Closing report for National Research, Development and Innovation Office, Hungary (NKFIH) OTKA PD project Nr. 131839

Summary

In the reporting period, I and my research partners prepared and published several manuscripts and reached several goals. These include most of the original project's goals and many connected emerging ones. Our specific results are as follows:

1. A manuscript about the "Global map of evolutionary dependencies between antibiotic resistance and virulence genes in *E. coli*": I provide a detailed description of the main project in the form of a manuscript, while the actual manuscript is still under preparation and is planned to be sent to a high-impact journal by the end of 2023.
2. I provide a detailed list of the dissemination activity of the main project in the Dissemination of the results chapter.
3. Results of related activity: I list the results of further connected research projects.

Global map of evolutionary dependencies between antibiotic resistance and virulence genes in *E. coli*

Eszter Ari[1,2,3], Ágoston Hunya[1,3], Enikő Kiss[1], Bálint Vásárhelyi[1], Gergely Fekete[1], Tamás Stirling[1], Gábor Grézal[1], Barnabás Pintér[1,4], Chryso Christodoulou[1,5], Bálint Kintses[1,6,7], Balázs Papp[1,2]

1. Synthetic and Systems Biology Unit, Institute of Biochemistry, HUN-REN Biological Research Centre, Temesvári krt. 62, 6726 Szeged, Hungary
2. HCEMM-BRC Metabolic Systems Biology Research Group, Temesvári krt. 62, 6726 Szeged, Hungary
3. Department of Genetics, ELTE Eötvös Loránd University, Pázmány P. stny. 1/C, 1117, Budapest, Hungary
4. University of Cambridge, Trinity Ln, Cambridge CB2 1TN, United Kingdom
5. Imperial College London, SW7 2AZ London, United Kingdom
6. HCEMM-BRC Translational Microbiology Research Group, Temesvári krt. 62, 6726, Szeged, Hungary
7. Department of Biochemistry and Molecular Biology, Faculty of Science and Informatics, University of Szeged, Közép fasor 52, 6726, Szeged, Hungary

Abstract

Genes conferring antibiotic resistance or virulence phenotypes frequently undergo horizontal gene transfer in bacteria, contributing to the emergence of new multidrug-resistant pathogenic variants. Mounting evidence indicates that the genome content of the host influences the successful acquisition of such genes. However, the underlying evolutionary dependencies among specific genes, *i.e.* when one gene facilitates or hinders the acquisition of a second gene, remain poorly understood. Here we chart a high-resolution map of evolutionary dependencies between resistance and

virulence genes and resistance- or virulence-conferring mutations by phylogenetic analysis of ~ 9,000 human-associated *Escherichia coli* genomes. Our map reveals that resistance genes generally facilitate each other's gain in several independent lineages. The same is true for the minority of the key virulence genes. But contrary to a reported trade-off between resistance and virulence genes in other bacteria, we found no overall negative dependency between the acquisitions of key virulence and resistance genes, indicating largely independent evolution between these two traits in *E. coli*. We found that the virulence gene *iroN* is a general facilitator of both resistance and virulence gene acquisition. We also observed that the presence of enzymatic modificator and folate pathway antagonist resistance genes in a genome increases the chance of acquiring various other classes of resistance genes, making these genes indicators of potentially emerging new multidrug-resistant strains. Our results indicate that the evolutionary paths of resistance and virulence determinants might be predictable and could help to identify high-risk strains.

Introduction

The human microbiota is a complex ecosystem containing hundreds of beneficial commensal, opportunistic, and pathogenic bacteria species. These bacteria respond to changes in their environment – such as taking antibiotics – with continuous evolutionary adaptation, driven by the ability to *donate* and *accept horizontally transferred genes* from other lineages. The emergence of bacterial resistance, resulting from antibiotics' overuse, and the horizontal gene transfer of resistance genes, is responsible for an estimated 700,000 deaths worldwide annually (O'Neill, 2023). This crisis is primarily driven by the emergence of multi-drug resistant ESKAPEE bacteria (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter* species, and *Escherichia coli*) (Murray *et al.*, 2022).

Escherichia coli (*E. coli*) is a commensal bacteria in vertebrate- and human gut microbiome, and also an opportunistic pathogen, that primarily threatens immunocompromised and chronically ill patients by causing intra- and extraintestinal infections. Antibiotic-resistant *E. coli* is the leading cause of bloodstream and urinary tract infections, both in the community and healthcare settings (Cassini *et al.*, 2019). These pathogen bacteria contain virulence factor(s), produced by virulence gene(s). Virulence factors enable bacteria to damage their host by facilitating their attachment to cells, inhibiting the host's immune response or obtaining nutrition from the host.

The evolutionary pathways of *E. coli* strains can be estimated by the Warwick University multilocus sequence typing (MLST or ST) by analyzing specific genetic loci, commonly referred to as housekeeping genes or core genome genes (Zhou *et al.*, 2020). This analysis helps to classify and categorize *E. coli* strains into genetically closely related isolates, based on their genetic relatedness and evolutionary history. In several STs, specific pathogens (such as enterohaemorrhagic *E. coli*, enteropathogenic *E. coli*, and enteroinvasive *E. coli*) have emerged independently and repeatedly, while others contain only a few (Denamur *et al.*, 2021). Sequence typing can also be used to study the genetic basis of antimicrobial resistance in *E. coli*. By analyzing sequence types and associated resistance genes, it is possible to track the spread of resistant strains.

However, due to horizontal gene transfer, resistance genes from resistant microbiota and virulence genes from pathogens can recombine into a single host bacterium. As an example, it has been shown in a mouse colitis model that a resident commensal *E. coli* acquired the virulence plasmid *colicin p2* from a pathogenic *Salmonella enterica* strain

and became virulent (Stecher *et al.*, 2012). Therefore, highly virulent and untreatable infections may emerge and may become increasingly common even in healthy populations (Beceiro *et al.*, 2013; Wyres *et al.*, 2019). Emerging resistance of pathogenic bacteria is predicted to become a leading cause of death by 2050 unless novel antimicrobial strategies are developed (O'Neill, 2023).

Previous case studies have demonstrated the joint spread of resistance determinants and virulence-modulating genes (O'Neill, 2023; Giraud *et al.*, 2017). In sharp contrast, other researchers observed a scarcity of resistance gene acquisitions by virulent *Klebsiella pneumoniae* (Wyres *et al.*, 2019). Furthermore, it remains debated whether multi-drug resistant bacteria tend to emerge among opportunistic pathogens or not (Wyres *et al.*, 2019; Zhang *et al.*, 2015). Although several works focused on the emergence of 'superbugs' (*i.e.* pathogens which are not only multi-drug resistant but also hypervirulent), these researches did not investigate the evolutionary relatedness of such traits (Wyres *et al.*, 2019; Lam *et al.*, 2019). We know that pre-existing genome content variations influence the successful acquisition of resistance genes (Denamur *et al.*, 2021; Press *et al.*, 2016), and even the appearance of new resistance mutations (Coluzzi *et al.*, 2023), but no comprehensive study has been maintained to investigate the effect of the accessory gene content on the acquisition of new resistance and virulence genes. To be able to overcome the threat of 'superbugs', it is critical to understand how the evolution of resistance and virulence traits interact with each other, how the underlying genes are exchanged and what limiting factors influence these processes.

With the advent of whole-genome sequencing, genomic data has been accumulated that provides the basis for a phylogenomic approach to understanding the evolution of resistance and virulence. Therefore, here, we examine the evolutionary dependencies between resistance and virulence genes, *i.e.* when one gene facilitates or hinders the acquisition of the other. The dependency map makes it possible to investigate and even predict the evolutionary paths of increased resistance and virulence. We employed a statistical test specifically designed to assess whether character state changes are concentrated on certain branches of the phylogenetic tree (Maddison, 1990). Therefore this test, referred to as the 'concentrated-changes test', can be used to infer whether the presence of specific genes in the genome ('background genes') facilitates or hinders the acquisition of other specific genes ('tested genes'). We applied this test on the reconstructed gene gain and loss events along the phylogenetic trees of 25 different *E. coli* sequence types (STs). This analysis resulted in a global 'gene dependency map' over the entire history of the *Escherichia coli* species. Based on this gene dependency map, we found several recurring gene dependencies across multiple STs which indicate general rules of gene content evolution.

Results and Discussion

The gain-loss patterns of resistance and virulence genes

To understand the underlying evolutionary dependencies among specific resistance and virulence genes, we investigated genome evolution within 25 phylogenetically distinct clades (sequence types) within the *E. coli* species. The 25 phylogenetic clades defined by STs provide multiple independent instances of evolutionary diversification that we can exploit to uncover recurring patterns in gene content evolution. Note that the investigated clades represent various isolates with different lifestyles and virulence properties, including the ExPEC and UPEC pathotypes (*e.g.* ST69, ST73, ST95, and ST131) causing extra-intestinal and urinary tract infections in humans, domestic

mammals and birds, human pneumonia-associated *E. coli* strains (*e.g.* ST127), the avian pathogenic APEC pathotype (*e.g.* ST88 and ST117) isolated from birds, the Shiga toxin-producing STEC pathotype (*e.g.* ST11 and ST29) containing the *stx* virulence gene, and the enteropathogenic EPEC pathotype (*e.g.* ST10 and ST32) attaching and effacing lesions on intestinal epithelial cells in humans and domestic mammals (Denamur *et al.*, 2021).

