MICROBIAL GENOMICS

RESEARCH ARTICLE

Kolenda *et al.*, *Microbial Genomics* 2021;7:000743

DOI 10.1099/mgen.0.000743

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Genome placement of alpha-haemolysin cluster is associated with alpha-haemolysin sequence variation, adhesin and iron acquisition factor profile of *Escherichia coli*

Rafał Kolenda[1,*], Katarzyna Sidorczuk[2], Mateusz Noszka[3], Adrianna Aleksandrowicz[1], Muhammad Moman Khan[4], Michał Burdukiewicz[5], Derek Pickard[6] and Peter Schierack[4,7]

Abstract

Since the discovery of haemolysis, many studies focused on a deeper understanding of this phenotype in *Escherichia coli* and its association with other virulence genes, diseases and pathogenic attributes/functions in the host. Our virulence-associated factor profiling and genome-wide association analysis of genomes of haemolytic and nonhaemolytic *E. coli* unveiled high prevalence of adhesins, iron acquisition genes and toxins in haemolytic bacteria. In the case of fimbriae with high prevalence, we analysed sequence variation of FimH, EcpD and CsgA, and showed that different adhesin variants were present in the analysed groups, indicating altered adhesive capabilities of haemolytic and nonhaemolytic *E. coli*. Analysis of over 1000 haemolytic *E. coli* genomes revealed that they are pathotypically, genetically and antigenically diverse, but their adhesin and iron acquisition repertoire is associated with genome placement of *hlyCABD* cluster. Haemolytic *E. coli* with chromosome-encoded alpha-haemolysin had high frequency of P, S, Auf fimbriae and multiple iron acquisition systems such as aerobactin, yersiniabactin, salmochelin, Fec, Sit, Bfd and hemin uptake systems. Haemolytic *E. coli* with plasmid-encoded alpha-haemolysin had similar adhesin profile to nonpathogenic *E. coli*, with high prevalence of Stg, Yra, Ygi, Ycb, Ybg, Ycf, Sfm, F9 fimbriae, Paa, Lda, intimin and type 3 secretion system encoding genes. Analysis of HlyCABD sequence variation revealed presence of variants associated with genome placement and pathotype.

DATA SUMMARY

Sequences of 220 *E. coli* sequenced for this study are freely available from the NCBI BioProject database under accession number PRJNA725000.

INTRODUCTION

Escherichia coli is a versatile bacterium that colonizes intestines of clinically healthy mammals and birds [1]. Some *E. coli* are pathogenic, causing various intestinal and extra-intestinal diseases affecting humans and animals worldwide [2]. Most investigations of pathogenic *E. coli* over the last decades focused on the characterization of genes/operons associated with virulence [virulence-associated genes/factors (VAGs/VAFs)] and molecular virulence mechanisms of pathogenic *E. coli* [3]. On the basis of their characteristic VAGs and infection phenotypes, *E. coli* are divided into several pathotypes (reviewed in detail in [4]). Adhesins, iron acquisition systems and toxins constitute the majority of VAGs determining the ability of a particular *E. coli* to colonize the host [1]. Adhesins are responsible not only for binding and invasion of various cell types but also for interactions between bacteria, and between bacteria and abiotic surfaces [5, 6]. These interactions enable bacteria to form microcolonies and biofilms, colonize surfaces and attach to receptors expressed at the cell surface and their subsequent invasion [7–9]. Another group of

Received 24 May 2021; Accepted 10 November 2021; Published 23 December 2021

Author affiliations: [1]Department of Biochemistry and Molecular Biology, Faculty of Veterinary Medicine, Wrocław University of Environmental and Life Sciences, Wrocław, Poland; [2]Department of Bioinformatics and Genomics, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland; [3]Department of Microbiology, Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, Wrocław, Poland; [4]Institute of Biotechnology, Faculty Environment and Natural Sciences, BTU Cottbus-Senftenberg, Senftenberg, Germany; [5]Clinical Research Centre, Medical University of Białystok, Białystok, Poland; [6]Cambridge Institute for Therapeutic Immunology & Infectious Disease, University of Cambridge Department of Medicine, Cambridge, UK; [7]Faculty of Health Sciences, Public Health Campus Brandenburg, Brandenburg, Germany.

***Correspondence:** Rafał Kolenda, rafal.kolenda@upwr.edu.pl

Keywords: adhesins; alpha-haemolysin; *Escherichia coli*; genomics; haemolysin; toxins; virulence-associated genes.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary tables and eight supplementary figures are available with the online version of this article.

factors that mediate adhesion and/or invasion of cells includes type 3 and 6 secretion systems (T3SS, T6SS) [10, 11]. Iron acquisition and storage genes are important for the survival of bacteria in an environment with limited access to this element [12]. Many iron acquisition and storage systems have been identified in *E. coli* and can be divided into siderophores, iron transporters, haem and hemin uptake systems [13]. Toxins – haemolysins and proteases – are hypothesized to damage host cells and tissues, which leads to haemoglobin and haem release increasing iron availability [14].

Five haemolysins have been described so far in the literature for *E. coli* [15]. Two of them – bacteriophage-associated enterohaemolysin (Ehly) and haemolysin F (HlyF) – turned out to be recombinase RecT [16] and SDR (short-chain dehydrogenase/reductase) family oxidoreductase [17], respectively. Therefore, these will not be considered as haemolysins in this study. The other three haemolysins are alpha-haemolysin, enterohaemolysin (Ehx/EHEC-hly) and silent haemolysin (HlyE) [18, 19]. Enterohaemolysin is a pore-forming toxin, mainly associated with Shiga-toxigenic and enterohaemorrhagic *E. coli* (STEC and EHEC) [20]. It is encoded on the plasmid as a part of an operon consisting of four genes: *ehxA* (or *EHEC-hlyA*, encoding functional toxin), *ehxC* (or *EHEC-hlyC*, responsible for post-translational modification of EhxA), *ehxB* and *ehxD* (also known as *EHEC-hlyB* and *EHEC-hlyD*, responsible for the transport of EhxA through the inner membrane) [21]. Silent haemolysin is a pore-forming toxin as well and may be also referred to as haemolysin E, cytolysin A (ClyA) and silent haemolysin locus A (SheA). HlyE expressing *E. coli* causes lysis of mammalian erythrocytes, cytotoxicity to cultured mammalian cells, apoptosis induction in macrophages and reduces intracellular Ca^{2+} oscillations in epithelial cells [22].

Alpha-haemolysin can be encoded on plasmids or chromosomal pathogenicity islands [23]. Similar to Ehx, alpha-haemolysin is part of an operon consisting of four genes: *hlyA* (encoding functional toxin HlyA), *hlyC* (responsible for post-translational modification of HlyA), *hlyB* and *hlyD* (responsible for the transport of HlyA through the inner membrane). The role and importance of alpha-haemolysin in *E. coli* pathogenesis remain unclear so far. Alpha-haemolysin has been detected in various *E. coli* pathotypes and commensal *E. coli* but is mainly associated with extraintestinal pathogenic *E. coli* (ExPEC), and there are indications to implicate alpha-haemolysin in the pathogenesis of ExPEC pathotypes. Possible roles of alpha-haemolysin in *E. coli* pathogenesis include exfoliation of epithelial cells to grant access to underlying tissue sites, modulation of host immune response or cell signalling subversion and induction of apoptosis [21]. Different modes of pathogenic actions assigned to alpha-haemolysin depend on cell type, infection site environment and concentration of secreted haemolysin. *In vivo* studies with uropathogenic *E. coli* (UPEC) strains CFT073 and UTI89 showed other genetic factors influencing the outcome of alpha-haemolysin loss-of-function models and finally, undermining the 'importance' of investigated VAG [24].

Impact Statement

Since the discovery of haemolysis, many studies have been undertaken to analyse the role of this phenotype in *E. coli* along with its association with other virulence genes, diseases and pathogenic actions in the host but the results have been quite often inconclusive. Analysis of over 1000 haemolytic *E. coli* genomes allowed us to show that isolates with chromosome- and plasmid-encoded haemolysin differ in virulence factor prevalence and alpha-haemolysin sequence. Taking into consideration that all investigations concerning the role of alpha-haemolysin were conducted with the use of *E. coli* isolates with chromosome-encoded haemolysin, this is the first study to examine *E. coli* genomes with plasmid-encoded haemolysin, which have fewer iron acquisition systems and possess a different set of adhesion factors than bacteria with chromosome-encoded hemolysin.

Since the discovery of haemolysis, many studies have been undertaken to analyse the role of this phenotype in *E. coli* along with its association with other virulence genes, diseases and pathogenic actions in the host but the results have been quite often inconclusive. The reason for these inconsistent results can be associated with limited availability of virulence-associated factors to test or inconsistent selection of virulence factors used in testing. Our previous study focused on VAGs' association with the presence of haemolysin in *E. coli* isolated from healthy pigs [25]. Taking advantage of developments in sequencing technologies and the presence of publicly available *E. coli* genome sequences, we decided to expand the scope of our study and provide a global picture of the haemolytic property and its associated contributing VAGs/VAFs differing in haemolytic and nonhaemolytic *E. coli*. The objectives of this study were to analyse and characterize the diverse collection of 220 haemolytic and nonhaemolytic *E. coli* based on their respective genomes. The analysis was expanded by utilization of already available genomes in the GenBank database to acquire a global overview allowing us to further explore various genetic and sequence variations involved in the haemolytic ability possessed by *E. coli* leading to the revelation of novel virulence-associated genes/factors linked with alpha-haemolysin.

