cmgh
CELLULAR AND MOLECULAR GASTROENTEROLOGY AND HEPATOLOGY

ORIGINAL RESEARCH

Genetic and Microbial Analysis of Invasiveness for *Escherichia coli* Strains Associated With Inflammatory Bowel Disease

Jungyeon Kim,[1,*] Jing Zhang,[2,3,*] Lisa Kinch,[4,5,*] Jinhui Shen,[3,*] Sydney Field,[4,5] Shahanshah Khan,[1] Jan-Michael Klapproth,[6] Kevin J. Forsberg,[7] Tamia Harris-Tryon,[8] Kim Orth,[4,5,9] Qian Cong,[2,3,10] and Josephine Ni[1,7]

[1]*Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas;* [2]*Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas;* [3]*Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas;* [4]*Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, Texas;* [5]*Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas;* [6]*Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania;* [7]*Department of Microbiology, University of Texas Southwestern Medical Center, Dallas, Texas;* [8]*Department of Dermatology, University of Texas Southwestern Medical Center, Dallas, Texas;* [9]*Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas; and* [10]*Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas*

SUMMARY

We established a pangenome of *Escherichia coli* isolates from patients with inflammatory bowel disease using whole genome sequencing. The genotypes were correlated with a newly developed measurement of invasion phenotype to identify cosegregating genes.

BACKGROUND & AIMS: The adherent-invasive *Escherichia coli* (AIEC) pathotype is implicated in inflammatory bowel disease (IBD) pathogenesis. AIEC strains are currently defined by phenotypic measurement of their pathogenicity, including invasion of epithelial cells. This broad definition, combined with the genetic diversity of AIEC across patients with IBD, has complicated the identification of virulence determinants. We sought to quantify the invasion phenotype of clinical isolates from patients with IBD and identify the genetic basis for their invasion into epithelial cells.

METHODS: A pangenome with core and accessory genes (genotype) was assembled using whole genome sequencing of 168 *E coli* samples isolated from 13 patients with IBD. A modified assay for invasion of epithelial cells (phenotype) was established with consideration of antibiotic resistance phenotypes. Isolate genotype was correlated to invasiveness phenotype to identify genetic factors that cosegregate with invasion.

RESULTS: Pangenome-wide comparisons of *E coli* clinical isolates identified accessory genes that can cosegregate with invasion phenotype. These correlations found the acquisition of antibiotic resistance genes in clinical isolates compromised the traditional gentamicin protection assays used to quantify invasion. Therefore, an alternate assay, based on amikacin resistance, identified genes cosegregating with invasion. These genes encode an arylsulfatase, a glycoside hydrolase, and genetic islands carrying propanediol utilization and sulfoquinovose metabolism pathways.

CONCLUSIONS: This study highlights the importance of incorporating antibiotic resistance screening for invasion assays used in AIEC identification. Accurately screened invasion phenotypes identified accessory genome elements among *E coli* IBD isolates that correlate with their ability to invade epithelial cells. These results help explain why single genetic markers for the AIEC phylotype are challenging to identify. *(Cell Mol Gastroenterol Hepatol 2025;19:101451; https://doi.org/10.1016/j.jcmgh.2024.101451)*

Keywords: Crohn's Disease; Inflammatory Bowel Disease; Ulcerative Colitis; Adherent-Invasive *Escherichia Coli*; Virulence Factor; Antibiotic Protection Assay; Cell Invasion; Antibiotic Resistance; Comparative Genomics.

This article has an accompanying editorial.

Escherichia coli are common colonizers of the vertebrate digestive tract and appear frequently as commensal constituents of the human gut microbiome.[1,2] Many *E coli* variants possess virulence factors that facilitate gut or urinary tract infections, even in otherwise healthy individuals.[3,4] The presence of these virulence factors and the replicative niche the bacteria colonize during infection forms the basis of the pathotype categorization of these *E coli*.[3] Although the *E coli* core genome, which determines the bacteria's phylogenetic classification, is highly conserved across both pathogenic and commensal strains, pangenomic analyses

*Authors share co-first authorship.

Abbreviations used in this paper: **AIEC, adherent-invasive *Escherichia coli*; AslA, arylsulfatase; CD, Crohn's Disease; GH127, glycoside hydrolase family 127; IBD, inflammatory bowel disease; IntI1, class 1 integron integrase; MIC, minimum inhibitory concentration; MLST, multilocus sequence typing; MSI, mean survival index; NCBI, National Center for Biotechnology Information; PBS, phosphate-buffered saline; PIS, percent invasion score; SQ, sulfoquinovose.**

Most current article

2352-345X
https://doi.org/10.1016/j.jcmgh.2024.101451

reveal diversity within a highly variable accessory genome that distinguishes pathogenic *E coli* from their commensal counterparts and from one another.[3,5–7] Analyses of the accessory genome have likewise illuminated how horizontal genetic acquisition, genetic diversification, and gene loss have blurred the lines differentiating pathotypes.[3,6]

The adherent-invasive *E coli* (AIEC) pathotype has classically been characterized by bacterial invasion of the host intestinal epithelium and propagation within macrophages without inducing cell death.[8–10] Although AIEC can also be identified in healthy individuals, they are consistently enriched in the microbiota of patients with inflammatory bowel disease (IBD), especially Crohn's disease (CD).[8,10] They correlate with elevated gut inflammation and dysbiosis of the intestinal microbiome.[11,12] Furthermore, they cause disease when inoculated into murine colitis models.[13] Many efforts have been made to identify AIEC strains genotypically, but they exhibit significant diversity across individual patients with CD.[14] This diversity across clinical isolates presents a challenge in linking the AIEC pathotype with specific genetic markers. Few virulence factors have been identified in AIEC accessory genomes, and no consensus exists for genetic markers within the accessory genome.[15,16] Because of this limitation, the gold standard for identifying AIEC is a phenotypic gentamicin protection assay where intracellular bacterial invasion is quantified.[17,18] Unfortunately, *E coli* isolates from patients with CD exhibit elevated frequencies of antibiotic resistance relative to those from healthy individuals, partially because patients with CD are immunocompromised and are frequently exposed to antibiotics as part of their treatment.[19,20]

Here, using 168 *E coli* samples isolated from 13 patients with IBD, we aimed to evaluate their invasiveness and correlate such ability with genetic markers. However, we found that gentamicin protection assays were not trustworthy because we observed frequent gains of gentamicin resistance genes, resulting in false-positives for invasion. We thus developed an alternative amikacin protection assay that proved more reliable. Based on these data, we recommend including an antibiotic resistance screen in the phenotypic identification of AIEC, and using the amikacin protection assays to complement the classic gentamicin protection assays, which may eliminate false-positives and provide a more stringent filter for pathotype constituents. Integrating results from our improved assays and comparative genomic analyses, we identified genes that cosegregate with the invasiveness. These include genomic neighborhoods encoding propanediol utilization and sulfoquinovose (SQ) metabolism, and an arylsulfatase (AslA) and a glycoside hydrolase (glycoside hydrolase family 127 [GH127]).

Results

Clinical Escherichia coli *Isolate Collection and Genomic Sequencing*

A total of 49 subjects undergoing colonoscopy consented to sample collection, and their demographics are summarized in Table 1. Bacterial growth on MacConkey agar plates was detected in the form of individual lactose-fermenting and nonfermenting colonies. A total of 830 bacterial isolates were further analyzed in triplicate by an in vitro gentamicin protection assay with differentiated Caco-2 epithelial cells, including 378 isolates from patients with CD, 111 from patients with ulcerative colitis, and 341 from normal control subjects. From these assays, 117 AIEC were tentatively identified based on significant epithelial invasion defined as a percent invasion of $>1\%$.[21] Consistent with previous studies, more AIEC strains were found in patients with CD (26%) versus patients with ulcerative colitis (18%) and normal control subjects (13%).[21]

We attempted to obtain whole genome sequences using both Nanopore and Illumina platforms for 213 isolates from this clinical collection that were characterized as *E coli* via biochemical testing with the Api-20E system. The Illumina short reads and Nanopore long reads were combined to assemble the genomes of these candidate *E coli* isolates. Analysis of these assembled genomes by National Center for Biotechnology Information (NCBI) Prokaryotic Genome Annotation Pipeline revealed that 45 (21%) of them were not *E coli* based on the genomic data, and instead, they were other microbes associated with humans, such as *Citrobacter rodentium* and *Klebsiella pneumoniae*.

Therefore, we focused on the remaining 168 *E coli* isolates in our study. On average, we obtained 350 Mbp and 420 Mbp Illumina and Nanopore reads, respectively, for each isolate (Supplementary Table 1). These reads allowed us to assemble high-quality genomes for nearly all isolates: 154 (92%) were each assembled into 1 long scaffold corresponding to the chromosome and several short scaffolds corresponding to plasmids or regions difficult to assemble into the chromosome (Supplementary Table 1). The assembled genomes for these isolates range between 4.7

Table 1. Subject Demographics

	CD	UC	Indeterminate colitis	Control subjects
Number of subjects	23	12	2	12
Female sex, %	60	75	50	57
Mean age	44	38	54	60
Mesalamine treatment, %	48	83	0	—
6-MP/AZA treatment, %	30	25	0	—
Steroid treatment, %	26	8	0	—
Anti-TNF treatment, %	17	17	0	—

6-MP, 6-mercaptopurine; AZA, azathioprine; CD, Crohn's disease; TNF, tumor-necrosis factor; UC, ulcerative colitis.