We inferred maximum likelihood phylogenetic trees based on the recombination-free regions of the ~ 10,000 genomes of the 25 investigated *E. coli* STs. Specifically, we inferred a separate tree for each sequence type and dated them based on the isolate sampling dates (see the Phylogenetic reconstruction section of the Materials and Methods chapter). Reassuringly, all the dated trees showed evidence of a temporal signal confirming that we have reliably reconstructed the evolution of strains within STs (*Supplementary Table 1*). After excluding samples positioned on extremely long branches of the tree, altogether 9,010 genomes were retained for further analysis. We then inferred gain and loss events of 893 gene families (orthogroups) and mutations that represent previously described resistance and virulence factors (*Table 1*). To annotate the *E. coli* genomes and reconstruct such gain-loss events we used the resistance genes of the *ResFinder* database (Zankari *et al.*, 2012), resistance-conferring mutations (Caroff *et al.*, 2000; Yu *et al.*, 2009; Zankari *et al.*, 2017), key virulence genes – are considered as the most important virulence factors in *E. coli* – (Johnson *et al.*, 2019; Marin *et al.*, 2022), other virulence genes – contributing to some pathogenic traits of *E. coli* – from the *VirulenceFinder* database (Malberg Tetzschner *et al.*, 2020), and virulence-conferring mutations (Kalas *et al.*, 2017). We also mapped 4,192 non-resistance and non-virulence, so-called 'reference genes' from the *E. coli* reference genome (NC_000913) to the genomes. The annotation of reference genes allows us to compare the number of gain-loss events of resistance and virulence genes to the transfer pattern of reference genes. Changes in gene content were reconstructed using the maximum parsimony method along each of the 25 phylogenetic trees. Only those gene families and mutations were included in the downstream analyses that were gained at least twice independently on at least two different ST trees (*Table 1*).

Table 1: The number of investigated gene families and resistance- or virulence-conferring mutations included in the genome annotations and in the gene dependency analyses.

Type	Nr. of elements	Nr. of elements having at least 2 gain events in at least 2 of the 25 ST trees
Resistance	292	53
Resistance mutation	37	14
Key virulence	68	49
Other virulence	492	198
Virulence mutation	4	3
Reference	4,192	566
Summa	*5,085*	*883*

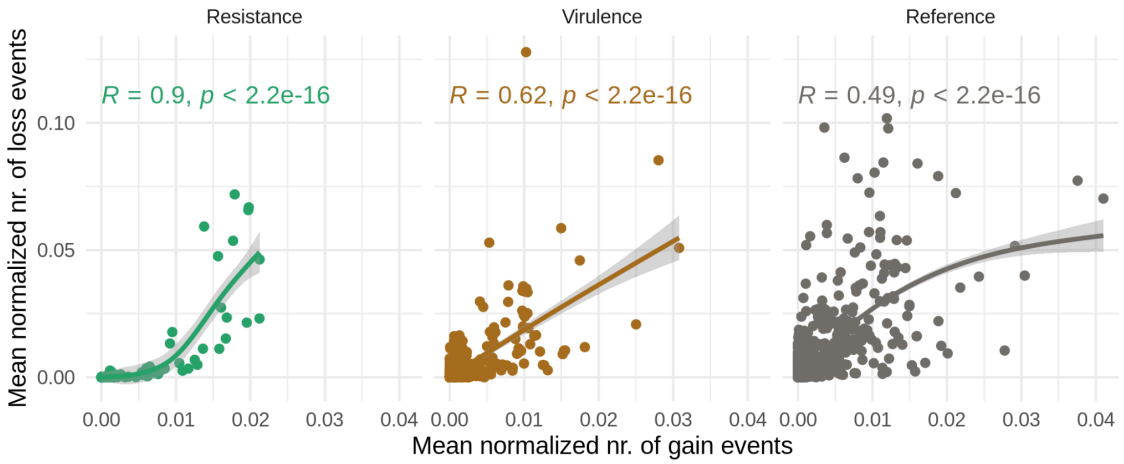

Figure 1: The mean – across 25 STs – number of inferred gene gain (*x*-axis) and loss (*y*-axis) events normalized by the size of the phylogenetic tree of each ST. The values of resistance genes are shown with green, virulence genes with brown and reference genes with grey points. Generalised additive models with integrated smoothness estimation regression fitting were applied for each gene type (curved lines with shading by confidence interval). The *R* and *p*-values of the Spearman correlation tests are written on the plot.

We found that the number of gain and loss events correlates (*Figure 1*), indicating that genes with high mobility undergo both many gains and many losses. Based on the mean number of gain and loss events normalised with the number of genomes appearing on each phylogenetic tree (*Table 2*), the overall mobility of genes in each sequence type can be estimated. We found large variations in mobility, with certain sequence types showing increased (*e.g.* ST10, ST48, ST88, and ST155), while others decreased overall gene mobility (*e.g.* in ST11, ST655, and ST2332). It is plausible that such variation in overall gene mobility is associated with the overall level of antibiotic resistance since we found increased mobility among the resistance genes compared to the other genes under investigation (*Table 2*).

We then asked whether different genes display different overall mobility. We found large variations in both gene loss and gene gain numbers across genes. The number of gains in the case of resistant genes was significantly higher compared to the number of gains of virulence (*p*-value = 5.791e-07) or reference genes (*p*-value < 2.2e-16), measured with the Mann-Whitney-U test (*Figure 2*). When comparing the number of losses we saw the opposite trend: the number of such events in the case of resistant genes was significantly lower compared to the number of losses of reference genes (*p*-value = 2.238e-08) and non-significantly but still lower (*p*-value = 0.203) compared to the virulence genes, indicating that once a resistance gene is acquired by the host it tends to remain there longer than the acquired virulence and reference genes. The increased number of gain and the decreased number of loss events among the resistance genes are in line with common knowledge about the spread of antibiotic resistance genes (Gladstone *et al.*, 2021; Murray *et al.*, 2022; Tadesse *et al.*, 2012).

Table 2: Number of genomes, normalised mean of gain and loss events of all, resistance (RES), virulence (VIR), and reference (REF) genes in each sequence type. We normalised the counts with the number of genomes appearing on the ST's phylogenetic tree.

Sequence type	Nr. of genomes	Normalized mean gain+loss events	Normalized mean gain+loss events among RES	Normalized mean gain+loss events among VIR	Normalized mean gain+loss events among REF
ST10	1175	0.011	0.020	0.011	0.003
ST11	939	0.004	0.002	0.001	0.001
ST16	144	0.006	0.011	0.003	0.001
ST21	585	0.006	0.011	0.004	0.001
ST29	160	0.006	0.006	0.003	0.001
ST32	216	0.005	0.007	0.002	0.001
ST38	249	0.009	0.017	0.007	0.003
ST48	159	0.017	0.030	0.010	0.005
ST58	264	0.010	0.018	0.015	0.004
ST69	411	0.007	0.009	0.007	0.002
ST73	665	0.006	0.009	0.004	0.001
ST88	127	0.014	0.027	0.014	0.003
ST95	604	0.006	0.012	0.003	0.001
ST101	206	0.008	0.024	0.011	0.003
ST117	194	0.008	0.022	0.009	0.002
ST127	173	0.007	0.016	0.005	0.003
ST131	1524	0.006	0.010	0.004	0.002
ST155	212	0.013	0.022	0.012	0.003
ST156	107	0.010	0.032	0.012	0.004
ST167	135	0.008	0.018	0.007	0.002
ST405	183	0.008	0.020	0.008	0.002
ST410	217	0.007	0.017	0.006	0.002
ST648	152	0.008	0.014	0.006	0.003
ST655	116	0.003	0.003	0.001	0.000
ST2332	93	0.002	0.003	0.001	0.001

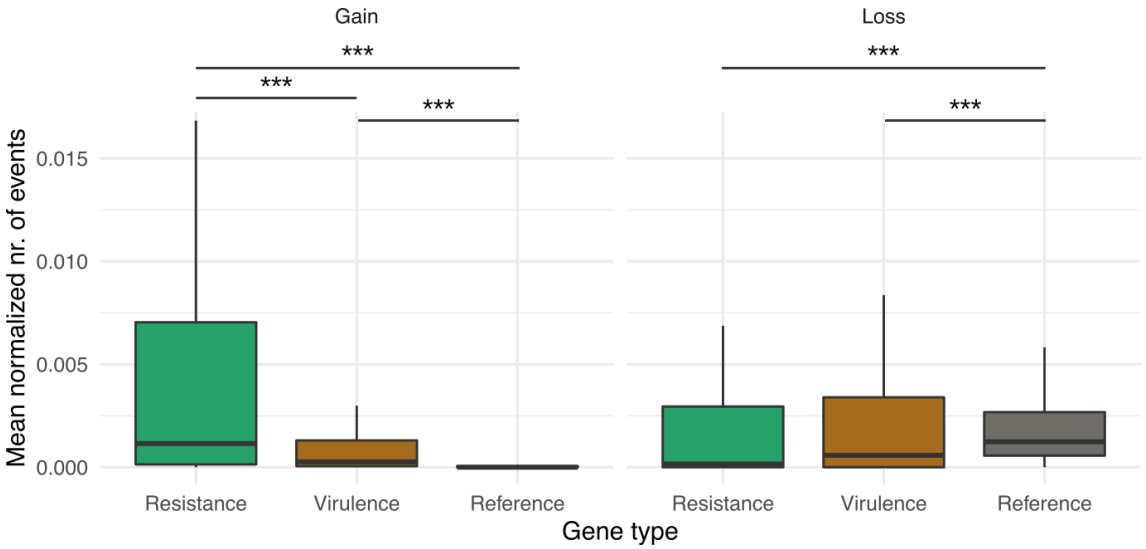

Figure 2: The mean – across 25 STs – number of inferred gene gain or loss events normalized by the size of the phylogenetic tree of each ST. The values of resistance genes are shown with green, virulence genes with brown and reference genes with grey. The levels of significant differences, calculated with the Mann-Whitney-U test are shown above the boxplots, where *** means that the *p*-value < 0.001.

