frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 13 December 2023
DOI 10.3389/fmicb.2023.1276332

OPEN ACCESS

EDITED BY
Daniel Yero,
Autonomous University of Barcelona, Spain

REVIEWED BY
Moataz Abd El Ghany,
The University of Sydney, Australia
Jess Vergis,
Kerala Veterinary and Animal Sciences University, India
Miquel Sánchez-Osuna,
Instituto de Investigación e Innovación Parc Taulí (I3PT), Spain

*CORRESPONDENCE
Michael Schroeder
✉ michael.schroeder@tu-dresden.de

RECEIVED 11 August 2023
ACCEPTED 16 November 2023
PUBLISHED 13 December 2023

CITATION
Malekian N, Sainath S, Al-Fatlawi A and Schroeder M (2023) Word-based GWAS harnesses the rich potential of genomic data for E. coli quinolone resistance.
Front. Microbiol. 14:1276332.
doi: 10.3389/fmicb.2023.1276332

Word-based GWAS harnesses the rich potential of genomic data for *E. coli* quinolone resistance

Negin Malekian[1], Srividhya Sainath[1], Ali Al-Fatlawi[1,2] and Michael Schroeder[1,3]*

[1]Biotechnology Center (BIOTEC), Technische Universität Dresden, Dresden, Germany, [2]ITRDC, University of Kufa, Najaf, Iraq, [3]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), TU Dresden, Dresden, Germany

Quinolone resistance presents a growing global health threat. We employed word-based GWAS to explore genomic data, aiming to enhance our understanding of this phenomenon. Unlike traditional variant-based GWAS analyses, this approach simultaneously captures multiple genomic factors, including single and interacting resistance mutations and genes. Analyzing a dataset of 92 genomic *E. coli* samples from a wastewater treatment plant in Dresden, we identified 54 DNA unitigs significantly associated with quinolone resistance. Remarkably, our analysis not only validated known mutations in *gyrA* and *parC* genes and the results of our variant-based GWAS but also revealed new (mutated) genes such as *mdfA*, the AcrEF-TolC multidrug efflux system, *ptrB*, and *hisI*, implicated in antibiotic resistance. Furthermore, our study identified joint mutations in 14 genes including the known *gyrA* gene, providing insights into potential synergistic effects contributing to quinolone resistance. These findings showcase the exceptional capabilities of word-based GWAS in unraveling the intricate genomic foundations of quinolone resistance.

KEYWORDS

genome-wide association studies (GWAS), microbial GWAS, word-based GWAS, unitig-GWAS, *E. coli*, quinolone, antibiotic resistance

1 Introduction

1.1 Conventional genome-wide association studies may not fully explore the potential of genomic data

While conventional GWAS methods have provided valuable insights, they may not fully explore the intricate genomic landscape underlying microbial phenotypes (Power et al., 2017), particularly in terms of variant interactions. These methods predominantly focus on individual single nucleotide polymorphisms (SNPs), potentially overlooking the significant influence of variant interactions and gene presence/absence on microbial phenotypic traits, including antibiotic resistance. To enhance our understanding and further complement conventional GWAS, embracing alternative strategies capable of simultaneously capturing the effects of genes and variants, whether they act individually or interactively, is crucial.

1.2 K-mers have the potential to broaden the horizons of conventional GWAS

K-mers, substrings of length k within biological sequences like DNA, RNA, or proteins, have the potential to broaden the horizons of conventional GWAS. By incorporating them

into genome analysis, we can expand beyond the scope of conventional GWAS, which typically centers on single nucleotide polymorphisms (SNPs) or genes as the unit of study. Unlike these traditional approaches, k-mers offer the advantage of capturing the combined effects of both SNPs and genes simultaneously, providing a more comprehensive view of the genomic landscape underlying diseases and traits. Moreover, k-mers enable the capture of cumulative effects from multiple genomic variants in a single analysis, including rare and structural variants. K-mer-based GWAS stands out prominently in studying microbial genomes, as exemplified by Sheppard et al.'s work on *Campylobacter* isolates using 30-bp DNA sequences (Sheppard et al., 2013; Power et al., 2017).

1.3 K-mer-based GWAS pose challenges that can be mitigated by adopting unitigs

While k-mer-based GWAS excels in identifying genomic variants undetectable by variant-based GWAS, interpreting results proves challenging due to mapping difficulties and high redundancy. Unitigs, as unique non-redundant genome sequences, mitigate these challenges (Chaguza et al., 2020, 2022a). Longer than the common k-mer size, unitigs cover more extensive genomic regions, providing additional context for identified genomic variants. Mapping unitigs back to the original genome is also simplified compared to k-mers, as each unitig represents a unique non-redundant region. Previous studies support the efficacy of unitig-based GWAS approaches, offering specific genomic information and facilitating functional annotation of associated loci across various bacterial genomes, including *Mycobacterium* (Jaillard et al., 2018; Hang et al., 2019; Yano et al., 2021), *Staphylococcus* (Jaillard et al., 2018; Chaguza et al., 2022b; Raineri et al., 2022), and *E. coli* (Denamur et al., 2022; Van Wonterghem et al., 2022).

1.4 Unitig-based GWAS can study quinolone resistance

Unitig-based GWAS can be applied to any phenotype, including quinolone resistance. Quinolones are a broad-spectrum family of antibiotics used to treat both gram-negative and gram-positive bacterial infections (Emmerson and Jones, 2003). Quinolone resistance is a significant concern as it can lead to urinary tract and intraabdominal infections. Therefore, a better understanding of quinolone resistance mechanisms is necessary to develop effective approaches to overcome this issue.

1.5 Quinolone resistance mainly results from chromosomal mutations in *gyrA* and *parC*

Quinolones primarily target bacterial DNA topoisomerase II and topoisomerase IV. These enzymes are crucial for DNA replication, transcription, and recombination, as well as for helping to under- and over-wind DNA (Naeem et al., 2016). Quinolones aim to inhibit DNA synthesis and cell growth by targeting these enzymes and inhibiting their activity. Topoisomerase II consists of GyrA and GyrB subunits, while topoisomerase IV comprises ParC and two ParE subunits. GyrA is homologous to ParC, and GyrB is homologous to ParE (Hooper and Jacoby, 2015). While known point mutations in these genes play a significant role, other biomarkers can also impact quinolone resistance.

