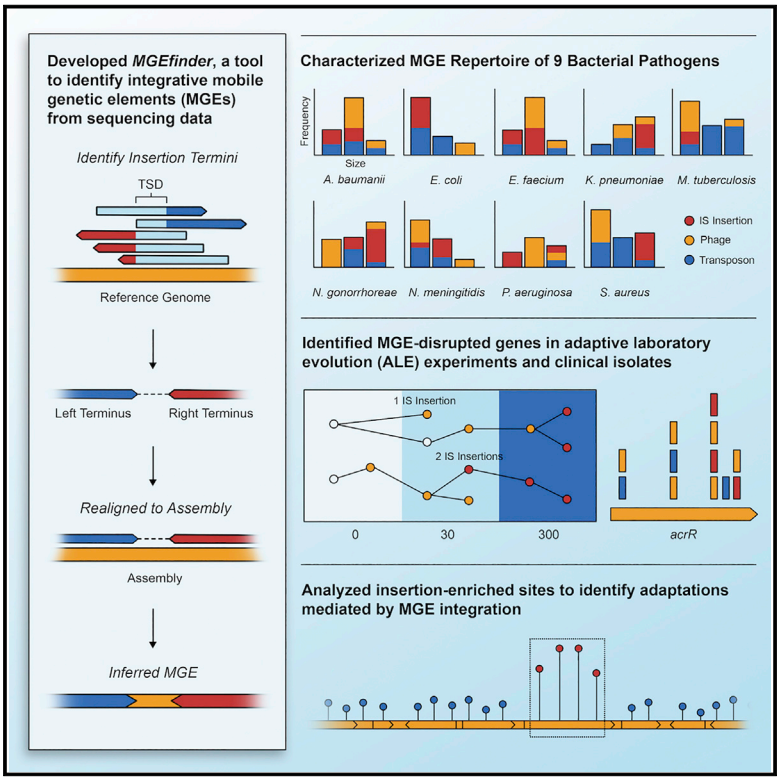
Resource

Cell Host & Microbe

A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation

Graphical Abstract



Highlights

- Introduces MGEfinder, a bioinformatic toolbox to detect MGE integrations
- Describes the MGE repertoire, from small repeats to prophages, of 9 pathogens
- MGE insertions affect antibiotic resistance, virulence, pathogenicity across species

Authors

Matthew G. Durrant, Michelle M. Li, Benjamin A. Siranosian, Stephen B. Montgomery, Ami S. Bhatt

Correspondence

asbhatt@stanford.edu

In Brief

Microbes evolve through DNA mutations and by the transfer of mobile genetic elements (MGEs). In this study, Durrant et al. describe MGEfinder, a tool that enables MGE localization and characterization. Using MGEfinder, they analyze 9 bacterial pathogens, providing insights into how MGEs might affect antibiotic resistance, virulence, and pathogenicity.

Durrant et al., 2020, Cell Host & Microbe *27*, 1–14
January 8, 2020
https://doi.org/10.1016/j.chom.2019.10.022

CellPress

Please cite this article in press as: Durrant et al., A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation, Cell Host & Microbe (2019), https://doi.org/10.1016/j.chom.2019.10.022

Cell Host & Microbe
Resource

A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation

Matthew G. Durrant,[1] Michelle M. Li,[1] Benjamin A. Siranosian,[1] Stephen B. Montgomery,[1,3] and Ami S. Bhatt[1,2,4,*]
[1]Department of Genetics, Stanford University, Stanford, CA 94305, USA
[2]Department of Medicine (Hematology, Blood and Marrow Transplantation) Stanford University, Stanford, CA 94305, USA
[3]Department of Pathology, Stanford University, Stanford, CA 94305, USA
[4]Lead Contact
*Correspondence: asbhatt@stanford.edu
https://doi.org/10.1016/j.chom.2019.10.022

SUMMARY

Mobile genetic elements (MGEs) contribute to bacterial adaptation and evolution; however, high-throughput, unbiased MGE detection remains challenging. We describe MGEfinder, a bioinformatic toolbox that identifies integrative MGEs and their insertion sites by using short-read sequencing data. MGEfinder identifies the genomic site of each MGE insertion and infers the identity of the inserted sequence. We apply MGEfinder to 12,374 sequenced isolates of 9 prevalent bacterial pathogens, including *Mycobacterium tuberculosis*, *Staphylococcus aureus*, and *Escherichia coli*, and identify thousands of MGEs, including candidate insertion sequences, conjugative transposons, and prophage elements. The MGE repertoire and insertion rates vary across species, and integration sites often cluster near genes related to antibiotic resistance, virulence, and pathogenicity. MGE insertions likely contribute to antibiotic resistance in laboratory experiments and clinical isolates. Additionally, we identified thousands of mobility genes, a subset of which have unknown function opening avenues for exploration. Future application of MGEfinder to commensal bacteria will further illuminate bacterial adaptation and evolution.

INTRODUCTION

Successful human pathogens can acquire adaptive phenotypes, such as antibiotic resistance, through single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), inversions, duplications, and the movement of mobile genetic elements (MGEs). Prokaryotic MGEs, such as insertion sequence (IS) elements, transposons, integrons, plasmids, and bacteriophages (Stokes and Gillings, 2011; Rankin et al., 2011) can mobilize and integrate in a site-specific or non-specific manner throughout the host genome.

MGEs can range in size from small elements, such as miniature inverted repeat transposable elements (MITEs), which are on the order of tens of base pairs (bps) in length, to prophages and transposons, which can be tens or hundreds of kilobase pairs (kbps) in length. Among the most well-studied integrative mobile elements are IS elements, relatively simple MGEs coding for only the transposase necessary for their transposition (Mahillon and Chandler, 1998). IS elements can transpose into genes, resulting in insertional mutagenesis and loss of function of the gene (Lerat and Ochman, 2004); alternatively, they can transpose into gene regulatory elements and influence expression of neighboring genes.

Several tools exist to identify MGE insertions. For example, whole-genome alignment tools, such as Mauve, can identify large insertions, also known as "genomic islands" (Darling et al., 2010; Bertelli et al., 2017). Other tools focus on IS elements in particular (Barrick et al., 2014; Lerat, 2010; Xie and Tang, 2017; Treepong et al., 2018; Jiang et al., 2015; Adams et al., 2016; Biswas et al., 2015; Hawkey et al., 2015) because their repetitive nature makes it unlikely for them to properly assemble in a draft assembly (see Table S3). Although these tools are useful, approaches based on whole-genome alignment fail to identify insertions of repetitive elements in draft genomes, and existing tools for identifying repetitive element insertions usually depend on homology with known MGEs or require well-annotated reference genomes. To our knowledge, few tools exist that can identify a wide range of mobile element insertions, both repetitive and non-repetitive, without relying on an external database of known elements.

Here, we develop and validate a computational toolbox that identifies complete MGEs and their insertion sites with respect to a reference genome from short-read sequencing data. Our workflow can reliably detect elements from as small as 70 base pairs (bps) in length to as large as hundreds of kbps. We use this approach to analyze 12,374 sequenced isolates of 9 pathogenic bacterial species and find large differences in the overall MGE repertoire and the rate of MGE insertion between species. By analyzing the location and genetic content of these elements, we infer their potential role in clinically relevant biological pathways, such as antibiotic resistance.

RESULTS

A Flexible Approach to Identify and Genotype MGE Insertions

We sought to develop a tool that would overcome the limitations of whole-genome alignment and homology-based approaches

to identify both repetitive and non-repetitive insertions ranging from 70 bps to 200 kbp in length, as well as their genomic position with respect to a reference genome. Our approach can identify such insertions without the need for a complete genome assembly or a database of known elements. Using this bioinformatic toolbox, which we call MGEfinder, we analyzed 12,374 publicly available sequenced isolates of nine bacterial pathogens: *Acinetobacter baumannii*, *Enterococcus faecium*, *Escherichia coli*, *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus*.

First, MGEfinder aligns short reads from isolates of a given species to a reference genome of that species (Figure 1A). The choice of reference genome is important because the sequence similarity of the reference genome and the isolate affect the sensitivity and precision of insertion detection (Figures S2A and S2B). In our analysis, we required isolates to share at least 98.5% nucleotide identity with the reference genome. Second, candidate insertion sites were identified by searching the alignments for reads that have clipped (unaligned) ends at the same genomic site. A read with a clipped end indicates the site where the inserted element begins with respect to the reference genome used, and the clipped ends are then used to build a consensus sequence of the terminal end (or terminus) of the insertion. Third, these high-quality terminal ends were paired with nearby oppositely oriented terminal ends to identify candidate insertions. Fourth, the genome of each bacterial isolate was then assembled to produce a draft assembly, candidate insertion termini were aligned back to the draft assembly, and the full inserted sequence was inferred from the alignments. Finally, a database of all inserted elements identified across all analyzed isolates was dynamically constructed, and we performed a final sequence inference step of all terminal end pairs by aligning to this accumulated database. By combining several inference approaches (Figures 1B and S1A–F), MGEfinder increases the overall sensitivity and confidence in the accuracy of the inferred sequence.

We compared MGEfinder with panISa, a tool that uses a similar approach to identify insertion junctions and build consensus sequences of insertion termini (Treepong et al., 2018), and progressiveMauve (Darling et al., 2010), a whole-genome alignment tool commonly used to identify genomic islands (Bertelli et al., 2017). We found that MGEfinder is more sensitive and has a lower overall false-positive rate than panISa (Figures S2E and S2H). We found that, compared with progressiveMauve, MGEfinder is much more sensitive to the detection of repetitive sequences inserted into the genome when using draft assemblies, which are much less contiguous than complete reference genomes, and sensitivity is further improved when using a dynamically constructed database of MGEs found across isolates (Figure S2I). Additionally, MGEfinder is better at identifying the precise boundaries of the inserted element than progressiveMauve (Figure S2J).

Characterizing the Integrative MGE Repertoire of Nine Bacterial Pathogens

We ran MGEfinder on 1,848 *E. faecium*, 1,646 *A. baumannii*, 1,570 *S. aureus*, 1,378 *M. tuberculosis*, 1,361 *K. pneumoniae*, 1,348 *P. aeruginosa*, 1,306 *N. meningitidis*, 1,026 *E. coli*, and 891 *N. gonorrhoeae* isolates. We limited our analysis to MGEs between 70 bp to 200 kbp in size and that generated a target site duplication of 20 bp or less. We identified 5,019 unique element clusters (clustered by 90% identity across 85% of each sequence) across all species analyzed (Figure 1C; Table S5). We observed significant differences in the distribution and composition of each species' putative MGE repertoire; for example, *N. gonorrhoeae* and *M. tuberculosis* were noticeably depleted of elements above 10 kbps in length (Figure 1C). We classified inserted elements into 11 categories, including IS elements (8.9% of all identified elements) (see Figure S3 for IS families), plasmids (0.4%), intact prophages (15.8%), questionable and/or incomplete prophages (6.1%), elements with a transposase and other predicted coding sequence (CDS) (20.4%), elements containing a protein with a predicted Group II intron domain (1.1%), elements containing a predicted serine or tyrosine recombinase (10.4%), elements with at least one CDS and terminal inverted repeats (TIRs) (3.4%), elements without any predicted CDS with TIRs (2.4%), elements with at least one predicted CDS but no TIR or transposase (19.1%), and elements without any of the previous annotations ("No CDS"; 12%) (see Figure S1E for schematic representation of categories). Additionally, 178 (3.5%) of these elements contain predicted conjugation systems, and 100 of these elements also contained transposases, indicating that additional classes of elements such as conjugative transposons exist in this collection.

To assess "transposability," we next counted the number of times we observed a given sequence element inserted at different loci across the reference genome (Figure 2A). We found that *E. coli*, *P. aeruginosa*, and *A. baumannii* had the highest number of highly transposable MGEs (elements found at > 10 genomic positions): 54 such elements in *E. coli*, 44 in *A. baumannii*, and 43 in *P. aeruginosa*. *N. gonorrhoeae* was on the opposite extreme: only one element was detected at > 10 positions, a correia repeat-enclosed element (CREE), which is a small non-autonomous MGE described previously (Liu et al., 2002). Among all MGEs, we define a transposable element (TE) as an element cluster found at more than three positions in the reference genome in sum across all analyzed isolates. In total, 516 elements were classified as TEs (10.3%). These differences across species suggest significant variability in MGE diversity and overall activity.

Different classes of putative MGEs vary considerably in their levels of transposability (Figure 2B). As expected, predicted IS elements are typically among the most transposable within each species, comprising 65.3% of all elements found at more than 10 loci. In *E. coli* and *P. aeruginosa*, several phage elements are highly transposable, and in *N. gonorrhoeae*, non-autonomous MITE elements are highly transposable. Less transposable elements are often categorized as "Contains CDS" and "No CDS," supporting the possibility that these types of elements are site-specific, less active in the population overall, or potential false positives.

