Evaluation of niche adaptation features by genome data mining approach of *Escherichia coli* urinary and gastrointestinal strains

Thiago GS Paim[1], Gustavo E Sambrano[1], Keli C Reiter[1], Fernando H Sant'Anna[2], Renata Soares[1], Mariana Mott[1], Samuel P Cibulski[2],Pedro A d'Azevedo[1].

[1] Universidade Federal de Ciências da Saúde de Porto Alegre, UFCSPA, Brazil.
[2] Universidade Federal do Rio Grande do Sul, UFRGS, Brazil.

Corresponding Author:
Thiago Paim[1]

Email address: thiagog@ufcspa.edu.bt

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

Abstract

Background. Urinary Tract Infections (UTIs) are among most common infections in humans. The vast majority are caused by *Escherichia coli*, occasionally responsible for severe clinical manifestations. Although the species frequently adheres and colonizes the bladder mucosa, its reservoir is the host gastrointestinal tract. Therefore, the study was designed to evaluate genomic features for niche adaptation of urinary and gastrointestinal strains of *E. coli* by data mining approach. **Results.** In the *E. coli* strains, the repertoire of genes was higher than those found in previous studies, and the majority of genes associated to primary metabolism did not depend of bacteria niche, with exception of cell cycle-division, cell motility and secondary metabolite metabolism. Urinary tract isolates of *E. coli* had great density of virulence and resistance genes carried by prophages. **Conclusion.** The urinary and gastrointestinal strains of *E. coli* evaluated in the study presented an open pan-genome, with groups of functional annotation genes associated to specific niches. In addition, gastrointestinal isolates of *E. coli* were demonstrated as important reservoir of resistance genes.

Keywords

Escherichia coli, niche adaptation, pangenome analysis, urinary tract infections

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

Background

Urinary tract infections (UTIs) are common bacterial infections that affect individuals of all ages (Flores-Mireles at al, 2015) and their recurrence can affect 40% of female patients (Ciani et al, 2013). The cost associated to UTIs are high in developed countries, approximately US$ 3 billions per year in the USA (Flores-Mireles at al, 2015) and € 58 million in France (François et al, 2016).

Escherichia coli is the most frequent microorganism recovered from UTIs, with pathotypes associated to enteric and extraintestinal infections (Bien et al, 2012). One important characteristic of this species is that its major niche reservoir is the gastrointestinal tract, where they are constitute the colon microbiota at low abundance, although in the infant gut this genus is rich (Jandhyala et al, 2015).

There are well characterized genes associated to pathogenesis of *E. coli* strains, which confer host adherence, colonization and bacterial survival in urinary tract (Bien et al, 2012). However, molecular features present in both enteroaggregative and uropathogenic strains *of E.* coli have been shown in the same isolate (Paim et al, 2016), supporting the hypothesis that others genetic determinants play important role on bacterial persistence in different host niche.

Under the circumstances, this study was planned to analyze *E. coli* strains recovered from gastrointestinal and urinary tract for data mining genomic features to better understand niche-specific adaptation and selection.

Methods

Ethical Application

The Universidade Federal de Ciências da Saúde de Porto Alegre granted Ethical approval to carry out the study within its facilities (Ethical Application Ref: 502.474).

Bacterial strains and Genomes

This study was performed with bacterial strains from Laboratory of Molecular Microbiology - UFCSPA and genomes deposited on NIH-NCBI database. The uropathogenic

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

strain *E. coli E2* had its genome sequence determined previously (Paim et al, 2016). Representative genomes were obtained from NCBI Microbial Genomes (https://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html), resulting in 5,295 genomes of *E. coli* that were downloaded to create a local database for bioinformatics analysis.

Only bacterial genomes containing information about source of isolation and recovering from human host were analyzed in the study. Among these, the genomes were grouped according to site of isolation: Urinary Tract (UT) and Gastrointestinal Tract (GI).

Genome Groups

Average Nucleotide Identity (ANI) confirmed the strains belonging to genomes groups at species level. The method was computed among microorganisms of the database using the module *pyani* for python computer programming language (available on https://pypi.python.org/pypi/pyani/) using BLASTN algorithm with fragment size of 1020 bp. The exception was *E. coli* GI group, whose tetranucleotide frequency correlation was computed due to computational limitation. Heatmaps were generated with command-line matplotlib package of python language for data visualization.

Pan-Genome Analysis

The genomes sequences were submitted to pan-genome analysis using the ROARY command-line software (Page et al, 2015). Briefly, the respective genomes were converted from GENBANK to GFF3 format file containing annotated coding regions and these were grouped in clusters based on homology. Representative sequence from each cluster was used for gene distribution among the isolates and to build the pan-genome. Core and accessory genome were defined by genes present in ≥ 99% and < 99% of the strains, respectively.

Functional Annotation of Pan-Genome

The pan-genome sequences were analyzed by functional characterization of the genes using the Clusters of Orthologous Groups of proteins (COGs) database (Galperin et al, 2015 - http://www.ncbi.nlm.nih.gov/COG/). The genes were *in silico* translated and a python script was

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

written to perform a BLASTP search (cutoff: e-value ≤ 10^{-10}) of each CDS from a local COG database. The functional assignments of the most significant sequences were extracted to compute the pan-genome functional profile of each isolates clusters.

Prediction of Prophages Sequences

To verify the contribution of integrative elements in the structure of bacterial genomes, prophages sequences were predicted using PHAge Search Tool (PHASTER), available on www.phaster.ca (Arndt et al, 2016).

Resistome

To evaluate the distribution of resistance-associated genes from genomes and prophage sequences, the command-line software ResFinder (Zankari et al, 2012) was used with the following parameters: at least alignment length of 60% and nucleotide identity of 95%. Informations about gene and antimicrobial class as well as frequency of the resistance genes were recovered for data analysis.

Virulence Factors Analysis

The prophage sequences were analyzed by presence of virulence factors (VF) using the Virulence Factor Database (VFDB) (Chen et al, 2012). The search of VF's was performed by an in-house python script using BLASTP algorithm (parameters: e-value ≤ 10^{-3}, alignment coverage ≥ 50% and identity ≥ 30%), with CDS recovered from prophages homologous sequences reported after PHASTER analysis.

Results

Pan-Genome analysis and Functional Annotation

The pan-genome analysis demonstrated different numbers of genes among *E. coli* genome groups. The biggest pan-genome was from GI group, with > 93,000 genes, compared with UT group, with 34,936 genes. On the other hand, the ratio core/pan-genome was higher for UT group (7.14% versus 1.29%), although the number of core genes was similar (1,200 for GI

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

and 2,493 for UT). Finally, the number of accessory genes for GI and UT was 92,022 and 32,443, respectively.

Genes from pan-genome were grouped in 25 functional annotations clusters (Figure 1). An elevated number of genes did not have homologous sequences in COG database, therefore, they were categorized as 'without functional annotation', ranging to 40% in *E. coli* strains. The association between the ratio of functional annotation and the site of isolation of each strains was analyzed by Pearson's Chi-squared test. There was a significant association on *E. coli* genomes ($p < 0.001$), in which UT strains had a superior ratio of genes associated to cell cycle-division, cell motility and secondary metabolite metabolism. Diversely, a lower percentage of genes associated to mobilome was found when compared to GI.

Prophages Analysis

A total of 13,131 prophage sequences were predicted for all genomes, with median of 8 prophages per genome [1st quartile: 6; 3rd quartile: 11]. For *E. coli* strains belonging to GI group (n=1190), the median was nine prophages, followed by *E. coli* UT (n=274, median = 7) ($p<0.001$, by Wilcoxon rank sum test) (Figure 2).

