particular interest, as we are the third study to observed such mutations arising in UPEC (both UTI-causing and asymptomatic colonizers). While our experimental evolution study suggests that *barA* mutations may not occur as adaptation to growth in urine, further study is needed to confirm this. It may also be that *barA* mutations slow down growth in the bladder, which assists *H*30 with immune evasion. Whether the resulting inability to store carbon impairs long-term bladder colonization, whether this kind of mutation occurs only in a certain host background, or whether these mutations are simply not selected against in the bladder requires further study. On the other hand, it may be that adhesion, an important component of bladder colonization that was absent in our experimental evolution, modulates the kinds of mutations that are beneficial. Along with longer-term studies with multiple urine samples, further studies would ideally involve metabolomic assessment of urine samples to determine the environment UPEC are replicating in, as well as simulation of urothelial adhesion ans testing of hosts for known SNPs that would predispose them to ABU. While the latter may be difficult to organize, even metabolomics of host urine alone would be useful, as the composition of urine can vary widely with host diet and water intake. Overall, the question of *H*30 bladder microevolution raises interesting questions about the importance of host factors, bacterial metabolism, and *E. coli* adaptation.

ACKNOWLEDGEMENTS

We thank the personnel of KPWARI for assistance in collection of samples, and Dr. Sifang Chen for proofreading of the manuscript.

This work was supported by the National Institutes of Health (grant numbers R01AI106007 and R42 AI116114-02 [to Evgeni Sokurenko])

Evgeni Sokurenko conceived the project and designed the experiments. Dagmara Kisiela performed predefine control sample sequencing and analysis. All other sequencing, validation, and analysis was performed by myself. Veronika Tchesnokova provided study data and samples. Matthew Radey programmed the algorithm; Matthew Radey and I tested and calibrated it.

I'd like to thank my advisor Evgeni Sokurenko, my committee members John Mittler, Dave Fredericks, Luke Hoffman, and Elhanan Borenstein, as well as Kevin Hybiske for agreeing to be my GSR and serving on my reading committee on such short notice. Your comments throughout my project were insightful and helped my development as a scientist. I'd like to thank Matthew Radey, who as invaluable for his help with PLAP, and Veronika and Dagmara for always being there to help me iron out my work. I'd like to thank my friends Brittany Ruhland and Renae Cruz, without whose support this work would have been much more difficult. I'd like to thank my partner Sifang Chen for taking care of me while I wrote this thesis and for emotionally supporting me through the last couple years of graduate school.

Finally, I'd like to thank my family, especially my mother, for fostering my interest in science, providing encouragement, and supporting me throughout this whole process. None of this would be possible without you. You are all PhDs with me, in a small way.

SUPPLEMENTAL DATA

Supplemental Table 1

List of all predicted CH types across all fecal and urine samples. Where no phylogroup is indicate, the CH type is novel. Where the indicated phylogroup is 'Unk', the CH type has been previously described but not placed in a phylogroup according to Enterobase. Italicized CH types and alleles are criterion CH types/alleles. Bolded CH types are those that have been confirmed by single-colony isolation and sequencing.

Participant	Sample	*fumC* allele	*fumC* allele prevalence (%)	*fimH* allele	*fimH* allele prevalence (%)	Predicted CH type (ST type)	Phylogroup
P9	First fecal	***40***	**100%**	***30***	**100%**	***40-30 (H30) 100% CIP-R***	B2
	Urine	***40***	**100%**	***30***	**100%**	***40-30 (H30) 100% CIP-R***	B2
	Second fecal	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
P12	First fecal	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
	Second fecal	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
F208	First fecal	*40* 95	98.6% 1.4	*30* 31	98.8% 1.2%	*40-30 (H30) 98.6% CIP-R* 95-31 (ST2014) 1.2%	B2 A
	Urine	**24** **6**	96.80% 3.20%	**2** **38**	95.40% 4.60%	**24-2 (ST104) 95.4%** **6-38 (ST2178) 3.2%**	B2 B1
	Second fecal	38 *40* 70	98.20% 1.80% 1.30%	27 *30*	92.30% 3.30%	38-27 (ST95) 92.3% *40-30 (H30) 1.8% CIP-S* 70-30 (Novel) 1.3%	B2 B2
F810	First fecal	*40* 4	96.2% 11.5%	*30* 65	77.5% 22.5%	*40-30 (H30) 77.5% CIP-S* 4-65 (ST1589) 11.5% 40-65 (Novel) 11%	B2 A
	Urine	24	100%	1 149	97.60% 2.40%	24-1 (ST80) 97.6% 24-149 (Novel) 2.4%	B2

	Second fecal	***40*** **58** **24**	94.70% 2.90% 4.80%	***30*** 65 **1**	90.30% 2.50% 7.20%	***40-30 (H30) 90.3% CIP-R,S*** **24-1 (ST80) 4.8%** **58-30 (ST182) 1.3%** 40-65 (Novel) 2.5%	B2 B2 E
P5	First fecal	*40* 7	52.10% 47.90%	*30* 54 32	43.60% 49.00% 7.40%	*40-30 (H30) 44.7% CIP-R* 7-54 (ST540) 47.9% 40-32 (Novel) 7.4%	B2 B1
	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
	Second fecal	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
P1	First fecal	**26** ***40***	56.50% 43.50%	**24** ***30*** 143	51.30% 47.70% 1.00%	**26-24 (ST38) 55.5%** ***40-30 (H30) 43.5% CIP-R*** 26-143 (Novel) 1%	D B2
	Urine	***40***	100%	***30***	100%	***40-30 (H30) 100% CIP-R***	B2
	Second fecal	***40*** **11**	50.30% 49.70%	***30*** **54**	46.20% 53.80%	***40-30 (H30) 46.2% CIP-R*** **11-54 (ST10) 49.7%**	B2 A
P4	First fecal	*40* 4 107 41	43.60% 29.50% 19.30% 7.50%	*30* 27	84.00% 27.00%	*40-30 (H30) 43.6% CIP-R* 4-27 (ST648) 27% 107-0 (ST536) 19.3% 41-0 (ST6754) 7.5%	B2 F C Unk
	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
	Second fecal	**Novel** **26** 40	70.70% 27.80% 1.50%	**47** **5**	74.10% 25.90%	**Novel-47 (Novel) 70.7%** **26-5 (ST38) 25.9%** 40-5 (ST131) 1.5%	 D B2
F24	First fecal	**53** ***40***	94.20% 5.80%	**158** ***30*** Novel	93.36% 5.50% 1.14%	**53-158 (ST219) 93.36%** ***40-30 (H30) 5.5% CIP-R*** 53-Novel (Novel) 1.1%	F B2
	Urine	38 23	98.50% 1.50%	18 Novel	98.80% 1.20%	38-18 (ST95) 98.5% 23-Novel (Novel) 1.2%	B2
	Second fecal	11 65 70	80.40% 15.00% 4.60%	39 27 Novel	50.60% 48.10% 1.30%	11-39 (ST5773) 50.6% 65-27 (ST58) 15% 11-27 (ST10) 29.8%	Unk B1 A

						70-27 (Novel) 3.3% 70-Novel (Novel) 1.3%	
P2	First fecal	**4** **214** ***40*** **24**	**81.2%** **13.1%** **4%** **1.6%**	**34** ***30*** **9**	**94.2%** **3.1%** **2.7%**	**4-34 (ST399) 80.2%** **214-0 (ST1316) 12.8%** ***40-30 (H30) 3.5%*** **24-9 (Novel) 1.2%**	C A B2
	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100%*	B2
	Second fecal	**40**	**100%**	**30**	**100%**	**40-30 (*H*30) 100%**	B2
F2002	First fecal	35 4 11 *14*	54.40% 31.50% 9.10% 5.00%	*30* 24	94.50% 5.50%	35-30 (ST394) 54.5% 4-30 (ST58) 31.5% 11-24 (ST10) 5.5% *14-30 (ST550) 5% CIP-S*	D B1 A B2
	Urine	35	100%	30 25	98.70% 1.30%	35-30 (ST394) 98.7% 0-25 (ST0) 1.3%	D
	Second fecal	14 11	93.60% 6.90%	30 215	94.60% 5.40%	14-30 (ST550) 93.6% CIP-S 11-215 (10) 5.4%	B2 A
P8	First fecal	41	100%	86	100%	41-86 (ST101) 100%	B1
	Urine	41 43	73.60% 26.40%	86 9	85.00% 15.00%	41-86 (ST101) 73.6% 43-0 (ST-38) 11.4% 43-9 (Novel) 15%	B1 Unk
	Second fecal	41	100%	86	100%	41-86 (ST101) 100%	B1
P11	First fecal	40 *4* 11	75.60% 12.90% 11.50%	5 *21* 35 34	49.90% 23.00% 13.50% 12.80%	40-5 (ST131) 49.9% *40-21 (ST357) 23%* 11-34 (ST10) 11.5% 4-35 (ST23) 13.5%	B2 B2 A C
	Urine	***40*** **38**	**80%** **20%**	***21*** **5** **35**	**78.10%** **16.40%** **5.50%**	***40-21 (ST357) 78.1%*** **38-5 (ST569) 16.4%** 38-35 (Novel) 5.5%	B2 B2
	Second fecal	23 *40*	98.50% 1.60%	38 *21*	98.20% 1.80%	23-38 (ST58) 98.2% *40-21 (ST357) 1.6%*	B1 B2

F846	First fecal	*14* 35	98.70% 1.30%	*64* 47 300	97.00% 1.70% 1.30%	*14-64 (ST1193) 97% CIP-R* 35-47 (ST69) 1.3% 14-300 (Novel) 1.3%	B2 D
	Urine	7 *14* 218	84.70% 10.60% 4.60%	41 82 47 135 87 *64*	55.60% 18.10% 14.50% 6.20% 4.70% 1.00%	7-41 (ST10) 55.6% 0-82 (ST-31) 18.1% 7-47 (ST6846) 9.9% 218-47 (ST1670) 4.6% *14-64 (ST1193) 1%* *CIP-R* 7-87 (Novel) 4.7% 14-135 (Novel) 6.2%	B1 Unk Unk D B2
	Second fecal	11 38 6	95.60% 2.50% 1.90%	27 33	93.00% 7.00%	11-27 (ST10) 95.6% 38-33 (Novel) 2.5% 6-33 (ST109) 1.9%	A B1
F510	First fecal	70 11 38	96.80% 1.20% 2.00%	305 18	94.10% 5.90%	70-305 (ST6701) 94.1% 11-18 (Novel) 1.2% 38-18 (ST95) 2%	Unk B2
	Urine	*35*	100%	*47*	100%	*35-47 (ST69) 100%* *CIP-S*	B2
	Second fecal	*35* 38 Novel	94.50% 4.00% 1.50%	*47* 64 30	92.20% 4.90% 2.90%	*35-47 (ST69) 92.2%* *CIP-S* 38-30 (ST95) 4% Novel-64 (Novel) 1.5%	D B2
F385	First fecal	*11* 901 38	97.80% 1.30% 0.90%	*215* 30 65 406	90.40% 5.60% 2.30% 1.70%	*11-215 (ST10) 90.4%* *CIP-S* 11-65 (ST6775) 2.3% 11-30 (ST665) 5.6% 901-406 (Novel)1.3% 38-30 (ST95) 0.9%	A Unk A B2
	Urine	*11* 38	98.60% 1.40%	*215* 15	96.90% 1.90%	*11-215 (ST10) 96.9%* *CIP-S* 11-30 (ST665) 1.2%	A A

				30	1.20%	38-15 (ST95) 1.4%	B2
	Second fecal	11 38 Novel	97.50% 1.40% 1.10%	54 30 27 Novel	93.10% 3.50% 1.60% 1.80%	11-54 (ST10) 93.1% 38-27 (ST95) 1.6% 11-30 (ST665) 3.5% Novel-Novel 1.1%	A B2 A
F26	First fecal	11 *14*	58.20% 41.80%	137 *64*	48.60% 51.40%	11-137 (ST10) 48.6% *14-64 (ST1193) 41.8% CIP-R*	A B2
	Urine	*14* 4 Novel	75.30% 21.80% 1.40%	*64*	100%	*14-64 (ST1193) 75.3% CIP-R* 4-0 (ST648) 21.8% Novel-64 (Novel) 1.4%	B2 F
	Second fecal	4 *14* 54 65	71.90% 21.90% 5.10% 1.10%	54 *64* 47	63.00% 36.00% 1.00%	4-54 (ST56) 63% *14-64 (ST1193) 21.9% CIP-R* 4-64 (Novel) 8.4% 54-47 (ST3057) 1% 54-0 (ST2562) 4.1% 65-0 (ST297) 1.1%	B1 B2 I I B1
F205	First fecal	*14*	100%	*64*	100%	*14-64 (ST1193) 100% CIP-R*	B2
	Urine	**40** ***14*** Novel	96.92% 2.20% 0.88%	**20** ***64***	99.00% 1.00%	**40-20 (ST429) 96.92%** ***14-64 (ST1193) 1% CIP-R*** Novel-20 (Novel) 0.88%	B2 B2
	Second fecal	***14*** **40** **29**	77.20% 18.70% 4.10%	***64*** **20** **31**	79.70% 18.10% 2.20%	***14-64 (ST1193) 77.2% CIP-R*** **40-20 (ST429) 18.1%** **29-31 (ST1248) 2.2%**	B2 B2 C
F236	First fecal	*65*	100%	*38*	100%	*65-38 (ST297) 100% CIP-S*	B2
	Urine	*65*	100%	*38*	100%	*65-38 (ST297) 100% CIP-S*	B2

