Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools

Xiaofang Jiang[1,2‡], Andrew Brantley Hall[2,3‡], Ramnik J. Xavier[1,2,3,4], Eric J. Alm[1,2,5]*

1 Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, United States of America, **2** Broad Institute of MIT and Harvard, Cambridge, MA, United States of America, **3** Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States of America, **4** Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States of America, **5** MIT Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States of America

‡ These authors are co-authors on this work.
* ejalm@mit.edu

Abstract

Mobile genetic elements (MGEs) drive extensive horizontal transfer in the gut microbiome. This transfer could benefit human health by conferring new metabolic capabilities to commensal microbes, or it could threaten human health by spreading antibiotic resistance genes to pathogens. Despite their biological importance and medical relevance, MGEs from the gut microbiome have not been systematically characterized. Here, we present a comprehensive analysis of chromosomal MGEs in the gut microbiome using a method that enables the identification of the mobilizable unit of MGEs. We curated a database of 5,219 putative MGEs encompassing seven MGE classes called ImmeDB. We observed that many MGEs carry genes that could confer an adaptive advantage to the gut environment including gene families involved in antibiotic resistance, bile salt detoxification, mucus degradation, capsular polysaccharide biosynthesis, polysaccharide utilization, and sporulation. We find that antibiotic resistance genes are more likely to be spread by conjugation via integrative conjugative elements or integrative mobilizable elements than transduction via prophages. Horizontal transfer of MGEs is extensive within phyla but rare across phyla, supporting phylum level niche-adaptive gene pools in the gut microbiome. ImmeDB will be a valuable resource for future studies on the gut microbiome and MGE communities.

Introduction

Horizontal gene transfer (HGT), the transfer of genes between organisms by means other than vertical transmission, allows for the rapid dissemination of genetic innovations between bacteria[1]. Ecology is an important factor shaping HGT, and the human gut in particular is a

hotspot for HGT[2,3]. HGT impacts public health through its role in spreading antibiotic resistance genes[4,5]. The biological importance of HGT is exemplified by a porphyranase identified in Bacteroides plebius that digests seaweed, which was horizontally transferred from marine bacteria to human gut bacteria[6]. However, a major contributor to horizontal transfer —mobile genetic elements (MGEs)—have not been systematically characterized in the human gut microbiome.

Canonical classes of MGEs includes prophages[7], group II introns[8], and transposons[9]. It has become increasingly apparent that the acquisition of a novel element class, genomic islands correspond to HGT events that differentiate commensal and pathogenic strains[10]. Genomic islands are non-canonical classes of MGEs that can transfer by conjugation or genomic regions derived from such MGEs. Integrative conjugative elements (ICEs) are a type of genomic island that can integrate into and excise from genomes using integrase, circularize using relaxase, replicate, and then transfer via conjugation[11,12]. Integrative mobilizable elements (IMEs) encode an integrase and relaxase for circularization like ICEs, but they have to hijack the conjugative machinery of co-resident ICEs or conjugative plasmids[13].

Conventionally, HGTs are computationally identified by searching for the inconsistencies in the evolutionary history of genes and species[14]. However, this method overlooks the fact the horizontal transfer of multiple genes from the same locus might be the result of a single HGT event. Rather than individual genes, it is critical to identify the mobilizable units, in other words, the entire sequence of MGEs. Determining the mobilizable unit of MGEs is crucial to identify the mechanism of transfer, the preference of insertion sites, and cargo genes as well as to track the frequency of horizontal transfer events. In addition, information on MGEs are also valuable in the context of metagenomic analysis, as MGEs confound many metagenomics workflows such taxonomic profiling, strain-level variation detection, and pangenome analysis.

The repetitive and mobile nature of MGEs confounds many types of studies in microbiome communities, such as taxonomic profiling, strain-level variation detection, and pan-genome analyses. However, unlike research in eukaryotes, where multiple repeats databases exist for masking and annotation of repetitive DNA[15], only a limited number of databases dedicated to the collection of MGE in prokaryotes[16–19]. Yet, these database are either limited one specific class of MGE or obsolete and not applicable for microbiome research. With the growing deluge of microbiome metagenomic sequencing data, a comprehensive MGE database of the gut microbiome is becoming increasingly critical.

In this study, we sought to characterize MGEs from the gut microbiome to understand how horizontal gene transfer by MGEs shapes the evolution of bacteria in the gut microbiome. First, we developed a method to identify the insertion unit of MGEs using whole metagenome sequencing data together with references genomes. Similar approaches have been used to detect well-known pathogenicity islands[20]. The algorithm implemented in SRID (Split Read Insertion Detection) is similar to that of Daisy[21], the first mapping-based HGT detection tool to our knowledge. We systematically identified MGEs with SRID and curated a database named ImmeDB (Intestinal microbiome mobile element database) dedicated to the collection, classification and annotation of these elements. The database is organized into seven MGE classes which can be searched using both text and BLAST. Each MGE entry provides a visualization of annotations and downloadable genomic sequence and annotations. We detected many MGEs carrying cargo genes that confer an adaptive advantage to the gut environment. This study provides insights into how the interplay of MGEs, bacteria, and the human host in the gut ecosystem lead to community-wide adaptations to the gut environment. The curated database of MGEs we have assembled here can be used by metagenomic workflows to improve future microbiome studies.

Materials and methods

Detection of putative insertions in reference genomes

A set of reference genomes of human gut/fecal isolates was curated from the NCBI (07-01-2018) based on metadata available from BioSamples, PATRIC, as well as literature searches. The quality of the genomes were assessed with CheckM [22]. Genomes with less than 90% completeness or more than 10% contamination were removed. Our target database included 5,103 genomes representing 1,896 taxa (834 species taxa). 274 stool metagenomes from the Human Microbiome Project (HMP)[23] were downloaded from the NCBI. We used Mash[24] to calculate the minhash distance between each genome and each metagenomic sample with the default sketch size of $s = 1000$ and $k = 21$. Metagenomic reads from HMP samples were aligned to each genome with more than 2 matching-hashes shared with the reads separately with bwa (version 0.7.12)[25]. To find genomic regions that differ in terms of insertions/deletions between strains in the individual samples and the reference genomes, we used split reads and information from pair-end reads from the alignment (Fig 1A). First, we identified putative deletion junctions using split reads, which we defined as reads that align to two distinct portions of a genome. Split reads were initially identified as those reads having multiple hits in the SAM output from bwa. If a split read alignment starts at one genomic location in the reference and then "jumps" to aligning to a distant site downstream in the same strand, it may indicate a potential deletion in the strain of bacteria from the metagenomic sample compared to the reference genome. For each putative deletion junction, we confirmed the presence of the junction by determining if paired-end reads flanked the junction. We considered a deletion junction to be valid if the read pairs flanking the junction were aligned in the correct orientation, and the distance between the pairs minus the junction size is within the range of +/- 2 times the standard deviation of the mean library insert size. Regions with more than ten split reads and more than ten read pairs supporting the deletion were considered as putative insertions in the reference genome. We selected for regions ranging in size between 1kbp and 150kbps to reduce the number of spurious results. The SRID method aimed to identify as many candidate MGEs as possible, especially in low-abundance species in metagenomic sequencing data. Therefore, there are no stringent requirements for filtering based on read or mapping quality at this stage. The putative insertions identified in the reference genomes were verified to be MGEs via additional stringent filtering steps detailed later in the methods. The code used to implement the SRID method, human gut related genome assembly accession numbers and HMP SRA accession numbers used in this study are available from GitHub (https://github.com/XiaofangJ/SRID) and the S2 File.