METHODS

Bacteria

Isolates (n=220) from diseased and clinically healthy humans, wild and domestic animals are listed in Table 1. To identify *E. coli* isolates, rectal (mammals) or cloacal (birds) swabs and urine samples were plated onto CHROMagar orientation agar [26]. After incubation at 37 °C, *E. coli* were initially identified as pink colonies. For further confirmation, these colonies were transferred onto Gassner Agar. Colonies appearing pink on CHROMagar and blue/yellow/green on Gassner Agar were

Table 1. Origin of *E. coli* used in this study

Species	Group name	No. of isolates	
		Haemolytic	Nonhaemolytic
Human			
Homo sapiens	Human, healthy*, †	12	19
Homo sapiens	Human, urinary tract infection‡	13	20
Domestic mammals			
Sus scrofa domestica	Domestic pig, healthy§	13	19
Wild mammals			
Capreolus capreolus	Roe deer¶, **	8	23
Martes sp.	Marten¶	6	15
Procyon lotor	Raccoon¶	10	17
Vulpes vulpes	Red fox¶	5	20
Wild birds			
Anas platyrhynchos	Mallard††	13	6
Total		81	139

*, †Sampled by Thomas Wex (Magdeburg, Germany) and Peter Schierack (Senftenberg, Germany).
‡Urine samples from patients with urinary tract infections collected by Steffen Vogel (Hoyerswerda, Germany) in the hospital in 2009.
§Samples from 18 different pig production units in eastern Germany collected by Peter Schierack from 2009 to 2010.
¶Samples collected from Lausitz (Lusatia), a region in south-eastern Germany, taken by Peter Schierack from the rectum (mammals) or cloaca (birds) of dead animals, which were collected directly as accident victims or delivered to Hermann Ansorge (Görlitz, Germany) and Olaf Zinke (Kamenz, Germany) between 2007 and 2011.
**Rectal samples were taken during several hunts by Peter Schierack between 2007 and 2010.
††Cloacal samples of wild birds were taken by Peter Schierack in the winter between 2007–2008 and 2010–2011.

assumed to be *E. coli* [27–29]. Haemolysis of each isolate was tested on 5% sheep blood agar by a clear (transparent) zone surrounding the colonies. All isolates from urine samples were considered as UPEC. A single *E. coli* colony (pink on CHROMagar orientation and haemolytic or nonhaemolytic on blood agar plates) was subcultured twice on CHROMagar orientation plates and stored in 15% glycerol at −80 °C.

Genome sequencing, assembly and annotation

Wizard Genomic DNA Purification Kit (Promega) was used to extract genomic DNA of 220 *E. coli* isolates. Extracted DNA was sequenced by HiSeq X platform at Sanger Institute Sequencing Facility. FastQ Screen was used to determine the quality of sequences [30]. Genome assemblies were acquired by utilising Shovill pipeline and assembly_improvement pipeline, followed by annotation using Prokka version 1.13.5 [31, 32]. Assembled genome sequences have been deposited at the NCBI under the BioProject ID: PRJNA725000.

Analysis of 220 *E. coli* genomes

Pangenomes were determined using Roary version 3.12 with a minimal percentage of identity for BLAST equal or higher than 95% and splitting the paralogues [33]. Core genome alignment provided by Roary analysis was used to obtain maximum-likelihood (ML) core-genome phylogeny with RAxML version 8.2.12 employing a GTRGAMMA substitution model and 100 bootstrap replications [34]. *In silico* serotyping was performed with SRST2 [35]. *E. coli* phylotypes were assigned by ClermonTyping [36]. Bayesian analysis of population structure (BAPS) was carried out by RhierBAPS [37]. Genome placement of *hlyCABD* operon was analysed with the use of BLAST by comparing reference sequences of 5′-upstream of *hlyC* and 3′-downstream of *hlyD* (Table S1, available in the online version of this article) [38]. Prevalence of VAGs and multilocus sequence typing (MLST) were determined by ARIBA version 2.13 and Virulence Factor Database Core Collection (VFDB) [39, 40]. Multiple genes encoding the same VAF were reduced to a single factor as shown in Table S2. Additionally, PCRs were carried out for detection of ten VAGs (*sitchr, traT, hra, sitep, malX, cvi/cva, iss, tia, ireA, csgA*) [41]. Genome-wide association of gene presence between haemolytic and nonhaemolytic *E. coli* was conducted with Scoary software [42]. For all genes with Bonferroni-adjusted *P*-value lower than 0.001 in Scoary analysis, the genes belonging to the following groups: iron acquisition, outer membrane, LPS, secretion systems, multidrug efflux transporter, flagella, toxins, adhesion, biofilm, stress and transcriptional regulators were overrepresented in one bacterial group and counted. Read mapping and SNP calling for *fimH, ecpD* and *csgA* genes were achieved using bwa, samtools and bcftools [43].

Analysis of GenBank genome collection

In order to select genomes containing alpha-haemolysin, a collection of *E. coli* genomes (*n*=22,752) was downloaded from GenBank database and blasted against *hlyA, hlyB, hlyC* and *hlyD* sequences of isolate UTI89 (GenBank accession number: CP000243.1). Genomes that contained all four genes were selected for further analysis. Additionally, genomes with no information about coverage (with the exception of complete genome sequences), coverage less than 50 times, number of contigs higher than 400 and third-generation sequencing platforms, i.e. 'Oxford Nanopore' and 'Pacific Biosciences', were excluded from analysis.

In order to select genomes of nonpathogenic *E. coli*, GenBank collection of *E. coli* was blasted against VAGs and pathotyped according to the presence of VAGs listed in Table S3. Genomes with none of the VAGs or only one of *fimH*, *fyuA*, *iucC*, *neuC*, *sitA* and *yfcV* were considered nonpathogenic *E. coli*. Additionally, part of the nonpathogenic *E. coli* genomes was filtered out like in the case of *hlyCABD*-positive genomes as described above. Genomes were annotated with the use of Prokka version 1.13.5 [32]. BAPS was performed by fastBAPS [44]. *E. coli* phylogroups and *hlyCABD* genome placement were determined as mentioned in the previous paragraph. MLST was examined with the use of mlst and PubMLST database [45, 46]. Serotypes were assigned with the use of ABRicate software and EcOH database [35]. Adhesin, toxin and iron acquisition genes were selected by literature review and sequences were manually collected from GenBank (Table S4). The prevalence of adhesin coding genes or toxin and iron acquisition genes listed in Table S4, was tested with ABRicate. Multiple genes encoding one adhesin, iron acquisition and toxin system were reduced to one VAF as shown in Table S4.

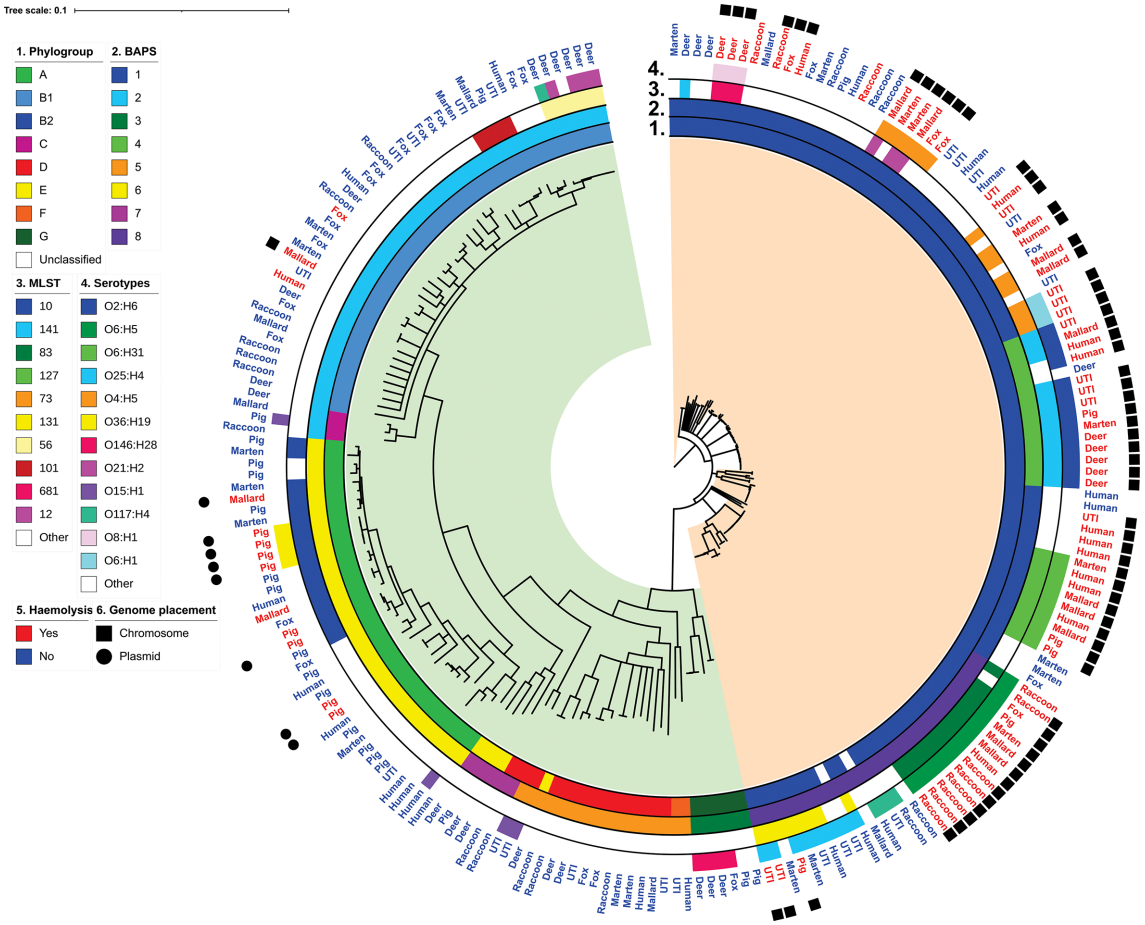

Fig. 1. Phylogenetic relationship between 220 haemolytic and nonhaemolytic *E. coli*. Core-genome phylogenetic tree for 81 haemolytic and 139 nonhaemolytic *E. coli* based on 2628 genes. *E. coli* isolates diverge into two lineages, 1 and 2, marked with orange and green background, respectively. Lineage, phylogroup, BAPS group, ST, serotype, host/disease and haemolysis and alpha-haemolysin genomic placement were annotated on the tree with the use of iTOL.

