Unlocking the Secrets of Urine Survival

An Exploration of Bacterial Genetic Adaptations

Louise Spekking (6201563)

December 12, 2023

Urinary tract infections (UTIs) are one of the most common bacterial infections and can be caused by a diverse range of bacterial species. Here we studied the relatedness and changes in gene presence and absence profiles of five species with uropathogenic potential, *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Enterococcus faecalis* and *Staphylococcus haemolyticus*, in UTI and healthy states. Relatedness was analysed using k-mer-based, 16s and core genome phylogenies, showing no indication of relatedness between these species in the urinary tract. To analyse the potential functional changes in bacteria isolated from the urinary tract, we constructed a pan-genome for each species, revealing gene clusters unique to urine-derived strains. These unique gene clusters are involved in pathways known to increase bacterial virulence in the urinary tract as well as genes attributed to antibiotic resistance. In addition to individual genes and metabolic pathways, we analysed potential functional changes in gene clusters predicted to encode metabolites that are non essential for bacterial growth. While no clusters were found to be unique to urine-derived strains, several were enriched for their presence in urine-derived strains and were predicted to facilitate competition between bacteria in the urinary tract. Preliminary analysis of the presence and absence of metabolites before and after bacterial growth revealed a potential for nutrient competition for amino acids. Together, these findings show unique adaptations of bacteria living within the urinary tract, with functions influencing virulence and antibiotic resistance.

Layman's summary

Urinary tract infections are common infections that are often caused by bacteria and can be treated with antibiotics. These bacteria are present not only when the urinary tract is infected but also in healthy individuals. Urine can be a difficult environment for bacteria to live in, as the conditions in urine can vary greatly between people but also

within the same person over time. In addition to this variation, many substances that are essential for bacterial growth are absent or in limited supply in urine, making urine a unique environment, thereby leading to the question if bacteria adapt their genome to these environmental circumstances. This study first looks into the question of whether this unique environment influences the relatedness of the bacterial species that can live in the urinary tract and can cause infection, making their genomes more similar, but finds no relatedness. This absence of relatedness does not have to indicate an absence of genetic adaptations to live in urine. Therefore, we also analysed if there are changes in presence and absences of genes and their functions. This was done using a pan-genome, which is an overview of all genes present in a set of members of one species for all different environments in this set. Here, several genes were found that could help these species cause disease and live in the environment of urine by making it easier for them to get nutrients that are scarce in urine, resisting antibiotics that are often used to treat urinary tract infections and adjusting to how the immune system of the host reacts. Bacteria in the urinary tract often do not live alone but form a community with other species where they can communicate and help or harm each other. One way bacteria within these communities can facilitate communication is by excreting secondary metabolites, small molecules not needed to survive but whose production can provide a competitive advantage when living in a community. Here, no predicted secondary metabolites were unique to the urine environment, but some were more present in bacteria isolated from urine than would be randomly expected. These are predicted to mediate competitive interactions between species, and for one species provided protection against the strong osmotic pressure present in urine. An additional method by which bacteria can compete within a community is through the consumption of nutrients. This was analysed using mass spectrometry analysis, measuring the presence of molecules in the medium before and after bacterial growth. From this analysis we identified the overlap of consumed and excreted molecules for each species. This showed that one amino acid, lysine, was consumed by all species studied here, which could indicate that these species compete for this nutrient when living in the urinary tract. In conclusion, this study shows that even though no evidence of relatedness between strains isolated from urine was found, bacteria in the urinary tract do adapt their genomes to the urine environment and compete with other members of their bacterial community.

Introduction

Urine was thought to be sterile up until a decade ago, however, new techniques have revealed hundreds of bacterial species to reside within the urinary tract[1–6]. Bacterial infections of the urinary tract, also known as urinary tract infections (UTIs), are one of the most common bacterial infections, with more than half of women and up to 12% of men experiencing a UTI in their lifetime and affecting approximately 150 million people worldwide[7,8]. UTIs are commonly treated with antibiotics, even in absence of true infections, which could increase antibiotic resistance in bacterial populations of the urinary tract[9,10].

A range of both gram-positive and gram-negative bacterial pathogens have been implicated in UTIs, such as *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Enterococcus faecalis* and *Staphylococcus haemolyticus*, which are studied here[11–18]. *E. coli* has been the most studied uropathogen as it is the cause of up to 80% of UTIs, in addition to being the model bacterium in molecular biology[19,20]. As polymicrobial infections are common in the elderly as well as in patients with underlying risk factors for UTIs, and uropathogens can inhabit the urinary tract in absence of infection, a study of a broader scope of possible uropathogens is warranted[20–22]. The urinary microbiome is hypothesised to originate from the gut, the change from an intestinal to a urinary environment requires quick adaptations, as environmental factors such as pH and nutrient availability vary greatly between the two environments[2,23]. Understanding the adaptations required for bacterial fitness and survival in the urinary tract is essential to our understanding of UTIs and the healthy urinary microbiome.

The aim of this study is to better understand the interspecies variation and commonalities of five potentially pathogenic species living in the urinary tract by studying the relatedness and functional changes of these microbes in relation to their host environment. Changes hypothesised to be present as urine is a harsh environment where several essential nutrients are limited and pH and chemical diversity can vary greatly from host to host as well as within one host over time[24,25]. Moreover, the urinary environment is iron-limited, moderately oxygenated and has high osmolarity, in addition to containing mostly amino acids and small peptides but low carbohydrate availability[23,26–32]. Due to these environmental features, we hypothesise that a selective pressure is exerted on the microbes living in the urinary tract, resulting in phylogenetic relatedness between the microbes isolated from the urinary environment as well as potential functional changes. In this study, five potentially uropathogenic species with roles in UTIs were studied[14]. Phylogenetic analysis revealed no signal of relatedness between urine-derived strains. To analyse potential functional changes, a pan-genome approach was taken, where gene clusters contributing to virulence and antimicrobial resistance (AMR) were found to be unique to strains isolated from urine. In addition to functional changes, a preliminary analysis of potential interactions showed a role for newly excreted metabolites and amino acid nutrient competition between species.

Results and Discussion

Dataset

After initial filtering as described in Methods section Dataset, the dataset for *S. haemolyticus* was deemed too small for analysis. It contained 17 genomes in total, with two of a urinary origin of isolation. To increase the dataset size and number of urine genomes therein, incomplete genomes for which Prokka annotated a minimum number of 2000 genes were added to the dataset to use in the current analysis. A second alteration to the initial filtering was applied for *K. pneumoniae*, where several long outliers were trimmed from the tree produced by the k-mer-based analysis, removing an additional

Table 1: Dataset sizes for all species after filtering on genome length, GC content, average nucleotide identity > 95% to reference, genome quality, known origin of isolation and CheckM genome completeness and contamination. Numbers of genomes with urine and UTI origins of isolation are a subset of total genome counts.

Species	Entries PATRIC	No. genomes after filtering	No. urine isolated genomes filtered set	No. UTI isolated genomes filtered set
S. haemolyticus	656	31	3	1
P. aeruginosa	8573	401	30	1
E. faecalis	2966	119	7	1
E. coli	45869	1642	142	20
K. pneumoniae	20027	793	110	3

19 genomes from the dataset (Figure S1, Table S1). No additional filtering steps were applied to the datasets of the other species. Final numbers of genomes are shown in Table 1.

Given the low number of urine and UTI genomes for each species, genomes with either annotation were combined and labelled as "urine" in further analysis. Moreover, the labelling as urine origin of isolation does not exclude the presence of UTI in these subjects. Given the low number of genomes, analysing the two groups separately might not yield meaningful results. Only for *E. coli* all metadata and associated papers, if available, were analysed to make a distinction between genomes isolated from healthy urine and UTIs, facilitating a preliminary analysis into the differences between healthy and UTI states.

No distinction was made between hosts, allowing for a comprehensive analysis of the urinary environment as a whole. The results of this study are therefore not specific to the human host and must be interpreted with caution in relation to specific hosts.

All pan-genomes are open

The pan-genome is considered to be the collection of all gene clusters or orthologous groups (OGs) found to be present in a set of genomes, here the filtered dataset for each species[33–35]. A second subsection of the pan-genome is the dispensable genome, divided in the OGs shared between two or more strains, but not all strains, and the OGs unique to one strain, named the accessory and unique genome respectively[33–35]. Pan-genomes were constructed with the filtered datasets for each species using Roary (Table 2).

The data from Table 2 shows that the number of strains included in the analysis has a great influence on the total number of OGs. This trend is not linear, which is expected as genome sizes and genetic diversities between species differ. Given the different number of strains for the species direct comparison of genetic diversity is difficult, as the lower number of genomes can indicate undersampling of the possible environments for

Table 2: Pan-genome results as constructed by Roary

Species	No. genomes	No. total OGs	No. core OGs	No. accessory OGs	No. unique OGs
S. haemolyticus	31	7522	1448 (19%)	2967 (39%)	3107 (41%)
P. aeruginosa	401	46275	1150 (3%)	29301 (63%)	15824 (34%)
E. faecalis	119	11044	1557 (14%)	6381 (58%)	3106 (28%)
E. coli	1642	104684	100 (0.1%)	67292 (64%)	37292 (36%)
K. pneumoniae	793	50735	869 (2%)	32991 (65%)	16875 (33%)

these species. Sub-sampling of the datasets could help in assessing the genetic diversity between species. In this study, we attempted to construct pan-genomes as comprehensively as possible to find all potential genetic adaptations to the urinary environment. The genetic diversity was therefore not assessed in depth here.

A noticeable difference between the species is the number of core genes as a percentage of the total number of OGs. The pan-genome of *E. coli* contains a core genome of only 100 OGs, this number is significantly lower than expected based on previous literature, where a pan-genome of 2247 *E. coli* strains yielded a core of over 1500 genes[36]. These results were obtained using a different pan-genome construction approach, applying only CD-HIT and a lower identity (70%) to determine OGs. Where the usage of CD-HIT alone will increase the number of OGs, the lower identity will lead to the grouping of not-true orthologs in one OG, decreasing the number of OGs[37]. It is therefore important to compare pan-genome sizes between pan-genomes constructed in a similar manner. A second study on the pan-genome in *E. coli* using Roary reported similar results as found here on a set of more than 1300 genomes of this species, finding 104 core OGs and 223 when using highly similar strains[38]. These numbers are significantly smaller than the core genomes of the other species analysed here. It is unlikely that the pan-genome construction approach here is the cause of the small core genome for *E. coli*, as the core genomes found for the other four species are in line with previous literature[39–42]. One possible reason for this small core genome of *E. coli* could be the inclusion of genomes from the *Shigella* species mistakenly annotated as *E. coli*. *Shigellae* are phylogenetically *E. coli*, having been initially classified as part of the same species but later recognised as separate species and sharing 80%-90% nucleotide similarity[43–45]. This potential inclusion would increase the diversity within the dataset for this species and thereby reduce the core genome size.

As discussed previously, the pan-genome construction method and parameters are of influence on the OGs found. Moreover, Prokka annotations of a high number of genomes could lead to an increased number of annotation errors, leading to an inflated accessory genome and a reduced core genome[46,47]. Therefore, a second pan-genome construction strategy utilising a graph-based approach with the ability to correct for some of the false positive and negative annotation sources was applied using Panaroo. Panaroo utilises gene re-finding where the amino acid sequences of re-found genes are not easily accessible. Moreover, the construction of the pan-genomes for the larger datasets of *E.*

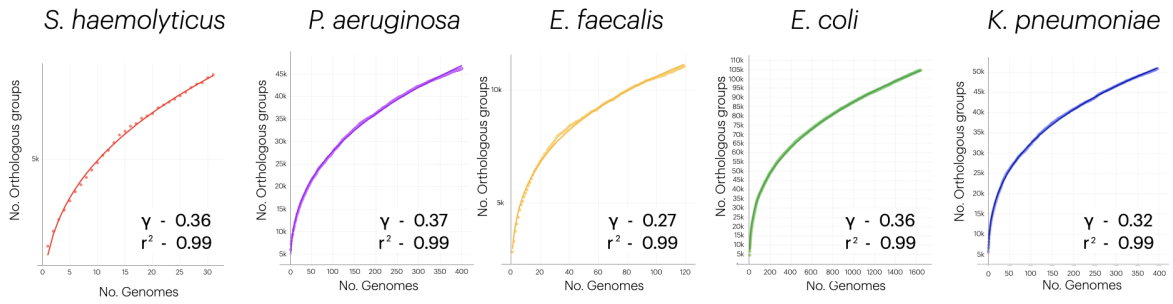


Figure 1: Pan-genome plots for all five species, with the number of genomes used for pan-genome construction on the x-axis and the the number of identified orthologous groups on the y-axis. A line was fitted using the formula of Heaps law, $n = kN^{\gamma}$, fitted γ values indicate all open pan-genomes. r^2 values indicate goodness of fit.

coli and *K. pneumoniae* was estimated to last more than 400 days. The pan-genome results of this tool were therefore not used in further analysis.

After pan-genome construction, the openness of the pan-genome for each species was calculated using Heaps law, $n = kN^{\gamma}$.[48]. Here n represents the number of orthologous groups found, N the number of genomes analysed and k and γ represent parameters to fit the function, where γ is a measure of pan-genome openness[48]. Fitting Heaps law to the pan-genome collection graphs resulted in a γ value for each species above 0, indicating an open pan-genome. Sequencing additional strains will therefore increase the number of OGs found, as well as potentially including OGs from the unique genome into the accessory genome and reducing the core[48].

The open pan-genomes are in line with expectations, as the strains analysed here were isolated from a broad range of environments, an indication that these species are generalists and therefore have open pan-genomes[49]. Furthermore, previous literature on the pan-genomes of these species has shown open pan-genomes with comparable γ values, except for *S. haemolyticus*, where no reference values were available[38,41,50–52]. The calculated γ value for *E. faecalis* was lower than for the other species, 0.27 versus 0.36, 0.37, 0.36 and 0.32 for *S. haemolyticus, P. aeruginosa, E. coli* and *K. pneumoniae* respectively, which is an indication of lower genetic diversity within this species.

Annotation uniformity validates orthologous gene clusters

Choosing the right parameters is of importance for the correct clustering of genes in orthologous groups within the pan-genome. To verify if the gene clusters as identified by Roary contain true orthologs, the uniformity of the COG annotations of the sequences within each group was analysed for the gene clusters identified as urine unique (Figure 2A). This analysis was performed on the urine unique gene clusters only due to the computational time required to annotate every sequence within the pan-genome with EggNOG. COG annotations were used as this was the most common annotation type for the sequences. Results show a general COG annotation uniformity within OGs, with

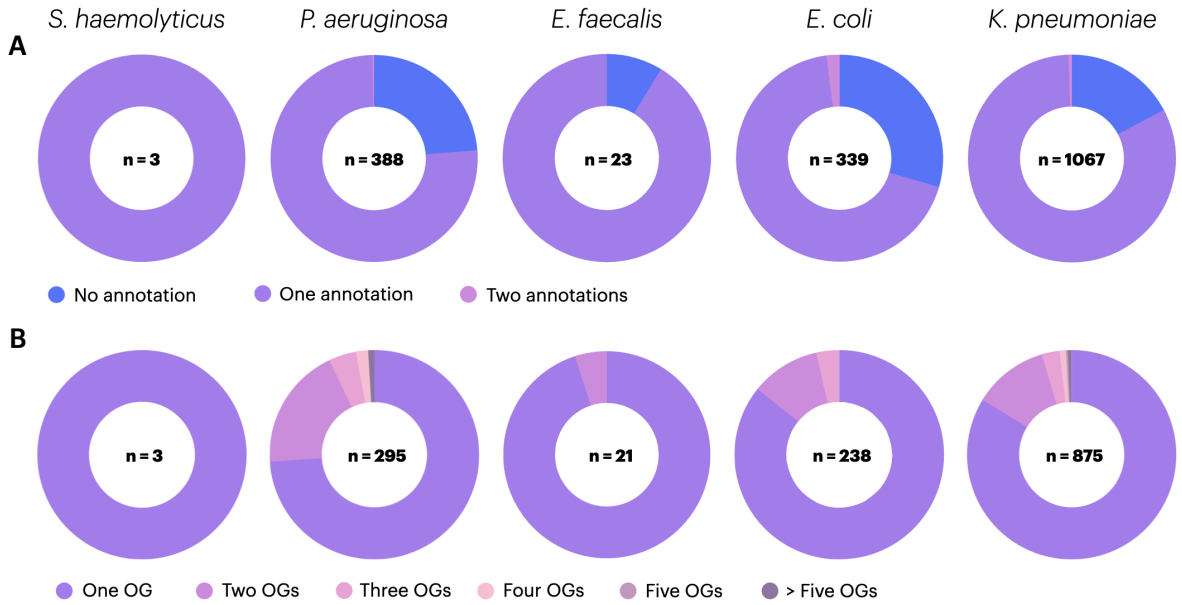


Figure 2: (A) Uniformity and (B) uniqueness of the Orthologous group (OG) annotations, (A) based on number of COG annotations within one group or (B) the number of occurrences of the COG of the OGs containing one annotation.

at most two annotations per gene cluster, as well as a low percentage of gene clusters with more than one annotation or none for *S. haemolyticus* and *E. faecalis*. This result indicates that the gene clusters mostly contain true orthologs and that any sequence within a cluster could be chosen as representative of its cluster.

Alongside uniformity of annotation within one group, the uniqueness of each COG assigned to one OG was tested for all annotated OGs (Figure 2B). These results showed that several OGs had been split, with the majority of OG COG annotations being unique within each species. The most erroneously split gene clusters were present in the *P. aeruginosa* pan-genome, where 19% of the annotations were shared between two or more gene clusters and a further 7% between three or more clusters, indicating that the sequence identity for OG clustering was possibly set too high at 95% for this species. Future studies could analyse the optimal sequence identity for the grouping of gene sequences into clusters. These more fragmented clusters could increase the number of OGs found for *P. aeruginosa* in comparison to the other species and relative to the number of urine genomes within the dataset, making additional analysis of duplicate or closely related annotations warranted. In conclusion, most groups contain true orthologs and a representative sequence can reliably be chosen from one group. For the purpose of this study, duplicate KEGG KO annotations were treated as a single entry due to the splitting of OGs. The settings for pan-genome construction might have to be altered in future analyses to prevent this splitting.