MGE signature detection

Genes from the genomes with putative insertions identified were predicted with Prodigal (version 2.6.3)[26]. Protein sequences were functionally annotated with interproscan (version 5.30–69.0) using the default settings[27]. Then, we used the interproscan annotations to identify serine and tyrosine integrases as well as group II intron proteins from all genomes. Prophage-related genes were identified by searching for genes with phage-related or phage-specific Pfam signatures. The list of phage signatures used were a combination of those identified in phage_finder[28] as well as a list of manually curated terms (S2 File). Serine integrases were identified as genes annotated with one of the Pfam identifiers: PF00239 (Resolvase: resolvase, N terminal domain), PF07508 (Recombinase: recombinase), and PF13408 (Zn_ribbon_-recom: Recombinase zinc beta ribbon domain). Tyrosine integrases were identified as genes annotated with one of the identifiers: PF00589 (Phage_integrase: site-specific recombinase,
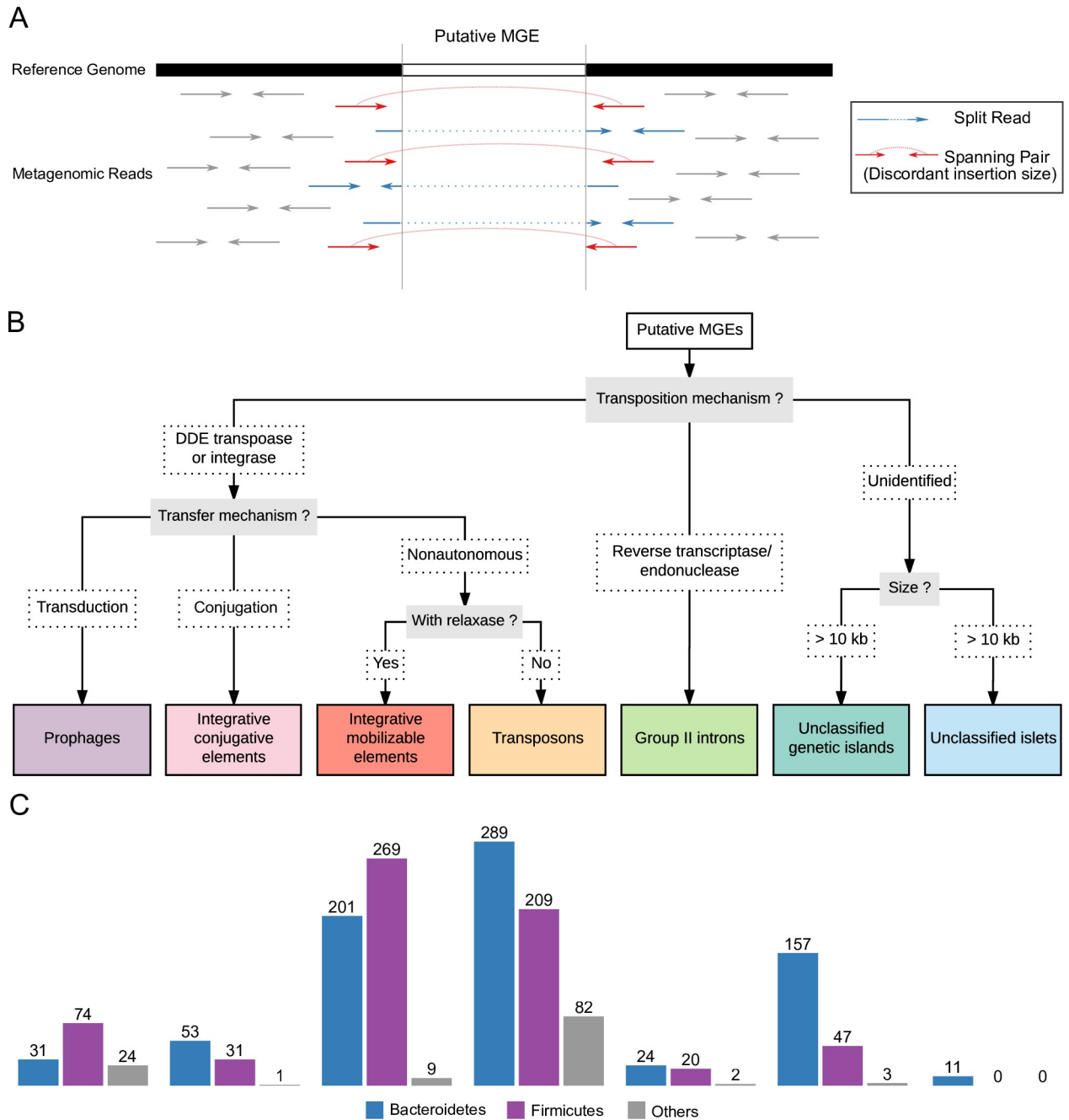
Fig 1. Identification and classification of gut microbiome MGEs. (A) The method used to identify putative MGEs using split reads and discordantly-mapped paired-end reads. Split reads are colored blue, and discordantly-mapped paired-end reads are colored red. (B) The method used to classify MGEs based on gene signatures. (C) The number of MGE clusters identified stratified by phyla and MGE classification.

prophage integrase family), PF02899 (Phage_integr_N: prophage integrase, N-terminal SAM-like domain), PF09003 (Phage_integ_N: bacteriophage lambda integrase, N-terminal domain), TIGR02225 (recomb_XerD: tyrosine recombinase XerD), TIGR02224 (recomb_XerC: tyrosine recombinase XerC), and PF13102 (Phage_int_SAM_5: prophage integrase SAM-like domain). Group II intron proteins were identified as genes annotated with the identifier: TIGR04416 (group_II_RT_mat: group II intron reverse transcriptase maturase). To identify

genes involved in mobilization and conjugation of MGEs, we used ConjScan via a Galaxy web server (https://galaxy.pasteur.fr/)[29]. We identified transposases using blastp against the ISFinder database with an e-value 1-e10[17]. The best hit for each protein was used to annotate the family of transposases.

Identification and classification of MGEs

MGEs are typically identified by two methods: the presence of MGE signatures[16,30], and evidence of phylogenetic incongruence such highly-identical copies of same element present in multiple species[2]. We used both methods in our study. Putative insertions were annotated as an ICE if they contained complete conjugation and relaxase modules and an integrase or transposase at the boundary of the element. Putative MGEs were annotated as prophages if there is an integrase or transposase at the boundary of the element and more than five genes were annotated with prophage-related Pfams. Putative MGEs were annotates as IMEs when no VirB4 is present and when an integrase is found in the vicinity of a relaxase and did not contain genes involved in conjugation. Putative MGEs were annotated as transposons if they contained transposase and were not previously annotated as an IME. We limited the size of IMEs to 30kb and transposons to 10kb to decrease the number of false positives. Putative MGEs were annotated as group II introns if the element was less than 10kb, contained a protein with the TIGR04416 signature, and did not contain a gene annotated as transposase. The remaining putative MGEs were then divided into two groups based on their sizes: unclassed genomic islands (>10kb), and islets (<10kb). To eliminate spurious MGEs, we only report genomic islands that contain an integrase or transposase, or those that are related to prophage/ICEs, and islets that exist in more than two species with identity higher than 98%. We confirmed that the SRID method accurately identified known MGEs and their boundaries using both synteny and comparative genomics for CTnDOT and Tn916/Tn1549. To ensure the identified boundary of identified MGEs is correct, we manually inspected the elements using additional methods, including comparative genomics and visualization of alignments of the reads supporting the insertion (Figure B in S1 File). After classification and verification, we identified 5,219 MGEs in 1,199 genomes (S2–S4 Files). The MGEs identified with the SRID method are limited to chromosomal MGEs. Thus, plasmids and extrachromosomal prophages were not characterized in this study.

Clustering each class of MGEs

Pairwise alignment of elements from the same class of MGEs was performed with nucmer (version 3.1)[31]. Elements with more than 50 percent of the sequence aligned to each other are grouped in the same cluster. For ICEs, we additionally require that elements in the same cluster should have the same types of integrase, relaxase and conjugation modules. For IMEs, we required that each cluster has the same types of integrases and relaxases for all elements. For transposons, the same cluster should have the same type and number of IS genes. If a transposon is a "nested" or composite transposon, the family names of all IS contained within were used to annotate the transposon.