	Second fecal	*65*	100%	*38*	100%	*65-38 (ST297) 100% CIP-S*	B2
F67	First fecal	*14*	100%	*64*	100%	*14-64 (ST1193) 100% CIP-R*	B2
	Urine	*14*	100%	*64*	100%	*14-64 (ST1193) 100% CIP-R*	B2
	Second fecal	*14*	100%	*64*	100%	*14-64 (ST1193) 100% CIP-R*	B2
P3	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
	Second fecal	7	100%	93 30	64.20% 35.60%	7-93 (Novel) 64.2% 7-30 (Novel) 35.6%	
P7	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
	Second fecal	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
P6	First fecal	***40*** **108**	96.80% 3.20%	***30*** **75**	97.20% 2.80%	***40-30 (H30) 96.8% CIP-R*** **108-75 (ST636) 2.8%**	B2 B2
	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2
F486	First fecal	**38** ***40***	45.30% 54.70%	**41** ***30***	55.20% 44.80%	**38-41 (ST95) 45.3%** ***40-30 (H30) 44.8% CIP-R***	B2 B2
	Urine	**36** ***40*** 47	83.80% 14.90% 1.80%	**24** ***30***	88.10% 11.80%	**36-24 (ST349) 82%** ***40-30 (H30) 11.8% CIP-R*** 47-24 (Novel) 1.8%	D B2
P10	First fecal	**38** **36** ***40***	71.50% 17.20% 8.10%	**41** **93** ***30***	77.20% 15.50% 7.30%	**38-41 (ST95) 71.5%** ***40-30 (H30) 7.3%*** **36-93 (ST349) 15.5%**	B2 B2 D
	Urine	*40*	100%	*30*	100%	*40-30 (H30) 100% CIP-R*	B2

Supplemental Table 2

List of all SNPs detected in same-host fecal and urinary *H*30 single-colony isolates. Stop codons are indicated with an asterisk. Presence of SNP is indicated with an 'x'. Lack of SNP is indicated with an '-'. Synonymous mutations are indicated with '(syn)'.

Gene	Mutation	Gene product	P1 isolates		
			First fecal	Second fecal	Urinary
fixB	A174S	electron transfer flavoprotein subunit alpha/FixB family protein	-	x	x
aroP	Upstream	aromatic amino acid transporter AroP	-	x	-
dtd	*85E	aminoacyl-tRNA hydrolase	-	x	-
P423_RS01525	Y101D	hypothetical protein	-	x	x
pdeL	L218 (syn)	cyclic di-GMP phosphodiesterase	-	x	x
mhpA	E142 (syn)	3-(3-hydroxy-phenyl)propionate/3-hydroxycinnamic acid hydroxylase	-	x	x
panE	T251A	2-dehydropantoate 2-reductase	-	x	x
acrR	L203P	transcriptional regulator acrR	-	x	x
mscK	M960T	mechanosensitive channel MscK	-	x	x
htpG	S444 (syn)	molecular chaperone HtpG	-	x	x
P423_RS02600	V713A	ABC transporter permease	-	x	x
wcaG	P115 (syn)	NAD(P)-dependent oxidoreductase wcaG	-	x	x
pepN	A601 (syn)	aminopeptidase N	-	x	x
ghrA	E256D	glyoxylate/hydroxypyruvate reductase A	-	x	x
plsX	S82I	phosphate acyltransferase	-	x	x
P423_RS06870	G76V	hypothetical protein	-	x	x
narK	Upstream	Nitrate/nitraite transporter NarK	-	x	x
narG	N1186 (syn)	nitrate reductase subunit alpha	-	x	x
rssB	A145T	two-component system response regulator RssB	-	x	x
P423_RS07870	Upstream	CMD domain-containing protein	-	x	x
pspE	A19 (syn)	thiosulfate sulfurtransferase PspE	-	x	x
sapA	Upstream	peptide ABC transporter substrate-binding protein SapA	-	x	x
P423_RS07975	L88P	sucrose phosphorylase	-	x	x
P423_RS08365	Upstream	DUF2554 domain-containing protein	-	x	x

cfa	P88L	cyclopropane-fatty-acyl-phospholipid synthase	-	x	x
P423_RS09380	E227*	acyl-CoA dehydrogenase	-	x	x
P423_RS11770	R545Q	TonB-dependent siderophore receptor	-	x	x
manC	I355 (syn)	mannose-1-phosphate guanylyltransferase 1	-	x	x
mglB	L326 (syn)	galactose ABC transporter substrate-binding protein	-	x	-
P423_RS13095	Upstream	MFS transporter	-	x	x
nuoE	L134 (syn)	NADH-quinone oxidoreductase subunit E	-	x	x
folX	F57Y	dihydroneopterin triphosphate 2'-epimerase	-	x	x
pdxB	P148 (syn)	4-phosphoerythronate dehydrogenase PdxB	-	x	x
pdxB	K86N	4-phosphoerythronate dehydrogenase PdxB	-	x	x
purL	A170 (syn)	phosphoribosylformylglycinamidine synthase	-	x	x
pdxJ	E240 (syn)	pyridoxine 5'-phosphate synthase	-	x	x
sdaB	G288 (syn)	L-serine ammonia-lyase	-	x	x
yggF	C165R	fructose-bisphosphatase class II	-	x	x
P423_RS17755	I8 (syn)	TIGR00156 family protein	-	x	x
qseC	I240F	two-component system sensor histidine kinase QseC	-	x	x
P423_RS18155	Upstream	YgjV family protein	-	x	x
aaeB	H285R	p-hydroxybenzoic acid efflux pump subunit AaeB	-	x	-
tldD	A29P	metalloprotease TldD	-	x	-
P423_RS19480	P219R	hydrolase	-	x	-
P423_RS19500	G49R	hypothetical protein	-	x	x
livH	A242V	branched-chain amino acid ABC transporter permease	-	x	x
P423_RS20115	Upstream	GntR family transcriptional regulator	-	x	x
hdfR	Upstream	transcriptional regulator HdfR	-	x	x
fabR	L87V	TetR family transcriptional regulator FabR	-	x	x
soxR	G145 (syn)	redox-sensitive transcriptional activator SoxR	-	x	x
qorB	Upstream	NAD(P)-dependent oxidoreductase	-	x	x
P423_RS25160	K393N	DNA cytosine methyltransferase	-	x	x

Gene	Mutation	Gene product	P4 isolates		
			First fecal	Second fecal	Urinary

P423_RS09705	D103A	CDP-alcohol phosphatidyltransferase family protein	-	N/A	x
Gene	Mutation	Gene product	P5 isolates		
			First fecal	Second fecal	Urinary
ahr	G181 (syn)	aldehyde reductase Ahr	-	x	x
ppc	Upstream	phosphoenolpyruvate carboxylase	-	x	-
Gene	Mutation	Gene product	P6 isolates		
			First fecal	Second fecal	Urinary
artQ	G80D	arginine ABC transporter permease ArtQ	-	x	-
plsC	A46 (syn)	1-acylglycerol-3-phosphate O-acyltransferase	-	x	-
Gene	Mutation	Gene product	P7 isolates		
			First fecal	Second fecal	Urinary
murQ	E247G	N-acetylmuramic acid 6-phosphate etherase	-	-	x
P423_RS20935	Upstream	glycosyltransferase family 1 protein	-	-	x
Gene	Mutation	Gene product	P8 isolates		
			First fecal	Second fecal	Urinary
cysH	A280G	phosphoadenosine phosphosulfate reductase	-	x	-
fldB	I133V	flavodoxin-2	-	-	x
potB	F57L	spermidine/putrescine ABC transporter permease PotB	-	x	x
Gene	Mutation	Gene product	P9 isolates		
			First fecal	Second fecal	Urinary
P423_RS00715	H207	carbonic anhydrase	-	x	-
araH	T52K	arabinose ABC transporter permease	-	-	x
flhA	N449D	formate hydrogenlyase transcriptional activator FlhA	-	-	x
aspA	L12 (syn)	aspartate ammonia-lyase	-	-	x
frdD	D111V	fumarate reductase subunit D	-	-	x
P423_RS24930	Upstream	TetR family transcriptional regulator	-	-	x
caiA	R352 (syn)	crotonobetainyl-CoA dehydrogenase	-	-	x
P423_RS27245	T57I	hypothetical protein	-	-	x
pbp2	D312G	penicillin-binding protein 2	-	-	x
P423_RS04315	M49I	MFS transporter	-	-	x

P423_RS05890	Upstream	lipoprotein	-	-	x
P423_RS06380	T3A	RusA family crossover junction endodeoxyribonuclease	-	-	x
acnA	G500C	aconitate hydratase	-	-	x
rnfD	T187 (syn)	electron transport complex subunit D	-	-	x
P423_RS11780	R787 (syn)	hypothetical protein	-	-	x
yegW	Upstream	GntR family transcriptional regulator	-	-	x
psuG	P202S	pseudouridine-5'-phosphate glycosidase	-	-	x
barA	S106*	two-component sensor histidine kinase BarA	-	-	x
P423_RS17215	Y97 (syn)	DUF3987 domain-containing protein	-	-	x
uxaA	Upstream	altronate dehydratase	-	-	x
garL	P81 (syn)	5-keto-4-deoxy-D-glucarate aldolase	-	-	x
mtr	T356S	tryptophan permease	-	-	x
lptA	L144 (syn)	lipopolysaccharide ABC transporter substrate-binding protein LptA	-	-	x
torZ	G406S	molybdopterin guanine dinucleotide-containing S/N-oxide reductase	-	-	x
gppA	E339K	guanosine-5'-triphosphate-2C3'-diphosphate diphosphatase	-	-	x
P423_RS22030	V53A	cytosine permease	-	-	x
frdD	M87I	fumarate reductase subunit D	-	-	x
P423_RS25320	Y21 (syn)	hypothetical protein	-	-	x
P423_RS25740	Q138*	type II toxin-antitoxin system HipA family toxin	-	-	x
Gene	Mutation	Gene product	P10 isolates		
			First fecal	Second fecal	Urinary
P423_RS16510	N28 (syn)	6-phospho-beta-glucosidase	-	-	x
nagC	S341A	N-acetylglucosamine repressor	-	-	x
entE	N489I	2,2C3-dihydroxybenzoate-AMP ligase	-	-	x
xanQ	I94F	xanthine permease XanQ	-	-	x
hipB	W26L	antitoxin HipB	-	x	-
glnG	D61N	nitrogen regulation protein NR(I)	-	x	-
P423_RS07325	R55K	hypothetical protein	-	-	x
P423_RS12995	V270D	DUF2300 domain-containing protein	-	-	x
rpiA	D74 (syn)	ribose-5-phosphate isomerase	-	-	x

frdB	F236Y	fumarate reductase iron-sulfur subunit	-	-	x

Supplemental Table 3

List of all polymorphic positions detected in same-host fecal and urinary *H*30 population genomes (positions are not listed in genomic order). Stop codons are indicated with an asterisk. Synonymous polymorphisms are indicated with '(syn)'. Frequencies of polymorphic positions are indicated in percent of total aligned reads. Coverage in all population genomes was at least 100X per position on the genome.

Gene	Polymorphism	Gene product	Frequency in P1 samples		
			First fecal	Urinary	Second fecal
pdhR	Upstream	pyruvate dehydrogenase complex repressor	0%	0%	58.2%
dtd	E85*	aminoacyl-tRNA hydrolase	96.5%	0.8%	0%
pdeL	L218 (syn)	cyclic di-GMP phosphodiesterase pdeL	96.2%	0%	0%
mhpA	E142 (syn)	3-(3-hydroxy-phenyl)propionate/3-hydroxycinnamic acid hydroxylase mhpA	100.0%	0%	0%
panE	A251T	2-dehydropantoate 2-reductase	99.1%	0%	0%
acrR	L203P	transcriptional regulator acrR	0%	99.5%	98.5%
mscK	T960M	mechanosensitive channel MscK	80.9%	0%	1.0%
htpG	S444 (syn)	molecular chaperone HtpG	0%	99.3%	100.0%
ybbP	V713A	ABC transporter permease ybbP	0%	100.0%	97.9%
bioA	L45 (syn)	adenosylmethionine--8-amino-7-oxononanoate transaminase bioA	98.4%	0%	0%
wcaG	P115 (syn)	NAD(P)-dependent oxidoreductase wcaG	0.7%	99.2%	100.0%
pepN	A601 (syn)	aminopeptidase N pepN	96.7%	0%	0%
rutR	A83S	transcriptional regulator rutR	0%	6.2%	1.5%
ghrA	E256D	glyoxylate/hydroxypyruvate reductase A ghrA	0.6%	99.3%	100.0%