Nucleotide sequences of *hlyA, hlyB, hlyC* and *hlyD* from haemolytic *E. coli* were extracted from genomes with use of BLAST and *hlyA, hlyB, hlyC* and *hlyD* sequences of isolate UTI89 (GenBank accession number: CP000243.1) were used as reference. Next, nucleotide sequences were translated into amino acid sequences with use of UGENE [47]. Amino acid sequences with 100% identity and 100% coverage were clustered together with use of CD-HIT [48]. The protein sequence was considered a variant if it differed from other proteins with at least one amino acid. Variant numbers were given in order of the viariant prevalence (the lowest number '0' was given to the variant with the highest prevalence). HlyA phylogenetic tree was constructed with the use of alignment of variants that were at least 99% of length of reference UTI89 HlyA protein (which is 1024 amino acids long) and RAxML version 8.2.12 employing HIVW substitution model with 500 bootstrap replications. HlyA protein features were downloaded from InterPro protein families and domains database [49]. Average number of variable sites was calculated by dividing variable sites in feature by all sites in feature.

Figures and statistical analysis

All figures were generated with use of ggplot2 and ggalluvial packages implemented in R software [50–52]. Gene prevalence was compared with Chi-squared test of independence implemented in R stats package [53]. Phylogenetic trees were annotated with iTOL software [54].

RESULTS

Characterization of 220 haemolytic and nonhaemolytic *E. coli* genomes

In total, 81 haemolytic and 139 nonhaemolytic *E. coli* were isolated to generate a collection of bacteria covering various hosts and different clinical status (Table 1). Phylogenetic analysis revealed that all isolates could be categorized into two different lineages (Fig. 1). Lineage 1 comprised of 68 haemolytic (84%) and 38 nonhaemolytic isolates (27%), whereas lineage 2 had only 13 haemolytic (16%) and 101 nonhaemolytic (73%) *E. coli*. Nearly all isolates from

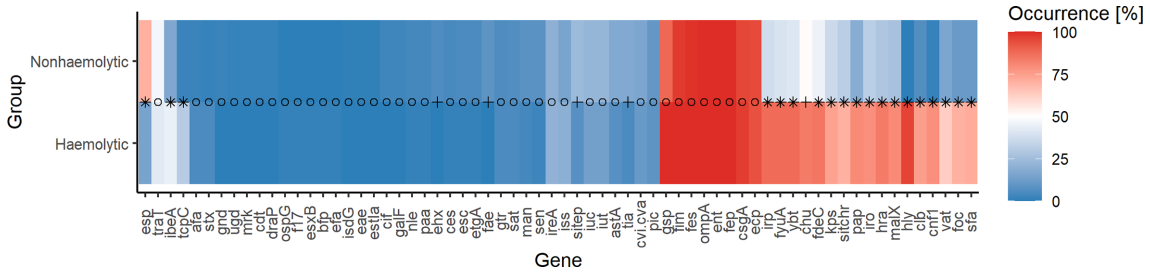

Fig. 2. Virulence factor prevalence in 220 haemolytic and nonhaemolytic *E. coli*. Heatmap with virulence factor sequence prevalence in genomes of haemolytic and nonhaemolytic *E. coli*. Groups 'Haemolytic' and 'Nonhaemolytic' are shown on the *y*-axis and refer to haemolytic and nonhaemolytic *E. coli* isolates, respectively. Names of virulence factors detected in at least one genome are shown on the *x*-axis. The colour gradient is proportional to the prevalence of each virulence factor in each group. Results of statistical analysis with Chi-squared test of independence are shown as symbols: 'o' - not statistically significant, '+' - 0.01<*P*<0.05, '*' - *P*<0.01.

lineage 1 belonged to the phylogroup B2, whereas lineage 2 included isolates from all phylogroups, except for B2. Haemolytic *E. coli* isolates belonged to groups A, B1 and B2, while nonhaemolytic *E. coli* had representatives in each phylogroup (Table S5). Investigation of the genetic structure with BAPS revealed eight clusters, from which three clusters belonged to lineage 1 and phylogroup B2. In the case of isolates from lineage 2, BAPS clusters corresponded well with phylogroups.

Sequence types (STs) and serotypes were assigned to each isolate. Haemolytic *E. coli* were represented by nearly four times less STs and serotypes when compared with nonhaemolytic isolates (Figs 1 and S1). Both groups shared only eight STs and ten serotypes. Haemolytic *E. coli* had 17 unique STs and serotypes, while nonhaemolytic had 85 and 94 unique STs and serotypes, respectively.

Operon *hlyCABD* encoding alpha-haemolysin was detected in 78 out of 81 haemolytic *E. coli*. All haemolytic isolates did not possess the genes encoding enterohaemolysin and silent haemolysin. All nonhaemolytic isolates were devoid of genes encoding alpha-haemolysin, but 92 isolates (66.2%) were positive for *hlyE* encoding silent haemolysin and eight were positive for *ehxCABD* encoding enterohaemolysin operon. Analysis of *hlyCABD* operon placement revealed that for all isolates from lineage 1, these genes were carried on the chromosome, whereas all isolates from lineage 2, with the exception of one isolate, had these genes located on the plasmid. All isolates with chromosome-encoded *hlyCABD* belonged to phylogroup B1 and B2, while plasmid-encoded alpha-haemolysin was found in isolates from phylogroup A, isolated nearly exclusively from pigs.

Virulence-associated factors profiling revealed high prevalence of adhesins, iron acquisition genes and toxins in haemolytic *E. coli*

In order to investigate the association of virulence-associated factors with haemolytic and nonhaemolytic *E. coli*, the prevalence of VAGs was examined. In total, 255 VAGs coding for 66 different VAFs were detected in at least one isolate, but prevalence differed for only 24 VAFs between haemolytic and nonhaemolytic *E. coli* (Figs 2 and S2). The majority (i.e. 19 of 24) of VAFs were detected significantly more often in haemolytic *E. coli* (*P*<0.05), and their functions were adhesion, invasion, iron acquisition or toxicity. Four VAFs with similar functions as mentioned earlier were detected more frequently in nonhaemolytic *E. coli*. Surprisingly, one VAF also associated with this group contained VAGs encoding for the type three secretion system. Eight VAFs were present nearly in all *E. coli* isolates, out of which four were associated with iron acquisition and another three with adhesion.

Genome-wide comparison of 220 haemolytic and nonhaemolytic *E. coli* genomes showed differences in the prevalence of adhesin, outer membrane and iron acquisition genes

Analysis of VAGs prevalence was initially limited to genes present in the VFDB_core database. To find other genetic traits associated with haemolytic and nonhaemolytic *E. coli* groups, the pangenome was inferred, and differences in gene content between haemolytic and nonhaemolytic *E. coli* were investigated (for 'gene' please take into consideration settings of Roary). GWAS analysis revealed 1080 genes overrepresented in one of the *E. coli* groups (*P*<0.001) and the majority associated with metabolism. Overall, 184 genes were identified as VAGs and assigned to one of the ten functional groups (Table 2). Adhesion as the VAG group had the highest frequency of genes associated with one of the *E. coli* groups and there were 13 additional adhesion VAGs identified in haemolytic *E. coli* (*P*<0.1). Genes encoding for P, Yfc, Yde (F9), Ecp, Yeh, Yad fimbriae and autotransporters YeeJ, YfaL were more often found in haemolytic *E. coli* (*P*<0.001) (Table S5). Another large functional group covers genes associated with outer membrane, where genes responsible for LPS and flagella synthesis, secretion systems and multidrug efflux transporters are categorized. The presence of secretion systems and LPS genes were associated with haemolytic phenotype (*P*<0.001 and *P*<0.05, respectively). The majority of genes assigned to secretion systems group were coding for type

Table 2. Comparison of virulence-associated genes presence between 81 haemolytic and 139 nonhaemolytic *E. coli*

VAG group	Haemolytic	Nonhaemolytic	All	*P*-value
Iron acquisition	25	2	27	0.000001
Outer membrane:	36	14	50	0.002
• LPS	4	0	4	0.05
• Secretion systems	14	1	15	0.001
• Multidrug efflux transporter	7	4	11	0.37
• Flagella	2	4	6	0.42
Toxins	16	8	24	0.1
Adhesion	39	26	65	0.11
Biofilm	3	6	9	0.32
Stress	7	4	11	0.37
Transcriptional regulators	8	8	16	1
Total	161	77	238	

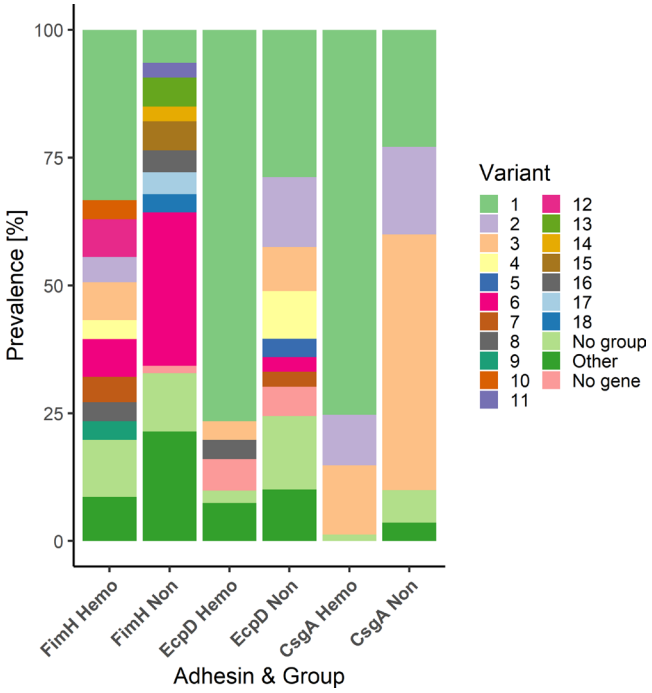

Fig. 3. Comparison of adhesin frequency between 81 haemolytic and 139 nonhaemolytic *E. coli*. Barplot showing FimH, EcpD and CsgA adhesin frequency in genomes of haemolytic and nonhaemolytic *E. coli*. The frequency of different adhesin variants is shown on the *y*-axis. Names of adhesins and group ('Haemo' refers to haemolytic *E. coli*, 'Non' refers to nonhaemolytic *E. coli*) are shown on the *x*-axis. Various colours represent different adhesin variants and are described on the legend. Each number represents one adhesin variant, 'No Group' contains frequency of adhesin variants present in only one isolate in both groups, 'Other' contains frequency of adhesin variants present in less than 2.5% of isolates. 'No gene' refers to genomes without investigated gene.