1.6 Known and new biomarkers from our previous variant-based GWAS serve as a baseline for this study

In our previous variant-based GWAS study of quinolone resistance (Malekian et al., 2021), we confirmed known mutations in *gyrA* and *parC*, and we also identified new mutations, mainly in *valS* and *bdcA* genes. Using these findings as our baseline, we explore the efficacy of unitig-based GWAS to discover additional single and joint resistance mutations for quinolone resistance using the same dataset.

1.7 Approach overview

This study comprehensively analyzes 92 *Escherichia coli* (*E. coli*) genomes from a wastewater treatment plant in Dresden, Germany, and their corresponding quinolone resistance data. The approach includes extracting unitigs from the genomic data of the samples, applying quality control measures, and investigating their association with quinolone resistance labels for levofloxacin, norfloxacin, ciprofloxacin, and nalidixic acid. Subsequently, the unitigs are mapped back to a reference genome to determine the genes and mutations they encompass. Finally, the identified genes and mutations are investigated for their potential role in conferring resistance. An overview of our study is shown in Figure 1.

2 Methods

2.1 Sequence data and resistance phenotype

The dataset comprised 92 *E. coli* genomes from a municipal wastewater treatment plant in Dresden, Germany, and their antibiotic resistance values (i.e., the diameters of inhibition zones in the agar disk diffusion method) against 20 widely prescribed antibiotics, including four quinolones (levofloxacin, norfloxacin, ciprofloxacin, and nalidixic acid). The genomic data can be accessed directly from NCBI's assembly database under the project identifier PRJNA380388 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA380388/). Antibiotic resistance data is available within the Biosample field of the same dataset. For further details, (see Mahfouz et al., 2018).

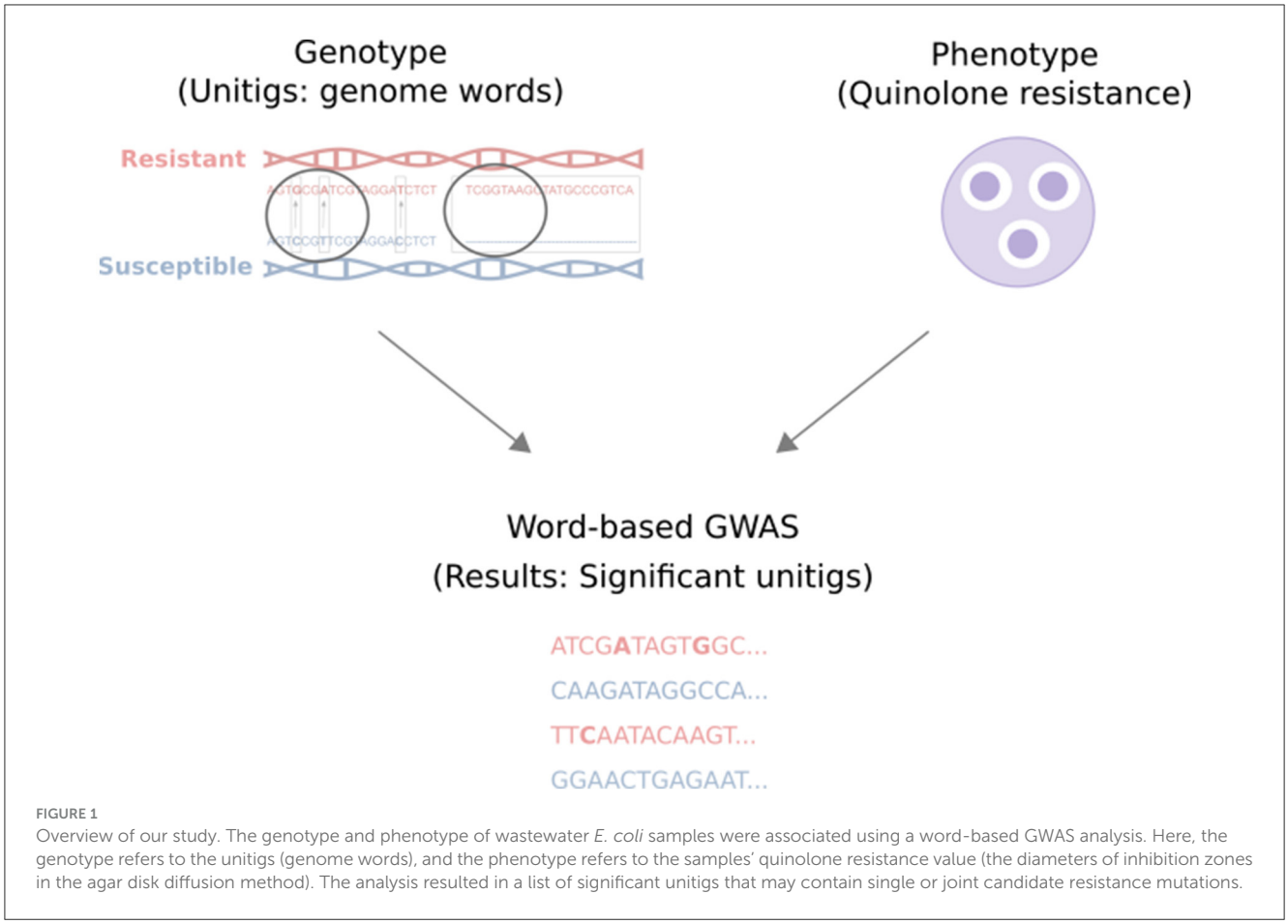

FIGURE 1
Overview of our study. The genotype and phenotype of wastewater *E. coli* samples were associated using a word-based GWAS analysis. Here, the genotype refers to the unitigs (genome words), and the phenotype refers to the samples' quinolone resistance value (the diameters of inhibition zones in the agar disk diffusion method). The analysis resulted in a list of significant unitigs that may contain single or joint candidate resistance mutations.