Mbp and 5.5 Mbp, and they encode 4.3k to 5.1k proteins, typical for *E coli*. We further assessed the quality of these assembled genomes by CheckM,[22] which evaluates the completeness by the presence of a set of universal single-copy genes expected to be present in a lineage. CheckM also detects potential contaminations in genome assemblies by the presence of more than 1 copy of such universal single-copy genes. Our genome assemblies show a median completeness of 99.2% and contamination of 0.37%; both are better than the median values (99.0% and 0.59%) of high-quality bacterial genomes in NCBI.

Clonal Nature of Patient Isolates

Phylogenetic analysis of all sequenced isolates highlighted their clonal nature among patients (Figure 1*A*). Similar to previous reports for *E coli* isolates from patients with CD,[14,20,23] the isolates collected from different patients with CD, ulcerative colitis, and indeterminate colitis are phylogenetically diverse and span 6 phylogroups (A, B1, B2, D, E, and G) (Figure 1*A*). However, isolates collected from the same patient were mostly clonal, regardless of the biopsy location. The clonal nature of isolates from the same patient is also reflected by the average sequence identity between orthologous segments (Figure 1*B*). Although isolates from different patients show sequence identity between 96.3% and 99.1% (green dots in Figure 1*B*), pairs from the same patient mostly show 100% sequence identity (orange dots in Figure 1*B*).

The shared gene content among isolate genomes positively correlates with the sequence identity between isolates (Figure 1*B*). However, compared with sequence identity, the gene content of these isolates shows a much larger variability: isolates from different patients share 74%–94% of genes, and isolates showing 100% sequence identity in orthologs may still possess slightly different sets (<5%) of genes (Figure 1*B*). Such variability is common among bacterial isolates[24] and highlights the important role of horizontal gene transfer in the evolution of microbes. As a result of horizontal gene transfer, the sequenced isolates process different sets of accessory genes in addition to the shared core genomes.

To characterize the diverse set of proteins encoded by the sequenced *E coli* isolates, we identified orthologous groups among these proteomes and assembled a pan-genome reference protein set that includes 1 representative protein per orthologous group (Supplementary Table 2). The presence or absence of each orthologous group in each isolate is shown as a heatmap in Figure 1*C*. In this heatmap, the reference proteins are ordered to approximate their genomic distances, placing each gene near its genomic neighbors. The core genes shared by all isolates are mostly encoded by the chromosome, and they cluster in the middle of this heatmap. In contrast, the accessory genes, frequently encoded on plasmids, are placed around the core genome in the heatmap. This distribution underscores that although core genes are shared by all isolates, the set of accessory genes varies between isolates. As expected, isolates from the same patients tend to share a similar set of accessory genes, which clearly distinguishes each patient's isolates from the rest. Because of the clonal nature of the patient isolates, we selected a less redundant set of 32 reference isolates, which displays a similar distribution of sequence identity and shared gene content as the complete set (Figure 1*B*, metadata summarized in Table 2).

Despite the clonal tendency of isolates, different *E coli* strains (according to multilocus sequence typing [MLST]) existed in 3 of the patients with CD (patient II, XI, and XXXIII), which has been reported for other patients with CD.[23] When multiple strains existed within a single patient, they belonged to alternate phylogroups. For example, isolates from patient XI belong to 2 clonal populations represented by phylogroup B2 (closest to the NRG 857c AIEC strain) and phylogroup G. The phylogroup G isolates from patient XI belong to the main sequence type (ST117) for the group. Previous epidemiologic studies on ST117 isolates suggested their lineage is associated with poultry, but they can cause extraintestinal disease in humans.[25] Similarly, 2 patients presented with isolates from both phylogroups A (ST216 in patient II and ST607 in patient XXXIII) and B1 (ST224 in patient II and ST8492 in patient XXXIII), with 1 of the B1 sequence types (ST224) reported as a high-risk lineage with antibiotic resistance found in healthy chickens.[26]

Antibiotic Resistance Compromises Gentamicin Protection Assays for AIEC Invasion

All AIEC isolates were initially characterized by field-standard gentamicin protection assays to measure their ability to invade epithelial cells (Figure 2*A*). The mean survival index for this assay (MSI_{gent}) displayed a broad range of values (0.022–35) across all isolates (Supplementary Table 3), with the highest invasion scores stemming from clonal isolates of a single patient (IV). To assess if any protein-coding genes segregate with bacterial invasion, we correlated the genetic differences between isolates with the MSI_{gent} scores (Supplementary Table 4). Two of these correlations are shown in Figure 2*B*, with the top-ranked protein annotated as aminoglycoside N-acetyltransferase AAC(3)-VIa (AAC-VIa). This enzyme inactivates aminoglycoside antibiotics, including gentamicin, by acetylating the 3-amino group, providing gentamicin resistance in bacteria that express the protein.[27] The identification of this cosegregating gentamicin resistance gene suggests the inference of invasion phenotype from the MSI_{gent} scores is compromised. However, we provide functional annotations for proteins encoded by all genes correlating with MSI_{gent} scores for completeness (Supplementary Table 5).

The AAC-VIa gene is present in *E coli* isolates from 2 patients (IV and XLII), with each located near a class 1 integron integrase (IntI1) (Figure 2*B*). These integrons enable the horizontal spread of genetic determinants across clinical isolates and play an important role in the transmission of drug resistance.[28,29] The AAC-VIa neighborhood of patient IV represents a classic arrangement of the IntI1 adjacent to an array of gene cassettes, which include the gentamicin resistance gene and another aminoglycoside antibiotic that nucleotidylylates streptomycin at the position 3'' hydroxyl group (ANT [3'']), according to the Comprehensive Antibiotic Resistance Database.[30] Comprehensive Antibiotic Resistance Database also helped us find another