Genes encoding resistance to antimicrobials located inside of prophages sequences were predicted in 9.8% of *E.coli* strains. There was no significant correlation between prophages number and acquisition of resistance genes by *E. coli* genomes [Pearson correlation coefficient (r) = 0.0955, p-value = 0.2545, CI95% (-0.0691 to 0.2552)]. However, the GI group had a significant, but weak correlation [r = 0.2178, p-value = 0.0183, CI95% (0.0378 to 0.3841)]. Regarding density of resistance genes by prophage, UT isolates had a superior median (Figure 3). Four-hundred and nine genes were found, containing genetic determinants that confer resistance to trimethoprim and sulphonamide most frequently found between prophages (GI – 103 and 84, UT – 23 and 16, respectively) (Figure 4). From these, co-occurrence of *dfrA7* gene (dihydrofolate reductase - trimethoprim resistance) and *sul1* gene (dihydropteroate synthase - sulphonamide resistance) in same prophage was frequently recovered from *E. coli* GI group (45 co-occurrences). While in *E. coli* UT group the most common co-ocurrence found was *dfrA14* and *mphA* gene (macrolide 2'-phosphotransferase I – macrolide resistance) (5 co-occurrence). To both groups, co-occurrence of three resistance genes in a same prophage predicted was 14 (*dfrA7*,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

sul1 and *catA1* - chloramphenicol acetyltransferase) (Figure 5).

A total of 13,606 virulence genes were found inside of prophage sequences, and these were recovered in 1,431 of 1,464 genomes (97.7%). The median of virulence factors in prophages was 7 [1st quartile: 4; 3rd quartile: 11] and the density of virulence genes per prophage was equal 0.86 [1st quartile: 0.53; 3rd quartile: 1.30]. By isolation source, the median of density was 0.82 and 1.13 to GI and UT, respectively (p-value < 0.001 by Wilcoxon rank sum test). Strong and moderate correlation between number of prophages and virulence factors carried by these integrative elements was found in GI and UT groups, respectively (Figure 6). The most frequent genes recovered by *E. co*li genomes were *r*ck, *ipaH*, *ipaH2*.5, *om*pD, *perC/bf*pW and *flm*H (Table 1).

Resistome

For *E. coli* isolates, 771 of 1,475 (52.3%) had resistance genes predicted by ResFinder tool. Regarding to genome groups, 635 of 1,199 (53%) and 136 of 276 (49.3%) showed at least one resistance determinant in GI and UT clusters, respectively (Figure 7). To both groups, the most prevalent phenotype class which confers resistance to antimicrobial agents were aminoglycoside, sulphonamide and beta-lactam (Figure 8).

There was a significant correlation between isolation date and the number of resistance genes predicted from genomes of *E. coli* (Figure 9). The analysis by class of resistance genes, showed significant correlations among aminoglycoside (r = 0.1657, p-value < 0.001, CI95% (0.0686 to 0.2596)), phenicol (r = 0.2486, p-value = 0.0055, CI95% (0.0749 to 0.4077)), sulphonamide (r = 0.1455, p-value = 0.0027, CI95% (0.0749 to 0.4077)) and beta-lactam (r = 0.1258, p-value = 0.0210, CI95% (0.0191 to 0.2297)). For the variables, correlation was examined separately for *E. coli* group (GI and UT). Although significant correlation was demonstrated on the phenotype of aminoglycoside and phenicol resistance for all *E. coli* genomes, only GI group was significant. In addition, only in the GI group a significant negative correlation was found to phenotype of trimethoprim resistance (Figure 10-12).

Discussion

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

Uropathogenic microorganisms are an important cause of morbidity in humans, affecting female and male of all ages, considering gram-negative bacteria as the most common agents of these infections, predominantly *E. coli*. The gene repertoire that confers adherence, colonization and ability to survive at the urinary tract is important to uropathogens, and some genetic determinants are known to execute these functions (Flores-Mireles at al, 2015).

Pan-genome analysis is defined by the repertoire of genes based on set of genomes, and information about closed or open pan-genome can be determined analyzing the pattern of acquisition of new genetic elements (Rouli et al, 2015). Furthermore, the bacteria lifestyle is affected by capacity of niche adaptation, which sympatric species develops in microbial community and has an open pangenome (Georgiades; Raoult, 2010). The *E. coli* genomes of the study had low ratio core/pan-genome for both isolation sources (GI or UT), typical of sympatric species. In addition, based on pan-genome size, it increased when new genomes were added, showing that this species represents an open pan-genome model independently of the niche. Similar conclusion of pan-genome model was obtained by a previous study (Rasko et al, 2008). Although, in our study, isolates from UT and GI demonstrated pan-genomes with more than 30,000 and 90,000 genes, respectively, the number of core genes were similar (2,493 *versus* ~2,200).

Analysing the gene-content of secondary metabolite metabolism from *E. coli* in this study, the UT strains had a high frequency of genes associated to non-ribosomal peptide synthetase component F, which is a homologous gene to the enterobactin component F by COG id. Iron uptake is essential for bacterial metabolism as a biocatalyst or as an electron carrier in reactions associated to carbon metabolism, replication and DNA repair and energy production (Caza; Kronstad, 2013). Enterobactin system is enconded by operon *entCDEBAHF*, that synthesizes siderophore protein from shikimic acid pathway (Ma; Payne, 2012; Peralta et al, 2016). In mammalian hosts, iron is available to bacteria strains from various sources, but dietary iron in the gastrointestinal tract is important to colonization and commensalism and is limited in extraintestinal infection sites (Russo et al, 2002; Caza; Kronstad, 2013). The presence of sixty-four homologous *entF* genes shows the diversity of this enzime in the pangenome of *E. coli* urinary tract isolates. This enzime and others of operon convert 2,3-dihydroxybenzoate to enterobactin, which is secreted by bacteria (Ma; Payne, 2012). Currently, is not well clear if

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

variant proteins of EntF modify enterobactin secreted by these strains, unless the chemical structure and compositon of this siderophore are experimentaly determinated. However, in *Salmonella* species the production of modified enterobactins is relevant to infection processes by preserving the iron-binding activity but evading the siderocalin activity. This protein is expressed by immune cells of the host that bind to enterobactin-iron complex and present an antibacterial activity (Cherayil, 2011). In addition, in urinary tract infections the oxidative stress is elevated (Belge Kurutas et al, 2005), which induces the up-regulation of enterobactin synthesis (Peralta et al, 2016). These findings suggest that homologous enterobactin F genes have a relevant role on uropathogenesis of *E. coli* strains.

Prophage sequences, besides mobile DNA elements, are common source of variability among isolates of same species. Integrated in bacterial chromosome, genetic elements as resistance genes, virulence factor and genes for niche adaptation can be interchanged among different strains or replicons (ex. plasmids) (Canchaya et al, 2003). The elevated diversity of prophage predicted in genomes of *E. coli* shows that these genetics elements are important for chromosome structure and potential selective adaptation. In the bacterial genomes of this study, the distribution of prophages was evaluated by isolation source (GI and UT). The genome database had a high number of GI strains of *E. coli*, alike UT strains. For all, there was an important variability of prophage recovered, with strains showing number of predicted sequences higher 25 integrative elements. In addition, the median of prophages in GI group was superior of UT.

There was an increase of number genes that confer antimicrobial resistance at the proportion that the number of prophages incorporated in bacterial genome o*f E. co*li recovered from gastrointestinal tract increased, although for urinary tract strains this correlation was not valid. However, for GI group, it was found a higher number of integrative elements carrying genetic determinants other than resistance genes. These facts are straightforward for the analysis of density of resistance genes by prophage sequences predicted in the study. Even though strains isolated from GI tract may have a superior genomic plasticity to acquisition of prophage sequences, these elements did not carry genes that confer selective advantages as antimicrobial resistance very often. UT strains had higher number of homologous genes carrying antimicrobial resistance and virulence factors. Consequently, these isolates were more cost-effective in relation

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

to the acquisition of integrative elements, while showed superior selective advantages per prophage when compared to GI strains.

The resistome recovered from the study demonstrated that the number of resistance genes carried by isolates of *E. coli* increased over the last decades. Independently of isolation source, the *E. coli* isolates had an important increase of genes that confer resistance to aminoglycoside, phenicol, sulphonamide and beta-lactam antimicrobial agents. In a retrospective study that evaluated the historical changes of *E. coli* resistance in US, Tadesse et al (2012) sampled strains isolated from humans and animals during 1950-2002 and performed phenotypic antimicrobial susceptibility test (AST) to 15 antimicrobial agents. From human source, the study demonstrated that 65% of the strains were pan-susceptible (Tadesse et al, 2012). Our data had genomes isolated from 1970 to 2016, and the antimicrobial susceptibility based on absence of resistance gene predicted was 48%. In addition, our data supports the increasing trend in resistance to ampicillin (beta-lactam) and sulphonamide, but not to tetracycline (Tadesse et al, 2012). The most frequent beta-lactamase gene found in genomes of *E. coli* was *blaTEM-1b*, which is often associated to ampicillin resistance in this specie (Brinas et al, 2002).