plsX	S82I	phosphate acyltransferase plsX	0.7%	97.2%	100.0%
tmk	G31R	thymidylate kinase tmk	0%	8.3%	0%
P423_RS06870	V76G	hypothetical protein	94.4%	1.0%	2.5%
narK	Upstream	nitrate/nitrite transporter NarK	0%	100.0%	100.0%
narG	N1186(syn)	nitrate reductase subunit alpha narG	99.4%	0%	0.6%
rssB	A145T	two-component system response regulator RssB	0.7%	99.4%	98.9%
P423_RS07180	Upstream	CMD domain-containing protein	99.4%	1.3%	0%
pspE	A19 (syn)	thiosulfate sulfurtransferase PspE	99.4%	0.7%	1.0%
sapA	Upstream	peptide ABC transporter substrate-binding protein SapA	98.6%	0%	0%
P423_RS07975	P88L	sucrose phosphorylase	99.2%	0%	0%
P423_RS08365	Upstream	DUF2554 domain-containing protein	99.5%	0%	0%
cfa	L88P	cyclopropane-fatty-acyl-phospholipid synthase	99.3%	0%	0%
P423_RS09380	E227*	acyl-CoA dehydrogenase	0%	98.1%	100.0%
P423_RS11770	R545Q	TonB-dependent siderophore receptor	0%	97.8%	100.0%
manC	I355 (syn)	mannose-1-phosphate guanylyltransferase 1	0%	100.0%	98.4%
mglB	L326 (syn)	galactose ABC transporter substrate-binding protein mglB	0%	0%	49.0%
fruK	A226 (syn)	1-phosphofructokinase	0%	13.6%	0%
P423_RS13095	Upstream	MFS transporter	0%	96.7%	100.0%
nuoE	L134 (syn)	NADH-quinone oxidoreductase subunit E	100.0%	0%	0%

folX	Y57F	dihydroneopterin triphosphate 2'-epimerase	100.0%	0%	0%
dedD	P24L	cell division protein DedD	100.0%	0%	0%
pdxB	P148 (syn)	4-phosphoerythronate dehydrogenase PdxB	0%	98.5%	99.0%
P423_RS 13820	G438 (syn)	sensor domain-containing phosphodiesterase	0%	100.0%	100.0%
talA	K86N	transaldolase A	0%	100.0%	100.0%
purL	A170 (syn)	phosphoribosylformylglycinamidine synthase	93.9%	0%	0%
pdxJ	E240 (syn)	pyridoxine 5'-phosphate synthase	0.6%	99.3%	98.8%
sdaB	G288 (syn)	L-serine ammonia-lyase sdaB	0%	97.4%	98.7%
yggF	R165C	fructose-bisphosphatase class II yggF	66.9%	0%	0%
lptG	T92 (syn)	LPS export ABC transporter permease LptG	99.3%	0%	0%
P423_RS 17520	A157D	YtfJ family protein	99.0%	0%	0%
P423_RS 17755	I8 (syn)	TIGR00156 family protein	0.8%	97.9%	98.8%
qseC	F240I	two-component system sensor histidine kinase QseC	99.1%	0%	0.9%
P423_RS 18155	Upstream	YgjV family protein	0%	98.0%	100%
tldD	A29P	metalloprotease TldD	0%	0%	95.9%
P423_RS 19480	P219R	hydrolase	0.6%	0%	97.8%
P423_RS 19500	G49R	hypothetical protein	0%	99.5%	100.0%
livH	A242V	branched-chain amino acid ABC transporter permease LivH	0%	92.4%	95.3%

P423_RS20115	Upstream	GntR family transcriptional regulator	100.0%	0.6%	0.8%
hdfR	Upstream	transcriptional regulator hdfR	100.0%	0%	0%
fabR	V87L	TetR family transcriptional regulator fabR	71.9%	0.8%	0%
metH	Upstream	methionine synthase	0%	61.1%	0%
soxR	G145 (syn)	redox-sensitive transcriptional activator SoxR	97.1%	0%	0%
qorB	Upstream	NAD(P)-dependent oxidoreductase qorB	1.2%	100.0%	100.0%
P423_RS25160	K393N	DNA cytosine methyltransferase	0%	99.2%	99.2%
P423_RS01960	V469 (syn)	autotransporter outer membrane beta-barrel domain-containing protein	97.8%	99.5%	100.0%
P423_RS02110	Upstream	maltodextrin glucosidase	99.0%	99.5%	100.0%
P423_RS02355	V386 (syn)	ammonium transporter	97.7%	97.6%	98.7%
P423_RS03305	L143 (syn)	apolipoprotein N-acyltransferase	97.4%	100.0%	100.0%
P423_RS04100	Upstream	DksA/TraR family C4-type zinc finger protein	98.1%	100.0%	100.0%
wcaD	H111Y	putative colanic acid polymerase WcaD	97.8%	100.0%	98.9%
P423_RS12590	Upstream	DUF418 family protein	98.4%	99.3%	100.0%
lrhA	F189L	transcriptional regulator LrhA	97.1%	97.1%	96.5%
P423_RS14420	T182 (syn)	IscS subfamily cysteine desulfurase	96.9%	100.0%	100.0%
P423_RS14470	A204S	ABC transporter permease	98.3%	100.0%	98.6%
hycE	D168Y	hydrogenase large subunit	96.8%	96.6%	100.0%
P423_RS15585	Upstream	MBL fold metallo-hydrolase	98.5%	98.5%	100.0%

aslB	Upstream	anaerobic sulfatase maturase AslB	98.4%	97.0%	98.8%
P423_RS23340	Upstream	ADP-forming succinate--CoA ligase subunit beta	95.4%	97.5%	99.2%
Gene	Polymorphism	Gene product	Frequency in P2 samples		
			First fecal	Urinary	Second fecal
mipA	S94N	MltA-interacting protein	0%	18.4%	0%
murQ	G247E	N-acetylmuramic acid 6-phosphate etherase murQ	0%	16.9%	96.1%
purL	G623 (syn)	phosphoribosylformylglycinamidine synthase	0%	5.3%	100%
uacT	V211I	Urate/xanthine permease	0%	1.7%	100%
fbaA	D215(syn)	ketose-bisphosphate aldolase	0%	64.7%	0%
uvrD	G669D	DNA-dependent helicase	11%	0%	97.3%
recQ	G372C	DNA helicase RecQ	0.0%	0.0%	97.3%
P423_RS00480	E324 (syn)	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	97.5%	98.7%	98.6%
hrpB	R106Q	ATP-dependent helicase HrpB	95.7%	98.0%	97.1%
lacI	E297G	DNA-binding transcriptional repressor LacI	98.0%	97.6%	98.7%
P423_RS02305	V196A	peptidylprolyl isomerase	98.3%	98.4%	98.8%
putA	G1288S	trifunctional transcriptional regulator/proline dehydrogenase/L-glutamate gamma-semialdehyde dehydrogenase	98.0%	100.0%	98.5%
lrhA	F189L	transcriptional regulator LrhA	98.1%	96.7%	100.0%
hycE	D168Y	hydrogenase large subunit	97.9%	98.1%	98.8%
P423_RS15585	Upstream	MBL fold metallo-hydrolase	98.2%	100.0%	100.0%
lptG	T92	LPS export ABC transporter permease LptG	98.1%	98.5%	100.0%

P423_RS17755	I8 (syn)	TIGR00156 family protein	98.3%	100.0%	100.0%
dsbL	A165 (syn)	thiol:disulfide interchange protein DsbL	98.4%	97.7%	97.6%
P423_RS19030	T302 (syn)	multidrug efflux RND transporter permease subunit	95.1%	98.4%	100.0%
P423_RS22445	Y430 (syn)	glycoporin	97.5%	100.0%	100.0%
P423_RS23035	Upstream	hypothetical protein	98.1%	100.0%	100.0%
P423_RS23340	Upstream	ADP-forming succinate--CoA ligase subunit beta	96.8%	97.8%	92.8%
nrfD	T2 (syn)	cytochrome c nitrite reductase subunit NrfD	97.1%	98.7%	97.6%
frdB	I477 (syn)	fumarate reductase flavoprotein subunit	97.1%	98.7%	98.9%
P423_RS03305	L143 (syn)	apolipoprotein N-acyltransferase	99.0%	100.0%	97.2%
nrdI	S41P	class Ib ribonucleoside-diphosphate reductase assembly flavoprotein NrdI	100.0%	98.3%	96.2%
torZ	N237S	molybdopterin-dependent oxidoreductase Mo/Fe-S-binding subunit	100.0%	98.8%	97.8%
Gene	Polymorphism	Gene product	Frequency in P3 samples		
			First fecal	Urinary	Second fecal
fadE	S786R	acyl-CoA dehydrogenase	100%	100%	93.6%
bioA	P407S	adenosylmethionine--8-amino-7-oxononanoate transaminase	92.1%	95.5%	98.0%
zapE	Upstream	cell division protein ZapE	98.6%	100.0%	91.8%
aaeB	R504 (syn)	p-hydroxybenzoic acid efflux pump subunit AaeB	11.3%	1.3%	0%
yiaA	Q140*	Inner membrane protein YiaA	10.8%	0%	0%
sgcC	Upstream	putative PTS enzyme IIC component SgcC	19.2%	6.7%	6.7%

phoE	Upstream	phosphoporin PhoE	97.5%	96.8%	96.6%
P423_RS02110	Upstream	maltodextrin glucosidase	96.9%	100.0%	99.0%
mutS	L796 (syn)	DNA mismatch repair protein MutS	97.2%	97.3%	97.1%
P423_RS15585	Upstream	MBL fold metallo-hydrolase	95.0%	99.1%	99.0%
P423_RS23125	G110V	sugar-binding transcriptional regulator	97.2%	100.0%	100.0%
caiC	Q109 (syn)	crotonobetaine/carnitine-CoA ligase	98.2%	97.7%	98.2%
leuA	D14 (syn)	2-isopropylmalate synthase	100.0%	100.0%	97.2%
rpoA	G840 (syn)	DNA polymerase III subunit alpha	100.0%	97.8%	96.2%
brnQ	V210 (syn)	branched-chain amino acid transporter 2 carrier protein BrnQ	100.0%	100.0%	98.3%
P423_RS02920	H471P	carbohydrate kinase	98.4%	98.7%	97.9%
P423_RS03305	L143 (syn)	apolipoprotein N-acyltransferase	100.0%	100.0%	97.4%
astE	A58 (syn)	succinylglutamate desuccinylase	100.0%	98.4%	96.6%
P423_RS12665	Upstream	pseudouridine kinase	100.0%	100.0%	97.1%
gdpG	L38 (syn)	type II secretion system protein GspG	98.0%	99.0%	96.7%
P423_RS17440	G396 (syn)	DUF4092 domain-containing protein	98.0%	100.0%	98.4%
P423_RS20935	Upstream	glycosyltransferase family 1 protein	100.0%	97.6%	97.9%
atpD	R107C	ATP synthase subunit beta	100.0%	99.2%	95.5%
rmuC	S395N	DNA recombination protein RmuC	100.0%	98.8%	98.1%
P423_RS23340	Upstream	ADP-forming succinate--CoA ligase subunit beta	98.6%	96.4%	96.4%

P423_RS09305	G340D	cysteine desulfurase	100.0%	96.0%	100.0%
wcaD	H111Y	putative colanic acid polymerase WcaD	97.9%	97.6%	100.0%
mdtB	A384V	multidrug resistance protein MdtB	98.1%	98.1%	100.0%
P423_RS12590	Upstream	DUF418 family protein	100.0%	97.3%	100.0%
nuoG	C573R	NADH-quinone oxidoreductase subunit NuoG	96.5%	96.9%	100.0%
P423_RS14595	E229G	elongation factor 4	100.0%	97.2%	100.0%
vgrG	A636 (syn)	type VI secretion system tip protein VgrG	94.7%	100.0%	98.9%
P423_RS17880	A157D	YtfJ family protein	98.7%	97.8%	100.0%
P423_RS20085	S451P	nickel ABC transporter nickel/metallophore periplasmic binding protein	100.0%	98.1%	100.0%
P423_RS22840	P203A	UDP-N-acetylmuramate dehydrogenase	100.0%	98.0%	100.0%