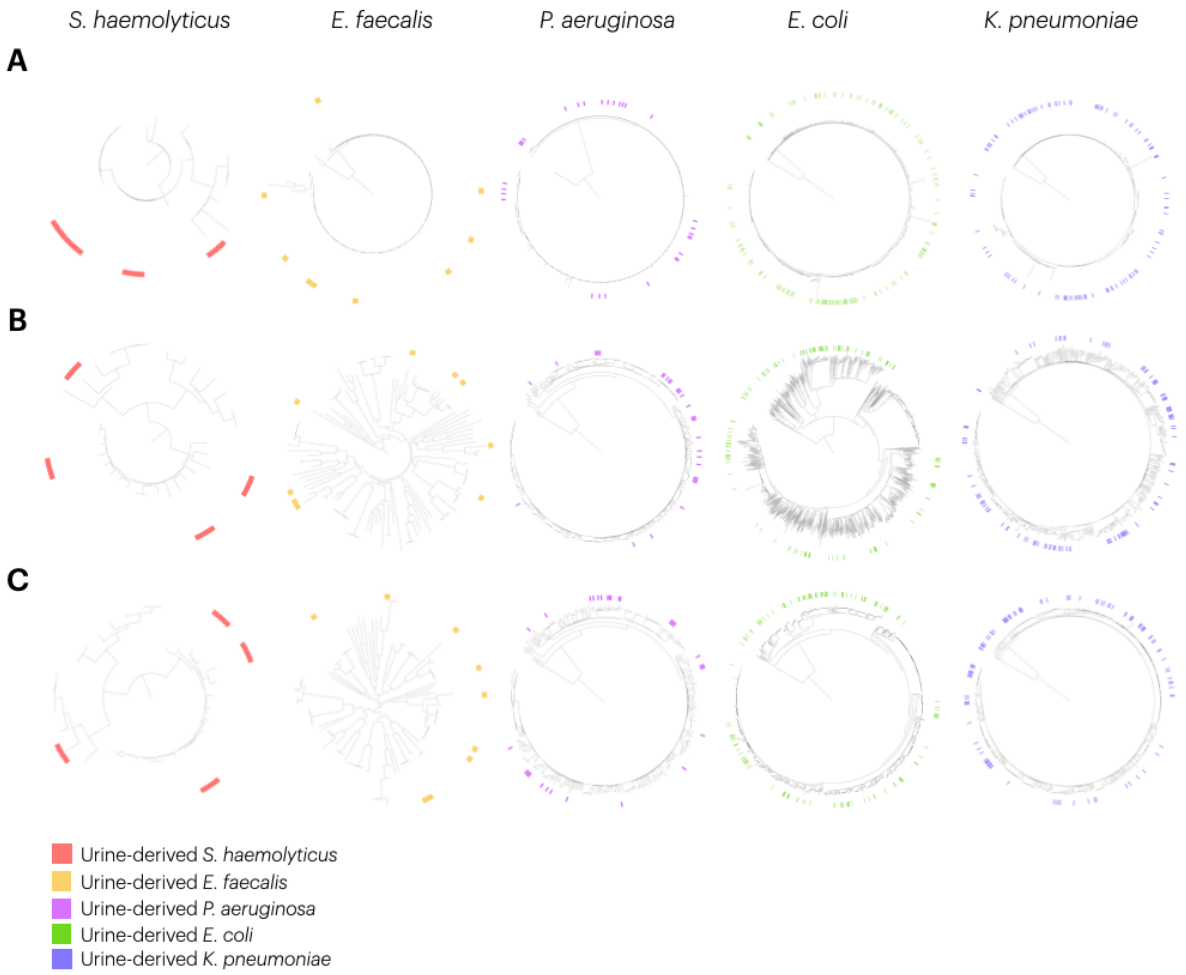


Figure 3: Phylogenetic relationship of five species of interest, colors indicate location of genomes with a urinary origin of isolation for each species. (A) Phylogenies of the longest predicted 16s gene for each strain, identifying one outlier for the species *P. aeruginosa*, *E. coli* and *K. pneumoniae* that was kept in the analysis dataset, showing no clustering of genomes isolated from urine. (B) *k*-mer and (C) core genome phylogenetic analysis showed no clustering of urine genomes for all species.

No strong phylogenetic signal for the urine origin of isolation

The general relatedness between the strains was analysed using the predicted 16s gene from each strain, if one was available (Figure 3A). For *E. faecalis*, *E. coli*, *K. pneumoniae* and *P. aeruginosa* similar results were observed, a tree containing short branches and bootstrap values as low as 1%, indicating little signal and differences between these clades. For *E. coli*, *K. pneumoniae* and *P. aeruginosa* one outlier was observed in the 16s tree that was not trimmed at the filtering step, as the FastANI calculations of the Average Nucleotide Identity (ANI) showed ANI above 95% between almost all genomes

within the dataset and the found outlier (Figure S2). Moreover, predicted 16s sequences were not available for all genomes, these outliers might therefore not be true outliers but merely an indication of missing surrounding clades. Results for *S. haemolyticus* showed a slightly altered pattern (Figure 3A). Distinct clades are observed and two strains isolated from urine group together in one clade, however, with a bootstrap value of 9%, no strong conclusions can be drawn from this tree. To quantify potential clustering of 16s genes from urine-derived strains, a permutation test of all 16s phylogenies was performed[53,54]. No significant clustering of the metadata was found in the 16s phylogenies for *E. faecalis* and *S. haemolyticus* ($p = 0.001$, permutations = 10 000). Significant clustering was observed for the other three species, however, cluster purities were low (0.22 - 0.34) and no clusters unique to urine were observed, as well as the absence of clusters containing the majority of 16s genes with a urinary isolation origin. In conclusion, the resulting 16s phylogenies showed no visual or statistical clustering of the 16s genes from strains isolated from urine, indicating that the urine environment does not influence the relatedness of the 16s gene within these species (Figure 3A).

Subsequently, the hypothesis of potential clustering of urine-derived strains was tested using a *k*-mer-based approach. MashTree was used to create an overview of the general relationship between strains and analyse if the potential relatedness of urine-derived strains is detectable on genome level in a fast manner[55]. Visual and permutation test analysis of the clustering of urine strains on the Mashtrees shows no urine unique clades, only clades where urine is absent (Figure 3B). One of these clades for *E. coli* was investigated in more detail, as this clade was the largest clade where urine was absent, thereby increasing the numbers for analysis. The majority of isolation origins within this clade were faeces or faecal-related sites, the most prevalent isolation origins in the dataset for this species. Other isolation sites, such as cattle or food, were also present within this clade. Additionally, host or environmental isolation, genome length and GC content were analysed and no apparent bias was found between this clade on GC content or host association isolation (Figure S3). Genomes within this clade were found to be on average 450 kb longer than genomes in the rest of the tree ($p < 0.0001$), which could indicate the presence of genes that are lost in strains isolated from urine.

It should be noted that the MashTree approach, although fast, does not result in a tree of high accuracy. MashTree is a *k*-mer-based approach that uses a Bloom filter to filter the top *k*-mers to use in the MinHash algorithm[55]. Thereby not taking into account all genomic information as well as the fact that a Bloom filter does allow for false positives, reducing the accuracy of the resulting tree. Moreover, Mashtree does not infer phylogeny. Therefore, more accurate analysis of the possible clustering is warranted.

As no apparent clustering signal was found with a *k*-mer-based approach as well as on the conserved 16s genes, the core genes of all species were analysed. Hereto the construction of a maximum likelihood phylogeny using IQtree was attempted and proved to be unsuccessful for the species, *P. aeruginosa*, *E. coli* and *K. pneumoniae*, due to unknown reasons. Therefore, after careful consideration, this attempt was halted and a less accurate approach was taken using Fasttree[56]. All resulting phylogenies show no clustering of genomes isolated from urine, for both the visual as well as the permutation test analysis (Figure 3C). For future analysis with different alignments or species, IQtree

Table 3: Number of urine related orthologous groups (OGs) identified by Scoary ($p < 0.05$) split in under- and overrepesented groups

Species	No. Overrepresented OGs / (Urine Unique)	No. Underrepresented OGs / (Absent urine)
S. haemolyticus	4 / (3)	29 / (2)
P. aeruginosa	583 / (388)	2173 / (53)
E. faecalis	68 / (23)	227 / (4)
E. coli	629 / (339)	11350 / (1592)
K. pneumoniae	1165 / (1067)	3152 / (113)

will be preferred, as it has been shown that IQtree yields more accurate phylogenies[56]. Moreover, FastTree is not the optimal method for alignments of closely related sequences, as present in this set, due to the fact that FastTree does not account for recombination or gene conversion and the absence of traditional bootstrapping using the Shimodaira-Hasegawa test to assess the reliability of the tree topology in each split[56,57]. In conclusion, the phylogenetic analysis on 16s genes and core genome phylogenies as well as the *k*-mer-based analysis does not show relatedness between urine-derived strains.

Accessory genome analysis suggests niche adaptation in urinary bacterial strains

The absence of a strong phylogenetic signal does not exclude the possibility of a genetic relation between strains and the urine environment. It can therefore be hypothesised that there are genes over- or underrepresented in bacterial strains isolated from the urine environment. The accessory genome is the part of the pan-genome shared between a subset of two or more strains[33,35]. Here, we study the functions encoded by the genes in this set, as they are known to be involved in non-essential processes that can give a selective advantage, such as niche adaptations and antibiotic resistance[34].

We used the feature selection method Scoary, a microbial pan-GWAS approach identifying over- and underrepresented OGs within the presence-absence profile of the accessory genome. Splitting the identified OGs in these groups showed more underrepresented OGs in strains isolated from urine than overrepresented OGs (Table 3). This could be due to the absence of several nutrients in urine, such as most proteins and often glucose[25]. Possessing genes facilitating the catabolism of these nutrients will therefore most likely not increase fitness when growing within the urinary environment. Additionally, the group of underrepresented OGs is likely to contain genes important for growth or fitness in other environments present within the dataset. This together with the finding of longer genomes in one clade where urine strains were absent for the *k*-mer-based tree of *E. coli* (Figure S3), the larger number of underrepresented OGs could point to a role for reductive evolution of strains living in the urinary tract. It could therefore be hypothesised that due to the nutrient-limited environment, bacteria living in the urinary tract

adapt by reductive evolution as excess genes have a fitness cost to the bacterium, enhancing positive selection for gene loss[58]. Studies have found that reductive evolution might play a role in the change of uropathogenic *E. coli* to a commensal lifestyle[59,60]. However, here we made no distinction between strains isolated from healthy urine and UTIs, and positive selection is reported to mostly occur in bacteria living in stable nutrient-rich environments[58]. Future studies investigating the role of reductive evolution of urinary tract-based strains can highlight the underlying evolutionary changes.

Protein-Protein interaction networks show genetic adaptations to urine host environment

Potential urinary tract bacteria-specific protein-protein interactions of urine-unique OGs were predicted using STRING[61]. *S. haemolyticus* showed no connections within STRING. Gene annotation revealed two unnamed proteins and a third gene annotated as Uracil-DNA glycosylase (*udg*), a DNA repair enzyme initiating the uracil base excision repair pathway[62]. Loss of *udg* increases mutation rate in *E. coli* while growth is unaffected, only reducing growth in bacterial species with high CG content under conditions of increased reactive nitrogen intermediates production[63,64]. Increased production and availability of reactive nitrogen intermediates is known to be a part of the host defence during UTI, as well as within macrophages that are present in urine during infection[65,66]. It could therefore be hypothesised that *udg* plays a role in pathogenicity of *S. haemolyticus* in urine. As no previous studies have investigated the *Staphylococcus* genus, future studies should investigate this novel finding. Moreover, *udg's* potential role in *S. haemolyticus* living in urine is based on four genomes in a set where not all genomes are complete. This, combined with *S. haemolyticus'* open pan-genome, could alter the finding of urine uniqueness for *udg*.

For the subsequently analysed species, *E. faecalis*, 17 of the 23 urine unique OGs were identified by STRING, forming two clusters (Figure 4A). The genes of the first cluster play a role in the uptake of β-glucosides, an alternative carbon source for glycolysis, an important mechanism as glucose is scarce in urine (Figure 4A (yellow))[67]. The second cluster contains genes, in addition to two unconnected genes, that are reported to be functionally enriched for the WxL domain and LPXTG cell wall anchor motif (4 out of 8 present in this network) (Figure 4A (dark blue)). The WxL domain is the characteristic domain of a family of cell surface proteins, of which one of its members has been identified to contribute to *E. faecalis* pathogenicity in UTI[68–70]. Together these adaptations may increase the fitness and pathogenicity of *E. faecalis* in the urinary tract, however, as the WxL domain family has many members future studies are warranted.

Two clusters with annotated functions were identified in the STRING network for *P. aeruginosa* (Figure 4B). The first cluster was densely interconnected and functionally enriched for conjugation, a type of horizontal gene transfer (HGT) that requires bacterial

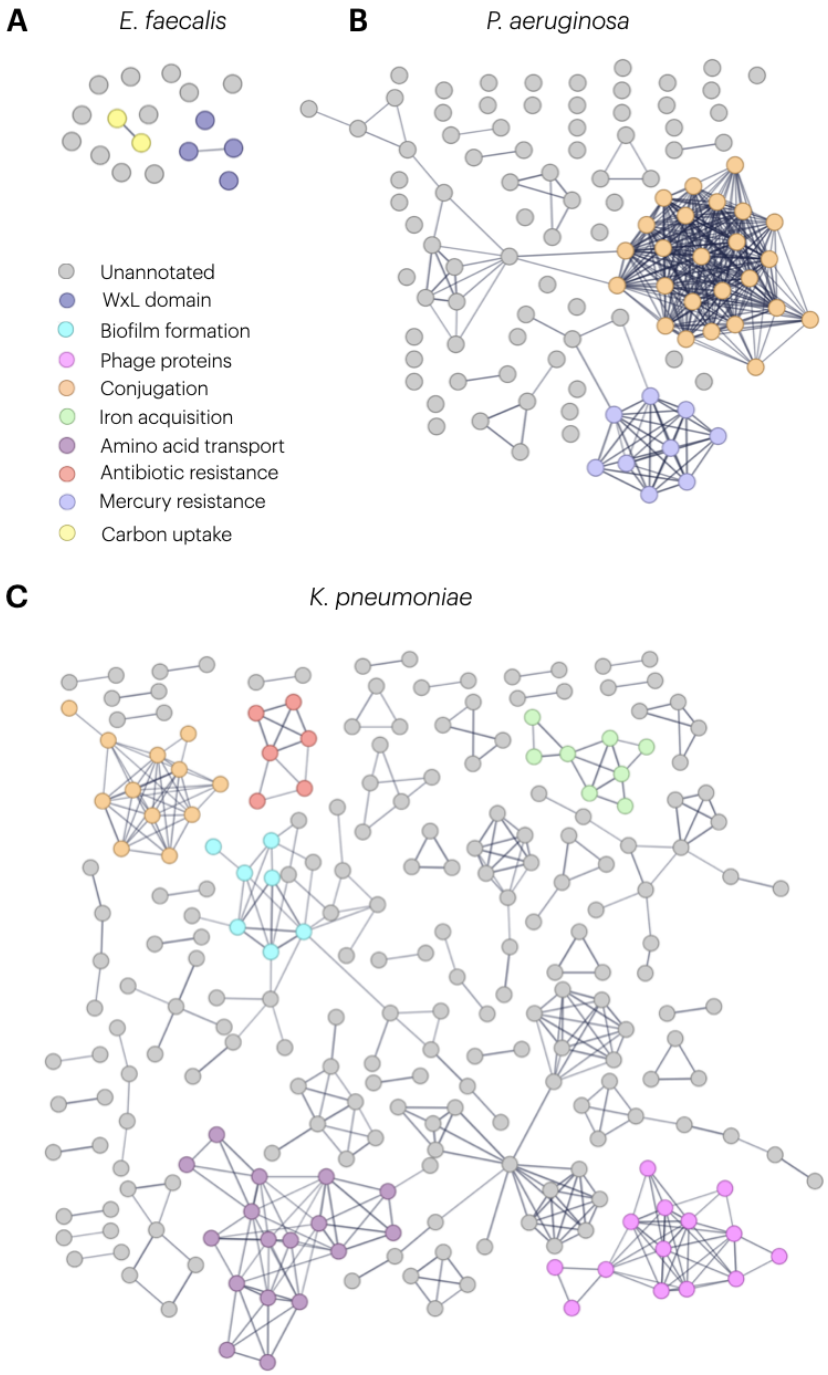


Figure 4: STRING networks of genes uniquely present in urine. (A) *E. faecalis*, alternative carbon source acquisition (yellow), WxL domain containing proteins (dark blue). (B) *P. aeruginosa* showing clusters for conjugation (orange) and mercury resistance (blue). (C) Network clusters for *K. pneumoniae* where 272 non connected nodes are hidden. The network shows distinct clusters with phage related proteins (pink), a cluster representing conjugation (orange) and clusters indicating biofilm formation (light blue), iron acquisition (green), amino acid transport (purple) and antimicrobial resistance (red). Interaction score > 0.007.

contact (Figure 4B (orange)). This bacterial contact in urine can be facilitated by biofilm formation, which is reported to be a hotspot for HGT in bacteria, amongst which *P. aeruginosa*[71–75]. HGT might play a role in the transfer of virulence and antibiotic resistance genes in *P. aeruginosa* and could be the origin of the second cluster (Figure 4B (blue))[76–78].

This second cluster's genes play a role in the response to mercury, a response less expected to be present in urinary tract bacteria as the urinary mercury concentration is low in healthy individuals[79]. A genetic linkage has been reported between mercury resistance and AMR, a resistance more expected in urinary tract bacteria as UTIs are commonly treated with antibiotics. The reported rise in AMR genes when bacteria are subjected to high mercury environments could therefore be bidirectional, mercury resistance could co-arise with AMR genes in an environment under antibiotic treatment[80]. Additionally, a plasmid carrying mercury and antimicrobial resistance genes has been identified, pointing to HGT as a possible origin of these genes[81]. This hypothesis needs further investigation as no clusters for antimicrobial resistance or biofilm formation were found for *P. aeruginosa* in the urine unique set nor by a preliminary analysis of the over-represented set.

Where datasets for *E. faecalis* and *S. haemolyticus* were small for analysis, the dataset for *K. pneumoniae* was the largest, where 48% of sequences mapped to a reference in STRING. Several clusters with distinct functionalities can be identified within this network, one being a cluster pointing to the presence of phage-related genes within urinary *K. pneumoniae*, indicating previously reported phage infections (Figure 4C (pink))[82]. Similar to *P. aeruginosa*, a conjugation cluster, here in addition to a biofilm formation cluster, was observed (Figure 4C (orange) and (light blue)), supporting the hypothesis that biofilms are a hotspot for HGT among urinary tract bacteria. In addition to facilitating HGT, biofilms increase antimicrobial tolerance for bacteria residing within them, as antibiotics poorly penetrate the biofilm. The subinhibitory dose of antibiotics could further increase antibiotic resistance, as selection for resistant bacteria can occur at concentrations several hundred folds below the lethal concentrations[12,83,84]. This hypothesis of an increase in AMR is strengthened by a cluster containing genes implicated in resistances against a commonly used antibiotic treatment, β-lactam and antimicrobial defence by the host, Cationic Antimicrobial Peptides (Figure 4C (red))[85,86].

The largest cluster present corresponds to amino acid uptake and could be hypothesised to improve fitness, as several branched amino acids are scarce in urine and improved uptake can aid in competition, however, this has not been studied to date[87]. It has been reported that several components of amino acid transport could increase virulence in *K. pneumoniae*, and amino acid transport is upregulated in uropathogenic *E. coli*[28,88]. Pointing to the relevance of further testing this hypothesis.

The last cluster to point out for *K. pneumoniae* corresponds to iron acquisition (Figure 4C (green)). Iron is an essential nutrient for all life, known to be highly sequestered by the host to fight bacterial infection, bacteria colonising the urinary tract therefore need acquisition strategies to be successful pathogens[89,90]. Moreover, iron is an important nutrient in nutrient competition between *E. coli* and *K. pneumoniae* in a urinary biofilm[91].