The background-dependent gene acquisition map

Next, we inferred the background-dependent gene acquisition map by the calculating concentrated-changes test (see the Inferring the chronological order of ancestral gene gain events section of the Materials and Methods chapter) for gene gain events of more than 23 million gene pairs across the 25 investigated *E. coli* sequence types (*Supplementary Figures 1*). We found several gene dependencies that occur in at least two sequence types, indicating general rules of gene content evolution (*Figure 3*). Overall, the majority of such recurring gene dependencies were *positive*, suggesting that it is more common that the presence of a particular gene in a lineage is associated with an *increased* rather than a *decreased* likelihood of gaining a second specific gene or mutation. For example, where the *aph(3'')* aminoglycosides resistance genes were present on the phylogenetic trees, the *tet(B)* tetracycline-resistant efflux pump gene appeared in the descendants more frequently than expected, so there is a *positive interaction* (also referred to 'more gains than expected') between these genes (*Figure 4* and *Figure 5*). We were able to observe the very same positive dependency of the *aph(3'')* and the *tet(B)* genes in seven different sequence types. In fact, a statistically significant association was also found between these antimicrobial resistance genes by Gow *et al.* (2008).

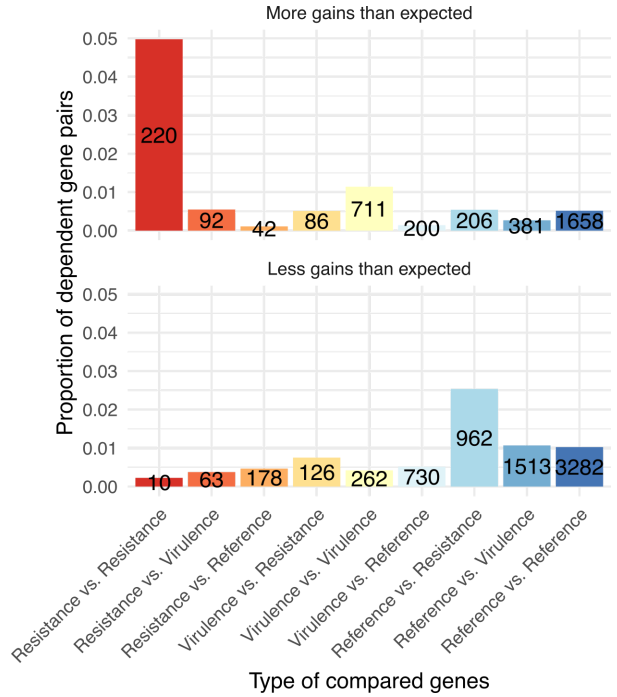

Figure 3: The proportion of dependent gene and mutation pairs recurring in at least two different sequence types. On the *x*-axis, we indicated the type of the background (first in the pair) and the tested (second in the pair) elements. To calculate the proportions we divided the counts of the recurring individual pairs (these numbers are written to the columns on the plot) with the number of possible pairs in the category pair using only those genes and mutations having at least two gain events in at least two of the 25 STs (as occurring in the third column of *Table 1*). *E.g.* the possible number of resistance background *vs.* resistance tested gene or mutation pairs is 67×67-67=4,422, and resistance background *vs.* virulence tested gene or mutation pairs is 67×250=16,750. The distribution of the number of sequence types the dependency pair recurring in is shown in Supplementary Figure 2.

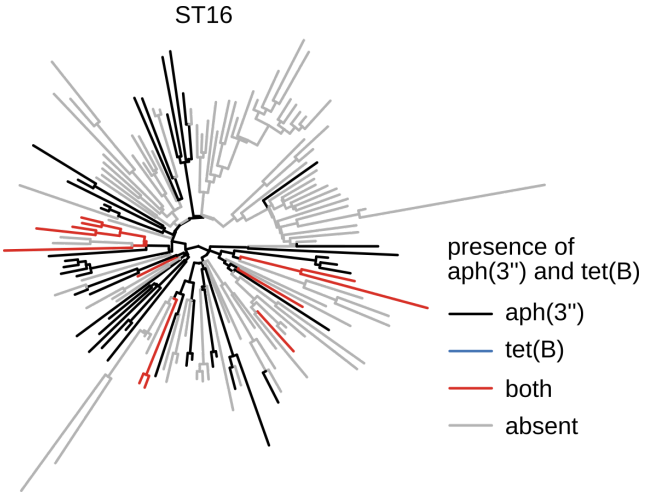

Figure 4: The inferred presence-absence pattern of the dependent resistance gene pair *aph(3'')* and *tet(B)* on the phylogenetic tree of ST16. Edges of the tree where the *aph(3'')* background gene was present without the *tet(B)* tested gene are coloured black. There are no edges where the *tet(B)* tested gene appeared without the *aph(3'')* background gene (would be blue). Edges where both investigated genes were present are coloured red, and edges where none of the investigated

genes were present are coloured grey. According to the concentrated-changes test, the *tet(B)* gene was gained more frequently in lineages where the *aph(3")* gene was already present in their genomes (*p*-value of more gains than expected = 0.039, odds ratio = 7.052). From the six independent gains of the *tet(B)* gene, five happened after the *aph(3")* gene appeared and in one case the two genes gained together.

A trade-off between the acquisition of resistance and virulence gene pairs

Prior studies focusing on *Klebsiella pneumoniae* isolates found a general negative relationship between resistance and virulence gene number (Lam *et al.*, 2019; Wyres *et al.*, 2019). This has been interpreted as a trade-off between the evolution of enhanced resistance *versus* enhanced virulence phenotype. We reasoned that the gene dependency map we provide offers an excellent opportunity to examine whether such a trade-off shapes the acquisition of resistance and virulence genes in *E. coli*. In the majority of cases, positive gene dependencies were observed between resistance-resistance genes, while resistance-virulence or virulence-resistance gene pairs were less frequent (*Figure 3*). Hence, the likelihood of an *E. coli* 'superbug' arising during evolution is lower than that of a multidrug-resistant or hypervirulent strain arising during evolution. Although, contrary to a reported trade-off between resistance and virulence genes in other bacteria, we found no overall *negative* dependency between the acquisitions of key virulence and resistance genes, indicating largely independent evolution between these two traits in *E. coli*. However, we found that the virulence gene *iroN* is a general facilitator of both resistance and virulence gene acquisition (*Figure 5*) and can be applied as an indicator of bacteria having the potential to become a superbug.

The interplay between antibiotic resistance and virulence determinants has been under investigation by multiple research groups. Selection acting on one trait can adversely affect other traits in evolutionary trade-offs. Antibiotic resistance is hypothesized to trade off with fitness in pathogenic microbes in the absence of antibiotics. Although studies of single resistance mutations support this hypothesis, it remains unclear whether trade-offs continue over time due to compensatory evolution and broader effects of genetic background (Basra *et al.*, 2018). Genomic evolutionary analyses indicated that the hypervirulent *Klebsiella pneumoniae* clones are subject to some sort of constraint for horizontal gene transfer (Lam *et al.*, 2019; Wyres *et al.*, 2019). Such examples of trade-off patterns between the appearance, expression, and function of resistance and virulence determinants were already demonstrated in multiple pathogenic bacterial species. *E.g.* mutations in efflux pumps of *Pseudomonas aeruginosa* cause fitness costs during growth, and these costs are linked to global effects on bacterial physiology, including reduced production of virulence factors; multi-resistant *Staphylococcus aureus* strains with intermediate-level resistance to vancomycin, decreased the expression of *surface protein A*, and reduced innate immune activation, and display lowered virulence (Geisinger and Isberg, 2017).