Although significant correlation has been found for acquisition of aminoglycoside resistance genes over time in all *E. coli* genomes, only for GI strains the trend was significant. There were genes found only in GI isolates, as *aph(3')-Ic*, *aph(3')-IIa*, *aph(3')-XV*, *aph(4)-Ia*, *aac(3)-IVa* and *aadA12*, but also there were genes found only in UT strains as *aadA16*, *rmtC*, *aph(3')-VIa*, *rmtE* and *aadD*. This scenario corroborates the hypothesis that GI isolates have an important role as reservoir of resistance genes (Huddleston, 2014; Van Schaik, 2015) and the presence of these molecular determinants are cumulative on *E. coli* genome. However, some selective pressure are fixing determined genes based on adaptation to niche of *E. coli* isolates.

Conclusion

The study showed that all species have an open pangenome independently of isolation source. For *E. coli* strains, the repertoire of genes is higher than previous studies and the ratio of genes associated to primary metabolism does not depend of bacteria niche, however, homologous genes to siderophore proteins in urinary tract isolates could contribute – as fimbrial-like proteins

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

– for uropathogenesis of *E. coli* isolates. These same genomes showed often prophages sequences carrying genetic determinants of resistance and virulence, although the gene-content of these mobile elements did not show differences with gastrointestinal isolates.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

References

Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*. 2016;44(Web Server issue):W16-W21.

Belge Kurutas E, Ciragil P, Gul M, Kilinc M. The Effects of Oxidative Stress in Urinary Tract Infection. *Mediators of Inflammation*. 2005;2005(4):242-244. doi:10.1155/MI.2005.242.

Bien J, Sokolova O, Bozko P. Role of Uropathogenic *Escherichia coli* Virulence Factors in Development of Urinary Tract Infection and Kidney Damage. *International Journal of Nephrology*. 2012;2012:681473. doi:10.1155/2012/681473.

Briñas L, Zarazaga M, Sáenz Y, Ruiz-Larrea F, Torres C. Beta-lactamases in ampicillin-resistant Escherichia coli isolates from foods, humans, and healthy animals. Antimicrob Agents Chemother. 2002 Oct;46(10):3156-63.

Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H. Prophage Genomics. *Microbiology and Molecular Biology Reviews*. 2003;67(2):238-276. doi:10.1128/MMBR.67.2.238-276.2003.

Caza M, Kronstad JW. Shared and distinct mechanisms of iron acquisition by bacterial and fungal pathogens of humans. *Frontiers in Cellular and Infection Microbiology*. 2013;3:80. doi:10.3389/fcimb.2013.00080.

Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Research*. 2012;40(Database issue):D641-D645.

Cherayil BJ. The role of iron in the immune response to bacterial infection. *Immunologic research*. 2011;50(1):1-9. doi:10.1007/s12026-010-8199-1.

Ciani O, Grassi D, Tarricone R. An economic perspective on urinary tract infection: the "costs of resignation". Clin Drug Investig. 2013 Apr;33(4):255-61. doi: 10.1007/s40261-013-0069-x

Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature reviews Microbiology*. 2015;13(5):269-284. doi:10.1038/nrmicro3432.

François M, Hanslik T, Dervaux B, et al. The economic burden of urinary tract infections in women visiting general practices in France: a cross-sectional survey. *BMC Health Services Research*. 2016;16:365. doi:10.1186/s12913-016-1620-2.

Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*. 2015;43(Database issue):D261-D2

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

Georgiades K, Raoult D. Defining Pathogenic Bacterial Species in the Genomic Era. *Frontiers in Microbiology*. 2010;1:151. doi:10.3389/fmicb.2010.00151.

Huddleston JR. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. *Infection and Drug Resistance*. 2014;7:167-176. doi:10.2147/IDR.S48820.

Jandhyala SM, Talukdar R, Subramanyam C, Vuyyuru H, Sasikala M, Reddy DN. Role of the normal gut microbiota. *World Journal of Gastroenterology : WJG*. 2015;21(29):8787-8803. doi:10.3748/wjg.v21.i29.8787.

Ma L, Payne SM. AhpC Is Required for Optimal Production of Enterobactin by Escherichia coli. *Journal of Bacteriology*. 2012;194(24):6748-6757. doi:10.1128/JB.01574-12.

Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3693.

Paim TGS, Pieta L, Prichula J, et al. Draft Genome Sequence of Brazilian Escherichia coli Uropathogenic Strain E2. Genome Announcements. 2016;4(5):e01085-16. doi:10.1128/genomeA.01085-16.

Peralta DR, Adler C, Corbalán NS, Paz García EC, Pomares MF, Vincent PA. Enterobactin as Part of the Oxidative Stress Response Repertoire. Semsey S, ed. *PLoS ONE*. 2016;11(6):e0157799. doi:10.1371/journal.pone.0157799.

Rasko DA, Rosovitz MJ, Myers GSA, et al. The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. Journal of Bacteriology. 2008;190(20):6881-6893. doi:10.1128/JB.00619-08.

Rouli L, Merhej V, Fournier P-E, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*. 2015;7:72-85. doi:10.1016/j.nmni.2015.06.005.

Russo TA, McFadden CD, Carlino-MacDonald UB, Beanan JM, Barnard TJ, Johnson JR. IroN Functions as a Siderophore Receptor and Is a Urovirulence Factor in an Extraintestinal Pathogenic Isolate of *Escherichia coli* . *Infection and Immunity*. 2002;70(12):7156-7160. doi:10.1128/IAI.70.12.7156-7160.2002.

Tadesse DA, Zhao S, Tong E, et al. Antimicrobial Drug Resistance in *Escherichia coli* from Humans and Food Animals, United States, 1950–2002. *Emerging Infectious Diseases*. 2012;18(5):741-749. doi:10.3201/eid1805.111153.

Van Schaik W. The human gut resistome. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2015;370(1670):20140087. doi:10.1098/rstb.2014.0087.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

Zankari E, Hasman H, Cosentino S, et al. Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*. 2012;67(11):2640-2644.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