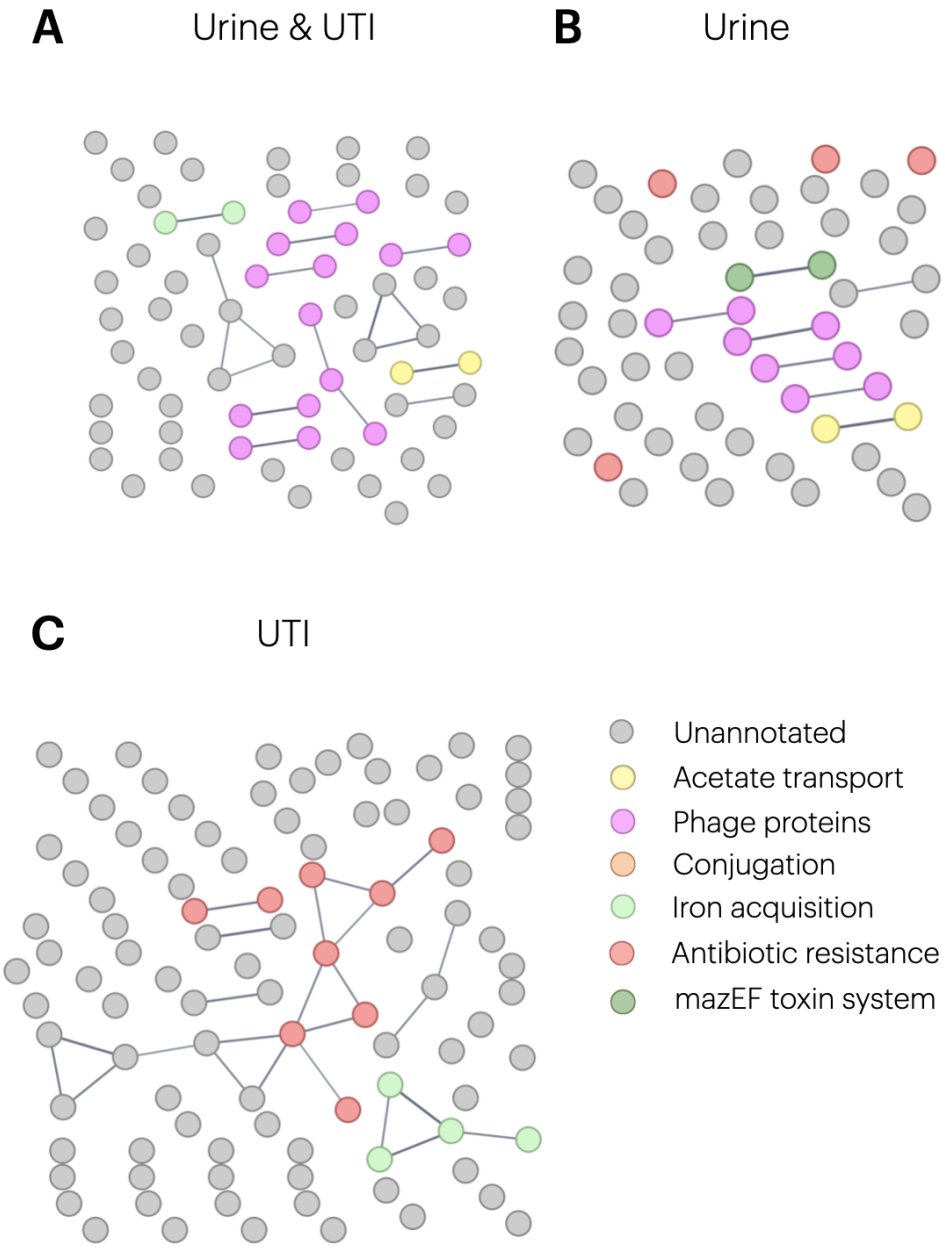


Figure 5: STRING networks of genes uniquely present *E. coli* strains isolated from urine. (A) *E. coli* STRING network of unique genes urine and UTI strains combined with clusters corresponding to phage infection (pink). Additionally clusters corresponding to iron acquisition (green) and acetate transport (yellow). Analysis for *E. coli* in healthy urine samples (B) and UTI samples (C), both analysis contain antimicrobial resistance genes (red). Additionally in strains isolated from healthy urine show clusters for acetate transport; yellow and mazEF toxin system: dark green. Iron acquisition was only observed in UTI strains ((C); green). Interaction score > 0.007.

Therefore unsurprisingly, a cluster corresponding to iron acquisition was also observed in the network for *E. coli* (Figure 5A (green)). In this species successful iron acquisition is reported to promote virulence and to be upregulated during UTIs. Moreover, iron acquisition strategies evading iron sequestering by the host are more frequently identified in uropathogenic *E. coli* than in faecal commensal *E. coli*, indicating a role for iron acquisition in pathogenic bacteria in urine[92,93].

The organism most often studied in UTIs or related infections is *E. coli*. It is therefore surprising that here only 20% of sequences could be mapped to a reference in STRING. The low number of mapped sequences is likely due to the absence of these genes in the reference strain rather than false positives in gene predictions by Prokka. This is supported by the findings that the trend for the number of genes per genome and gene lengths does not notably differ for *E. coli* in comparison to the other four species studied (Figure S4), as well as 55% of OGs being annotated by STRING in the absent in urine group for *E. coli*.

Multiple clusters corresponding to phage infections are present in the network for *E. coli* (Figure 5A (pink)). Phages have been proposed as a novel therapeutic strategy to treat multi-drug-resistant uropathogenic bacteria[94,95]. Identifying phages infecting uropathogenic bacteria could aid the development of this therapeutic strategy. Alongside these phage-related clusters, a cluster unique to urinary *E. coli* was detected. This cluster corresponded to acetate transport, a compound whose concentrations are increased in urine after antibiotic treatment, as well as increasing virulence of uropathogenic *E. coli*[96,97]. In conclusion, this STRING analysis shows that bacteria isolated from the urinary tract show genetic adaptations increasing virulence and pathogenicity as well as adaptations enhancing AMR and HGT.

Indication for antibiotic resistance in healthy urine strains of E. coli

To establish which traits can be unique to pathogens isolated from UTIs in comparison to healthy urine, *E. coli* strains isolated from urine and UTI were analysed separately. This analysis showed that iron acquisition genes were uniquely present in UTI-isolated *E. coli* (Figure 5C (green)). This does not indicate that strains isolated from other sources do not contain genes for iron acquisition, but rather that more unique genetic components for iron acquisition are present in strains isolated from UTI over other isolation sites. This is in correspondence with the findings that evasion of iron sequestration by the host increases virulence and pathogenicity[92,93]. Additionally, one new cluster unique to strains isolated from healthy urine was observed corresponding to the MazEF toxin-antitoxin system, a system mediating growth arrest and persister cell formation during antibiotic treatment, thereby increasing bacterial survival (Figure 5B (dark green)). This might suggest that the *E. coli* in healthy urine samples have been previously exposed to antibiotics, which could also account for the presence of the four non-clustered antibiotic resistance genes detected in these strains (Urine; Figure 5B (red), UTI; Figure 5C (red)). This hypothesis could be supported by the finding of the acetate transporter cluster in strains isolated from healthy urine.

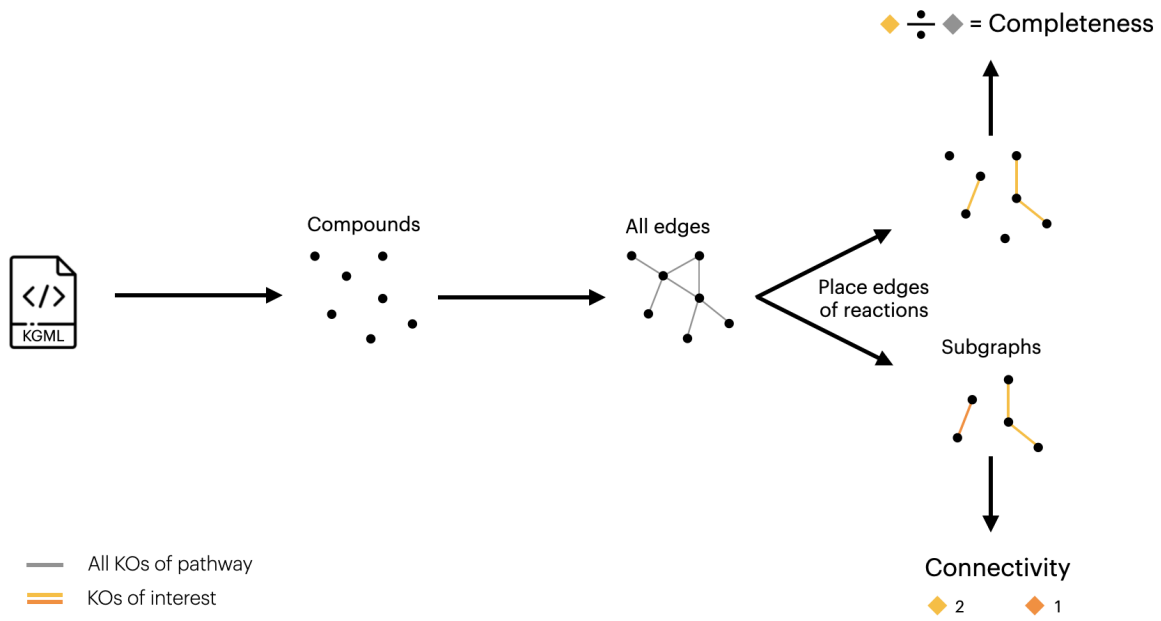


Figure 6: Graphical overview of the calculation of pathway completeness and connectivity from KEGG KGML files and the KOs of interest.

Urine unique genes are present in pathways promoting virulence and antibiotic resistance

STRING analysis shows protein-protein interactions, however, genes and their encoded proteins can be part of larger pathways. Analysing these could highlight metabolic pathways and functions enriched or underrepresented in bacterial strains isolated from urine. Hereto the urine unique OGs were annotated with KEGG KO numbers and mapped to their respective pathways (Table S3). Subsequently, the percentage of urine unique genes in each pathway was analysed, showing no pathways unique to the urinary lifestyle, only pathways containing urine unique genes (Figure 6; Figure 7A).

No pathways with urine unique genes were shared by all five species, suggesting a species-specific urinary host adaption profile. For the species with the smallest dataset, *S. haemolyticus*, only one gene mapped to a pathway, the previously identified *udg* mapping to base excision repair. One urine unique KO for one other species, *K. pneumoniae*, was present in this pathway. Where a possible role of *udg* in infection was discussed previously, the urine unique gene of *K. pneumoniae*, *nei*, was not implicated in changes of phenotype for this species and was, by contrast, uniquely absent from *E. coli* isolated from urine. Since the base excision repair pathway is the principal pathway for repairing small base lesions in DNA, it is unlikely that the interspecies differences within this pathway are indicative of any urine unique function.

Protein-protein interactions pointed to the urine unique alternative carbon source uptake of β-glucoside by *E. faecalis*. This pathway analysis shows that the β-glucoside transporter found in *E. faecalis* is part of the phosphotransferase system (PTS), a system

Figure 7: KEGG pathways with genes uniquely present in bacteria isolated from the urinary tract. (A) Pathway completeness defined as KOs uniquely present in urine-isolated bacteria as a percentage of total KOs present in the complete pathway. (B) Connectivity of genes uniquely present in urine-isolated bacteria, in size of connected components defines as number of KOs connected in pathway, dot size indicates number of connected components of given size.

where urine unique genes of *K. pneumoniae* were also mapped to (Figure 7A). The genes of the latter species form a complete complex that facilitates cellobiose uptake, a sugar compound known to be present in urine, and can be metabolised to glucose[98]. One component of this complex was found to be unique to *E. faecalis* strains isolated from urine and the other complex components were found to be present within the core genome of this species.

Alongside uniquely present genes for cellobiose uptake, one component of the cellobiose transport complex was absent in *E. coli* strains isolated from the urinary tract, indicating that this species does not use this carbon source to grow in the urinary tract. Together, these results point to a role for cellobiose as carbon source for some bacterial species growing in the urinary tract, and possible nutrient competition between *E. faecalis* and *K. pneumoniae*, but not for *E. coli*. Additionally, the presence of one of the components of the cellobiose transport complex, *celB*, has been reported to increase virulence and aid biofilm formation in *K. pneumoniae*, making cellobiose possibly more than just a carbon source[99].

In addition to nutrient consumption, analysis revealed two species with urine unique genes mapping to the pathway for cellular motility mediated by flagella, which is known to increase fitness in uropathogenic *E. coli*[100,101]. For both *E. coli* and *K. pneumoniae*, one urine unique gene was annotated, being *flgN* and *fliY*, respectively. The *FlgN* protein regulates flagellar assembly, which needs tight regulation as flagella are down-regulated in chronic *E. coli* infection and biofilms but are needed for bacterial ascension from the bladder to the kidneys[28,102,103]. Moreover, flagella play a role in *E. coli*'s ability to form intracellular bacterial communities within bladder epithelial cells[104–106]. Bacteria residing within these intracellular communities are shielded from antibiotics, washing out and the host immune system, establishing a quiescent reservoir of pathogenic cells, thereby contributing to infection recurrence and possibly to resistance to antimicrobial treatments[107,108]. When comparing urine and UTI-isolated *E. coli* strains, the role of flagella in uropathogenicity becomes more apparent as urine unique genes within the flagellar assembly pathway were only found in strains isolated from UTIs (Figure 8A).

The finding of a gene regulating flagellar assembly in *K. pneumoniae* is surprising, as this species is considered to be a non-motile and non-flagellated bacterium[109]. A recent study found flagella in a strain isolated from a neonatal sepsis patient, suggesting that flagella-mediated motility may be a novel way for *K. pneumoniae* to increase virulence[110,111]. This possible role for flagella in this species for life in the urinary tract remains to be studied further as the *fliY* gene also functions as a transporter component facilitating cystine uptake.

Solely interpreting pathway completeness and drawing conclusions on this metric only would yield a biased result, as reactions are often catalysed by several alternative enzymes. Moreover, urine unique enzymes scattered over a metabolic pathway with alternative routes might have less weight than a urine unique connected component. Hereto the connected component sizes of urine unique genes were calculated for each species (Figure 6; Figure 7B). This analysis shows that for many pathways the urine unique genes do not form a connected component of a size larger than one, meaning no urine unique connection of subsequent reactions, thereby possibly decreasing the urine unique

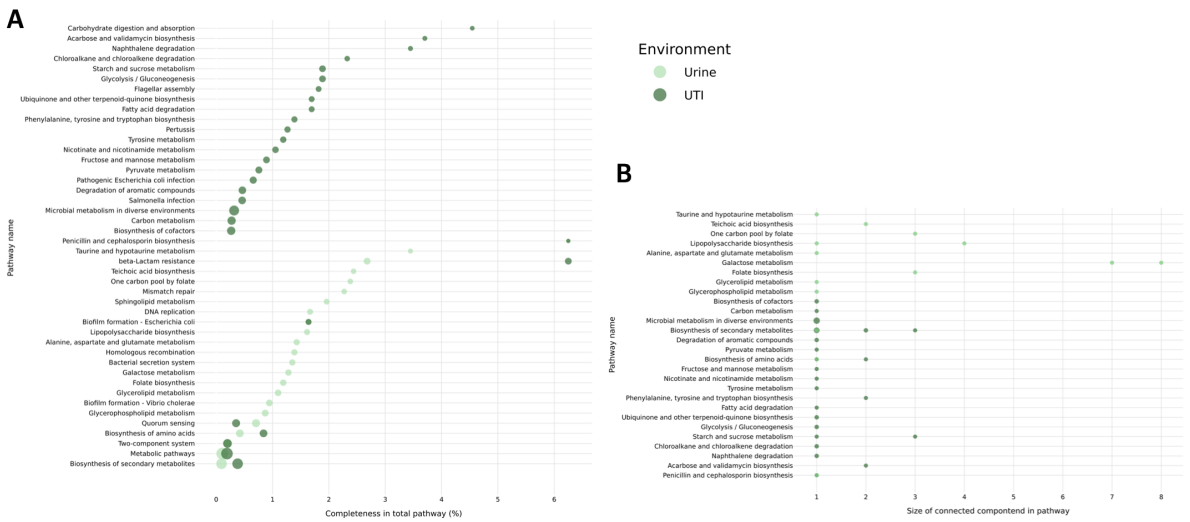


Figure 8: KEGG pathways with genes uniquely present in *E. coli* isolated from the urinary tract in healthy and urinary tract infections (UTI). (A) Pathway completeness defined as KOs uniquely present in urine-isolated of UTI-isolated *E. coli* as a percentage of total KOs present in the complete pathway. (B) Connectivity of genes uniquely present in urine-isolated of UTI-isolated *E. coli*, in size of connected components defines as number of KOs connected in pathway, dot size indicates number of connected components of given size.

effect on the bacterial metabolism.

One pathway with a connected component of a size larger than one is folate biosynthesis. Bacteria require folate to synthesise nucleic acids, however most bacteria cannot transport folic acid across their cell walls, making it an essential pathway[112,113]. Urine unique genes were identified for *K. pneumomiae*, *E. coli* and *P. aeruginosa*, all unable to transport folate. *P. aeruginosa* and *K. pneumomiae* formed a connected component with the *sul1* gene, a target for the sulfonamide class of antibiotics, often used in the treatment of UTIs[114,115]. However, *sul1* is a mutated form of the original target providing resistance against sulfonamide treatment[116]. The second connected component within this pathway was formed by the enzyme catalysing the conversion reactions between folate, di- and tetrahydrofolate and was uniquely present in urine for *E. coli* and *K. pneumoniae*. This conversion too is a target for a class of antibiotics often used in the treatment of UTI, here trimethoprim. However, only the urine unique gene for *E. coli* is a known AMR gene, while annotation for *K. pneumoniae* was unclear. When analysing if the AMR gene was present in urine- or UTI-derived *E. coli* strains, the gene was found to be unique to strains isolated from healthy urine. Indicating that AMR is also found in non-pathogenic *E. coli* populations.

Atrazine degradation is a pathway implicated in recurrent UTIs by facilitating the formation of urinary stones, urolithiasis, by means of the urease enzyme[117]. The *urease* gene is present in the core genome of *P. aeruginosa*, indicating possible urolithiasis when *P. aeruginosa* is present in the urine microbiome. Two urine unique genes were

Table 4: Percentage of KEGG KO annotations with antimicrobial resistance.

Species	**AMR overrepresented**	**AMR underrepresented**
S. haemolyticus	0	0
P. aeruginosa	9.4%	7.5%
E. faecalis	0	4.4%
E. coli	14.7%	2.6%
K. pneumoniae	7.8%	5.9%

identified in *K. pneumoniae* facilitating urolithiasis in an alternative two-step process. The presence of urinary stones within the urinary tract allows for bacteria to migrate into these stones where they are shielded from washing out as well as from host defences and antibiotic treatment, forming a quiescent reservoir of pathogenic cells[118,119]. After initial antibiotic treatment bacteria can recede into the urine environment, now containing a subinhibitory dose of antibiotics, thereby possibly selecting for AMR[83,84,120].