II secretion system. The highest difference in gene presence between haemolytic and nonhaemolytic *E. coli* was found in the group containing iron acquisition genes ($P<0.000001$). *HlyCABD* genes were part of this VAG group and the first four genes associated with haemolytic phenotype with the lowest *P*-values ($P<10^{-48}$).

Sequence variation shows differences between haemolytic and nonhaemolytic *E. coli* in adhesins with a very high prevalence

When we analysed the prevalence of different VAFs, we observed that three adhesins, i.e. type 1 fimbriae, *Escherichia* common pili and curli fimbriae were present in nearly all haemolytic and nonhaemolytic *E. coli*. On the other hand, some of the genes encoding for these adhesins were found differentially represented in GWAS analysis. Taking these observations together with the overwhelming amount of research showing the contribution of *fimH* sequence polymorphism to adhesive properties of type 1 fimbriae in *E. coli* lead to the conclusion that the analysis of these adhesins should not only include the presence of genes coding for adhesin clusters but also a comparison of variant prevalence. It was found that FimH, EcpD and CsgA variants differed in prevalence between haemolytic and nonhaemolytic *E. coli* (Fig. 3). In the case of FimH, four variants (no. 1, 3, 6 and 13) were found statistically more often in one of the groups ($P<0.05$), and 15 variants (out of 17) were present only in one of these groups. Nonhaemolytic *E. coli*, had 2.5 times more FimH variants present in less than 2.5% isolates. When sequence variation in EcpD adhesin was tested, six variants were present in only one group (i.e. no. 2, 4, 5, 6, 7 and 8), and there were 5.8 times more variants present in only one isolate of nonhaemolytic *E. coli* in comparison to haemolytic *E. coli* ($P<0.05$). One EcpD variant was found statistically more often in haemolytic *E. coli* ($P<0.001$, no. 1) and two in nonhaemolytic *E. coli* ($P<0.01$, no. 2, 4). The lowest number of variants were found during analysis of CsgA. One dominant CsgA variant (i.e. no. 1) with 75% prevalence was found in haemolytic *E. coli* and detected 3.3 times more often compared to nonhaemolytic *E. coli* ($P<0.001$). The most prevalent CsgA variant (i.e. no 3) in nonhaemolytic *E. coli* (50%) was found 3.7 times less often in haemolytic *E. coli* ($P<0.001$, no. 1). To summarize, there were significant differences between haemolytic and nonhaemolytic *E. coli* in the prevalence of diverse adhesin variants.

Global diversity of haemolytic *E. coli*

To get a global overview of haemolytic *E. coli* population structure, 1041 genomes were downloaded from the GenBank database. The downloaded genomes along with 81 genomes sequenced in this study were analysed. Phylogenetic analysis revealed that the isolates clustered according to their respective phylogroup placement (Fig. 4). Majority of isolates belonged to phylogroup B2 (65%), B1 (17%) and A (9.5%). Altogether, 2 and 5% of isolates belonged to the phylogroup

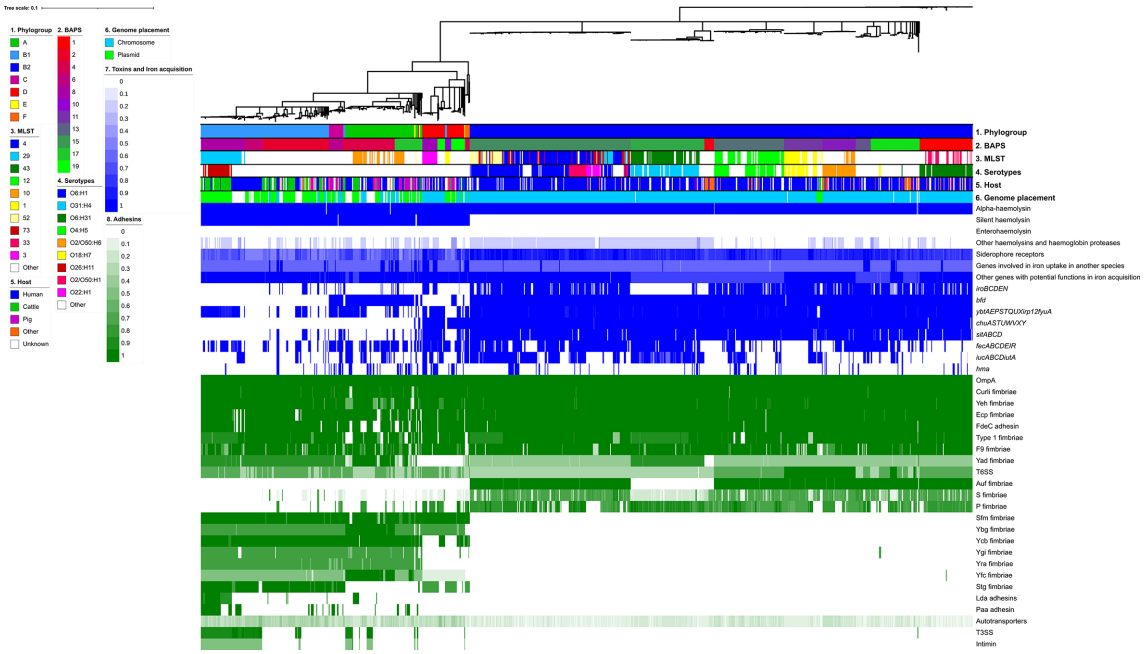

Fig. 4. Phylogenetic relationship and VAGs' distribution in global population of haemolytic *E. coli*. Core-genome phylogenetic tree for 1122 haemolytic *E. coli* based on 3032 genes. Phylogroup, BAPS group, ST, serotype and alpha-haemolysin genomic placement, iron acquisition, toxin and adhesin gene prevalence were annotated on the tree with the use of iTOL.

C and D, respectively, and only a few isolates were clustered to phylogroups E and F. BAPS revealed 19 clusters that aligned well within phylogroups, further dividing them into subgroups (Figs 4 and S3a). To test the genetic and antigenic diversity in a collection of haemolytic *E. coli* genomes, all isolates were analysed for STs and serotypes. All 1122 isolates represented 149 STs and 189 serotypes, indicating high diversity in this set of genomes (Fig. S3C, E). Such high diversity is supported by the fact that all isolates originate from 46 countries across six continents (Fig. S3B). When genome placement for operon *hlyCABD* was tested, the majority of the isolates from phylogroup B2 and 8 (out of 9) BAPS subgroups had alpha-haemolysin encoded on a chromosome (97.8%) (Figs S3D and S4). Similarly, 69.5% of isolates from phylogroup D encoded the operon on the chromosome. Isolates belonging to phylogroups A and C had comparable distribution of isolates between chromosome- and plasmid-encoded *hlyCABD*. For group B1, most isolates had plasmid-encoded *hlyCABD* (69.8%). Overall, only 4% of isolates with plasmid-encoded alpha-haemolysin were detected in phylogroup B2. The majority of isolates belonging to phylogroups B2 and D were isolated from humans (Fig. S3f). Human isolates were also present in other phylogroups, but their prevalence was lowered by animal isolates, mainly from pigs and cattle. When genome placement was compared with the host origin of the isolate, it was observed that animal isolates possessed plasmid-encoded alpha-haemolysin, whereas human isolates encoded this operon on the chromosome (P<0.001)(Fig. S3G).

As the presence of adhesins, iron acquisition and toxin genes have a direct linkage to the haemolytic and virulence abilities of *E. coli* (Fig. 2, Table 2), they were investigated to elucidate global diversity in haemolytic and nonhaemolytic *E. coli*. As there is no comprehensive database containing *E. coli* adhesins, iron acquisition and toxin genes, they were downloaded from GenBank nucleotide collection. A database was set up with a total of 124 toxin and iron acquisition and 443 adhesin genes, providing functionality for 24 iron acquisition and toxin systems and 75 adhesins or adhesion-related molecules.

In total, 124 toxin and iron acquisition and 275 adhesin genes were detected in at least one genome (Fig. S5A, B). Alpha-haemolysin was detected in all except three isolates. Entero-haemolysin was not present in any of the tested genomes, whereas silent haemolysin was present in isolates from all phylogroups except B2. Opposite observation was made for group 'Other haemolysins and haemoglobin proteases', which were more prevalent in phylogroup B2 compared to A and B1 groups (P<0.001).