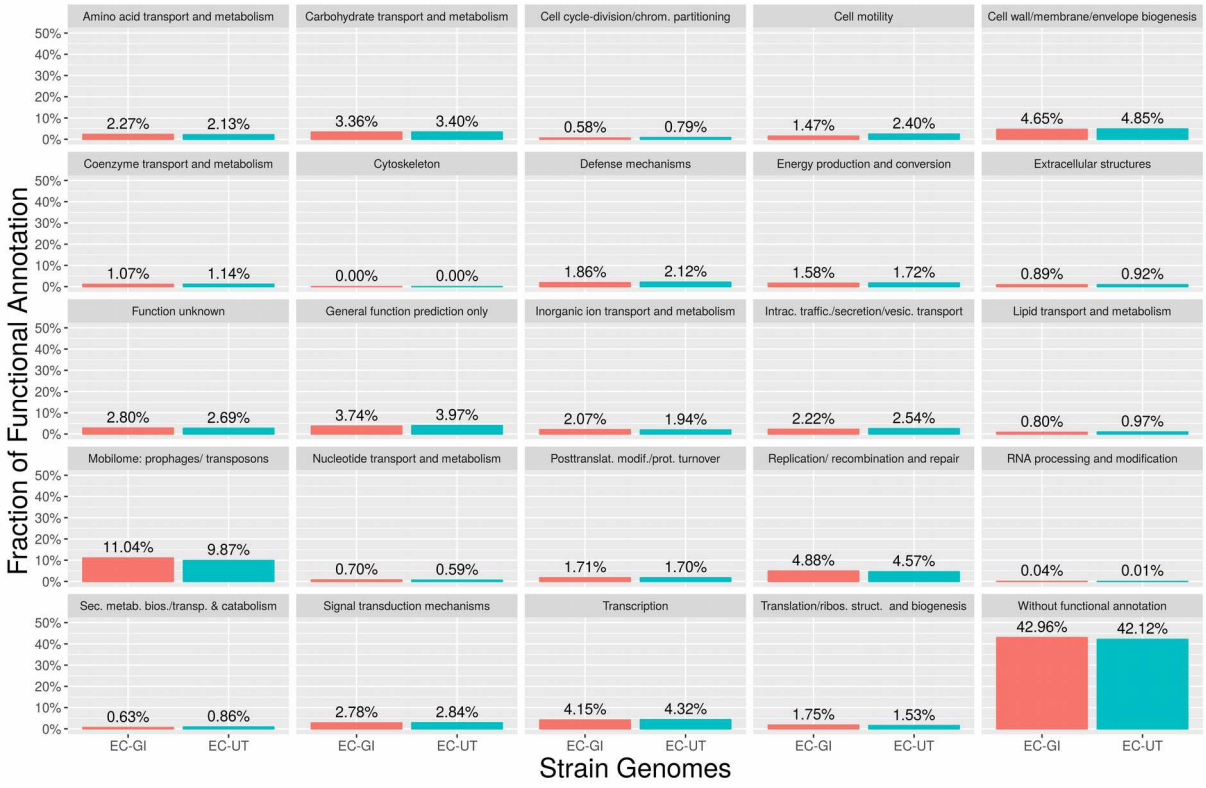


Figure 1. Functional annotation of pan-genome sequences. Legend: EC-GI (*Escherichia coli* genomes from gastrointestinal tract) and EC-UT (*Escherichia coli* genomes from urinary tract).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

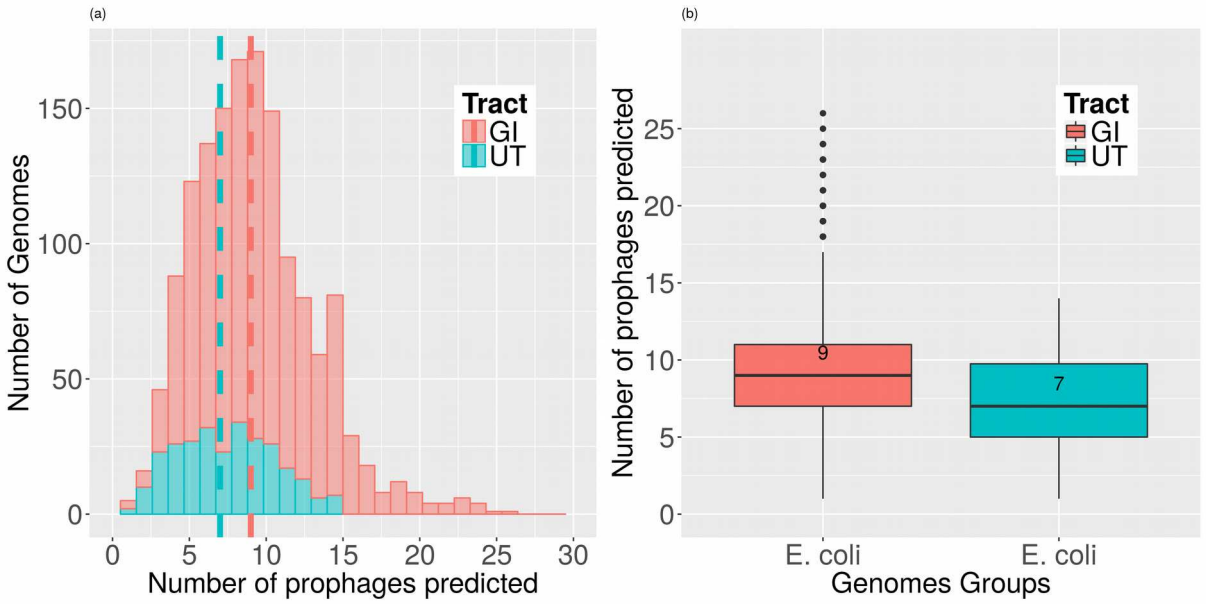
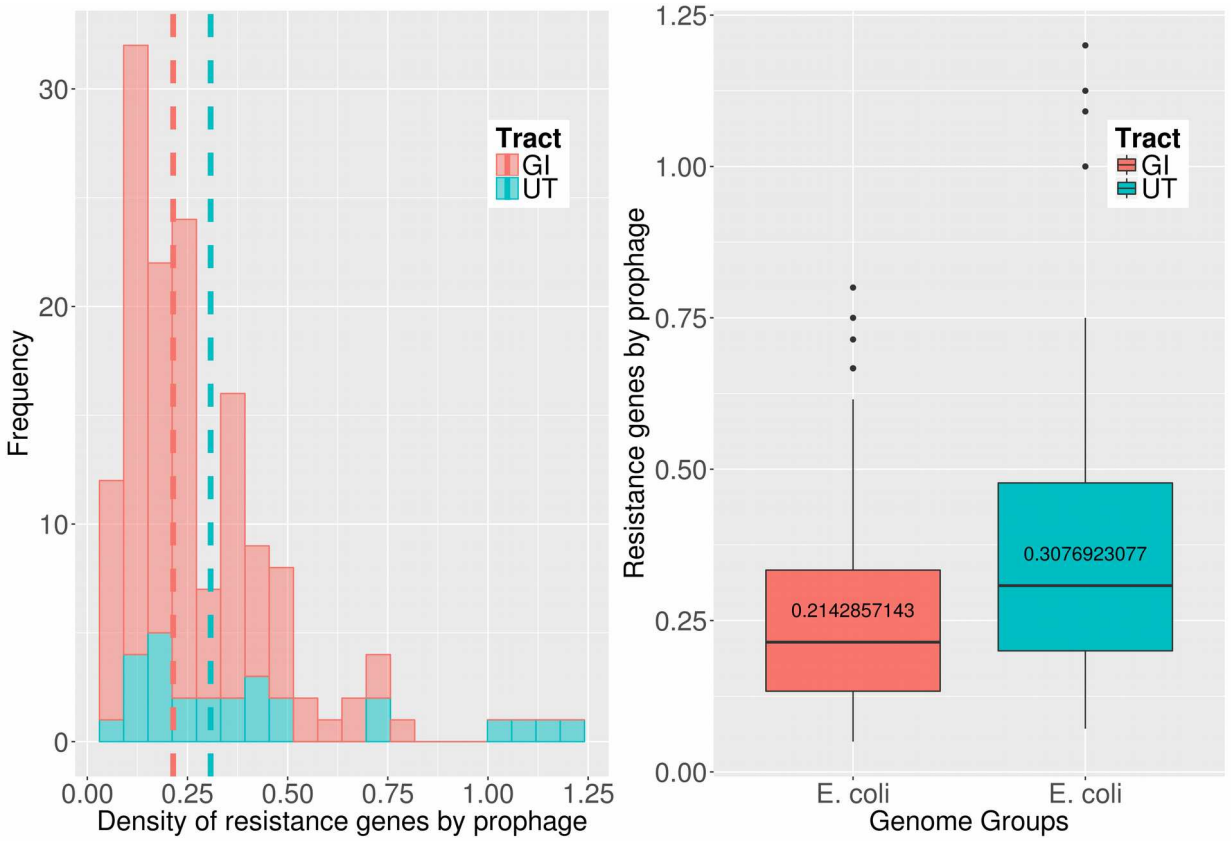


Figure 2. Distribution of prophages predicted *in silico* according to tract and bacterial species of the study. Lines: median of prophage in gastroinstestinal isolates (GI) [9, red line] and Urinary Tract (UT) [7, blue line].