2.2 Identifying unitigs

To identify unitigs, we used the unitig-counter v1.1.0 software tool (Lees et al., 2020). The input for the unitig-counter was the genomic assembly of all our isolates in the FASTA format. The algorithm of this tool involves constructing a compressed de Bruijn graph from all the input genome sequences and identifying contiguous sequences of nucleotides, known as unitigs, from the resulting graph.

2.3 Association analysis

We used the fixed-effect generalized linear model (SEER) of the Pyseer tool v1.3.9 (Lees et al., 2018) to associate the presence/absence of unitigs with continuous antibiotic resistance values, similar to the approach used in our previous variant-based GWAS (Malekian et al., 2021). Population structure was controlled by adding covariates to the linear regression model. We utilized multidimensional scaling (MDS) on distances from the phylogenetic tree constructed from the VCF file using VCF-kit v0.1.6 (Cook and Andersen, 2017). Selected components from the MDS model, the number of components at the knee of a scree plot, served as covariates to control for population structure. Significant unitigs were retained after applying the Bonferroni correction. To calculate the Bonferroni-corrected threshold, we used the Pyseer tool, which divides the standard significance level of 0.05 by the number of distinct unitig patterns, resulting in a p-value threshold of 8.06×10^{-8}.

2.4 Unitig annotation

We mapped significant unitigs to our reference genome (*E. coli* K-12 MG1665, accession NC_000913.3) using BWA v0.7.17-r1188 (Li, 2013) and performed variant calling with Samtools v1.7 and Bcftools v1.8 (Danecek et al., 2021). The variants were annotated using SnpEff v5.1 (Cingolani et al., 2012). For variant-free unitigs associated with antibiotic susceptibility instead of resistance, we searched for resistance mutations in the complementing isolate set. The complementing isolate set comprises unitigs covering the same genomic regions and associated with antibiotic resistance (p-value < 0.05). Our final set of resistance mutations included mutations in both the mutated unitigs and the complementing isolate set of unmutated unitigs.

2.5 Functional analysis of annotated unitigs

We investigated the function of genes and variants mapped to our significant unitigs using the UniProt knowledgebase

(The UniProt Consortium, 2023) and the EcoCyc *E. coli* database (Keseler et al., 2010).

2.6 Network analysis of new genes

To investigate connections among new (mutated) genes not captured in our previous variant-based GWAS, we employed STRING-db (Szklarczyk et al., 2015). Default settings were utilized, which incorporate multiple sources of evidence such as text mining, experimental data, databases, co-expression, neighborhood, gene fusion, and co-occurrence. The edge thickness setting was adjusted to indicate the strength of data support.

2.7 Antibiotic resistance analysis of unitigs

To assess the association between the genes containing significant unitigs and antibiotic resistance, we employed the CARD database (McArthur et al., 2013) and relevant literature.

3 Results and discussion

This study examined correlations between unitigs and antibiotic resistance labels for four quinolones. The first objective was to validate the coverage of the positive controls, which consisted of known mutations in *gyrA* and *parC*, by the unitigs. The word-based GWAS was also expected to confirm previously identified mutations from the variant-based GWAS, particularly the mutations in the *bdcA* and *valS* genes. Moreover, the study aimed to explore the possibility of discovering new mutations or genes that the variant-based GWAS had not detected. Finally, but most importantly, the investigation aimed to uncover joint mutations that could potentially interact with each other.

3.1 A total of 54 highly-quality significant unitigs were identified

We extracted a total of 1,491,067 unitigs from our genomes. Initially, we excluded unitigs that appeared in over 99% of the dataset, resulting in a reduction to 1,209,817 unitigs, representing a 19% decrease. We then associated the presence/absence of these unitigs with resistance values for the four quinolones: levofloxacin, norfloxacin, ciprofloxacin, and nalidixic acid. Applying a Bonferroni-corrected threshold (p-value: 8.06×10^{-08}), we identified 100 highly significant unitigs: 33 for levofloxacin, 65 for norfloxacin, 17 for ciprofloxacin, and 1 for nalidixic acid. Some unitigs were significant for more than one antibiotic, resulting in a total count that differs from 100.

To understand the biological significance of the unitigs, we mapped them to the reference genome (*E. coli* K-12 substr. MG1655) to determine the associated genes and variants. Among the 100 highly significant unitigs, 54 possessed clear and high-quality annotations, constituting our final set of significant unitigs: 18 for levofloxacin, 36 for norfloxacin, 12 for ciprofloxacin, and 1 for nalidixic acid (see Supplementary material).

The 54 significant unitigs captured both some of the previously identified mutations and novel mutations in unexplored genes. Additionally, we observed joint mutations within unitigs, and their combinations exhibited significant correlations with quinolone resistance. A summary of our findings is presented in Table 1, indicating the potential utility of word-based GWAS in identifying genomic mutations that may be associated with antibiotic resistance. We will elaborate on each of our new findings in the following.

3.2 Significant unitigs captured known mutations

By mapping our genomes to the reference, *E. coli* K-12 substr. MG1655, we successfully identified significant unitigs harboring known biomarkers of quinolone resistance. Notably, these unitigs encompassed mutations in key genes such as *gyrA* and *parC*. For levofloxacin resistance, we observed mutations including *parC* S80I and *gyrA* S83L. Similarly, for norfloxacin resistance, we found *parC* S80I, *parE* L416F, *gyrA* S83L, and *gyrA* D87N. In the case of ciprofloxacin resistance, the identified known mutations were *parC* S80I, *gyrA* S83L, and *gyrA* D87N. Additionally, *gyrA* S83L was associated with resistance to nalidixic acid. Detailed information about these unitigs, including p-values, effect sizes, frequencies, and more, can be found in Supplementary material.

Additionally, our top significant unitigs were located within the quinolone resistance-determining regions (QRDRs) of GyrA. The QRDRs are amino acids between 67 and 106, which are conserved regions involved in DNA binding and are well-known to cause quinolone resistance when mutations occur (Varughese et al., 2018). The QRDR is near tyrosine 122, which binds to DNA during DNA double-strand breaks or DNA single-strand rejoins. Non-synonymous mutations (*gyrA* S83L, *gyrA* D87N) and synonymous mutations (*gyrA* R91R, *gyrA* V85V) were observed in this region. Among the unitigs in this region, those without any mutations were correlated with antibiotic susceptibility, while those with mutations were correlated with antibiotic resistance.