Construction of phylogenetic trees

To build phylogenetic trees of ICE and prophage integrases, we selected a representative integrase sequence from each cluster. Representative sequences were chosen as examples with the highest identity to their homologs in the same cluster. We performed alignment of each group of sequences with mafft(v7.123b)[32] (parameter "—maxiterate 1000"). We used trimal (version 1.4.rev15)[33] to remove region with gaps representing more than 20% of the total

alignments (parameter "-gt 0.8"). RAxML(version 8.2.10)[34] was used to build phylogenetic trees from the alignments using the LG substitution matrix and a gamma model of rate heterogeneity (parameter "-m PROTGAMMALGF"). Phylogenetic trees were plotted with the R package phytools[35].

Functional enrichment analysis of cargo genes

Cargo genes were identified by excluding genes involved in transposition and transfer from all genes on MGEs. To understand the function of cargo genes, we performed enrichment analysis based on gene ontology (GO), antibiotic resistance (Resfam), and protein families (Pfam). The enrichments were performed with all genes present in the genomes as background reference. We used hmmer(version 3.1b2) [36] to search Resfam[37] core database to annotate antibiotic resistant gene. The "—cut_ga" parameters were used to set the threshold. The best hits to each gene from the Resfam database were used to annotate antibiotic resistant genes. GO terms and Pfam signature of the same genes sets were extracted from interproscan result. R package GOStat[38] was used for GO enrichment analysis for GO and Pfam. The R package clusterProfiler[39] was used for the enrichment analysis of cargo genes based on Resfam and Pfam signatures. P-value of 0.05 were used as cutoff for all enrichment analysis. Samples size, p-value, and exact test statistics of these tests are provided in S4 Table.

Results

Prevalence of MGEs in species of the gut microbiome

We systematically identified MGEs from species of the human gut microbiome using mapping information from metagenomic reads from the Human Microbiome Project (HMP)[23]. MGEs are actively inserted and deleted from genomes, causing differences between strains of bacteria. We identified cases where the reference genome of a bacterial strain differed from strains in the individual samples from the HMP. To find the sequences responsible for these differences, we mapped HMP metagenomic reads to available gut-associated bacterial reference genomes and identified genomic regions flanked by split reads and discordantly-aligned paired-end reads (Fig 1A). These regions potentially are recent insertions of active MGEs. The MGEs identified with the SRID method are limited to chromosomal MGEs. Thus, plasmids and extrachromosomal prophages were not characterized in this study. By searching for MGE-specific gene signatures, we manually verified and classified these MGEs (See Fig 1B and Methods).

We identified 5,219 putative MGEs from gut microbiome representatives of 101 strains of Actinobacteria (7 species), 283 strains of Bacteroidetes (95 species), 214 strains of Firmicutes (149 species), 574 strains of Proteobacteria (13 species), and 27 strains of Verrucomicrobia (3 species) (S2 Table). Then, we classified the identified MGEs based on their transfer and transposition mechanisms into seven classes: ICEs, prophages, IMEs, group II introns, transposons, unclassified islets, and unclassified genomic islands (Fig 1C). Most of the MGEs identified (4113/5219) were from the phyla Bacteroidetes and Firmicutes because these two phyla tend to dominate the gut microbiome of healthy adults[23] (Figure A in S1 File).

Different strains of the same species often share identical or nearly-identical MGEs. To eliminate this redundancy, we collapsed MGEs into clusters based on overall nucleotide identity (Fig 1C). Phylum-level differences in the diversity of MGEs were revealed. For example, Bacteroidetes had more diversity of ICEs than Firmicutes (53 vs. 31 respectively), while Firmicutes had more diversity of prophages than Bacteroidetes (71 vs. 31 respectively).

Most of our understanding of genomic islands such as ICEs comes from pathogens and a small number of tractable model species. In gut species, the knowledge of ICEs is limited to a

few families of ICEs such as CTnDOT and Tn916/Tn1549. In this study, we successfully identified exact copies CTnDOT with the correct boundary in multiple species (ICE cluster 1), and several Tn916/Tn1549 family elements using the SRID method. Due to limited knowledge of genomic islands in gut microbes, most genomic islands described in this study are novel.

Diversity of MGE modules in gut microbiota

Although it has been known that ecology is important in shaping the diversity of MGEs in the gut microbiome, this study is the first to systematically characterize the putative mechanisms of transposition and transfer for MGEs of the gut microbiome[2]. We annotated the genes in MGEs involved in their transposition and transfer, and then classified the elements into groups based on these annotations (S2 Table).

There are four major protein families responsible for transposition of gut MGEs: serine integrases, tyrosine integrases, transposases, and group II intron proteins conferring reverse transcriptase and endonuclease activity. Serine and tyrosine integrases are the most prevalent protein families responsible for transposition in ICEs, IMEs, and prophages. In the gut microbiome MGE clusters we identified, we found 482 MGEs with tyrosine integrases (64 from ICEs, 285 from IMEs and 133 from prophages) and 237 MGEs with serine integrases (21 from ICEs, 184 from IMEs and 32 from prophages). Interestingly, while tyrosine integrases are found in several phyla, serine integrases of ICEs and prophages were exclusively found in the phylum Firmicutes. In IMEs, most serine integrases were identified in Firmicutes (171), but 18 clusters of serine integrases were found in Bacteroidetes, Verrucomicrobia and Actinobacteria (12,5 and 1 respectively). Unlike prophages and ICEs, 12 of 197 clusters of IMEs use serine integrases to transpose in Bacteroidetes. This implies that although integration via serine integrases occurs in Bacteroidetes, it occurs much less frequently than integration via tyrosine integrases. No ICEs and IMEs with transposase were identified in our study. Twelve prophage clusters were found with transposase from IS families: IS30, IS256, and IS110.All transposons we identified utilized DDE/DED/DDD-transposase. We identified 23 families of transposase. Most of the transposase clusters we identified are present in insertion sequences. Thirty-nine clusters (164 copies) of transposons are composite transposons flanked by two different insertion sequences families.

ICEs and IMEs encode relaxases (MOB) to initiate DNA mobilization and transfer. We used the CONJscan-T4SSscan server to classify relaxases identified in MGEs[40]. Six types of relaxase were identified in ICEs and IMEs. In ICEs, MOB_T was identified only in Firmicutes, MOB_V was identified only in Bacteroidetes, and MOB_{P1} was identified in Firmicutes, Bacteroidetes, and Actinobacteria. IMEs have a more diverse reservoir of relaxases. Besides the three types of relaxase found in ICEs, we also identified IMEs with MOB_{P3}, MOB_B, and MOB_Q type relaxases.

ICEs are capable of conjugation via mating pair formation systems. Six types of mating pair formation systems for conjugation have been described[40]. We found three types of mating pair formation system: typeB, typeFA, and typeFATA, in ICEs from the gut microbiome. Consistent with previous findings, type FA systems were identified in 6 ICE clusters from Firmicutes, type B systems were identified in 53 ICE clusters from Bacteroidetes, and type FATA systems were identified in 25 Firmicutes ICE clusters and one Actinobacteria ICE cluster[41].

MGEs carry niche-adaptive genes

Although fundamentally selfish, MGEs often carry genes other than those necessary for their transposition and transfer, sometimes referred to as cargo genes[42]. We found that smaller elements like transposons generally carry zero or only a few cargo genes. Genetic islands like ICEs

and IMEs often carry numerous cargo genes (median cargo genes 40 and 11 respectively). One example is an ICE found in *Bacteroides fragilis* 3_1_12 (NZ_EQ973213.1:1794893–1944846) which carries 143 cargo genes. We performed functional annotation on the cargo genes, and enrichment analysis using gene ontology (GO), Pfam, and Resfam[37,43,44] (S4 Table). Several classes of enriched genes are well-known to be associated with the maintenance of MGEs such as restriction-modification systems (GO:0009307,PF07669) and toxin-antitoxin pairs (PF02452, PF07927,PF15738) (S4 Table). Many other gene families carried by MGEs may confer an adaptive advantage to colonize the gut.