We define a "unique insertion" as an insertion event of a single element at a single insertion site. In each of the species studied here, the elements responsible for the most unique insertions are IS elements, with the exception of the non-autonomous CREE element in *N. gonorrhoeae* and *N. meningitidis* (Figure S5C). As more isolates are analyzed, more unique insertions attributable to TEs are identified, but the rate of increase varies from species

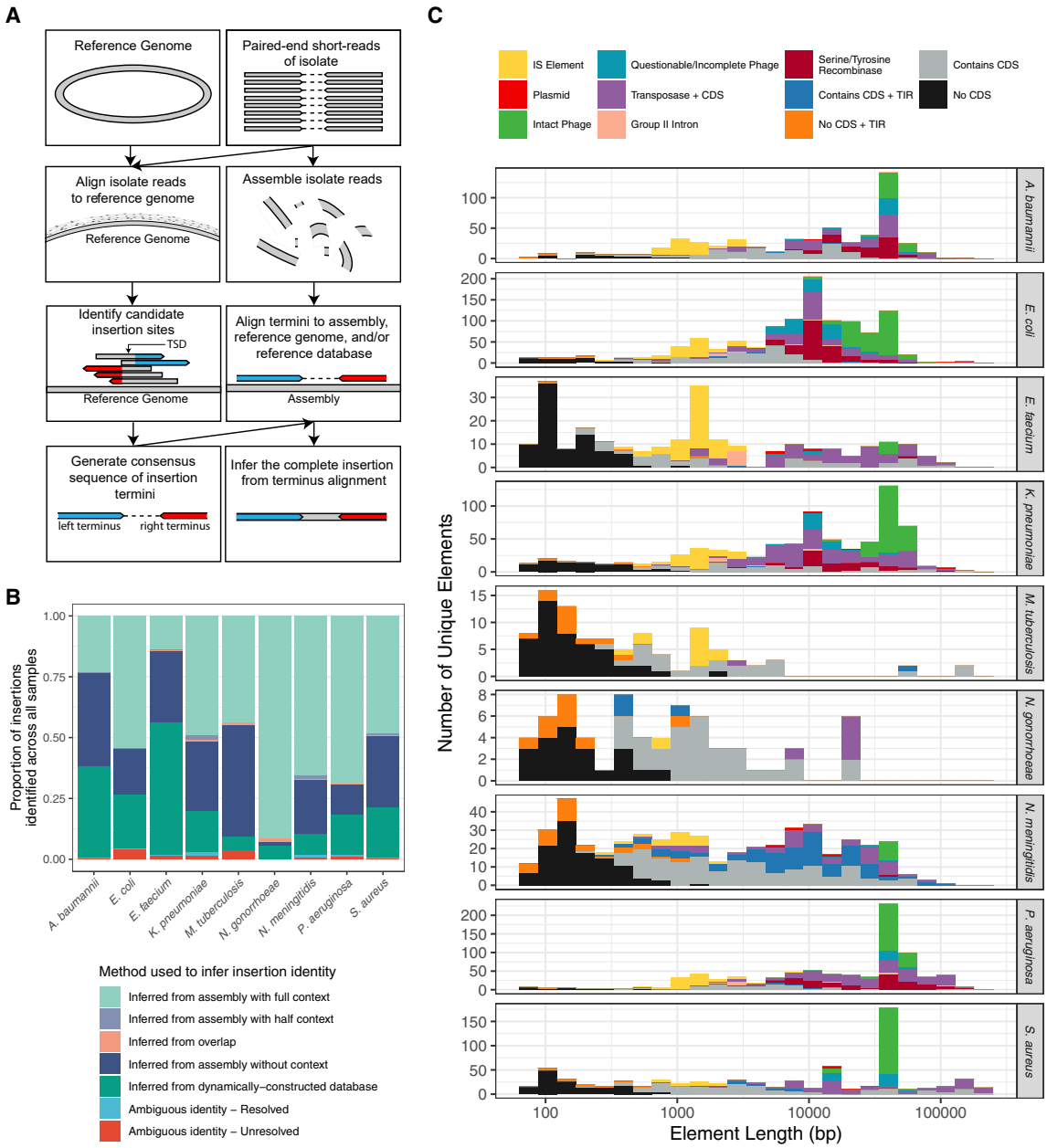


Figure 1. An Approach to Identify a Variety of Integrative MGEs from Short-Read Sequencing Data

(A) MGEfinder workflow schematic. See main text and STAR Methods for details of each step. See Figure S2 for MGEfinder performance metrics. Abbreviation is as follows: TSD, target site duplication.

(B) The proportion of elements identified by each inference method for the downloaded isolates of nine bacterial pathogens. See STAR Methods and Figure S1 for description of each inference technique and how ambiguous insertions were resolved.

(C) An analysis of the types of elements identified in the MGEfinder workflow when applied to nine bacterial pathogens; element length is on the x axis (log-scale), and element count is on the y axis. A "unique element" refers to a unique cluster of elements (see STAR Methods for details). See Figure S1E for a schematic representation of each type of element. See Figure S3 for identified IS families. Abbreviations are as follows: IS, insertion sequence; CDS, coding sequence; TIR, terminal inverted repeat.

to species (Figure 2C). *E. coli*, with 5,883 unique TE insertions identified after analyzing 1,026 isolates, is on the high end of the distribution, suggesting that MGE transposition is relatively common in this species. In contrast, only 28 unique TE insertions were detected in *N. gonorrhoeae* across all 891 analyzed isolates, suggesting that TE insertions might not be a common source of mutation for this species. We also generated accumulation curves for the number of unique sequence elements within each MGE category, and in many cases across species, certain MGE categories appear to be nearly saturated, indicating that

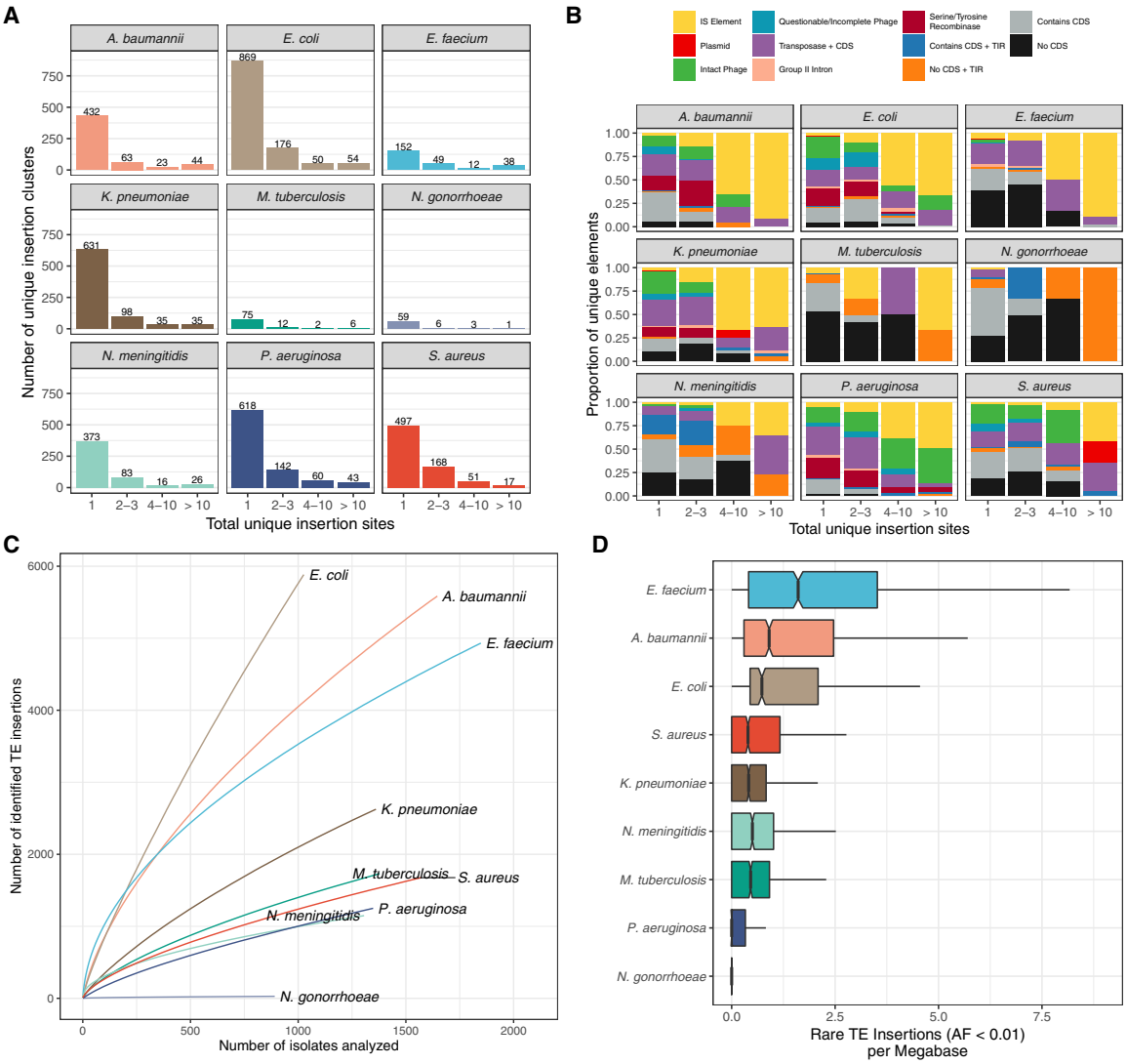


Figure 2. Bacterial Species Vary Considerably in Their Overall MGE Repertoire and Rate of Insertion

(A) The number of unique sequence elements identified, binned by species and total unique insertion sites per element. "Total unique insertion sites" refers to the number of unique sites where members of a given element cluster can be found across all isolates for each species.

(B) Types of MGEs at different levels of transposability. The bins along the x axis are the same as in (A). Element categories are indicated by the colors in the legend.

(C) An accumulation curve of the number of new TE insertions identified as additional isolates are analyzed. See also Figure S4.

(D) Notched boxplots of the number of rare TE insertions detected across all samples for each species. A rare TE insertion is defined as a TE insertion identified in < 1% of all samples. Rare insertions are adjusted by the number of genomic sites in the sequenced isolates with non-zero coverage, and then multiplied by 1 megabase. Notches indicate 1.58 × interquartile range / sqrt(n), a rough 95% confidence interval for comparing medians (Mcgill et al., 1978). Outliers are excluded from this figure. See also Figure S5.

further isolate collection would not substantially increase the number of detected integrative MGEs (Figure S4). For example, the number of detected *A. baumannii* intact phage elements appears to plateau at ~50 elements, and the slope of increase is quite small near the end of the curve. This differs considerably from the number of intact phage elements detected in *E. coli*, which maintains a high slope of increase even after analyzing 1,000 isolates. These accumulation curves do not, however, account for differences in genome size across species, which likely partially explains differences in the rate of insertion.

To adjust for genome size, we analyzed the number of rare TE insertions (those with allele frequency < 0.01 in the population) per megabase for each isolate (Figure 2D). By analyzing only rare TE insertions, our goal is to approximate the relative rate of insertion across isolates and species. *E. faecium* has the highest number of rare TE insertions per megabase and has a mean of 2.44 (95% confidence interval [CI], 2.32–2.56) across all analyzed samples, whereas *N. gonorrhoeae* and *P. aeruginosa* have means of 0.02 (95% CI, 0.01–0.03) and 0.26 (95% CI, 0.23–0.29), respectively. The number of rare TE insertions is

higher in intergenic regions than in predicted coding sequences (Figure S5A). Realizing that the differences in mutation rate here could simply reflect differences in genetic diversity or sequencing quality across analyzed isolates, we further interrogated this possibility by using ANOVA. The average pairwise average nucleotide identity (ANI) of each isolate to all other samples explains 3.1% of the variance in the number of detected rare TE insertions ($p < 2e{-}16$), whereas species explains 16.3% ($p < 2e{-}16$), and sequencing library fragment length explains 0.8% ($p = 9.8e{-}11$). *E. faecium* has the highest number of rare TE insertions despite being the third most genetically homogeneous species (Figure S5B). This suggests that genetic diversity and sequencing artifacts influence our estimates, but characteristics intrinsic to the species itself have a much larger effect on the number of TE insertions. Here, we see that the number of TE insertions continues to increase as more isolates are analyzed and estimates for the relative rate of insertion vary considerably across species even when correcting for genetic diversity. These estimates can inform future efforts to sequence and analyze additional pathogenic isolates belonging to these species.

Characterizing MGE Passenger Proteins and Prophages

In our previous analyses, we have focused on MGEs primarily at the nucleotide level, but we can also characterize MGEs by using a gene-centric approach. Using predicted CDS regions, we clustered all protein sequences across species at 50% identity. This resulted in a total of 27,718 unique gene clusters across all 9 species. We then filtered these gene clusters to only those appearing at over 10 unique insertion sites across all species, which resulted in a list of 1,239 high-confidence "mobility gene" clusters (Figure 3A). Predicted transposases make up 18.7% of these mobility genes and 37.3% of genes appearing at over 100 genomic locations. Phage-associated and conjugation system proteins make up 43% and 4.4% of all mobility genes, respectively. A further 20.2% of mobility genes have no annotated function when using Prokka or common mobility gene hidden Markov models (HMMs), including 5 such genes that appear at over 100 genomic locations in *S. aureus*. Further investigating mobility genes of unknown function might lead to the discovery of new genes important for DNA mobility and horizontal gene transfer.

This curated list of mobility genes can be further used to characterize elements in greater detail. Of the 5,019 elements found across all species, 2,584 (51.4%) contain at least one mobility gene, and an additional 1,440 (28.6%) share high nucleotide BLAST similarity (e value < 1e−4) with an element that contains a mobility gene, meaning 80.2% of identified elements bear some genetic relationship to this curated set of mobility genes. In total, 61.5% of poorly characterized elements (Contains CDS + TIR, No CDS + TIR, Contains CDS, and No CDS elements) contain mobility genes or are homologous to other elements containing mobility genes; up to 83.6% of poorly characterized in *S. aureus* elements met these criteria (Figure 3B). In the case of *M. tuberculosis*, very few elements met these criteria, and this might be because *M. tuberculosis* is the only member of the Actinobacteria phylum that we analyzed. Including more species from this phylum in the future will likely improve our ability to characterize the mobile elements belonging to this species.

Several elements contain known antibiotic resistance genes as annotated by ResFinder (Zankari et al., 2012) (Table S1). We detect resistance to "drugs of last resort" encoded on MGEs. For example, the resistance gene *vanHBX*, which confers resistance to vancomycin, is found in four different MGEs in *E. faecium*; in another example, carbapenem-hydrolyzing beta-lactamases *blaKPC-2* and *blaKPC-3* are found in two different MGEs in *K. pneumoniae*. In addition to antibiotic resistance genes, MGEs contain other genetic "cargo" that are thought to benefit the bacterial host, such as streptomycin 3′-adenylyltransferases, colicin V secretion proteins, bicyclomycin resistance proteins, genes involved in copper resistance (cation efflux pumps and copper resistance proteins *copA*, *copB*, *copD*, and *copR*) (Hamlett et al., 1992), and those involved in restriction modification systems (*hsdR* and *hsdM*) (Murray et al., 1982) (Figure 3C and 3D). In summary, our workflow identified MGEs containing passenger proteins of known and unknown function, shedding light on the phenotypic changes that likely accompany these insertion events.

Although a subset of these identified elements is likely transferred between hosts through conjugative mechanisms as naked DNA, other elements might be encoded by and transferred through phages. Using PHASTER, a previously described tool that identifies candidate prophages on the basis of input DNA sequence (Arndt et al., 2016), we identified 792 "Intact Phage" sequence clusters. Many identified phage element clusters correspond to a single record in the PHASTER database. For example, in *E. coli* we find that 55 site-specific phage elements are similar to Enterobacteria phage Fels-2 (NC_010463.1), a 33 kbp phage first identified in *Salmonella enterica* serovar Typhimurium. The average pairwise nucleotide identity between these Fels-2-like phage clusters is 96.4%, and an average of 82.9% of bases are aligned, suggesting the Fels-2-like phages that integrate at this site are genetically diverse. Analyzing the protein sequences found in these phages, we identify 19 core proteins (present in at least 90% of these phages), and 112 accessory proteins (present in less than 90% of phages). One of these accessory proteins is beta-lactamase CTX-M-97, indicating that this particular antibiotic resistance gene might have entered the cell by way of this phage and that these phages are a valuable source of novel genetic material. Carefully characterizing these integrative phage elements by using the MGEfinder workflow can help to understand the role phages play in horizontal gene transfer across different species.