Several previously identified results point to a role for AMR in urinary tract bacteria. Further indication in this direction is the presence of urine unique genes in the β-lactam resistance pathway. Separate analysis between urine and UTI-derived *E. coli* strains reveals that β-lactamases are present in both conditions. To determine if AMR genes were overrepresented in urine-derived strains, the number of AMR KOs in the over- and underrepresented groups were analysed, showing a higher percentage of AMR genes in OGs overrepresented in urine-derived strains for three of the five species (Table 4). For *S. haemolyticus* no AMR genes were detected in either group, however the number of annotated KOs in each group was small, so no conclusions can be drawn on the overrepresentation of AMR for this species (Table 3). The trend of overrepresentation of AMR genes was also not followed for *E. faecalis*, where no AMR genes were identified within the overrepresented group but were in the underrepresented group. This could again be due to the low number of KOs in the overrepresented group or the low number of urine-derived strains. An alternative hypothesis could be that these results are due to the fact that *E. faecalis* and *S. haemolyticus* are gram-positive bacteria whereas the other three species are gram-negative, or due to species differences, as *E. faecalis* is naturally resistant to several antibiotics commonly used in UTI[121].

In conclusion, pathway analysis showed the presence of urine unique genes in pathways promoting virulence, adaptations to the host environment as well as an increase in AMR genes in strains isolated from urine. It should however be noted that the urine unique genes are often present in only a small number of genomes (2-5 genomes), making further investigation into these findings warranted.

Competitive interactions in the urinary tract mediated by secondary metabolites

Bacteria in the urinary tract live in communities, and interactions between community members can alter infection severity and complicate treatment[122–124]. Interbacterial interactions can, among other strategies, be mediated by the excretion of secondary metabolites, which are biologically active small molecules that are not required for viability within the host environment but can provide a competitive advantage[125,126]. The genes regulating assembly and transport of secondary metabolites are often grouped on the bacterial genome, forming biosynthetic gene clusters (BGCs)[127]. Here BGCs were clustered into Gene Cluster Families (GCFs) and classes to identify GCFs enriched for urine, as no GCFs or classes were unique to the urine origin of isolation.

Urine enrichment analysis at the BGC class level revealed two classes with significant changes in the number of BGCs of urinary origin. The first class, the ribosomally synthesised and post-translationally modified peptides (RiPPs) ($p = 0.02$), was found for *E. coli* and had an equal number of GCFs with more or less urinary BGCs. One GCF of this class was enriched for urinary BGCs (7 urine, 25 total, 2 expected) and predicted as agrD-like cyclic lactone autoinducer peptide, a class of peptides of which about one third are known virulence factors and might play a role in the switching from an adhesive commensal to a pathogenic lifestyle[128,129]. Interestingly, GCFs of another class predicted to encode siderophores were found to contain more as well as less urinary-related clusters than anticipated. This finding can be explained by the fact that *E. coli* can possess many redundant iron acquisition strategies, with none being essential for virulence, suggesting a greater role for some siderophores in urinary *E. coli*[89,90,130–134].

The second class showing significant changes was the not type 1 polyketide syntesases (PKS Other) class for *P. aeruginosa* ($p = 0.03$), however no GCF showed an altered number of urinary-related clusters. Thus, the relationship between this class and the urine environment requires further investigation. Another BGC class for *P. aeruginosa* contained a GCF predicted as N-acetylglutaminylglutamine amide (NAGGN) (3 urine, 9 total, 0.6 expected). NAGGN protects against osmotic stress, which is beneficial for survival in urine as urinary osmolarity can fluctuate, protection could therefore potentially improve fitness[135–137].

S. haemolyticus was the only species with the BGC class of Terpenes, containing two GCFs of which one was enriched for urinary BGCs. Terpenes are a large class of compounds known to have antimicrobial properties[138,139]. A whole pan-genome pathway analysis pointed to possible synthesis of the terpene geraniol, a known antimicrobial against *E. coli*, *P. aeruginosa*, *K. pneumoniae* and *E. faecalis*[140,141]. The excretion of a geraniol by *S. haemolyticus* might explain the observed negative interactions in the study by de Vos et al. (2017) (Figure S7)[14].

Another species where a predicted secondary metabolite was found to potentially influence interspecies competition was *E. faecalis*. Here one of the RiPPs class's GCFs was enriched for the urine origin of isolation and predicted to be regulating a class II lanthipeptide (4 urine, 20 total, 1 expected), a metabolite with antimicrobial properties against gram-positive bacteria with greater effectiveness against those closely related

to the producing organism, thereby potentially eliminating competition from bacteria within the same genus[142,143]. This could be a beneficial strategy in the nutrient-limited urinary environment, as resource overlap is believed to be a major contributor to intra-genus competition[144].

In conclusion, the predicted secondary metabolites that are found to potentially play a role for urinary tract bacteria might explain the interactions between members of the microbial community, as well as indicating strategies by which the species studied here might improve their fitness in the hostile urinary environment. However, genomic analysis does not directly correlate with environmental presence of secondary metabolite. Bacteria live in microbial communities, which can greatly affect their metabolic behaviour and thereby the secondary metabolites in the environment[145]. Caution when generalising results is therefore warranted.

Indication of resource competition among urinary microbial communities

To identify commonly and uniquely consumed and excreted compounds and thereby infer potential interactions, the changes in metabolites in media before and after growth were measured using DART metabolomics. Filtering on m/z accuracy and relative intensities revealed a different number of peaks in the artificial urine medium (AUM) measurements before culturing between the species (Table S5). This is possibly caused by residual noise in the measurements. Hereto only peaks common in all reference AUM measurements were analysed, and changes in relative intensities were calculated for these peaks. This resulted in 249 peaks for each measurement mode, positive and negative, for four species. Subsequently, common peaks were used to identify commonly consumed and excreted peaks.

Analysis of commonly consumed metabolites revealed two compounds, indicating possible resource competition (Figure 9A). One peak was identified as lysine, the second as possibly N6-Acetyl-L-lysine, an intermediate in lysine degradation. This corresponds with findings for *E. coli* indicating that small peptides and amino acids are the main carbon source for strains living in the urinary tract[146,147]. Given that lysine is shared between all species here, this could indicate that a similar carbon metabolism is used in *S. haemolyticus*, *P. aeruginosa* and *E. faecalis*. A transcriptomics analysis of these species similar to the one by Paudel et al. (2021) could confirm this[148]. The largest overlap in consumed metabolites was observed between *P. aeruginosa*, *E. coli* and *S. haemolyticus*, followed by the overlap between *P. aeruginosa* and *E. coli* (Figure 9A). This could indicate that resource competition is strongest between these species, and that *E. faecalis* is least impacted. However, in this analysis, *E. faecalis* was reported to consume only 11 metabolites, whereas the other species were found to consume upwards of 100 metabolites. This could have been caused by the smaller changes in relative intensities of consumed peaks for *E. faecalis*, thereby possibly falling below the selection limit of three standard deviations. Alternatively, this difference could be explained by *E. faecalis*' poor growth in AUM, thus requiring less nutrients.

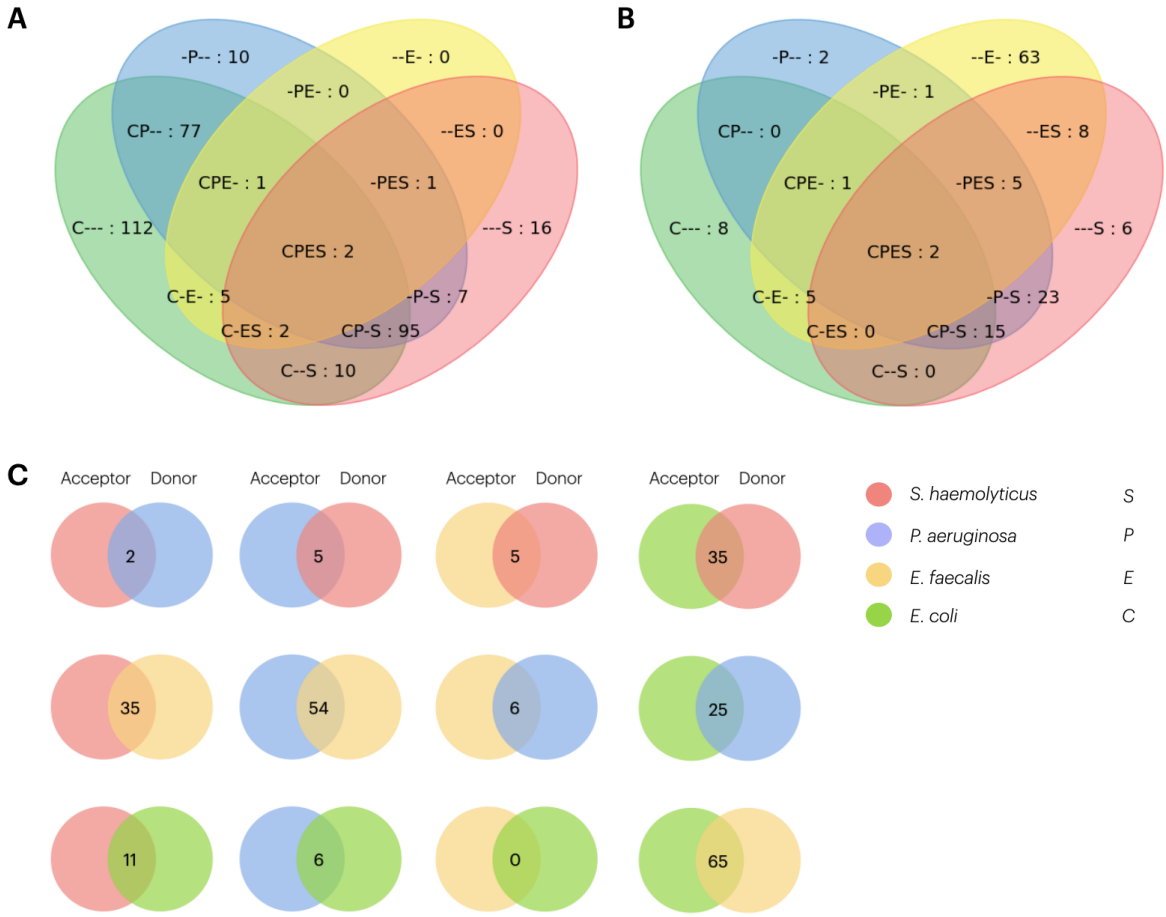


Figure 9: Overview of commonly metabolites present in the reference media as measured by DART-MS for (A) consumed and (B) excreted metabolites. (C) Possible cross-feeding interactions between species of measured metabolites present in reference media.

Analyses of excreted metabolites and their interspecies overlaps showed a different pattern where *E. faecalis* excreted most metabolites and *E. coli* the least, the difference was however less prominent, with 85 versus 31 respectively (Figure 9B). Of the two peaks identified as commonly excreted, one remains unannotated as none of the annotation strategies yielded a plausible result. Identification of the second peak yielded two potential candidates, one being 2,4-Diaminobutyric acid, a di-aminated form of butyric acid that is produced by bacteria in the catabolism of aspartate, an amino acid with a broad range of concentrations in urine[25]. The excretion of this compound could further point to amino acids as a carbon source for urinary tract bacteria. Alternatively, the peak could be identified as homo-cysteïne, a metabolite occurring as a side product of cellular cystine import. Upon import of cystine, cystine can transfer its disulfide bonds to proteins in the cytoplasm, altering their functionality. To avoid this, the imported cystine is rapidly reduced to cysteine or homo-cysteïne, which can be exported from the

cell[149].

After identifying potential commonly consumed and excreted peaks the cooperative strategy of cross-feeding was analysed (Figure 9C). Results show that *E. coli* could benefit most from cross-feeding, while *E. faecalis* would benefit the least but help others the most. These results contradict the interaction reported by de Vos et al. (2017), where *E. faecalis* was found to benefit the most from others without returning the favour[14] (Figure S7). This could be due to the fact that newly excreted compounds are excluded from this analysis. Additionally, the metabolomics measured here are obtained from monocultures, and *E. faecalis* is known to grow poorly in monoculture in AUM, improving growth when cultured on spent media, pointing to a potential role for newly excreted compounds. In conclusion, these results point to a resource competition between *P. aeruginosa*, *E. coli* and *S. haemolyticus* and point to an important role for newly excreted metabolites in interspecies interactions in the urinary tract community.

Where DART has many benefits, as it is a rapid mass spectrometry (MS) technique that can be used at atmospheric pressure, the metabolomics analysis as performed here has several caveats[150]. Firstly, the range of m/z values that can be detected by DART is limited to 60 - 990, however after filtering only values between 100 - 400 remained. This excludes analysis of larger proteins such as enterobactin or other siderophores, as well as smaller molecules including urea and sulphate. Thus, future metabolomics analysis could use alternative measurement approaches such as ICP-MS to identify metal uptake, as iron and other heavy metals like copper are known to contribute to bacterial virulence, and NMR to identify urea and other small molecules[89,90,151–154]. For larger proteins a method such as MALDI or LESA could be used[155].

A second caveat is that peak picking was done manually, limiting accuracy. Peak picking is necessary as one metabolite can result in multiple peaks. Many different peak picking algorithms are in use, but none are optimised for DART raw data, and all have their own biases. Making peak picking for DART data not optimal, however necessary. Additionally, the use of molecular networking tools to identify classes of products, such as GNPS, are not suitable for DART data[156]. All combined complicating analysis and thereby potentially reducing accuracy of the drawn conclusions.

Thirdly, when using DART data for quantitative analysis, the use of a reference with known metabolite concentrations is common practice[157,158]. The absence hereof in this study complicates analysis. In addition to quantification, identification of peaks is complicated by the fact that not all metabolites in the reference are known due to the presence of yeast extract in the AUM.

Lastly, DART raw data can only be processed using FreeStyle (v1.8 sp2, Thermo Fisher Scientific), a software package only available on Windows. Recompiling this for MacOS resulted in the loss of several functionalities. Identification of peaks also led to erroneous annotations as non-existent molecules, complicating peak annotations for peaks with no prior knowledge, hence only peaks common in the reference were taken into account. For future studies processing the data on a Windows operating system might aid this analysis, as a complete range of functionalities will be available, and might prevent the erroneous annotations of nonexistent molecules.

Conclusion

In this study of five potential uropathogens no signal of selective pressure in the urinary host environment was reported, as no clustering of urine-derived strains was observed. This study does find functional changes in bacteria living in the urinary tract to adapt to the unique urine environment. The urine unique gene clusters mediate functions such as increasing virulence, alternative nutrient uptake as well as antibiotic resistance genes. The presence of urine unique antibiotic resistance genes and the reported overrepresentation thereof in gram-negative bacteria might indicate that antibiotic treatment has an important role in shaping the urine environment and could be interpreted as a warning against over use of antibiotics in UTI treatment. It should however be noted that the urine unique genes were only present in a low number of genomes, these finding must therefore be validated in future studies.

Method

Dataset

Five new isolates of the species *Staphylococcus haemolyticus, Pseudomonas aeruginosa, Enterococcus faecalis, Escherichia coli* and *Klebsiella pneumoniae*, with respective identifiers 36, 9, 18, 20 and 1, were sequenced from hosts with polymicrobial UTI's in a pairwise interaction study[14]. To verify the correct annotations of the species, the average nucleotide identity (ANI) percentage for the newly sequenced genomes in comparison to NCBI reference genomes isolated from urine, were calculated using OrthoANI and genomes were only accepted as the indicated species if the ANI percentage was 95% or above[159]. To increase the dataset size for analysis, all available genome metadata for all genomes of each individual species were downloaded on April 19th 2022 via the PATRIC API[160].

To ensure quality of the genomes for future analysis, several filtering steps were conducted. Firstly, all genomes for which the genome quality was annotated as "Poor" were removed from the dataset. Secondly, the genome length of each genome was checked against the expected range for the species in the NCBI database and Genomes with a genome length outside the expected range were subsequently removed from the dataset[161]. Thereafter, if no CheckM completeness and contamination scores were available in the metadata, these score were calculated using CheckM (v1.2.0) in batches to increase processing speed[162]. Genomes with a CheckM completeness $< 90\%$ or contamination $> 10\%$ were excluded from further analysis[162]. Subsequently, all genomes except those from urine with a GC content outside of the normal distribution (3 * z-score) were removed from the analysis. Lastly, all genomes with an unknown or "other" origin of isolation or that were not marked by PATRIC as "Complete" were excluded from analysis.

Gene predictions

Before further analysis, all genomes were annotated using Prokka (v1.14.6) with genus and species tags corresponding to the species being annotated, for all other options default settings were used[163]. Prokka was chosen over the PATRIC GFF3 annotations as the RAST GFF3 annotation format provided by PATRIC resulted in discrepancy between manually curated and non-curated annotations between strains.

Pan-genome construction and analysis

The pangenome was created using Roary (v3.13.0) with 95% identity[164]. The maximum number of clusters was increased for *E. coli* and *K. pneumoniae* to 100.000 and 90.000 respectively, for all other species the default number of 50.000 clusters was used. Alongside the construction of a pan-genome, Roary was used to create a core gene alignments using MAFFT[165].

Prokka annotations of a high number of genomes could lead to a higher number of annotation errors, either false positive or negative annotations, leading to an inflated accessory genome and a reduced core genome[46,47]. Therefore a second pan-genome construction strategy, utilizing a graph-based approach with the ability to correct for some of the false positive and negative annotation sources was applied using Panaroo (v1.2.10)[166]. Panaroo was run using the GFF3 files created by Prokka on a strict threshold, a core sample threshold of 0.95, removing all invalid genes, and setting the proportion of an accessory gene that must be found in order to consider it a match to 0.75 with a 1000 bp search radius.

After pan-genome construction, the openness of the pan-genome for each species was calculated using Heaps law, $n = kN^{\gamma}$[48]. Heaps law was fit to the number of orthologous groups identified when adding each genome individually by Roary, both k and γ were set as free parameters without constraints.

Phylogenetic analysis

Three different phylogenetic strategies were attempted to identify the role of the urine environment on genome evolution for each individual species. All methods resulted in a Newick tree that was visualized using iTol (v6) and rooted at mid point[167]. Firstly, Mashtree (v1.2.0) was utilised to create a neighbour-joining tree, using Mash distances on all genomic FASTA files for each species[55]. Mashtree was run using accuracy mode, thereby ignoring rare and singleton k-mers, increasing the accuracy of the resulting tree.

The second phylogenetic strategy was based on the evolutionary conserved bacterial 16S RNA. The 16S RNA sequences were predicted from each genomic sequence using Barrnap (v0.9), if more than one 16S sequence was predicted the longest sequence was kept for further analysis[168]. For each species the selected 16S sequences were aligned using ssu-align (v0.1.1) to account for RNA 3D structure and transformed from Stockholm to FASTA file format using ssu-mask[169]. A maximum likelihood tree was inferred from the ssu multiple sequence alignment using IQ-tree (v2.2.0.3) with 1000 bootstrap itera-

tions, 1000 replicates for SH approximate likelihood ratio test and utilizing ModelFinder for optimal model selection, other settings were set to default[170,171].