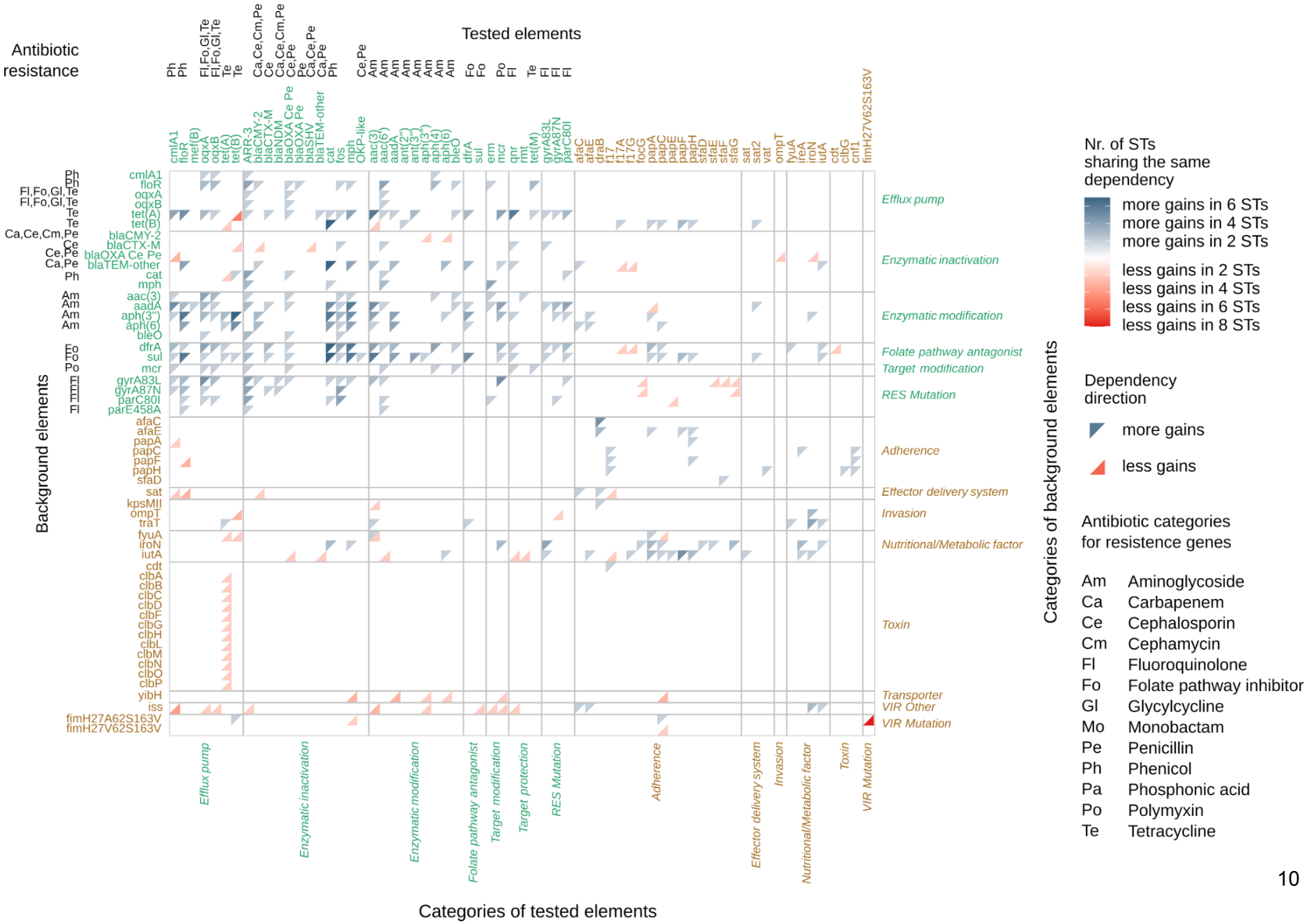
Antibiotic resistance
Tested elements
Ph Ph Fl,Fo,Gl,Te Fl,Fo,Gl,Te Te Te Ca,Ce,Cm,Pe Ce Ca,Ce,Cm,Pe Ce,Pe Pe Ca,Ce,Pe Ca,Pe Ph Ce,Pe Am Am Am Am Am Am Am Am Fo Fo Po Fl Te Fl Fl Fl
cmlA1 floR mef(B) oqxA oqxB tet(A) tet(B) ARR-3 blaCMY-2 blaCTX-M blaNDM blaOXA Ce Pe blaOXA Pe blaSHV blaTEM-other cat fos mph OKP-like aac(3) aac(6') aadA ant(2'') ant(3'') aph(3'') aph(4) aph(6) bleO dfrA sul erm mcr qnr rmt tet(M) gyrA83L gyrA87N parC80I afaC afaE draB f17 f17A f17G focG papA papC papE papF papH sfaD sfaE sfaF sfaG sat sat2 vat ompT fyuA ireA iroN iutA cdt clbG cnf1 fimH27V62S163V
Background elements
Ph Ph Fl,Fo,Gl,Te Fl,Fo,Gl,Te Te Te Ca,Ce,Cm,Pe Ce Ce,Pe Ca,Pe Ph Am Am Am Am Fo Fo Po Fl Fl Fl Fl
cmlA1 floR oqxA oqxB tet(A) tet(B) blaCMY-2 blaCTX-M blaOXA Ce Pe blaTEM-other cat mph aac(3) aadA aph(3'') aph(6) bleO dfrA sul mcr gyrA83L gyrA87N parC80I parE458A afaC afaE papA papC papF papH sfaD sat kpsMII ompT traT fyuA iroN iutA cdt clbA clbB clbC clbD clbF clbG clbH clbL clbM clbN clbO clbP yibH iss fimH27A62S163V fimH27V62S163V
Efflux pump
Enzymatic inactivation
Enzymatic modification
Folate pathway antagonist
Target modification
RES Mutation
Adherence
Effector delivery system
Invasion
Nutritional/Metabolic factor
Toxin
Transporter
VIR Other
VIR Mutation
Categories of tested elements
Categories of background elements
Nr. of STs sharing the same dependency
more gains in 6 STs
more gains in 4 STs
more gains in 2 STs
less gains in 2 STs
less gains in 4 STs
less gains in 6 STs
less gains in 8 STs
Dependency direction
more gains
less gains
Antibiotic categories for resistence genes
Am Aminoglycoside
Ca Carbapenem
Ce Cephalosporin
Cm Cephamycin
Fl Fluoroquinolone
Fo Folate pathway inhibitor
Gl Glycylcycline
Mo Monobactam
Pe Penicillin
Ph Phenicol
Pa Phosphonic acid
Po Polymyxin
Te Tetracycline

Figure 5: Resistance and key virulence gene acquisition dependencies recurred in at least two of the 25 investigated sequence types of *E. coli*. The names of the background genes and mutations are listed on the *y-axis* (left), and the names of the tested genes and mutations on the *x-axis* (top). These elements are grouped by their mechanism of resistance indicated by the ResFinder database, in the case of resistance genes (green) and by their VFDB category in the case of key virulence genes (brown) (right and bottom axis labels). The colour of box shading indicates the direction of the dependency where blue upper triangles indicate more gain events of the tested element than expected and red lower triangles indicate less gain events than expected. The intensity of box shading indicates the number of sequence types harbouring the same dependency, as per inset legends. In the case of resistance genes, the antibiotic categories of their resistance profile are also listed on the *y* and *x* axes, next to the gene names.

In a recent research, Coluzzi *et al.* (2023) investigated the evolution of resistance to quinolones antibiotics in *E. coli* genomes. They screened the genomes for quinolone-conferring resistance mutations in *gyrA*, *gyrB*, *parC* and *parE* genes and also inferred the presence of some resistance and iron uptake virulence genes on the phylogenetic tree. We were able to recover nearly half of the gene and mutation pair dependencies they found in at least two different sequence types. *E.g.* when either of the following genes: *iroA, B, C, D, E*, *N* (metal uptake virulence genes), *cea* (colicin toxin producing virulence gene), *cvaC* (microcin C toxin producing virulence gene), *racC, ralR, rcbA, recT, sgcQ* or *xisR* (reference genes) was present in the descendant genomes it increased the chance for gaining the *gyrA83L* fluoroquinolone-conferring resistance mutation in many STs (*Figure 5* and Supplementary figure 3).

Positive dependencies between the gain of resistance genes

We identified numerous positive interactions between resistance genes, meaning that if a resistance gene appears in the genome it increases the chance of gaining other resistance genes (*Figure 5*). This indicates that once a bacterium becomes resistant to an antibiotic, this event paves the way for the emergence of multidrug-resistant offspring. Our map reveals that resistance genes generally facilitate each other's gain (overrepresentation test $p = 0$, odds ratio = 13.3). Differences in the genetic background influence the evolution of antibiotic resistance, even under strong selection. However, few studies have identified chronologies between specific changes in the genetic background and the acquisition of antibiotic resistance (Wong, 2017). For example, the presence of the efflux pump *norA* in *Staphylococcus aureus* potentiated the subsequent evolution of point mutation conferring resistance to quinolones (Papkou *et al.*, 2020). Our results is in line with this finding and indicate that various other efflux pumps, like *cmlA1 (provides resistance to phenicols), floR (phenicols), oqxA* and *oqxB (fluoroquinolones, folate pathway inhibitors, glycylcyclines), tet(A)* and *tet(B)* (tetracyclines) increase the chance of the acquisition of further diverse resistance genes and *gyrA* mutations causing resistance to fluoroquinolones (*Figure 5*). Jangir el. al (2022) showed that those *E. coli*-s that evolved high-level colistin resistance by point mutations in the *lpxC* gene, increased the presence of a plasmid carrying the *mcr-1* colistin resistance gene also. We did not find these point mutations in the *lpxC* genes of the genomes we investigated but we observed that the presence of several other resistance genes (efflux pump *tet(A)*, enzymatic inactivator *blaTEM*, enzymatic modificators *aadA* and *aph(3")*, folate pathway antagonists *dfrA* and *sul*) and a resistance-conferring mutation of the *gyrA* gene increased the chances of the acquisition of the *mcr* gene in numerous STs. We also recovered the dependencies among well-known quinolone-conferring resistance mutation combinations of *gyrA83L*, *parC80I*, and *gyrA87N* (Marcusson *et al.*, 2009).