Analysis of Insertion-Enriched Sites Reveals Their Role in Microbial Adaptation

The approach we have taken allows us to not only identify MGEs, but it also identifies their sites of insertion with respect to the reference genome. This allows us to investigate the role MGEs play in genomic evolution, identifying sites that are enriched for unique insertions (insertion-enriched sites) that might indicate functionally important genes and pathways. We performed an analysis of TE-insertion-enriched sites by using all unique TE insertions in each species analyzed (Figure 4A). With the exception of *N. gonorrhoeae*, which had too few unique TE insertions to analyze, we identified several TE-insertion-enriched sites for all species, and 654 were identified in total (false

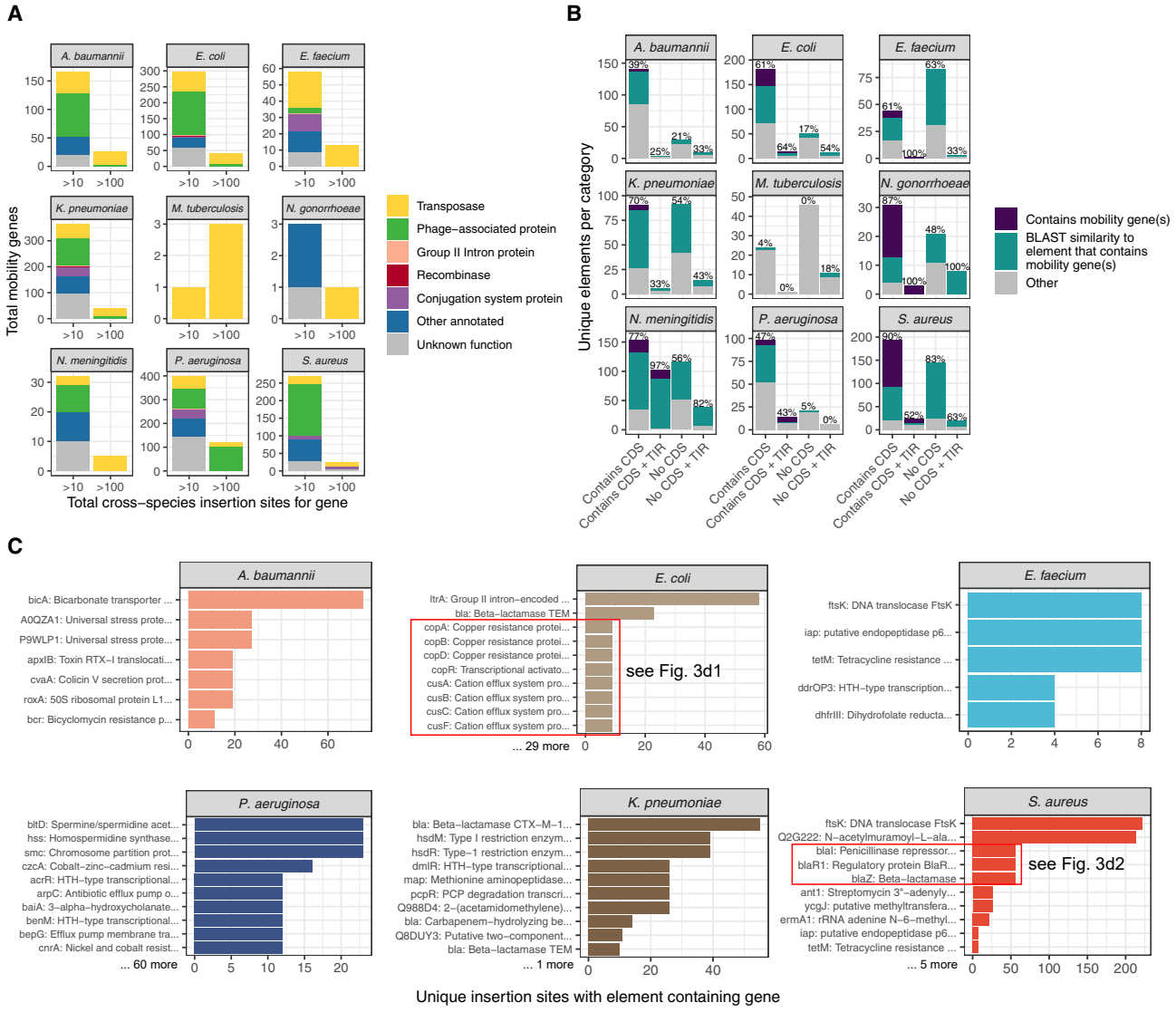


Figure 3. Many Identified MGEs Include Passenger Genes, Some of Which Are Largely Uncharacterized

(A) Mobility genes identified when analyzing coding sequences found in mobile elements across species. Mobility genes are grouped into those found at >10 or >100 unique insertion sites. See STAR Methods for description of each term in the legend.

(B) Bar plots depict the number of elements falling in the four lowest-confidence element categories: "Contains CDS," "Contains CDS + TIR," "No CDS," and "No CDS + TIR." In each category, the proportion of elements with a predicted mobility gene (purple), no predicted mobility gene but high BLAST similarity to another element with a mobility gene (blue), and all other elements (gray) are indicated. Percentages on the top of the bars indicate the percentage of elements in each bin falling in the purple or blue categories.

(C) The number of unique insertion sites identified where a mobile element containing each labeled passenger protein was found. Shown are only proteins that occur on MGEs that contain at least one predicted transposase. Ties in the number of unique sites generally indicates the passenger proteins are on the same element.

(D) Two examples of mobile elements carrying passenger proteins: (1) a copper resistance mobile element found in *E. coli*; (2) a mobile element containing a beta-lactamase found in *S. aureus*. Yellow rectangles represent coding sequences with predicted transposase activity; purple rectangles represent all other predicted coding sequences.

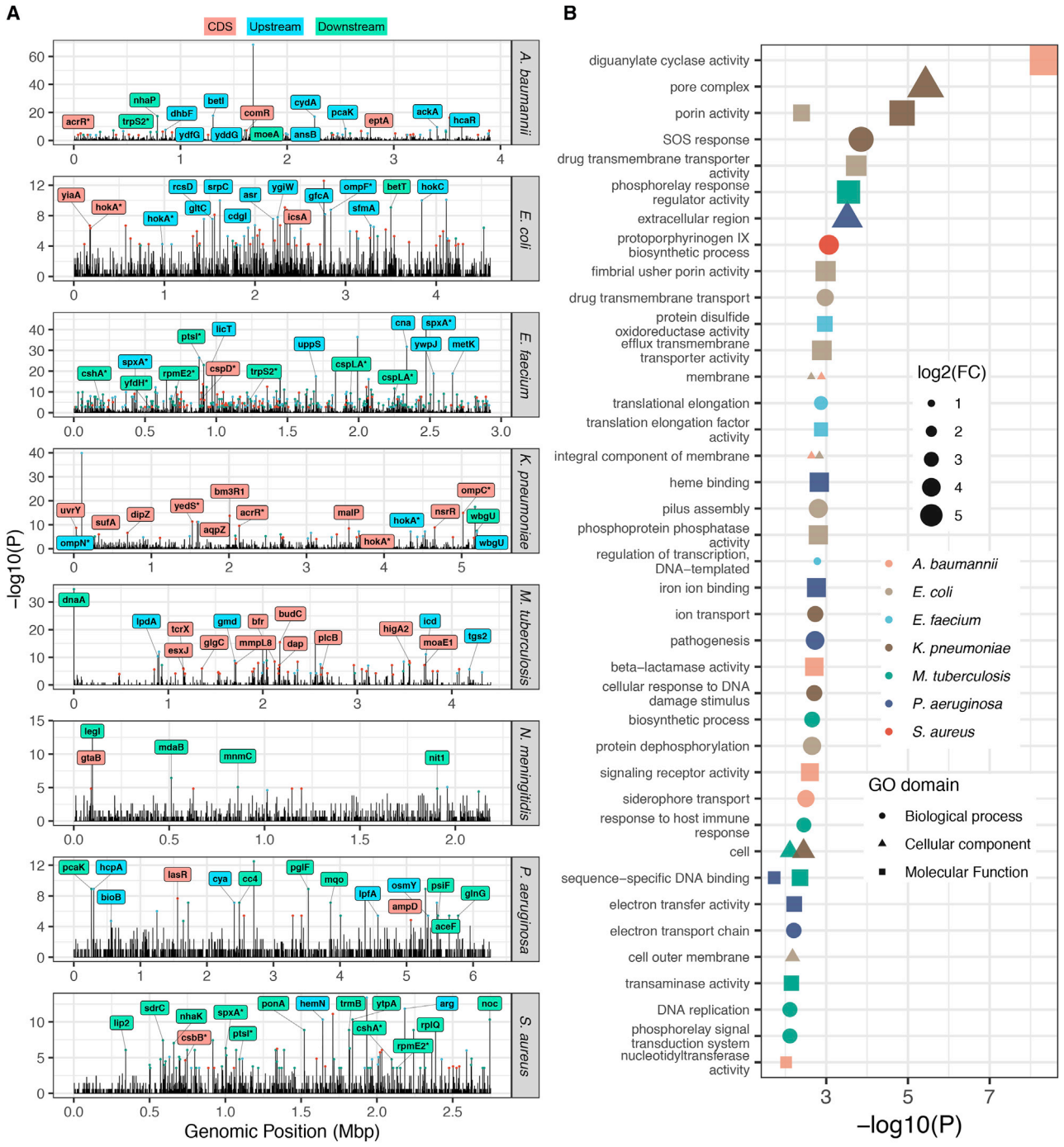


Figure 4. MGE-Insertion-Enriched Sites Occur near Functionally Important Genes and Pathways

(A) An analysis of MGE-insertion-enriched sites found on each species' chromosome. Insertion-enriched sites were assigned to coding sequences by choosing the coding sequence closest to the center of each insertion-enriched site. All insertion-enriched sites meeting FDR < 0.05 are indicated with the colored points. The 15 most significant insertion-enriched sites associated with well-annotated coding sequences are shown in the text labels. The insertion-enriched site is shown to be upstream (blue), within (red), or downstream of the nearest coding sequence (green). p values were calculated using a one-sided exact poisson test. (B) GO enrichment analysis of predicted coding sequences near MGE-insertion-enriched sites. All coding sequences near significant insertion-enriched sites were tested for enrichment of each GO term by using a hypergeometric test; all 44 significantly enriched GO terms are presented.

discovery rate [FDR] $\leq$ 0.05) (Table S6). We find that 227 of the insertion-enriched sites appear to directly overlap with predicted coding sequences, indicating loss-of-function for the disrupted gene. Interestingly, 192 are upstream of the nearest gene, and 235 are downstream, which makes the functional consequence of these insertions more difficult to predict.

Motif	Species	Element	# Targets	% Targets	% BG	-log10(P)
	*P. aeruginosa*	ISPa11 (IS110 Family)	45	95.74	0.69	94
	*K. pneumoniae*	ISEc33 (IS630 Family)	34	89.47	4.46	41
	*P. aeruginosa*	Novel (IS1182 Family)	24	96.00	3.46	33
	*A. baumannii*	ISAba45 (IS3 Family)	11	84.62	0.70	21
	*A. baumannii*	Novel (IS3 Family)	12	84.62	0.70	21

Figure 5. An Analysis of MGE Insertion Sites Reveals Their Target Site Specificity

Examples of target-sequence motifs identified for five different MGEs with high target-sequence specificity. "# Targets" refers to the number of unique insertion sites analyzed for each MGE. "% Targets" indicates the percentage of target sites containing the motif. "% BG" indicates the percentage of randomly chosen background sequences containing the motif. p values shown were calculated directly by the HOMER motif analysis software.

Next, we sought to identify genes that are frequently near TE insertion-enriched sites both within and across species (Table S2). The gene *acrR*, a drug efflux pump repressor, is adjacent to significant insertion-enriched sites in *K. pneumoniae* and *A. baumannii*, and a nominally significant insertion-enriched site in *E. coli* (uncorrected p = 5.2e−4; FDR-adjusted p = 0.14), indicating that disruption of this drug efflux pump repressor by MGE insertions is an adaptive strategy shared across multiple species. *Hok/sok* components are frequently targeted by unique MGE insertions in *E. coli* and *K. pneumoniae*, which indicates that disruption of these toxin-antitoxin systems by MGE insertion might be a common adaptive strategy in both of these species (Hayes 2003). Several transposases themselves are near MGE-insertion-enriched sites in *E. faecium*, indicating that these regions are already disrupted by nearby IS insertions in the reference genome, or that existing IS elements are frequently disrupted by TE insertions. Other genes repeatedly located near insertion-enriched sites include outer membrane porins *ompC*, *ompN*, *ompF*, and *yedS*; cold shock proteins *cspD* and *cspLA*; and general stress protein *glsB*, among others. Considering the repeated targeting of these homologous genes by unique TE insertions both within and across species, they are likely good candidate genes to investigate further for functional and adaptive significance.

To determine what types of genes are adjacent to significant insertion-enriched sites, we performed a Gene Ontology (GO) enrichment analysis of all genes near these sites. In this analysis, we identified 44 GO term pathways enriched near these sites (FDR $\leq$ 0.05) (Figure 4B; Table S6). Several of these enriched pathways are clearly associated with antibiotic resistance, including "pore complex" in *K. pneumoniae*, "porin activity" in *K. pneumoniae* and *E. coli*, both "efflux transmembrane transporter activity" and "drug transmembrane transporter activity" in *E. coli*, and "beta-lactamase activity" in *A. baumannii*. Other enriched pathways are associated with host infection and virulence, including "siderophore transport" in *A. baumannii*, "fimbrial usher porin activity" in *E. coli*, "pilus assembly" in *E. coli*, and "pathogenesis" in *P. aeruginosa*. However, the most significantly enriched pathway is diguanylate cyclase activity in *A. baumannii* (hypergeometric test; p = 4.5e−9), which is driven by insertion-enriched sites near response regulator *pleD*, diguanylate cyclase *dgcN*, and two homologs of diguanylate cyclase *dgcM*. Previous research suggests that disruption of this pathway might influence biofilm formation (Sarenko et al., 2017). Altogether, these results suggest that insertion-enriched sites in these populations of pathogenic isolates might specifically modulate cellular functions such as antibiotic resistance, virulence, pathogenesis, and biofilm formation across species.

Finally, we analyzed the sequence context of unique MGE insertions to determine their target-site specificity. Using HOMER (Heinz et al., 2010), a motif analysis software, we identified 63 elements that have significant target sequence motifs (p < 1e−11) (Table S6). Motifs for five MGEs with particularly high target sequence specificity are highlighted in Figure 5. Seven elements have similar (>70% pairwise similarity) CTAG target-site motifs, a motif that has been described previously for other IS elements (Fournier et al., 1993). The highly specific 12-base motif identified for ISPa11 corresponds with previous studies demonstrating that this element targets repetitive extragenic palindromic (REP) sequences throughout the genome (Tobes and Pareja, 2006). When we search for this 12-base motif throughout the *P. aeruginosa* reference genome, 32% of all matching motifs are occupied by an MGE insertion in at least one isolate, with 0.31 sites occupied per isolate on average, and a maximum of 6 sites occupied in 2 separate isolates. The target sequence specificity of the MGEs described here might be of interest to the field of genome engineering. In summary, regions frequently disrupted by unique (convergent) MGE insertions are associated with important biological functions such as antibiotic resistance, and by analyzing these regions we can determine the target-site specificity of these MGEs.