Gene	Polymorphism	Gene product	Frequency in P4 samples		
			First fecal	Urinary	Second fecal
P423_RS06275	Upstream	DUF4222 domain-containing protein	0%	6.8%	N/A
P423_RS09705	A103D	CDP-alcohol phosphatidyltransferase family protein	0%	83.1%	
eutA	A383V	ethanolamine ammonia-lyase reactivating factor EutA	49.6%	0.9%	
lysM	Upstream	peptidoglycan-binding protein LysM	53.8%	0%	
ttdT	R161H	L-tartrate:succinate antiporter	84.0%	90.9%	

caiE	G188R	carnitine operon protein CaiE	99.2%	97.6%	
P423_RS00480	V271L	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	99.1%	96.3%	
P423_RS01035	Upstream	VOC family protein	99.2%	98.5%	
P423_RS02355	V386 (syn)	ammonium transporter	98.2%	95.4%	
P423_RS02660	A241 (syn)	allantoin permease	100.0%	98.2%	
P423_RS09385	A16V	AraC family transcriptional regulator	100.0%	98.3%	
P423_RS09595	I341L	PTS diacetylchitobiose transporter subunit IIC	97.4%	98.0%	
P423_RS13325	F15Y	DUF412 domain-containing protein	100.0%	98.2%	
qseE	V33A	two component system sensor histidine kinase QseE/GlrK	100.0%	98.6%	
P423_RS18015	R161H	anion permease	84.0%	90.9%	
P423_RS18050	Upstream	siderophore-interacting protein	99.1%	98.2%	
P423_RS20935	Upstream	glycosyltransferase family 1 protein	100.0%	98.3%	
aslB	Upstream	anaerobic sulfatase maturase AslB	99.0%	98.1%	
ntrC	P48T	nitrogen regulation protein NR(I)	100.0%	98.5%	
P423_RS25605	Upstream	primosomal protein 1	100.0%	95.2%	
Gene	Polymorphism	Gene product	Frequency in P5 samples		
			First fecal	Urinary	Second fecal
rclA	V275F	pyridine nucleotide-disulfide oxidoreductase	1.0%	0.5%	4.5%

sucA	A468 (syn)	2-oxoglutarate dehydrogenase E1 component	0%	0%	73.3%
rsxC	E572 (syn)	electron transport complex subunit RsxC	8.9%	5.9%	6.4%
sppA	T149S	signal peptide peptidase SppA	98.6%	94.9%	98.3%
astE	A58 (syn)	succinylglutamate desuccinylase astE	94.5%	98.3%	98.4%
hexR	L124 (syn)	MurR/RpiR family transcriptional regulator hexR	1.0%	0%	31.1%
lrhA	F189L	transcriptional regulator LrhA	92.4%	98.0%	97.4%
yfcV	Upstream	fimbrial yfcV	0%	0%	20.5%
cysK	K55N	cysteine synthase A cysK	1.1%	3.1%	2.6%
csiR	A86V	transcriptional regulator	94.4%	95.1%	94.4%
P423_RS15585	Upstream	MBL fold metallo-hydrolase	93.3%	98.2%	97.0%
yhfT	A91 (syn)	membrane protein	1.9%	0.5%	2.9%
atpA	F126 (syn)	ATP synthase subunit alpha	0%	2.2%	7.9%
P423_RS23100	V149 (syn)	PTS mannose transporter subunit IID	5.3%	4.3%	8.2%
sucC	Upstream	ADP-forming succinate--CoA ligase subunit beta	93.8%	96.8%	95.2%
adiA	Upstream	arginine decarboxylase	3.3%	2.6%	3.8%
prfC	A17 (syn)	peptide chain release factor 3	13.1%	61.5%	1.1%
surA	L307 (syn)	chaperone SurA	97.0%	99.3%	98.5%
P423_RS01710	A334	autotransporter outer membrane beta-barrel domain-containing protein	98.1%	97.7%	100.0%
P423_RS01740	E211 (syn)	betaine-aldehyde dehydrogenase	95.7%	98.2%	100.0%
P423_RS01960	V469 (syn)	autotransporter outer membrane beta-barrel domain-containing protein	95.9%	99.0%	99.2%

P423_RS03285	T39I	glutamate ABC transporter permease	97.9%	96.8%	99.3%
P423_RS05710	T254 (syn)	GGDEF domain-containing protein	96.3%	97.6%	100.0%
chaB	D45 (syn)	cation transport regulator	97.2%	98.9%	97.5%
P423_RS08610	A536V	oxidoreductase	98.5%	99.6%	98.1%
uidC	Upstream	glucuronide uptake porin UidC	98.0%	100.0%	96.9%
P423_RS10810	P594L	excinuclease ABC subunit C	98.1%	100.0%	98.7%
P423_RS12425	A23 (syn)	VWA domain-containing protein	98.6%	98.9%	97.7%
P423_RS12760	A27 (syn)	Hypothetical protein	93.1%	98.8%	99.6%
arnA	A309D	bifunctional UDP-4-amino-4-deoxy-L-arabinose formyltransferase/UDP-glucuronic acid oxidase ArnA	97.2%	96.9%	98.9%
nuoJ	G77S	NADH-quinone oxidoreductase subunit J	95.1%	95.7%	95.9%
P423_RS13770	Upstream	exoaminopeptidase	98.2%	97.6%	99.1%
qseE	V33A	two component system sensor histidine kinase QseE/GlrK	98.0%	97.3%	98.4%
P423_RS17440	G396	DUF4092 domain-containing protein	98.0%	99.0%	98.5%
lptG	T92 (syn)	LPS export ABC transporter permease LptG	96.9%	98.2%	99.4%
P423_RS17520	A157D	YtfJ family protein	98.5%	98.7%	99.5%
dsbL	A165 (syn)	thiol:disulfide interchange protein DsbL	95.9%	97.4%	98.6%
P423_RS19615	Upstream	phosphoglycolate phosphatase	97.8%	99.3%	100.0%

P423_RS20365	Upstream	transporter	96.0%	96.7%	99.0%
P423_RS20555	R89H	DUF3053 domain-containing protein	98.1%	99.5%	98.7%
emrD	P74 (syn)	multidrug transporter EmrD	96.2%	96.9%	97.8%
caiD	T161 (syn)	carnitinyl-CoA dehydratase	98.6%	99.2%	98.1%
mraZ	Upstream	division/cell wall cluster transcriptional repressor MraZ	98.6%	100.0%	98.6%
lacI	E297G	DNA-binding transcriptional repressor LacI	98.7%	97.2%	98.5%
P423_RS02355	V386 (syn)	ammonium transporter	95.7%	98.8%	96.6%
P423_RS02415	E142K	transcriptional regulator	100.0%	100.0%	98.8%
efeU	I56N	ferrous iron permease EfeU	100.0%	99.2%	98.1%
fabG	A75V	beta-ketoacyl-ACP reductase	100.0%	98.6%	98.8%
P423_RS09765	Y203C	bifunctional nicotinamidase/pyrazinamidase	99.0%	98.9%	97.7%
nuoG	C573R	NADH-quinone oxidoreductase subunit NuoG	98.5%	99.3%	97.6%
P423_RS14495	A88 (syn)	serine hydroxymethyltransferase	100.0%	100.0%	98.9%
hycE	D168Y	hydrogenase large subunit	100.0%	95.6%	98.4%
aslB	Upstream	anaerobic sulfatase maturase AslB	100.0%	96.4%	98.3%
P423_RS25820	C233Y	DNA-binding response regulator	99.2%	99.4%	96.8%
htpG	I433V	molecular chaperone HtpG	100.0%	98.8%	100.0%
Gene	Polymorphism	Gene product	Frequency in P6 samples		
			First fecal	Urinary	Second fecal
yaaA	T89 (syn)	peroxide stress protein YaaA	0.0%	0.0%	7.1%

alsT	L343 (syn)	sodium:alanine symporter family protein	0.0%	0.0%	12.9%
caiD	S236 (syn)	carnitinyl-CoA dehydratase caiD	0.0%	0.0%	8.0%
caiA	A129 (syn)	crotonobetainyl-CoA dehydrogenase caiA	0.0%	0.0%	14.7%
caiT	G354 (syn)	L-carnitine/gamma-butyrobetaine antiporter	0.0%	0.0%	8.3%
araA	P379 (syn)	L-arabinose isomerase	0.0%	0.0%	9.8%
sgrR	R194 (syn)	transcriptional regulator SgrR	0.0%	0.0%	6.7%
hpt	A19 (syn)	hypoxanthine phosphoribosyltransferase hpt	0.0%	0.0%	17.2%
yadE	A368 (syn)	polysaccharide deacetylase family protein	1.4%	0.0%	15.0%
thpR	A111 (syn)	RNA 2',2C3'-cyclic phosphodiesterase thpR	0.0%	0.0%	10.8%
fhuC	T240N	iron-hydroxamate transporter ATP-binding protein fhuC	0.0%	1.6%	9.8%
P423_RS01540	P110L	hypothetical protein	99.2%	98.4%	88.8%
prpB	T25 (syn)	methylisocitrate lyase	0.0%	0.0%	12.8%
fadM	Upstream	thioesterase	0.0%	0.0%	9.2%
allD	A53 (syn)	ureidoglycolate dehydrogenase	0.0%	0.0%	6.4%
fdrA	R317 (syn)	acyl-CoA synthetase FdrA	0.0%	0.0%	9.6%
ylbE	Q350 (syn)	DUF1116 domain-containing protein	0.0%	0.9%	13.6%
fes	A45T	enterochelin esterase fes	0.0%	0.0%	8.0%
uspG	Upstream	universal stress protein G	0.0%	0.0%	9.4%
cobC	A29 (syn)	adenosylcobalamin/alpha-ribazole phosphatase cobC	0.0%	0.0%	9.2%
ybfF	P116 (syn)	acyl-CoA esterase	0.0%	0.0%	13.2%
dinG	A39P	ATP-dependent DNA helicase DinG	100.0%	98.6%	93.7%

ybiX	V60M	PKHD-type hydroxylase	100.0%	100.0%	90.1%
artQ	G80D	arginine ABC transporter permease ArtQ	0.0%	0.0%	10.4%
acnA	T850M	aconitate hydratase	0.0%	0.0%	11.1%
ugpB	R430H	carbohydrate ABC transporter substrate-binding protein	0.0%	0.0%	16.7%
ydcO	T102(syn)	BenE family transporter YdcO	3.9%	12.4%	5.6%
ydeA	A177T	L-arabinose exporter	98.6%	96.9%	90.1%
ppsA	V575(syn)	phosphoenolpyruvate synthase ppsA	0.0%	1.5%	14.8%
entE	A170T	2,2C3-dihydroxybenzoate-AMP ligase	0.0%	0.0%	4.3%
waaF	Upstream	lipopolysaccharide heptosyltransferase family protein	0.0%	0.0%	10.2%
P423_RS11995	D185E	glycosyltransferase family 2 protein	0.0%	100.0%	98.4%
wcaD	L128V	putative colanic acid polymerase WcaD	0.0%	0.0%	88.9%
yegI	Upstream	helix-hairpin-helix domain-containing protein	0.0%	1.2%	8.0%
mdtA	P37 (syn)	multidrug transporter subunit MdtA	0.0%	0.0%	7.7%
P423_RS12400	D769G	DUF4132 domain-containing protein	97.1%	100.0%	93.9%
nrfE	V360 (syn)	heme lyase CcmF/NrfE family subunit	1.6%	2.2%	10.3%
yfaS	K630T	alpha-2-macroglobulin family protein	0.0%	1.1%	6.9%
lrhA	F189L	transcriptional regulator LrhA	98.3%	94.0%	95.8%
eutB	I224V	ethanolamine ammonia-lyase heavy chain eutB	0.0%	0.0%	8.1%
glrR	I304T	two-component system response regulator GlrR	100.0%	100.0%	94.3%
pheA	N167(syn)	prephenate dehydratase	0.0%	0.0%	9.8%

mutS	L796 (syn)	DNA mismatch repair protein MutS	100.0%	97.8%	95.0%
P423_RS15720	K192N	FAD-binding oxidoreductase	0.0%	0.9%	6.6%
syd	E114*	protein Syd	0.0%	0.0%	7.0%
queF	V85A	NADPH-dependent 7-cyano-7-deazaguanine reductase QueF	0.0%	0.0%	7.3%
P423_RS16160	Y48H	PRD domain-containing protein	0.0%	0.0%	11.4%
argA	P120 (syn)	N-acetylglutamate synthase	0.0%	0.0%	7.1%
P423_RS16155	Upstream	PTS glucose transporter subunit IIBC	0.0%	1.1%	9.1%
gcvP	I747T	aminomethyl-transferring glycine dehydrogenase	0.0%	0.8%	6.5%
gshB	T128 (syn)	glutathione synthetase	0.0%	0.0%	5.8%
P423_RS16860	S338 (syn)	nucleoside permease	0.0%	0.0%	4.4%
P423_RS16995	T234 (syn)	TonB-dependent receptor	0.0%	0.0%	5.2%
P423_RS17215	Upstream	DUF3987 domain-containing protein	0.0%	0.0%	5.1%
ag43	R815 (syn)	autotransporter domain-containing protein	96.3%	94.8%	95.6%
ygiR	Upstream	YgiQ family radical SAM protein	0.0%	0.0%	9.4%
plsC	A46 (syn)	1-acylglycerol-3-phosphate O-acyltransferase plsC	0.0%	0.0%	98.3%
fhuD	L111 (syn)	iron-siderophore ABC transporter substrate-binding protein	0.8%	0.0%	7.2%
P423_RS17815	G64 (syn)	ABC transporter ATP-binding protein	0.0%	0.0%	10.5%
cpdA	L246 (syn)	phosphodiesterase	0.0%	0.0%	8.6%
ygiD	G234 (syn)	4,2C5-DOPA dioxygenase extradiol	0.0%	0.7%	5.5%
bacA	L71 (syn)	undecaprenyl-diphosphatase	0.9%	1.0%	7.8%

ttdR	R99 (syn)	transcriptional activator TtdR	0.0%	0.0%	8.0%
rpoD	P263 (syn)	RNA polymerase sigma factor RpoD	1.6%	0.0%	7.4%
aer	R87H	PAS domain S-box protein	0.0%	0.0%	7.2%
exuT	S123 (syn)	MFS transporter	0.8%	0.0%	5.7%
yhbV	L207 (syn)	U32 family peptidase	0.0%	0.0%	8.4%
elbB	A119E	glutamine amidotransferase	1.3%	0.0%	8.7%
yhfW	Y309D	phosphopentomutase	2.0%	2.1%	4.4%
hslR	H37 (syn)	heat-shock protein Hsp15	0.0%	0.8%	6.2%
feoB	Q341 (syn)	ferrous iron transporter B	0.0%	0.9%	5.9%
glpG	H254 (syn)	rhomboid family intramembrane serine protease GlpG	0.0%	0.0%	6.7%
glpE	R98 (syn)	thiosulfate sulfurtransferase	0.0%	0.0%	5.7%
ugpB	R112G	sn-glycerol-3-phosphate ABC transporter substrate-binding protein UgpB	0.0%	0.0%	8.5%
tusA	T29 (syn)	sulfurtransferase TusA	0.0%	0.9%	5.3%
ftsY	Upstream	signal recognition particle-docking protein FtsY	0.8%	0.0%	6.2%
nikC	A248V	nickel ABC transporter permease subunit NikC	99.2%	100.0%	94.2%
P423_RS20115	Upstream	GntR family transcriptional regulator	0.0%	0.0%	5.2%
hlyD	Upstream	membrane protein	0.0%	0.0%	7.0%
uspB	V60M	universal stress protein B	0.0%	0.0%	6.3%
chuT	S72G	ABC transporter substrate-binding protein	96.4%	97.6%	87.0%
chuW	T72 (syn)	heme anaerobic degradation radical SAM methyltransferase ChuW/HutW	0.0%	0.0%	5.5%
P423_RS20275	V218 (syn)	iron ABC transporter permease	0.0%	0.0%	7.6%
mdtE	L349 (syn)	multidrug resistance protein MdtE	0.0%	0.0%	8.1%