The last phylogenetic method attempted was focused on the multiple sequence core gene alignment produced by Roary. The core multiple sequence alignment was trimmed using ClipKIT (v1.3.0) and used to infer a maximum likelihood phylogeny using IQ-tree with the same setting previously described[172]. Additionally a maximum likelihood phylogeny approximation was inferred using Fasttree (v2.1) using the clipped core gene nucleotide alignments for all species[57].

Clustering of genomes

To analyse the effect of origins of isolation on possible clustering of the genomes, a permutation test of the metadata on the phylogenies is applied using the Clustering significance test script by Y. Wijesekara[53,54]. Here a p-value of 0.01 was set with 10000 permutation replicates for all species for the 16s, core and Mash phylogenies. Origins of isolation were counted for each cluster proven significant after permutation.

Feature selection

The presence and absence patterns of the orthologous groups of the accessory pan-genome as identified for each species by Roary, were associated to the origin of isolation using Scoary (v1.6.16) ($p < 0.05$), a bacterial pan-GWAS approach[173]. The origins of isolation were grouped, to reduce the number of singleton isolation origins, and for the per species analysis the urine and UTI origins of isolation were both labeled as urine for all species. In a secondary analysis of the urine and UTI origins of isolation in *E. coli* the two origins of isolation were separately annotated where all other annotations and settings remained unchanged.

Hereafter the orthologous groups that were identified as being associated with the urine or UTI origin of isolation, were further filtered for presence rather than absence in these isolation sources. Hereto the final Scoary results for urine were sub-selected for a specificity of 100, resulting in orthologous groups only present in urine, creating a set of urine unique orthologous groups. Additionally the urinary associated orthologous groups were split in two groups, being overrepresented and underrepresented in urine or UTI. Hereto the presence-absence profile of the urinary associated orthologous groups were grouped on origin of isolation and the number of occurrences was divided by to the number of genomes with the origin of isolation, resulting in a relative occurrence for each origin of isolation. Subsequently, the relative occurrences were normalized from 0 and 1 per orthologous group. Orthologous groups with normalized values above 0,5 were classified as overrepresented in urine, groups with values below 0,5 were classified as underrepresented. Both filtering strategies were applied to the Scoary results associating urine and UTI separately in *E. coli*.

Functional annotations

To map the filtered orthologous groups to KEGG pathways, three annotation methods were applied. Firstly, one representative sequence was selected from each filtered orthologous group and concatenated in one FASTA file per species. These files were uploaded to the BlastKOALA and KofamKOALA servers for annotation of KEGG KO identifiers[174,175]. For BlastKOALA annotations taxonomy ID's in accordance with the NCBI Taxonomy browser were used for each species and the "species prokaryotic" KEGG genes database was searched. Of the resulting annotations only the first annotated KO for each sequence was used in further analysis. KofamKOALA annotations were run using the default E-value of 0.01.

Lastly, all sequences from the filtered urine unique, overrepresented and core orthologous group were annotated with EggNOG (v2.1.9) using HMMER searches, with the database for bacteria, taxID 2, for each species[176–178]. A KEGG KO, CAZy and TC identifier was appointed to each orthologous group, if available, based on a majority voting system. In cases where two or more KO, TC or CAZy identifiers had an identical number of votes, these identifiers were both kept and the double annotation was stored in a separate file. For orthologous groups filtered as underrepresented in urine one representative sequence of each group was annotated with the same settings.

Additionally EggNOG annotations were used to analyse the grouping of sequences within the orthologous groups, by analysing the uniformity of the annotations within one orthologous group. Hereto the number of unique COG annotations for each group were counted. Alongside the uniformity of the orthologous groups, splitting of the orthologous groups was also analysed. Hereto the occurrences of each COG assigned to one orthologous group were analysed, if two or more COGs were present within one orthologous group, the occurrences for all COGs was analysed.

Protein-protein interactions using STRING

The selected representative sequence from each filtered orthologous group per species were loaded into STRING for each species and the full species name was selected as Organism[61]. When multiple strains of the organism were present as reference in STRING, the organism with the highest number of matches was chosen. Chosen reference strains were *Staphylococcus haemolyticus JCSC1435*, *Enterococcus faecalis V583*, *Pseudomonas aeruginosa*, *Klebseilla pneumoniae*, *Escherichia coli CFT073*, the resulting networks were manually analysed. To avoid bias between analyses the *Escherichia coli O157H7* strain was selected as reference for the urine vs UTI analysis, albeit not being top rated for UTI. The minimum required interaction score was set as high (0.700) for all analyses.

KEGG pathway analysis

The by EggNOG KO annotated core and urine specific orthologous groups for each species were mapped to KEGG pathways and modules using the KEGG API[179–181]. The KOs, pathways and modules for the urine specific orthologous groups were then

analysed on uniqueness and overlap between species. Additionally, completeness of the urine specific pathways and modules was calculated, for each species separately, with complete pathway and module data from the KEGG API.

Alongside pathway completeness, the connectivity of the urine specific KOs within the pathways was analysed. Hereto for each pathway the KGML was downloaded via the KEGG API, all compounds were extracted and loaded as nodes in a graph. Thereafter, the KO annotations for all reactions were extracted from the KGML file and only reactions with KO annotations present in the annotated orthologous groups were placed as edges in the graph, resulting in a graph with only urine specific reactions. To analyse the urine specific graph connectivity, the size and number of connected components was calculated for each graph.

With the core orthologous groups being present in all genomes of one species the connectivity of the core for each species was also analysed. All KOs of the orthologous groups of the core that had a KO annotation were added to the urine specific graphs of the corresponding pathways as edges. If no urine specific graph was present a new graph for this pathway was initialized as before. Graph connectivity was then reanalysed, recalculating the size and number of the connected components for each pathway.

Biosynthetic gene clusters

To not only study the individual genes and pathways, the Biosyntethic Gene Clusters (BGC) were analysed. Hereto the GBK files for all genomes as annotated by Prokka were used as input files for antiSMASH (v6.1.0), using the settings "–cb-general", "–cb-knownclusters" and "–cb-subclusters" to compare identified clusters against a database of antiSMASH clusters, against known subclusters responsible for synthesising precursors and the MIBiG database respectively[182]. Additionaly the options "–asf", to run active site finder analysis, "–pfam2go", to use Pfam to Gene Ontology mapping module and "–genefinding-tool none", were used.

Predicted BGC gbk files were grouped into Gene Cluster Families (GCF) and classes using BiG-SCAPE (v1.1.5) with settings to include analysis for mixing all classes, including singleton BGCs and including BGCs from MIBiG database and mode set to auto, all other default settings were used[183]. To analyse the BGCs per annotated class and family number, the total number of clusters and the number of clusters with a urine isolation origin per family number were counted. If the number of clusters in one family was the same as the number of clusters with a urine origin of isolation this cluster was labeled as urine specific.

Subsequently, the expected number of urine-related clusters in a given family was calculated. Firstly, the percentage of clusters identified in urine genomes in the total number of identified clusters from all isolation origins was calculated for each species. Secondly, this percentage was used as an expected urine percentage for all GCF and BGC classes for the specific species, giving the expected number of urine-related clusters of that family or class. A Mann Whitney U test was used to test statistical differences between the actual number and the expected number of clusters with urinary origin for class of BGCs present per species.

Additionally, families within each BGC class per species that exhibited an overrepresentation of urinary related clusters, but were not exclusive to urine, were selected by applying a threshold of a 50% increase of the actual number of clusters with urinary origin compared to the expected number of urine-related clusters. Predicted products for all families within one class were analysed and potential products were inferred in correspondence with the whole pan-genome mapped to KEGG[179,181].

Metabolomics

DART measurements

To measure the changes in metabolites when the five species grow in AUM, reference and spent media, after 48h of bacterial growth, were measured using direct analysis in real time (DART). AUM and spent media were obtained for *S. haemolyticus, P. aeruginosa, E. faecalis, E. coli* and *K. pneumoniae*, with respective identifiers 36, 9, 18, 20 and 1, as described in a previous study[14]. For each species, fresh AUM was measured as a reference in positive and negative mode, whereafter the spent medium was twice measured in negative mode succeeded by two measurements in positive mode. Measurement times are shown in Table 5. No timestamps were noted for *K. pneumoniae* AUM reference measurements and positive mode spent medium measurements, *K. pneumoniae* was therefore excluded from further metabolomics analysis.

Table 5: Times of DART measurements in fractions of minutes for the measured species.

Species	**Positive AUM**	**Negative AUM**	**Negative spent medium 1**	**Negative spent medium 2**	**Positive spent medium 1**	**Positive spent medium 2**
S. haemolyticus	0.56	2.6	4.5	6.1	8.55	10.8
P. aeruginosa	1.03	2.7	4.4	6.2	8.0	10.15
E. faecalis	0.66	3.5	5.6	7.5	10.9	12.9
E. coli	1.03	4.9	5.7	6.2	9.7	11.6

DART data analysis

The by DART produced .RAW files were analysed using FreeStyle (v1.8 sp2, Thermo Fisher Scientific). Firstly, an average spectrum was created of an interval of 0.15 minute fractions, corresponding to 33 scans, for each measurement, and peak lists of these average spectra were created and saved as csv files.

Secondly, the average spectrum data were filtered on mass accuracy, all m/z values with relative intensity 100 were rounded from 12 to 0 decimals, calculating the standard deviation between the m/z values at each number of decimals. The m/z values for all measurements were rounded to 3 decimals, as the standard deviation between the maximum relative intensity peaks at this m/z accuracy was zero. When two peaks with

the same rounded m/z value were present, the highest relative intensity was assigned to the m/z value. Analysis of these maximum relative intensity peaks revealed that the m/z values at relative intensity 100 for two *E. coli* measurements were different from all other measurements, these measurements, being "Negative spent medium 2" and "Positive spent medium 1", where therefore excluded from further analysis. Average spectra of spent media of one mode for each species were joined on m/z values and relative intensity was averaged if m/z value was detected in both measurements, otherwise the once measured intensity value was assigned to m/z value.

Alongside the m/z accuracy filtering, peaks were filtered on relative intensity values to reduce noise. Histograms of the m/z filtered peaks of the reference AUM spectra were plotted and showed a drop in peak count after relative intensity 0.006 except for *E. coli*. Thus, the assumption was made that the noise level was 0.006 in most measurements of this experiment and therefore, all peaks were filtered with a relative intensity threshold of 0.006 (Figure S6).

The resulting reference peaks were further filtered by selection of peaks with m/z values common in all species' positive and negative mode reference measurements separately. Subsequently the difference in relative intensities between average spent medium and common reference peaks was calculated for each mode individually, if no peak in spent medium was detected at specific m/z value and mode, the spent medium relative intensity was assumed to be 0, and the compound was assumed to be depleted. Similarly if no reference peak was detected at specific m/z value and mode the reference relative intensity was assumed to be 0, and the compound was assumed to be newly excreted. Per measurement mode and per species the differences were sub-selected between -0.2 and 0.2 to exclude long tails from the distribution and the standard deviations were calculated. Changes were labeled as consumed when the relative intensity difference of average spent medium minus the reference were below minus 3 times the calculated standard deviation, and labeled as excreted when this difference was greater than 3 standard deviations. Common and unique peaks for each species were identified, for the depleted, consumed, excreted and newly excreted category and mode separately.

DART peak identification

Three methods were applied to identify the compounds represented by the peaks in the spectra. Firstly, all compounds known to be present in AUM as described in the original methods were attempted to be linked to a m/z peak, using their known chemical structure, the mode by which they were detected and a conversion table to convert the known masses to possible m/z values, $M - H$ conversions were used for negative mode, $M + H$ and $M + NH_4$[27,184].

Secondly, for peaks left unannotated by the first strategy, an additional annotation method was attempted. This method utilises the "Elemental Composition" function of FreeStyle to predict possible molecular formulas for each peak. The peak m/z value was entered in the Mass tab, using a mass tolerance of 5 ppm, a charge of 1 and selecting the top 5 predicted candidates. Candidate molecular formulas were then based on their ranking checked and against PubChem. PubChem structures and annotations were

reviewed based on likeliness to be present in yeast extract and only molecular formulas consisting of covalent bound molecules were allowed[161].

In parallel to this second method a third method was utilized, here all unannotated m/z peak values were converted to possible masses using the previously used conversion based on their mode and the assumption was made that most unannotated peaks originated from yeast extract present in the AUM. The predicted monoisotopic masses were checked against the The Yeast Metabolome Database (YMDB), unless they were common in *E. coli* then matching metabolites were searched in the E. coli Metabolome Database (ECMDB) as this second database specifies compounds present in *E. coli*, reducing false positive or erroneous annotations[185–188]. the converted mass values were entered as a mass search query for the monoisotopic mass, allowing a mass change of 1%. These possible yeast metabolites were cross-referenced against the molecular formulas as predicted by FreeStyle.

Data and Code availability

Code to reproduce the results and PATRIC genome identifiers are available at `https://gitlab.com/LMSpekking/uti_project.git`

References

1. Hilt, E. E. *et al.* Urine Is Not Sterile: Use of Enhanced Urine Culture Techniques To Detect Resident Bacterial Flora in the Adult Female Bladder. *Journal of Clinical Microbiology* **52,** 871–876 (2014).
2. Dubourg, G. *et al.* Deciphering the Urinary Microbiota Repertoire by Culturomics Reveals Mostly Anaerobic Bacteria From the Gut. *Frontiers in Microbiology* **11** (2020).
3. Wolfe, A. J. *et al.* Evidence of Uncultivated Bacteria in the Adult Female Bladder. *Journal of Clinical Microbiology* **50,** 1376–1383 (2012).
4. Whiteside, S. A., Razvi, H., Dave, S., Reid, G. & Burton, J. P. The microbiome of the urinary tract–a role beyond infection. *Nature Reviews Urology* **12,** 81–91 (2015).
5. Lewis, D. A. *et al.* The human urinary microbiome; bacterial DNA in voided urine of asymptomatic adults. *Frontiers in Cellular and Infection Microbiology* **3** (2013).
6. Morand, A. *et al.* Human Bacterial Repertoire of the Urinary Tract: a Potential Paradigm Shift. *Journal of Clinical Microbiology* **57,** e00675–18 (2019).
7. Stamm, W. E. & Norrby, S. R. Urinary tract infections: disease panorama and challenges. *The Journal of Infectious Diseases* **183 Suppl 1,** S1–4 (2001).
8. Foxman, B. & Brown, P. Epidemiology of urinary tract infections: transmission and risk factors, incidence, and costs. *Infectious Disease Clinics of North America* **17,** 227–241 (2003).

9. Chardavoyne, P. C. & Kasmire, K. E. Appropriateness of Antibiotic Prescriptions for Urinary Tract Infections. *Western Journal of Emergency Medicine* **21,** 633–639 (2020).

10. Ben, Y. *et al.* Human health risk assessment of antibiotic resistance associated with antibiotic residues in the environment: A review. *Environmental Research* **169,** 483–493 (2019).

11. Flores-Mireles, A. L., Walker, J. N., Caparon, M. & Hultgren, S. J. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature reviews. Microbiology* **13,** 269–284 (2015).

12. Tabibian, J. H. *et al.* Uropathogens and Host Characteristics. *Journal of Clinical Microbiology* **46,** 3980–3986 (2008).

13. Kline, K. A. & Bowdish, D. M. E. Infection in an aging population. *Current Opinion in Microbiology* **29,** 63–67 (2016).

14. De Vos, M. G. J., Zagorski, M., McNally, A. & Bollenbach, T. Interaction networks, ecological stability, and collective antibiotic tolerance in polymicrobial infections. *PNAS* **114,** 10666–10671 (2017).

15. Haider, J. S. Frequency of Urinary Tract Bacterial Infection and their Susceptibility Patterns among Hemodialysis Patients in Zliten Hospital. *Journal of Microbiology & Experimentation* **3** (2016).

16. Fourcade, C., Canini, L., Lavigne, J.-P. & Sotto, A. A comparison of monomicrobial versus polymicrobial Enterococcus faecalis bacteriuria in a French University Hospital. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology* **34,** 1667–1673 (2015).

17. Cottalorda, A. *et al.* Within-Host Microevolution of Pseudomonas aeruginosa Urinary Isolates: A Seven-Patient Longitudinal Genomic and Phenotypic Study. *Frontiers in Microbiology* **11,** 611246 (2021).

18. Eltwisy, H. O., Twisy, H. O., Hafez, M. H., Sayed, I. M. & El-Mokhtar, M. A. Clinical Infections, Antibiotic Resistance, and Pathogenesis of Staphylococcus haemolyticus. *Microorganisms* **10,** 1130 (2022).

19. Ruiz, N. & Silhavy, T. J. How Escherichia coli Became the Flagship Bacterium of Molecular Biology. *Journal of Bacteriology* **204,** e00230–22 (2022).

20. Ronald, A. The etiology of urinary tract infection: traditional and emerging pathogens. *Disease-a-month: DM* **49,** 71–82 (2003).

21. Croxall, G. *et al.* Increased human pathogenic potential of Escherichia coli from polymicrobial urinary tract infections in comparison to isolates from monomicrobial culture samples. *Journal of Medical Microbiology* **60,** 102–109 (2011).

22. Vayssier-Taussat, M. *et al.* Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. *Frontiers in Cellular and Infection Microbiology* **4,** 29 (2014).

23. Reitzer, L. & Zimmern, P. Rapid Growth and Metabolism of Uropathogenic Escherichia coli in Relation to Urine Composition. *Clinical Microbiology Reviews* **33,** e00101–19 (2019).

24. Clarkson, M. R., Magee, C. N. & Brenner, B. M. in *Pocket Companion to Brenner and Rector's The Kidney (Eighth Edition)* (eds Clarkson, M. R., Magee, C. N. & Brenner, B. M.) Eighth Edition, xi (W.B. Saunders, Philadelphia, 2011). ISBN: 978-1-4160-6640-8.

25. Bouatra, S. *et al.* The Human Urine Metabolome. *PLoS ONE* **8,** e73076 (2013).

26. Alteri, C. J., Hagan, E. C., Sivick, K. E., Smith, S. N. & Mobley, H. L. T. Mucosal Immunization with Iron Receptor Antigens Protects against Urinary Tract Infection. *PLoS Pathogens* **5,** e1000586 (2009).

27. Brooks, T. & Keevil, C. A simple artificial urine for the growth of urinary pathogens. *Letters in Applied Microbiology* **24,** 203–206 (1997).

28. Snyder, J. A. *et al.* Transcriptome of Uropathogenic Escherichia coli during Urinary Tract Infection. *Infection and Immunity* **72,** 6373–6381 (2004).

29. Alteri, C. J. & Mobley, H. L. T. Quantitative Profile of the Uropathogenic Escherichia coli Outer Membrane Proteome during Growth in Human Urine. *Infection and Immunity* **75,** 2679–2688 (2007).

30. Brumfitt, W., Hamilton-Miller, J. M., Cooper, J. & Raeburn, A. Relationship of urinary pH to symptoms of 'cystitis'. *Postgraduate Medical Journal* **66,** 727–729 (1990).