MGE Insertions Likely Contribute to Antibiotic Resistance in Laboratory Evolution Experiments and in Clinical Isolates

To definitively demonstrate a role for MGEs in evolution, orthogonal experimental approaches are necessary. Adaptive laboratory evolution (ALE) experiments are powerful tools to understand how drug resistance emerges. Given that our approach can be used to analyze MGE insertions in an experimental context from short-read sequencing data, we sought to determine how frequently MGE insertions contribute to antibiotic resistance in a controlled laboratory experiment. Studies have shown in laboratory grown *E. coli* that the rate of IS insertion is about one-third the rate of point mutation, but it is unclear how frequently these mutations would actually affect gene function

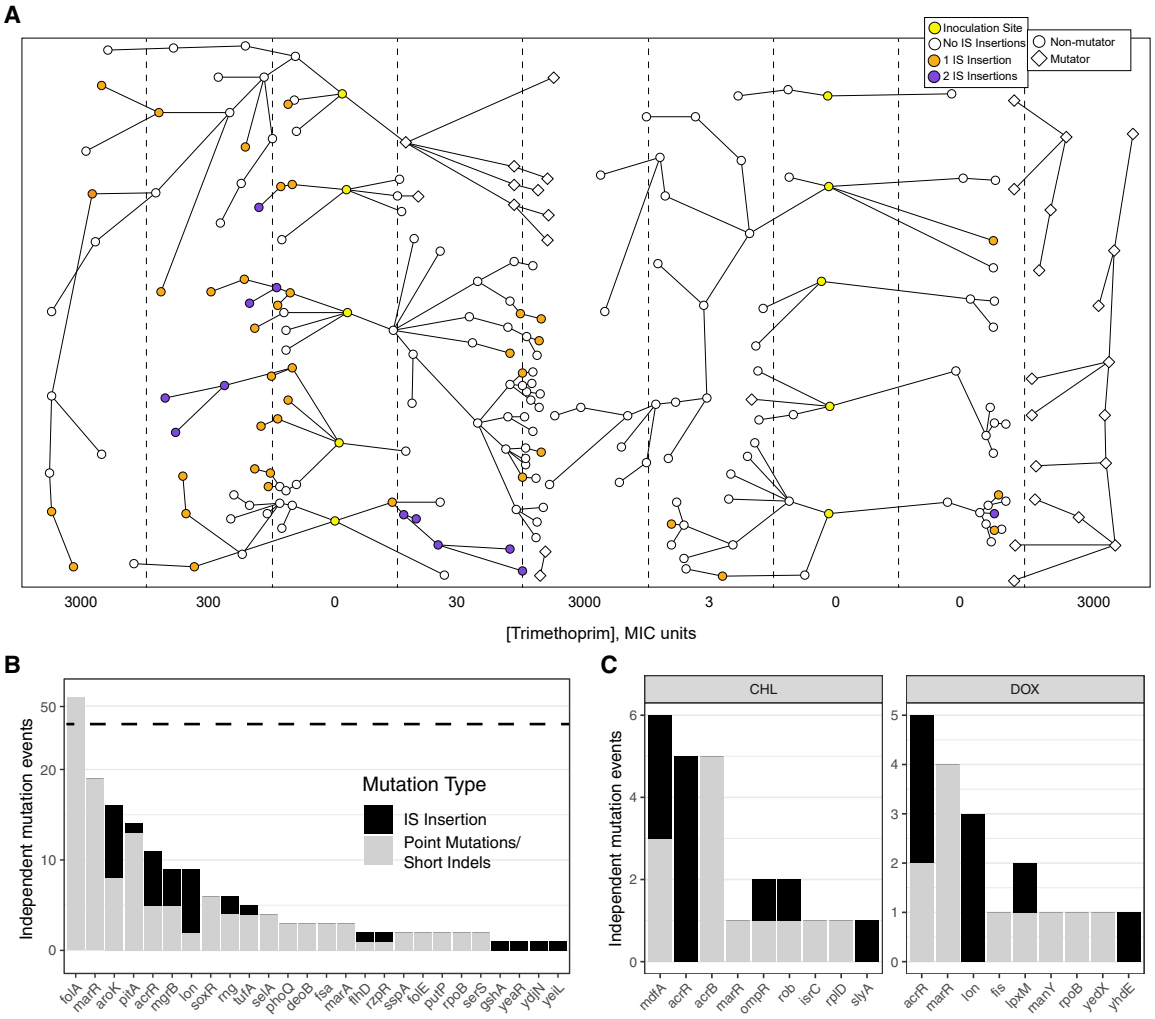


Figure 6. MGE Insertions Contribute to Antibiotic Resistance in Adaptive Laboratory Evolution Experiments

(A) A schematic representation of the intermediate-step trimethoprim megaplate experiment conducted by Baym et al. (2016), demonstrating the MGE insertion count in each sequenced isolate collected from the noted position on the megaplate. See also Figure S6.

(B) The number of independent mutation events assumed to affect each gene listed in the intermediate-step megaplate experiment conducted by Baym et al. (2016). The number of independent mutations is grouped according to mechanism: black, by MGE insertion; gray, by point mutations or short indels. This visualization includes MGE insertions affecting a gene only once (yeiL, ydjN, gshA, and yeaR) and excludes all point mutations/short indels affecting a gene only once.

(C) An analysis of the results of the chloramphenicol (CHL) and doxycycline (DOX) morbidostat experiment conducted by Toprak et al. (2011), supplemented with IS insertion information. The legend in (B) also applies to (C).

given that they seem to selectively target intergenic regions (Lee et al., 2016). Using the MGEfinder workflow, we re-analyzed data from two ALE experiments investigating the mechanisms by which *E. coli* adapts to prolonged exposure to antibiotics. These included the megaplate experiment conducted by Baym et al. (2016), and the morbidostat experiment conducted by Toprak et al. (2011).

We first analyzed the whole-genome sequencing (WGS) data collected by Baym et al. (2016), a study that introduced the megaplate as a means of visually observing a migrating bacterial front across a landscape of varying antibiotic concentrations (Baym et al., 2016). We found that MGE insertions played a significant role as a source of adaptive loss-of-function mutations. Among all sequenced isolates in the intermediate-step trimethoprim (TMP) experiment, we identified 35 independent IS insertions (Figures 6A, 6B, and S6; Table S7). Most of these insertions disrupted genes that were also found to be mutated by SNPs and indels in the originally reported study, including *acrR*, *aroK*, *pitA*, *mgrB*, *tufA*, and *rng*. Other genes disrupted by IS insertion were not reported in the original study. In comparison to the adaptive SNP and indel mutations originally reported for the TMP experiment, insertion sequences account for 17.6% (95% CI, 12.0%–23.2%) more adaptive mutations (defined as the mutations occurring in genes that are mutated independently at least twice) and 24.8% (95% CI, 17.2%–32.4%) when *folA* (the direct target of TMP) substitutions and indels are excluded.

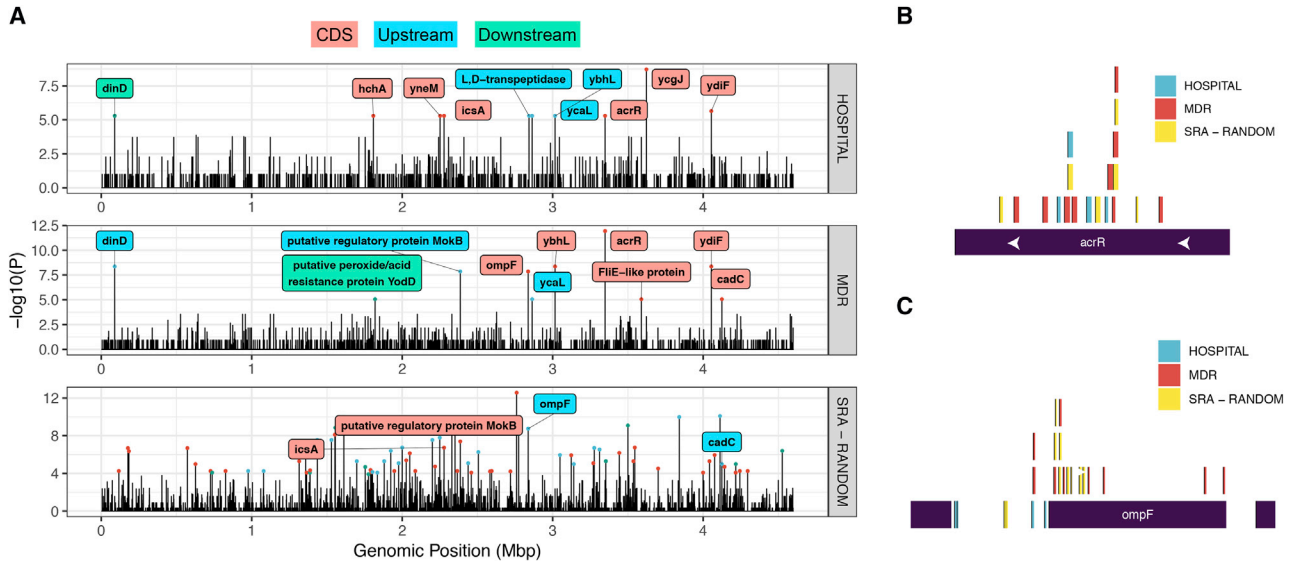


Figure 7. MGE Insertions Disrupt Known Antibiotic Resistance Genes in Clinical Isolates

(A) An analysis of MGE-insertion-enriched sites for two collections of clinical *E. coli* isolates. The third graph is the same as the *E. coli* graph shown in Figure 4A (insertion-enriched site analysis on randomly downloaded *E. coli* isolates from SRA), highlighting the insertion-enriched sites shared with either of the clinical isolate collections. See also Figure S7.

(B) Unique acrR MGE insertions found in the Hospital collection, the MDR collection, and the randomly downloaded *E. coli* isolates from the SRA database.

(C) All unique ompF MGE insertions found in the same three isolate collections.

Next, we analyzed the WGS data generated by Toprak et al., 2011, where they introduced the morbidostat, a selection device that continuously monitors bacterial growth and dynamically regulates drug concentrations to constantly challenge the bacterial population (Toprak et al., 2011). Toprak et al., 2011 treated drug-sensitive MG1655 *E. coli* with chloramphenicol (CHL), doxycycline (DOX), and TMP and performed whole-genome sequencing on 5 independently treated cultures. When considering only the mutated genes that were originally reported by Toprak et al., IS insertions caused 11 out of 25 (44%) and 8 out of 19 (42%) of the antibiotic resistance mutations detected in the CHL- and DOX-treated populations, respectively (Figure 6C). In the originally reported results for this study, two independent nonsense mutations were identified in the *acrR* gene across the ten CHL and DOX samples. When accounting for IS insertions, all ten CHL- and DOX-treated populations had a disrupted *acrR* gene, suggesting that *acrR* disruption is a strong contributor to CHL and DOX resistance in the context of this morbidostat experiment. This shows that IS insertions are key mutations that confer antibiotic resistance and that their inclusion in the analysis of such experiments is critical to form a more complete understanding of mechanisms of adaptation.

Although these *in vitro* findings suggest that MGE insertions contribute to antibiotic resistance, we also wanted to determine whether MGE insertions associated with antibiotic resistance also occur in clinical isolates. To address this directly, we chose to analyze two collections of clinical *E. coli* isolates with available antibiotic resistance phenotype information. The first is a collection of 241 *E. coli* bacteremia isolates previously investigated in a bacterial genome-wide association study (GWAS) (Stoesser et al., 2013; Earle et al., 2016). This collection was obtained from patients at the Oxford University Hospitals NHS Trust, and will be referred to as the "Hospital" collection. The second collection includes 260 *E. coli* clinical isolates collected from various locations across the United States. Several of these isolates were sequenced and analyzed in connection with the Federal Drug Administration (FDA)-Center for Disease Control and Prevention (CDC) Antimicrobial Resistance Bank (referred to as the multi-drug resistant [MDR] collection) (Table S4). Antibiotic resistance phenotypes were available for several drugs in both cohorts (Figure S7); of the two collections, the MDR collection contains more multi-drug resistant organisms (Figure S7B). Phylogenetic analysis indicates that isolates from both collections can be found in most major lineages (Figures S7C and S7D).

We ran MGEfinder on both of the isolate collections and performed an analysis to identify TE-insertion-enriched genomic sites (Figure 7A). We compared the insertion-enriched sites in each collection with the insertion-enriched sites found among the randomly downloaded Sequence Read Archive (SRA) isolates. We identified 9 insertion-enriched sites that replicated across the MDR, Hospital, and/or randomly downloaded SRA collections. The *acrR* insertion-enriched site replicated across the two clinical isolate collections (Figures 7A and 7B), again highlighting the importance of disruption of this gene in a collection of antibiotic-resistant isolates. An insertion-enriched site near outer membrane transporter *icsA* replicated between the hospital collection and the random SRA collection; insertion-enriched sites near the outer membrane porin *ompF* (Figure 5A and 5C), DNA-binding transcriptional activator *cadC*, and putative regulatory protein *mokB* (upstream of methyl-accepting chemotaxis protein *trg*) replicate between the MDR collection and the random SRA collection (Figure 7A). Insertion-enriched sites near DNA damage-inducible protein *dinD*, periplasmic protease *ycaL*, Bax1-I

family protein *ybhL*, and putative acetate-CoA transferase *ydiF* replicate between the hospital and MDR collections. We propose that the effect of the MGE insertions within other sites should be investigated further for functional significance.

The fact that these insertion-enriched sites replicate across isolate collections and occur near genes involved in antibiotic resistance indicates that MGE insertions confer important adaptations and are not merely sinks for random, functionally insignificant insertions. The *acrR* and *ompF* mutations, for example, are well-described antibiotic resistance mutations (Harder et al., 1981; Jellen-Ritter and Kern, 2001), and the effect of the MGE insertions within other sites should be investigated further for functional significance.