mdtF	A47 (syn)	multidrug efflux RND transporter permease subunit	0.0%	0.0%	8.4%
gadE	Upstream	transcriptional regulator GadE	0.0%	0.0%	4.8%
yhjJ	T150I	insulinase family protein	0.0%	0.0%	7.9%
bcsG	G444D	cellulose biosynthesis protein BcsG	97.2%	100.0%	92.7%
dppD	A22 (syn)	dipeptide ABC transporter ATP-binding protein	0.0%	0.0%	8.0%
dppB	K339 (syn)	peptide ABC transporter permease	0.0%	0.0%	6.1%
ulaE	D123 (syn)	L-xylulose 5-phosphate 3-epimerase	0.0%	0.0%	6.5%
P423_RS20710	Upstream	AraC family transcriptional regulator	0.0%	0.0%	9.3%
yibH	P274 (syn)	HlyD family secretion protein	0.0%	0.0%	5.8%
gltS	Y320 (syn)	sodium/glutamate symport carrier protein	0.8%	0.0%	6.4%
xanP	T250 (syn)	xanthine permease XanP	0.0%	0.0%	4.5%
agaY	R117K	aldolase	0.0%	0.0%	9.5%
uhpT	Upstream	hexose-6-phosphate:phosphate antiporter	0.0%	0.0%	6.6%
P423_RS21320	G61 (syn)	hypothetical protein	0.0%	0.0%	8.0%
recF	R79 (syn)	DNA replication and repair protein RecF	1.2%	1.1%	7.0%
atpH	G150 (syn)	ATP synthase subunit delta	0.0%	0.0%	7.6%
atpB	G84 (syn)	ATP synthase subunit A	0.9%	1.0%	8.8%
mnmG	L33(syn)	tRNA uridine(34) 5-carboxymethylaminomethyl synthesis enzyme MnmG	0.0%	0.0%	6.5%
asnC	R11 (syn)	AsnC family transcriptional regulator	0.0%	0.0%	9.9%

rbsB	S34 (syn)	ribose ABC transporter substrate-binding protein RbsB	0.0%	0.9%	6.9%
rbsR	R106 (syn)	ribose operon repressor rbsR	0.0%	0.0%	8.1%
IlvM	T80 (syn)	acetolactate synthase isozyme 2 small subunit	0.0%	0.0%	5.0%
ilvY	Upstream	HTH-type transcriptional activator IlvY	0.0%	0.0%	9.3%
yigA	L123M	DUF484 domain-containing protein	0.9%	0.0%	5.6%
uvrD	G660(syn)	DNA-dependent helicase	0.0%	0.0%	5.7%
rhtC	V8 (syn)	threonine export protein RhtC	0.0%	0.0%	15.9%
rmuC	S395N	DNA recombination protein RmuC	99.0%	98.6%	93.8%
ubiD	R167 (syn)	3-octaprenyl-4-hydroxybenzoate decarboxylase UbiD	0.0%	0.0%	9.9%
yihI	T77A	GTPase-activating protein	0.0%	1.0%	8.0%
hemN	I436 (syn)	oxygen-independent coproporphyrinogen III oxidase	0.0%	0.0%	6.4%
glnA	D19 (syn)	glutamine synthetase	1.0%	0.0%	4.7%
P423_RS22365	Upstream	alpha/beta hydrolase	0.0%	0.0%	12.7%
rhaR	R237 (syn)	HTH-type transcriptional activator RhaR	0.0%	0.0%	9.2%
cysZ	Upstream	sulfate transporter cysZ	3.6%	12.8%	1.7%
malE	Upstream	maltose/maltodextrin ABC transporter substrate-binding protein MalE	1.2%	0.0%	12.6%
PlsB	K547E	glycerol-3-phosphate 1-O-acyltransferase PlsB	0.0%	0.0%	5.6%
lpdA	Upstream	dihydrolipoamide dehydrogenase	98.0%	97.7%	93.8%
actP	A370 (syn)	cation acetate symporter	0.0%	0.0%	7.8%

fdhF	E572 (syn)	formate dehydrogenase H subunit alpha selenocysteine-containing	0.0%	0.0%	8.5%
alsA	D441 (syn)	cytochrome c6	1.1%	0.0%	7.9%
eptA	S384 (syn)	phosphoethanolamine transferase EptA	0.0%	0.0%	8.3%
adiA	D606 (syn)	arginine decarboxylase	0.0%	0.0%	7.6%
P423_RS 23810	G1149(syn)	autotransporter outer membrane beta-barrel domain-containing protein	0.0%	0.0%	6.6%
P423_RS 24155	G342 (syn)	PTS ascorbate transporter subunit IIC	0.0%	0.0%	7.3%
fklB	A188 (syn)	FKBP-type peptidyl-prolyl cis-trans isomerase	0.0%	0.0%	9.6%
P423_RS 24250	Upstream	3-ketoacyl-ACP reductase	0.0%	0.0%	9.2%
P423_RS 24280	L41 (syn)	EamA/RhaT family transporter	0.0%	0.0%	6.7%
msrA	T27(syn)	peptide-methionine (S)-S-oxide reductase	0.0%	0.0%	7.1%
ytfQ	Upstream	ABC transporter substrate-binding protein	0.0%	0.0%	6.9%
P423_RS 24530	L120 (syn)	DUF853 domain-containing protein	0.0%	0.0%	11.4%
ahr	G67 (syn)	aldehyde reductase Ahr	0.0%	0.8%	9.9%
P423_RS 25270	P75 (syn)	hypothetical protein	0.0%	0.0%	10.2%
sgcE	Upstream	protein SgcE	0.0%	1.0%	8.8%
opgB	T88 (syn)	phosphatidylglycerol--membrane-oligosaccharide glycerophosphotransferase	0.0%	1.9%	9.7%
gpmB	V195 (syn)	phosphoglycerate mutase GpmB	0.0%	0.0%	6.8%
creC	L8 (syn)	two-component system sensor histidine kinase CreC	0.0%	0.0%	7.6%
Gene	Polymorphism	Gene product	Frequency in P7 samples		
			First fecal	Urinary	Second fecal

lnt	L143 (syn)	apolipoprotein N-acyltransferase lnt	94.7%	100.0%	98.1%
P423_RS03925	L46 (syn)	anion permease	1.0%	0%	21.6%=
astE	A58 (syn)	succinylglutamate desuccinylase astE	98.3%	100.0%	94.6%
P423_RS13820	Upstream	sensor domain-containing phosphodiesterase	4.8%	1.4%	8.3%
barA	R477H	two-component sensor histidine kinase BarA	0%	32.4%	0.7%
torZ	Q681R	molybdopterin guanine dinucleotide-containing S/N-oxide reductase	6.3%	0%	5.7%
thiC	A543G	phosphomethylpyrimidine synthase ThiC	5.0%	11.9%	1.4%
leuA	D14 (syn)	2-isopropylmalate synthase	97.5%	97.6%	100.0%
lacI	E297G	DNA-binding transcriptional repressor LacI	96.2%	100.0%	96.9%
P423_RS01980	V293A	peptide transporter	98.2%	97.3%	99.1%
P423_RS02180	A10V	thiamine-monophosphate kinase	98.0%	99.3%	98.1%
P423_RS02355	V386 (syn)	ammonium transporter	97.3%	100.0%	100.0%
P423_RS02600	W635*	ABC transporter permease	94.9%	100.0%	100.0%
gsiB	C401G	glutathione ABC transporter substrate-binding protein GsiB	98.3%	100.0%	99.0%
asmA	QRDL196L	outer membrane assembly protein AsmA	97.3%	99.1%	98.9%
P423_RS12590	Upstream	DUF418 family protein	97.5%	100.0%	100.0%
P423_RS12675	G101 (syn)	1-phosphofructokinase	95.2%	100.0%	96.2%
P423_RS13020	L70 (syn)	bifunctional 3-demethylubiquinone 3-O-methyltransferase/2-	97.4%	99.4%	97.2%

		octaprenyl-6-hydroxy phenol methylase			
lrhA	F189L	transcriptional regulator LrhA	97.4%	100.0%	96.4%
argT	T151I	lysine/arginine/ornithine ABC transporter substrate-binding protein ArgT	98.0%	99.3%	98.0%
P423_RS14420	T182 (syn)	IscS subfamily cysteine desulfurase	98.5%	99.4%	98.0%
qseE	V33A	two component system sensor histidine kinase QseE/GlrK	98.1%	100.0%	100.0%
tyrA	T93 (syn)	bifunctional chorismate mutase/prephenate dehydrogenase	98.4%	100.0%	99.2%
hycE	D168Y	hydrogenase large subunit	97.9%	98.5%	98.1%
mutS	L796 (syn)	DNA mismatch repair protein MutS	97.5%	100.0%	97.0%
P423_RS15585	Upstream	MBL fold metallo-hydrolase	96.5%	99.4%	99.1%
zapA	E64K	cell division protein ZapA	97.8%	100.0%	99.3%
lptG	T92 (syn)	LPS export ABC transporter permease LptG	98.2%	99.4%	97.4%
P423_RS17520	A157D	YtfJ family protein	98.8%	98.5%	97.2%
P423_RS20085	S451P	nickel ABC transporter nickel/metallophore periplasmic binding protein	98.8%	100.0%	99.3%
uhpB	A122 (syn)	signal transduction histidine-protein kinase/phosphatase UhpB	97.0%	98.8%	98.4%
gyrB	A456E	DNA gyrase subunit B	97.3%	100.0%	97.1%
aslB	Upstream	anaerobic sulfatase maturase AslB	97.4%	100.0%	99.0%
rmuC	S395N	DNA recombination protein RmuC	98.0%	100.0%	98.4%

P423_RS22515	R80C	DNA-binding response regulator	97.7%	99.4%	96.9%
P423_RS22550	G234S	triose-phosphate isomerase	97.9%	100.0%	99.3%
P423_RS23035	Upstream	hypothetical protein	97.6%	98.0%	97.9%
P423_RS23115	A27 (syn)	PTS sorbose transporter subunit IIA	97.3%	98.3%	100.0%
P423_RS23340	Upstream	ADP-forming succinate--CoA ligase subunit beta	95.1%	97.5%	94.2%
P423_RS23685	I289V	hypothetical protein	97.9%	100.0%	100.0%
P423_RS23910	Upstream	hypothetical protein	96.9%	100.0%	98.4%
P423_RS24245	R240 (syn)	sugar phosphate isomerase/epimerase	97.2%	100.0%	100.0%
P423_RS03505	Upstream	putrescine-ornithine antiporter	100.0%	98.6%	98.2%
uidC	Upstream	glucuronide uptake porin UidC	100.0%	100.0%	95.9%
P423_RS14680	R708 (syn)	protein lysine acetyltransferase	100.0%	100.0%	97.6%
P423_RS17440	G396 (syn)	DUF4092 domain-containing protein	100.0%	100.0%	98.6%
P423_RS22700	L178R	fructose-6-phosphate aldolase	100.0%	100.0%	99.1%
P423_RS24250	C89Y	3-ketoacyl-ACP reductase	100.0%	100.0%	98.6%
Gene	Polymorphism	Gene product	Frequency in P8 samples		
			First fecal	Urinary	Second fecal
apaG	A75(syn)	Co2+/Mg2+ efflux protein ApaG	0.0%	0.7%	23.0%
gsk	R335C	inosine/guanosine kinase gsk	35.5%	0.0%	0.0%
flgA	Q184L	flagellar basal body P-ring formation protein FlgA	0.0%	4.8%	0.0%