3.3 Significant unitigs captured some of the mutations previously identified in our variant-based GWAS study

Our previous variant-based GWAS analysis (Malekian et al., 2021) discovered significant correlations between specific variants in the *valS* and *bdcA* genes, which are involved in translation and biofilm formation, respectively, and quinolone resistance. Building upon this finding, our subsequent GWAS analysis at the unitig level unveiled additional insights. Specifically, we

10.3389/fmicb.2023.1276332

TABLE 1 Summary results of word-based GWAS for quinolone resistance.

Description	Levofloxacin	Norfloxacin	Ciprofloxacin	Nalidixic acid
Known mutations	*parC* S80I *gyrA* S83L	*parC* S80I *parE* L416F *gyrA* S83L *gyrA* D87N	*parC* S80I *gyrA* S83L *gyrA* D87N	*gyrA* S83L
Mutations identified by variant-based GWAS	*valS* R733	*bdcA* G135S *valS* R733		
New (mutated) genes	*cheA* *flhB* *cheZ* *stfR* *yecF* *gnd*	*nrdD* *yicI* *yicH* *acrF* *yhdV* *acrE* *gspA* *ymfL* *lldP* *sgbU* *yiaM* *yiaN* *aldB* *mutM* *dut* *ptrB* *mdfA*	*hisI* *stfR* *ymfL*	
Genes with joint mutations	*cheA* *cheZ* *stfR* *valS* *gnd*	*gyrA* *nrdD* *valS* *yicH* *ymfL* *lldP* *yiaN* *yiaM* *aldB* *dut*	*gyrA* *stfR*	

Using word-based GWAS, we identified known mutations as well as some of the mutations previously discovered through our variant-based GWAS. Additionally, we found new mutations in previously unexplored genes. We also observed joint mutations within unitigs, and their combinations were significantly correlated with quinolone resistance.

found that two unitigs within the *valS* gene were strongly associated with resistance to both levofloxacin and norfloxacin, while one unitig within the *bdcA* gene exhibited a significant correlation with norfloxacin susceptibility. These results further reinforce the importance of these genomic regions in developing quinolone resistance.

Within the unitigs associated with the *valS* gene, one unitig contained two synonymous mutations, namely *valS* R733R and *valS* A730A. Notably, the variant *valS* R733R had also been identified in our previous variant-based GWAS analysis. In contrast, the other unitig related to *valS* did not exhibit any variants, and thus, it displayed a correlation with antibiotic susceptibility rather than resistance. Similarly, the unitig linked to the *bdcA* gene did not include any variants and was associated with antibiotic susceptibility. It is intriguing to observe that this unitig encompassed the region harboring the G135S variant previously identified in *bdcA* among resistant isolates. These findings highlight the potential of word-based GWAS in identifying potential antibiotic resistance targets, expanding our understanding of the genomic landscape of quinolone resistance. For more comprehensive information on the significant unitigs in the *bdcA* and *valS* genes, including p-values, effect sizes, frequencies, and other details, please refer to Supplementary material.

3.4 Significant unitigs captured new mutations

Additionally, the word-based GWAS revealed mutations in new genes linked to quinolone resistance. The list of these new (mutated) genes can be found in Table 1. For more in-depth information regarding the unitigs containing these mutations, including p-values, effect sizes, frequencies, and other relevant details (see Supplementary material).

Some of these new (mutated) genes, such as *mdfA*, the AcrEF-TolC multidrug efflux system, *ptrB*, and *hisI* have been previously associated with antibiotic or quinolone resistance in the literature.

The *mdfA* gene encodes a multidrug efflux pump that, when upregulated, is strongly linked to resistance to several antibiotics, including quinolones (Yasufuku et al., 2011; Gu et al., 2021; Li and Ge, 2023). In this study, we found a region within the *mdfA* gene that displayed a strong association with norfloxacin susceptibility. This region was mostly unmutated in susceptible samples but mutated in resistant samples, with a synonymous variant, A42A.

The AcrEF-TolC multidrug efflux system is a homolog of the well-known AcrAB-TolC multidrug efflux system and is composed of three genes, *acrF*, *acrE*, and *tolC*. Prior research has demonstrated that overexpression of the *acrF* and *acrE* genes

TABLE 2 The (potential) role of new (mutated) genes in antibiotic resistance (AR).

Gene(s)	(Potential) role in AR	Description
mdfA	Multidrug efflux	Overexpression of the gene confers broad-spectrum antibiotic resistance, including resistance to fluoroquinolones, by actively pumping out antibiotics (Edgar and Bibi, 1997; Alcock et al., 2020).
acrE, *acrF*, *yhdV*	Multidrug efflux	The *acrF* and *acrE* genes are part of the AcrEF-TolC multidrug efflux system, and its overexpression is associated with multidrug resistance (Okusu et al., 1996; Alcock et al., 2020). The *yhdV* gene is in the same operon as *acrF* gene.
ptrB	Bacterial tolerance	Mutation in this gene enhances bacterial tolerance to stresses like ciprofloxacin by reducing pyocin production (Sun et al., 2014). Pyocins are bacteriocins, proteins bacteria produce to combat similar bacteria.
aldB	Bacterial persistence	The gene knockdown decreased *E. coli* persistence under some conditions (Kawai et al., 2018). Persistence refers to the ability of bacteria to survive lethal doses of antibiotics without genetic mutations.
yicI, *yicH*	Fitness to stress	These genes can improve the overall fitness of bacteria under stress conditions, but their role in antibiotic resistance is not explored yet (Répérant et al., 2011).
cheA, *cheZ*, *flhB*	Biofilm formation	Part of the *flhAB cheZYBR tap tarc cheWA motBA flhCD* gene cluster, which is involved in chemotaxis and biofilm formation (Tirumalai et al., 2019).
ymfL, *stfR*	Biofilm formation	The *ymfL* gene is within the prophage element e14, while the *stfR* gene is within the prophage element rac. Both e14 and rac phage remnants affect biofilm formation, as removing them from *E. coli* K-12 impairs biofilm production (Mehta et al., 2004; Fortier and Sekulovic, 2013).
yecF	Biofilm formation	Mutation in the *sdiA* gene, which belongs to the same operon as *yecF*, helps form thicker biofilm and higher motility than the wild type and complemented strains (Culler et al., 2018).
hisI	Biofilm formation	The upregulation of this gene, involved in amino-acid and metabolite transport, along with other genes, likely contributes to antibiotic resistance in *E. coli* biofilms (Ranjith et al., 2017).
nrdD	Biofilm formation	This gene, which is essential for DNA synthesis and repair, was upregulated after prolonged exposure to biocides that caused biofilm development and inhibited motility (Merchel Piovesan Pereira et al., 2020).
gpsA	Biofilm formation	The gene's role in antibiotic resistance remains unclear but has been studied in biofilm formation. Mutants lacking this gene showed a significant negative impact on biofilm formation (Qin et al., 2019).
yiaM, *yiaN*	Biofilm formation	These genes belong to the *yiaMNO* gene cluster. Deleting the *yiaMNO* genes in *E. coli* led to significant alterations in its growth pattern, ability to survive in high-salt conditions, and the formation of biofilms (Plantinga et al., 2005).