Antibiotic resistance genes

Many classes of antibiotics consumed orally are incompletely absorbed in the small intestine, and therefore proceed to the large intestine where they can kill the resident microbes[45]. Therefore, genes that confer antibiotic resistance can be adaptive to the gut environment. In total, we identified 246 antibiotic resistance genes encompassing 19 distinct classes carried by MGEs. Classes of MGEs varied in their carriage of antibiotic resistance genes. Of 14,632 pro-phage cargo genes, only 2 were found to be antibiotic resistance genes. The carriage rate of antibiotic resistance genes normalized by total cargo genes in prophages is more than thirty times lower than that identified in ICEs (140/20269) and IMEs (66/9302) (S1 Table). This sug-gests that conjugation via ICE/IME may be more important than transduction in the spread of antibiotic resistance genes, consistent with previous findings[46,47]. GO analysis revealed that cargo genes involved in antibiotic resistance are enriched in both Bacteroidetes (GO:0015307; GO:0017001;GO:0008800") and Firmicutes (GO:0015307;GO:0016999;GO:0008811). Resfam enrichment analysis also supported this, as RF0135 (tetracycline resistance ribosomal protec-tion protein), and RF0067 (Emr 23S ribosomal RNA methyltransferase) were enriched. One example of an MGE responsible for the transmission of antibiotic resistance is the ICE CTnDOT, the spread of which dramatically increased the prevalence of tetracycline-resistant Bacteroidetes species[48]. CTnDOT-like ICEs were clustered in ICE1. Elements in this cluster typically confer resistance to tetracycline via the tetQ antibiotic resistance gene (Fig 2A). In addition, ICE1 elements have multiple sites where antibiotic resistance genes can be inserted or substituted. We characterized 5 insertions of antibiotic resistance genes into ICE1 (Fig 2A). Insertion sites 1, 2, and 5 are between operons; therefore they do not interrupt the function of crucial genes. We observed one insertion and two substitutions of antibiotic resistance genes around the tetQ operon, suggesting that this site is likely a "hotspot" for insertions and substi-tutions of antibiotic resistance genes. Our analysis reveals the surprising extent to which MGEs in species of the gut microbiome contribute to the phenomenon of antibiotic resistance and that the insertion of antibiotic resistance genes into MGEs is an active and ongoing process.

Bile salt hydrolase and bile transporters

Bile acids are found in high concentrations in the human intestines[49] and can be toxic to bacteria[50]. Therefore, gut microbes have developed strategies to deal with bile acids by actively pumping bile acids out of the cell, or via deconjugation, which is hypothesized to diminish the toxicity of bile acids[49,50]. The high identity of archaeal and bacterial bile salt hydrolases strongly suggests the horizontal transfer of this gene[51]. A sodium bile acid sym-porter family (PF01758), which could help to pump bile acids out of the cell, was found to be enriched in the cargo genes of MGEs. Furthermore, 18 examples of bile salt hydrolases were identified as cargo genes of MGEs (S3 Table). Thus, MGEs carry genes that help microbes to overcome a specific challenge of colonizing the human gut.

Glycoside hydrolases for mucus utilization

The colon is lined with a layer of mucus composed of the glycoprotein MUC2[52]. The glycans that decorate MUC2 have a core structure composed of galactose, N-acetylglucosamine, N-acetylgalactosamine, with terminal residues of fucose and sialic acid[53]. These specific glycans are a major energy source for members of the gut microbiota[54]. Therefore, it may benefit members of the gut microbiota to degrade these specific glycans[55]. We identified 43 glycoside hydrolases capable of degrading mucosal glycans carried by MGEs from the categories: sialidases (GH33), fucosidases (GH95), α-N-acetylgalactosaminidases (GH109), β-galactosidases (GH35) and β-hexosaminidase (GH20)[54,56] (S3 Table). Thus, MGEs carry genes to unlock a key energy source available to gut microbes.

Polysaccharide utilization loci

Gut Bacteroidetes can utilize a wide variety of polysaccharides via the products of polysaccharide utilization loci, which collectively make up large proportions of Bacteroidetes genomes[57]. Each polysaccharide utilization locus contains a copy of the gene SusC, a sugar transporter, and SusD, a glycan binding protein[58]. Due to the wide range of polysaccharides available to gut microbes, it is hypothesized that the possession of a large repertoire of polysaccharide utilization loci confers an adaptive advantage in Bacteroidetes[57]. We found 47 polysaccharide utilization loci containing both SusC and SusD carried by MGEs suggesting that the ability to degrade complex polysaccharides may be readily transferred between members of the gut microbiota (S3 Table).

Capsular polysaccharide biosynthesis loci

Many bacterial species produce capsules, an extracellular structure made up of polysaccharides [59]. However, gut Bacteroidetes species have a large repertoire of capsular polysaccharide
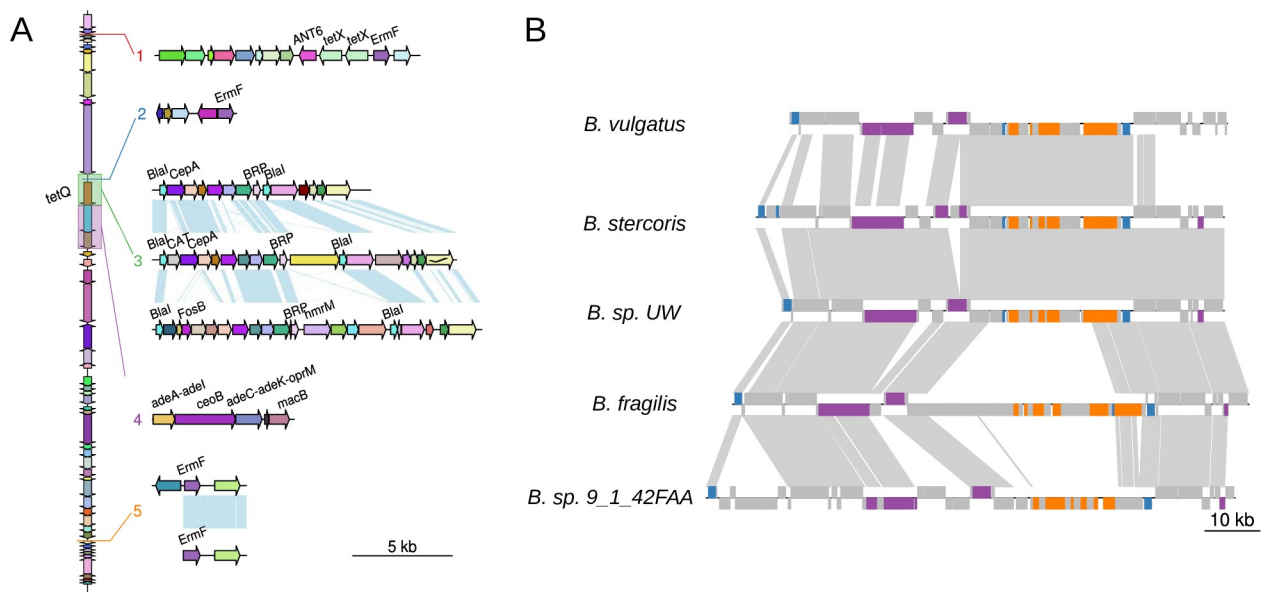

Fig 2. Examples of niche-adaptive genes. (A) CTn-DOT-like elements have acquired antibiotic resistance genes on multiple, independent occasions. Here, we show insertion sites of antibiotic resistant genes in CTnDOT-like elements. A CTnDOT-like ICE is shown on the left. Orthologs between elements are visualized using genoPlotR (light blue connections) and are the same color. Numbers in the top panel represent the insertion site of the numbered elements below. Antibiotic resistance genes are labeled. (B) ICEs are involved in the transfer of capsular polysaccharide biosynthesis loci between Bacteroidetes species. Here, we show examples of ICEs containing capsular polysaccharide biosynthesis loci. Orthologs between elements are plotted with GenoPlotR. Genes involved in capsular polysaccharide biosynthesis are colored orange, integrases are colored blue, and genes involved in conjugation are colored purple. Grey links indicate orthologs between elements.