potB	L57F	spermidine/putrescine ABC transporter permease PotB	25.0%	0.0%	0.0%
narK	F97 (syn)	NarK family nitrate/nitrite MFS transporter	44.9%	2.2%	0.0%
cysH	A280G	phosphoadenosine phosphosulfate reductase CysH	0.0%	0.6%	99.1%
cysZ	P168L	sulfate transporter CysZ	0.0%	89.5%	0.0%
tas	D287A	NADP(H)-dependent aldo-keto reductase	0.0%	18.5%	0.0%
fldB	I133V	flavodoxin-2	0.0%	95.5%	2.4%
P423_RS16700	A192S	hypothetical protein	0.0%	0.0%	73.5%
P423_RS20790	R220Q	hypothetical protein	0.0%	0.0%	11.9%
P423_RS00690	S115 (syn)	polyamine aminopropyltransferase	98.3%	97.1%	97.3%
rnhA	R27H	ribonuclease HI	98.1%	100.0%	100.0%
lacI	E297G	DNA-binding transcriptional repressor LacI	98.2%	99.3%	97.6%
P423_RS02355	V386 (syn)	ammonium transporter	98.1%	98.2%	93.1%
P423_RS08810	I75S	DUF1161 domain-containing protein	98.4%	98.3%	100.0%
P423_RS12290	P44 (syn)	tagatose bisphosphate family class II aldolase	97.8%	97.1%	100.0%
P423_RS12705	N282K	GTP-binding protein	98.0%	98.8%	96.0%
lrhA	F189L	transcriptional regulator LrhA	97.1%	99.2%	98.8%
bamB	Q41K	outer membrane protein assembly factor BamB	98.2%	99.1%	100.0%
hycE	D168Y	hydrogenase large subunit	98.6%	97.4%	98.6%
glgC	L176 (syn)	transcriptional regulator GlcC	98.6%	98.2%	99.0%
P423_RS17520	A157D	YtfJ family protein	98.5%	100.0%	100.0%

gpr	G31S	glyceraldehyde 3-phosphate reductase	96.8%	97.4%	98.5%
emrD	P74 (syn)	multidrug transporter EmrD	98.4%	99.2%	98.7%
dgoD	L247 (syn)	D-galactonate dehydratase	96.6%	98.5%	96.4%
P423_RS04280	Upstream	glutathione S-transferase	100.0%	96.8%	98.6%
P423_RS12590	Upstream	DUF418 family protein	100.0%	98.3%	97.4%
P423_RS13770	Upstream	exoaminopeptidase	99.2%	98.6%	98.1%
P423_RS13980	L69V	NAD(P)-dependent oxidoreductase	99.1%	97.9%	98.5%
P423_RS15585	Upstream	MBL fold metallo-hydrolase	99.1%	100.0%	97.3%
P423_RS21485	G264 (syn)	phosphate ABC transporter permease	99.4%	100.0%	96.7%
gppA	E344K	guanosine-5'-triphosphate 3'-diphosphate diphosphatase	100.0%	100.0%	97.8%
P423_RS22225	Upstream	hypothetical protein	100.0%	100.0%	96.3%
gspG	L38 (syn)	type II secretion system protein GspG	100.0%	98.9%	92.9%
P423_RS25595	T4A	DUF2501 domain-containing protein	100.0%	98.2%	100.0%
Gene	Polymorphism	Gene product	Frequency in P9 samples		
			First fecal	Urinary	Second fecal
caiA	R352 (syn)	crotonobetainyl-CoA dehydrogenase caiA	0.7%	96.6%	0.0%
P423_RS01750	T57I	proton-motive-force-driven choline transporter	0.0%	98.8%	0.0%
P423_RS02305	V196A	peptidylprolyl isomerase	97.7%	98.9%	100.0%
pbp2	D312G	penicillin-binding protein 2	0.0%	91.5%	0.0%
fur	Upstream	transcriptional repressor FUR	0.0%	0.9%	9.6%

P423_RS04210	C58F	glycyl-radical enzyme activating protein	5.5%	5.7%	5.1%
ybjJ	M49I	MFS transporter	0.0%	94.2%	0.0%
yceB	Upstream	lipoprotein	1.9%	93.2%	0.0%
yciK	Upstream	NAD(P)-dependent oxidoreductase	0.8%	0.0%	12.8%
P423_RS05920	Upstream	virulence factor	0.0%	93.2%	0.0%
acnA	G500C	aconitate hydratase	0.7%	95.4%	1.8%
prr	R417G	aminobutyraldehyde dehydrogenase	0.0%	6.5%	0.0%
rnfD	T187(syn)	electron transport complex subunit D	0.0%	98.2%	0.0%
P423_RS09665	G236A	TVP38/TMEM64 family protein	2.1%	1.1%	7.6%
copD	T286S	copper homeostasis membrane protein	7.8%	0.6%	9.4%
araH	K52T	arabinose ABC transporter permease	94.2%	2.2%	98.3%
fliH	Q152*	flagellar assembly protein FliH	0.0%	92.0%	0.8%
yegW	Upstream	GntR family transcriptional regulator	0.0%	100.0%	0.0%
P423_RS12415	Upstream	hypothetical protein	10.2%	7.0%	3.3%
cirA	P304L	catecholate siderophore receptor CirA	4.7%	0.0%	17.8%
psuG	P202S	pseudouridine-5'-phosphate glycosidase psuG	0.0%	93.7%	1.0%
P423_RS14965	Upstream	peptidase S14	19.0%	0.0%	17.1%
fhlA	D449N	formate hydrogenlyase transcriptional activator FhlA	94.3%	1.9%	89.4%
barA	S106*	two-component sensor histidine kinase BarA	0%	98.0%	0.6%
uxaA	Upstream	altronate dehydratase	0.0%	93.2%	0.0%

garL	P81 (syn)	5-keto-4-deoxy-D-glucarate aldolase garL	0.6%	96.0%	0.0%
mtr	T356S	tryptophan permease	0.0%	97.5%	0.0%
lptA	L144 (syn)	lipopolysaccharide ABC transporter substrate-binding protein LptA	0.5%	98.8%	0.0%
P423_RS19030	T302 (syn)	multidrug efflux RND transporter permease subunit	98.4%	99.4%	100.0%
gph	P143S	phosphoglycolate phosphatase gph	0.5%	1.0%	5.8%
yhiD	Upstream	MgtC/SapB family protein	7.3%	0.0%	10.0%
bcsA	D285A	UDP-forming cellulose synthase catalytic subunit bcsA	1.9%	96.7%	0.0%
dppF	I307V	dipeptide ABC transporter ATP-binding protein	0.0%	5.8%	0.0%
torZ	Q681R	molybdopterin guanine dinucleotide-containing S/N-oxide reductase	7.1%	5.0%	1.8%
torZ	G406S	molybdopterin guanine dinucleotide-containing S/N-oxide reductase	0.0%	92.7%	0.7%
waaH	A316P	glycosyltransferase	5.5%	50.0%	12.2%
tdh	V154M	L-threonine 3-dehydrogenase tdh	0.0%	0.0%	12.2%
waaZ	L116S	3-deoxy-D-manno-oct-2-ulosonate III transferase WaaZ	0.6%	97.9%	0.0%
gppA	E339K	guanosine-5'-triphosphate 3'-diphosphate diphosphatase gppA	1.1%	97.0%	0.0%
P423_RS22030	C58F	cytosine permease	0.0%	95.9%	0.0%
cadC	G464R	transcriptional regulator CadC	0.8%	0.0%	20.7%
aspA	L12(syn)	aspartate ammonia-lyase aspA	99.5%	1.3%	99.4%

frdD	V111D	fumarate reductase subunit D	97.4%	0%	99.5%
frdD	M87I	fumarate reductase subunit D	0.0%	95.2%	0.0%
P423_RS25615	L237R	membrane protein	5.5%	4.6%	0.8%
yjjJ	Q138*	type II toxin-antitoxin system HipA family toxin	0.7%	96.7%	0.0%
P423_RS00085	G107C	hypothetical protein	97.8%	99.5%	100.0%
dapB	S64R	4-hydroxy-tetrahydrodipicolinate reductase	96.6%	98.2%	100.0%
lptD	Y63S	LPS-assembly protein LptD	98.8%	99.5%	99.4%
leuA	D14 (syn)	2-isopropylmalate synthase	98.4%	97.5%	100.0%
ponB	T617 (syn)	penicillin-binding protein 1B	98.6%	98.0%	98.6%
P423_RS01960	V469 (syn)	autotransporter outer membrane beta-barrel domain-containing protein	98.3%	98.6%	99.0%
lnt	L143 (syn)	apolipoprotein N-acyltransferase	98.1%	98.2%	100.0%
kdpB	S369*	K(+)-transporting ATPase subunit B	98.3%	98.0%	100.0%
glnP	G111S	glutamine ABC transporter permease	97.8%	97.3%	99.3%
gsiB	C401G	glutathione ABC transporter substrate-binding protein GsiB	98.7%	100.0%	100.0%
opgH	A35D	glucan biosynthesis protein H	96.7%	97.5%	97.7%
mdtG	T189A	multidrug resistance protein MdtG	96.5%	98.7%	99.1%
fhuE	Upstream	TonB-dependent siderophore receptor	97.7%	100.0%	99.4%
wcaI	A297D	colanic acid biosynthesis glycosyltransferase WcaI	97.7%	98.3%	99.1%
wcaD	H111Y	putative colanic acid polymerase WcaD	98.8%	98.9%	99.3%

P423_RS 12490	D95G	D-lactate dehydrogenase	98.5%	98.8%	98.7%
P423_RS 12590	Upstream	DUF418 family protein	98.8%	97.0%	99.0%
InaA	V137M	hypothetical protein	98.5%	97.8%	99.3%
lrhA	F189L	transcriptional regulator LrhA	97.5%	97.7%	97.9%
murQ	G185S	N-acetylmuramic acid 6-phosphate etherase	96.8%	98.1%	98.4%
P423_RS 14470	A204S	ABC transporter permease	97.5%	98.0%	98.2%
lepA	E229G	elongation factor 4	99.0%	99.2%	100.0%
nrdH	Upstream	glutaredoxin-like protein NrdH	39.5%	0.0%	27.9%
P423_RS 15585	Upstream	MBL fold metallo-hydrolase	97.7%	96.7%	100.0%
P423_RS 15730	A107T	MFS transporter	98.9%	97.9%	100.0%
P423_RS 17520	A157D	YtfJ family protein	98.5%	97.9%	98.5%
P423_RS 17755	I8 (syn)	TIGR00156 family protein	98.2%	99.4%	100.0%
dsbL	A165 (syn)	thiol:disulfide interchange protein DsbL	97.7%	98.9%	99.3%
mlaE	S8L	phospholipid ABC transporter permease	97.7%	99.0%	100.0%
nikA	S451P	nickel ABC transporter nickel/metallophore periplasmic binding protein	98.4%	99.5%	100.0%
P423_RS 20130	P433L	PTS galactitol transporter subunit IIC	98.2%	97.7%	100.0%
P423_RS 20935	Upstream	glycosyltransferase family 1 protein	98.1%	98.9%	100.0%
P423_RS 21460	L320 (syn)	PTS beta-glucoside transporter subunit IIABC	97.4%	98.4%	100.0%
rep	A498 (syn)	DNA helicase Rep	97.9%	97.2%	97.9%
Gene	Polymorphism	Gene product	Frequency in P10 samples		
			First fecal	Urinary	Second fecal

P423_RS01520	D81V	hypothetical protein	1.3%	1.1%	98.6%
nagC	A341S	N-acetylglucosamine repressor nagC	50.0%	87.3%	100.0%
arcA	W154C	aerobic respiration control protein ArcA	0.0%	0.0%	55.6%
hipB	W26L	antitoxin HipB	0.0%	0.0%	92.0%
yedA	L145*	amino acid exporter for phenylalanine-2C threonine	0.0%	8.2%	0.0%
entE	I489N	2,2C3-dihydroxybenzoate-AMP ligase	21.2%	89.8%	97.2%
P423_RS12995	V270D	DUF2300 domain-containing protein	38.5%	8.9%	0.0%
xanQ	F94I	xanthine permease XanQ	43.3%	92.3%	100.0%
rpiA	D74 (syn)	ribose-5-phosphate isomerase	0.0%	13.2%	0.0%
atpD	L389 (syn)	ATP synthase subunit beta	0.0%	21.3%	0.0%
glnG	D61N	nitrogen regulation protein NR(I)	0.0%	0.0%	97.0%
frdB	F236Y	fumarate reductase iron-sulfur subunit	40.8%	7.4%	0.0%
dnaT	Upstream	primosomal protein 1	0.0%	16.1%	95.7%
lacI	E297G	DNA-binding transcriptional repressor LacI	96.8%	100.0%	100.0%
lrhA	F189L	transcriptional regulator LrhA	98.2%	100.0%	99.0%
P423_RS22560	L41I	DUF805 domain-containing protein	96.7%	100.0%	100.0%
P423_RS23080	L455I	Na/Pi cotransporter family protein	97.8%	97.0%	100.0%
P423_RS23340	Upstream	ADP-forming succinate--CoA ligase subunit beta	96.8%	97.8%	96.1%
P423_RS08855	Upstream	LysR family transcriptional regulator	100.0%	97.5%	97.9%
uidC	Upstream	glucuronide uptake porin UidC	100.0%	98.0%	100.0%