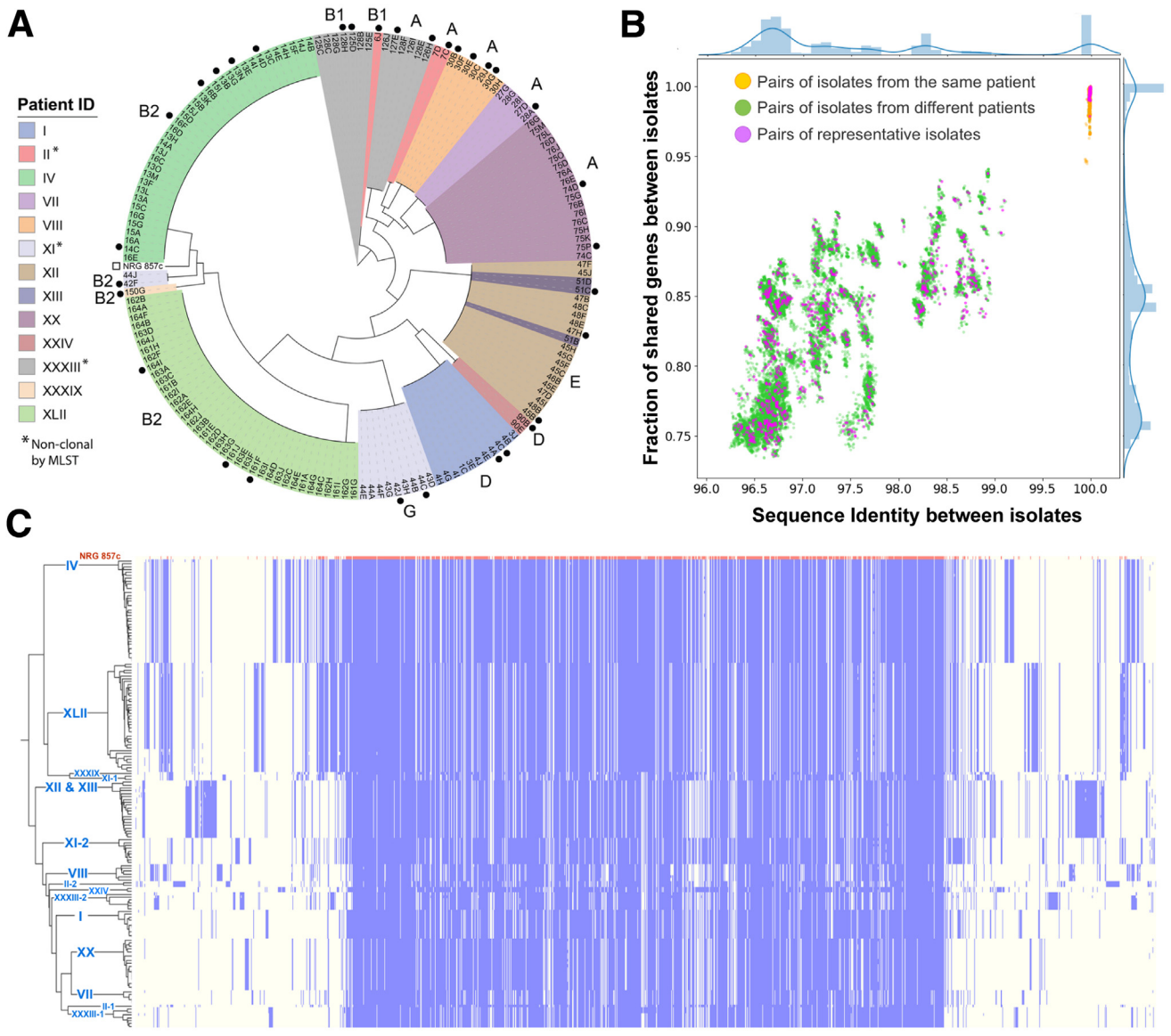


Figure 1. Pangenome of clonal patient *Escherichia coli* isolates. (*A*) Maximum-likelihood phylogenetic tree based on the concatenated protein multiple sequence alignment of orthologous groups. Isolates are colored by patient ID, and patients with nonclonal isolates by MLST typing are indicated by *asterisk*. Selected reference isolates (*black circle*) and the public AIEC strain (*open square*) are indicated outside the tree. Phylotypes for each clonal group are labeled outside the tree. (*B*) The similarity between *E coli* isolates is measured by the average nucleotide identity (x-axis) and the fraction of shared genes (y-axis). Each *dot* in the scatterplot represents a pair of isolates, and the *bar plots* by the scatterplots show the distribution of values on these axes. (*C*) A heatmap showing the presence (*blue*) and absence (*light yellow*) of genes (x-axis) across different *E coli* isolates (y-axis). The genes are ordered along the axes based on their average genomic distances in all the isolates shown in the graph. The first row of the heatmap represents the reference strain, NRG 857c, whereas other rows show the *E coli* isolates sequenced in this study. These isolates are ordered based on their phylogeny on the *left* of the heatmap. Patient IDs associated with these isolated are labeled on the tree branch.

gene related to a known sulfonamide-resistant dihydropteroate synthase and an N-acetyltransferase (gnat) that may also modify an antibiotic in the upstream region to AAC-VIa. Thus, isolates from patient IV likely acquired gentamicin resistance through IntI1-mediated recombination into the integron array of gene cassettes. The AAC-VIa neighborhood of patient XLII, despite being near an IntI1 and its cassettes, is flanked by IS6 transposases (Figure 2*B*). Thus, the IS6 (or adjacent IS4) transposase could have mediated the acquisition of gentamicin resistance as a passenger gene in this isolate.[31]

Given the identification of gentamicin and other potential antibiotic-resistance cassettes in the isolates of 2 patients, we comprehensively searched for known antibiotic-resistance genes encoded by all the sequenced genomes to predict each isolate's drug resistance (Supplementary Table 6). Isolates from all patients are predicted to be resistant to beta-lactam, and all but patient VII are likely

Table 2. Selected Metadata/MLST for Reference Isolates

Patient ID	MSI	Strain	Inflamed	Biopsy location	Phylo group	MLST
I	0.98	4A	No	Colon	D	38
I	0.11	4C	No	Colon	D	38
II-1	0.05	6J	No	Colon	B1	224
II-2	0.08	7C	No	Colon	A	216
IV	19.67	13B	Yes	Terminal ileum	B2	141
IV	30.18	13N	Yes	Terminal ileum	B2	141
IV	21.04	14C	Yes	Terminal ileum	B2	141
IV	29.88	14D	Yes	Terminal ileum	B2	141
IV	4.93	16B	No	Cecum	B2	141
IV	26.19	16F	No	Cecum	B2	141
VII	0.04	28A	No	Colon	A	10
VIII	2.84	29J	No	Colon	A	46
VIII	2.48	30B	No	Colon	A	46
VIII	1.58	30E	No	Colon	A	46
VIII	0.19	30G	No	Colon	A	46
XI-1	0.22	42F	No	Colon	B2	144
XI-2	0.23	42J	No	Colon	G	117
XI-2	1.53	44C	Yes	Colon	G	117
XII	1.68	45B	Yes	Pouch	E	11
XII	1.50	47H	Yes	Pouch	E	11
XIII	0.38	51C	No	Colon	E	11
XX	0.02	74D	Yes	Cecum	A	2223
XX	2.36	75P	No	Colon	A	2223
XXIV	0.10	90B	No	Colon	D	130
XXXIII-1	0.21	125D	No	Terminal ileum	B1	8492
XXXIII-1	4.13	128H	No	Colon	B1	8492
XXXIII-2	0.13	126H	No	Anastomosis	A	607
XXXIII-2	10.02	127E	No	Colon	A	607
XLII	1.56	161F	No	Colon	B2	131
XLII	1.20	161J	No	Colon	B2	131
XLII	2.72	164I	No	Terminal ileum	B2	131

MLST, multilocus sequence typing; MSI, mean survival index.

resistant to fosfomycin. In addition to their resistance to gentamicin, isolates from patient XLII include probable resistance to 10 of the 15 antibiotics considered. Given this propensity for antibiotic resistance in our isolates, we screened a subset of these strains with additional antibiotics to assess for their minimum inhibitory concentration (MIC) required to prevent growth. Consistent with the predicted gentamicin resistance, patients IV and XLII exhibit MICs for gentamicin higher than the concentrations used in the protection assays (Figure 2*C*). Even more, many other isolates exhibited intermediate resistance (>2 mg/mL) to gentamicin.

Development of an Alternate Antibiotic Protection Assay to Estimate Invasion

The expanded antibiotic screen confirmed resistance among the isolates to additional drugs (Table 3). Patient VIII isolates, which were predicted to be resistant to kanamycin by genomic sequences, were also resistant experimentally. Alternately, patient IV isolates were predicted to be resistant to ampicillin, but the bacteria exhibited low MIC when grown in the presence of the antibiotic. A similar difference between prediction and experimental measurement was seen with patient XLII isolates, which were resistant to ampicillin by experimental results but not predictable by sequence. These discrepancies highlight the complicated nature of antibiotic resistance of bacteria and the necessity for experimental screening before choosing antibiotics in protection assays for invasiveness.

Using the expanded antibiotic screen, we initially tested alternative antibiotics for their ability to penetrate the host cell, because this would compromise our invasion assays. Using multiple classes of antibiotics, including ceftazidime, cefepime, ertapenem, and amikacin, we found that only