We found a potential link to AR for 19 new (mutated) genes out of 25, presented here.

results in antibiotic resistance, including fluoroquinolones (Ma et al., 1993; Okusu et al., 1996; Lau and Zgurskaya, 2005). Our study found one unitig in the *acrE* gene and one unitig in the *acrF* gene strongly associated with norfloxacin susceptibility.

Overexpression of the *hisI* and *ptrB* genes has been shown to increase ciprofloxacin resistance in *Escherichia coli* (Ranjith et al., 2017) and *Pseudomonas aeruginosa* (Sun et al., 2014), respectively. This study found a non-synonymous mutation, L46I, and a synonymous mutation, T52T, in the *hisI* gene correlated with ciprofloxacin resistance. However, for the *ptrB* gene, we found a non-synonymous mutation, V629I, was linked to norfloxacin resistance, which is in the relative vicinity of the predicted active sites at positions 617 and 652; refer to UniProt ID P24555.

Furthermore, while the direct association of the remaining new (mutated) genes with antibiotic resistance was not explored in the literature, we identified evidence indicating potential links for 19 out of the 25 new (mutated) genes (refer to Table 2). As the table shows, these genes are actively involved in crucial functions such as multidrug efflux, bacterial tolerance, bacterial persistence, fitness to stress, and biofilm formation. Thus, they present plausible candidates that could potentially be linked to antibiotic resistance.

Moreover, the involvement of new (mutated) genes in antibiotic resistance through interactions with other genes is plausible, as demonstrated in Figure 2. The interaction network in this figure shows the potential collaborative relationships among these genes. For instance, genes like *gnd*, *dut*, and *mutM*, which lacked evidence of a direct link to antibiotic resistance, might still confer resistance when cooperating with the *hisI* and *nrdD* genes, possibly aiding in biofilm formation. Similarly, the *sgbU* gene could contribute to antibiotic resistance in partnership with the *yiaM* and *yiaN* genes, potentially through biofilm formation mechanisms.

3.5 Significant unitigs captured joint mutations

Significant unitigs captured joint mutations that could interact with each other to drive quinolone resistance. Unlike conventional variant-based GWAS, which only considers individual mutations, word-based GWAS analyzes larger genome portions, enabling the detection of mutation interactions. The word-based GWAS identified highly significant unitigs for quinolone resistance that contain multiple variants, suggesting that the interaction between these mutations contributes to the development of antibiotic resistance. Some of these variants were found in well-known targets of quinolone resistance, such as *gyrA* and *parC*, while others were located in new genes, including *galF*, *cheA*, *yiaM*, and *cheZ*. Notably, these variations were in crucial positions, such as the quinolone-resistance determining regions for *gyrA* (Varughese et al., 2018), the substrate binding site neighborhood for *gnd* (UniProt ID P00350), or an essential catalytic domain for the *cheA* (UniProt ID P07363), *yiaM* (UniProt ID P37674), and *cheZ*