wzzB	N171I	LPS O-antigen chain length determinant protein WzzB	100.0%	97.9%	98.2%
P423_RS21980	E243D	lysophospholipase	100.0%	98.0%	99.2%
Gene	Polymorphism	Gene product	Frequency in P12 samples		
			First fecal	Urinary	Second fecal
P423_RS00110	A366E	glycoside hydrolase family 31 protein	13.6%	10.3%	9.1%
mdoH	P98S	glucan biosynthesis protein H mdoH	0.7%	0%	7.7%
entE	V307F	salicylate synthase	96%	92.1%	97.3%
nac	Upstream	LysR family transcriptional regulator nac	0.7%	0%	6.8%
cysZ	P168L	sulfate transporter CysZ	0.7%	24.2%	0.6%
tdcD	Upstream	propionate kinase	21.1%	31.3%	7.5%
ulaD	A93T	3-dehydro-L-gulonate-6-phosphate decarboxylase UlaD	0%	0%	6.9%
bsmA	L28R	lipoprotein BsmA	0.7%	0.9%	30.8%
lafA	Q56K	lateral flagellin LafA	96.3%	100.0%	100.0%
P423_RS01960	V469 (syn)	autotransporter outer membrane beta-barrel domain-containing protein	97.7%	99.1%	99.6%
P423_RS01980	V293A	peptide transporter	97.9%	100.0%	99.3%
P423_RS02305	V196A	peptidylprolyl isomerase	98.5%	97.7%	100.0%
P423_RS02355	V386 (syn)	ammonium transporter	96.8%	93.8%	97.7%
P423_RS02430	A134T	primosomal replication protein N	97.1%	98.0%	95.6%
P423_RS02660	A241 (syn)	allantoin permease	97.6%	97.8%	97.4%
P423_RS03305	L143 (syn)	apolipoprotein N-acyltransferase	98.3%	100.0%	98.2%
P423_RS03575	R3*	hypothetical protein	98.2%	99.0%	98.8%
P423_RS08330	G260R	acetyltransferase	97.7%	97.5%	97.5%
uidC	Upstream	glucuronide uptake porin UidC	98.8%	99.1%	99.4%

astD	D416E	succinylglutamate-semialdehyde dehydrogenase	97.3%	100.0%	98.7%
P423_RS09775	T98A	DeoR/GlpR transcriptional regulator	95.9%	98.0%	98.5%
wcaD	H111Y	putative colanic acid polymerase WcaD	97.8%	98.9%	98.4%
cirA	G11R	catecholate siderophore receptor CirA	97.0%	92.9%	98.0%
lrhA	F189L	transcriptional regulator LrhA	98.6%	97.6%	97.8%
P423_RS13770	Upstream	exoaminopeptidase	98.7%	97.7%	99.5%
P423_RS14420	T182 (syn)	IscS subfamily cysteine desulfurase	97.7%	97.9%	99.3%
P423_RS14595	E229G	elongation factor 4	98.3%	98.0%	99.3%
hycE	D168Y	hydrogenase large subunit	98.3%	96.9%	98.0%
mutS	L796 (syn)	DNA mismatch repair protein MutS	97.7%	100.0%	97.6%
P423_RS15585	Upstream	MBL fold metallo-hydrolase	98.3%	98.6%	99.5%
fucI	A476 (syn)	L-fucose isomerase	97.1%	100.0%	99.0%
P423_RS17440	G396 (syn)	DUF4092 domain-containing protein	94.4%	97.4%	97.4%
P423_RS17520	A157D	YtfJ family protein	98.7%	96.5%	98.4%
dsbL	A165 (syn)	thiol:disulfide interchange protein DsbL	97.7%	100.0%	99.4%
P423_RS19030	T302 (syn)	multidrug efflux RND transporter permease subunit	98.4%	98.6%	99.4%
uhpB	A122 (syn)	signal transduction histidine-protein kinase/phosphatase UhpB	97.6%	100.0%	99.2%
P423_RS22225	Upstream	hypothetical protein	91.8%	90.0%	89.7%
P423_RS22405	L39 (syn)	hypothetical protein	98.2%	97.6%	99.5%
P423_RS23035	Upstream	hypothetical protein	98.4%	98.7%	96.6%

P423_RS23340	Upstream	ADP-forming succinate--CoA ligase subunit beta	96.9%	94.1%	99.3%
P423_RS24285	H187L	NAD(P)-dependent oxidoreductase	95.7%	98.1%	96.9%
P423_RS09805	Upstream	alcohol dehydrogenase	100.0%	100.0%	97.6%
P423_RS12590	Upstream	DUF418 family protein	100.0%	97.6%	98.4%
nuoG	C573R	NADH-quinone oxidoreductase subunit NuoG	100.0%	96.7%	97.6%
lptG	T92 (syn)	LPS export ABC transporter permease LptG	100.0%	100.0%	98.4%
mreD	S6 (syn)	rod shape-determining protein MreD	100.0%	98.6%	98.7%
rmuC	S395N	DNA recombination protein RmuC	100.0%	100.0%	96.7%
P423_RS23685	I289V	hypothetical protein	100.0%	100.0%	98.4%
lacI	E297G	DNA-binding transcriptional repressor LacI	99.2%	97.8%	100.0%
qseE	V33A	two component system sensor histidine kinase QseE/GlrK	100.0%	97.3%	100.0%
P423_RS20935	Upstream	glycosyltransferase family 1 protein	99.4%	97.8%	100.0%

REFERENCES

1. Jandhyala SM, Talukdar R, Subramanyam C, Vuyyuru H, Sasikala M, Reddy DN. 2015. Role of the normal gut microbiota. World J Gastroenterol 21:8836–8847.

2. Hooper L V., Wong MH, Thelin A, Hansson L, Falk PG, Gordon JI. 2001. Molecular analysis of commensal host-microbial relationships in the intestine. Science (80-).

3. Macfarlane S, Macfarlane GT. 2003. Regulation of short-chain fatty acid production. Proc Nutr Soc.

4. Cash HL, Whitham C V., Behrendt CL, Hooper L V. 2006. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. Science (80-).

5. Hooper L V., Stappenbeck TS, Hong C V., Gordon JI. 2003. Angiogenins: A new class of microbicidal proteins involved in innate immunity. Nat Immunol.

6. Cresci GA, Bawden E. 2015. Gut microbiome: What we do and don't know. Nutr Clin Pract. SAGE Publications Inc.

7. DiGiulio DB. 2012. Diversity of microbes in amniotic fluid. Semin Fetal Neonatal Med.

8. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. 2014. The placenta harbors a unique microbiome. Sci Transl Med.

9. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A.

10. De Palma G, Collins SM, Bercik P, Verdu EF. 2014. The microbiota-gut-brain axis in gastrointestinal disorders: Stressed bugs, stressed brain or both? J Physiol.

11. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris

M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature.

12. Lloyd-Price J, Abu-Ali G, Huttenhower C. 2016. The healthy human microbiome. Genome Med. BioMed Central Ltd.

13. Parfrey LW, Walters WA, Lauber CL, Clemente JC, Berg-Lyons D, Teiling C, Kodira C, Mohiuddin M, Brunelle J, Driscoll M, Fierer N, Gilbert JA, Knight R. 2014. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. Front Microbiol 5.

14. Neis EPJG, Dejong CHC, Rensen SS. 2015. The role of microbial amino acid metabolism in host metabolism. Nutrients.

15. Gominak SC. 2016. Vitamin D deficiency changes the intestinal microbiome reducing B vitamin production in the gut. The resulting lack of pantothenic acid adversely affects the immune system, producing a “pro-inflammatory” state associated with atherosclerosis and autoimmun. Med Hypotheses.

16. Magwira CA, Kullin B, Lewandowski S, Rodgers A, Reid SJ, Abratt VR. 2012. Diversity of faecal oxalate-degrading bacteria in black and white South African study groups: Insights into understanding the rarity of urolithiasis in the black group. J Appl Microbiol.

17. Cardona F, Andrés-Lacueva C, Tulipani S, Tinahones FJ, Queipo-Ortuño MI. 2013. Benefits of polyphenols on gut microbiota and implications in human health. J Nutr Biochem.

18. Shi N, Li N, Duan X, Niu H. 2017. Interaction between the gut microbiome and mucosal immune system. Mil Med Res. BioMed Central Ltd.

19. Cani PD, Possemiers S, Van De Wiele T, Guiot Y, Everard A, Rottier O, Geurts L, Naslain D, Neyrinck A, Lambert DM, Muccioli GG, Delzenne NM. 2009. Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. Gut.

20. Yan F, Cao H, Cover TL, Washington MK, Shi Y, Liu LS, Chaturvedi R, Peek RM, Wilson KT, Polk DB. 2011. Colon-specific delivery of a probiotic-derived soluble protein ameliorates intestinal inflammation in mice through an EGFR-dependent mechanism. J Clin Invest.

21. Conway T, Cohen PS. 2015. Commensal and Pathogenic Escherichia coli Metabolism in the Gut. Microbiol Spectr 3.

22. Sassone-Corsi M, Nuccio SP, Liu H, Hernandez D, Vu CT, Takahashi AA, Edwards RA, Raffatellu M. 2016. Microcins mediate competition among Enterobacteriaceae in the inflamed gut. Nature.

23. Vahjen W, Cuisiniere T, Zentek J. 2017. Protective effects of indigenous Escherichia coli against a pathogenic E. coli challenge strain in pigs. Benef Microbes 8:779–783.

24. Dzidic M, Boix-Amorós A, Selma-Royo M, Mira A, Collado M. 2018. Gut Microbiota and Mucosal Immunity in the Neonate. Med Sci.

25. Sewell AK, Han M, Qi B. 2018. An unexpected benefit from E. Coli: How enterobactin benefits host health. Microb Cell.

26. Breton J, Tennoune N, Lucas N, Francois M, Legrand R, Jacquemot J, Goichon A, Guérin C, Peltier J, Pestel-Caron M, Chan P, Vaudry D, Do Rego JC, Liénard F, Pénicaud L, Fioramonti X, Ebenezer IS, Hökfelt T, Déchelotte P, Fetissov SO. 2016. Gut commensal E. coli proteins activate host satiety pathways following nutrient-induced

bacterial growth. Cell Metab.

27. Secher T, Brehin C, Oswald E. 2016. Early settlers: Which E. coli strains do you not want at birth? Am J Physiol - Gastrointest Liver Physiol.

28. Mirsepasi-Lauridsen HC, Vallance BA, Krogfelt KA, Petersen AM. 2019. Escherichia coli pathobionts associated with inflammatory bowel disease. Clin Microbiol Rev.

29. Cools P. 2017. The role of Escherichia coli in reproductive health: state of the art. Res Microbiol.

30. Rang CU, Licht TR, Midtvedt T, Conway PL, Chao L, Krogfelt KA, Cohen PS, Molin S. 1999. Estimation of growth rates of Escherichia coli BJ4 in streptomycin- treated and previously germfree mice by in situ rRNA hybridization. Clin Diagn Lab Immunol.

31. Miranda RL, Conway T, Leatham MP, Chang DE, Norris WE, Allen JH, Stevenson SJ, Laux DC, Cohen PS. 2004. Glycolytic and Gluconeogenic Growth of Escherichia coli O157:H7 (EDL933) and E. coli K-12 (MG1655) in the Mouse Intestine. Infect Immun.

32. Barroso-Batista J, Sousa A, Lourenço M, Bergman ML, Sobral D, Demengeot J, Xavier KB, Gordo I. 2014. The First Steps of Adaptation of Escherichia coli to the Gut Are Dominated by Soft Sweeps. PLoS Genet.

33. Lourenço M, Ramiro RS, Güleresi D, Barroso-Batista J, Xavier KB, Gordo I, Sousa A. 2016. A Mutational Hotspot and Strong Selection Contribute to the Order of Mutations Selected for during Escherichia coli Adaptation to the Gut. PLoS Genet.

34. Giraud A, Arous S, De Paepe M, Gaboriau-Routhiau V, Bambou JC, Rakotobe S, Lindner AB, Taddei F, Cerf-Bensussan N. 2008. Dissecting the genetic components of adaptation of Escherichia coli to the mouse gut. PLoS Genet.

35. Lescat M, Launay A, Ghalayini M, Magnan M, Glodt J, Pintard C, Dion S, Denamur E,

Tenaillon O. 2017. Using long-term experimental evolution to uncover the patterns and determinants of molecular evolution of an Escherichia coli natural isolate in the streptomycin-treated mouse gut. Mol Ecol.

36. Gordo I, Demengeot J, Xavier K. 2014. Escherichia coli adaptation to the gut environment: A constant fight for survival. Future Microbiol.

37. Biggest Threats and Data | Antibiotic/Antimicrobial Resistance | CDC.

38. Bevan ER, McNally A, Thomas CM, Piddock LJV, Hawkey PM. 2018. Acquisition and loss of CTX-M-producing and non-producing escherichia coli in the fecal microbiome of travelers to South Asia. MBio.

39. Anderson MA, Whitlock JE, Harwood VJ. 2006. Diversity and distribution of Escherichia coli genotypes and antibiotic resistance phenotypes in feces of humans, cattle, and horses. Appl Environ Microbiol.

40. Bailey JK, Pinyon JL, Anantham S, Hall RM. 2010. Commensal Escherichia coli of healthy humans: A reservoir for antibiotic-resistance determinants. J Med Microbiol.

41. Manges AR, Smith SP, Lau BJ, Nuval CJ, Eisenberg JNS, Dietrich PS, Riley LW. 2007. Retail meat consumption and the acquisition of antimicrobial resistant Escherichia coli causing urinary tract infections: A case-control study. Foodborne Pathog Dis.

42. Gurnee EA, Ndao IM, Johnson JR, Johnston BD, Gonzalez MD, Burnham CAD, Hall-Moore CM, McGhee JE, Mellmann A, Warner BB, Tarr PI. 2015. Gut colonization of healthy children and their mothers with pathogenic ciprofloxacin-resistant Escherichia coli. J Infect Dis.

43. Blake DP, Hillman K, Fenlon DR, Low JC. 2003. Transfer of antibiotic resistance between commensal and pathogenic members of the Enterobacteriaceae under ileal conditions. J

Appl Microbiol.