Besides these examples which are in line with the results of other studies, we detected several more dependent gene pairs among resistance genes that should be further investigated under experimental evolution conditions. For example, genes providing

resistance to aminoglycosides through enzymatic modifications (*aac(3)*, *aadA*, *aph(3")*, and *aph(6)*) emerge as general facilitators of further resistance gene acquisitions by showing positive interactions with efflux pumps (*cmlA1, floR, mef(B), oqxA, oqxB, tet(A)*, and *tet(B)*) and enzymatic inactivator genes (*ARR-3*, more beta-lactamases, *cat, fos,* and *mph*) in several STs.

Dependencies between virulence genes

According to our map, the gains of virulence genes are generally facilitated by each other, although the effect size is much lower (p = 1.28*10-51, odds ratio = 1.8) compared to the interdependency level of resistance genes. We have to mention that no other studies were found where gene acquisition dependencies of virulence gene pairs were examined, therefore our findings for example, the increased potential of gaining adherence and metabolic factor virulence genes when the *iutA* ferric aerobactin receptor gene is present, are unique.

Conclusion

In this study, we have created a comprehensive map of gene and mutation acquisition patterns of resistance, virulence and reference genes in *E. coli.* We performed the same analysis on 25 STs to test the idea that the genetic background could influence the success of gene acquisitions. We found many recurring positive dependencies among resistance genes and no general dependency pattern between resistance and virulence genes, providing a rich list of gene and mutation pairs for future experimental analysis. We observed that the presence of aminoglycoside resistance enzymatic modificator and folate pathway antagonist resistance genes in a genome increases the chance of acquiring various other classes of resistance genes, making these genes indicators of potentially emerging new multidrug-resistant strains. Overall, our results indicate that the evolutionary paths of resistance and virulence determinants might be predictable and could help to identify high-risk strains.

Throughout natural populations, bacterial genomes continually gain and lose genes, leading to a wide variety of latent phenotypes, which may enable certain lineages to acquire specific novel traits (such as antibiotic resistance) more quickly than others. The acquisition of certain resistance genes may increase the probability of resistance, either because they increase mutation probability or because they provide a less costly genetic background. It is well known that bacteria colonize diverse environments that differ in terms of selective pressures. Epistatic interactions lead to complex evolutionary patterns of adaptation while horizontal gene transfer occurs in bacteria with very different genetic backgrounds. Besides the epistatic interactions, certain dependencies may be explained by environmental factors, such as the chronological order of acquisition of resistance genes and the temporal order of antibiotic pressure. Alternatively, the effects might be ecological: certain lineages are more likely to encounter antibiotic pressure and hence tend to accumulate multiple resistance genes.

Making whole-genome sequencing data available to millions of bacterial strains is fundamentally changing clinical microbiology. Using such data combined with suitable methods can help us to understand the genetic basis of evolvability. With this knowledge in hand novel therapies can be developed to prevent the evolution of resistant and virulent bacterial strains.

Materials and Methods

Genomic data

Sources

We downloaded all available *E. coli* RefSeq genome sequences and metadata from NCBI RefSeq (O'Leary *et al.*, 2016) (18 815 genomes) and from the JGI IMG/M Integrated Microbial Genomes & Microbiomes (Chen *et al.*, 2023) (3 638 genomes) databases on 29 January 2020. We also added genome sequences and metadata of *E. coli* strains for which resistance and virulence phenotypes were measured and published in these two articles Moradigaravand *et al.* (2018) (1 936 unassembled genomes) and Johnson *et al.* (2019) (292 unassembled genomes). To assemble the genomes of the two later resources the *fastq* files were downloaded with *API* from the *European Nucleotide Archive* using their *BioSample* IDs. The reads were trimmed using the *Cutadapt* software (v3.2) (Martin, 2011) and assembled with the *SPAdes* program (v3.14.1) (Prjibelski *et al.*, 2020).

Filtering

The BUSCO software (v5.0.0) (Manni *et al.*, 2021) was used to exclude genome sequences with less than 95% of the enterobacterales_odb10.2019-04-24 BUSCO genes. Then, we merged genomes from the above four sources into a single database and excluded duplicates (having the same sequence or BioSample ID) by combining their connected metadata while keeping the genomes with higher BUSCO scores, longer sequences, and fewer contigs. We also excluded potentially non-*E. coli* genomes based on extremely long branches of a phylogenetic tree we inferred using the *RapidNJ* software (v.2.2.2) (Simonsen *et al.*, 2008) based on their core genes (ORFs were predicted by the *Prodigal* program (v2.6.3) (Hyatt *et al.*, 2010)) which were aligned using the MAFFT software (v7.475) (Katoh and Standley, 2013) and then concatenated. In total, we had 20,814 genomes to analyse (*Supplementary Table 2*).

We also downloaded further metadata of the samples (*e.g.* host disease and health state, isolation source and geographic location, serotype, serovar, and collection date) based on the *BioSample* IDs of the genomes *efetch* command of the *EDirect* software (v15.0) (Kans, 2023), as well as their antibiotic resistance profiles from the PATRIC database of the Bacterial and Viral Bioinformatics Resource Center (Wattam *et al.*, 2017).

Multilocus sequence typing

We applied the *MLST* software (v2.0.4) (Larsen *et al.*, 2012) in conjunction with the *ecoli* database (v2.0.0, available at: https://bitbucket.org/genomicepidemiology/mlst_db/src/master/ecoli/) to identify the multilocus sequence type (ST) of the downloaded *E. coli* genomes. Only those genomes belonging to major sequence types (that were represented in our dataset with over 100 genomes) were subjected to further phylogenetic analyses: ST10, ST11, ST12, ST16, ST17, ST21, ST29, ST32, ST38, ST48, ST58, ST69, ST73, ST88, ST93, ST95, ST101, ST117, ST127, ST131, ST155, ST156, ST167, ST405, ST410, ST443, ST648, ST655, and ST2332. The identifiers, sources, MLST types, and further metadata about the genomes and the samples are listed in *Supplementary Table 2*.

Annotation

To annotate the resistance, virulence and other, so-called 'reference genes' on the *E. coli* genomes we used the *blastx* command of the *Diamond* software (v2.1.0.154) (Buchfink *et al.*, 2021) with 'sensitive' searching algorithm, 90% coverage, and 90% identity thresholds. Resistance query proteins were downloaded from the *ResFinder* database (v2022.08.25) (Zankari *et al.*, 2012), and additional features of the genes were included in the metadata table (*Supplementary Table 3*) based on the *CARD Comprehensive Antibiotic Resistance Database* (v3.2.4) (Alcock *et al.*, 2020).

Virulence genes specific to *E. coli* were downloaded from the *VirulenceFinder* database (2 Dec. 2022) (Malberg Tetzschner *et al.*, 2020). Further sequences were added based on two publications: Johnson *et al.* (2019) and Marin *et al.* (2022), for the latter sequences were downloaded from the *VFDB* virulence factor database (18 Sep 2022) (Liu *et al.*, 2019). We identified the so-called 'key virulence genes' based on the Johnson *et al.* (2019) study. If needed, the downloaded DNA sequences were translated to proteins using a custom *Biopython* (Cock *et al.*, 2009) script. If a gene or protein was represented with more than one sequence in the downloaded dataset we merged its metadata to a single representative orthogroup (*Supplementary Table 3*) after the sequence similarity search step. The creation of these orthogroups was mostly based on the names of the genes and proteins and literature searching. When the case was not clear we also considered the results of the sequence similarity clustering calculated by the *MMseqs2* software (version: bdd169b3e285299cab792e62d60eb1f4e4e434d2) (Steinegger and Söding, 2017) (using the *easy-cluster* with minimum sequence identity 0.9).

To be able to estimate the baseline of the gene gain-loss events in the investigated *E. coli* genomes we annotated the orthologs of the protein-coding genes of the *E. coli* reference genome (NC_000913, GCF_000005845.2_ASM584v2). We referred to these genes as 'reference genes'. We added additional gene features to the reference genes (*Supplementary table 3*) by blasting their sequences to the *COG* database (25 Nov. 2020) (Galperin *et al.*, 2021) by the *Diamond* software.

We also aimed to identify and analyse the most important allelic variants causing resistant or virulent phenotypes. Therefore, we identified the following fluoroquinolone resistance-conferring mutations in the *gyrA* gene: S83A, S83L, D87G, D87N; in the *parC* protein: A56T, S57T, S80I, E84G; and in the *parE* gene: I355T, S458A, I529L (Zankari *et al.*, 2017); a β-lactam antibiotics (except carbapenems and cefepime) resistance-conferring mutation in the promoter region of the *ampC* gene: T-32A (Caroff *et al.*, 2000; Yu *et al.*, 2009); and allelic variants in the *fimH* virulence gene increasing the adherence ability of *E. coli*: 27A + 62A + 163V, 27A + 62S + 163A, 27A + 62S +163V, 27V + 62S + 163V (Kalas *et al.*, 2017). Here we indicated all allelic variations at the protein level, except for the promoter mutation of the *ampC* gene.

Based on the annotated genomes, we created a prevalence table (*Supplementary Table 4*) containing the number of each resistance, virulence, and reference orthogroup appearing in each *E. coli* genome.