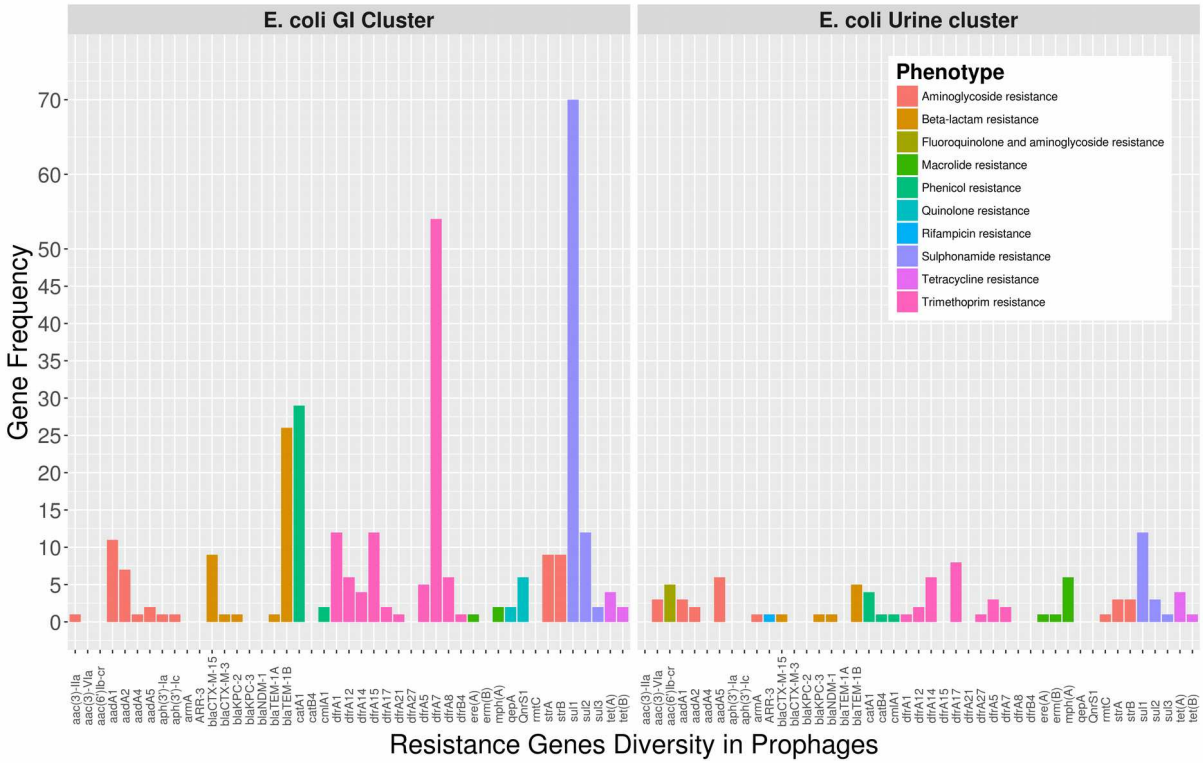
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019



384 Figure 3. Density of resistance genes by prophage. There was significant difference between GI
385 and UT isolates (Wilcoxon rank sum test, p = 0.01746).
386
387
388

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

389



390 Figure 4. Diversity of resistance genes (n=409) recovered inside of prophage sequences, all in *E.*

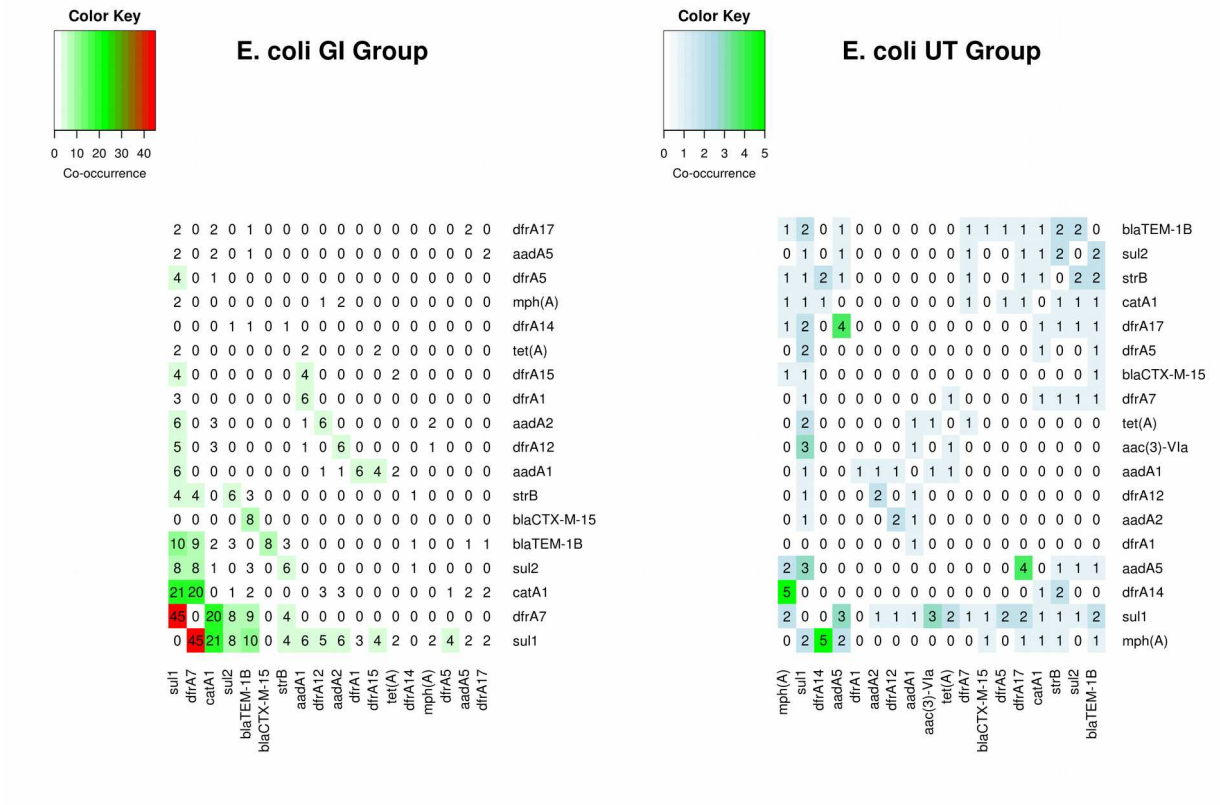
391 *coli* genomes.

392

393

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

394



395 Figure 5. Co-occurrence of resistance genes carried by same prophage sequence.

396

397

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

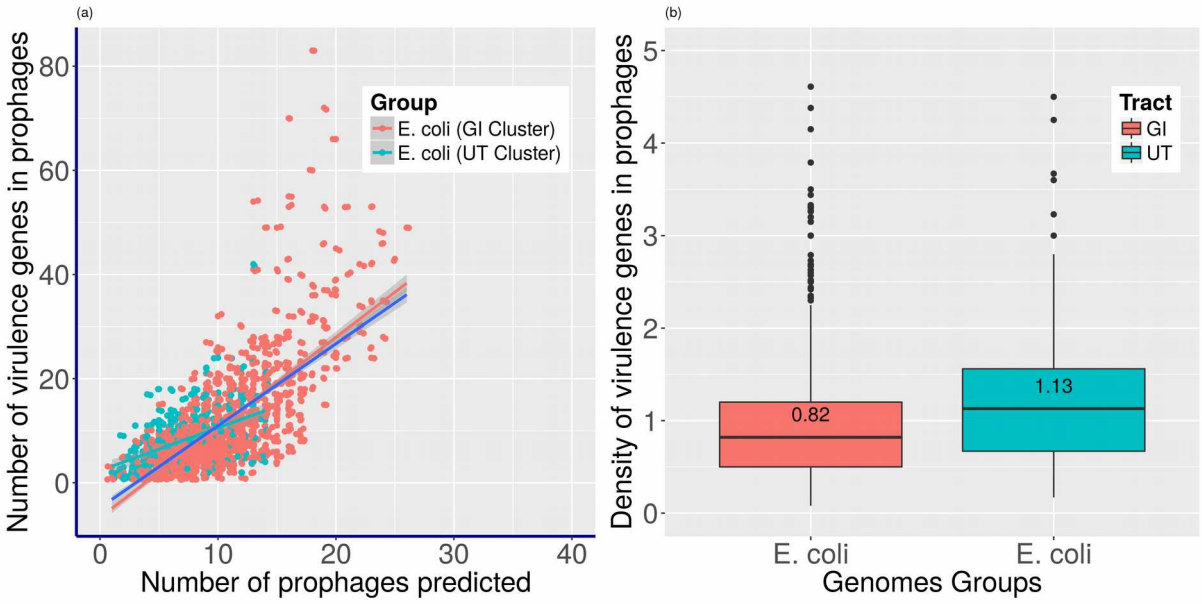


Figure 6. Distribution of virulence factors carried by prophage sequences. Density is the rario of number of virulence genes to number of prophage for each genome. Lines: red line – Pearson's product-moment correlation for GI isolates [r = 0.7269, CI95% (0.6988 to 0.7528), p-value < 0.001]; light blue line – Pearson's product-moment correlation for UT isolates [r = 0.4868, CI95% (0.3929 to 0.5705), p-value < 0.001]; dark blue line – Pearson's product moment correlation for all isolates [r = 0.6983, CI95% (0.6710 to 0.7236), p-value < 0.001].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