https://doi.org/10.1371/journal.pone.0223680.g002

biosynthesis loci (up to 8) compared to other bacterial species and even Bacteroidetes from other sites such as the mouth[60]. Furthermore, capsular polysaccharide biosynthesis loci have been reported to be the most polymorphic region of *Bacteroides* genomes[61,62]. Multiple capsular polysaccharide biosynthesis loci are necessary to competitively colonize the gut, and are therefore considered to be gut adaptive genes in gut Bacteroidetes[63]. Capsular polysaccharide biosynthesis loci are large and complex; many contain upwards of 20 genes[59]. We found 22 complete or fragmented capsular polysaccharide biosynthesis loci containing at least 8 genes carried by MGEs (S3 Table). For example, almost identical copies of ICE9 containing a capsular polysaccharide biosynthesis locus were found in two species, *B. stercoris* and *B. sp. UW*. The same capsular polysaccharide biosynthesis locus was also found in *B. vulgatus*, but the ICE9 copy was slightly divergent. Two other copies of ICE9 likely containing an orthologous capsular polysaccharide biosynthesis locus were found in *B. fragilis*, *B. cellulosilyticus*, *B. vulgatus*, *Parabacteroides merdae* and *B. sp. 9_1_42FAA* (Fig 2B; S3 Table). Additionally, many GO-terms related to capsular polysaccharide biosynthesis are enriched in Bacteroidetes MGEs including GO:0033037, GO:0015774, GO:0008653, GO:0009103, GO:0045226, GO:0046379 and GO:0015159 The transfer of large segments of capsular polysaccharide biosynthesis loci by MGEs may help to explain the incredible diversity of capsular polysaccharide biosynthesis loci observed in the genomes of gut Bacteroidetes[64].

Sporulation

The gut is an anaerobic environment colonized by many classes of strictly anaerobic organisms[65,66]. However, to transmit between hosts, gut microbes must be exposed to oxygen. Recent work has shown that many more gut microbes form spores than previously thought, likely enabling transmission between hosts[67]. In Firmicutes, 30 genes involved in sporulation (GO:0030435) were found to be enriched in MGEs. In addition, PF08769 (Sporulation initiation factor Spo0A C terminal) and PF04026 (SpoVG) were also enriched in our Pfam analysis. One example is GI153, a genetic island from *Faecalibacterium prausnitzii A2-165*, which contains a series of spore formation-related genes in an operon: SpoVAC, SpoVAD, spoVAEb, gpr (spore protease), and spoIIP. Another example is GI175, a genetic island derived from a degenerate prophage in *Roseburia intestinalis* L1-82. In one operon of GI175, there are three genes: SpoVAEb, SpoVAD, and one unknown gene with Cro/C1-type HTH DNA-binding domain. SpoVAC, SpoVAD, spoVAEb homologs were previously found to be carried by a Tn1546-like ICE and conferred heat resistance to spores in the model spore forming organism *Bacillus subtilis*[68]. Thus, MGEs may help to transfer genes involved in sporulation between gut microbiota which may prove adaptive for colonizing new hosts.

Summary of cargo genes

Many additional gene families were found to be enriched in MGEs that could plausibly be niche adaptive including: histidine sensor kinases, and genes involved in vitamin B biosynthesis (S4 Table). Notably, MGEs from Firmicutes and Bacteroidetes have different types of genes enriched reflecting the differences in physiology between the phyla. Antibiotic resistance genes and genes involved in the detoxification of bile acids are enriched in MGEs from both phyla. Glycoside hydrolases for mucus utilization, and capsular polysaccharide biosynthesis loci are enriched only in MGEs from Bacteroidetes, while genes for sporulation are enriched in MGEs only from Firmicutes. Overall, the transfer of niche adaptive genes by MGEs likely has a large impact on the fitness of species of the gut microbiome.

Host ranges and evolution of MGEs

Although MGEs readily transfer between species, there has not been a systematic analysis of the host range of MGEs in the gut microbiome. The host ranges of different classes of MGEs is variable, and even within a class, different elements have variable host ranges. Understanding the host range of gut MGEs is of particular importance because gut MGEs carry many cargo genes, and the host range of the MGE defines how widely these cargo genes can be distributed. For example, the gut microbiome is a reservoir of antibiotic resistance genes, and many antibiotic resistance genes are located within MGEs[69]. Therefore, it is important to understand the probability of the transfer of MGEs with antibiotic resistance genes from commensals to pathogens[69].

First, we studied the host range of MGEs from the same cluster. MGEs in the same cluster that exist in at least two species generally represent recent horizontal transfer. Some MGE clusters are present in a wide range of species indicative of active horizontal transfer. One example is the ICE1 cluster, a representative of the CTnDOT-like ICEs, which is found in 35 species of Bacteroidetes from the genera: *Bacteroides*, *Parabacteroides*, *Allistipes*, and *Paraprevotella* (Figure A in S1 File). The entirety of the 49kb element is found at more than 99 percent nucleotide identical to 10 *Bacteroides*, *Parabacteroides*, *and Allistipes* species, indicative of very recent horizontal transfer. This cluster also includes other CTnDOT-like elements with more variability such as CTnERL, which has an additional insertion of an IME conferring erythromycin resistance[70]. Another example is the Firmicutes ICE cluster ICE10, which is found in 10 species of the families *Lachnospiraceae*. This ICE10 cluster belongs to Tn916/Tn1549 family of ICEs, some members of which carry the medically-important VanB gene conferring resistance to vancomycin[71]. We found no examples of ICEs from the same cluster present in multiple phyla. Clusters of prophage, IMEs, group II introns, and transposons were also found in many species but were again limited to a single phylum. Our results support that although the recent horizontal transfer of MGEs is common within phyla, cross-phyla horizontal transfer is rare, as we did not observe any cross-phyla horizontal transfer events for elements of the same cluster.

Here we generated phylogenetic trees of tyrosine and serine integrases from ICEs and prophages identified to study the evolutionary history of the recombination module of MGEs. To contrast the phylogeny of the tyrosine and serine integrases with host species lineages we plotted tanglegrams (Fig 3 and Fig 4). The phylogeny of both serine and tyrosine integrases is incongruent with the host species lineages which is indicative of extensive past horizontal transfer of ICEs and prophages between species of the gut microbiome.

The tyrosine integrases can be divided into two clades: the first is associated with the phylum Bacteroidetes, the second clade is associated with the phyla Proteobacteria, Actinobacteria, Firmicutes and Verrucomicrobia (Fig 3). Tyrosine integrases from Bacteroidetes show no evidence of close inter-phyla transfer but ancient transfers of tyrosine integrases between the phyla Proteobacteria, Actinobacteria, and Firmicutes likely occurred several times during evolution. Serine integrases from ICEs and prophages were only found in the phylum Firmicutes. Therefore, we found no evidence of inter-phyla transfer for ICEs and prophages with serine integrases suggesting a phylum-level restriction in host range.