31. Ipe, D. S., Horton, E. & Ulett, G. C. The Basics of Bacteriuria: Strategies of Microbes for Persistence in Urine. *Frontiers in Cellular and Infection Microbiology* **6,** 14 (2016).

32. Kucheria, R., Dasgupta, P., Sacks, S., Khan, M. & Sheerin, N. Urinary tract infections: new insights into a common problem. *Postgraduate Medical Journal* **81,** 83–86 (2005).

33. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *PNAS* **102,** 13950–13955 (2005).

34. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Current Opinion in Genetics & Development* **15,** 589–594 (2005).

35. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Current Opinion in Microbiology* **23,** 148–154 (2015).

36. Yang, M.-R. & Wu, Y.-W. Enhancing predictions of antimicrobial resistance of pathogens by expanding the potential resistance gene repertoire using a pan-genome-based feature selection approach. *BMC Bioinformatics* **23,** 131 (2022).

37. Souza Costa, S., Guimarães, L. C., Silva, A., Castro Soares, S. & Azevedo Baraúna, R. First Steps in the Analysis of Prokaryotic Pan-Genomes. *Bioinformatics and Biology Insights* **14,** 1–9 (2020).

38. Yang, T. & Gao, F. High-quality pan-genome of *Escherichia coli* generated by excluding confounding and highly similar strains reveals an association between unique gene clusters and genomic islands. *Briefings in Bioinformatics* **23,** bbac283 (2022).

39. Hesse, C. *et al.* Genome-based evolutionary history of Pseudomonas spp. *Environmental Microbiology* **20.** ISSN: 1462-2912 (2018).

40. He, Q. *et al.* Comparative genomic analysis of Enterococcus faecalis: insights into their environmental adaptations. *BMC Genomics* **19,** 527. ISSN: 1471-2164 (2018).

41. Pain, M., Hjerde, E., Klingenberg, C. & Cavanagh, J. P. Comparative Genomic Analysis of Staphylococcus haemolyticus Reveals Key to Hospital Adaptation and Pathogenicity. *Frontiers in Microbiology* **10** (2019).

42. Flores-Valdez, M. *et al.* Whole Genome Sequencing of Pediatric Klebsiella pneumoniae Strains Reveals Important Insights Into Their Virulence-Associated Traits. *Frontiers in Microbiology* **12.** ISSN: 1664-302X (2021).

43. Connor, T. R. *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri. *eLife* **4,** e07335. ISSN: 2050-084X (2015).

44. Brenner, D. J., Fanning, G. R., Steigerwalt, A. G., Ørskov, I. & Ørskov, F. Polynucleotide Sequence Relatedness Among Three Groups of Pathogenic Escherichia coli Strains. *Infection and Immunity* **6,** 308–315. ISSN: 0019-9567 (1972).

45. Khot, Prasanna D. & Fisher, Mark A. Novel Approach for Differentiating Shigella Species and Escherichia coli by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry. *Journal of Clinical Microbiology* **51,** 3711–3716 (2013).

46. Denton, J. F. *et al.* Extensive error in the number of genes inferred from draft genome assemblies. *PLoS computational biology* **10,** e1003998 (2014).

47. Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biology* **20,** 92 (2019).

48. Heaps, H. S. *Information retrieval, computational and theoretical aspects* ISBN: 978-0-12-335750-2 (Academic Press, New York, 1978).

49. Von Meijenfeldt, F. A. B., Hogeweg, P. & Dutilh, B. E. A social niche breadth score reveals niche range strategies of generalists and specialists. *Nature Ecology & Evolution* (2023).

50. Tantoso, E. *et al.* To kill or to be killed: pangenome analysis of Escherichia coli strains reveals a tailocin specific for pandemic ST131. *BMC Biology* **20,** 146 (2022).

51. Park, S.-C., Lee, K., Kim, Y. O., Won, S. & Chun, J. Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. *Frontiers in Microbiology* **10,** 834 (2019).

52. Wyres, K. L. *et al.* Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. *Pros Genetics* **15,** e1008114 (2019).

53. Wijesekara, Y. *Clustering-significance-test* July 2023. `https://github.com/Yasas1994/Clustering-significance-test` (2023).

54. Balaban, Metin, Moshiri, Niema, Mai, Uyen, Jia, Xingfan & Mirarab, Siavash. TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS ONE* **14,** e0221068 (2019).

55. Katz, L. S. *et al.* Mashtree: a rapid comparison of whole genome sequence files. *Journal of open source software* **4,** 10.21105/joss.01762 (2019).

56. Zhou, X., Shen, X.-X., Hittinger, C. T. & Rokas, A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution* **35,** 486–503 (2018).

57. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5,** e9490 (2010).

58. D'Souza, G. *et al.* Less Is More: Selective Advantages Can Explain the Prevalent Loss of Biosynthetic Genes in Bacteria. *Evolution* **68,** 2559–2570 (2014).

59. Zdziarski, J., Svanborg, C., Wullt, B., Hacker, J. & Dobrindt, U. Molecular Basis of Commensalism in the Urinary Tract: Low Virulence or Virulence Attenuation? *Infection and Immunity* **76,** 695–703 (2008).

60. Zdziarski, J. *et al.* Host Imprints on Bacterial Genomes—Rapid, Divergent Evolution in Individual Patients. *PLoS Pathogens* **6,** e1001078 (2010).

61. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* **51,** D638–D646 (2022).

62. Lindahl, T. An N-Glycosidase from Escherichia coli That Releases Free Uracil from DNA Containing Deaminated Cytosine Residues. *PNAS* **71,** 3649–3653 (1974).

63. Venkatesh, J., Kumar, P., Krishna, P. S. M., Manjunath, R. & Varshney, U. Importance of uracil DNA glycosylase in Pseudomonas aeruginosa and Mycobacterium smegmatis, G+C-rich bacteria, in mutation prevention, tolerance to acidified nitrite, and endurance in mouse macrophages. *The Journal of Biological Chemistry* **278,** 24350–24358 (2003).

64. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287,** 560–561 (1980).

65. Fang, F. C. & Vázquez-Torres, A. Reactive Nitrogen Species in Host-Bacterial Interactions. *Current opinion in immunology* **60,** 96–102 (2019).

66. Lacerda Mariano, L. *et al.* Functionally distinct resident macrophage subsets differentially shape responses to infection in the bladder. *Science Advances* **6,** eabc5739 (2020).

67. Lund, G. S. & Wolf, C. G. L. The Glucose Content of Normal Urine. *Biochemical Journal* **19,** 538–540 (1925).

68. Brinster, S., Furlan, S. & Serror, P. C-Terminal WxL Domain Mediates Cell Wall Binding in Enterococcus faecalis and Other Gram-Positive Bacteria. *Journal of Bacteriology* **189,** 1244–1253 (2007).

69. Brinster, S. *et al.* Enterococcal Leucine-Rich Repeat-Containing Protein Involved in Virulence and Host Inflammatory Response. *Infection and Immunity* **75,** 4463–4471 (2007).

70. Jamet, A. *et al.* The Enterococcus faecalis virulence factor ElrA interacts with the human Four-and-a-Half LIM Domains Protein 2. *Scientific Reports* **7,** 1–13 (2017).

71. Katongole, P., Nalubega, F., Florence, N. C., Asiimwe, B. & Andia, I. Biofilm formation, antimicrobial susceptibility and virulence genes of Uropathogenic Escherichia coli isolated from clinical isolates in Uganda. *BMC Infectious Diseases* **20,** 453 (2020).

72. Ballash, G. A. *et al.* Pathogenomics and clinical recurrence influence biofilm capacity of Escherichia coli isolated from canine urinary tract infections. *PLoS ONE* **17,** e0270461 (2022).

73. Soto, S. M. *et al.* Implication of biofilm formation in the persistence of urinary tract infection caused by uropathogenic Escherichia coli. *Clinical Microbiology and Infection* **12,** 1034–1036 (2006).

74. Madsen, J. S., Burmølle, M., Hansen, L. H. & Sørensen, S. J. The interconnection between biofilm formation and horizontal gene transfer. *FEMS immunology and medical microbiology* **65,** 183–195 (2012).

75. Niveditha, S., Pramodhini, S., Umadevi, S., Kumar, S. & Stephen, S. The Isolation and the Biofilm Formation of Uropathogens in the Patients with Catheter Associated Urinary Tract Infections (UTIs). *Journal of Clinical and Diagnostic Research* **6,** 1478–1482 (2012).

76. Juhas, M. Horizontal gene transfer in human pathogens. *Critical Reviews in Microbiology* **41,** 101–108 (2015).

77. Van der Zee, A. *et al.* Spread of Carbapenem Resistance by Transposition and Conjugation Among Pseudomonas aeruginosa. *Frontiers in Microbiology* **9,** 2057 (2018).

78. Wozniak, R. A. F. & Waldor, M. K. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews. Microbiology* **8,** 552–563 (2010).

79. Nuttall, K. L. Interpreting Mercury in Blood and Urine of Individual Patients. *Annals of Clinical & Laboratory Science* **34,** 235–250 (2004).

80. Yazdankhah, S., Skjerve, E. & Wasteson, Y. Antimicrobial resistance due to the content of potentially toxic metals in soil and fertilizing products. *Microbial Ecology in Health and Disease* **29,** 1548248 (2018).

81. Schlüter, A. *et al.* The 64 508 bp IncP-1beta antibiotic multiresistance plasmid pB10 isolated from a waste-water treatment plant provides evidence for recombination between members of different branches of the IncP-1beta group. *Microbiology (Reading, England)* **149,** 3139–3153 (2003).

82. Tan, D. *et al.* Characterization of Klebsiella pneumoniae ST11 Isolates and Their Interactions with Lytic Phages. *Viruses* **11,** 1080 (2019).

83. Stanton, I. C., Murray, A. K., Zhang, L., Snape, J. & Gaze, W. H. Evolution of antibiotic resistance at low antibiotic concentrations including selection below the minimal selective concentration. *Communications Biology* **3,** 1–11 (2020).

84. Gullberg, E. *et al.* Selection of resistant bacteria at very low antibiotic concentrations. *PLoS pathogens* **7,** e1002158 (2011).

85. Reddy, K. V. R., Yedery, R. D. & Aranha, C. Antimicrobial peptides: premises and promises. *International Journal of Antimicrobial Agents* **24,** 536–547 (2004).

86. Sutton, J. D. *et al.* Oral -Lactam Antibiotics vs Fluoroquinolones or Trimethoprim-Sulfamethoxazole for Definitive Treatment of Enterobacterales Bacteremia From a Urine Source. *JAMA Network Open* **3,** e2020166 (2020).

87. Vejborg, R. M. *et al.* Identification of genes important for growth of asymptomatic bacteriuria Escherichia coli in urine. *Infection and Immunity* **80,** 3179–3188 (2012).

88. Chung The, H. *et al.* A high-resolution genomic analysis of multidrug-resistant hospital outbreaks of Klebsiella pneumoniae. *EMBO Molecular Medicine* **7,** 227–239 (2015).

89. Barber, M. F. & Elde, N. C. Buried Treasure: Evolutionary Perspectives on Microbial Iron Piracy. *Trends in genetics: TIG* **31,** 627–636 (2015).

90. Cassat, J. E. & Skaar, E. P. Iron in Infection and Immunity. *Cell host & microbe* **13,** 509–519 (2013).

91. Juarez, G. E. & Galván, E. M. Role of nutrient limitation in the competition between uropathogenic strains of Klebsiella pneumoniae and Escherichia coli in mixed biofilms. *Biofouling* **34,** 287–298 (2018).

92. Vigil, P. D. *et al.* Presence of Putative Repeat-in-Toxin Gene tosA in Escherichia coli Predicts Successful Colonization of the Urinary Tract. *mBio* **2,** e00066–11 (2011).

93. Henderson, J. P. *et al.* Quantitative Metabolomics Reveals an Epigenetic Blueprint for Iron Acquisition in Uropathogenic Escherichia coli. *PLoS Pathogens* **5,** e1000305 (2009).

94. Sybesma, W. *et al.* Bacteriophages as Potential Treatment for Urinary Tract Infections. *Frontiers in Microbiology* **7,** 465 (2016).

95. Chegini, Z. *et al.* Bacteriophage therapy for inhibition of multi drug-resistant uropathogenic bacteria: a narrative review. *Annals of Clinical Microbiology and Antimicrobials* **20,** 30 (2021).

96. Anfora, A. T., Halladin, D. K., Haugen, B. J. & Welch, R. A. Uropathogenic Escherichia coli CFT073 Is Adapted to Acetatogenic Growth but Does Not Require Acetate during Murine Urinary Tract Infection. *Infection and Immunity* **76,** 5760–5767 (2008).

97. Racine, S. X. *et al.* N-acetyl functions and acetate detected by nuclear magnetic resonance spectroscopy of urine to detect renal dysfunction following aminoglycoside and/or glycopeptide antibiotic therapy. *Nephron. Physiology* **97,** p53–57 (2004).

98. Cobden, I., Hamilton, I., Rothwell, J. & Axon, A. T. Cellobiose/mannitol test: physiological properties of probe molecules and influence of extraneous factors. *Clinica Chimica Acta* **148,** 53–62 (1985).

99. Wu, M.-C., Chen, Y.-C., Lin, T.-L., Hsieh, P.-F. & Wang, J.-T. Cellobiose-Specific Phosphotransferase System of Klebsiella pneumoniae and Its Importance in Biofilm Formation and Virulence. *Infection and Immunity* **80,** 2464–2472 (2012).

100. Lane, M. C. *et al.* Role of motility in the colonization of uropathogenic Escherichia coli in the urinary tract. *Infection and Immunity* **73,** 7644–7656 (2005).

101. Wright, K. J., Seed, P. C. & Hultgren, S. J. Uropathogenic Escherichia coli flagella aid in efficient urinary tract colonization. *Infection and Immunity* **73,** 7657–7668 (2005).

102. Lane, M. C., Alteri, C. J., Smith, S. N. & Mobley, H. L. T. Expression of flagella is coincident with uropathogenic Escherichia coli ascension to the upper urinary tract. *PNAS* **104,** 16669–16674 (2007).

103. Guttenplan, S. B. & Kearns, D. B. Regulation of flagellar motility during biofilm formation. *FEMS microbiology reviews* **37,** 849–871 (2013).

104. Rosen, D. A., Hooton, T. M., Stamm, W. E., Humphrey, P. A. & Hultgren, S. J. Detection of intracellular bacterial communities in human urinary tract infection. *PLoS medicine* **4,** e329 (2007).

105. De Nisco, N. J. *et al.* Direct Detection of Tissue-Resident Bacteria and Chronic Inflammation in the Bladder Wall of Postmenopausal Women with Recurrent Urinary Tract Infection. *Journal of Molecular Biology* **431,** 4368–4379 (2019).

106. Hirakawa, H., Suzue, K., Kurabayashi, K. & Tomita, H. The Tol-Pal System of Uropathogenic Escherichia coli Is Responsible for Optimal Internalization Into and Aggregation Within Bladder Epithelial Cells, Colonization of the Urinary Tract of Mice, and Bacterial Motility. *Frontiers in Microbiology* **10,** 1827 (2019).

107. Anderson, G. G., Martin, S. M. & Hultgren, S. J. Host subversion by formation of intracellular bacterial communities in the urinary tract. *Microbes and Infection* **6,** 1094–1101 (2004).

108. Song, J. *et al.* TLR4-mediated expulsion of bacteria from infected bladder epithelial cells. *PNAS* **106,** 14966–14971 (2009).

109. Ewing W.H. *Edwards and Ewing's Identification of Enterobacteriaceae.* 4th ed. (Elsevier, New York, 1986).

110. Carabarin-Lima, A. *et al.* First evidence of polar flagella in Klebsiella pneumoniae isolated from a patient with neonatal sepsis. *Journal of Medical Microbiology* **65,** 729–737 (2016).

111. Schwan, W. R. Flagella allow uropathogenic Escherichia coli ascension into murine kidneys. *International journal of medical microbiology* **298,** 441–447 (2008).

112. Bermingham, A. & Derrick, J. P. The folic acid biosynthesis pathway in bacteria: evaluation of potential for antibacterial drug discovery. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* **24,** 637–648 (2002).

113. in. *Veterinary Medicine (Eleventh Edition)* (eds Constable, P. D., Hinchcliff, K. W., Done, S. H. & Grünberg, W.) 153–174 (W.B. Saunders, 2017). ISBN: 978-0-7020-5246-0.

114. Burman, L. G. The antimicrobial activities of trimethoprim and sulfonamides. *Scandinavian Journal of Infectious Diseases* **18,** 3–13 (1986).

115. Lavanya, R. Sulphonamides: A Pharmaceutical Review. *International Journal of Pharmaceutical Science Invention* **6,** 1–3.

116. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research* **51,** D690–D699 (2023).

117. Griffith, D. Urease stones. *Urological Research* **7** (1979).

118. Li, X. *et al.* Visualization of Proteus mirabilis within the Matrix of Urease-Induced Bladder Stones during Experimental Urinary Tract Infection. *Infection and Immunity* **70,** 389–394 (2002).

119. Hedelin, H. Uropathogens and urinary tract concretion formation and catheter encrustations. *International Journal of Antimicrobial Agents* **19,** 484–487 (2002).

120. Nielubowicz, G. R. & Mobley, H. L. T. Host–pathogen interactions in urinary tract infection. *Nature Reviews Urology* **7,** 430–441 (2010).

121. Hollenbeck, B. L. & Rice, L. B. Intrinsic and acquired resistance mechanisms in enterococcus. *Virulence* **3,** 421–569 (2012).

122. Peters, B. M., Jabra-Rizk, M. A., O'May, G. A., Costerton, J. W. & Shirtliff, M. E. Polymicrobial interactions: impact on pathogenesis and human disease. *Clinical Microbiology Reviews* **25,** 193–213 (2012).

123. Gaston, J. R., Johnson, A. O., Bair, K. L., White, A. N. & Armbruster, C. E. Polymicrobial Interactions in the Urinary Tract: Is the Enemy of My Enemy My Friend? *Infection and Immunity* **89,** e00652–20 (2021).

124. Byrd, A. L. & Segre, J. A. Adapting Koch's postulates. *Science* **351,** 224–226 (2016).

125. Fredrickson, A. & Stephanopoulos, G. Microbial Competition. *Science* **213,** 972–979 (1981).

126. Straight, P. D. & Kolter, R. Interspecies chemical communication in bacterial development. *Annual Review of Microbiology* **63,** 99–118 (2009).

127. Chevrette, M. G. *et al.* Evolutionary dynamics of natural product biosynthesis in bacteria. *Natural Product Reports* **37,** 566–599 (2020).

128. George, E. A. & Muir, T. W. Molecular Mechanisms of agr Quorum Sensing in Virulent Staphylococci. *ChemBioChem* **8,** 847–855 (2007).

129. Roux, A., Payne, S. M. & Gilmore, M. S. Microbial Telesensing: Probing the Environment for Friends, Foes, and Food. *Cell Host & Microbe* **6,** 115–124 (2009).