Iron acquisition genes encoding for salmochelin (*iroBCDEN*) and SitABCD system (*sitABCD*) were found more prevalent in isolates from phylogroup B2 than in other groups (P<0.001, Fig. 4). Yersiniabactin (*ybtAEPSTQUXirp12fyuA*) was detected around two times more often in isolates from group B2 than groups A, B1 and D (P<0.001). Gene *bfd* involved in iron storage was not detected in group B1 but had nearly 100% prevalence in group B2. Fec system (*fecABCDEIR*) and aerobactin (*iucABCDiutA*) were more prevalent in B2 phylogroup than in group B1 (P<0.001), but less prevalent when compared to groups C and D (P<0.01).

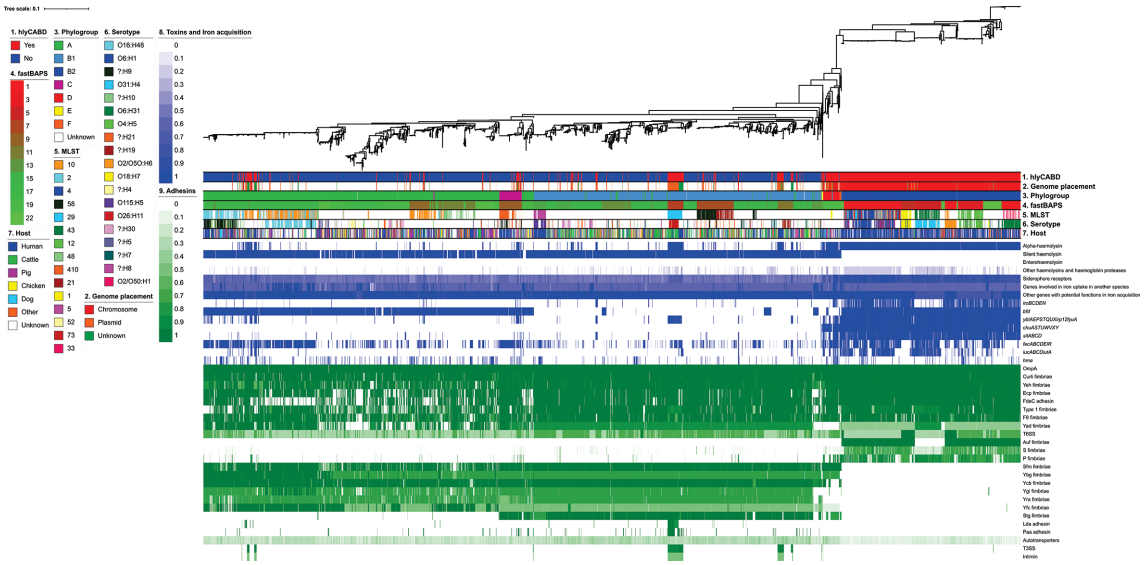

Fig. 5. Phylogenetic relationship and VAGs' distribution in haemolytic and nonpathogenic *E. coli*. Core-genome phylogenetic tree for 1122 haemolytic and 2257 nonpathogenic *E. coli* based on 2818 genes. Phylogroup, *hlyCABD* genome placement, BAPS group, ST, serotype, host, toxin and iron acquisition and adhesin gene prevalence were annotated on the tree with the use of iTOL.

Adhesin and adhesion-associated genes' prevalence analysis revealed a group of nine adhesins present in nearly all investigated genomes (Fig. 4). Auf, S and P fimbriae were found nearly exclusively in B2 phylogroup (P<0.001), whereas Sfm, Ybg and Yfc were present in all groups with the exception of B2. A similar observation could be made for Ycb, Ygi and Yra fimbriae, but the prevalence of these adhesins was significantly lower in group D when compared with phylogroups A, B1 and C (P<0.001). Stg fimbriae were detected only in phylogroups B1, C and D. In some isolates of B1 group belonging to BAPS subgroup 8, Lda and Paa adhesins were detected. In the same phylogroup, BAPS subgroup 6 and 8, type 3 secretion system and intimin genes were found.

When frequency of iron acquisition and toxin systems was compared with genome placement, only silent haemolysin was found eight times more frequently in isolates with plasmid-encoded *hlyCABD* operon (P<0.001). Eight systems associated with iron acquisition, i.e. salmochelin, SitABCD, hemin uptake (*chuASTUWVXY*), Fec, aerobactin, other haemolysins and haemoglobin proteases, iron storage-*bfd* and yersiniabactin were found 1.5-9.7 times more often in genomes of isolates with alpha-haemolysin encoded on the chromosome (P<0.001).

Adhesin genes prevalence was also associated with genome placement of alpha-haemolysin. Auf, S and P fimbriae were found 11–184 times more often when *hlyCABD* cluster was encoded on a chromosome (P<0.001), whereas Sfm, Stg, Ybg, Yfc, Ycb, Ygi, Yra and Paa adhesins were identified in genomes with plasmid-encoded alpha-haemolysin 8–71 times more often than in other isolates (P<0.001). Intimin, T3SS and Lda adhesin were found exclusively in isolates with plasmid-encoded alpha-haemolysin.

Pathotyping revealed that majority of isolates with chromosome-encoded *hlyCABD* belonged to UPEC (87%), whereas plasmid-encoded *hlyCABD* can be found mainly in nonpathogenic isolates and undetermined pathotype (50.5%) followed by atypical enteropathogenic *E. coli* (aEPEC, 29.6%), enterotoxigenic *E. coli* (ETEC, 14.4%), UPEC (2.3%), neonatal meningitis *E. coli* (NMEC, 1.85%), STEC (0.9%) and enteroinvasive *E. coli* (EIEC, 0.5%) (Fig. S6).

Taken together, the data suggest that haemolytic *E. coli* are pathotypically, genetically and antigenically diverse, but their adhesin, toxin and iron acquisition system repertoire is associated with genome placement of *hlyCABD* cluster.

E. coli with plasmid-encoded alpha-haemolysin has similar iron acquisition system and adhesin profile to nonpathogenic *E. coli*

Next, haemolytic *E. coli* genomes were compared with a set of *E. coli* genomes that did not possess VAFs typical for well-defined *E. coli* pathotypes (and therefore were considered as 'nonpathogenic *E. coli*') to assess virulence factors unique for haemolytic *E. coli* and which can be also found in nonpathogenic *E. coli*. Overall, 2257 genomes of nonpathogenic *E. coli* belonged to phylogroups A, B1, C and D, and 14 out of 22 BAPS groups (Fig. 5). Nonpathogenic *E. coli* were more diverse group in comparison to their haemolytic counterparts. Altogether, 509 STs and 671 serotypes were defined in 2257 nonpathogenic *E. coli* collected from 65 different hosts in 64 countries.

Analysis of iron acquisition and toxin systems' frequency revealed that salmochelin, SitABCD, hemin uptake, aerobactin, other haemolysins and haemoglobin proteases,

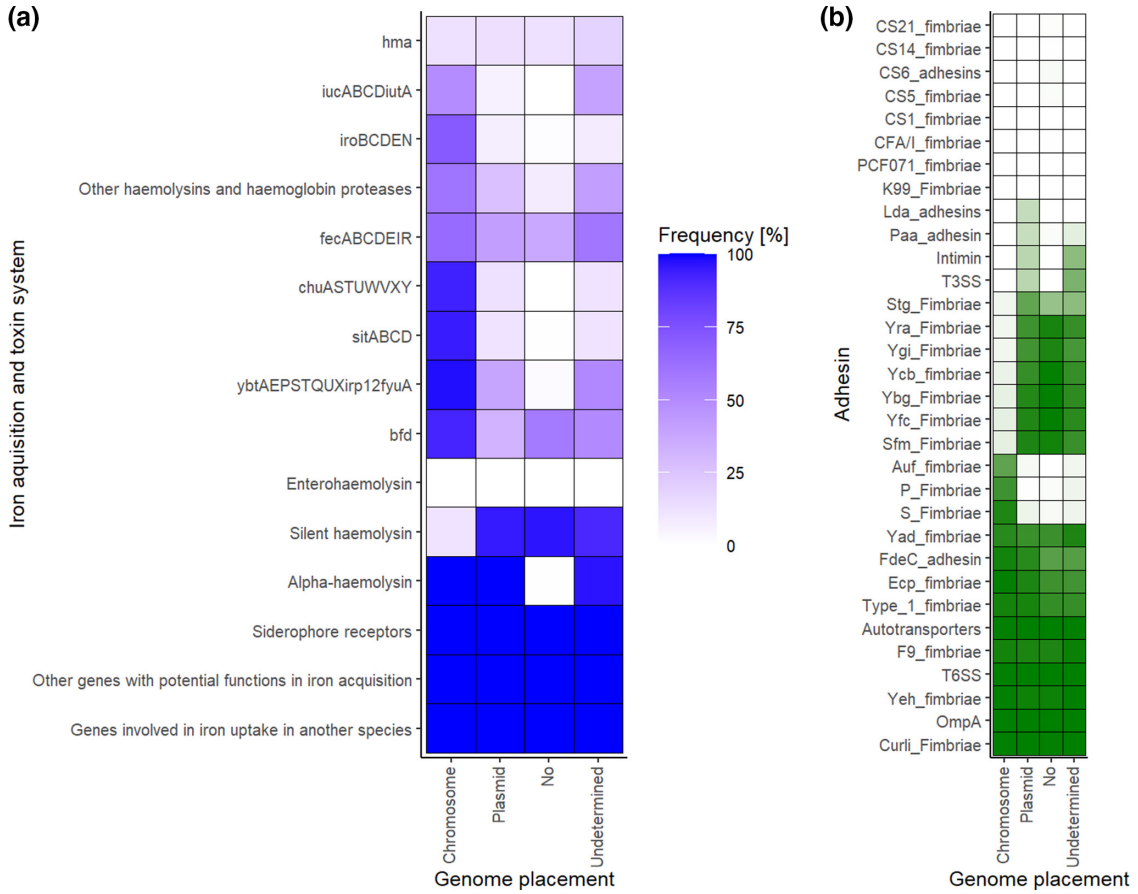

Fig. 6. Frequency of toxin, iron acquisition and adhesin systems in haemolytic and nonpathogenic *E. coli*. Frequency of 124 toxin and iron acquisition and 443 adhesin genes providing information about the presence of 24 toxins or iron acquisition systems and 75 adhesins or adhesion-related molecules was tested in 1122 haemolytic and 2257 nonpathogenic *E .coli*. Genome placement is shown on the *x*-axis. Iron acquisition and toxin systems (a) and adhesins (b) are listed on the *y*-axis. The colour gradient is proportional to the frequency of each system, and the colour scale is shown on the legends attached with each heatmap.

and yersiniabactin were found 3–41 times less prevalent in genomes of nonpathogenic isolates than in isolates with plasmid-encoded *hlyCABD* (P<0.001) (Fig. 6a). When compared with isolates with genome-encoded *hlyCABD*, the aforementioned systems were found 7–296 times less frequently in nonpathogenic isolates (P<0.001).