We also examined whether integrases derived from ICEs and prophages segregated into clades based on element type. Previous studies on the phylogenetic relationships of integrases from ICEs and prophages did not find strong evidence of intermingling between ICE and prophage integrases[72,73]. In our phylogeny of the tyrosine integrases, ICE and prophage integrases are extensively intermingled, suggesting that ICEs and prophages have exchanged integrases multiple times over the course of evolution. Moreover, in our phylogeny of serine
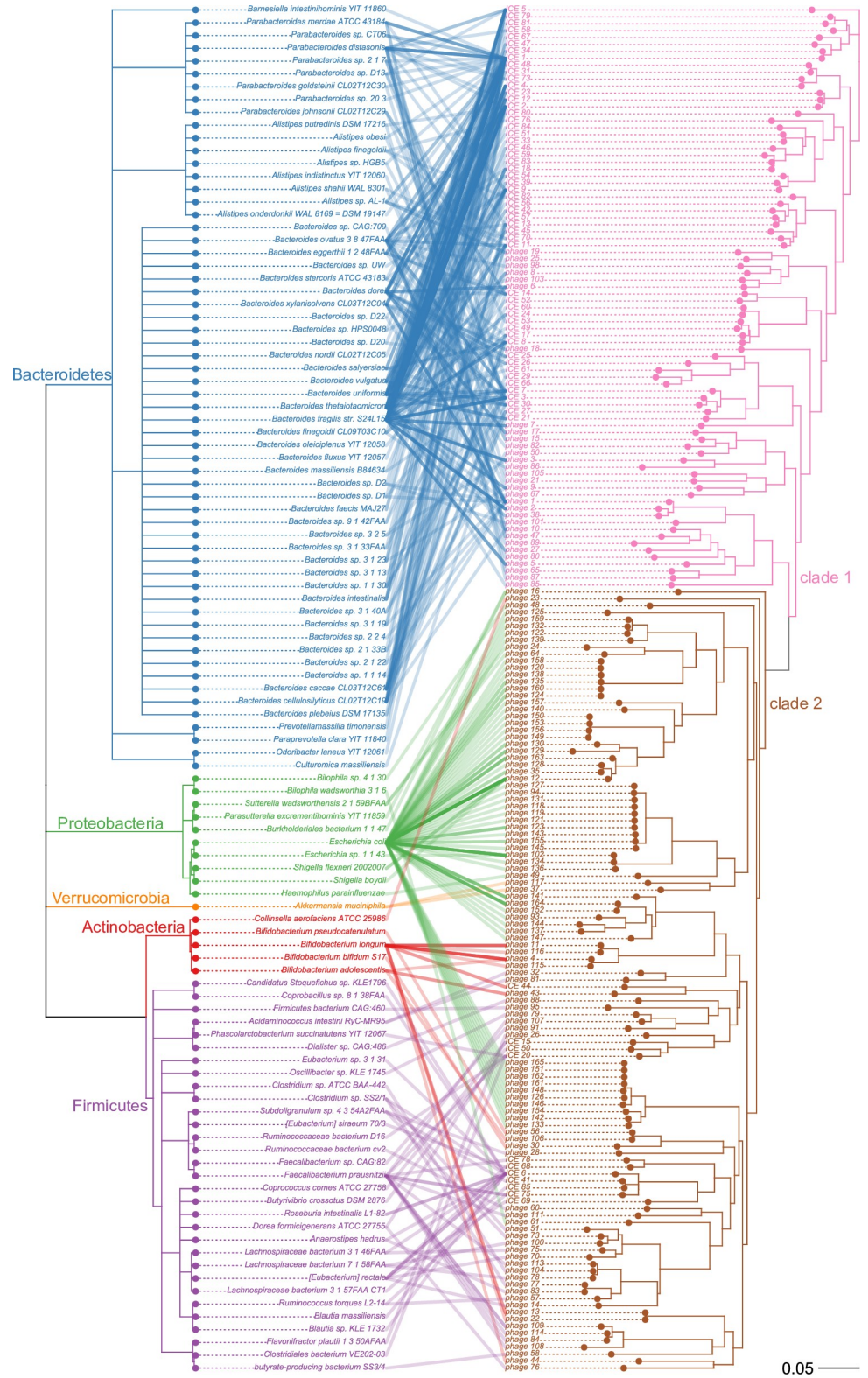
Fig 3. Tanglegram of host species lineages and phylogeny of the integrases in prophages and ICEs. A tanglegram of tyrosine integrases from ICEs and prophages with the species lineages plotted on the left and tyrosine integrase phylogeny plotted on the right. Connections are drawn between a species and the tyrosine integrase(s) found in that species and each connecting line is colored according to host bacteria phylum.
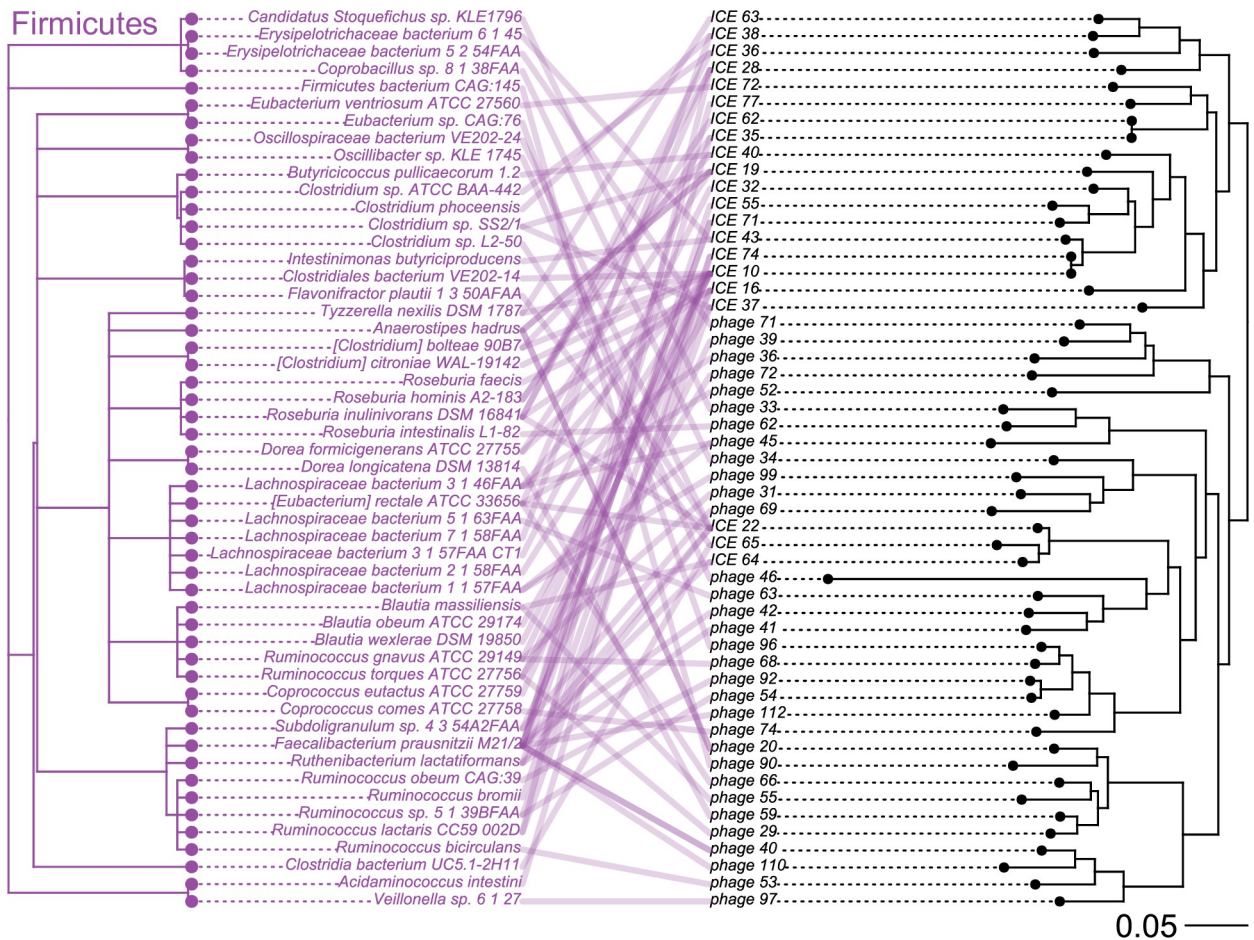
Fig 4. Tanglegram of host species lineages and phylogeny of serine integrases in prophages and ICEs. A tanglegram of serine integrases from ICEs and prophages with the species lineages plotted on the left and tyrosine integrase phylogeny plotted on the right. Connections are drawn between a species and the serine integrase(s) found in that species.

integrases, ICE22, ICE64 and ICE65 appear in a branch containing mostly prophages, suggesting that the integrase may have originated from a prophage integrase.