130. Russo, T. A. *et al.* IroN functions as a siderophore receptor and is a urovirulence factor in an extraintestinal pathogenic isolate of Escherichia coli. *Infection and Immunity* **70,** 7156–7160 (2002).

131. Hagan, E. C. & Mobley, H. L. T. Haem acquisition is facilitated by a novel receptor Hma and required by uropathogenic Escherichia coli for kidney infection. *Molecular Microbiology* **71,** 79–91 (2009).

132. Torres, A. G., Redford, P., Welch, R. A. & Payne, S. M. TonB-dependent systems of uropathogenic Escherichia coli: aerobactin and heme transport and TonB are required for virulence in the mouse. *Infection and Immunity* **69,** 6179–6185 (2001).

133. Johnson, J. R. *et al.* The IrgA homologue adhesin Iha is an Escherichia coli virulence factor in murine urinary tract infection. *Infection and Immunity* **73,** 965–971 (2005).

134. Russo, T. A., Carlino, U. B. & Johnson, J. R. Identification of a new iron-regulated virulence gene, ireA, in an extraintestinal pathogenic isolate of Escherichia coli. *Infection and Immunity* **69,** 6209–6216 (2001).

135. Sagot, B. *et al.* Osmotically induced synthesis of the dipeptide N-acetylglutaminylglutamine amide is mediated by a new pathway conserved among bacteria. *PNAS* **107,** 12652–12657 (2010).

136. Siregar, P. & Setiati, S. Urine osmolality in the elderly. *Acta Medica Indonesiana* **42,** 24–26 (2010).

137. Dias, F. C., Boilesen, S. N., Tahan, S., Melli, L. C. & Morais, M. B. Prevalence of voluntary dehydration according to urine osmolarity in elementary school students in the metropolitan region of São Paulo, Brazil. *Clinics (Sao Paulo, Brazil)* **74,** e903 (2019).

138. Inoue, Y. *et al.* The antibacterial effects of terpene alcohols on Staphylococcus aureus and their mode of action. *FEMS microbiology letters* **237,** 325–331 (2004).

139. Nogueira, J. O. E. *et al.* Mechanism of action of various terpenes and phenylpropanoids against Escherichia coli and Staphylococcus aureus. *FEMS microbiology letters* **368,** fnab052 (2021).

140. Ilić, B. S., Kocić, B. D., Ćirić, V. M., Cvetković, O. G. & Miladinović, D. L. An *In Vitro* Synergistic Interaction of Combinations of *Thymus glabrescens* Essential Oil and Its Main Constituents with Chloramphenicol. *The Scientific World Journal* **2014,** e826219 (2014).

141. Miladinović, D. L., Ilić, B. S., Kocić, B. D. & Miladinović, M. D. An in vitro antibacterial study of savory essential oil and geraniol in combination with standard antimicrobials. *Natural Product Communications* **9,** 1629–1632 (2014).

142. Rahman, I. R. *et al.* Substrate Recognition by the Class II Lanthipeptide Synthetase HalM2. *ACS chemical biology* **15,** 1473–1486 (2020).

143. Wang, J., Ge, X., Zhang, L., Teng, K. & Zhong, J. One-pot synthesis of class II lanthipeptide bovicin HJ50 via an engineered lanthipeptide synthetase. *Scientific Reports* **6,** 38630 (2016).

144. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews. Microbiology* **8,** 15–25 (2010).

145. Chevrette, M. G. *et al.* Microbiome composition modulates secondary metabolism in a multispecies bacterial community. *PNAS* **119,** e2212930119 (2022).

146. Alteri, C. J. & Mobley, H. L. T. Metabolism and Fitness of Urinary Tract Pathogens. *Microbiology spectrum* **3,** 10.1128/microbiolspec.MBP–0016–2015 (2015).

147. Alteri, C. J., Smith, S. N. & Mobley, H. L. T. Fitness of Escherichia coli during Urinary Tract Infection Requires Gluconeogenesis and the TCA Cycle. *PLoS Pathogens* **5,** e1000448 (2009).

148. Paudel, S., Bagale, K., Patel, S., Kooyers, N. J. & Kulkarni, R. Human Urine Alters Methicillin-Resistant Staphylococcus aureus Virulence and Transcriptome. *Applied and Environmental Microbiology* **87,** e00744–21. eprint: `https://journals.asm.org/doi/pdf/10.1128/AEM.00744-21`. `https://journals.asm.org/doi/abs/10.1128/AEM.00744-21` (2021).

149. Korshunov, S., Imlay, K. R. C. & Imlay, J. A. Cystine import is a valuable but risky process whose hazards Escherichia coli minimizes by inducing a cysteine exporter. *Molecular Microbiology* **113,** 22–39 (2020).

150. Cody, R. B., Laramée, J. A. & Durst, H. D. Versatile New Ion Source for the Analysis of Materials in Open Air under Ambient Conditions. *Analytical Chemistry* **77,** 2297–2302 (2005).

151. Choe, K.-Y. & Gajek, R. Determination of trace elements in human urine by ICP-MS using sodium chloride as a matrix-matching component in calibration. *Analytical Methods* **8,** 6754–6763. https://pubs.rsc.org/en/content/articlelanding/2016/ay/c6ay01877g (2016).

152. Bogaerts, A. & Aghaei, M. Inductively coupled plasma-mass spectrometry: insights through computer modeling. *Journal of Analytical Atomic Spectrometry* **32,** 233–261 (2017).

153. Hyre, A. N., Kavanagh, K., Kock, N. D., Donati, G. L. & Subashchandrabose, S. Copper Is a Host Effector Mobilized to Urine during Urinary Tract Infection To Impair Bacterial Colonization. *Infection and Immunity* **85,** e01041–16 (2017).

154. Liu, L., Mo, H., Wei, S. & Raftery, D. Quantitative analysis of urea in human urine and serum by 1H nuclear magnetic resonance. *The Analyst* **137,** 595–600. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4758351/ (2012).

155. Vimer, S., Ben-Nissan, G. & Sharon, M. Mass Spectrometry Analysis of Intact Proteins from Crude Samples. *Analytical Chemistry* **92,** 12741–12749 (2020).

156. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34,** 828–837 (2016).

157. Jagerdeo, E. & Abdel-Rehim, M. Screening of cocaine and its metabolites in human urine samples by direct analysis in real-time source coupled to time-of-flight mass spectrometry after online preconcentration utilizing microextraction by packed sorbent. *Journal of the American Society for Mass Spectrometry* **20,** 891–899 (2009).

158. Nilles, J. M., Connell, T. R. & Durst, H. D. Quantitation of Chemical Warfare Agents Using the Direct Analysis in Real Time (DART) Technique. *Analytical Chemistry* **81,** 6744–6749 (2009).

159. Lee, I., Ouk Kim, Y., Park, S.-C. & Chun, J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology* **66,** 1100–1103 (2016).

160. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Research* **45,** D535–D542 (2017).

161. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **50,** D20–D26 (2021).

162. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25,** 1043–1055 (2015).

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

164. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3693 (2015).

165. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30,** 772–780 (2013).

166. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21,** 180 (2020).

167. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49,** W293–W296 (2021).

168. Seemann, T. *Barrnap 0.9: Rapid ribosomal RNA prediction* `https://github.com/tseemann/barrnap`.

169. Nawrocki, E. *Structural RNA Homology Search and Alignment Using Covariance Models* PhD thesis (Washington University School of Medicine, 2009).

170. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37,** 1530–1534 (2020).

171. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14,** 587–589 (2017).

172. Steenwyk, J. L., Iii, T. J. B., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLOS Biology* **18,** e3001007 (2020).

173. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology* **17,** 238 (2016).

174. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* **428,** 726–731 (2016).

175. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36,** 2251–2252 (2019).

176. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47,** D309–D314 (2019).

177. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38,** 5825–5829 (2021).

178. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7,** e1002195 (2011).

179. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28,** 27–30 (2000).

180. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Science: A Publication of the Protein Society* **28,** 1947–1951 (2019).

181. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research* **51,** D587–D592 (2023).

182. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49,** W29–W35 (2021).

183. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* **16,** 60–68 (2020).

184. Huang, N., Siegel, M. M., Kruppa, G. H. & Laukien, F. H. Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data. *Journal of the American Society for Mass Spectrometry* **10,** 1166–1173 (1999).

185. Jewison, T. *et al.* YMDB: the Yeast Metabolome Database. *Nucleic Acids Research* **40,** D815–820 (2012).

186. Ramirez-Gaona, M. *et al.* YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Research* **45,** D440–D445 (2017).

187. Guo, A. C. *et al.* ECMDB: the E. coli Metabolome Database. *Nucleic Acids Research* **41,** D625–630 (2013).

188. Sajed, T. *et al.* ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Research* **44,** D495–501 (2016).

Supplementary

Louise Spekking (6201563)

December 12, 2023

Results

Additional filtering of *Klebsiella pneumoniae*

Pan-genomes construction of *Klebsiella pneumoniae* initially yielded no core genomes, indicating large genetic variation between strains. To remove the strains with the largest variation from analysis a Mashtree was used to remove all outliers from the dataset (Figure S1). Removed PATRIC gene ids are displayed in Table S1.

Table S1: PATRIC gene ids removed from analysis for *K. pneumoniae*

Ids	
573.4104	573.5649
573.4033	573.2003
573.4035	573.12495
573.4132	507522.9
573.2314	72407.164
573.4193	72407.165
573.4107	72407.166
573.27692	72407.83

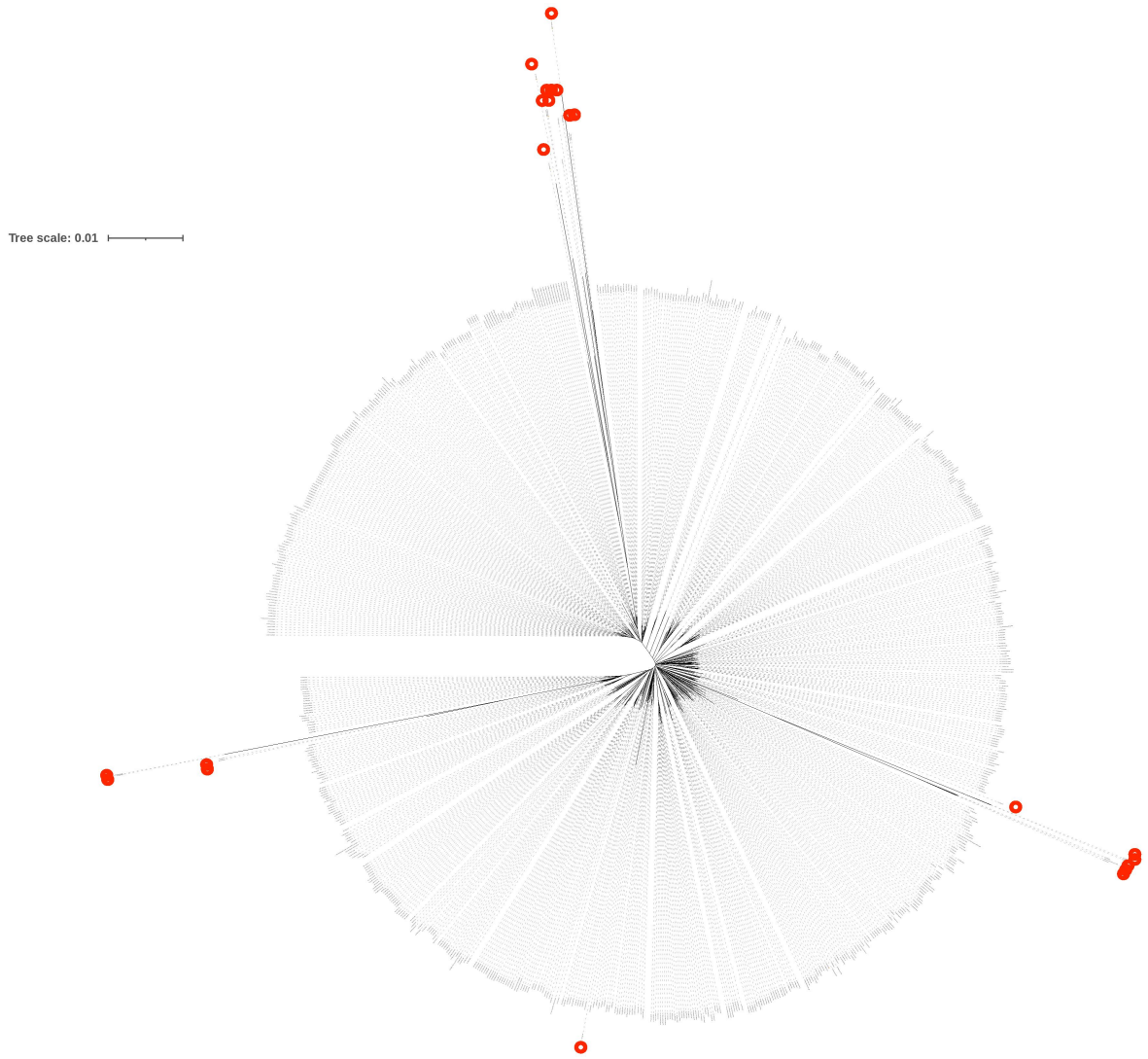


Figure S1: Mashtree as visualised by iTOL (v5.0) of all *Klebsiella pneumoniae* genomes annotated as complete in PATRIC. Strains indicated with red circle were removed from analysis.

Phylogenetic analysis

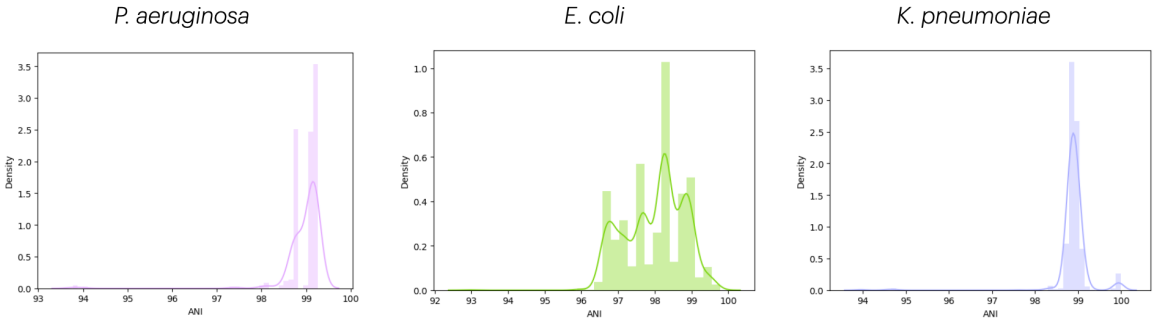


Figure S2: Distribution of ANI scores for outlier genomes as identified by 16s phylogeny. Showing distributions with only incidental outliers below 95%.

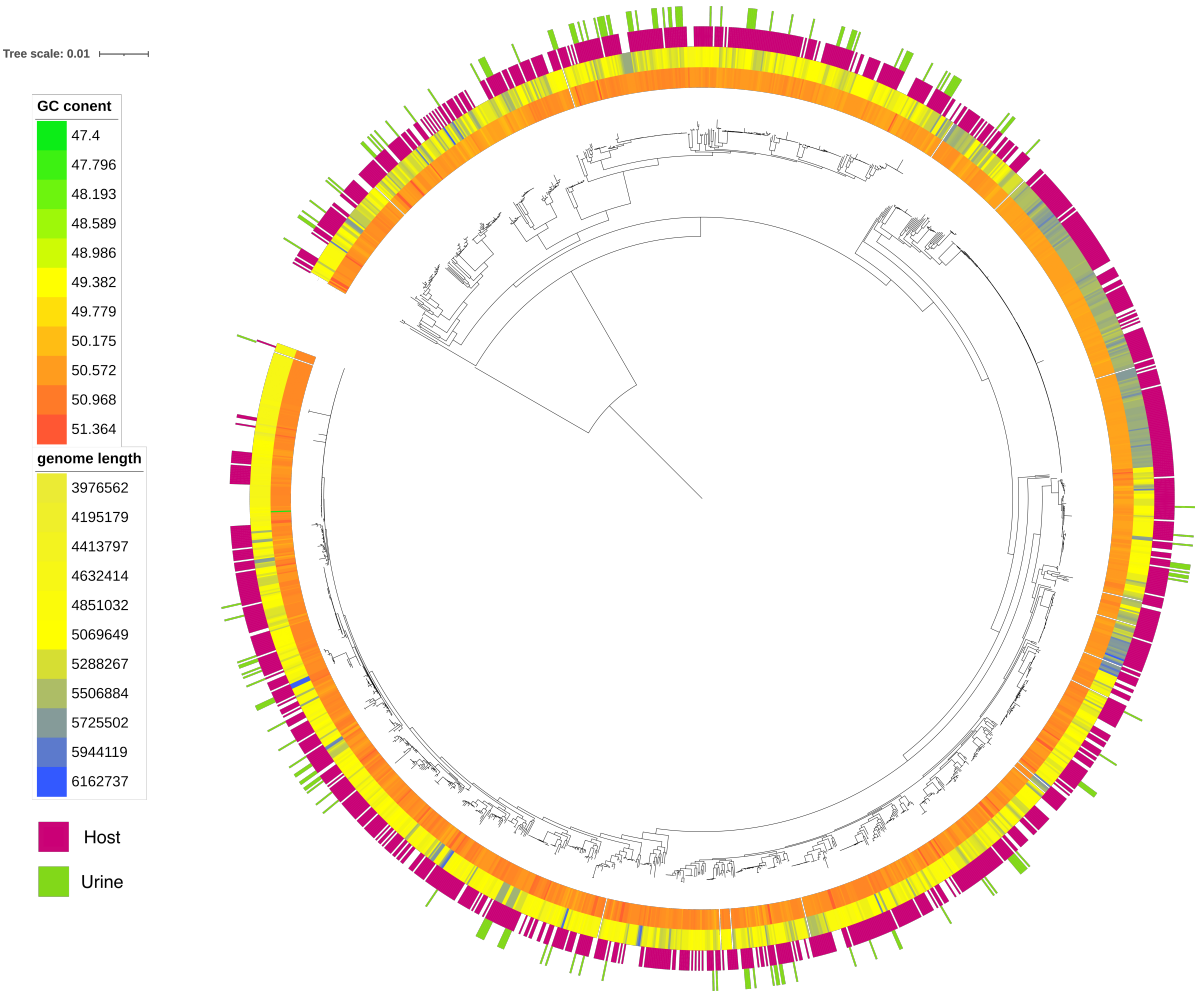
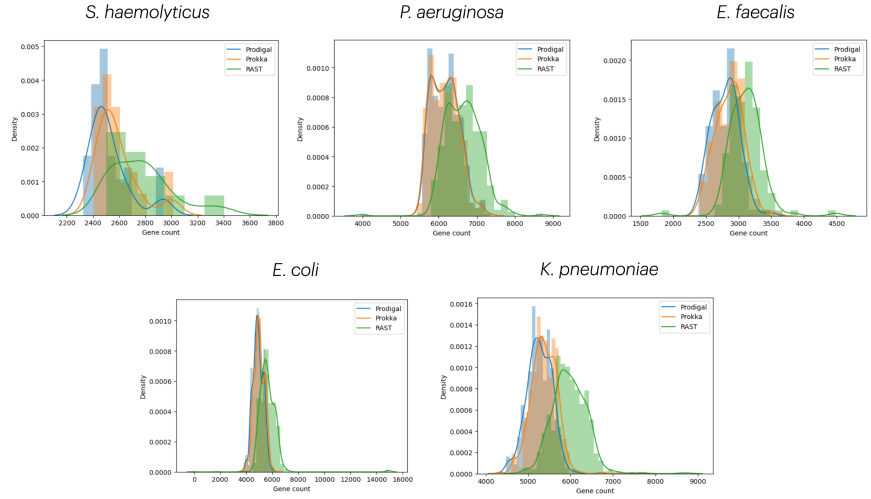


Figure S3: Distribution metadata for *k*mer based phylogeny of *E. coli* produced by Mashtree. Showing no bias for GC content and host/environmental isolation, and longer genomes for clade where the urine origin of isolation in absent.