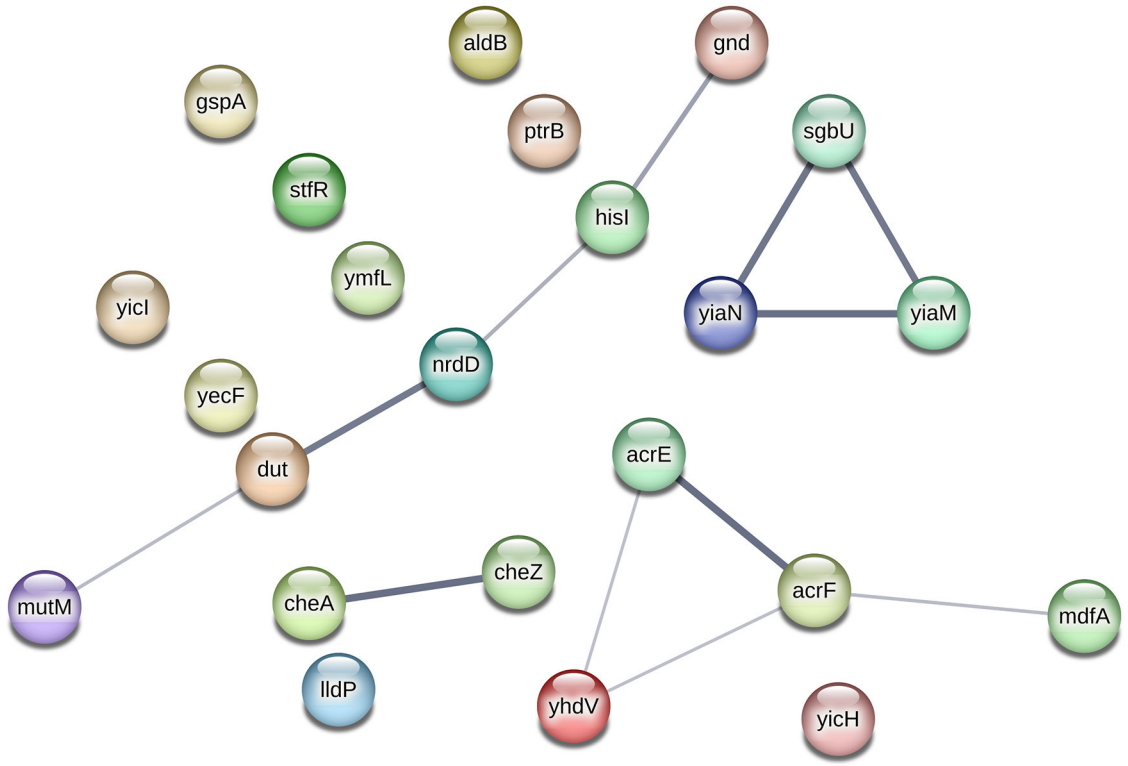

FIGURE 2
Network analysis of new (mutated) genes discovered by word-based GWAS, using STRING database (Szklarczyk et al., 2015). The edge thickness indicates the strength of data support. The network analysis shows evidence of interactions between four groups of genes: (1) the *gnd*, *hisI*, *nrdD*, *dut*, and *mutM* genes, (2) the *cheA* and *cheZ* genes, (3) *sgbU*, *yiaN*, and yiaM, and (4) the *acrE*, *yhdV*, *acrF*, and *mdfA* genes. However, there is no evidence of interactions among other genes.

(UniProt ID P0A9H9) genes. For a full list of genes that contain joint mutations, refer to Table 1. For further details on unitigs that contain such mutations, such as *p*-value, effect size, frequency, etc., (see Supplementary material).

3.6 General discussion

3.6.1 Word-based GWAS using unitigs yielded significant quinolone resistance findings

The word-based GWAS using unitigs was highly effective for analyzing quinolone resistance. Comparing it to k-mer level analyses (results not provided here), we found unitig analysis superior in terms of interoperability and significant findings. We identified 54 unitigs containing regions covering known mutations in *gyrA* and *parC* genes, as well as previously identified mutations using our variant-based GWAS in *bdcA* and *valS*.

Additionally, we discovered new variants in previously unexplored genes, some of which have been linked to antibiotic resistance. However, further investigation is required to confirm these associations. Notably, these new genes were missed by both our previous variant-based GWAS (Malekian et al., 2021) and the positive selection (Malekian et al., 2022) analysis of *E. coli* for antibiotic resistance. Therefore, our unitig-based GWAS allowed for a comprehensive analysis of quinolone resistance, showcasing the value of this approach in identifying genomic factors associated with antibiotic resistance.

3.6.2 Word-based and variant-based GWAS do not fully overlap in their detection of single mutations

Word-based GWAS and variant-based GWAS did not identify exactly the same list of single mutations. The reason behind this relies on the way that unitigs are built. In the context of variant-based GWAS, cases consist of isolates harboring particular mutations, while controls encompass isolates devoid of these mutations. Conversely, in word-based GWAS, the landscape is more complex, encompassing unitigs that contain a specific single mutation, unitigs devoid of the mutation, and unitigs with the specific mutation alongside other mutations. Consequently, when focusing on the detection of single mutations, the results from variant-based GWAS are considered more reliable.

3.6.3 A significant portion of the identified resistance mutations are synonymous

The list of significant resistant mutations contain many synonymous mutations. Unlike non-synonymous mutations that directly impact the protein product, structure, and function, synonymous mutations exert their influence indirectly by affecting

processes such as splicing, RNA stability, RNA folding, translation, and cotranslational protein folding (Sharma et al., 2019). As a result, these synonymous mutations play an indirect yet vital role in shaping the phenotype of interest.

3.6.4 Word-based GWAS suggests specific mutations and their interactions drive quinolone resistance

Our study emphasizes the significance of mutations and their interactions over individual resistance gene presence in quinolone resistance. However, the factors influencing resistance may vary depending on the antibiotic under study. We detected joint variants in the known target *gyrA* and new genes (*galF*, *cheA*, *yiaM*, and *cheZ*), all situated in critical regions. These mutations occupy essential positions, such as quinolone-resistance determining regions for *gyrA*, substrate binding site neighborhood for *gnd*, and catalytic domains for *cheA*, *yiaM*, and *cheZ*.

3.6.5 While word-based GWAS demonstrates computational power, biological experimental validation remains essential to confirm findings

We conducted our analysis on an *E. coli* dataset of 92 samples from a wastewater treatment plant, which might initially appear modest in size. However, the dataset's capability to affirm the presence of positive controls, notably the *gyrA* and *parC* genes, instilled confidence in its reliability for conducting thorough statistical analyses and evaluating the efficacy of the word-based GWAS approach. This robust dataset, anchored in unbiased sequencing and antibiotic resistance measurement methods (Mahfouz et al., 2018), underscores the comprehensiveness of our study. Word-based GWAS in bacterial studies can unveil genomic sequence patterns associated with bacterial phenotypes beyond single-nucleotide variations, providing efficient phenotype predictions and valuable functional insights. However, experimental validation is necessary to affirm the biological findings.

4 Conclusion

This study utilized a word-based GWAS to overcome the limitations of conventional variant-based GWAS analyses for genomic data. By examining genome words in 92 wastewater *E. coli* genomes, the study identified 54 significant words strongly associated with quinolone resistance. Positive controls, including known mutations in *gyrA* and *parC*, were validated, along with previously identified mutations in *bdcA* and *valS* from variant-based GWAS. Additionally, novel (mutated) genes such as *mdfA*, the acrEF-TolC multidrug efflux system, *ptrB*, and *hisI* were discovered, which are known to contribute to antibiotic resistance. Notably, the study revealed potentially interacting mutations in 14 genes, one of them being the well-known quinolone target *gyrA*. These mutations are located in critical sites, including quinolone-resistance determining regions in *gyrA*, the neighborhood of the substrate binding site in *gnd*, and the catalytic domains of *cheA*, *yiaM*, and *cheZ*. This finding suggests that quinolone resistance may not only result from individual mutations identified by variant-based GWAS but also from potential interactions between mutations.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA380388/.