In summary, although ancient cross-phyla horizontal transfers did occur during the evolution of MGEs, we did not observe recent cross-phyla horizontal transfer of MGEs. Therefore, the gene pools that are shared within the gut microbiome are likely limited to the phylum-level.

Modular evolution of gut MGEs

Genes in MGEs are typically organized in functionally related modules which can be readily exchanged between MGEs. Type of modules found in MGEs include conjugation, integration, regulation, and adaptation. Deletion, acquisition, and exchanges of these modules can lead to immobilization, adaptation, and shifts in insertion specificity and host ranges of MGEs[13]. Here, we detail examples of each of these types of events.

Many unclassified genetic islands are likely remnants of ICEs or prophages due to the presence of only a subset of genes necessary for autonomous transfer. In many cases, the integrases have been lost while other genes for conjugation or capsid formation are maintained. One example is GI73, which appears to have formed when a CTnDOT-like element lost its

conjugation and mobilization modules to a large deletion (Fig 5A). We also observed many examples of the acquisition of new modules by insertions. CTnDOT-like elements have obtained adaptive modules via insertions of a group II intron together with the antibiotic resistance gene ErmF, an IME containing multiple antibiotic resistance genes including: ANT6, tetX, and ErmF[70,74], and other unidentified insertions containing many antibiotic resistance genes (Fig 2A; S5 File; S5 Table). Other examples are GI90, where ICE7 (CTnBST) inserted into a CTnDOT-like element (Fig 5A), and GI46, a genomic island formed when two types of ICEs (ICE43 and ICE56) inserted in tandem (Fig 5B). We observed that the exchange of recombination modules is common. Integrases have frequently been exchanged between ICEs and prophages during the evolution of MGEs (Figs 3 and 4). Exchanges also occur in the same class of MGE. For example, we observed that two clusters of ICEs, ICE15 and ICE16, share nearly identical sequences and the same typeFA conjugation module, but have different integrases: ICE15 has a tyrosine integrase while ICE16 has a serine integrase (Fig 5C). Overall, the modular nature of MGEs enables the formation of new mosaic elements, leading to the diversification of MGEs, and increasing the dynamics of the gene pools in the gut microbiome.

Discussion

In this study, we systematically characterized MGEs from the gut microbiome using a novel method to identify the mobilizable unit of active MGEs. We dramatically expanded the number of annotated MGEs from gut microbial species by identifying 5,219 putative MGEs. The MGEs we identified allow for the understanding of several fundamental questions about the role of MGEs and their importance to the evolution of species of the gut microbiome.

Comparison of SRID and other MGE detection methods

Two approaches are commonly used in the identification of MGEs. The first approach was to search for genomic loci enriched with mobile gene signatures and was used in tools such as ICEfinder[75], Conjscan[40], Phaser[16]. However, this approach cannot precisely determine the exact boundaries of MGEs, which is crucial to identify the functional units of MGEs. In addition, these tools are generally designed to identify one class of MGEs, not systematically identify all classes of MGEs. The second approach is done via comparative genomics. It requires two or more completely sequenced genomes of different strains or closely related species. In the metagenome field, metagenomic samples containing a species often far outnumber the number isolate genomes for that species. Therefore, SRID can leverage the thousands of metagenomic samples available and only requires one reference genome.

A limitation of the SRID is that it requires a match between the metagenomic data and the reference genome. The presence of a different strain of the same species as the reference genomes in the metagenomic samples is required. It also requires this strain to be relatively high abundance. This study used the HMP metagenomic sequencing data, where the samples were from healthy donors. In the stool of a healthy individual, Proteobacteria are of low abundance[76]. Therefore, the number of MGEs detected in the Proteobacteria is at least partially limited by the sequencing depth. Another limitation is that SRID alone cannot distinguish insertions due to MGEs or other random events. It often needs to use the MGE signatures to verify and classify the MGEs. Given no available tools for GIs/Islets gene signature identification, we require the presence of identical sequences present in two different species to confidently classify sequences as GI/Islets. There are only a few species of Proteobacteria that reach the necessary threshold of coverage. As a result, it is likely we are unable to identify some GI/Islets in these Proteobacteria species.
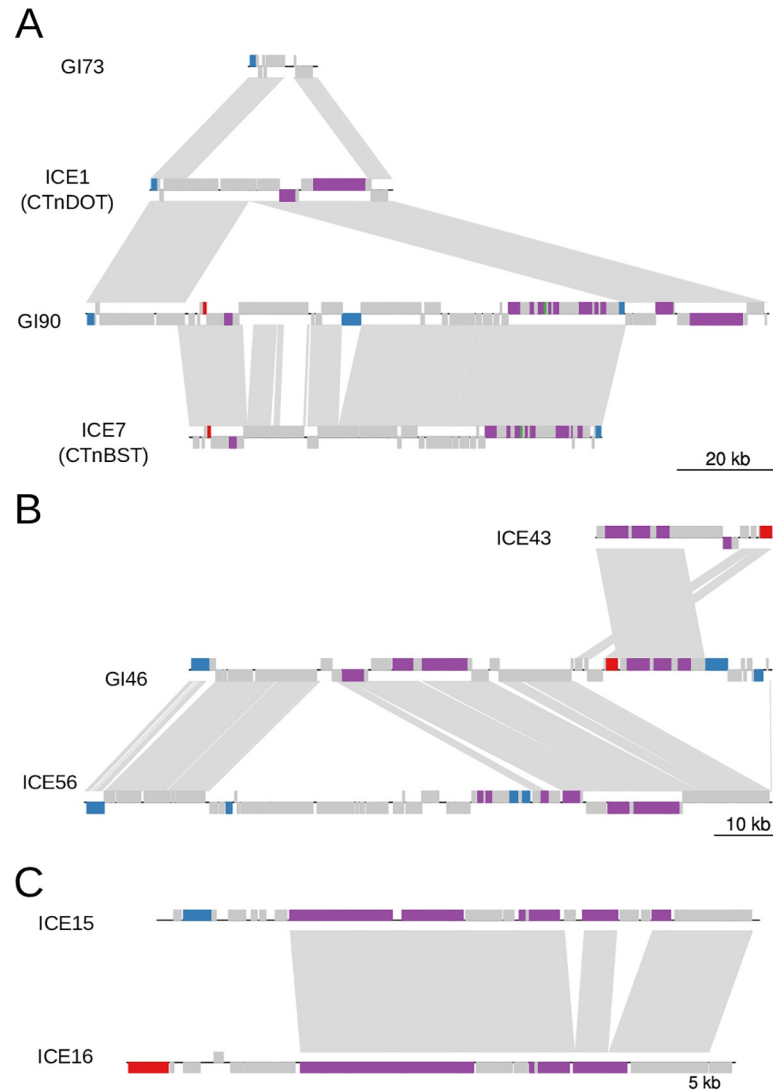
Fig 5. Modular evolution of MGEs. Examples of deletion, acquisition, and exchange of gene modules between MGEs. Orthologous genes between elements are shown with grey connections and are plotted with genoPlotR. Tyrosine integrases are colored blue, serine integrases are colored red, and genes involved in conjugation are colored purple. (A) GI73 was likely formed via a deletion of a CTnDOT-like ICE. GI90 was formed from an insertion of the ICE CTnBST into a CTnDOT-like ICE to form a large, composite GI that transfers as a unit. (B) An example of the tandem insertion of two ICEs to form a larger GI that moves as a unit. (C) An example of recombination module exchanges between ICE15 and ICE16.