Adhesin profile of nonpathogenic *E. coli* was very similar to the haemolytic *E. coli* with plasmid-encoded alpha-haemolysin (Fig. 6b). These two groups only differed in frequency of intimin, T3SS, Lda and Paa adhesins (P<0.001). Whereas in haemolytic *E. coli* with chromosome-encoded alpha-haemolysin, Auf, S and P fimbriae were found 25–73 times more often in comparison to nonpathogenic *E. coli* (P<0.001). However, adhesins like Paa, Stg, Sfm, Yfc, Ybg, Ycb, Ygi, Yra were identified in nonpathogenic *E. coli* 6–15 times more often than in haemolytic *E. coli* with chromosome-encoded *hlyCABD* cluster (P<0.001).

Overall, our data suggests that isolates with plasmid-encoded alpha-haemolysin are similar to *E. coli* genomes that do not possess VAFs typical for well-defined *E. coli* pathotypes.

Variant prevalence of HlyCABD is associated with genome placement and pathotype of *E. coli*

To assess the diversity within *hlyCABD* cluster, the aforementioned sequences were collected and submitted to amino acid sequence variation analysis. Highest number of variants were detected in HlyA (122), followed by HlyB (74) and HlyD (56) and the lowest number of variants were found for HlyC (37) (Figs 7a and S7A). Interestingly, different protein variants were associated with genome placement of alpha haemolysin (Figs 7b and S7B). Similarly, in the case of pathotypes, selected variants were specific to one pathotype only (Figs 7c and S7C).

Phylogenetic analysis with the use of HlyA alignment revealed clustering of HlyA variants encoded on chromosome (mainly from UPEC strains) together (Fig. 8). In the case of plasmid-encoded HlyA variants, it is visible that variants from aEPEC and ETEC cluster together, which is associated with the presence of variable sites, specific for plasmid-encoded HlyA. Distribution of variable sites was not equal in HlyA (Fig. S8). Taking into consideration InterPro database features, the

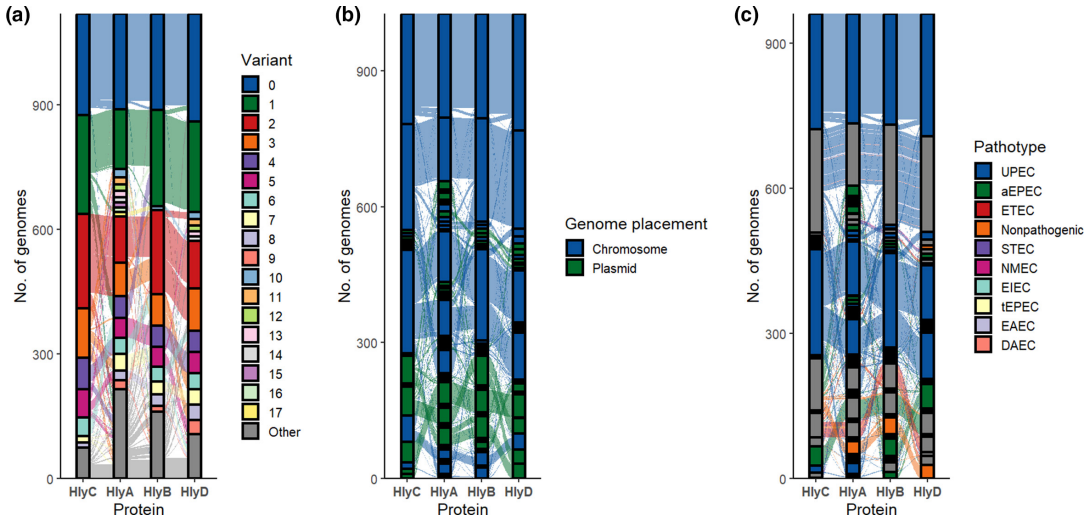

Fig. 7. HlyCABD variant frequency in *E. coli*. Alluvial plots with HlyCABD variant prevalence contextualized with information about alpha-haemolysin genome placement and pathotype of alpha-haemolysin-positive *E. coli*. Names of proteins are shown on the *x*-axis and number of genomes is shown on the *y*-axis. Colour of the bars and streams connecting them represent variants (a), genome placement (b) or pathotypes (c) and are described on the legend on the right side for each plot separately. Variants with more than one pathotype (c) are coloured grey. All variants detected in less than ten isolates were joint together in one group 'Other' on plot (a) (see Fig. S7A for alluvial plot on which all variants are shown). All genomes with undetermined genome placement are not shown on plot (b) (see Fig. S7B for alluvial plot on which all genomes are shown). All genomes with Not determined ('NA') pathotype are not shown on plot (c) (see Fig. S7C for alluvial plot on which all genomes are shown).

average number of variable sites in RTX C-terminal, RTX N-terminal, RTX toxin determinant A, RTX calcium-binding nonapeptide repeat was 0.32, 0.17, 0.1, 0.09, respectively, when in the rest of the sites the average number of variable sites was 0.18.

DISCUSSION

E. coli can colonize various niches in the host body and the environment [2]. Its success depends on the presence of various genetic factors that provide the bacterium with the

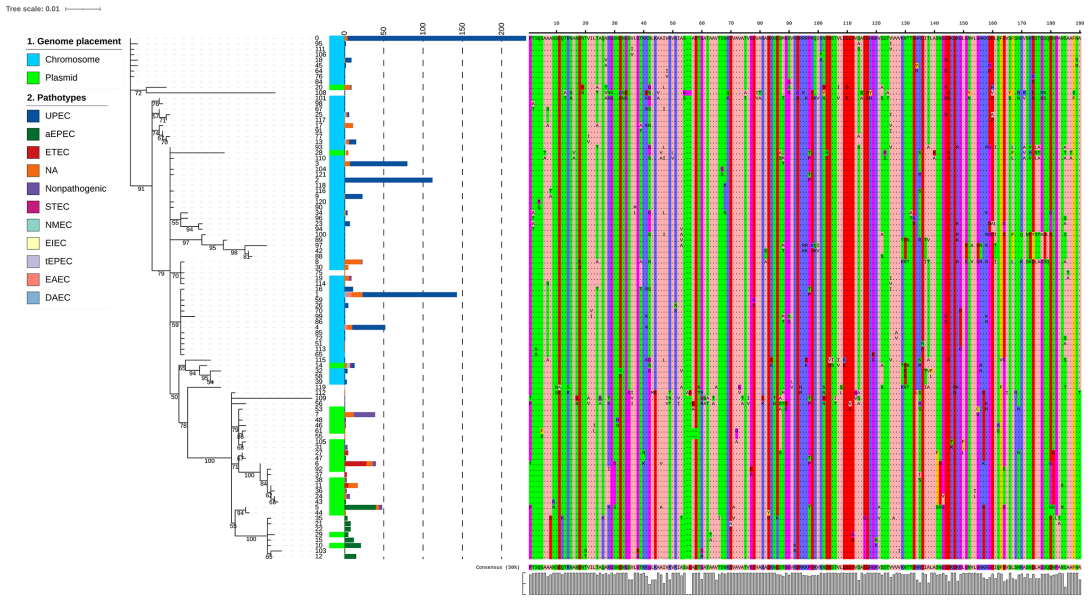

Fig. 8. Phylogenetic analysis of HlyA in *E. coli.* Phylogenetic tree for 96 HlyA variants from alpha-haemolysin-positive *E. coli* genomes. Genome placement, pathotype prevalence and amino acid variable sites (190) were annotated on the tree with the use of iTOL. Numbers on the alignment do not reflect actual site numbers in the whole protein alignment.