Author contributions

NM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing. SS: Data curation, Formal analysis, Investigation, Validation, Writing—original draft, Writing—review & editing. AA-F: Writing—review & editing, Formal analysis. MS: Formal analysis, Writing—review & editing, Project administration, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by ACRAS-R and SNRT projects. The funder of both project is Bundesministeriums für Bildung und Forschung (BMBF).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1276332/full#supplementary-material

References

Alcock, B. P., Raphenya, A. R., Lau, T. T., Tsang, K. K., Bouchard, M., Edalatmand, A., et al. (2020). Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucl. Acids Res.* 48, D517–D525. doi: 10.1093/nar/gkz935

Chaguza, C., Jamrozy, D., Bijlsma, M. W., Kuijpers, T. W., van de Beek, D., van der Ende, A., et al. (2022a). Population genomics of group b streptococcus reveals the genetics of neonatal disease onset and meningeal invasion. *Nat. Commun.* 13, 4215. doi: 10.1038/s41467-022-31858-4

Chaguza, C., Smith, J. T., Bruce, S. A., Gibson, R., Martin, I. W., and Andam, C. P. (2022b). Prophage-encoded immune evasion factors are critical for staphylococcus aureus host infection, switching, and adaptation. *Cell Genom.* 2, 100194. doi: 10.1016/j.xgen.2022.100194

Chaguza, C., Yang, M., Cornick, J. E., Du Plessis, M., Gladstone, R. A., Kwambana-Adams, B. A., et al. (2020). Bacterial genome-wide association study of hyper-virulent pneumococcal serotype 1 identifies genetic variation associated with neurotropism. *Commun. Biol.* 3, 559. doi: 10.1038/s42003-020-01290-9

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPS in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Cook, D. E., and Andersen, E. C. (2017). Vcf-kit: assorted utilities for the variant call format. *Bioinformatics* 33, 1581–1582. doi: 10.1093/bioinformatics/btx011

Culler, H. F., Couto, S. C., Higa, J. S., Ruiz, R. M., J., Yang, M., et al. (2018). Role of sdia on biofilm formation by atypical enteropathogenic Escherichia coli. *Genes* 9, 253. doi: 10.3390/genes9050253

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of samtools and bcftools. *Gigascience*, 10, giab008. doi: 10.1093/gigascience/giab008

Denamur, E., Condamine, B., Esposito-Farèse, M., Royer, G., Clermont, O., Laouenan, C., et al. (2022). Genome wide association study of *Escherichia coli* bloodstream infection isolates identifies genetic determinants for the portal of entry but not fatal outcome. *PLoS Genet.* 18, e1010112. doi: 10.1371/journal.pgen.1010112

Edgar, R., and Bibi, E. (1997). Mdfa, an *Escherichia coli* multidrug resistance protein with an extraordinarily broad spectrum of drug recognition. *J. Bacteriol.* 179, 2274–2280. doi: 10.1128/jb.179.7.2274-2280.1997

Emmerson, A., and Jones, A. (2003). The quinolones: decades of development and use. *J. Antimicr. Chemother.* 51, 13–20. doi: 10.1093/jac/dkg208

Fortier, L.-C., and Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence* 4, 354–365. doi: 10.4161/viru.24498

Gu, Y., Huang, L., Wu, C., Huang, J., Hao, H., Yuan, Z., et al. (2021). The evolution of fluoroquinolone resistance in salmonella under exposure to sub-inhibitory concentration of enrofloxacin. *Int. J. Molec. Sci.* 22, 12218. doi: 10.3390/ijms222212218

Hang, N. T. L., Hijikata, M., Maeda, S., Thuong, P. H., Ohashi, J., Van Huan, H., et al. (2019). Whole genome sequencing, analyses of drug resistance-conferring mutations, and correlation with transmission of mycobacterium tuberculosis carrying katg-s315t in hanoi, vietnam. *Scient. Rep.* 9, 1–14. doi: 10.1038/s41598-019-51812-7

Hooper, D. C., and Jacoby, G. A. (2015). Mechanisms of drug resistance: quinolone resistance. *Ann. NY. Acad. Sci.* 1354, 12. doi: 10.1111/nyas.12830

Jaillard, M., Lima, L., Tournoud, M., Mahé, P., Van Belkum, A., Lacroix, V., et al. (2018). A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. *PLoS Genet.* 14, e1007758. doi: 10.1371/journal.pgen.1007758

Kawai, Y., Matsumoto, S., Ling, Y., Okuda, S., and Tsuneda, S. (2018). ALDB controls persister formation in *Escherichia coli* depending on environmental stress. *Microbiol. Immunol.* 62, 299–309. doi: 10.1111/1348-0421.12587

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Mu niz-Rascado, L., et al. (2010). ECOCYC: a comprehensive database of *Escherichia coli* biology. *Nucl. Acids Res.* 39, D583–D590. doi: 10.1093/nar/gkq1143

Lau, S. Y., and Zgurskaya, H. I. (2005). Cell division defects in *Escherichia coli* deficient in the multidrug efflux transporter acref-tolc. *J. Bacteriol.* 187, 7815–7825. doi: 10.1128/JB.187.22.7815-7825.2005

Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., and Corander, J. (2018). PYSEER: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* 34, 4310–4312. doi: 10.1093/bioinformatics/bty539

Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J., et al. (2020). Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio* 11, e01344–e01320. doi: 10.1128/mBio.01344-20

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-mem. *arXiv preprint arXiv:1303.3997*.