Finally, we wanted to estimate the frequency of gene disruption by MGE insertion compared with other nonsense mutations in MDR and Hospital isolate collections. We estimate that 12.7% (95% CI, 13.7%–18.1%) and 16.9% (95% CI, 15.7%–18.1%) of nonsense mutations are caused by MGE insertion in the hospital and MDR collections, respectively (Figures S7E and S7F). For *acrR*, 4 out of 17 (23.5%) and 8 out of 20 (40%) of nonsense mutations are mediated by MGE insertion in the Hospital and MDR collections, respectively. For *ompF*, 0 out of 8 (0%) and 5 out of 16 (31.2%) of nonsense mutations are mediated by MGE insertion in the Hospital and MDR collections, respectively. Altogether, these findings suggest that a significant proportion of nonsense mutations are mediated by MGE insertion in these clinical isolates and that the inclusion of MGE insertion data are critical to any comprehensive mutation analysis.

DISCUSSION

Genetic variation is produced by a variety of molecular mechanisms, and measuring all types of variation is critical when trying to understand how bacteria adapt and evolve. Here, we have presented MGEfinder, a sensitive and precise approach to genotyping large insertions from short-read sequencing datasets. We applied this approach to an analysis of several thousand bacterial isolates, identifying thousands of MGEs, and found MGE mutational signatures that highlight genes involved in antibiotic resistance.

As more isolates of a given species are analyzed, more MGE insertions are identified, suggesting that many of these elements are active in each of the species analyzed, although the level of activity of these elements varies between organisms. Of note, certain species, such as *E. faecium*, have particularly high MGE insertional activity. This suggests that certain species might rely more heavily on MGE movement as a mechanism of adaptation and evolution than others.

Using MGEfinder, we identified insertion sequences, MITEs, integrative plasmids, phage elements, transposons, integrons, group II introns, and other classes of MGEs in this analysis. Although some MGEs do not contain any coding sequence, many of the MGEs we identified contained genes encoding transposases as well as passenger genes coding for known functions, such as antibiotic resistance genes. Interestingly, many MGEs contained largely uncharacterized proteins, which we anticipate might provide yet undescribed adaptive advantages to the host bacterium. The passenger genes encoded in these MGEs can spread rapidly between organisms, and selective forces likely affect the retention versus loss of these elements in individual organisms and in communities of organisms, such as microbiomes. Indeed, recent work has demonstrated that individuals with very similar gut microbiomes can harbor very different MGE repertoires (Brito et al., 2016). This suggests that monitoring the MGE potential of individual bacterial species and microbiomes will inform our understanding of the extent of and consequences of MGE-derived genetic variation.

Because MGEfinder allows us to both identify MGEs and their insertion sites, we are able to find genes that are repeatedly "hit" by insertional mutagenesis, such as *acrR*, a gene involved in sensitivity to many different antibiotics (Jellen-Ritter and Kern, 2001). Insertional loss of function of the gene is identified at a high rate in *E. coli*, *K. pneumoniae*, and *A. baumannii* and likely correlates with increased antibiotic resistance of these organisms. In addition to identifying genes known to be involved in antibiotic resistance, we identify additional genic "insertion-enriched sites" that might represent genes involved in antibiotic resistance and pathogenicity.

When we apply this method to previously published adaptive laboratory evolution experiments on *E. coli*, we find that insertions comprise a large proportion of the acquired mutations in laboratory *E. coli*. By including MGE insertions in these analyses, the importance of *acrR* loss-of-function mutation as an adaptation to antibiotics is enhanced, re-prioritizing these mechanisms of resistance. Finally, we found that certain target sites appear to be hit by specific IS elements. These sequence-specific transposases might have value in genetic engineering applications. Thus, by identifying the locations where IS elements accumulate, we can identify genes important for bacterial fitness and the molecular specificity of transposase genes.

Although MGEfinder enables a detailed analysis of MGE insertions, there are several limitations of this approach. First, the inference approach used to identify large structural variants has limited precision when elements cannot be fully assembled in context by using short-read sequencing. With the advent of more accessible read cloud and long-read sequencing approaches (Bishara et al., 2018; Nicholls et al., 2019), we anticipate that such inferences can be readily orthogonally validated. Because our approach is based on comparative genomics, the choice of reference genome can affect the analysis in various ways. In this study, we chose a percent identity cutoff of 98.5% between all compared species, but others might choose more lenient cutoffs, realizing sensitivity will be diminished. Also, very small insertions (< 70 bp), large insertions (> 200 kbp), or insertions that create large target site duplications (> 20 bp) were not investigated in this study. We limited ourselves to only those insertion events that do not already exist within the reference genome, which emphasizes the role of more highly transposable mobile elements and elements of probable horizontal origin, potentially at the expense of more site-specific mobile elements. Finally, although the identified associations between gene disruption and antibiotic resistance are statistically significant, the predictions that these genes are associated with true antibiotic resistance must be validated by orthogonal *in vitro* testing.

In conclusion, we have developed a sensitive and precise approach to characterize a wide variety of MGEs and their sites of insertion from short-read sequencing data. We have

demonstrated the utility of this approach when analyzing large publicly available datasets, experimental data, and clinical isolates. Our analysis highlights the importance of thoroughly investigating MGE insertions in prokaryotic genomes. We anticipate that applying our workflow to a wide variety of different bacterial species will greatly enhance our understanding of MGEs and their role in bacterial adaptation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Data Sources, Preprocessing, and Quality Control
  - Randomly Selected Short-Read Data - Sample Selection, Preprocessing, Quality Control
  - Baym et al. and Toprak et al. Datasets - Sample Selection, Preprocessing, Quality Control
  - Clinical *E. coli* Isolate Collections with Antibiogram Data - Sample Selection, Preprocessing, Quality Control
  - MGE Identification Workflow
  - Identifying the Candidate Insertion Sites
  - Inferring the Complete Sequence of the Inserted Elements
  - Inferring Sequences from a Dynamically-Constructed Element Database
  - Clustering Elements across Isolates
  - Assigning Final Insertion Genotypes to Isolates
  - Simulations of Key Steps in the Pipeline
  - Unique Insertion and Elements Accumulation Curves
  - Calculating Rare TE Insertions per Megabase
  - Annotating Identified Elements
  - Identifying a Set of High-Confidence Mobility Genes
  - Describing MGEs with Annotated Passenger Genes
  - Annotating Reference Genomes
  - Analysis of Insertion-Enriched Sites
  - Insertion-Enriched Sites near Homologous Genes within and across Species
  - Insertion-Enriched Site Gene Ontology Enrichment Analysis
  - Target-Sequence Motif Discovery
  - Analysis of Baym et al. Megaplate Experiment Sequencing Data
  - Analysis of Toprak et al. Morbidostat Experiment Sequencing Data
  - Inferring Ancestral States and Identifying Independent IS Mutation Events in Clinical Isolates
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.chom.2019.10.022.

ACKNOWLEDGMENTS

We thank Duncan MacCannell at the CDC for his assistance with identifying clinical isolates to analyze in this study, Christopher T. Walsh at Stanford University and Kaitlin Tagg at the CDC for their helpful feedback and critiques, Brunilda Balliu and other members of the Montgomery Lab for their feedback and guidance, and Eli Moss and other members of the Bhatt Lab for their feedback and guidance and Ryan Brewster for his design of the graphical abstract. We also thank the anonymous reviewers whose feedback was critical in improving this manuscript. This work was supported by NIH R01AI148623 to A.S.B., a Donald E. and Delia B. Baxter Foundation Faculty Scholar award to A.S.B., the National Science Foundation Graduate Research Fellowship to M.G.D., and in part by NIH grant P30 CA124435, which supports the following Stanford Cancer Institute Shared Resource: the Genetics Bioinformatics Service Center.

AUTHOR CONTRIBUTIONS

M.G.D. conceived the study and analyses, designed the software, performed formal analyses, visualized the data, wrote the manuscript, and coordinated the project. M.M.L. helped conceive of the study and analyses, performed formal analysis, and helped to write the software and edit the manuscript. B.A.S. helped with data curation and editing the manuscript. S.B.M. provided expertise and helped with statistical analysis. A.S.B. helped with conceptualization, writing, and editing the manuscript, and funding acquisition.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 3, 2019
Revised: June 18, 2019
Accepted: October 29, 2019
Published: December 17, 2019

REFERENCES

Abby, S.S., Cury, J., Guglielmini, J., Néron, B., Touchon, M., and Rocha, E.P. (2016). Identification of protein secretion systems in bacterial genomes. Sci. Rep. *6*, 23080.

Adams, M.D., Bishop, B., and Wright, M.S. (2016). Quantitative assessment of insertion sequence impact on bacterial genome architecture. Microbial. Genomics *2*, e000062.

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. *44* (W1), W3–W10.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. *44* (W1), W16-21.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

Barrick, J.E., Colburn, G., Deatherage, D.E., Traverse, C.C., Strand, M.D., Borges, J.J., Knoester, D.B., Reba, A., and Meyer, A.G. (2014). Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. BMC Genomics *15*, 1039.

Baym, M., Lieberman, T.D., Kelsic, E.D., Chait, R., Gross, R., Yelin, I., and Kishony, R. (2016). Spatiotemporal microbial evolution on antibiotic landscapes. Science *353*, 1147–1151.

Bertelli, C., Laird, M.R., Williams, K.P., Lau, B.Y., Hoad, G., Winsor, G.L., and Brinkman, F.S.L.; Simon Fraser University Research Computing Group (2017).

IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res. *45*, W30–W35.

Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., and Bhatt, A.S. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat. Biotechnol. https://doi.org/10.1038/nbt.4266.

Biswas, A., Gauthier, D.T., Ranjan, D., and Zubair, M. (2015). ISQuest: finding insertion sequences in prokaryotic sequence fragment data. Bioinformatics *31*, 3406–3412.

Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. Nature *535*, 435–439.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Dabney, A., Storey, J.D., and Warnes, G.R. (2010). qvalue: Q-value estimation for false discovery rate control. ftp://ftp.uni-bayreuth.de/pub/math/statlib/R/CRAN/src/contrib/Descriptions/qvalue.html.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147.

Earle, S.G., Wu, C.H., Charlesworth, J., Stoesser, N., Gordon, N.C., Walker, T.M., Spencer, C.C.A., Iqbal, Z., Clifton, D.A., Hopkins, K.L., et al. (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat. Microbiol. *1*, 16041.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat. Protoc. *7*, 1728–1740.

Fournier, P., Paulus, F., and Otten, L. (1993). IS870 requires a 5′-CTAG-3′ target sequence to generate the stop codon for its large ORF1. J. Bacteriol. *175*, 3151–3160.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2019). PLSDB: a resource of complete bacterial plasmids. Nucleic Acids Res. *47* (D1), D195–D202.

Garrison, E. (2010). Others (FreeBayes. Marth Lab).

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv, Prepr. arXiv1207.

Hamlett, N.V., Landale, E.C., Davis, B.H., and Summers, A.O. (1992). Roles of the Tn21 merT, merP, and merC gene products in mercury resistance and mercury binding. J. Bacteriol. *174*, 6377–6385.

Harder, K.J., Nikaido, H., and Matsuhashi, M. (1981). Mutants of Escherichia coli that are resistant to certain beta-lactam compounds lack the ompF porin. Antimicrob. Agents Chemother. *20*, 549–552.

Hawkey, J., Hamidian, M., Wick, R.R., Edwards, D.J., Billman-Jacobe, H., Hall, R.M., and Holt, K.E. (2015). ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC Genomics *16*, 667.

Hayes, F. (2003). Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. Science *301*, 1496–1499.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass C, K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell *38*, 576–589.

Homer, N. (2017). DWGSIM. https://github.com/nh13/DWGSIM.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 5114.

Jellen-Ritter, A.S., and Kern, W.V. (2001). Enhanced expression of the multidrug efflux pumps AcrAB and AcrEF associated with insertion element transposition in Escherichia coli mutants Selected with a fluoroquinolone. Antimicrob. Agents Chemother. *45*, 1467–1472.

Jiang, C., Chen, C., Huang, Z., Liu, R., and Verdier, J. (2015). ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. BMC *16*, 72.

Jiang, X., Hall, A.B., Xavier, R.J., and Alm, E. (2017). Comprehensive analysis of mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. bioRxiv https://www.biorxiv.org/content/10.1101/214213v2.

Kathiresan, N., Temanni, M.R., and Al-Ali, R. (2014). Performance improvement of BWA MEM algorithm using data-parallel with concurrent parallelization. In 2014 International Conference on Parallel, Distributed and Grid Computing 406–411.

Köster, J., and Rahmann, S. (2012). Building and documenting workflows with python-based snakemake. In German Conference on Bioinformatics 2012. http://drops.dagstuhl.de/opus/volltexte/2012/3717/.

Krueger, F. (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lee, H., Doak, T.G., Popodi, E., Foster, P.L., and Tang, H. (2016). Insertion sequence-caused large-scale rearrangements in the genome of Escherichia coli. Nucleic Acids Res. *44*, 7109–7119.

Lees, J.A., Harris, S.R., Tonkin-Hill, G., Gladstone, R.A., Lo, S.W., Weiser, J.N., Corander, J., Bentley, S.D., and Croucher, N.J. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Research *29*, 304–316.

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity *104*, 520–533.

Lerat, E., and Ochman, H. (2004). Psi-Phi: exploring the outer limits of bacterial pseudogenes. Genome Research *11*, 2273–2278.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Liu, S.V., Saunders, N.J., Jeffries, A., and Rest, R.F. (2002). Genome analysis and strain comparison of correia repeats and correia repeat-enclosed elements in pathogenic Neisseria. Journal of Bacteriology *184*, 6163–6173.

Mahillon, J., and Chandler, M. (1998). Insertion sequences. Microbiology and Molecular Biology Reviews *62*, 725–774.

Mcgill, R., Tukey, J.W., and Larsen, W.A. (1978). Variations of box plots. Am. Stat. *32*, 12–16.

Murray, N.E., Gough, J.A., Suri, B., and Bickle, T.A. (1982). Structural homologies among type I restriction-modification systems. EMBO J. *1*, 535–539.

Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience *8*, giz043.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. (2011). Vegan: community ecology package. R package version 1.17.2, pp. 117–118.

Petersen, K.R., Streett, D.A., Gerritsen, A.T., Hunter, S.S., and Settles, M.L. (2015). Super deduper, fast PCR duplicate detection in Fastq Files. In Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA. 491–492.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rankin, D.J., Rocha, E.P., and Brown, S.P. (2011). What traits are carried on mobile genetic elements, and why? Heredity *106*, 1–10.

Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods in ecology and evolution. British. Ecol. Soc. *3*, 217–223.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Sarenko, O., Klauck, G., Wilke, F.M., Pfiffer, V., Richter, A.M., Herbst, S., Kaever, V., and Hengge, R. (2017). More than enzymes that make or break cyclic di-GMP-local signaling in the interactome of GGDEF/EAL domain proteins of *Escherichia coli*. MBio *8*, e01639-17.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Stoesser, N., Batty, E.M., Eyre, D.W., Morgan, M., Wyllie, D.H., Del Ojo Elias, C., Johnson, J.R., Walker, A.S., Peto, T.E., and Crook, D.W. (2013). Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J. Antimicrob. Chemother. *68*, 2234–2244.