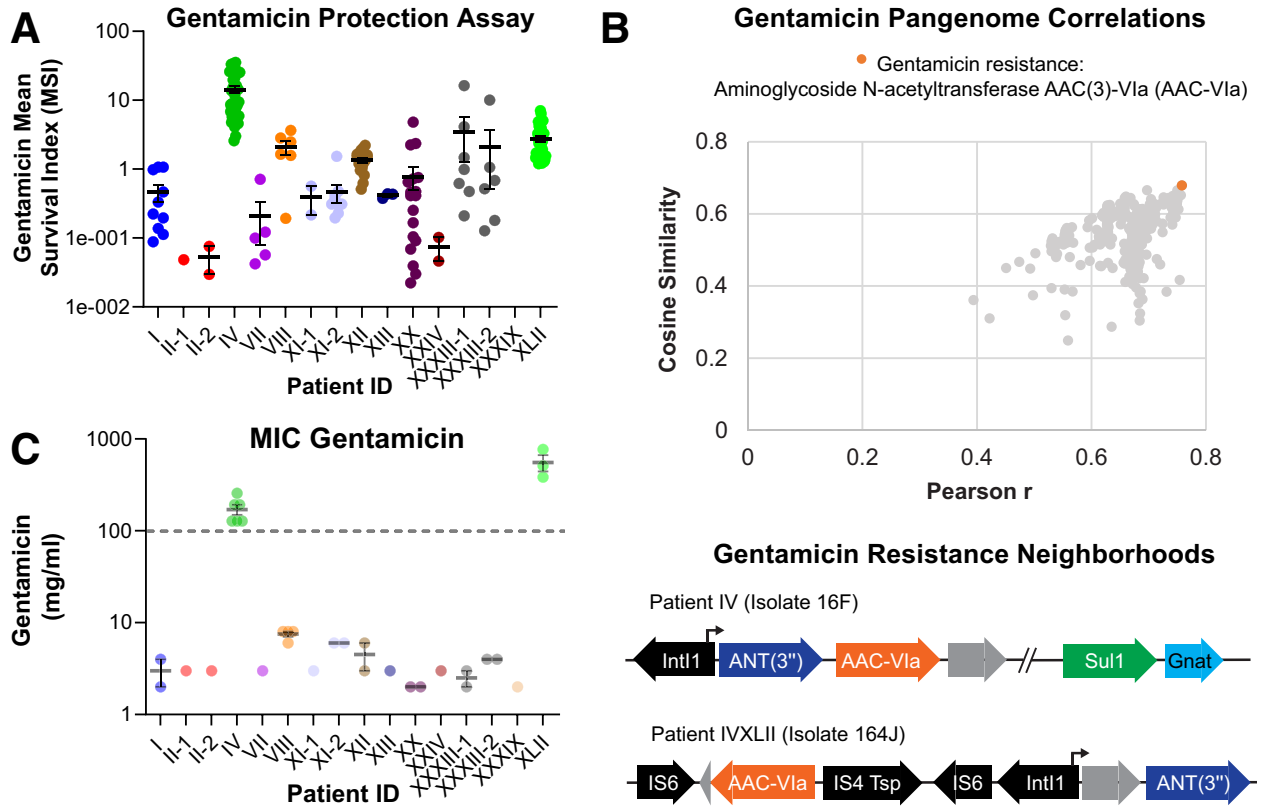


Figure 2. Acquisition of gentamicin resistance by *Escherichia coli* isolates hinders invasion assay. (*A*) MSI measured by gentamicin protection assay is depicted for patient isolates (*colored circles* by patient, as in Figure 1*A*). Isolates were grouped according to MLST type. The mean (*horizontal line*) and standard error of the mean are depicted for all grouped scores. (*B*) Pangenome distance correlation with LogMSI, as measured by cosine similarity (x-axis) is compared with that measured by Pearson correlation (y-axis). Only significantly correlated proteins (by Pearson and quantitative trait locus statistics; see Methods) are shown. Proteins with similar functions are colored and labeled *above* the graph. The top-ranked protein is labeled with its genome neighborhood from patients IV and XVII, illustrated *below*. (*C*) MIC for gentamicin measured for representative patient isolates (*colored circles*, as in *A*). The *dashed line* indicates the concentration used in antibiotic protection assays.

amikacin did not penetrate Caco-2 cells at 10 x MIC (Supplementary Table 7). We also observed a lack of resistance to amikacin among our isolates and a lack of toxicity for Caco-2, with average cell viability of 90% assessed by trypan blue staining after incubation at 10 x MIC for 7 hours. We therefore developed a protection assay using this antibiotic with representative strains defined by the phylogenetic tree and MLST. Interestingly, the patient isolates with acquired gentamicin resistance were on antibiotic treatments at the time of their colonoscopy (Supplementary Table 8). Representative isolates from 6 patients (XI-1, XI-2, XII, XIII, XX, and XXIV) and the positive control, NRG857c, were significantly different from the negative controls in this amikacin protection assay (Figure 3*A*, Supplementary Table 9), suggesting these strains are invasive. Amikacin protection measured for isolates from another patient (II-1) were borderline (0.035 adjusted *P* value, noted by a single asterisk), which we conservatively classified as noninvasive. An isolate from an asymptomatic control patient (XXXIX) with a normal colonoscopy and without evidence of inflammation or colonic polyps was not significantly different from the negative control subjects for invasion (Figure 3*A*).

Cosegregation of Virulence Factors with Invasion Phenotype

We sought to identify genes that correlate with 1 of the phenotypes used to define AIEC strains, namely their invasiveness (MSI_{amk}) as measured by the amikacin protection assays. These gene correlations (Supplementary Table 10) were less significant than those for gentamicin (likely caused by the lowered number of isolate genomes), and the significant gene set included fewer proteins (Supplementary Table 11 and Figure 3*B*). Among these, the most significantly correlated proteins are unknown phage proteins, a GH127 family protein, and AslA. Two sets of proteins functioning in propanediol utilization and SQ metabolism were also correlated to the invasion of Caco2 cells. The propanediol utilization proteins belong to a genome neighborhood (Figure 3*C*) present in the positive control strain for invasiveness (ie, NRG857c, which is commonly found in ileal lesions of patients with CD). This neighborhood was

Table 3. Antibiotic Resistance Screen of AIEC Isolates

Strain	Patient ID	Antibiotic resistance profile[a]			
		Gentamicin	Ceftazidime	Ampicillin	Kanamycin
4A	I	I	S	S	S
4C	I	S	S	S	S
6J	II-1	I	S	S	S
7C	II-2	I	S	S	S
13B	IV	R	S	S	I
13N	IV	R	S	S	S
14C	IV	R	S	S	I
14D	IV	R	S	S	I
16B	IV	R	S	S	I
16F	IV	R	S	S	I
28A	VII	I	S	S	S
29J	VIII	I	S	S	R
30B	VIII	I	S	S	R
30E	VIII	I	S	S	R
30G	VIII	I	S	S	R
42F	XI-1	I	S	S	S
42J	XI-2	I	S	S	S
44C	XI-2	I	S	S	S
45B	XII	I	S	S	S
47H	XII	I	S	S	S
51C	XIII	I	S	S	S
74D	XX	R	S	R	R
75I	XX	R	S	R	R
90B	XXIV	R	S	R	R
125D	XXXIII-1	S	S	S	S
128H	XXXIII-1	S	S	S	S
126H	XXXIII-2	I	S	S	S
127E	XXXIII-2	I	S	S	S
161F	XLII	I	S	S	S
161J	XLII	S	S	S	S
164I	XLII	I	S	S	S
150G	XXXIX	S	S	S	S

AIEC, adherent-invasive *Escherichia coli*; I, intermediate; R, resistant; S, susceptible.
[a]Minimum inhibitory concentration $\geq$ susceptible threshold and $\leq$ resistance threshold.

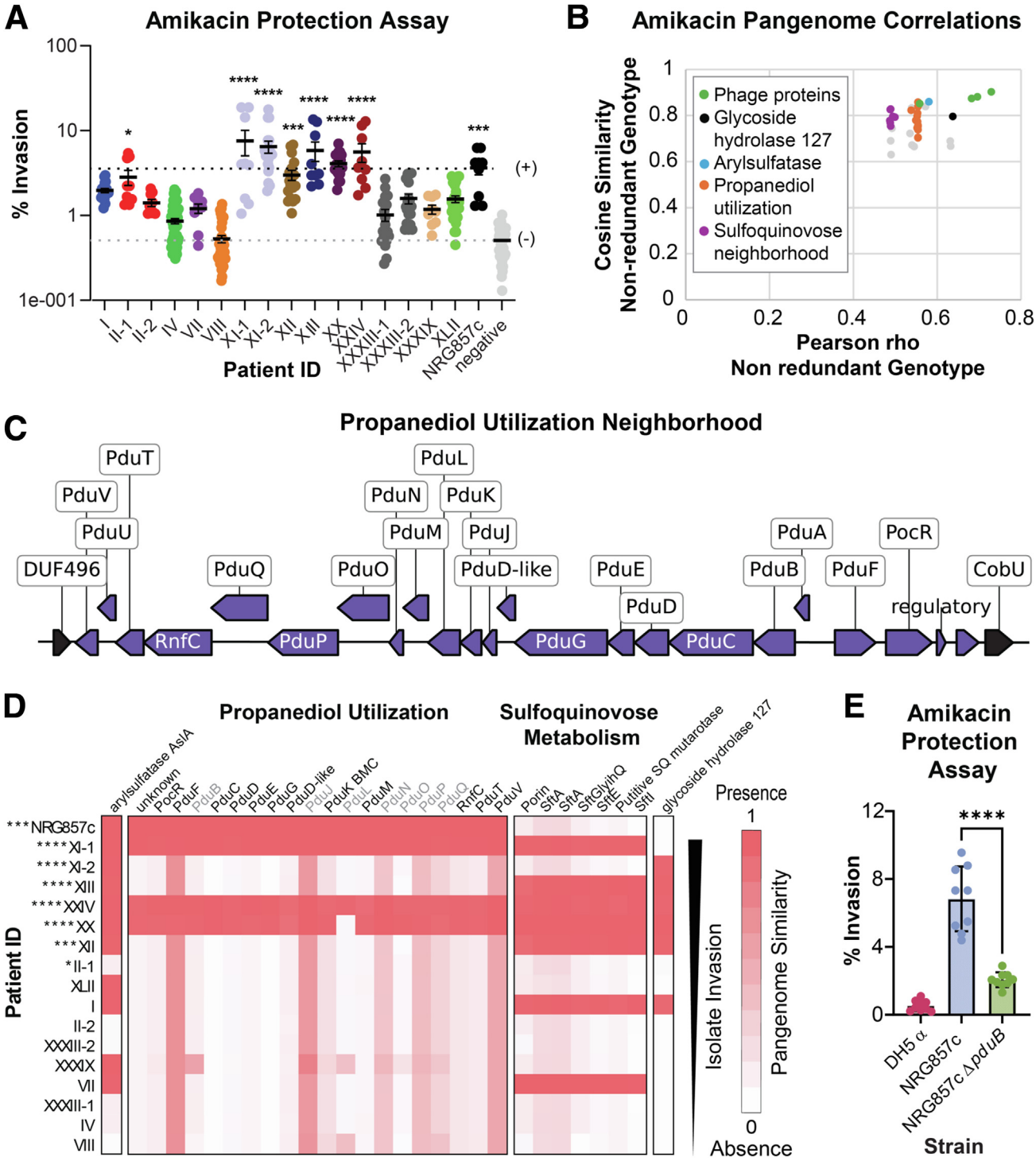
noted as unique to the NRG857c strain by comparative genomic analysis to other pathogenic and commensal *E coli*.[32] Most propanediol utilization proteins are identified as correlated with the invasion phenotype, because they belong to the same genomic neighborhood; exceptions (eg, PduQ and PduJ) exhibit intermediate similarity to proteins participating in ethanolamine utilization, representing a paralogous metabolic pathway to propanediol utilization that is present as a core component of all *E coli* isolates.