408

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

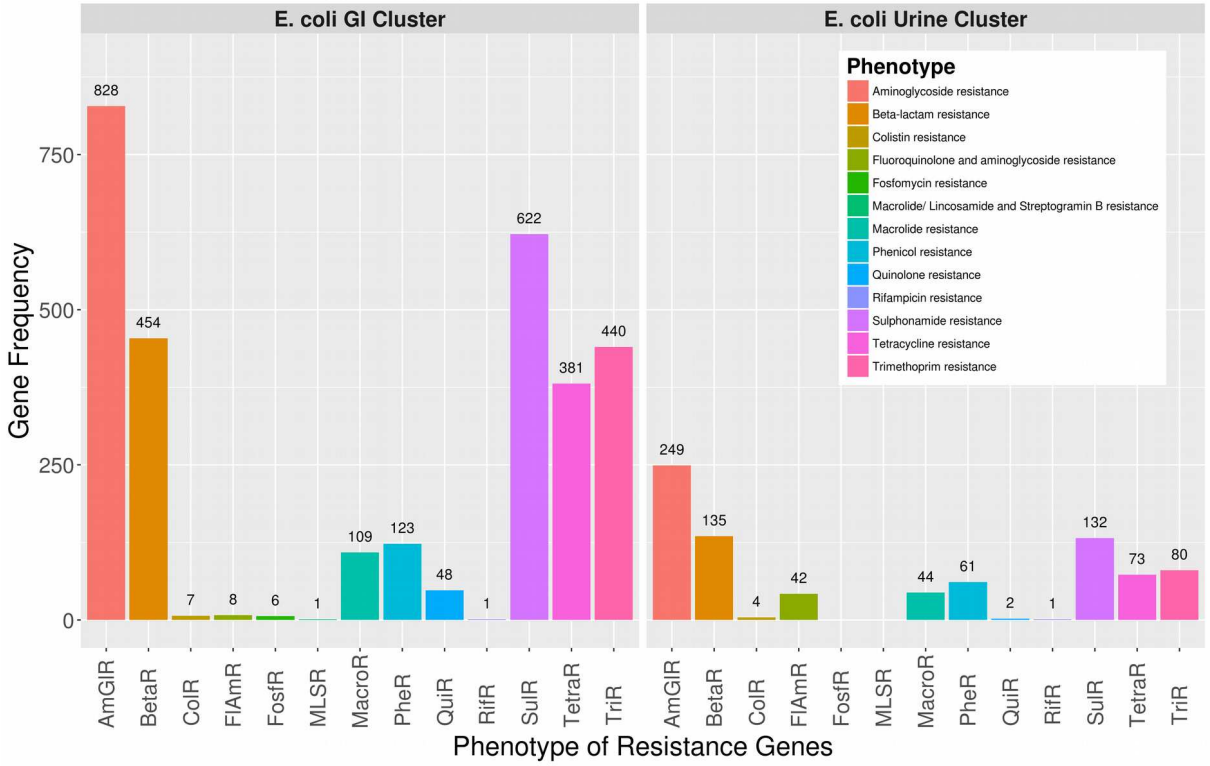
409 Table 1. Virulence factor *in silico* predicted and carried by prophage sequences. The most ten frequent genes by genome groups were recovered with
410 respective number of occurrences. The diversity of virulence factors by genome group was*: E. coli* GI = 518 *and E. coli* UT = 213
411

Group	Virulence Factor Predicted	N	%
E. coli (GI Cluster)	(rck) resistance to complement killing [Rck (VF0108)]	1091	9,6%
	(ipaH) hypothetical prophage protein [Mxi-Spa TTSS effectors controlled by MxiE (CVF465)]	971	8,6%
	(mlr6326) putative DNA invertase [T3SS (SS026)] [Mesorhizobium loti MAFF303099]	967	8,5%
	(ipaH2.5) invasion plasmid antigen, fragment [Mxi-Spa TTSS effectors controlled by MxiE (CVF465)]	732	6,5%
	(ompD) outer membrane porin precursor [Hek (AI384)]	590	5,2%
	(perC/bfpW) transcriptional regulator BfpW [Per (VF0190)]	475	4,2%
	(sclB) Collagen-like surface protein [Streptococcal collagen-like proteins (CVF116)]	232	2,0%
	(lpg2644) hypothetical protein [Legionella collagen-like protein (Lcl) (AI343)]	176	1,6%
	(flmH) 3-oxoacyl-ACP reductase [Polar flagella (VF0473)]	157	1,4%
	(eps4) exopolysaccharide biosynthesis protein [Capsule (CVF186)]	153	1,3%
E. coli (UT Cluster)	(mlr6326) putative DNA invertase [T3SS (SS026)]	313	13,8%
	(rck) resistance to complement killing [Rck (VF0108)]	298	13,1%
	(ipaH) hypothetical prophage protein [Mxi-Spa TTSS effectors controlled by MxiE (CVF465)]	189	8,3%
	(ompD) outer membrane porin precursor [Hek (AI384)]	153	6,7%
	(ipaH2.5) invasion plasmid antigen, fragment [Mxi-Spa TTSS effectors controlled by MxiE (CVF465)]	152	6,7%
	(nleK) hypothetical protein [T3SS (SS022)]	97	4,3%
	(flmH) 3-oxoacyl-ACP reductase [Polar flagella (VF0473)]	65	2,9%
	(mll6352) transposase [T3SS (SS026)]	64	2,8%
	(perC/bfpW) transcriptional regulator BfpW [Per (VF0190)]	55	2,4%
	(aaiW) transposase [AAI/SCI-II (SS182)]	44	1,9%

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27720v1 | CC BY 4.0 Open Access | rec: 10 May 2019, publ: 10 May 2019

413

414



415 Figure 7. Frequency of resistance genes predicted in *E. coli* genomes. A total of 3,851 gene
416 sequences were recovered by *in silico* search (GI – n=3,028; UT – n=823).