Accuracy of identifying antibiotic resistance determinants from whole genome sequences

We validated our study's antibiotic resistance gene annotation with the phenotypic resistance data from Moradigaravand et al. (2018). Specifically, we assumed that the presence of a resistance determinant (i.e. annotated resistance gene or specific point mutation) to a given antibiotic class would predict resistance phenotype to the same

antibiotic class. Note that this corresponds to a rule-based approach to predict resistance phenotypes based on known causal resistance-conferring genes (Moradigaravand et al., 2018). Annotations of resistance genes belonging to five antimicrobial categories (aminoglycosides, cephalosporins, fluoroquinolones, folate pathway antagonists and penicillins) were compared with the phenotypic data in 1936 genomes shared between the two datasets. Precision and recall values of predicting resistance for each antibiotic class were calculated (Table 3).

Table 3: Precision and recall values for antibiotic resistance gene annotations

Antimicrobial category	True positives	True negatives	False positives	False negatives	Precision	Recall
aminoglycosides	309	1519	54	51	0.85	0.86
cephalosporins	381	1340	64	148	0.86	0.72
fluoroquinolones	437	1201	283	12	0.61	0.97
folate pathway antagonists	337	1150	425	21	0.44	0.94
penicillins	723	692	289	229	0.71	0.76

Multidrug-resistant (MDR) states were also inferred for the genomes. A genome was defined as MDR if it was non-susceptible to antimicrobial agents belonging to more than three antimicrobial categories: aminoglycosides, cephalosporins (including anti-MRSA cephalosporins, non-extended spectrum cephalosporins, 1st and 2nd generation cephalosporins, extended-spectrum cephalosporins, 3rd and 4th generation cephalosporins), carbapenems, cephamycins, fluoroquinolones, folate pathway inhibitors, glycylcyclines, monobactams, penicillins (including antipseudomonal penicillins + β-lactamase inhibitors and penicillins + β-lactamase inhibitors), phenicols, phosphonic acids, polymyxins, and tetracyclines (Magiorakos et al., 2012). All categories containing penicillins were merged into one ‘penicillin’ antimicrobial category, and all categories containing cephalosporins were merged into the ‘cephalosporin’ antimicrobial category, which is more in line with the less complex CARD Drug Class classification (Alcock et al., 2023) that we used to predict the MDR state for the annotated resistance genes and genomes.

Phylogenetic reconstruction

Excluding recombinant genomic regions

To reconstruct reliable phylogenetic trees – which were used as the backbone for further analyses – recombinant genomic regions had to be excluded. Therefore, we applied the Gubbins software (v3.3.0) (Croucher et al., 2015) with FastTree tree inference (Price et al., 2010), 100 iteration steps, and extensive search to identify recombinant regions of genomes belonging to each major sequence type. Then pseudo-whole-genome alignments were prepared using the Snippy software (v4.6.0) (Seemann, 2020), meaning that we aligned the genomes of each ST to a reference genome. We chose the genome with the longest contig as the reference genome in each ST. Duplicate sequences were then removed with the SeqKit software (v2.3.0) (Shen et al., 2016). The aligned pseudo-whole-genomes were masked with the maskfasta command of the BEDTools toolkit (v2.30.0) (Quinlan and Hall, 2010) using the coordinates of the recombinant genomic regions calculated by the Gubbins software. The masked alignments were further filtered with the SNP-sites program

(v2.5.1) (Page et al., 2016) such that we kept only those sites where at least one nucleotide was different from the majority consensus.

Inferring, pruning and rooting phylogenetic trees of each ST

To infer maximum likelihood phylogenetic trees from the nucleotide alignments we applied the *FastTree* software (v2.1.11) with the GTR-gamma nucleotide substitution model. To improve the reliability of the phylogenetic reconstructions we aimed to prune the extremely long branches – representing potentially misclassified genomes or genomes with undetected recombinant regions – from the trees. We tried the *Treeshrink* software (v1.3.9) (Mai and Mirarab, 2018) but some outlier long branches still remained on 19 (66%) of the ST trees. Therefore, we implemented the *IQR Tree Pruner R* script (https://github.com/barizona/iqr_tree_pruner) with which we were able to exclude extreme outlier tips having longer branches or root-to-tip distances than the upper fence $Q3 + 3 \times interquartile\ ranges$.

To find the root of each phylogenetic tree we implemented a three-step approach. In the first step, we used a modified version of the non-exported *.multi.rtt* function from the *treedater R* package (Volz and Frost, 2017) to identify a number of root position candidates using frequently applied root-to-tip objectives, *i.e.* highest correlation coefficient, highest R^2 value, lowest residual standard error. We collected the three best root positions for each objective, resulting in altogether 9 rooted trees for each sequence type. In the second step, we used the *dater* function from the *treedater* package to date each of the 9 trees – using the sampling dates of genomes – and calculated their log-likelihoods (without rerooting, using a strict molecular clock). In the third step, we selected the non-dated tree for which its dated counterpart had the highest log-likelihood among the 9 rooted trees.

After the rooting, we calculated and plotted the correlation between genome sampling dates and the distances between each tip and the root using the *BactDating* (v1.1.1) (Didelot *et al.*, 2018) *R* package. To check the reliability of the root we visually inspected all the trees and root-to-tip correlation plots then we excluded the phylogenetic trees of four STs (ST12, ST17, ST93, and ST443) due to spurious root position which coincided with unreliable correlation values (*p*-val > 0.01).

Inferring the chronological order of ancestral gene gain events

The gene gain and loss events were estimated along each phylogenetic tree of the remaining 25 STs using genome-gene prevalence tables and the *count* (v10.04) (Csűrös, 2010) software. Here the asymmetric Wagner parsimony method was applied (Csűrös, 2008) with the gain/loss penalty ratio = 2 (Mirkin *et al.*, 2003).

To identify gene pairs with a coupled acquisition order along the phylogenetic tree we re-implemented the method for testing the correlated evolution of two binary characters described in the study of Maddison (1990). The code was written in *R* and *Java* and is available as a *docker* file here: https://github.com/stitam/ecoli-hgt. During the so-called 'concentrated-changes test', we call one member of the investigated gene pair as *background* gene (orthogroup or mutation) and *tested* gene (orthogroup or mutation), referring to their function in the test. The concentrated-changes test is testing if gains of the tested gene unexpectedly concentrated on branches of the tree where the background gene was present (*Figure 6*). The test compares the actual number of gains of the tested gene to randomly distributed gains along the tree. Therefore, we obtained two *p*-values for each background and tested gene pairs: one reflects the

probability of more gains of the tested gene in the area where the background gene was present than expected and the other for fewer gains than expected. Odds ratios were also calculated to infer effect size.

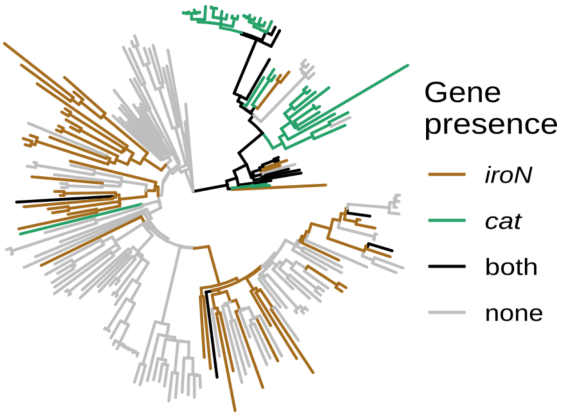

Figure 6: An example of the gain of the tested gene has concentrated those branches where the background gene was present on a phylogenetic tree. The phylogenetic tree of ST101 with the evolution of the *iroN*, a virulence factor involved in iron uptake (brown) and *cat*, a resistance gene to chloramphenicol. The branches where the *iroN* gene is present are coloured brown, where the *cat* gene is present are coloured green, where both genes are present are coloured black, and where none of them is present are coloured grey. According to the results of the concentrated-changes test the *cat* gene has been gained more frequently than expected (p = 0.008, odds ratio = 12.587) in those parts of the phylogenetic tree where the *iroN* gene was present in the ancestral genomes.

To keep the reliable coupled acquisition order only we filtered the results of the concentrated-changes test using the following criteria: i) Only those elements were considered in the role of background or tested which had at least two independent gain events along the trees of at least two different sequence types. ii) The sample size of the simulation is above 100. iii) The p-value of the 'less gains of the tested genes' or 'more gains of the tested genes' than expected is less than 0.05. iv) The absolute value of the difference between the expected and the actual number of branches where the background gene was present and the tested gene was gained is more than 1.5.

To further explore the genes with coupled acquisition we grouped them using their main characteristics (*e.g.* resistance or virulence genes, resistance mechanisms, virulence properties) and applied Fisher's exact tests.

We used the Nextflow pipeline management system (Di Tommaso *et al.*, 2017) to run the analysis for each ST from excluding the recombinant genomic region to the analysis of the results of the concentrated-changes test. The code is available at this repository: https://github.com/stitam/ecoli-hgt.

References

Alcock,B.P. *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **48**, D517–D525.