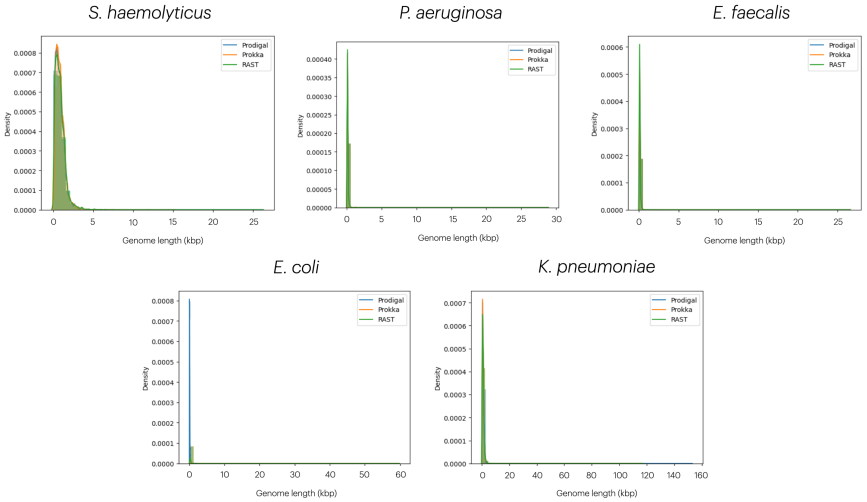
Gene annotations

Three annotation strategies were tested, Prokka, Prodigal (v2.6.3) and the RAST annotations as available on the PATRIC server, analysing the annotated gene numbers and lengths [1, 2, 3, 4]. Distributions of gene counts per genome showed a general trend where Prodigal predicted the smallest number of genes per genome, followed by Prokka (Figure S4a). This result could be explained by Prokka's gene finding method where Prodigal is used for prediction of coding genes, with the possibility of adding more non-coding genes at further annotation steps [3]. Most genes per genome were annotated in the GFF3 files as downloaded from PATRIC, mostly annotated with RAST. This would make these predicted gene files the preferred data source for further analysis, however gene predictions from PATRIC can be manually curated for some genomes, creating a disparity between the predicted genes. Moreover, GFF3 and GBK files were not avail-

able for all strains, further increasing the disparity between annotations by adding new gene predictions for the unannotated genomes. Gene lengths of the predicted genes all had approximately the same distribution, indicating that all three predictions have approximately the same bias on length of predicted genes. Given the disparity created by the use of PATRIC supplied gene annotation files and Prokka predicting coding and non-coding genes, Prokka was chosen as a gene prediction tool.



(a) Gene counts



(b) Gene lengths

Figure S4: Distribution of (A) gene counts and (B) gene lengths of genes predicted for all genomes of each species in the dataset. Prodigal (blue), Prokka (orange) and RAST (green)

Annotations

To map the gene clusters of the pan-genome that are overrepresented or unique to urine to KEGG pathways and modules, KEGG K0 annotations were needed for each gene cluster. Hereto three annotation methods were compared, BlastKOALA, KofamKOALA and EggNOG for the urine unique gene clusters (Table S2, Figure S6). Annotations for comparison were preformed on urine unique gene clusters only to reduce computational time of annotations for analysis.

Total numbers of unique K0 annotations show that EggNOG annotated the most gene clusters with a unique KO for all species except *E. coli*, where KofamKOALA annotated most unique KOs. Analysing the overlaps although there is an overlap in KO annotations, the different tools almost always annotate unique KOs as well, the annotation method chosen will therefore be of influence on the final result. Here EggNOG was chosen as annotation method as EggNOG can be run locally, thereby facilitating the annotation of all sequencers within one gene cluster in an more efficient manner and resulted in the most annotation for all species except *E. coli*.

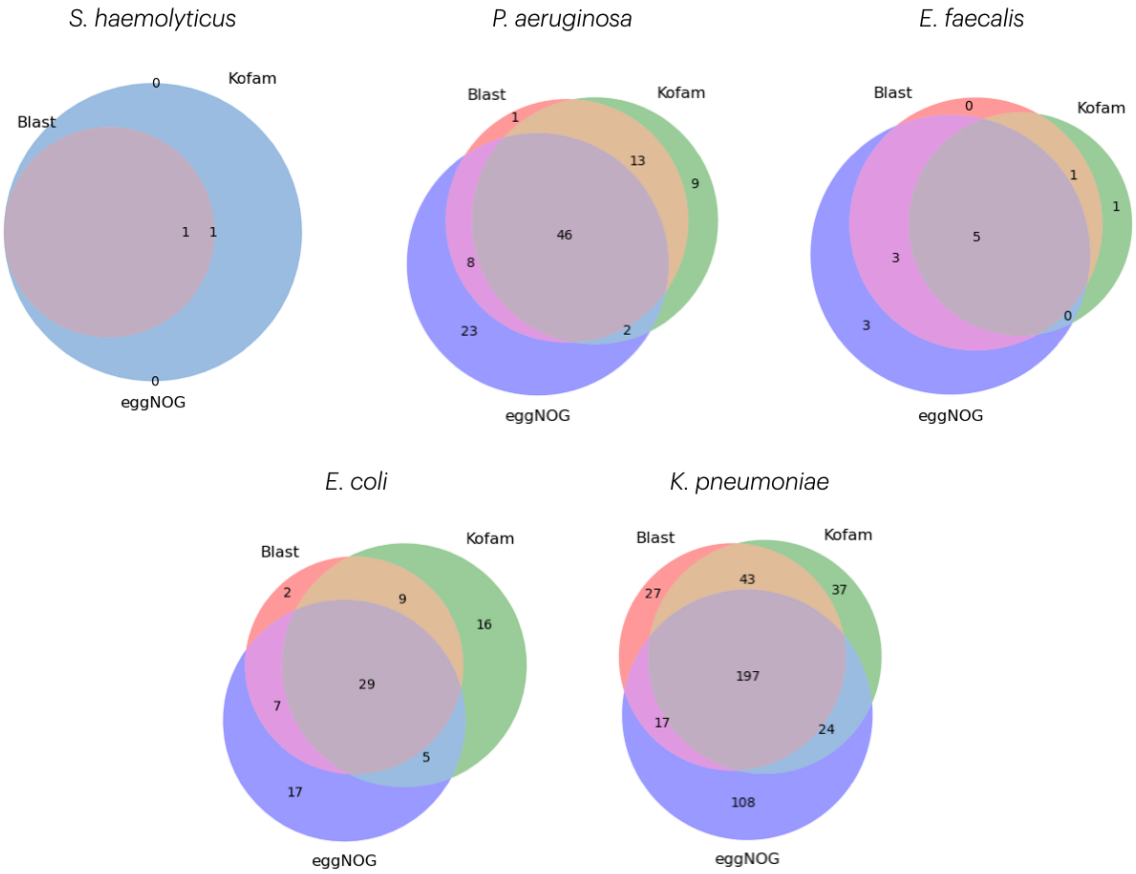


Figure S5: Overlaps of K0 annotations of urine unique sets for blastKOALA, kofamKOALA and eggNOG. Showing most unique and common annotated K0s by eggNOG.

Table S2: KEGG K0 annotations per annotation tool

Species	**BlastKOALA**	**KofamKOALA**	**EggNOG**
S. haemolyticus	1	2	2
P. aeruginosa	68	70	79
E. faecalis	9	7	11
E. coli	47	59	58
K. pneumoniae	284	301	346

Table S3: KEGG K0 annotations by eggNOG, total and mapped to a pathway

Species	KEGG K0	K0 in Pathway
S. haemolyticus	2	1
P. aeruginosa	79	34
E. faecalis	11	4
E. coli	58	20
K. pneumoniae	364	169

KEGG pathway analysis

Several KEGG pathways were excluded from analysis as these pathways are know to be absent in bacteria or are non functional within the urinary tract, thereby having no additive value to the analysis. List of excluded KEGG pathways are displayed in Table S4

Table S4: KEGG pathways excluded from analysis

KEGG pathway id	Pathway name
map00603	Glycosphingolipid biosynthesis - globo and isoglobo series
map00710	Carbon fixation in photosynthetic organisms
map00980	Metabolism of xenobiotics by cytochrome P450
map00981	Insect hormone biosynthesis
map00982	Drug metabolism - cytochrome P450
map00999	Biosynthesis of various plant secondary metabolites
map01524	Platinum drug resistance
map03250	Viral life cycle - HIV-1
map04011	MAPK signaling pathway - yeast
map04013	MAPK signaling pathway - fly
map04016	MAPK signaling pathway - plant
map04066	HIF-1 signaling pathway
map04080	Neuroactive ligand-receptor interaction
map04113	Meiosis - yeast
map04141	Protein processing in endoplasmic reticulum
map04211	Longevity regulating pathway
map04212	Longevity regulating pathway - worm
map04213	Longevity regulating pathway - multiple species
map04214	Apoptosis - fly
map04217	Necroptosis
map04361	Axon regeneration
map04626	Plant-pathogen interaction

(To be continued)

KEGG pathway id	Pathway name
map04727	GABAergic synapse
map04940	Type I diabetes mellitus
map04970	Salivary secretion
map04972	Pancreatic secretion
map04973	Carbohydrate digestion and absorption
map04975	Fat digestion and absorption
map05010	Alzheimer disease
map05012	Parkinson disease
map05014	Amyotrophic lateral sclerosis
map05016	Huntington disease
map05017	Spinocerebellar ataxia
map05020	Prion disease
map05022	Pathways of neurodegeneration - multiple diseases
map05030	Cocaine addiction
map05031	Amphetamine addiction
map05131	Shigellosis
map05134	Legionellosis
map05146	Amoebiasis
map05166	Human T-cell leukemia virus 1 infection
map05200	Pathways in cancer
map05204	Chemical carcinogenesis - DNA adducts
map05206	MicroRNAs in cancer
map05207	Chemical carcinogenesis - receptor activation
map05208	Chemical carcinogenesis - reactive oxygen species
map05225	Hepatocellular carcinoma
map05230	Central carbon metabolism in cancer
map05231	Choline metabolism in cancer
map05340	Primary immunodeficiency
map05415	Diabetic cardiomyopathy
map05418	Fluid shear stress and atherosclerosis

Metabolomics

Table S5: Number of m/z values in reference measurements after filtering at a relative intensity of 0.006

Species	Positive mode	Negative mode
S. haemolyticus	351	397
P. aeruginosa	391	489
E. faecalis	402	671
E. coli	990	1518

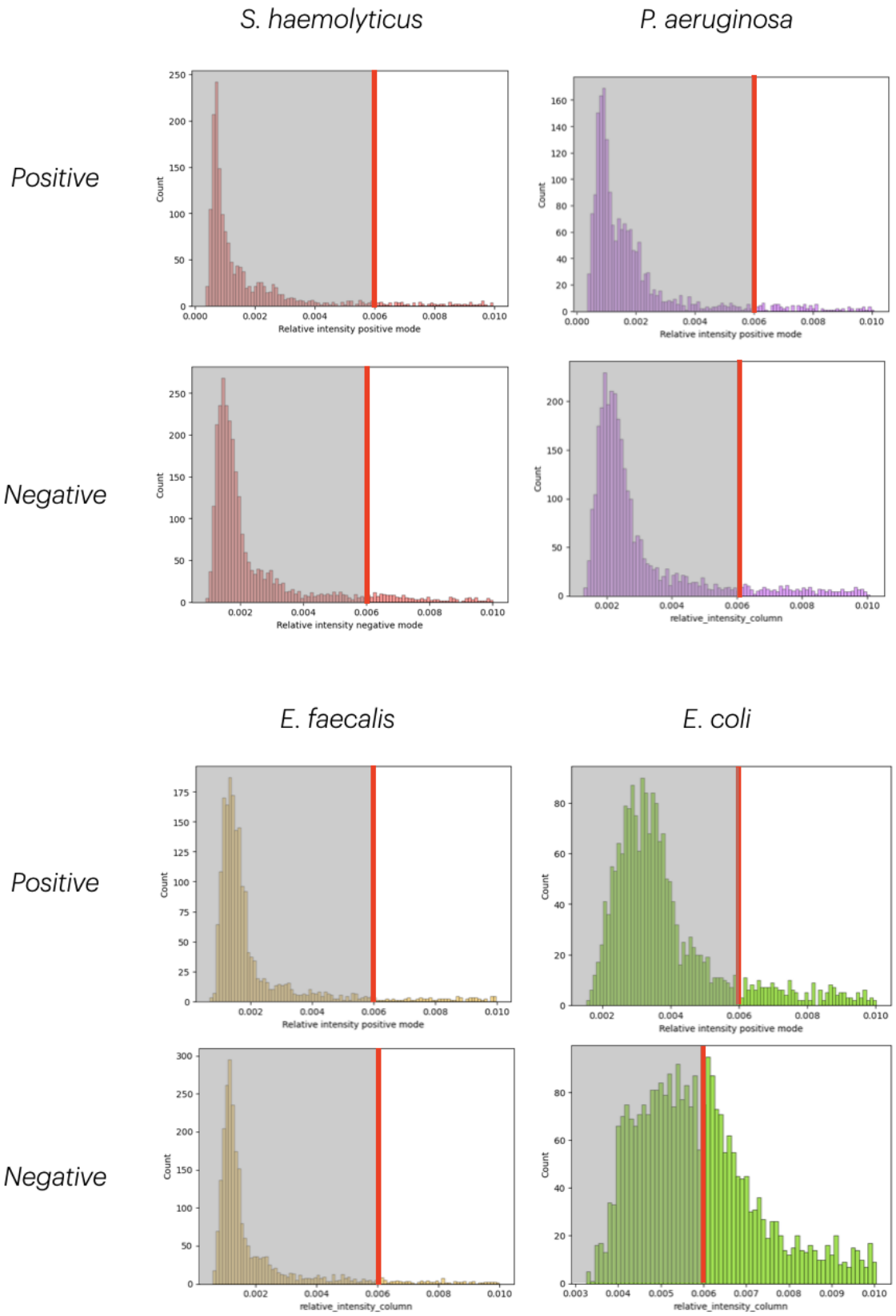


Figure S6: Histograms of relative intensities of reference measurements in positive and negative by DART. Red line indicated filtering threshold of 0.006, gray shaded was removed from analysis.

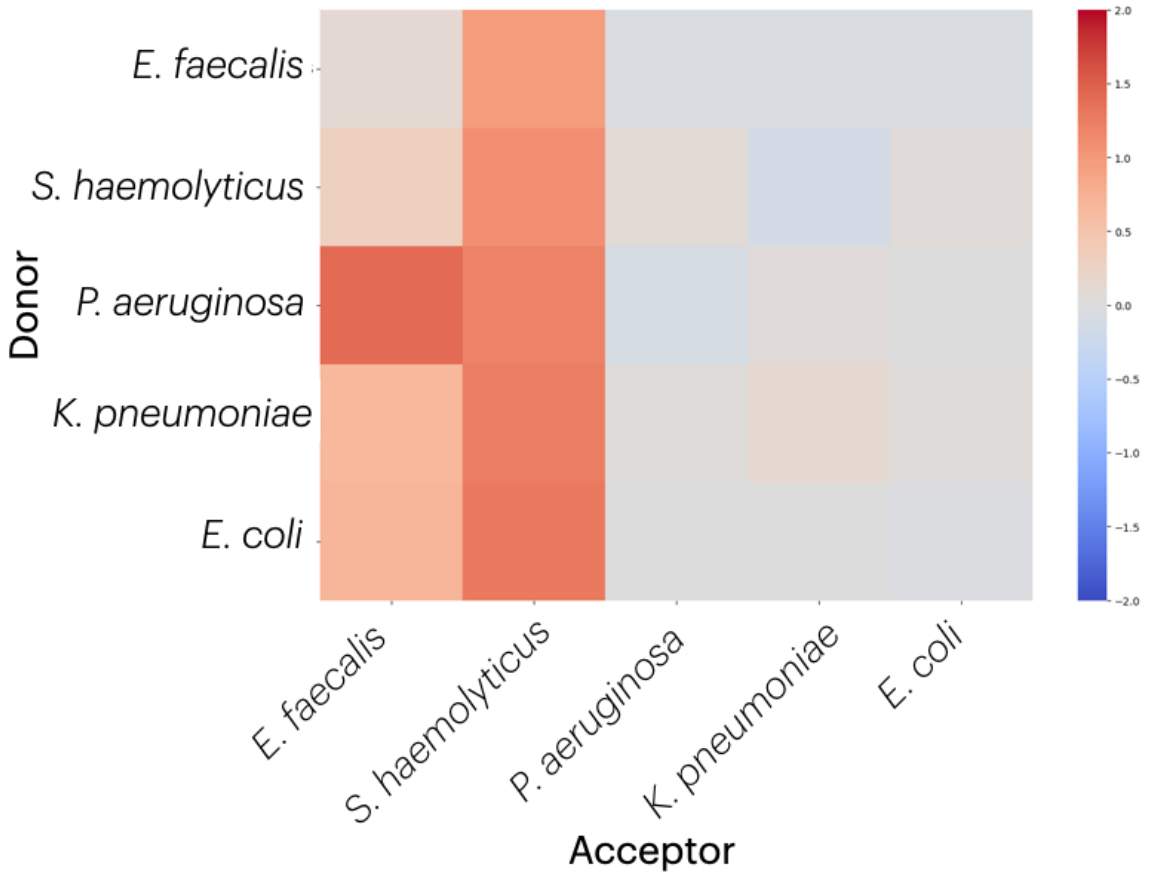


Figure S7: Pairwise interaction matrix depicting the interactions in yield (maximum *OD600*) of 5 UTI isolates of which metabolomics were measured in conditioned medium prepared from these same isolates. The interaction measure, -2 indicates positive interaction (blue), 2 negative interactions (red). The acceptor strains (columns) are grown in the conditioned medium of the donor strains (rows). The upper left to lower right diagonal represents the self-interactions. Modified from [5].

References

[1] Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**, 75 (2008).

[2] Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

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

[4] Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Research* **45**, D535–D542 (2017).

[5] de Vos, M. G. J., Zagorski, M., McNally, A. & Bollenbach, T. Interaction networks, ecological stability, and collective antibiotic tolerance in polymicrobial infections. *PNAS* **114**, 10666–10671 (2017).