Stokes, H.W., and Gillings, M.R. (2011). Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. FEMS Microbiology Reviews *35*, 790–819.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932.

Tobes, R., and Pareja, E. (2006). Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. BMC Genomics *7*, 62.

Toprak, E., Veres, A., Michel, J.B., Chait, R., Hartl, D.L., and Kishony, R. (2011). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat. Genet. *44*, 101–105.

Treepong, P., Guyeux, C., Meunier, A., Couchoud, C., Hocquet, D., and Valot, B. (2018). panISa: ab initio detection of insertion sequences in bacterial genomes from short read sequence data. Bioinformatics *34*, 3795–3800.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.

Xie, Z., and Tang, H. (2017). ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics *33*, 3340–3347.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. Genetics *141*, 1641–1650.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. In Methods in ecology and evolution, 8, G. McInerny, ed. (British Ecological Society), pp. 28–36.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M., and Larsen, M.V. (2012). Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. *67*, 2640–2644.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
All data used in this study was public or generated *in silico* for simulations.	NCBI Sequence Read Archive (SRA)	See Table S4
Software and Algorithms		
MGEfinder	This study	github.com/bhattlab/MGEfinder
SuperDeduper (v0.3.2)	Petersen et al., 2015	github.com/dstreett/Super-Deduper
Trim Galore (v0.5.0)	Krueger, 2015	www.bioinformatics.babraham.ac.uk/projects/trim_galore/
BWA MEM (v0.7.17-r1188)	Li and Durbin, 2009	bio-bwa.sourceforge.net
Bowtie2 (2.3.4.3)	Langmead and Salzberg, 2012	bowtie-bio.sourceforge.net
emboss	Rice et al., 2000	emboss.sourceforge.net
biopython python package	Cock et al., 2009	biopython.org
FASTQC (v0.11.7)	Andrews, 2010	www.bioinformatics.babraham.ac.uk/projects/fastqc/
panISa	Treepong et al., 2018	github.com/bvalot/panISa
progressiveMauve	Darling et al., 2010	darlinglab.org/mauve/user-guide/progressivemauve.html
dwgsim (v.0.1.11-3)	Homer, 2017	github.com/nh13/DWGSIM
CD-HIT	Fu et al., 2012	weizhongli-lab.org/cd-hit/
vegan R package (v2.5-3)	Oksanen et al., 2011	cran.r-project.org/web/packages/vegan/index.html
prokka (v1.13)	Seemann, 2014	github.com/tseemann/prokka
PopPUNK	Lees et al., 2019	https://github.com/johnlees/PopPUNK
HOMER	Heinz et al., 2010	http://homer.ucsd.edu/homer/
ISEScan	Xie and Tang, 2017	github.com/xiezhq/ISEScan
ResFinder	Zankari et al., 2012	cge.cbs.dtu.dk/services/ResFinder/
PHASTER	Arndt et al., 2016	phaster.ca/
bedtools	Quinlan and Hall, 2010	bedtools.readthedocs.io/en/latest/
DIAMOND	Buchfink et al., 2015	github.com/bbuchfink/diamond
phytools (v0.6.44)	Revell, 2012	cran.r-project.org/web/packages/phytools/index.html
FreeBayes	Garrison, 2010	github.com/ekg/freebayes
ggtree R package	Yu et al., 2017	https://doi.org/10.18129/B9.bioc.ggtree
snakemake	Köster and Rahmann, 2012	snakemake.readthedocs.io/en/stable/
fastANI	Jain et al., 2018	github.com/ParBLiSS/FastANI
SPAdes	Bankevich et al., 2012	github.com/ablab/spades
hmmsearch	Eddy, 2011	hmmer.org
ConjScan	Abby et al., 2016	research.pasteur.fr/en/software/conjscan-t4ssscan/
Pilon	Walker et al., 2014	https://github.com/broadinstitute/pilon/wiki
Other		
PLSDB	Galata et al., 2019	https://ccb-microbe.cs.uni-saarland.de/plsdb/
Toprak et al. Dataset	Toprak et al., 2011	BioProject PRJNA274794
Baym et al. Dataset	Baym et al., 2016	BioProject PRJNA259288
MDR Isolate Collection	Multiple Sources	BioProjects PRJNA278886, PRJNA288601, PRJNA292901, PRJNA292902, PRJNA292904, PRJNA296771, and PRJNA316321
Hospital Isolate Collection	Stoesser et al., 2013; Earle et al., 2016	BioProject PRJNA306133

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information regarding the data and code presented in this study is available through the Lead Contact, Ami S. Bhatt (asbhatt@stanford.edu). This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study was a computational analysis of existing bacterial isolates and datasets. This included sequencing data downloaded at random from the NCBI SRA database, the study performed by Baym et al. (2016), the study performed by Toprak et al. (2011), the MDR *E. coli* isolate collection, and the Hospital *E. coli* collection (Earle et al., 2016). See Table S4 for additional details.

METHOD DETAILS

Data Sources, Preprocessing, and Quality Control

All of the data analyzed in this study were downloaded directly from public databases. The datasets are divided into three categories for clarity: (1) Randomly selected Illumina short-read datasets of bacterial isolates, (2) Short-read *E. coli* sequencing datasets generated by Baym et al. and Toprak et al. for adaptive laboratory evolution experiments (Toprak et al., 2011; Baym et al., 2016), and (3) Clinical *E. coli* isolate collections with antibiogram data. The sample selection, preprocessing, and quality control for each dataset is described below.

Randomly Selected Short-Read Data - Sample Selection, Preprocessing, Quality Control

The Sequence Read Archive (SRA) SQL database was downloaded on Sep. 25th, 2018. Potential sequencing datasets were filtered initially by available metadata to only include those samples with an estimated coverage between 50-150x per isolate, and only including samples annotated as paired-end, whole-genome sequencing samples.

In simulations, as the genetic distance between the simulated samples and the reference genome increases, both sensitivity and precision are reduced. To maintain high performance across samples, it was necessary to choose a reference genome that was closely related to the randomly downloaded SRA samples. All complete genomes were downloaded from the NCBI RefSeq database for each of the nine species of interest. The tool PopPUNK was used to rapidly estimate the core genome distances between the reference genomes (Lees et al., 2019). This distance matrix was clustered by the k-means clustering algorithm, and the genomes closest to the center of each cluster were chosen. 100-200 randomly downloaded isolates from each species were aligned to all of these selected genomes and the relevant NCBI reference genome. Quality control filters were applied, and all genomes with a genetic distance greater than 0.015 mutations per covered bp were removed. According to simulations, this cutoff should put a lower bound on the overall sensitivity of MGEfinder at around 0.85. Reference genome(s) that resulted in a high number of downloaded isolates meeting the quality filters were chosen. The final reference genomes used are: *Escherichia coli* NCTC9084 (NCBI Reference Sequence: NZ_LR134075.1), *Pseudomonas aeruginosa* PAO1 (NCBI Reference Sequence: NC_002516.2), *Neisseria gonorrhoeae* RIVM0610 (NCBI Reference Sequence: NZ_CP019466.1), *Neisseria meningitidis* M22811 (NCBI Reference Sequence: NZ_CP016654.1), *Staphylococcus aureus* subsp. *aureus* LGA251 (NCBI Reference Sequence: NC_017349.1), *Enterococcus faecium* E7933 (NCBI Reference Sequence: NZ_LR135384.1), *Mycobacterium tuberculosis* GG-137-10 (NCBI Reference Sequence: NZ_CP025606.1), *Klebsiella pneumoniae* subsp. *pneumoniae* ATCC 43816 KPPR1 (NCBI Reference Sequence: NZ_CP009208.1), and *Acinetobacter baumannii* strain XH856 (NCBI Reference Sequence: NZ_CP014541.1).

Up to 2000 samples for each species of interest were randomly selected and downloaded. The reads in the downloaded FASTQ files were deduplicated using SuperDeduper v0.3.2 (Petersen et al., 2015) and adaptor sequences were trimmed using Trim Galore v0.5.0 (Krueger, 2015). Samples were then aligned to their respective reference genomes using BWA MEM v0.7.17-r1188 (Li and Durbin, 2009). Samples were excluded if they met the following filters: greater than 0.015 mutations per covered bp, median coverage less than 40, estimated average read length less than 95 or greater than 305, estimated average fragment length less than 150 or greater than 750, a FASTQC v0.11.7 ''Per base sequence quality'' quality control failure, a FASTQC ''Per sequence GC content'' quality control failure, and a FASTQC ''Per base N content'' failure (Andrews, 2010). Those sequence isolates that passed these filtering steps were analyzed further to identify MGEs and their sites of insertion.

Baym et al. and Toprak et al. Datasets - Sample Selection, Preprocessing, Quality Control

Baym et al. and Toprak et al. datasets were downloaded from the NCBI Sequencing Read Archive (Bioproject accessions PRJNA259288 and PRJNA274794, respectively) (Baym et al., 2016; Toprak et al., 2011). Samples were processed by removing duplicate sequences using SuperDeduper v0.3.2 (Petersen et al., 2015) and adaptor sequences were trimmed using Trim Galore v0.5.0 (Krueger 2015). The Baym et al. and the Toprak et al. short-read sequences were aligned to the *E. coli* K12 U00096.2 and NCBI Reference Sequence: NC_000913.2 reference genomes, respectively, as was done in the original studies.

Clinical *E. coli* Isolate Collections with Antibiogram Data - Sample Selection, Preprocessing, Quality Control

Two collections of pathogenic clinical *E. coli* isolates were used in this study. The first is referred to as the Hospital collection as it represents *E. coli* isolates collected at a single hospital, the Oxford University Hospitals NHS Trust. This included 241 isolates in total,

all of which were bacteremia samples. They were downloaded from the NCBI database by querying all *E. coli* samples in BioProject PRJNA306133.

The second collection, referred to in this study as the Multi-Drug Resistant (MDR) collection, includes 260 isolates collected from multiple BioProjects, including PRJNA278886, PRJNA288601, PRJNA292901, PRJNA292902, PRJNA292904, PRJNA296771, and PRJNA316321 (See Table S4). Most of these isolates come from the projects PRJNA278886 (173 isolates, Antimicrobial Surveillance from Brigham & Women's Hospital, Boston MA) and PRJNA288601 (52 isolates, CDC's Emerging Infections Program (EIP)). Antibiograms were collected for samples from NCBI using the search key term ''antibiogram[filter].'' This collection represents samples taken from several locations throughout the country as part of many pathogen surveillance programs. They were isolated from a variety of sources, including 70 isolated from blood, 142 isolated from urine, and 27 isolated from other sources.

Samples were processed first by removing duplicate sequences using SuperDeduper v0.3.2 (Petersen et al., 2015), and adaptor sequences were trimmed using Trim Galore v0.5.0 (Krueger 2015). All isolates from both collections were then aligned to both the *E. coli* NCTC9084 genome (NCBI Reference Sequence: NZ_LR134075.1) and the *E. coli* FDAARGOS_144 (NCBI Reference Sequence: NZ_CP014111.1). The NCTC9084 reference genome was used when making comparisons between these collections and the randomly downloaded SRA collection, and the FDAARGOS_144 reference genome was used otherwise. BWA MEM was used for alignments (Kathiresan et al., 2014), with default settings in paired-end mode.

MGE Identification Workflow

A combination of several previously published tools and custom tools, detailed below, were used to identify MGE insertions and their sequence from short-read sequencing data. This pipeline is summarized in five steps: (1) Identifying the candidate insertion sites, (2) inferring the complete sequence of the inserted element, (3) inferring sequences from a dynamically constructed database, (4) clustering elements across isolates and (5) assigning final insertion genotypes to isolates.

Identifying the Candidate Insertion Sites

This step is fully implemented in the MGEfinder software package under the find command. The approach taken to identify candidate insertion sites was developed independently but is similar to one taken recently by another group who published their tool under the name panISa (Treepong et al., 2018). First, the alignment is parsed to identify sites where reads are clipped according to the BWA MEM alignment software, filtering out reads with a mapping quality less than 20 (min_alignment_quality parameter) or those that are clipped on both sides and have an alignment length less than 21 bps (min_alignment_inner_length parameter). Second, clipped sites where the longest clipped end falls below a total length of 8 are excluded (min_softclip_length parameter). Third, clipped sites that have fewer than 2 supporting reads in total are excluded (min_softclip_count parameter). Fourth, sites that are not within 22 bps of an oppositely oriented read clipped site are excluded (min_distance_to_mate parameter).

In the next step, information about the un-clipped reads overlapping each of the candidate insertion sites is calculated. This is an important quality control step that filters out small indels, which commonly cause false positives. First, reads at the insertion site are classified as clipped reads, reads containing small insertions near the insertion site (small insertion reads), reads containing large insertions near the insertion site (large insertion reads), reads containing deletions near the insertion site (deletion reads), and reads that completely span the insertion site (run through reads). Small insertion reads are defined as those reads that contain insertions smaller than 30 bps (large_insertion_cutoff), and large insertion reads are those reads with insertions that exceed 30 bps. Clipped reads and large insertion reads are used to support the existence of an insertion, and small insertion reads and deletion reads are used to filter out sites that are likely small insertions or deletions. Sites where the ratio of the insertion-supporting reads to total reads falls below a value of 0.15 are excluded (min_softclip_ratio parameter). Then sites where the ratio of the deletion reads plus small insertion reads to the total number of reads at the site exceeds 0.03 are then excluded (max_indel_ratio parameter). This is repeated for deletions located at the base pairs directly adjacent to the insertion site. These filters were identified largely by iterative attempts to optimize sensitivity and precision, and could be further improved in the future. The result of this step is a filtered list of candidate sites.

With this filtered list of candidate sites, consensus sequences for the termini of the inserted element at each site are determined. It is possible to observe two distinct terminal sequences at a single site, and an approach that could filter out poor quality reads to produce a high-quality consensus was taken. The clipped ends are first added to a trie data structure. All of the unique paths from the parent node to the leaves are traversed, resulting in a list of unique sequences seen at an individual site. These sequences are then clustered with each other in a pairwise manner by truncating the longer sequence to the length of the shorter sequence, and by calculating a similarity metric as the edit distance divided by the total length of the shorter sequence. This matrix of similarity scores is then analyzed to identify all connected components, with connections existing between all sequences with a similarity greater than 0.75. Each component of sequences forms its own cluster, and this cluster is then analyzed to determine a consensus sequence.

Next, a consensus sequence for a cluster of sequences is constructed by traversing down the trie data structure and taking the base with the highest average quality score at each level of the trie as the consensus. The trie structure is traversed until only until one read supports a given site, and the consensus sequence is terminated at this point. Consensus sequences that fall below 8 bps in length (min_softclip_length parameter) and that have fewer than 4 clipped ends supporting their existence (min_softclip_count parameter) are excluded, and the most well-supported consensus sequence is chosen as the representative. Finally, the list of candidate sites are filtered again to only include sites that are not within 22 bps of an oppositely oriented read clipped site (min_distance_to_mate parameter).