Alcock,B.P. *et al.* (2023) CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance

Database. *Nucleic Acids Res.*, **51**, D690–D699.
Basra,P. *et al.* (2018) Fitness tradeoffs of antibiotic resistance in extraintestinal pathogenic *Escherichia coli*. *Genome Biol. Evol.*, **10**, 667–679.
Beceiro,A. *et al.* (2013) Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clin. Microbiol. Rev.*, **26**, 185–230.
Buchfink,B. *et al.* (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.
Caroff,N. *et al.* (2000) Analysis of the effects of -42 and -32 ampC promoter mutations in clinical isolates of *Escherichia coli* hyperproducing AmpC. *J. Antimicrob. Chemother.*, **45**, 783–788.
Cassini,A. *et al.* (2019) Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.*, **19**, 56–66.
Chen,I.-M.A. *et al.* (2023) The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.*, **51**, D723–D732.
Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
Coluzzi,C. *et al.* (2023) Chance favors the prepared genomes: horizontal transfer shapes the emergence of antibiotic resistance mutations in core genes. 2023.06.20.545734.
Croucher,N.J. *et al.* (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.*, **43**, e15.
Csűrös,M. (2008) Ancestral reconstruction by asymmetric Wagner parsimony over continuous characters and squared parsimony over distributions. In, Nelson,C.E. and Vialette,S. (eds), *Comparative Genomics*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 72–86.
Csűrös,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.
Denamur,E. *et al.* (2021) The population genetics of pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.*, **19**, 37–54.
Di Tommaso,P. *et al.* (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
Didelot,X. *et al.* (2018) Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.*, **46**, e134.
Galperin,M.Y. *et al.* (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, **49**, D274–D281.
Geisinger,E. and Isberg,R.R. (2017) Interplay between antibiotic resistance and virulence during disease promoted by multidrug-resistant bacteria. *J. Infect. Dis.*, **215**, S9–S17.
Giraud,E. *et al.* (2017) Editorial: Antimicrobial resistance and virulence common mechanisms. *Front. Microbiol.*, **8**.
Gladstone,R.A. *et al.* (2021) Emergence and dissemination of antimicrobial resistance in *Escherichia coli* causing bloodstream infections in Norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. *Lancet Microbe*, **2**, e331–e341.
Gow,S.P. *et al.* (2008) Associations between antimicrobial resistance genes in fecal generic *Escherichia coli* Isolates from cow-calf herds in Western Canada. *Appl. Environ. Microbiol.*, **74**, 3658–3666.
Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
Jangir,P.K. *et al.* (2022) Pre-existing chromosomal polymorphisms in pathogenic *E. coli*

potentiate the evolution of resistance to a last-resort antibiotic. *eLife*, **11**, e78834.
Johnson,J.R. *et al.* (2019) Accessory traits and phylogenetic background predict *Escherichia coli* extraintestinal virulence better than does ecological source. *J. Infect. Dis.*, **219**, 121–132.
Kalas,V. *et al.* (2017) Evolutionary fine-tuning of conformational ensembles in *FimH* during host-pathogen interactions. *Sci. Adv.*, **3**, e1601944.
Kans,J. (2023) Entrez Direct: e-utilities on the Unix command line. In, *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).
Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
Lam,M.M.C. *et al.* (2019) Convergence of virulence and MDR in a single plasmid vector in MDR *Klebsiella pneumoniae* ST15. *J. Antimicrob. Chemother.*, **74**, 1218–1222.
Larsen,M.V. *et al.* (2012) Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.*, **50**, 1355–1361.
Liu,B. *et al.* (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.
Maddison,W.P. (1990) A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution*, **44**, 539–557.
Magiorakos,A.-P. *et al.* (2012) Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.*, **18**, 268–281.
Mai,U. and Mirarab,S. (2018) TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, **19**, 272.
Malberg Tetzschner,A.M. *et al.* (2020) *In silico* genotyping of *Escherichia coli* isolates for extraintestinal virulence genes by use of whole-genome sequencing data. *J. Clin. Microbiol.*, **58**, 10.1128/jcm.01269-20.
Manni,M. *et al.* (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of Eukaryotic, Prokaryotic, and Viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.
Marcusson,L.L. *et al.* (2009) Interplay in the selection of fluoroquinolone resistance and bacterial fitness. *PLoS Pathog.*, **5**, e1000541.
Marin,J. *et al.* (2022) The population genomics of increased virulence and antibiotic resistance in human commensal *Escherichia coli* over 30 years in France. *Appl. Environ. Microbiol.*, **88**, e00664-22.
Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
Mirkin,B.G. *et al.* (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **3**, 2.
Moradigaravand,D. *et al.* (2018) Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLOS Comput. Biol.*, **14**, e1006258.
Murray,C.J.L. *et al.* (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, **399**, 629–655.
O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
O'Neill,J. (2023) Review on antimicrobial resistance. *Rev. Antimicrob. Resist.*
Page,A.J. *et al.* (2016) SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics*, **2**, e000056.

Papkou,A. *et al.* (2020) Efflux pump activity potentiates the evolution of antibiotic resistance across *S. aureus* isolates. *Nat. Commun.*, **11**, 3970.
Press,M.O. *et al.* (2016) Evolutionary assembly patterns of prokaryotic genomes. *Genome Res.*, **26**, 826–833.
Price,M.N. *et al.* (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, **5**, e9490.
Prjibelski,A. *et al.* (2020) Using SPAdes de novo assembler. *Curr. Protoc. Bioinforma.*, **70**, e102.
Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
Seemann,T. (2020) Snippy: fast bacterial variant calling from NGS reads.
Shen,W. *et al.* (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE*, **11**, e0163962.
Simonsen,M. *et al.* (2008) Rapid neighbour-joining. In, Crandall,K.A. and Lagergren,J. (eds), *Algorithms in Bioinformatics*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 113–122.
Stecher,B. *et al.* (2012) Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proc. Natl. Acad. Sci.*, **109**, 1269–1274.
Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
Tadesse,D.A. *et al.* (2012) Antimicrobial drug resistance in *Escherichia coli* from humans and food animals, United States, 1950–2002. *Emerg. Infect. Dis.*, **18**, 741–749.
Volz,E.M. and Frost,S.D.W. (2017) Scalable relaxed clock phylogenetic dating. *Virus Evol.*, **3**, vex025.
Wattam,A.R. *et al.* (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.*, **45**, D535–D542.
Wong,A. (2017) Epistasis and the evolution of antimicrobial resistance. *Front. Microbiol.*, **8**.
Wyres,K.L. *et al.* (2019) Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLOS Genet.*, **15**, e1008114.
Yu,W. *et al.* (2009) AmpC promoter and attenuator mutations affect function of three *Escherichia coli* strains. *Curr. Microbiol.*, **59**, 244–247.
Zankari,E. *et al.* (2012) Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.*, **67**, 2640–2644.
Zankari,E. *et al.* (2017) PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J. Antimicrob. Chemother.*, **72**, 2764–2768.
Zhang,L. *et al.* (2015) Effects of selection pressure and genetic association on the relationship between antibiotic resistance and virulence in *Escherichia coli*. *Antimicrob. Agents Chemother.*, **59**, 6733–6740.
Zhou,Z. *et al.* (2020) The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res.*, **30**, 138–152.

Dissemination of the results

I have participated in the following international conferences, where I have presented and discussed the results of the main project in poster form.

1. Ari E, et al. (2023) Global map of evolutionary dependencies between antibiotic resistance and virulence genes in E. coli, EMBO Workshop: Predicting evolution, 11-14 July, Heidelberg, Germany
2. Ari E, et al. (2022) Global map of evolutionary dependencies between antibiotic resistance and virulence genes in *E. coli*. Lake Arrowhead Microbial Genomics Conference, 11-15 September, Lake Arrowhead, CA, USA
1. Ari E, et al. (2023) Global map of evolutionary dependencies between antibiotic resistance and virulence genes in E. coli, EMBO Workshop: Plasmids as vehicles of AMR spread, 12-18 September, Trieste, Italy, (online)

I gave oral presentations about the main project in national conferences:

1. Ari E, et al. (2022) Global map of evolutionary dependencies between antibiotic resistance and virulence genes in *E. coli*. 1st Bioinformatics and Data Science in Genomic Studies (BDG2022) online conference, November 25, University of Debrecen (online)

2. Ari E, et al. (2022) Global map of evolutionary dependencies between antibiotic resistance and virulence genes in *E. coli*. Bioinformatics, Celebrating the Hungarian Science, 11 Nov, HAS, Research Centre for Natural Sciences, Budapest (in Hungarian)

3. Ari E*, Kintses B (2020) Methods to investigate the microbiome and understand the results. Meeting of Hungarian Society for Gastroenterology, Section Colon, 6-7 March, Visegrád (in Hungarian; *invited talk)

I gave an invited talk about the project at “ELTEFeszt TTK” in Hungarian for high school students who are interested to learn biology at Eötvös Loránd University. The title of my presentation was: Will the *E. coli* become a superbug? (Lesz-e a coliból szuperbaktérium?) 2022, Budapest. The video is available on YouTube.

Results of related activity

In the reported research period I was involved in research projects in which we used the results of the main project, or in which I learned and applied new methods (*e.g.* the inferrance of dated phylogenetic trees and the creation of large databases including sequences and metadata) that later were applied in the main project.