44. Mann R, Mediati DG, Duggin IG, Harry EJ, Bottomley AL. 2017. Metabolic adaptations of Uropathogenic E. coli in the urinary tract. Front Cell Infect Microbiol 7.

45. Subashchandrabose S, Mobley HLT. 2015. Virulence and Fitness Determinants of Uropathogenic Escherichia coli. Microbiol Spectr.

46. Brumbaugh AR, Mobley HL. 2012. Preventing urinary tract infection: Progress toward an effective Escherichia coli vaccine. Expert Rev Vaccines.

47. Moreno E, Johnson JR, Pérez T, Prats G, Kuskowski MA, Andreu A. 2009. Structure and urovirulence characteristics of the fecal Escherichia coli population among healthy women. Microbes Infect.

48. Tchesnokova VL, Rechkina E, Chan D, Haile HG, Larson L, Ferrier K, Schroeder DW, Solyanik T, Shibuya S, Hansen K, Ralston JD, Riddell K, Scholes D, Sokurenko E V. 2019. Pandemic Uropathogenic Fluoroquinolone-resistant Escherichia coli Have Enhanced Ability to Persist in the Gut and Cause Bacteriuria in Healthy Women. Clin Infect Dis.

49. Nielsen KL, Dynesen P, Larsen P, Frimodt-Møller N. 2014. Faecal Escherichia coli from patients with E. coli urinary tract infection and healthy controls who have never had a urinary tract infection. J Med Microbiol.

50. Sasabe J, Suzuki M. 2018. Emerging role of D-Amino acid metabolism in the innate defense. Front Microbiol.

51. Hancock V, Seshasayee AS, Ussery DW, Luscombe NM, Klemm P. 2008. Transcriptomics and adaptive genomics of the asymptomatic bacteriuria Escherichia coli strain 83972. Mol Genet Genomics.

52. Nielsen KL, Stegger M, Godfrey PA, Feldgarden M, Andersen PS, Frimodt-Møller N. 2016. Adaptation of Escherichia coli traversing from the faecal environment to the urinary tract. Int J Med Microbiol.

53. Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, Riddell K, Rogers P, Qin X, Butler-Wu S, Price LB, Aziz M, Nicolas-Chanoine MH, Debroy C, Robicsek A, Hansen G, Urban C, Platell J, Trott DJ, Zhanel G, Weissman SJ, Cookson BT, Fang FC, Limaye AP, Scholes D, Chattopadhyay S, Hooper DC, Sokurenko E V. 2013. Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. J Infect Dis.

54. Mathers AJ, Peirano G, Pitout JDD. 2015. Escherichia coli ST131: The Quintessential Example of anInternational Multiresistant High-Risk Clone. Adv Appl Microbiol.

55. Burgess MJ, Johnson JR, Porter SB, Johnston B, Clabots C, Lahr BD, Uhl JR, Banerjee R. 2015. Long-term care facilities are reservoirs for antimicrobial-resistant sequence type 131 Escherichia coli. Open Forum Infect Dis.

56. Johnson JR, Porter S, Thuras P, Castanheira M. 2017. The pandemic H30 subclone of sequence type 131 (ST131) as the leading cause of multidrug-resistant Escherichia coli infections in the United States (2011-2012). Open Forum Infect Dis.

57. Tchesnokova V, Riddell K, Scholes D, Johnson JR, Sokurenko E V. 2019. The uropathogenic Escherichia coli subclone sequence type 131-H30 is responsible for most antibiotic prescription errors at an urgent care clinic. Clin Infect Dis.

58. Johnson JR, Thuras P, Johnston BD, Weissman SJ, Limaye AP, Riddell K, Scholes D, Tchesnokova V, Sokurenko E. 2016. The Pandemic H30 Subclone of Escherichia coli Sequence Type 131 Is Associated with Persistent Infections and Adverse Outcomes Independent from Its Multidrug Resistance and Associations with Compromised Hosts.

Clin Infect Dis.

59. Kisiela DI, Radey M, Paul S, Porter S, Polukhina K, Tchesnokova V, Shevchenko S, Chan D, Aziz M, Johnson TJ, Price LB, Johnson JR, Sokurenko E V. 2017. Inactivation of transcriptional regulators during withinhousehold evolution of Escherichia coli. J Bacteriol.

60. Heintz-Buschart A, Wilmes P. 2018. Human Gut Microbiome: Function Matters. Trends Microbiol.

61. Metwaly A, Haller D. 2019. Strain-Level Diversity in the Gut: The P. copri Case. Cell Host Microbe.

62. Galardini M, Koumoutsi A, Herrera-Dominguez L, Varela JAC, Telzerow A, Wagih O, Wartel M, Clermont O, Denamur E, Typas A, Beltrao P. 2017. Phenotype inference in an Escherichia coli strain panel. Elife.

63. Gupta M, Didwal G, Bansal S, Kaushal K, Batra N, Gautam V, Ray P. 2019. Antibiotic-resistant Enterobacteriaceae in healthy gut flora: A report from north Indian semiurban community. Indian J Med Res.

64. Gorrie CL, Mirceta M, Wick RR, Judd LM, Wyres KL, Thomson NR, Strugnell RA, Pratt NF, Garlick JS, Watson KM, Hunter PC, McGloughlin SA, Spelman DW, Jenney AWJ, Holt KE. 2018. Antimicrobial-resistant klebsiella pneumoniae carriage and infection in specialized geriatric care wards linked to acquisition in the referring hospital. Clin Infect Dis.

65. Robin F, Beyrouthy R, Bonacorsi S, Aissa N, Bret L, Brieu N, Cattoir V, Chapuis A, Chardon H, Degand N, Doucet-Populaire F, Dubois V, Fortineau N, Grillon A, Lanotte P, Leyssene D, Patry I, Podglajen I, Recule C, Ros A, Colomb-Cotinat M, Ponties V, Ploy MC, Bonnet R. 2017. Inventory of extended-spectrum-β-lactamase-producing

Enterobacteriaceae in France as assessed by a multicenter study. Antimicrob Agents Chemother.

66. Johnson JR, Johnston B, Clabots C, Kuskowski MA, Castanheira M. 2010. Escherichia coli Sequence Type ST131 as the Major Cause of Serious Multidrug-Resistant E. coli Infections in the United States . Clin Infect Dis.

67. Fischer M, Strauch B, Renard BY. 2017. Abundance estimation and differential testing on strain level in metagenomics dataBioinformatics.

68. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res.

69. Weissman SJ, Johnson JR, Tchesnokova V, Billig M, Dykhuizen D, Riddell K, Rogers P, Qin X, Butler-Wu S, Cookson BT, Fang FC, Scholes D, Chattopadhyay S, Sokurenko E. 2012. High-resolution two-locus clonal typing of extraintestinal pathogenic Escherichia coli. Appl Environ Microbiol.

70. Scholz M, Ward D V., Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods.

71. Richter TKS, Hazen TH, Lam D, Coles CL, Seidman JC, You Y, Silbergeld EK, Fraser CM, Rasko DA. 2018. Temporal Variability of Escherichia coli Diversity in the Gastrointestinal Tracts of Tanzanian Children with and without Exposure to Antibiotics . mSphere.

72. Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, Tenaillon O. 2007. Extraintestinal virulence is a coincidental by-product of commensalism in b2 phylogenetic

group Escherichia coli strains. Mol Biol Evol.

73. Smati M, Clermont O, Le Gal F, Schichmanoff O, Jauréguy F, Eddi A, Denamur E, Picard B. 2013. Real-time PCR for quantitative analysis of human commensal Escherichia coli populations reveals a high frequency of subdominant phylogroups. Appl Environ Microbiol.

74. Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I. 2010. Pathogenicity-associated islands in extraintestinal pathogenic Escherichia coli are fitness elements involved in intestinal colonization. J Bacteriol.

75. Moreno E, Andreu A, Pérez T, Sabaté M, Johnson JR, Prats G. 2006. Relationship between Escherichia coli strains causing urinary tract infection in women and the dominant faecal flora of the same hosts. Epidemiol Infect.

76. Glover M, Moreira CG, Sperandio V, Zimmern P. 2014. Recurrent urinary tract infections in healthy and nonpregnant women. Urol Sci. Elsevier B.V.

77. Mabeck CE. 1972. Treatment of uncomplicated urinary tract infection in non-pregnant women. Postgrad Med J 48:69–75.

78. Birder LA. 2014. Urinary bladder, cystitis and nerve/urothelial interactions. Auton Neurosci Basic Clin 182:89–94.

79. Martin GS, Mannino DM, Eaton S, Moss M. 2003. The epidemiology of sepsis in the United States from 1979 through 2000. N Engl J Med 348:1546–1554.

80. Levy MM, Artigas A, Phillips GS, Rhodes A, Beale R, Osborn T, Vincent JL, Townsend S, Lemeshow S, Dellinger RP. 2012. Outcomes of the Surviving Sepsis Campaign in intensive care units in the USA and Europe: A prospective cohort study. Lancet Infect Dis 12:919–924.

81. Mazzariol A, Bazaj A, Cornaglia G. 2017. Multi-drug-resistant Gram-negative bacteria causing urinary tract infections: a review Antimicrobial Original Research Paper Multi-drug-resistant Gram-negative bacteria causing urinary tract infections: a review. J Chemother 29.

82. Zdziarski J, Brzuszkiewicz E, Wullt B, Liesegang H, Biran D, Voigt B, Grönberg-Hernandez J, Ragnarsdottir B, Hecker M, Ron EZ, Daniel R, Gottschalk G, Hacker J, Svanborg C, Dobrindt U. 2010. Host imprints on bacterial genomes-rapid, divergent evolution in individual patients. PLoS Pathog.

83. Alteri CJ, Smith SN, Mobley HLT. 2009. Fitness of Escherichia coliduring urinary tract infection requires gluconeogenesis and the TCA cycle. PLoS Pathog 5.

84. Palaniyandi S, Mitra A, Herren CD, Lockatell CV, Johnson DE, Zhu X, Mukhopadhyay S. 2012. BarA-UvrY two-component system regulates virulence of uropathogenic E. coli CFT073. PLoS One 7.

85. Reitzer L, Zimmern P. 2019. Rapid Growth and Metabolism of Uropathogenic Escherichia coli in Relation to Urine Composition. Clin Microbiol Rev. NLM (Medline).

86. Ghalayini M, Launay A, Bridier-Nahmias A, Clermont O, Denamur E, Lescat M, Tenaillona O. 2018. Evolution of a dominant natural isolate of Escherichia coli in the human gut over the course of a year suggests a neutral evolution with reduced effective population size. Appl Environ Microbiol 84.

87. Simmering JE, Tang F, Cavanaugh JE, Polgreen LA, Polgreen PM. 2017. The increase in hospitalizations for urinary tract infections and the associated costs in the United States, 1998-2011. Open Forum Infect Dis.

88. Dobrindt U, Wullt B, Svanborg C. 2016. Asymtomatic bacteriuria as a model to study the

coevolution of hosts and Bacteria. Pathogens.

89. Nielsen KL, Stegger M, Godfrey PA, Feldgarden M, Andersen PS, Frimodt-Møller N. 2016. Adaptation of Escherichia coli traversing from the faecal environment to the urinary tract. Int J Med Microbiol 306:595–603.

90. Pernestig AK, Georgellis D, Romeo T, Suzuki K, Tomenius H, Normark S, Melefors Ö. 2003. The Escherichia coli BarA-UvrY two-component system is needed for efficient switching between glycolytic and gluconeogenic carbon sources. J Bacteriol.

91. GitHub - FelixKrueger/TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data.

92. Clausen PTLC, Aarestrup FM, Lund O. 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics.

93. Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics.

94. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

95. Tchesnokova V, Avagyan H, Billig M, Chattopadhyay S, Aprikian P, Chan D, Pseunova J, Rechkina E, Riddell K, Scholes D, Fang FC, Johnson JR, Sokurenko E V. 2016. A Novel 7-single nucleotide polymorphism-based clonotyping test allows rapid prediction of antimicrobial susceptibility of extraintestinal Escherichia coli directly from urine specimens. Open Forum Infect Dis.

96. Stata: Software for Statistics and Data Science.

97. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis

Version 7.0 for Bigger Datasets. Mol Biol Evol.

98. seaborn: statistical data visualization — seaborn 0.9.0 documentation.

99. Roer L, Tchesnokova V, Allesoe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen MCF, Lund O, Hansen F, Hammerum AM, Sokurenko E, Hasman H. 2017. Development of a web tool for Escherichia coli subtyping based on fimh alleles. J Clin Microbiol.

100. Gardner SN, Hall BG. 2013. When whole-genome alignments just won't work: KSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PLoS One 8.

101. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

102. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–5.

103. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

104. BLAST+ features - BLAST® Command Line Applications User Manual - NCBI Bookshelf.

105. Kim Y, Oh T, Nam YS, Cho SY, Lee HJ. 2017. Prevalence of ST131 and ST1193 among bloodstream isolates of Escherichia coli not susceptible to ciprofloxacin in a tertiary care university hospital in korea, 2013-2014. Clin Lab.

106. Johnson JR, Johnston BD, Porter SB, Clabots C, Bender TL, Thuras P, Trott DJ, Cobbold

R, Mollinger J, Ferrieri P, Drawz S, Banerjee R. 2019. Rapid emergence, subsidence, and molecular detection of Escherichia coli sequence type 1193-fimH64, a new disseminated multidrug-resistant commensal and extraintestinal pathogen. J Clin Microbiol.