These candidate sites are then paired with other sites that are within a specified distance and oriented in the opposite direction using the MGEfinder command pair. These oppositely oriented termini are allowed to be up to 20 bases away from each other, as many MGEs create a target site duplication upon insertion. Since many MGEs, such as IS elements, have terminal inverted repeats, termini that share inverted repeats near their termini are prioritized. Ties are then broken first by pairing termini that have similar number of supporting clipped reads, and then by the difference in the length of the consensus sequences. If ties still exist, the pairs are randomly assigned to each other, but this is a rare occurrence. If no terminal inverted repeats exist between any of the pairs, then only the number of clipped reads and the difference in consensus sequence length are used to pair insertion termini.

The Baym et al. data included several isolates that were sequenced at low coverage (less than 20). To increase sensitivity for these low-coverage samples, the consensus sequence at each site was built from all available clipped reads by setting the min_count_consensus parameter to 1, and a minimum consensus length of 4 was used (min_softclip_length parameter).

Inferring the Complete Sequence of the Inserted Elements

Once candidate terminus pairs have been identified, the next step is to infer the full inserted element. This approach combines a variety of methods for inferring the identity of each insertion. The approach taken is described here in detail.

First, the identity of the inserted sequence is inferred from the assembly of the sequenced isolate using the inferseq-assembly command in MGEfinder. For each pair of termini identified, each terminus is aligned to the assembly using 25 bases of genomic sequence context for each terminus in single-end mode (Figure S1A). If these consensus termini with genomic sequence context align to the assembled isolate with the proper orientation, one can assume with high confidence that the intervening sequence is the complete inserted sequence. This sequence is described as ''inferred from assembly with full context.'' If only one terminus aligns with context, and the other partially aligns to the edge of the same assembled contig, this sequence is described as ''inferred from assembly with half context.'' These are the highest quality inferred sequences, and they are prioritized above all others when genotyping an insertion.

Next, termini are aligned to the assembly without any context sequence. Often, these termini align to small assembled fragments, with short parts of the terminal ends being clipped off at the ends of the assembled contig. The full sequence is inferred by including the clipped terminal ends, and the full intervening sequence (Figure S1B).

The next approach to infer the sequence identity is implemented in the inferseq-overlap command of MGEfinder. This infers the identity of the sequence by attempting to find high-confidence overlaps between the two termini. This can only identify inserted elements that can be spanned by the two termini, which makes it a good option for identifying smaller insertions.

Next, termini are aligned to the reference genome using the inferseq-reference command, and inserted sequences are inferred in a similar manner to those inferred from assemblies without sequence context (Figure S1C). At each of these sequence inference steps, all candidate insertions where the alignment scores of both termini are equally high are returned. For example, if a given IS element is found in multiple locations throughout the reference genome, all of these locations will be reported as inferred elements (Figure S1E). Only inferred elements between 30 bps and 200 kbps are returned in each of these inference steps (min_inferseq_size and max_inferseq_size).

Inferring Sequences from a Dynamically-Constructed Element Database

All of these inferred elements can then be combined into a single FASTA database using the makedatabase command. This step generates a database of elements that is filtered by 99 percent nucleotide similarity using CD-HIT-EST (Li and Godzik, 2006; Fu et al., 2012).

This final sequence inference approach, implemented in the MGEfinder package as the inferseq-database command, is similar to the sequence inference approaches implemented in the MGEfinder commands inferseq-assembly and inferseq-reference, but with two differences: a minimum percent identity of the aligned termini of 90% (as opposed to the 95% minimum identity required for other inference commands), and the requirement that the aligned termini map within 10 bases of the true termini of each element in the database to be considered a candidate for the inserted element at a given position (max_edge_distance parameter). This serves as the most sensitive inference approach of all, but the results depend on how the query database is constructed and should not be considered to be especially high-quality relative to other methods of sequence inference.

Clustering Elements across Isolates

This next step is performed using the clusterseq command in MGEfinder. At this point in the pipeline, thousands of insertion positions have been identified, and the exact identity of those insertions could be any number of hundreds of different inferred sequence elements. Much of this information is surely redundant, however, as the differences between elements may amount to a few nucleotide variations. To collapse this small level of heterogeneity, sequences were clustered using CD-HIT-EST at 90% similarity across 85% of each sequence (Li and Godzik, 2006; Fu et al., 2012). For example, if elements X and Y are found at position A in the reference genome in two different isolates, it is assumed that elements X and Y are in essence the same sequence if they cluster at 90% similarity across 85% of their sequence, and it is assumed they are different elements otherwise. At this step, only elements between 70 bps and 200 kbps are kept by default (min_inferseq_size and max_inferseq_size). The lower cutoff could be reduced to 30 bps under certain conditions (see Figure S2E), but to ensure sufficient sensitivity to detect all elements in the size range of interest, we used a lower cutoff of 70 bps.

This step results in two different elements set, which are referred to as element "clusters" and element "groups." Element clusters include those sets of inferred elements that meet sequence similarity cutoffs as calculated by CD-HIT-EST above. Element groups are sets of element clusters that occasionally cannot be distinguished from each other at the inference step. For example, if 10 unique elements are inferred across the inferseq commands for a given insertion, they may be clustered by CD-HIT-EST to produce 2 distinct element clusters, Cluster1 and Cluster2. Because these two element clusters are indistinguishable at this insertion site, they both become members of a single element group, referred to as Group1. If at a different insertion site, Cluster2 and Cluster3 are inferred together, then Cluster3 is also added to Group1. In this sense, for a given insertion there may be some ambiguity at the level of element cluster, but there will never be ambiguity at the level of element group. It should be noted that the strand orientation of the inserted element is ignored in this analysis. For example, if element X was inserted at position A in isolate 1, and was also inserted at position A in isolate 2 but with the reverse orientation, this will be considered an identical genotype. These two different element sets can be used for different tasks at the discretion of the researcher, depending on the level of ambiguity that is considered acceptable for the task at hand.

Assigning Final Insertion Genotypes to Isolates

Once elements were filtered by size and organized into clusters, insertions in all isolates are assigned a final genotype using the genotype command. Insertions are assigned to element clusters by prioritizing the results of the inference methods. Element clusters inferred from "assembly with full context" are given first priority, those inferred from "assembly with half context" are given second priority, those inferred from "overlap" are given third priority, those inferred from "assembly without context" are given fourth priority, and both those inferred from "dynamically constructed database" and those inferred from the reference genome given fifth and lowest priority.

In the next step, we seek to resolve ambiguous position-cluster assignments. Even after clustering elements and assigning clusters to insertion sites in this manner, ambiguities may exist. If the ends of two different elements are very similar to each other, and yet they do not cluster together at the 90% similarity threshold, a given insertion may have mapped to both of these clusters. It is important that these ambiguities are sufficiently resolved for many of the analyses conducted in this study.

First, the number of non-ambiguous positions within the reference genome where each element cluster is found are counted. This is first done for only insertions that are inferred from the isolate's assembly with sequence context, the highest confidence set. If a given cluster is found at more than two of these high-confidence insertion positions, it is classified as a "high-confidence MGE." These requirements are then relaxed to include all non-ambiguous position-cluster assignments, the next highest confidence set, and all clusters that are found at more than one position in the reference genome are classified as "low-confidence MGEs."

Next, ambiguous element assignments are counted. The ambiguous element assignments are first resolved by calculating the frequency of each cluster at each position within our collection of isolates and assigning the ambiguous insertion to the most frequent cluster at that position. For example, if an ambiguous insertion at position A in isolate 1 maps to both clusters X and Y, and throughout the population cluster X is found more frequently than cluster Y at position A, then cluster X is assumed to be the correct assignment at position A in isolate 1.

Finally, if any ambiguous position-cluster assignments remain, they are then resolved by prioritizing "high-confidence MGEs" described above. Our assumption is that if the insertion at position A maps to both clusters X and Y, and X is an MGE whereas Y is not, then position A is assigned to cluster X. If any ambiguous position-cluster assignments remain after this step, clusters are assigned using the "lenient MGEs." If a given insertion is still ambiguous, it is left in as an ambiguous insertion, but still considered in analyses where cluster ambiguity is irrelevant.

As a final quality control measure, elements that are never successfully inferred from the sequence assembly of any analyzed isolate are removed (filter-clusters-inferred-assembly parameter). This should reduce the number of false-positives that are inferred from the reference genome by chance. In some cases, such as when analyzing ALE experiment sequencing data, the sequences inferred from the reference genome may be considered high-quality, and this filter can be disabled (no-filter-clusters-inferred-assembly parameter).

Simulations of Key Steps in the Pipeline

To test the sensitivity and precision of our pipeline, we carried out various simulations. Here we demonstrate that the candidate insertion identification and sequence inference steps are both sensitive and precise. The sequences used for the simulations were: *Escherichia coli* NCTC9084 (NCBI Reference Sequence: NZ_LR134075.1), *Pseudomonas aeruginosa* PAO1 (NCBI Reference Sequence: NC_002516.2), *Neisseria gonorrhoeae* RIVM0610 (NCBI Reference Sequence: NZ_CP019466.1), *Neisseria meningitidis* M22811 (NCBI Reference Sequence: NZ_CP016654.1), *Staphylococcus aureus* subsp. *aureus* LGA251 (NCBI Reference Sequence: NC_017349.1), *Enterococcus faecium* E7933 (NCBI Reference Sequence: NZ_LR135384.1), *Mycobacterium tuberculosis* GG-137-10 (NCBI Reference Sequence: NZ_CP025606.1), *Klebsiella pneumoniae* subsp. *pneumoniae* ATCC 43816 KPPR1 (NCBI Reference Sequence: NZ_CP009208.1), and *Acinetobacter baumannii* strain XH856 (NCBI Reference Sequence: NZ_CP014541.1). For each genome, the mutation rate, coverage, and library read length were combinatorially selected and ten replicates of each combination of parameters were carried out. A total of 32 MGE insertions were simulated per isolate by randomly selecting insertion sites, using 16 species-specific IS elements downloaded from ISfinder and chosen at random, and 16 randomly generated sequences of lengths 30, 70, 100, 200, 300, 400, 500, 1000, 2000, 5000, 10000, 50000, and 200000 bps in length. We simulated four mutation rates (0.001, 0.005, 0.01, and 0.02 mutations per bp, with 10% of mutations being indels) using DWGSIM

v.0.1.11-3 (https://github.com/nh13/DWGSIM). Theses simulations included various genome coverages (5, 10, 20, 40, 60, 80, and 100x), and various read lengths (100, 150, 300 bp), and reads were aligned to genomes in paired-end mode with BWA MEM (Kathiresan et al., 2014). In total, this resulted in 840 simulated genomes per species. ANOVA tests implemented in R were used to analyze which factors influence sensitivity and precision. The MGEfinder tool was run on these simulated samples with the default parameters.

For comparison, these simulated genomes were also analyzed with panISa, and the results were compared directly to those of MGEfinder. The parameters used for panISa were chosen to most closely reflect the MGEfinder default parameters, and included setting the minimum number of clipped reads to predict an insertion to 2 using the"minimun" parameter.

In Figures S2A–S2D and S2F–S2H, the sensitivity and precision curves shown are based on whether or not the recovered termini are found near the expected insertion site, with termini being considered true positives if they are 90% similar to the termini of the true inserted elements. Similarity here is calculated as the edit distance divided by the length of the terminus. Additionally, other sequence inference techniques implemented in MGEfinder were tested to determine their sensitivity (Figure S2E).

MGEfinder was also compared to the progressiveMauve algorithm, a commonly used tool to identify genomic islands (Figures S2I and S2J). For each of the 9 species, 10 of the insertion simulations performed at a mutation rate of 0.01 with 300 bp reads were assembled using SPAdes. These simulated assemblies were then used to compare the ability of MGEfinder to progressiveMauve, including comparing the sensitivity of recovery of repetitive insertions, non-repetitive insertions, and the improvement in sensitivity achieved when sharing insertion information across isolates using MGEfinder.

Unique Insertion and Elements Accumulation Curves

A unique insertion is defined as a specific element group assigned to a specific insertion position. Element groups are used rather than element clusters (See ''Clustering elements across isolates'') to prevent the double-counting of sites with ambiguously assigned element clusters. Using the unique insertions that were identified, the number of new insertions identified as additional samples were analyzed was calculated. The ''specaccum'' function provided by the vegan package v2.5-3 was used to estimate the accumulation curve for these insertions (Oksanen et al., 2011). The ''random'' method was used to estimate the curve, with 100 permutations. These same steps were repeated for different element types (Figure S4).

Calculating Rare TE Insertions per Megabase

For each species, the number of rare TE insertions per megabase of covered genome was calculated. A transposable element (TE) is defined as any element with more than three predicted insertion sites within the reference genome in sum across all analyzed isolates. The allele frequency of each insertion was calculated by dividing the number of isolates where the insertion is observed by the total number of isolates analyzed for the species. This list was then filtered to only include insertions with an allele frequency less than 0.01 (1% of isolates), and we then calculated the number of such rare insertions per isolate. The number of megabases for each sample with non-zero read coverage was then calculated by analyzing the alignment files for each isolate, and the number of rare variants per isolate was divided by this number of megabases. This was repeated for rare TE insertions falling in both intergenic and coding regions (Figure S5).

The ANI between each isolate and its respective reference genome was calculated using fastANI (Jain et al., 2018), as well as the ANI between all pairs of isolates for each species. The average pairwise ANI for each isolate compared to all other isolates analyzed for the species was calculated. This tool failed for 312 isolates (2.5% of total) due to low contiguity (N50) of their respective assemblies. As a proxy for these draft assemblies, the reads for these isolates were aligned to their respective reference genomes, and those reference genomes were corrected to contain all of the same SNPs and indels as the isolate. Regions in the reference that had zero coverage or were unmappable were then masked. Using these corrected reference genomes with fastANI produced good estimates of the ANI for these isolates. An ANOVA model including species, reference ANI, average pairwise ANI, sequencing library read length, sequencing library fragment length, and median depth of coverage as groups were used to understand how each covariate related to the rate of rare TE insertion.

Annotating Identified Elements

All identified elements were annotated using the Prokka v1.13 annotation software (Seemann 2014). This approach uses a rapid hierarchical approach to classify proteins, with the databases being derived from UniProtKB. Default settings were used, with the exception of the added flags ''–kingdom Bacteria'' and ''–metagenome''. The ''metagenome'' flag was used to improve prediction of genes in short contigs. All unique identified elements were annotated, not just the cluster representatives.