Antibiotics of the future are prone to resistance in Gram-negative pathogens

With the Papp, Kintses and Pál groups (BRC, Szeged) we prepared a manuscript about antibiotic candidates currently under development. We proved that these new drugs are as prone to resistance evolution in Gram-negative pathogens (including *E. coli*) as clinically employed antibiotics. Here Gábor Grézal and I annotated resistance-conferring mutations on the very same *E. coli* genomes we used in the main project. Then we included this result in the main project as well. The manuscript is currently under review in Nature Microbiology. The preprint version can be found here:

Daruka L, Czikkely MS, Szili P, Farkas Z, Balogh D, Maharramov E, Vu TH, Sipos L, Vincze BD, Grézal G, Juhász Sz, Dunai A, Daraba A, Számel M, Sári T, Stirling T, Vásárhelyi BM, Ari E, Christodoulou C, Manczinger M, Enyedi MZs, Jaksa G, van Houte S, Pursey E, Papp CG, Szilovics Z, Pintér L, Haracska L, Gácser A, Kintses B, Papp B, Pál Cs: Antibiotics of the future are prone to resistance in Gram-negative pathogens, bioRxiv, 23 July, 10.1101/2023.07.23.550022, 2023

Retrospective evaluation of past waves of the SARS-CoV-2 epidemic in 2020 in Hungary

I have gained new skills by learning the inference of dated phylogenetic trees in a project where with my colleagues, we have analysed genome sequences of SARS-CoV-2 from the first two waves of the epidemic in 2020 in Hungary. This skill was applied when I inferred dated trees to root the phylogenetic trees of the 25 STs in the main project. The analysis of SARS-CoV-2 genomes reveals that the two waves markedly differed in viral diversity and transmission patterns. Despite the introduction of multiple viral lineages, extensive community spread was prevented by a timely national lockdown in the first wave. In sharp contrast, the majority of the cases in the much larger second wave can be linked to a single transmission lineage of the pan-European B.1.160 variant. Thus, despite its massive case number, the second wave showed lower viral diversity as compared to the first wave. We published the manuscript in the Virus Evolution journal:

Ari E, Vásárhelyi BM, Kemenesi G, Tóth GE, Zana B, Somogyi B, Lanszki Z, Röst G, Jakab F, Papp B, Kintses B (2022) A single early introduction governed viral diversity in the second wave of SARS-CoV-2 epidemic in Hungary. *Virus Evolution*, 8(2): veac069. **D1, IF: 5.3, IC: 1**

Population genetic and genomic analyses of Bronze Age communities from Western Hungary

Together with the talented Ph.D. student, Dániel Gerber under my and Anna Szécsényi-Nagy's supervision we examined 21 ancient shotgun genomes from Western Hungary, spanning the Late Copper Age Baden to the Bronze Age Somogyvár–Vinkovci, Kisapostag, and Encrusted Pottery cultures. Notably, the Kisapostag group displayed a significant Mesolithic hunter-gatherer ancestry despite expectations of dilution by the Early Bronze Age. This ancestry likely originated from two sources, including a previously unrecognized one in Eastern Europe, and contributed to various Bronze Age Central European and Baltic populations. I applied my general knowledge of evolution and analysis of genome sequences in this project, as I did to the main research. I also taught Dániel Gerber how to infer and plot phylogenetic trees. We published the manuscript with my shared correspondence in the Molecular Biology and Evolution journal:

Gerber D, Szeifert B, Székely O, Egyed B, Gyuris B, Giblin JI, Horváth A, Köhler K, Kulcsár G, Kustár Á, Major I, Molnár M, Palcsu L, Szeverényi V, Fábián S, Mende BG, Bondár M, Ari E*, Kiss V*, Szécsényi-Nagy A* (2023) Interdisciplinary analyses of Bronze Age communities from Western Hungary reveal complex population histories. *Molecular Biology and Evolution*, msad182 (preprint: *bioRxiv*, 2022, 10.1101/2022.02.03.478968) **D1, IF: 10.7, IC: 1 (*shared corresponding authorship)**

Effects of bowel cleansing on the composition of the gut microbiome

András Asbóth, a Ph.D. student under my and Bálint Kintses' supervision, and I contributed to a research project aimed to analyse the effect of before colonoscopy bowel preparation on the composition of the gut microbiome – containing more *E. coli* strains – in healthy people and patients having inflammatory bowel disease: Crohn's disease (CD) or ulcerative colitis (UC). Faecal microbiota structures of 19 healthy individuals, 9 CD patients, and 13 UC patients were determined by sequencing the V4 region of the 16S rRNA genes. The stool samples were collected right before the colonoscopy, 3 days, and 4 weeks after colonoscopy to assess the changes in the gut microbiota. We found that alpha diversity (reflecting the species richness of the gut microbiota) in CD patients was lower compared to the UC patients or the control group 3 days after the colonoscopy. While alpha diversity of UC patients was significantly higher than in CD patients or in the control group 4 weeks after colonoscopy. Our findings suggest that bowel preparation can result in a significant change in faecal microbial composition in patients with inflammatory bowel disease: microbial alterations recovering earlier in UC patients while reduced alpha diversity and altered abundance in CD patients may have a potential role in disease exacerbation after bowel cleansing. This study was published in a conference abstract book, and also in the Therapeutic Advances in Gastroenterology journal:

Rutka M, Szántó K, Bacsur P, Resál T, Jójárt B, Bálint A, Ari E, Kintses B, Fehér T, Asbóth A, Pigniczki D, Bor R, Fábián A, Farkas K, Maléth J, Szepes Z & Molnár T (2022) P713 Gut Microbiota Alterations after Bowel Preparation amongst Inflammatory Bowel Disease Patients. *Journal of Crohn's and Colitis*, 16 (Supplement_1): i609. **Q1, IF: 9.07**

Bacsur P, Rutka M, Resál T, Szántó K, Jójárt B, Bálint A, Ari E, Walliyulah A, Kintses B, Fehér T, Asbóth A, Pigniczki D, Bor R, Fábián A, Maléth J, Szepes Z, Farkas K, Molnár T (2023) Effects of bowel cleansing on the composition of the gut microbiome in inflammatory bowel disease patients and healthy controls. *Therapeutic Advances in Gastroenterology*, 16: 1-13. **Q1, IF: 4.2, IC: 1**

Downregulation of transposable elements extends lifespan

I was involved in a long going research with my colleagues at Eötvös Loránt University where we examined the role of transposable elements – that have some similar properties to the elements aiding the horizontal gene transfer in bacteria – in the lifespan of the nematode *Caenorhabditis elegans*. We revealed that the downregulation of active transposable element families and the ectopic activation of Piwi proteins in somatic cells can extend the lifespan of the nematode. Additionally, the increase in DNA N6-adenine methylation at transposable element stretches as the animal ages suggest a crucial role for this epigenetic modification in controlling the aging process, shedding light on the genetic factors influencing aging. We published the manuscript in the Nature Communications journal:

Sturm Á, Saskői É, Hotzi B, Tarnóci A, Barna J, Bodnár F, Sharma H, Kovács T, Ari E, Weinhardt N, Kerepesi C, Perczel A, Ivics Z, Vellai T (2023) Downregulation of transposable elements extends lifespan in Caenorhabditis elegans. *Nature Communications*, 14(1): 5278. **D1, IF: 16.6**

Creation of a transcription factor - target gene and signalling pathway databases

I was a leader of a project in which we created a transcription factor - target gene database, the TFLink (https://tflink.net/) together with a formal master student, Orsolya Liska who was under my supervision. Here I gained knowledge about how to use, create and maintain large databases containing sequences and information about the sequences. This knowledge was heavily applied in the main project when we downloaded and annotated 10,000 *E. coli* genomes. The TFLink gateway introduces a comprehensive resource that offers highly accurate information on transcription factor–target gene interactions (~12 million), as well as nucleotide sequences and genomic locations of transcription factor binding sites (~9 million sites) for human and six model organisms. This valuable database fills a crucial gap in providing user-friendly access to this wealth of data, offering interactive network visualizations, cross-links to other databases, and standardized regulatory information, making it a valuable tool for researchers across various fields, including functional genomics, evolutionary biology, and systems biology. We published the manuscript with my correspondence in the Database journal:

Liska O, Bohár B, Hidas H, Korcsmáros T, Papp B, Fazekas D, Ari E* (2022) TFLink: An integrated gateway to access transcription factor - target gene interactions for multiple species. Database, 2022, baac083. **D1, IF: 5.8, IC: 8 (*corresponding author)**

Together with Tamás Kadlecsik, a Ph.D. student under my and Tamás Korcsmáros' supervision, and other colleagues in the United Kingdom we created the Signalink3 database, which also directly used the data available in the TFLink. SignaLink3 is an extensive knowledge base that offers valuable insights into signalling pathways by providing manually curated data and integrating information from various databases for humans and three major animal model organisms. With over 400,000 newly added human protein-protein interactions, totalling 700,000 interactions for Homo sapiens, it stands as one of the largest integrated signalling network resources available. Notably, SignaLink3 uniquely incorporates gene expression data and subcellular localization information, enabling precise tissue- or compartment-specific pathway interaction analyses and more accurate modelling, making it a valuable resource for researchers in the field. With the shared first authorship of Tamás Kadlecsik, we published the manuscript in the Database issue of Nucleic Acid Research journal:

Csabai L*, Fazekas D*, Kadlecsik T*, Szalay-Bekő M, Bohár B, Madgwick M, Módos D, Ölbei M, Gul L, Sudhakar P, Kubisch J, Oyeyemi OJ, Liska O, Ari E, Hotzi B, Billes VA, Molnár E, Földvári-Nagy L, Csályi K, Demeter A, Pápai N, Koltai M, Varga M, Lenti K, Farkas IJ, Türei D, Csermely P, Vellai T, Korcsmáros T (2022) SignaLink3: A multi-layered resource to uncover tissue-specific signaling networks. *Nucleic Acids Research*, 50(D1): 701-709. **D1, IF: 14.9, IC: 13**