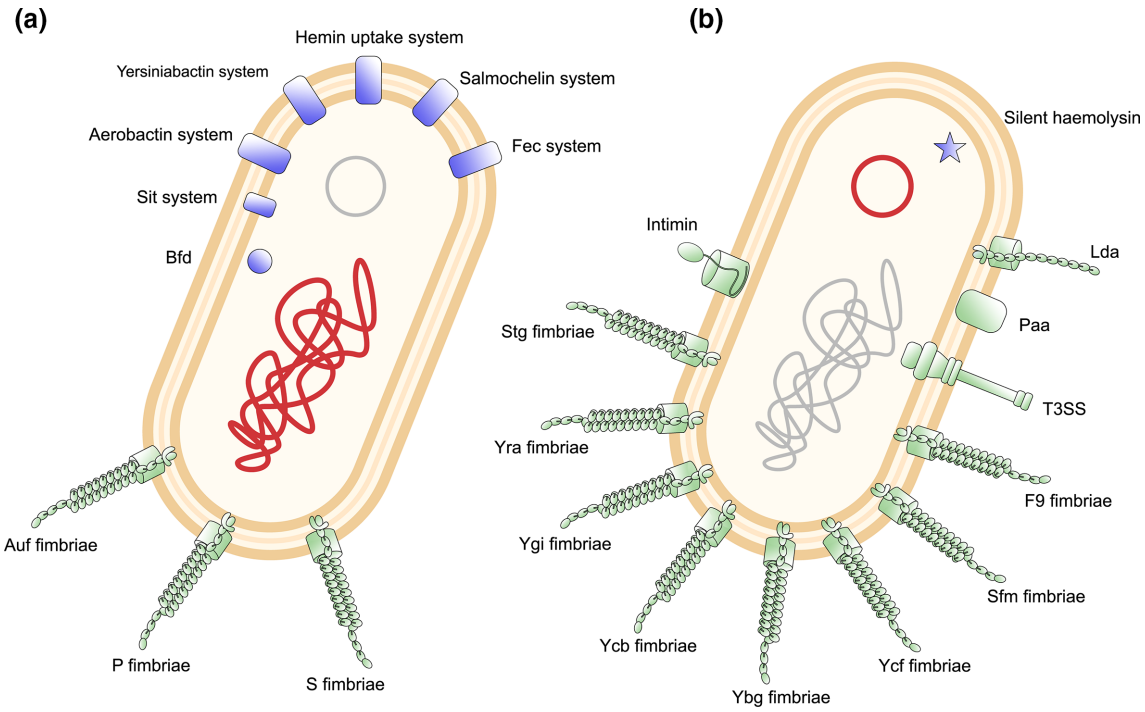

Fig. 9. Virulence-associated factors of haemolytic *E. coli*. Schematic depiction of differences in VAFs of *E. coli* with chromosome-encoded (a) and plasmid-encoded (b) alpha-haemolysin. Iron acquisition and toxin systems are marked in blue, while adhesins and adhesion-related molecules in green. VAFs present in haemolytic *E. coli* regardless of *hlyCABD* operon placement (i.e. alpha-haemolysin, siderophore receptors, Yad, type I, F9, Yeh and curli fimbriae, FdeC adhesin, ECP, autotransporters, T6SS, and OmpA) were omitted for graphic simplicity.

ability to survive and propagate in often harsh conditions. One of the VAFs with multiple proposed functions is alpha-haemolysin [24, 25]. In this work, we analysed the VAFs that can be associated with the presence of alpha-haemolysin in genomes to determine other factors that might influence utilization of alpha-haemolysin by *E. coli* during host colonization. First, we established a large collection of haemolytic and nonhaemolytic *E. coli* from various hosts and compared their phylogenetic relatedness. We noticed that several isolates collected from wild animals cluster together with human UTI-causing isolates. Similar results have been obtained in the previous studies, where wild animals have been identified as carriers of potential human pathogenic *E. coli* and our observations provide additional information about wild animals as the reservoir of haemolytic *E. coli* [55, 56]. The full assessment of the potential of these isolates to cause human infections can be only evaluated by functional analysis of virulence properties with the use of *in vitro* and *in vivo* animal experiments.

Analysis of VAFs' presence in haemolytic and nonhaemolytic *E. coli* revealed a group of VAFs with high prevalence in both groups, with three adhesins among them (Fig. 2). Our previous work and many other reports have shown that variations in FimH adhesin sequence have an impact on interaction with host cell receptors or biofilm formation [57–59]. ECP and Curli fimbriae also mediate binding of *E. coli* to host cells and take part in biofilm formation, but the effect of sequence variation in EcpD adhesin and CsgA to these processes have not been shown so far [60] [61, 62]. We were able to show that different FimH, EcpD and CsgA variant clusters are present in the analysed groups, which indicate different adhesive capabilities of haemolytic and nonhaemolytic *E. coli*. We hypothesize that different niche colonization of haemolytic and nonhaemolytic *E. coli* lead to point mutations that result in fimbrial functional diversification and altered binding capabilities to biotic or abiotic surfaces [63, 64]. It is also possible that observed mutations are based on different cluster membership in population or phylogenetic history, therefore our hypothesis requires experimental confirmation [65].

To further explore differences between haemolytic and nonhaemolytic *E. coli* we carried out GWAS and unveiled genes encoding toxins, iron acquisition and adhesion systems that have not been previously associated with the haemolytic *E. coli*. One of these genes encodes a protein belonging to the porin protein family functionally classified as an outer membrane protein with beta-barrel domain [66]. Porins are known to play a role in the regulation of outer membrane permeability, stress response and adhesion to epithelial cells [67, 68]; therefore we think that this gene requires further characterization to elucidate its role in haemolytic *E. coli* physiology. Furthermore, genes encoding

for type II secretion system were found nearly exclusively in haemolytic *E. coli*. As this system is utilized in the export of VAFs like proteases and adhesins aiding host colonization, our analysis shows that it can be important for haemolytic *E. coli* and may contribute in many virulence traits to gaining access and adhesion to host cells [69].

During the analysis of 81 haemolytic *E. coli* genomes, we noticed that majority of isolates have chromosome-encoded alpha-haemolysin, whereas only a few isolates possess plasmid-encoded alpha-haemolysin. Enriching our analysis with over 1000 genomes of haemolytic *E. coli* from GenBank allowed us to show that isolates with chromosome- and plasmid-encoded haemolysin differ in VAF prevalence (Fig. 9). Bacteria with plasmid-encoded haemolysin have fewer iron acquisition systems and possess different set of adhesion factors than bacteria with chromosome-encoded haemolysin. Moreover, different protein variants of HlyCABD were associated with genome placement of *hlyCABD*. Taking into consideration that: (1) all investigations concerning the role of alpha-haemolysin were conducted with the use of *E. coli* isolates with chromosome-encoded haemolysin [24, 70], (2) isolates with plasmid-encoded alpha-haemolysin have similar VAF profile to nonpathogenic *E. coli*, (3) the majority of isolates with plasmid-encoded alpha-haemolysin was isolated from farm animals and (4) influence of HlyCABD sequence variation on haemolytic properties of *E. coli* was investigated only in UPEC [71], we think that *E. coli* plasmid-encoded alpha-haemolysin requires further attention with a focus on phenotypic characterization of cell exfoliation potential, immunomodulatory properties and the possibility of farm animals as a source of alpha-haemolysin in human pathogenic *E. coli*.

Funding information
The work of AA was supported by the Polish National Science Centre Research Grant PRELUDIUM BIS [UMO-2019/35/O/NZ6/01590]. The work of RK and AA was supported by the Wrocław University of Environmental and Life Sciences (Poland) under the research program "Innowacyjny Naukowiec" [No. N060/0018/20].

Author contributions
Conceptualization: R.K. and P.S. Methodology: R.K., K.S., M.N., A.A., M.B. and D.P. Validation: R.K., K.S., M.N., A.A., M.M.K., M.B., D.P. and P.S. Formal analysis: R.K., K.S. and M.B. Investigation: R.K., K.S., M.N., A.A., M.M.K., M.B., D.P. and P.S. Resources: R.K., D.P. and P.S. Data curation: R.K., K.S. and M.B. Writing – original draft: R.K., K.S., M.N., A.A., M.M.K., M.B., D.P. and P.S. Writing - review and editing: R.K., K.S., M.N., A.A., M.M.K., M.B., D.P. and P.S. Visualization: R.K. and K.S. Supervision: R.K. and P.S. Project administration: R.K. and P.S. Funding acquisition: R.K., D.P. and P.S.

Conflicts of interest
The authors declare that there are no conflict of interest.