417

418

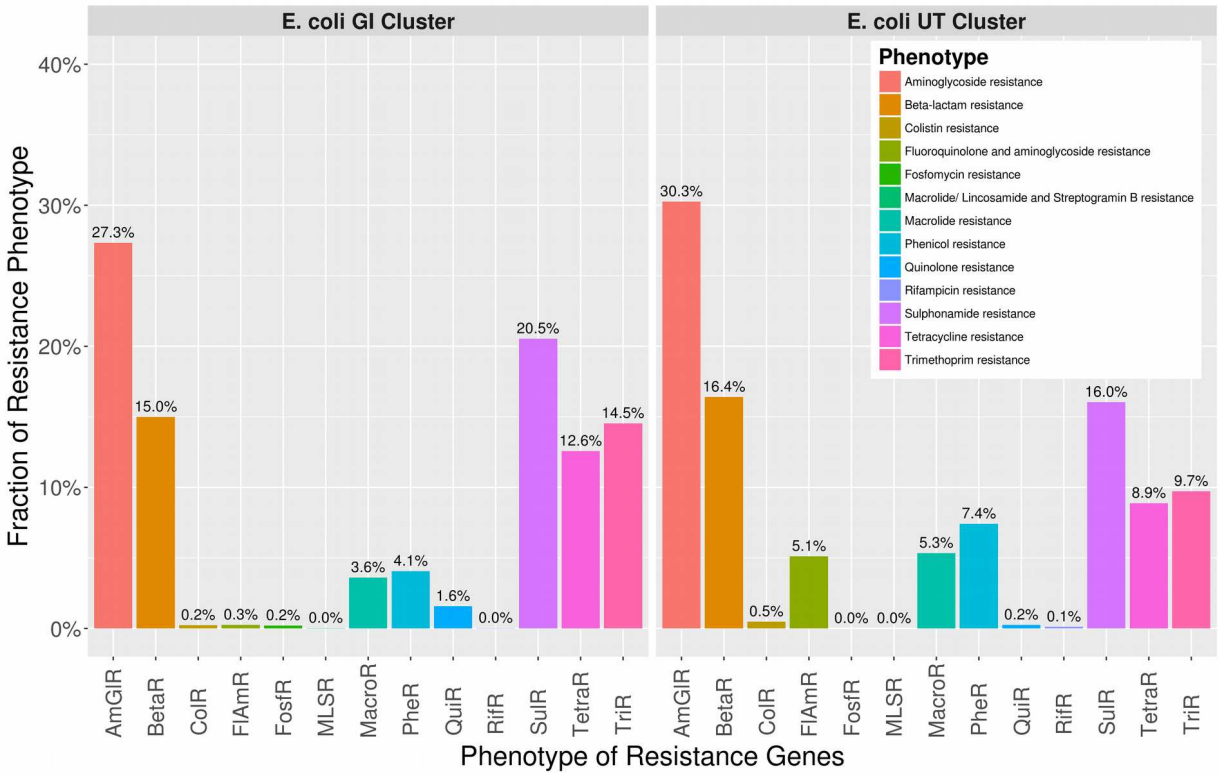
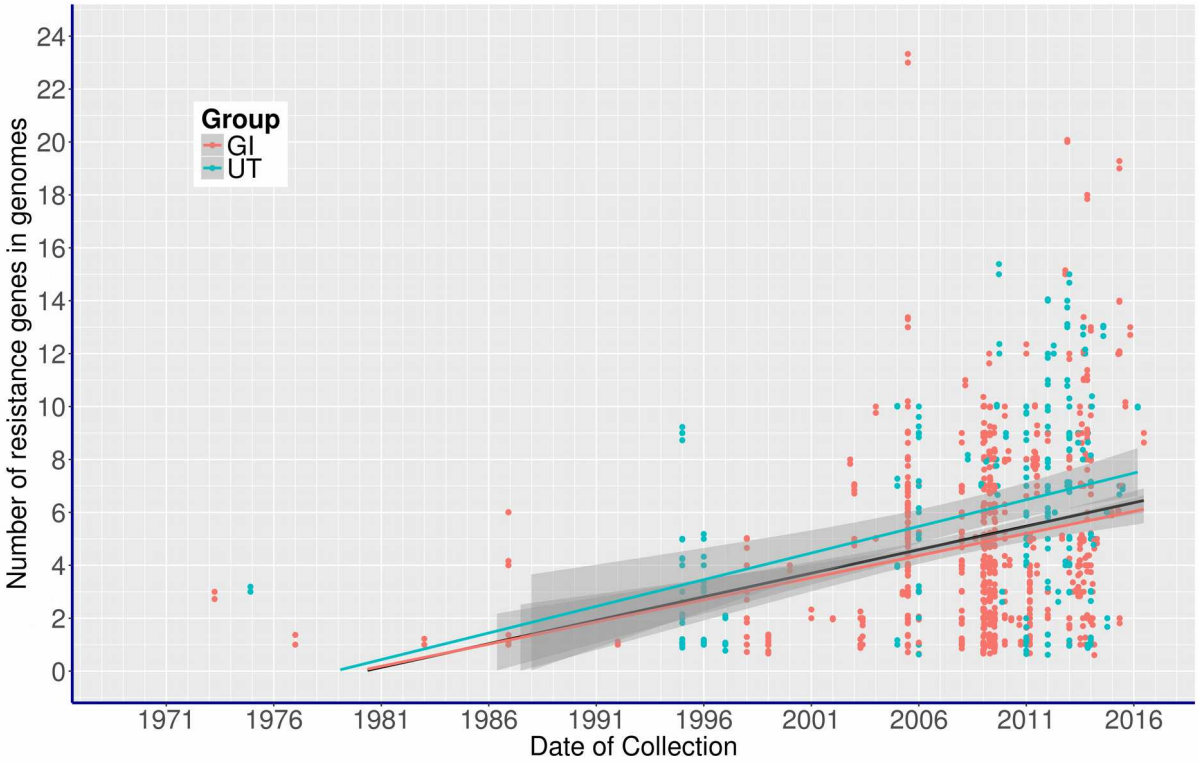
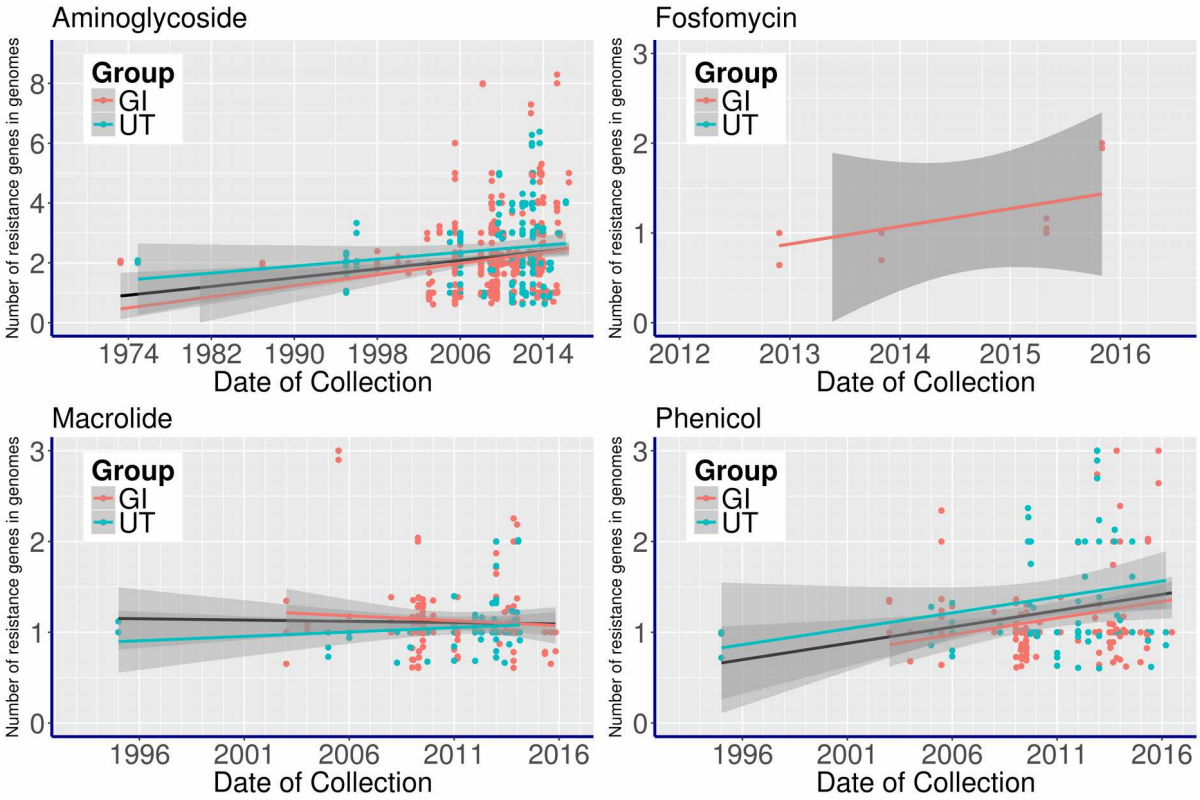


Figure 8. Prevalence of genes predicted in *E. coli* genomes by resistance phenotype. A total of 3,851 gene sequences were recovered by *in silico* search (GI – n=3,028; UT – n=823). There was significant association between class of resistance genes between groups by Pearson's Chi-squared test ($p < 0.001$), which urinary tract isolates had superior ratio of resistance genes of Fluoroquinolones and Aminoglycoside Resistance (FlAmR), Macrolide Resistance (MacroR) and Phenicol Resistance (PheR); and low ratio of Quinolone Resistance (QuiR), Sulphonamide Resistance (SulR), Tetracycline Resistance (TetraR) and Trimethoprim Resistance (TriR).