Li, Y., and Ge, X. (2023). Enhanced internal ionic interaction of MFS efflux pump MDFA contributes to its elevated antibiotic export. *Phys. Chem. Chem. Phys.* 25, 788–795. doi: 10.1039/D2CP05059E

Ma, D., Cook, D., Alberti, M., Pon, N., Nikaido, H., and Hearst, J. (1993). Molecular cloning and characterization of acra and acre genes of Escherichia coli. *J. Bacteriol.* 175, 6299–6313. doi: 10.1128/jb.175.19.6299-6313.1993

Mahfouz, N., Caucci, S., Achatz, E., Semmler, T., Guenther, S., Berendonk, T. U., et al. (2018). High genomic diversity of multi-drug resistant wastewater *Escherichia coli. Scient. Rep.* 8, 8928. doi: 10.1038/s41598-018-27292-6

Malekian, N., Agrawal, A. A., Berendonk, T. U., Al-Fatlawi, A., and Schroeder, M. (2022). A genome-wide scan of wastewater *E. coli* for genes under positive selection: focusing on mechanisms of antibiotic resistance. *Scient. Rep.* 12, 8037. doi: 10.1038/s41598-022-11432-0

Malekian, N., Al-Fatlawi, A., Berendonk, T. U., and Schroeder, M. (2021). Mutations in BDCA and VALS correlate with quinolone resistance in wastewater *Escherichia coli. Int. J. Molec. Sci.* 22, 6063. doi: 10.3390/ijms22116063

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The comprehensive antibiotic resistance database. *Antimicr. Agents Chemother.* 57, 3348–3357. doi: 10.1128/AAC.00419-13

Mehta, P., Casjens, S., and Krishnaswamy, S. (2004). Analysis of the *Lambdoid prophage* element e14 in the *E. coli* k-12 genome. *BMC Microbiol.* 4, 1–13. doi: 10.1186/1471-2180-4-4

Merchel Piovesan Pereira, B., Wang, X., and Tagkopoulos, I. (2020). Short-and long-term transcriptomic responses of *Escherichia coli* to biocides: a systems analysis. *Appl. Environ. Microbiol.* 86, e00708–e00720. doi: 10.1128/AEM.00708-20

Naeem, A., Badshah, S. L., Muska, M., Ahmad, N., and Khan, K. (2016). The current case of quinolones: synthetic approaches and antibacterial activity. *Molecules* 21, 268. doi: 10.3390/molecules21040268

Okusu, H., Ma, D., and Nikaido, H. (1996). Acrab efflux pump plays a major role in the antibiotic resistance phenotype of *Escherichia coli* multiple-antibiotic-resistance (mar) mutants. *J. Bacteriol.* 178, 306–308. doi: 10.1128/jb.178.1.306-308.1996

Plantinga, T. H., van der Does, C., Tomkiewicz, D., van Keulen, G., Konings, W. N., and Driessen, A. J. (2005). Deletion of the yiamno transporter genes affects the growth characteristics of *Escherichia coli* k-12. *Microbiology* 151, 1683–1689. doi: 10.1099/mic.0.27851-0

Power, R. A., Parkhill, J., and de Oliveira, T. (2017). Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* 18, 41. doi: 10.1038/nrg.2016.132

Qin, Y., He, Y., She, Q., Larese-Casanova, P., Li, P., and Chai, Y. (2019). Heterogeneity in respiratory electron transfer and adaptive iron utilization in a bacterial biofilm. *Nat. Commun.* 10, 3702. doi: 10.1038/s41467-019-11681-0

Raineri, E. J., Maaß, S., Wang, M., Brushett, S., Palma Medina, L. M., Sampol Escandell, N., et al. (2022). Staphylococcus aureus populations from the gut and the blood are not distinguished by virulence traits–a critical role of host barrier integrity. *Microbiome* 10, 239. doi: 10.1186/s40168-022-01419-4

Ranjith, K., Arunasri, K., Reddy, G. S., Adicherla, H., Sharma, S., and Shivaji, S. (2017). Global gene expression in *Escherichia coli*, isolated from the diseased ocular surface of the human eye with a potential to form biofilm. *Gut Pathog.* 9, 1–15. doi: 10.1186/s13099-017-0164-2

Répérant, M., Porcheron, G., Rouquet, G., and Gilot, P. (2011). The YICJI metabolic operon of *Escherichia coli* is involved in bacterial fitness. *FEMS Microbiol. Lett.* 319, 180–186. doi: 10.1111/j.1574-6968.2011.02281.x

Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., et al. (2019). A pan-cancer analysis of synonymous mutations. *Nat. Commun.* 10, 1–14. doi: 10.1038/s41467-019-10489-2

Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., et al. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in campylobacter. *Proc. Natl. Acad. Sci.* 110, 11923–11927. doi: 10.1073/pnas.1305559110

Sun, Z., Shi, J., Liu, C., Jin, Y., Li, K., Chen, R., et al. (2014). PRTR homeostasis contributes to pseudomonas aeruginosa pathogenesis and resistance against ciprofloxacin. *Infect. Immun.* 82, 1638–1647. doi: 10.1128/IAI.01388-13

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). String v10: protein-protein interaction

networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

The UniProt Consortium (2023). Uniprot: the universal protein knowledgebase in 2023. *Nucl. Acids Res*. 51, D523–D531. doi: 10.1093/nar/gkac1052

Tirumalai, M. R., Karouia, F., Tran, Q., Stepanov, V. G., Bruce, R. J., Ott, C. M., et al. (2019). Evaluation of acquired antibiotic resistance in *Escherichia coli* exposed to long-term low-shear modeled microgravity and background antibiotic exposure. *Mbio* 10, 10–1128. doi: 10.1128/mBio.02637-18

Van Wonterghem, L., De Chiara, M., Liti, G., Warringer, J., Farewell, A., Verstraeten, N., et al. (2022). Genome-wide association study reveals host factors affecting conjugation in *Escherichia coli. Microorganisms* 10, 608. doi: 10.3390/microorganisms10030608

Varughese, L. R., Rajpoot, M., Goyal, S., Mehra, R., Chhokar, V., and Beniwal, V. (2018). Analytical profiling of mutations in quinolone resistance determining region of GYRA gene among UPEC. *PLoS ONE* 13, e0190729. doi: 10.1371/journal.pone.0190729

Yano, H., Nishiuchi, Y., Arikawa, K., Ota, A., Miki, M., Maruyama, F., et al. (2021). Genome-wide association study reveals putative bacterial risk factors for cavitary mycobacterium avium complex lung disease. *bioRxiv* 2021–07. doi: 10.1101/2021.07.06.451401

Yasufuku, T., Shigemura, K., Shirakawa, T., Matsumoto, M., Nakano, Y., Tanaka, K., et al. (2011). Correlation of overexpression of efflux pump genes with antibiotic resistance in *Escherichia coli* strains clinically isolated from urinary tract infection patients. *J. Clin. Microbiol*. 49, 189–194. doi: 10.1128/JCM.00827-10