Transposases are poorly annotated using the Prokka annotation pipeline, and they are often described only as ''hypothetical'' proteins, so a different approach to identify potential transposases was necessary. The profile hidden Markov models (pHMMs) constructed by the creators of the ISEScan software were used to identify potential transposases (Xie and Tang, 2017), using Prokka predicted proteins as inputs. All such proteins with e value < 10-4 were considered to be transposases. We designated an element as an IS element if it only coded for predicted transposases.

The ResFinder web portal was used to identify antibiotic resistance genes (Zankari et al., 2012). The PHASTER URL API was used to predict which elements were considered intact, questionable, or incomplete phages (Arndt et al., 2016). We designated an element as a phage only if the predicted phage boundaries exceeded 85% of the length of the element. We ran BLAST on elements with the PLSDB plasmid database to determine which elements were plasmids (Galata et al., 2019), and only designated an element as a

plasmid if the BLAST alignments covered 85% of both elements and nucleotide identity exceeded 90%. Element clusters that contained predicted terminal inverted repeats (TIRs) at least 8 bp in length in at least 10% of cluster members were classified as elements with TIRs.

We assigned each element to a single category in the following order. If an element was a predicted IS element, it was assigned to the "IS element" category. If an element was a predicted intact phage, it was assigned to the "Intact Phage" category. If an element was either a questionable or an incomplete phage, it was assigned to the "Questionable/Incomplete Phage" category. If an element contained a predicted transposase and additional predicted non-transposase CDS regions, it was assigned to the "Transposase + CDS" category. If an element contained a gene with a predicted Group II Intron protein domain (matching pHMM TIGRFAMs: TIGR04416 at e-value < 1e-4), it was assigned to the "Group II Intron" category. If an element contained a gene with a predicted serine/tyrosine recombinase domain (matching HMM profiles Pfam: Arm-DNA-bind_1, Pfam: Phage_int_SAM_1, Pfam: Phage_int_SAM_5, Pfam: Phage_integrase, Pfam: Recombinase, Pfam: Resolvase, TIGRFAMs: TIGR02224, TIGRFAMs: TIGR02225, or Pfam: Zn_ribbon_recom at e-value < 1e-4), it was assigned to the "Serine/Tyrosine Recombinase" category. If an element contained other CDS regions and TIRs, it was assigned to the "Contains CDS + TIR" category. If an element did not contain any predicted CDS regions but it did contain TIRs, it was assigned to the "No CDS + TIR" category. If an element contained non-transposase CDS regions but no TIRs, it was assigned to the "Contains CDS" category. If an element contained no CDS regions and no TIRS, it was assigned to the "No CDS" category.

It should be noted that these categories were forced to be mutually exclusive, and that in reality these elements categories can overlap. Many predicted phages carry serine recombinases, for example. Additionally, elements containing conjugation systems were identified using ConjScan (Abby et al., 2016) through the Galaxy web server (Afgan et al., 2016). An element was considered to carry a conjugation system if it contained a predicted mobilization protein (MOB), a coupling protein (T4CP), and VIRB4, the only ubiquitous member of type IV secretion systems. The approach taken to identify recombinases, group II introns, and conjugation systems was inspired by a previous study (Jiang et al., 2017).

Identifying a Set of High-Confidence Mobility Genes

Prokka predicted genes in all identified elements were analyzed to identify a set of high-confidence mobility genes. All predicted proteins across species were clustered at 50% similarity using CD-HIT (Fu et al., 2012), using the same parameters that are used to cluster the UniRef50 database (Suzek et al., 2015). These protein clusters were assigned to a DNA element cluster if they appeared in over 75% of all unique sequences in the cluster. The number of unique sites (within and across species) where each protein cluster was found was calculated. All protein clusters that appeared in over 10 unambiguous sites were considered "mobility genes." Mobility genes include many different elements, including transposases, recombinases, and phage-related proteins. For the purposes of Figure 3A, mobility genes were designated as "Transposases" if the ISEScan HMM profiles identified them as such (Xie and Tang, 2017). Mobility genes were designated as "phage-related proteins" if they were not predicted transposases, and if they appeared in predicted phage sequences a plurality of the time. Mobility genes were designated as a "Group II intron protein" or "Recombinase" if they had significant HMM profile matches to domains of these proteins (e-value < 1e-4). Mobility genes were designated as a "Conjugation system protein" if they had a significant match to an HMM profile curated by ConjScan (Abby et al., 2016). Genes were described as "Other annotated" if Prokka assigned the genes to a described gene in their database. Genes were described as having "unknown function" if Prokka only predicted them to be hypothetical proteins with no assigned function. This set of high-confidence mobility genes was then used as markers to further characterize identified elements.

Describing MGEs with Annotated Passenger Genes

Figure 3C was generated to summarize the annotated passenger genes that were contained within predicted MGEs. If any member of a given element cluster was found to contain an annotated gene, then the entire cluster was predicted to contain the identified gene. The number of insertion sites where the predicted MGE at the specified site contained the annotated gene of interest was calculated. The passenger genes that appeared at the most unique locations throughout each organism's genome (excluding those found in phage elements) are presented in Figure 3C.

Annotating Reference Genomes

To ensure comparable and consistent gene-calling among different isolates, each organism's reference genome was annotated using Prokka v1.13 used with default settings. The gene names identified by Prokka were used to describe genes in Figures 4A and 7A (Table S6).

Analysis of Insertion-Enriched Sites

Regions of the genome with high numbers of unique insertions are described in this study as insertion-enriched sites. The approach taken here resembles approaches taken by ChIP-seq peak calling algorithms, such as MACS (Feng et al., 2012). All unique insertions found throughout each organism's genomes were extracted. Sliding windows of 500 bps, with 50 bp steps, were created across each organism's genome. The number of insertions found within each window was calculated using bedtools (Quinlan and Hall, 2010). To determine if a given window was enriched for insertions, a Poisson distribution was used to model the insertion distribution, with a dynamic parameter λ_{local}. This parameter is calculated as

$$\lambda_{local} = \max(\lambda_{BG}, \lambda_{1kbp}, \lambda_{10kbp})$$

where the term λ_{BG} describes the estimated rate of insertion calculated across the entire genome, λ_{1kbp} is the estimated rate of insertion within 1 kbp of the window being tested (500 bases on either side of the window), and λ_{10kbp} describes the estimated rate of insertion within 10 kbp of the window being tested. The estimated value of λ_{local} is then used in a one-sided exact Poisson test to determine if the observed insertion rate for the window exceeds the expectation, using the "poisson.test" function in R. Adjusted for the local insertion rate, this should account for biases in the local insertion rate, identifying windows that are significant above this background level.

Significant insertion-enriched sites were then calculated as windows that met FDR $\leq$ 0.05. Once these significant insertion-enriched sites were identified, overlapping insertion-enriched sites were merged into single regions. These insertion-enriched sites were then associated with nearby genes. This was done by first restricting the edges of each insertion-enriched site to begin and end at the exact site of the first and last insertions that they contain, effectively tightening the region edges to directly surround their insertions. The center point of each region was calculated and used to associate this region with the nearest gene. This resulted in a list of significant insertion-enriched sites associated with specific genes in the reference genome.

Insertion-Enriched Sites near Homologous Genes within and across Species

To determine if homologous genes within and across species were found near MGE insertion-enriched sites, all protein sequences that were mapped to insertion-enriched sites previously were extracted. These proteins were then clustered using the CD-HIT algorithm at 40% sequence identity across 70% of the sequence. If protein sequences clustered with each other across species, they were considered cross-species insertion-enriched sites. If multiple protein sequences near insertion-enriched sites from the same species clustered with each other, they were considered within-species insertion-enriched sites.

Insertion-Enriched Site Gene Ontology Enrichment Analysis

This approach was taken to determine if the genes near insertion-enriched sites were enriched for any particular function. Gene ontology (GO) terms were used for this purpose. To map genes to gene ontology terms, the Prokka-predicted protein sequences were mapped to the UniRef90 database (Suzek et al., 2015) using DIAMOND (Buchfink et al., 2015) with an e-value cutoff of 10^{-5}, choosing the top result as the representative. The database identifier mapping ('Retrieve/ID mapping') service provided by UniProt was used to map the IDs of each protein to all other proteins that clustered at the level of UniRef90, and then mapped these proteins to the GO terms associated with them in the UniProt database. In other words, the GO terms associated with each annotated protein include all of those assigned to it or any homologs that clustered with it in the UniRef90 database. All of the coding sequences near an insertion-enriched site were taken, and enriched GO terms were identified using the hypergeometric test, with all proteins with at least one GO term of any type used as a background. Only GO terms containing five or more genes, with two or more of these genes being found near an insertion-enriched site, were tested. A significance cutoff of FDR $\leq$ 0.05 was used to determine significant GO enrichments.

Target-Sequence Motif Discovery

The HOMER motif analysis tool was used to identify target-sequence motifs for MGEs with 10 or more different insertion sites (Heinz, 2010). HOMER findMotifsGenome.pl was executed with the default parameters, and searched for motifs of size 4, 6, 8, 10, and 12 within the reference of genome of each species. Randomly selected sequences from the reference genome were used as background sequences. All *de novo* motifs that met a p value cutoff of 1e-12 were reported, choosing the most significant motif for each MGE.

Analysis of Baym et al. Megaplate Experiment Sequencing Data

The large insertions found in the sequenced isolates from the intermediate-step trimethoprim megaplate experiment conducted by Baym et al. were analyzed. This included 231 sequenced isolates in total; of note, the sequenced isolates had widely varying sequence coverage. The sample sequencing coverage was calculated as the median sequencing coverage across the entire genome. Across all 231 isolates, the median isolate was covered by 19 reads, with 31% of isolates having median coverage of less than 15, and 5.6% being covered at a genome-wide median coverage of 0. As was done in the original publication, all sequenced isolates were analyzed, including these low-coverage samples, to see if any insertions could be identified. But it should be noted that, due to low coverage, not all insertions in all samples could be identified. This range of sequencing coverage may explain some of the unexpected patterns of inheritance observed in Figure 6A.

The samples in this experiment were analyzed with the complete insertion identification workflow. If an isolate was found to have coverage lower than 20, more lenient parameters were used by MGEfinder to identify insertions (See Identifying candidate insertion sites).

Baym et al. (Baym et al., 2016) inferred the pattern of inheritance from watching a video recording of the migrating bacterial front as it grew across the megaplate. Their visually inferred phylogeny was used to determine how many independent insertions occurred across all sequenced samples in this re-analysis of their data. In most cases, the inferred relationships between isolates was accurate; however, in a subset of cases, it appears that the inferred relationships between isolates were inaccurate. Thus, rather than

relying completely on the given phylogeny, the number of independent insertions was estimated conservatively. A visualization of the independent insertion events that were identified is provided in Figure S6.

Analysis of Toprak et al. Morbidostat Experiment Sequencing Data

Data from the trimethoprim, chloramphenicol, and doxycycline morbidostat experiments conducted by Toprak et al. were re-analyzed in this study. This included 20 samples in total, with 5 replicates per experiment, the wild type reference, and four additional samples that represented four additional colonies that grew from plating three of the replicates. All samples were covered at a median coverage between 21 and 30, suggesting that there was sufficient sensitivity to detect most insertions in all samples. These samples were sequenced using a single-end sequencing approach, which MGEfinder was able accommodate with some minor changes in the workflow.

These samples were analyzed using the same workflow used in the previous megaplate analysis, with a minimum clipped read count of 4 to support each insertion site. Independent mutations were identified and cross-referenced with genome annotations to determine the genes that were most likely impacted by each insertion. It should be noted that Toprak et al. (Toprak et al., 2011) were very stringent when initially calling SNPs and indels from their datasets, requiring validation by Sanger sequencing. The IS insertions found in this study were not subject to the same strict standards of evidence.

Inferring Ancestral States and Identifying Independent IS Mutation Events in Clinical Isolates

Using the insertions identified in the Hospital and MDR collections, we sought to infer the ancestral state of each insertion to determine the number of independent insertion events. The phylogenetic tree was rooted using an *Escherichia fergusonii* isolate, and multitomies were randomly resolved. The rerooting method developed by Yang et al. (Yang et al., 1995) was used through the "rerootingMethod" in the R package phytools, version 0.6.44. This technique uses maximum likelihood to estimate the marginal ancestral state for each node in the phylogenetic tree. We initialized isolate states for each insertion, with ambiguous states being assigned to any isolate where we did not detect the insertion and the number of reads at either the 5′ or 3′ clipped site being less than 10. After running this method, any node with a marginal probability of insertion greater than 0.5 was considered to contain the insertion.

We then sought to quantify how many times a given gene was disrupted by an independent insertion. For all of the insertions that overlapped with a given gene, we traversed the tree upward from each leaf node with at least one insertion until we reached a node whose parent was predicted to not have an insertion in that gene. We then used the total number of such intermediate nodes as the estimate for the number of independent insertions in that gene within the isolate collection. In the event that two insertions exist in the same gene in a single isolate, the approach would only count this as a single insertion event, since one event or the other would have been the initial nonsense mutation.

For genes disrupted by insertions and/or nonsense mutations, the aim was to infer which of the two mutations occurred first ancestrally. This was done using the same phylogenetic inference approach we used to identify independent insertions, but we also included other nonsense mutations identified by the SNP calling tool FreeBayes (Garrison and Marth, 2012). For each gene disrupted by an MGE insertion or other nonsense mutation, we traversed up the phylogenetic tree until we reached a node whose parent had no predicted MGE insertion or nonsense mutation. If the mutation at that node was found to be an MGE insertion, it was considered to be mutated initially by a MGE insertion, and likewise for other nonsense mutations. The proportion of genes disrupted by MGE insertion compared to other nonsense mutations was then calculated and visualized in Figure S7E.

QUANTIFICATION AND STATISTICAL ANALYSIS

All details of statistical analysis and software can be found in the method details, which we summarize here briefly. Statistical analyses were all conducted in the R programming language. Rarefaction curve analysis was done using the vegan package in R. For the rare TE insertion analysis, ANOVA was used to determine how species, average nucleotide identity (ANI) to reference, average pairwise ANI to population, sequencing read length, sequencing depth, and sequencing fragment length influenced the number of rare TE insertions detected for each species. Analysis of insertion-enriched sites was carried out using a procedure that was similar to MACS2 ChIP-seq peak caller. Briefly, a dynamic p value was calculated for each window of unique insertions using the Poisson distribution as a null, with the expected rate being the maximum rate of nearby genomic windows, or the whole genome. All p values were adjusted using FDR correction across all tested genomic windows. Enrichment of insertion-enriched sites near Gene Ontology pathways was performed using two-sided hypergeometric tests.

DATA AND CODE AVAILABILITY

The MGEfinder command-line toolbox is hosted on GitHub at https://github.com/bhattlab/MGEfinder. A detailed README and test dataset are included. The toolbox and all dependencies can be installed using conda (https://anaconda.org/mdurrant/mgefinder).