References

1. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2004;2:123–140.
2. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, *et al*. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin Microbiol Rev* 2013;26:822–880.
3. Aleksandrowicz A, Khan MM, Sidorczuk K, Noszka M, Kolenda R. Whatever makes them stick - Adhesins of avian pathogenic *Escherichia coli*. *Vet Microbiol* 2021;257:109095.
4. Croxen MA, Finlay BB. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol* 2010;8:26–38.
5. Kolenda R, Burdukiewicz M, Schierack P. A systematic review and meta-analysis of the epidemiology of pathogenic *Escherichia coli* of calves and the role of calves as reservoirs for human pathogenic E. coli. *Front Cell Infect Microbiol* 2015;5:23.
6. Le Bouguénec C. Adhesins and invasins of pathogenic *Escherichia coli*. *Int J Med Microbiol* 2005;295:471–478.
7. Larsonneur F, Martin FA, Mallet A, Martinez-Gil M, Semetey V, *et al*. Functional analysis of *Escherichia coli* Yad fimbriae reveals their potential role in environmental persistence. *Environ Microbiol* 2016;18:5228–5248.
8. Rödiger S, Kramer T, Frömmel U, Weinreich J, Roggenbuck D, *et al*. Intestinal *Escherichia coli* colonization in a mallard duck population over four consecutive winter seasons. *Environ Microbiol* 2015;17:3352–3361.
9. Schiebel J, Böhm A, Nitschke J, Burdukiewicz M, Weinreich J, *et al*. Genotypic and phenotypic characteristics associated with biofilm formation by human clinical *Escherichia coli* isolates of different pathotypes. *Appl Environ Microbiol* 2017;83:e01660-17.
10. Yao Y, Xie Y, Perace D, Zhong Y, Lu J, *et al*. The type III secretion system is involved in the invasion and intracellular survival of *Escherichia coli* K1 in human brain microvascular endothelial cells. *FEMS Microbiol Lett* 2009;300:18–24.
11. Zhou Y, Tao J, Yu H, Ni J, Zeng L, *et al*. Hcp family proteins secreted via the type VI secretion system coordinately regulate *Escherichia coli* K1 interaction with human brain microvascular endothelial cells. *Infect Immun* 2012;80:1243–1251.
12. Garénaux A, Caza M, Dozois CM. The Ins and outs of siderophore mediated iron uptake by extra-intestinal pathogenic *Escherichia coli*. *Vet Microbiol* 2011;153:89–98.
13. Andrews SC, Robinson AK, Rodríguez-Quiñones F. Bacterial iron homeostasis. *FEMS Microbiol Rev* 2003;27:215–237.
14. Welch RA. Uropathogenic *Escherichia coli*-Associated exotoxins. *Microbiol Spectr* 2016;4.
15. Lorenz SC, Son I, Maounounen-Laasri A, Lin A, Fischer M, *et al*. Prevalence of hemolysin genes and comparison of ehxA subtype patterns in Shiga toxin-producing *Escherichia coli* (STEC) and non-STEC strains from clinical, food, and animal sources. *Appl Environ Microbiol* 2013;79:6301–6311.
16. Oscarsson J, Westermark M, Beutin L, Uhlin BE. The bacteriophage-associated ehly1 and ehly2 determinants from *Escherichia coli* O26:H- strains do not encode enterohemolysins per se but cause release of the ClyA cytolysin. *Int J Med Microbiol* 2002;291:625–631.
17. Morales C, Lee MD, Hofacre C, Maurer JJ. Detection of a novel virulence gene and a *Salmonella* virulence homologue among *Escherichia coli* isolated from broiler chickens. *Foodborne Pathog Dis* 2004;1:160–165.
18. del Castillo FJ, Leal SC, Moreno F, del Castillo I. The *Escherichia coli* K-12 sheA gene encodes a 34-kDa secreted haemolysin. *Mol Microbiol* 1997;25:107–115.
19. Schmidt H, Beutin L, Karch H. Molecular analysis of the plasmid-encoded hemolysin of *Escherichia coli* O157:H7 strain EDL 933. *Infect Immun* 1995;63:1055–1061.
20. Schwidder M, Heinisch L, Schmidt H. Genetics, toxicity, and distribution of enterohemorrhagic *Escherichia coli* hemolysin. *Toxins (Basel)* 2019;11:E502.
21. Bielaszewska M, Aldick T, Bauwens A, Karch H. Hemolysin of enterohemorrhagic *Escherichia coli*: structure, transport, biological activity and putative role in virulence. *Int J Med Microbiol* 2014;304:521–529.
22. Hunt S, Green J, Artymiuk PJ. Hemolysin E (HlyE, ClyA, SheA) and related toxins. *Adv Exp Med Biol* 2010;677:116–126.
23. Burgos Y, Beutin L. Common origin of plasmid encoded alpha-hemolysin genes in *Escherichia coli*. *BMC Microbiol* 2010;10:193.
24. Ristow LC, Welch RA. Hemolysin of uropathogenic *Escherichia coli*: A cloak or a dagger? *Biochim Biophys Acta* 2016;1858:538–545.

25. Schierack P, Weinreich J, Ewers C, Tachu B, Nicholson B, *et al*. Hemolytic porcine intestinal Escherichia coli without virulence-associated genes typical of intestinal pathogenic *E. coli*. *Appl Environ Microbiol* 2011;77:8451–8455.
26. Merlino J, Siarakas S, Robertson GJ, Funnell GR, Gottlieb T, *et al*. Evaluation of CHROMagar Orientation for differentiation and presumptive identification of gram-negative bacilli and *Enterococcus* species. *J Clin Microbiol* 1996;34:1788–1793.
27. Ewers C, Guenther S, Wieler LH, Schierack P. Mallard ducks - a waterfowl species with high risk of distributing *Escherichia coli* pathogenic for humans. *Environ Microbiol Rep* 2009;1:510–517.
28. Guenther S, Filter M, Tedin K, Szabo I, Wieler LH, *et al*. Enterobacteriaceae populations during experimental Salmonella infection in pigs. *Vet Microbiol* 2010;142:352–360.
29. Schierack P, Römer A, Jores J, Kaspar H, Guenther S, *et al*. Isolation and characterization of intestinal Escherichia coli clones from wild boars in Germany. *Appl Environ Microbiol* 2009;75:695–702.
30. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* 2018;7:1338.
31. Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, *et al*. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb Genom* 2016;2:e000083.
32. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl* 2014;30:2068–2069.
33. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinforma Oxf Engl* 2015;31:3691–3693.
34. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma Oxf Engl* 2014;30:1312–1313.
35. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, *et al*. In silico serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genom* 2016;2:e000064.
36. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb Genom* 2018;4.
37. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. RhierBAPS: An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Res* 2018;3:93.
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
39. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, *et al*. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;3:e000131.
40. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* 2019;47:D687–D692.
41. Frömmel U, Lehmann W, Rödiger S, Böhm A, Nitschke J, *et al*. Adhesion of human and animal *Escherichia coli* strains in association with their virulence-associated genes and phylogenetic origins. *Appl Environ Microbiol* 2013;79:5814–5829.
42. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;17:238.
43. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinforma Oxf Engl* 2011;27:2987–2993.
44. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res* 2019;47:5539–5549.
45. Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:595.
46. Seemann T. *mlst*. GitHub; (n.d.). https://github.com/tseemann/mlst
47. Okonechnikov K, Golosova O, Fursov M, UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 2012;28:1166–1167.
48. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28:3150–3152.
49. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 2021;49:D344–D354.
50. Brunson JC. ggalluvial: layered grammar for alluvial plots. *J Open Source Softw* 2020;5:2017.
51. R Core Team. R: A language and environment for statistical computing. In: *R Foundation for Statistical Computing*. Vienna, Austria, 2021.
52. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. 2nd edn. Springer International Publishing, 2016.
53. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In: Kotz S and Johnson NL (eds). *Breakthroughs in Statistics: Methodology and Distribution Springer Series in Statistics*. New York, NY: Springer; 1992. pp. 11–28.
54. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
55. Murphy R, Palm M, Mustonen V, Warringer J, Farewell A, *et al*. Genomic epidemiology and evolution of *Escherichia coli* in wild animals in Mexico. *mSphere* 2021;6:e00738-20.
56. Osińska M, Nowakiewicz A, Zięba P, Gnat S, Łagowski D. Wildlife omnivores and herbivores as a significant vehicle of multidrug-resistant and pathogenic *Escherichia coli* strains in environment. *Environ Microbiol Rep* 2020;12:712–717.
57. Schembri MA, Klemm P. Biofilm formation in a hydrodynamic environment by novel fimh variants and ramifications for virulence. *Infect Immun* 2001;69:1322–1328.
58. Schierack P, Rödiger S, Kolenda R, Hiemann R, Berger E, *et al*. Species-specific and pathotype-specific binding of bacteria to zymogen granule membrane glycoprotein 2 (GP2). *Gut* 2015;64:517–519.
59. Sokurenko EV, Chesnokova V, Dykhuizen DE, Ofek I, Wu XR, *et al*. Pathogenic adaptation of Escherichia coli by natural variation of the FimH adhesin. *Proc Natl Acad Sci U S A* 1998;95:8922–8926.
60. Saldaña Z, Xicohtencatl-Cortes J, Avelino F, Phillips AD, Kaper JB, *et al*. Synergistic role of curli and cellulose in cell adherence and biofilm formation of attaching and effacing *Escherichia coli* and identification of Fis as a negative regulator of curli. *Environ Microbiol* 2009;11:992–1006.
61. Ali A, Kolenda R, Khan MM, Weinreich J, Li G, *et al*. Novel avian pathogenic *Escherichia coli* genes responsible for adhesion to chicken and human cell lines. *Appl Environ Microbiol* 2020;86:e01068-20.
62. Saldaña Z, De la Cruz MA, Carrillo-Casas EM, Durán L, Zhang Y, *et al*. Production of the Escherichia coli common pilus by uropathogenic *E. coli* is associated with adherence to HeLa and HTB-4 cells and invasion of mouse bladder urothelium. *PLoS One* 2014;9:e101200.
63. Barroso-Batista J, Sousa A, Lourenço M, Bergman M-L, Sobral D, *et al*. The first steps of adaptation of Escherichia coli to the gut are dominated by soft sweeps. *PLoS Genet* 2014;10:e1004182.
64. Kisiela DI, Chattopadhyay S, Libby SJ, Karlinsey JE, Fang FC, *et al*. Evolution of Salmonella enterica virulence via point mutations in the fimbrial adhesin. *PLoS Pathog* 2012;8:e1002733.
65. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, *et al*. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* 2016;1:16041.

66. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 2017;45:D200–D203.
67. Choi U, Lee C-R. Distinct roles of outer membrane porins in antibiotic resistance and membrane integrity in *Escherichia coli*. *Front Microbiol* 2019;10:953.
68. Hejair HMA, Zhu Y, Ma J, Zhang Y, Pan Z, *et al*. Functional role of ompF and ompC porins in pathogenesis of avian pathogenic Escherichia coli. *Microb Pathog* 2017;107:29–37.
69. Patrick M, Gray MD, Sandkvist M, Johnson TL. Type II Secretion in *Escherichia coli*. *EcoSal Plus* 2010;4.
70. Laestadius A, Richter-Dahlfors A, Aperia A. Dual effects of *Escherichia coli* alpha-hemolysin on rat renal proximal tubule cells. *Kidney Int* 2002;62:2035–2042.
71. Nhu NTK, Phan M-D, Forde BM, Murthy AMV, Peters KM, *et al*. Complex multilevel control of hemolysin production by uropathogenic *Escherichia coli*. *mBio* 2019;10:e02248-19.