The propanediol utilization neighborhood was found in invasive isolates from 3 of 6 patients tested (patients XI-1, XX, and XXIV), and these isolates displayed significantly higher invasion than the negative control subject (Figure 3*B* and *C*, significance marked by asterisk). The presence of this genomic island in patient isolates with different phylotypes (Figure 1*A*, phylotypes B2, A, and D) highlights its mobile nature and suggests its acquisition contributes to the invasion phenotype. Isolates from 3 other patients that were observed to invade Caco-2 cells in the amikacin protection assay lacked this propanediol utilization neighborhood (patients XI-2, XII, and XIII). Two of these patients encode SQ metabolism genes, which also significantly correlate with the invasion phenotype. Finally, all our sequenced invasive isolates have acquired AslA, and all but patient XI-1 and the positive control NRG 857C encode GH127. In sum, different isolates seem to have acquired alternative invasion strategies, and they collectively enable these isolates to colonize human epithelial cells and ultimately cause disease.

To verify that our identified correlated genes affect the invasion phenotype, we constructed NRG857c Δ*pduB* as proof of concept and tested the effect of this deletion on its invasiveness (MSI_{amk}) as measured by the amikacin protection assay. As predicted, deletion of *pduB*, a gene in the propanediol utilization operon, reduced the invasiveness of NRG857c (Figure 3*E*), providing genetic evidence that propanediol utilization is important for epithelial cell invasion.

Discussion

This study analyses bacterial isolates from patients with IBD using comparative genomics to identify factors that can



contribute to cell invasion. We used Illumina short reads and Nanopore long reads to obtain high-quality genome assemblies for our isolate collection. Consistent with previous studies, we show that there is clonality of isolates within each patient regardless of biopsy location, apart from 3 patients with IBD that had more than 1 dominant isolate. Because bacterial invasion of enteric cells can be a key feature of virulent strains of *E coli* in IBD, we further characterized the invasion capacity of the sequenced strains using the established gentamicin protection assay. Surprisingly, several of our strains carried genetic features of gentamicin resistance. Therefore, we developed an alternative antibiotic protection assay, based on sensitivity to amikacin. Through this approach, we were able to identify isolates that survived exposure to amikacin. By correlating each strain's amikacin protection assay score (a proxy for the invasiveness of each isolate) with their genomic sequence, we identified genes involved in propanediol utilization proteins and sulfur metabolism that might be novel *E coli* virulence factors required for invasion. We further validated 2 of our correlational predictions as proof of concept. First, identification of the aminoglycoside *N*-acetyltransferase AAC(3)-VIa as the top correlated predicted protein from our initial gentamicin protection correlations was verified by gentamicin MIC testing. Second, we performed mutational and functional experiments to show that chromosomal deletion of *pduB*, a gene in the propanediol utilization operation, reduced the invasiveness of an AIEC strain. Further validation and assessment of the effects of the other identified potential virulence factors remains a target for future studies.

The observation that a sulfur metabolizing pathway (SQ metabolism) and enzyme (AslA) are among the top cosegregating proteins with the invasion phenotype of *E coli* isolates suggests colonic sulfur metabolism might contribute to invasion and lead to inflammation associated with IBD. Although the extent to which the products of the SQ pathway and the AslA reaction lead to H_2S production by colonic microbiota remains undetermined, the gut microbiome seems to play a significant role in IBD pathology.[33] The relationship between H_2S and colonic health remains debated.[34,35] However, elevated H_2S can exhibit toxic properties to colonocytes, induce inflammation, and increase intestinal permeability.[34,36] Sulfur is provided to the human body in the diet, mainly through protein consumption. It is conceivable that exogenous sulfate levels in the gut may contribute to metabolism-linked AIEC in some genetic backgrounds.

The longest stretch of syntenic cosegregating genes that we identified in our invasive *E coli* isolates encoded propanediol utilization islands. Previous studies have also identified the propanediol utilization proteins as being enriched in pathogenic AIEC compared with control strains,[32] and the pathway seems to contribute to AIEC invasiveness and inflammation.[37,38] Propanediol utilization occurs in a bacterial microcompartment formed by a proteinaceous shell that sequesters the involved chemical reactions into a primitive organelle.[39] The propanediol bacterial microcompartment concentrates low levels of the metabolic enzymes, and volatile reaction intermediates, to enhance pathway flux from fucose and rhamnose precursors through 1,2 propanediol and keep the levels of toxic propionaldehyde low.[40] Such bacterial microcompartments also allow enteric pathogens, such as *Salmonella*, to colonize the mammalian gastrointestinal tract and outcompete native microbiota.[41,42] The presence of propanediol utilization islands in *E coli* isolates from 3 of our patients suggests these bacteria also might take advantage of this niche carbon source for successful colonization and invasion during inflammatory colitis.

The cosegregation of SQ degradation islands in some patient isolates lacking propanediol utilization suggests alternate genetic strategies might contribute to invasion phenotype. SQ is a sulfonated monosaccharide in green vegetables that serves as a nutrient source for select microbes in the gut and can lead to increased microbiota-generated H_2S levels.[43] Bacteria that typically degrade SQ from green vegetables in the human gut are generally associated with a positive impact on human health, bringing into question the pathobiology of SQ metabolism in AIEC invasion. However, microbiome-released H_2S from the products of SQ catabolism can have detrimental consequences on colonic health, including opening the mucus barrier in the colon and allowing bacteria access to the epithelium lining.[36] Another sulfur-metabolizing protein, AslA, cosegregates with invasion. Although AslA has not yet been defined as a contributing factor to AIEC colonic invasion, it is considered an invasion factor for *E coli* in brain microvascular endothelial cells.[44] Sulfatase activity like that potentially catalyzed by AslA can provide access for the bacteria to otherwise heavily sulfated colonic mucin glycans.[45]

Figure 3. (***See previous page***). **Amikacin protection assay for isolate invasion.** (*A*) Amikacin invasion assay for reference isolates (*colored circles*, as in Figure 2) with mean (*horizontal line*) and standard error of the mean depicted for MLST grouped reference isolates. *Horizontal dashed lines* represent the means for the NRG857c-positive control (*black*) and all negative controls (*gray*). Patients with percent invasion scores that are significantly different from the negative control subjects are indicated by *asterisk*. (*B*) Pangenome distance correlation with control scaled amikacin protection scores are measured by cosine similarity calculated for nonredundant isolates (x-axis) and Pearson correlation calculated for nonredundant isolates (y-axis). Only significantly correlated proteins (by Pearson and quantitative trait locus statistics; see Methods) are shown. Proteins with similar functions are colored and labeled in the graph insert. (*C*) The propanediol utilization gene neighborhood (*purple arrows* labeled *above*) is depicted for the NRG857c AIEC reference strain, with the flanking genes indicated by *black arrows*. (*D*) Pangenome similarity of patient isolates to AslA, propanediol utilization gene neighborhood, SQ metabolism neighborhood, and Gly127; with scaled scores between 0 (absent) and 1 (present). Patient isolates are ordered *below* the NRG857c AIEC positive strain from highest (*top*) to lowest (*bottom*) amikacin protection scores. (*E*) Amikacin invasion assay for NRG857c Δ*pduB*, NRG857c (positive control), and DH5α (negative control). Reported significance values for NRG857c Δ*pduB* were calculated by Mann Whitney test ($P < .0001$).

This activity might provide invading bacteria with glycans as a nutrient source, with access to the intestinal epithelium, or may increase sulfate levels for colonic sulfur metabolism. Another more universally encoded enzyme among invasive isolates, GH127, breaks down l-arabinofuranose-β1,2-l-arabinofuranose as a potential nutrient source, perhaps also acquired from plant polymers in the diet.[46]

Comparison of gene content between different isolates in this study also identified a considerably wider range of shared orthologs compared with sequence identity, with as little as 74% of shared genes among our isolates compared with 96%–99% average nucleotide identity. Genetic variability can result from horizontal genetic transfer, genetic diversification, and gene loss, especially under the pressure of antibiotic selection. Patients with IBD often have high antibiotic exposure and thus may be at risk for acquiring antibiotic-resistant organisms. Antibiotics, such as ciprofloxacin and metronidazole, have been used to treat IBD, and the prevalence of patients that have been exposed to these antibiotics is as high as 15%.[47–49] Furthermore, patients with IBD have a higher incidence of conditions that require antibiotics, including bacterial overgrowth manifesting from gut microbial dysbiosis and an increased prevalence of *Clostridium difficile* infections,[49,50] abscesses, and wound infections.[51,52] Biologics and immunomodulators, the current mainstay of CD therapy, have added additional exposure to antibiotics, because they predispose patients to infection.