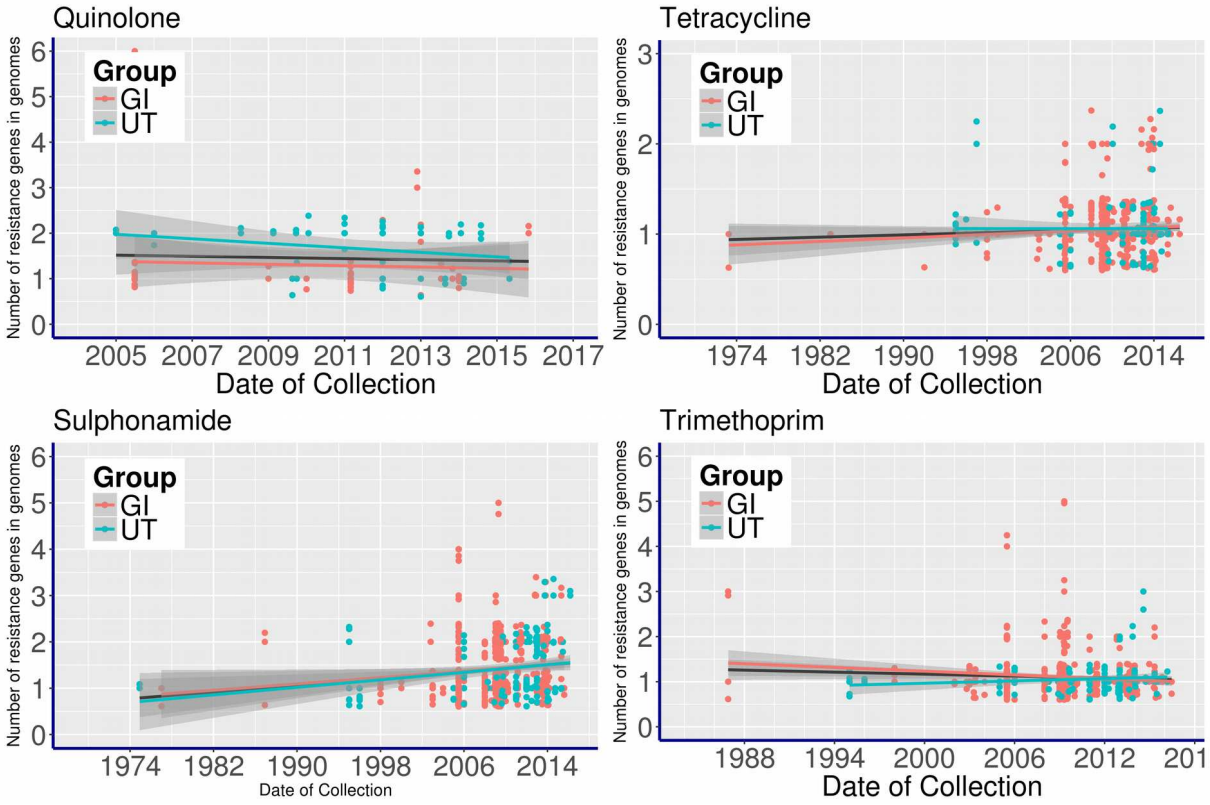
434



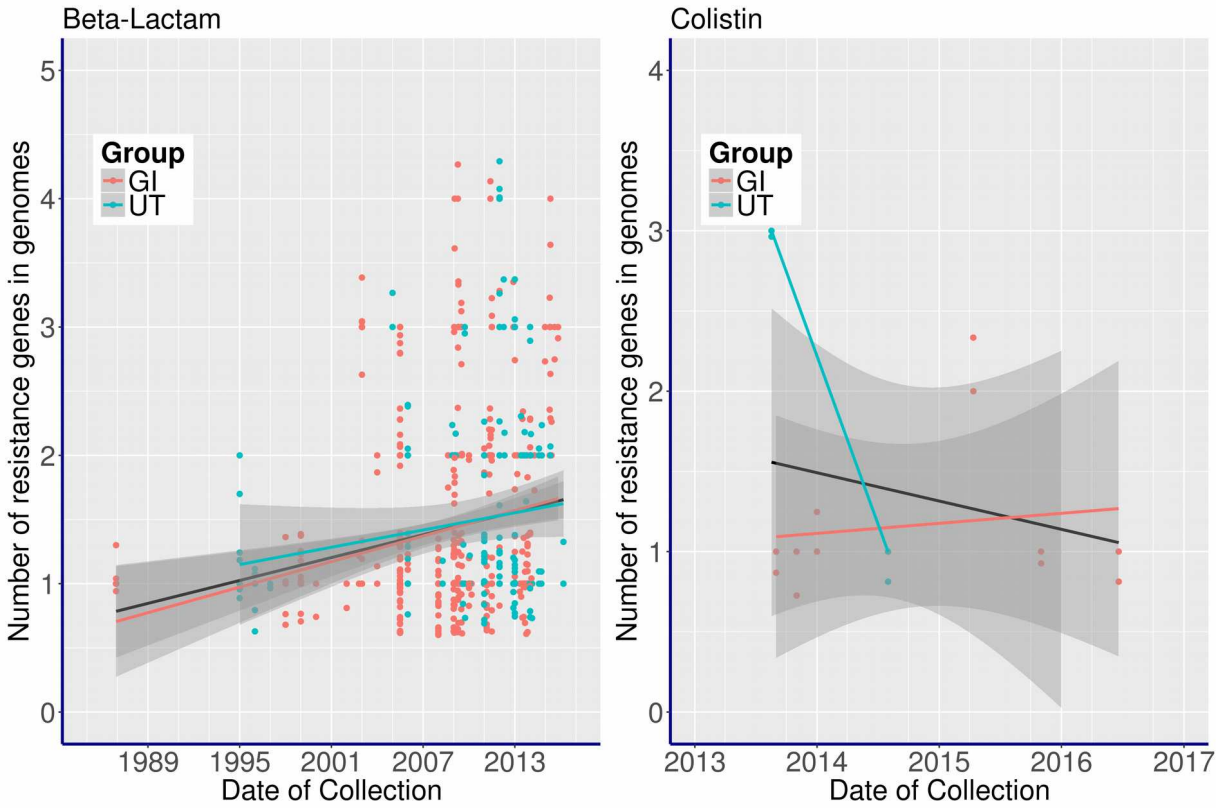
436 Figure 9. Scatter plot between number of resistant genes predicted in genomes of *E. coli* and
437 date of collection of isolates. Red points: Gastrointestinal Tract (GI); Blue points: Urinary
438 Tract (UT); Lines: Linear regression for both groups (GI + UT) (black); GI (red) and UT
439 (blue). Statistics: GI + UT – r = 0.2835, p-value < 0.001, CI95% (0.2097 to 0.3541); GI – r =
440 0.1140, p-value < 0.001, CI95% (0.0496 to 0.1774); UT – r = 0.4688, p-value < 0.001, CI95%
441 (0.3701 to 0.5570).



443 Figure 10. Scatter plots of number of resistance genes by class (aminoglycoside, fosfomycin,
444 macrolide and phenicol) and temporal isolation of *E. coli* strains. Red points: Gastrointestinal
445 Tract (GI); Blue points: Urinary Tract (UT); Lines: Linear regression for both groups (GI +
446 UT) (black); GI (red) and UT (blue).
447



449 Figure 11. Scatter plots of number of resistance genes by class (quinolone, tetracycline,
450 sulphonamide and trimethoprim) and temporal isolation of *E. coli* strains. Red points:
451 Gastrointestinal Tract (GI); Blue points: Urinary Tract (UT); Lines: Linear regression for both
452 groups (GI + UT) (black); GI (red) and UT (blue). Statistics: (r = -0.1193, p-value = 0.0352,
453 CI95% (-0.2273 to -0.0084).



455 Figure 12. Scatter plots of number of resistance genes by class (beta-lactam and colistin) and
456 temporal isolation of *E. coli* strains. Red points: Gastrointestinal Tract (GI); Blue points:
457 Urinary Tract (UT); Lines: Linear regression for both groups (GI + UT) (black); GI (red) and
458 UT (blue).
459
460
461
462
463
464