Gentamicin protection assays are the current standard of practice to test for invasion of bacterial isolates and phenotypically define them as AIEC.[17,18] However, in this study we found that 28.1% of the clinical isolates survived gentamicin protection. Using whole genome sequencing of this data set, we then searched for genes that correlate with the results of the gentamicin invasion assays and found AAC-VIa, an enzyme that provides resistance to gentamicin, to be the top-ranked gene. Based on these results, we concluded that it was likely that the survival of these strains in the presence of gentamicin was not caused by invasion, but by gentamicin resistance. With additional testing, we found resistance to multiple antibiotics among our clinical isolates, highlighting the necessity of determining MIC for a broad panel of antibiotics before performing cell protection assays as a proxy for invasion. This practice will contribute to eliminating false-positives for the identification of invasive phenotype for AIECs.

Taken together, these findings highlight the information gained through sequencing. Despite the ever-increasing availability of whole genome sequences and the power of comparative genomics, genetic markers for the AIEC pathotype are lacking. Only a few virulence factors have been described in AIEC, including adhesins, such as fimbriae,[53] siderophore-associated uptake proteins,[37] and capsule synthesis proteins.[54] However, none of these are specific to the pathotype. Therefore, our approach using long-read and short-read sequencing paired with accurate phenotypic characterization was able to identify putative virulence genes in strains of *E coli* from patients with IBD.

Materials and Methods

Collection of Bacterial Isolates and Measurement of Mean Invasion Index

This study involved the prospective collection of clinical *E coli* isolates from eligible subjects undergoing colonoscopy, ileoscopy, or flexible sigmoidoscopy at a single center as described.[21] Briefly, all participants signed informed consent, and the Emory Institutional Review Board approved this study. Patients with clinically confirmed IBD were recruited between July 2002 and August 2005, and 4 biopsies were obtained from each patient during outpatient colonoscopy. As previously described, biopsies were incubated in Dulbecco's Modified Eagle's Medium (Invitrogen, Carlsbad, CA) supplemented with 100 μg/mL gentamicin (Invitrogen) for 1 hour, washed 3 times in phosphate-buffered saline (PBS) and lysed in 1% Triton-X-100/PBS. Aliquots were cultured on MacConkey agar plates at 37°C overnight and lactose-fermenting colonies were enumerated and propagated in Luria-Bertani broth for 4 hours at 37°C under aerobic conditions. Individual clones were stored in 50% glycerol at −80°C until further use.

Isolate Sequencing

Total DNA was extracted from overnight cultures of clinical isolates (in Luria-Bertani broth at 37°C) using the Qiagen DNeasy Blood & Tissue kit. We prepared a paired-end library for Illumina sequencing and a long-DNA library for Nanopore sequencing for each isolate. After sheering the gDNA with NEBNext Ultra II FS kit, we prepared paired-end libraries with the NEBNext Ultra DNA Library Prep kit for Illumina. We pooled libraries of different samples and sequenced them using the Hiseq X ten sequencing service from Azenta, targeting a 100-fold coverage (0.5 Gb) per isolate. For long-read libraries, we used the LSK109 kit with Native Barcodes from kits EXP-NBD114, EXP-NBD114, and EXP-NBD196 for multiplexing. We made the following modifications to the standard protocol for the LSK109 kit: we extended the incubation time for DNA repair and end-preparation to 1 hour at 20°C; and we determined the needed Adapter Mix II volume based on the amount of pooled DNA during the adapter ligation step and used 6.0 μL Adapter Mix II per 1.0 μg DNA. We sequenced the pooled libraries using the MinION flow cell R9.4.1, targeting 100-fold coverage (0.5 Gb) per isolate.

Genome Assembly and Annotation

Illumina sequences were processed with Trimmomatic 0.38[55] to remove sequencing adapters, and low-quality bases (quality score $<$20) and discard reads shorter than 30 bp after such trimming. We performed base-calling, demultiplexing, and adaptor removal with Guppy, a software provided by Nanopore. Combining the Illumina and Nanopore reads, we assembled the genome of each isolate using Unicycler[56] with default settings. We used the NCBI Prokaryotic Genome Annotation Pipeline[57] with default settings to assign taxonomy to each isolate, predict protein-coding sequences in the genomes, and annotate each protein

with function. Additionally, we used the Artificial Intelligence tools ProteInfer,[58] and ProtNLM[59] to obtain additional function annotations.

Isolate Phylogenetics, MLST Assignment, and Genetic Diversity Estimation

The assembled genomic sequences from each isolate were submitted to the M1CR0B1AL1Z3R Web server[60] to extract orthologous sets of putative open reading frames and reconstruct a phylogenetic tree from the common set of orthologs. The phylogenetic tree was visualized using the interactive tree of life (iTOL) v5 online tool.[61] MLST analysis was carried out by searching[62] our genomes against protein sequences from the *Escherichia* sequence typing database at PubMLST[63] and comparing results to known *E coli* (Achtman) allelic profiles. To determine the phylogroup, genomic sequences of representative isolates from each MLST group were submitted to ClermonTyping.[25,64] To estimate the genetic diversity among isolates, we used fastANI[65] to estimate the average nucleotide identity and fraction of shared genes between each pair of isolates.

Building a Pangenome Proteome Reference

We built a pangenome reference protein set to enable comparative analysis. We classified proteins in all the *E coli* isolates by OrthoFinder (v2.5.4).[66] We used MAFFT[67] to align proteins in each orthologous group and identified positions shared by 25% of proteins. Proteins covering at least 95% of such shared positions were considered candidate reference proteins. In rare cases where no protein covered 95% or more of the shared positions, we considered proteins covering at least $C_{max} - 10$ shared positions as candidates, where C_{max} is the largest number of shared positions covered by a protein. Among these candidates, we selected 1 reference per orthologous group, prioritizing proteins from the reference strain NRG857c or isolates with higher invasiveness as determined by gentamicin protection assays. We calculated the number of other proteins between the coding genes of every protein pair in every *E coli* genome (ie, $dist_{x,y}$). We computed the genomic distance between each pair of reference proteins X and Y as $DIST_{X,Y} = median\,(dist_{x,y})$, where x and y are proteins from the same species and belong to the orthologous groups represented by X and Y, respectively. The high-dimensional matrix with $DIST_{X,Y}$ for all pairs of reference proteins was reduced to a 1-dimensional vector using the manifold.MDS function from the sklearn module of Python; values in this vector were used to order proteins in the pangenome reference in Figure 1.

Pangenome Protein Distances and Correlation with Invasion

We searched against the proteome of each isolate with the pangenome reference proteins as queries using BLAST[68] (e-value cutoff, 1000) and extracted the bitscore from the top hit to each isolate. For each pangenome reference against proteome, we calculated a similarity score based on the bitscore of the top hit (P_{bs}) and the bitscore of aligning the query against itself (M_{bs}) as $genoS = P_{bs}/M_{bs}$. The *genoS* scores of each reference protein over select isolates were then compared with their phenotypic scores (*phenoS*) using cosine similarity, Pearson correlation, and Spearman rank correlation implemented in Python Scipy module. In addition, we performed quantitative trait locus analyses to identify genes whose presence or absence in an isolate can predict its phenotype. For each protein, we partitioned the isolates into 2 groups based on whether it is present in the proteome; the correlation between this gene's presence/absence (genotype) with phenotype is evaluated by Student *t* tests on *phenoS* of the 2 groups defined by the genotypes.

For gentamicin protection assays, we transformed the mean survival index (MSI_{gent}) to a log scale to obtain *phenoS*. For the alternative invasion assay, the prescaled percent invasion scores (PIS) (MSI_{amk}, see later) were used as *phenoS*. Because such correlations are only meaningful if *genoS* and *phenoS* for different isolates have diverse values, we constrained our calculations to proteins satisfying 3 criteria: (1) *genoS* shows at least 3 different values; (2) $\max(genoS) - \min(genoS) > 0.01 \cdot mean(S)$; and (3) $\max(phenoS) - \min(phenoS) > 0.01 \cdot mean(phenoS)$. We were also concerned that isolates from the same patient cannot be treated as independent samples. Thus, in addition to analyzing the entire set of isolates with experimental data, we also assigned the isolates into groups based on phylogeny (Supplementary Tables 4 and 8), averaged both *distS* and *phenoS* scores within each group, and repeated these analyses on the average values. To find proteins showing significant correlation with the phenotypes, we required Pearson and quantitative trait locus Q-values (false discovery rate[69]) <0.05 for correlations using all isolates and the Pearson and quantitative trait locus P value $< .05$ for correlations using averages of phylogenetic groups of isolates.

Antibiotic Resistance Screening and Development of Alternate Invasion Assay

We used AMRFinderPlus[70] developed by NCBI to detect antibiotic resistance in the isolates using both proteins (-p) and genomes (-g), specifying *Escherichia* (-O) as the organism. MICs of antibiotics were determined using MIC Test Strips (Liofilchem). To perform antibiotic protection assays, Caco-2 cells (ATCC) were seeded on collagen-coated plates. To minimize basolateral infection, cells were allowed to differentiate for 3 days before infection. Each monolayer was infected in triplicate at a multiplicity of infection of 10 bacteria per epithelial cell.[71] After a 3-hour incubation period at 37°C, infected monolayers were washed with PBS to remove nonadherent bacteria. Extracellular bacteria were eliminated by 1-hour incubation with Dulbecco's Modified Eagle's Medium containing amikacin at 10 x MIC (240 μg/mL) of the isolates. Monolayers were washed, and monolayer integrity was confirmed microscopically after the final wash. Epithelial cells were lysed in 1% Triton-X 100 in PBS (Sigma). Samples were then serially diluted and plated on Luria-Bertani agar plates to determine the number of

bacterial colony-forming units recovered from the lysed monolayers. Significant epithelial invasion was determined by comparisons with pathogenic *E coli* strain NRG857c (positive control) and nonpathogenic strains DH5α, MP7, and MG1655 (negative controls).

Amikacin invasion scores were performed in triplicate, reporting the colony-forming units/milliliter and PIS for each reference isolate. PIS was defined as the percentage of intracellular bacteria at 1 hour after amikacin treatment relative to that of the original inoculum. Invasion assays were repeated 3 times for each isolate, and the averages were plotted as a group (Prism 10.1.1) according to phylogeny. Amikacin PISs were scaled according to the controls using 2 methods. The first method (premean scaled score) applies such scaling before averaging. The PIS for each assay (PIS_X) was scaled according to the following equation: $(PIS_X - PIS_{neg})/(PIS_{pos} - PIS_{neg})$, where PIS_{pos} is the PIS for the positive control (NRG 857c) and PIS_{neg} is the average PIS for the 4 negative controls (DH5α, MP7, MP13 and MG1655). The 3 scaled scores for each experiment were then averaged, reporting the prescaled mean for each reference isolate. For the second method (postmean scaled score), outliers were removed from PISs grouped by patient and phylogeny using the built-in ROUT method (Q = 1%; 5 outliers) of GraphPad. The mean PIS across replicates after outlier removal was calculated for each isolate, followed by rescaling by the same equation as the first method. Scores from both methods correlate with a Pearson correlation coefficient of 0.97. Both premean and postmean scaled scores identified similar cosegregating genes. We chose to illustrate the premean scaled scores (denoted as MSI_{amk}).

We tested the significance of invasion phenotype by ordinary 1-way analysis of variance in GraphPad, comparing all PISs grouped by patient and phylogeny (without removing outliers). Each group was compared with the mean of the negative controls using Dunnett multiple comparison test, and P values $< .05$ were reported.

Construction of Mutant NRG857c

E coli strain NRG857c deficient in *pduB* was generated using Lambda-Red recombineering as described previously.[72] Briefly, electrocompetent NRG857c cells were first electroporated with pTKRED (Addgene plasmid #41062) and selected aerobically on Luria-Bertani agar plates (Fisher) containing spectinomycin (Goldbio). Next, the electrocompetent NRG857c cells containing pTKRED were electroporated with polymerase chain reaction primers with 40–50 bases of homology (Supplementary Table 12) and pKD13 plasmid as a template. Selection was performed on 35 μg/mL kanamycin (Sigma) and the deletion was verified by polymerase chain reaction.

Supplementary Material

Note: To access the supplementary material accompanying this article, go to the full text version at https://doi.org/10.1016/j.jcmgh.2024.101451.

References

1. Gordon DM, Cowling A. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. Microbiology (Reading) 2003;149:3575–3586.
2. Tenaillon O, Skurnik D, Picard B, et al. The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol 2010;8:207–217.
3. Denamur E, Clermont O, Bonacorsi S, et al. The population genetics of pathogenic *Escherichia coli*. Nat Rev Microbiol 2021;19:37–54.
4. Hammer ND, Skaar EP. Molecular mechanisms of *Staphylococcus aureus* iron acquisition. Annu Rev Microbiol 2011;65:129–147.
5. Tantoso E, Eisenhaber B, Kirsch M, et al. To kill or to be killed: pangenome analysis of *Escherichia coli* strains reveals a tailocin specific for pandemic ST131. BMC Biol 2022;20:146.
6. Touchon M, Hoede C, Tenaillon O, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 2009;5:e1000344.
7. Yang ZK, Luo H, Zhang Y, et al. Pan-genomic analysis provides novel insights into the association of *E.coli* with human host and its minimal genome. Bioinformatics 2019;35:1987–1991.
8. Boudeau J, Glasser AL, Masseret E, et al. Invasive ability of an *Escherichia coli* strain isolated from the ileal mucosa of a patient with Crohn's disease. Infect Immun 1999;67:4499–4509.
9. Glasser AL, Boudeau J, Barnich N, et al. Adherent invasive *Escherichia coli* strains from patients with Crohn's disease survive and replicate within macrophages without inducing host cell death. Infect Immun 2001;69:5529–5537.
10. Palmela C, Chevarin C, Xu Z, et al. Adherent-invasive *Escherichia coli* in inflammatory bowel disease. Gut 2018;67:574–587.
11. Astley DJ, Masters N, Kuballa A, et al. Commonality of adherent-invasive *Escherichia coli* isolated from patients with extraintestinal infections, healthy individuals and the environment. Eur J Clin Microbiol Infect Dis 2021;40:181–192.
12. Martin HM, Campbell BJ, Hart CA, et al. Enhanced *Escherichia coli* adherence and invasion in Crohn's disease and colon cancer. Gastroenterology 2004;127:80–93.
13. Carvalho FA, Barnich N, Sauvanet P, et al. Crohn's disease-associated *Escherichia coli* LF82 aggravates colitis in injured mouse colon via signaling by flagellin. Inflamm Bowel Dis 2008;14:1051–1060.
14. Rakitina DV, Manolov AI, Kanygina AV, et al. Genome analysis of *E coli* isolated from Crohn's disease patients. BMC Genomics 2017;18:544.
15. Darfeuille-Michaud A, Boudeau J, Bulois P, et al. High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in Crohn's disease. Gastroenterology 2004;127:412–421.
16. Zhang Y, Rowehl L, Krumsiek JM, et al. Identification of candidate adherent-invasive *E coli* signature transcripts by genomic/transcriptomic analysis. PLoS One 2015;10:e0130902.

17. Elhenawy W, Hordienko S, Gould S, et al. High-throughput fitness screening and transcriptomics identify a role for a type IV secretion system in the pathogenesis of Crohn's disease-associated *Escherichia coli*. Nat Commun 2021;12:2032.
18. Mayorgas A, Dotti I, Martinez-Picola M, et al. A novel strategy to study the invasive capability of adherent-invasive *Escherichia coli* by using human primary organoid-derived epithelial monolayers. Front Immunol 2021;12:646906.
19. Martinez-Medina M, Strozzi F, Castillo BRD, et al. Antimicrobial resistance profiles of adherent invasive *Escherichia coli* show increased resistance to beta-lactams. Antibiotics (Basel) 2020;9:251.
20. Siniagina MN, Markelova MI, Boulygina EA, et al. Diversity and adaptations of *Escherichia coli* strains: exploring the intestinal community in Crohn's disease patients and healthy individuals. Microorganisms 2021;9:1299.
21. Sasaki M, Sitaraman SV, Babbin BA, et al. Invasive *Escherichia coli* are a feature of Crohn's disease. Lab Invest 2007;87:1042–1054.
22. Parks DH, Imelfort M, Skennerton CT, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;25:1043–1055.
23. Tyakht AV, Manolov AI, Kanygina AV, et al. Genetic diversity of *Escherichia coli* in gut microbiota of patients with Crohn's disease discovered using metagenomic and genomic analyses. BMC Genomics 2018;19:968.
24. Rodriguez RL, Conrad RE, Viver T, et al. An ANI gap within bacterial species that advances the definitions of intra-species units. mBio 2024;15:e0269623.
25. Clermont O, Dixit OVA, Vangchhia B, et al. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. Environ Microbiol 2019;21:3107–3117.
26. Benlabidi S, Raddaoui A, Lengliz S, et al. Occurrence of high-risk clonal lineages ST58, ST69, ST224, and ST410 among extended-spectrum beta-lactamase-producing *Escherichia coli* isolated from healthy free-range chickens (*Gallus Lenglen domesticus*) in a rural region in Tunisia. Genes (Basel) 2023;14:875.
27. Allmansberger R, Brau B, Piepersberg W. Genes for gentamicin-(3)-N-acetyl-transferases III and IV. II. Nucleotide sequences of three AAC(3)-III genes and evolutionary aspects. Mol Gen Genet 1985;198:514–520.
28. Collis CM, Hall RM. Expression of antibiotic resistance genes in the integrated cassettes of integrons. Antimicrob Agents Chemother 1995;39:155–162.
29. Li W, Ma J, Sun X, et al. Antimicrobial resistance and molecular characterization of gene cassettes from class 1 integrons in *Escherichia coli* strains. Microb Drug Resist 2022;28:413–418.
30. Alcock BP, Huynh W, Chalil R, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res 2023;51:D690–D699.
31. Varani A, He S, Siguier P, et al. The IS6 family, a clinically important group of insertion sequences including IS26. Mob DNA 2021;12:11.
32. Nash JH, Villegas A, Kropinski AM, et al. Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E coli* pathotypes. BMC Genomics 2010;11:667.
33. Manichanh C, Borruel N, Casellas F, et al. The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol 2012;9:599–608.
34. Carbonero F, Benefiel AC, Alizadeh-Ghamsari AH, et al. Microbial pathways in colonic sulfur metabolism and links with health and disease. Front Physiol 2012;3:448.
35. Wallace JL, Vong L, McKnight W, et al. Endogenous and exogenous hydrogen sulfide promotes resolution of colitis in rats. Gastroenterology 2009;137:569–578.e561.
36. Ijssennagger N, Belzer C, Hooiveld GJ, et al. Gut microbiota facilitates dietary heme-induced epithelial hyperproliferation by opening the mucus barrier in colon. Proc Natl Acad Sci U S A 2015;112:10038–10043.
37. Dogan B, Suzuki H, Herlekar D, et al. Inflammation-associated adherent-invasive *Escherichia coli* are enriched in pathways for use of propanediol and iron and M-cell translocation. Inflamm Bowel Dis 2014;20:1919–1932.
38. Viladomiu M, Metz ML, Lima SF, et al. Adherent-invasive *E coli* metabolism of propanediol in Crohn's disease regulates phagocytes to drive intestinal inflammation. Cell Host Microbe 2021;29:607–619.e608.
39. Yeates TO, Jorda J, Bobik TA. The shells of BMC-type microcompartment organelles in bacteria. J Mol Microbiol Biotechnol 2013;23:290–299.
40. Jakobson CM, Tullman-Ercek D, Slininger MF, et al. A systems-level model reveals that 1,2-propanediol utilization microcompartments enhance pathway flux through intermediate sequestration. PLoS Comput Biol 2017;13:e1005525.
41. Faber F, Thiennimitr P, Spiga L, et al. Respiration of microbiota-derived 1,2-propanediol drives *Salmonella* expansion during colitis. PLoS Pathog 2017;13:e1006129.
42. Kennedy NW, Mills CE, Abrahamson CH, et al. Linking the *Salmonella enterica* 1,2-propanediol utilization bacterial microcompartment shell to the enzymatic core via the shell protein PduB. J Bacteriol 2022;204:e0057621.
43. Hanson BT, Dimitri Kits K, Loffler J, et al. Sulfoquinovose is a select nutrient of prominent bacteria and a source of hydrogen sulfide in the human gut. ISME J 2021;15:2779–2791.
44. Hoffman JA, Badger JL, Zhang Y, et al. *Escherichia coli* K1 aslA contributes to invasion of brain microvascular endothelial cells in vitro and in vivo. Infect Immun 2000;68:5062–5067.
45. Luis AS, Jin C, Pereira GV, et al. A single sulfatase is required to access colonic mucin by a gut bacterium. Nature 2021;598:332–337.
46. Fujita K, Takashi Y, Obuchi E, et al. Characterization of a novel beta-L-arabinofuranosidase in *Bifidobacterium longum*: functional elucidation of a DUF1680 protein family member. J Biol Chem 2014;289:5240–5249.
47. Epidemiological and clinical features of Spanish patients with Crohn's disease. Spanish Epidemiological and Economic Study Group on Crohn's disease. Eur J Gastroenterol Hepatol 1999;11:1121–1127.

48. Colombel JF, Lemann M, Cassagnou M, et al. A controlled trial comparing ciprofloxacin with mesalazine for the treatment of active Crohn's disease. Groupe d'Etudes Therapeutiques des Affections Inflammatoires Digestives (GETAID). Am J Gastroenterol 1999; 94:674–678.
49. Nguyen GC, Kaplan GG, Harris ML, et al. A national survey of the prevalence and impact of *Clostridium difficile* infection among hospitalized inflammatory bowel disease patients. Am J Gastroenterol 2008; 103:1443–1450.
50. Shah A, Morrison M, Burger D, et al. Systematic review with meta-analysis: the prevalence of small intestinal bacterial overgrowth in inflammatory bowel disease. Aliment Pharmacol Ther 2019;49:624–635.
51. Hellers G, Bergstrand O, Ewerth S, et al. Occurrence and outcome after primary treatment of anal fistulae in Crohn's disease. Gut 1980;21:525–527.
52. Keighley MR, Eastwood D, Ambrose NS, et al. Incidence and microbiology of abdominal and pelvic abscess in Crohn's disease. Gastroenterology 1982; 83:1271–1275.
53. Barnich N, Carvalho FA, Glasser AL, et al. CEACAM6 acts as a receptor for adherent-invasive *E coli*, supporting ileal mucosa colonization in Crohn disease. J Clin Invest 2007;117:1566–1574.
54. Cieza RJ, Hu J, Ross BN, et al. The IbeA invasin of adherent-invasive *Escherichia coli* mediates interaction with intestinal epithelia and macrophages. Infect Immun 2015;83:1904–1918.
55. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30:2114–2120.
56. Wick RR, Judd LM, Gorrie CL, et al. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13: e1005595.
57. Tatusova T, DiCuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 2016;44:6614–6624.
58. Sanderson T, Bileschi ML, Belanger D, et al. ProteInfer, deep neural networks for protein functional inference. Elife 2023;12:e80942.
59. Gane A, Bileschi ML, Dohan D, et al. Model-based Natural Language Protein Annotation. Available at: https://storage.googleapis.com/brain-genomics-public/research/proteins/protnlm/uniprot_2022_04/protnlm_preprint_draft.pdf. Accessed February, 2025.
60. Avram O, Rapoport D, Portugez S, et al. M1CR0B1AL1Z3R-a user-friendly web server for the analysis of large-scale microbial genomics data. Nucleic Acids Res 2019;47:W88–W92.
61. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021;49:W293–W296.
62. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402.
63. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 2018;3:124.
64. Beghain J, Bridier-Nahmias A, Le Nagard H, et al. ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. Microb Genom 2018;4:e000192.
65. Jain C, Rodriguez RL, Phillippy AM, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018;9:5114.
66. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 2019; 20:238.
67. Katoh K, Misawa K, Kuma K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–3066.
68. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009; 10:421.
69. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57:289–300.
70. Feldgarden M, Brover V, Gonzalez-Escalona N, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 2021;11:12728.
71. Chimalapati S, Lafrance AE, Chen L, et al. *Vibrio parahaemolyticus*: basic techniques for growth, genetic manipulation, and analysis of virulence factors. Curr Protoc Microbiol 2020;59:e131.
72. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A. 2000;97: 6640–6645.

Received September 27, 2024. Accepted December 20, 2024.

Correspondence
Address correspondence to: Josephine Ni, MD, Division of Digestive and Liver Diseases, University of Texas, Southwestern, 5323 Harry Hines Boulevard K5.206, Dallas, Texas 75390-9151. e-mail: josephine.ni@utsouthwestern.edu; or Qian Cong, PhD, 5901 Forest Park Road, NB 10.110A, Dallas, Texas 75235-8591. e-mail: qian.cong@utsouthwestern.edu.

Acknowledgments
Jungyeon Kim, Jing Zhang, Lisa Kinch, and Jinhui Shen contributed equally.

CRediT Authorship Contributions
Jungyeon Kim (Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Methodology: Equal; Writing – review & editing: Supporting)
Jing Zhang (Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Methodology: Equal; Writing – review & editing: Supporting)
Lisa Kinch (Conceptualization: Equal; Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Visualization: Equal; Writing – original draft: Equal; Writing – review & editing: Equal)
Jinhui Shen (Data curation: Supporting; Investigation: Equal; Writing – review & editing: Supporting)
Sydney Field (Methodology: Supporting)
Shahanshah Khan (Investigation: Supporting)
Jan-Michael Klapproth (Project administration: Equal; Resources: Equal)
Kevin J. Forsberg (Investigation: Supporting; Writing – review & editing: Supporting)
Tamia Harris-Tryon (Investigation: Supporting; Writing – review & editing: Supporting)

Kim Orth (Funding acquisition: Lead; Investigation: Lead; Writing – review & editing: Equal)

Qian Cong (Conceptualization: Equal; Data curation: Equal; Formal analysis: Equal; Funding acquisition: Equal; Methodology: Equal; Writing – original draft: Equal; Writing – review & editing: Equal)

Josephine Ni (Conceptualization: Equal; Funding acquisition: Equal; Investigation: Lead; Methodology: Equal; Resources: Equal; Supervision: Equal; Writing – original draft: Equal; Writing – review & editing: Equal)

Conflicts of interest

The authors disclose no conflicts.

Funding

Supported by the Howard Hughes Medical Institute Emerging Pathogens Initiative (K.O., Q.C., K.J.F., T.A.H., J.N.), Welch Foundation grant I-1561 (K.O.), Once Upon a Time…Foundation (K.O.), NIH grant R35 GM134945 (K.O.), Welch Foundation grant I-2095-20220331 (Q.C.), NIAID grant 1K99AI180984-01A1 (J.Z.), NIH grant NIAMS K08AR076459 (T.A.H.), Burroughs Wellcome Fund- Pathogenesis of Infectious Diseases-1022777 (T.A.H.), Endowed Scholars Program at the University of Texas Southwestern Medical Center (Q.C., K.J.F.), Searle Scholars award (K.J.F.), NIH grant DP2-AI154402 (K.J.F.), NIH grant NIDDK K08-DK-123316 (J.N.), and Burroughs Wellcome Fund Career Awards for Medical Scientists - 1020904 (J.N.).