Implications for future gut metagenomic analysis

The database of MGEs we have curated will be a valuable resource for future studies on the gut microbiome, especially with the increasing importance of taxonomic profiling, strain-level variation detection, and pangenome analyses. Many metagenomic workflows for taxonomic profiling use marker genes or k-mers "unique" to a specific species, where uniqueness is constrained by the available reference genomes[77–79]. These marker gene should exclude MGEs, as the potential horizontal transfer of these elements invalidates their "unique" species-specific associations. Strain-level variation analyses that based on single nucleotide polymorphisms (SNPs) or copy number variation should also exclude SNPs from MGEs[80–83]. In

pangenome analysis, it is beneficial to distinguish the accessory genes unique to an individual species and the mobilome shared among multiple species. To address the problems posed by MGEs to metagenomic workflows, an approach common in eukaryotic genomics, repeat masking, can be applied[84,85]. The database of curated MGEs identified in this study can be used to mask gut microbiome reference genomes before metagenomic workflows such as species-level classification, strain-level detection, and pangenome analyses are performed.

Host ranges of MGEs and the spread of antibiotic resistance genes

In the United States alone, more than 23,000 people die each year from antibiotic-resistant infections[86]. Tracking antibiotic resistance is one of the key actions to fight the spread of antibiotic resistance. The human digestive tract is a major reservoir of antibiotic resistance genes and likely serves as a hub for the horizontal transfer of antibiotic resistance genes from commensals to pathogens[4,5,69]. MGEs play a significant role in the spread of antibiotic resistance genes, and we found that many MGEs in the gut microbiome contain antibiotic resistance genes. This study helps to define the host range of MGEs in the gut microbiome. Our results suggest that HGT occurs mostly within a phylum, and inter-phyla HGT is rare. These results underscore the risk posed by transfer of antibiotic resistance genes like the vancomycin-resistance conferring gene VanB between commensal Firmicutes and pathogenic Firmicutes, such as *Enterococcus faecalis*[87]. Due to the fact our analysis only focused on chromosomal MGEs, antibiotic resistance spread by plasmids is not included in this study. Further studies on plasmids will provide more insight in the spread of antibiotic resistance. Overall, our study advances the understanding of the host range of MGEs which is of critical importance to understand gene flow networks in the gut.

This study underestimates the extent of host range because only MGEs in sequenced genomes were detected. As more bacterial genomes are sequenced, the extent of host range of MGEs will be refined. The scope of our research is chromosomal MGEs. Future studies using a combination of molecular and computational approaches will be beneficial to further understand the rate and extent of horizontal gene transfer by MGEs, including plasmids and extra-chromosomal phages.

Niche-adaptive genes in the communal gene pool

The mammalian gut is a unique ecological niche vastly different from other environments due to the presence of IgA, antimicrobial peptides, bile acids, as well as specific polysaccharides available for utilization in the intestinal mucus. The microbes that inhabit the gut must develop mechanisms to cope with these challenges. We observed that MGEs transfer genes to help address the unique challenges of colonizing the human gut. MGEs influence the spread of gut adaptive genes in three ways. First, the spread of MGEs drives the expansion and diversification of protein families such as those involved in polysaccharide utilization, capsular polysaccharide biosynthesis, and sensing and responding to the environment[9]. Second, MGEs transfer successful innovations for colonizing the gut among distantly-related species from the same niche, such as bile salt hydrolases. Third, MGEs allow for the amplification and transfer of genes that are adaptive only under specific conditions, such as antibiotic resistance genes, and sporulation-related genes.

Cargo genes transferred by MGEs can have wide-ranging effects on the biology of the gut microbiome. They potentially involved in bacterial symbioses, sensing and responding to environmental stimuli, and metabolic versatility. The enriched classes of cargo genes we identified in this study are attractive targets for future studies to understand the underlying biology of the gut microbiome.

Opportunities to use MGEs to engineer gut microbes

Tools for genome editing only exist for a very limited number of species of the gut microbiome despite the exceptional basic and translational opportunities afforded by engineering gut species. Many of the tools for editing the genomes of species were originally derived from MGEs. For instance, the NBU system used to modify some *Bacteroides* species was originally derived from an IME[88], and the TargeTron system was originally derived from a group II intron [89]. The novel examples of MGEs identified in this study could be used to edit genomes from the gut microbiome, especially in currently intractable species such as *Faecalibacterium prausnitzii*. Unlike phages, whose cargo genes are limited by the capsid size, many novel ICEs and IMEs carry hundreds of genes that can confer selective advantages for the host, and are excellent candidate vectors for large genetic loci. Overall, the MGEs identified in this study could have translational applications for genome editing of species from the gut microbiome.

Supporting information

S1 File. Supplementary information.

- **Figure A**. Classification of gut microbiome MGEs at the phylum level

- **Figure B**. Comparative genomics and alignment visualization to inspect MGE boundaries (DOCX)

S2 File. Scripts to implement the SRID method, the genome assembly accession numbers, HMP SRA accession numbers used in this study and phage signature term list(zipped txt).
(ZIP)

S3 File. MGE sequences (zipped fasta).
(ZIP)

S4 File. Annotation of genes on MGEs (zipped gff3).
(ZIP)

S5 File. Sequences of CTnDOT-like elements with ARG insertions (zipped fasta).
(ZIP)

S1 Table. Antibiotic resistant genes identified in MGEs classes.
(DOCX)

S2 Table. Annotation and classification of MGEs.
(XLSX)

S3 Table. Niche adaptive cargo genes.
(XLSX)

S4 Table. Enrichment analysis of cargo genes.
(XLSX)

S5 Table. Annotation of regions with ARG insertions in CTnDOT-like elements.
(XLSX)

Author Contributions

Conceptualization: Xiaofang Jiang, Andrew Brantley Hall, Eric J. Alm.

Data curation: Xiaofang Jiang, Andrew Brantley Hall.

References

1. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015; 16: 472–482. https://doi.org/10.1038/nrg3962 PMID: 26184597

2. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011; 480: 241–244. https://doi.org/10.1038/nature10571 PMID: 22037308

3. Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, Jenkins AP, et al. Mobile genes in the human microbiome are structured from global to individual scales. Nature. Nature Research; 2016; 535: 435–439.

4. Huddleston JR. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. Infect Drug Resist. 2014; 7: 167–176. https://doi.org/10.2147/IDR.S48820 PMID: 25018641

5. Roberts AP, Mullany P. Oral biofilms: a reservoir of transferable, bacterial, antimicrobial resistance. Expert Rev Anti Infect Ther. 2010; 8: 1441–1450. https://doi.org/10.1586/eri.10.106 PMID: 21133668

6. Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature. 2010; 464: 908–912. https://doi.org/10.1038/nature08937 PMID: 20376150

7. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev. 2011; 75: 610–635. https://doi.org/10.1128/MMBR.00011-11 PMID: 22126996

8. Lambowitz AM, Zimmerly S. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb Perspect Biol. 2011; 3: a003616. https://doi.org/10.1101/cshperspect.a003616 PMID: 20463000

9. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 2011; 7: e1001284. https://doi.org/10.1371/journal.pgen.1001284 PMID: 21298028

10. Hacker J, Carniel E. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. EMBO Press; 2001; 2: 376–381.

11. Wozniak RAF, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol. 2010; 8: 552–563. https://doi.org/10.1038/nrmicro2382 PMID: 20601965

12. Johnson CM, Grossman AD. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annu Rev Genet. 2015; 49: 577–601. https://doi.org/10.1146/annurev-genet-112414-055018 PMID: 26473380

13. Bellanger X, Payot S, Leblond-Bourget N, Guédon G. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol Rev. 2014; 38: 720–760. https://doi.org/10.1111/1574-6976.12058 PMID: 24372381

14. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. PLoS Comput Biol. 2015; 11: e1004095. https://doi.org/10.1371/journal.pcbi.1004095 PMID: 26020646

15. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. Mob DNA. 2015; 6: 13. https://doi.org/10.1186/s13100-015-0044-6 PMID: 26244060

16. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016; 44: W16–21. https://doi.org/10.1093/nar/gkw387 PMID: 27141966

17. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006; 34: D32–6. https://doi.org/10.1093/nar/gkj014 PMID: 16381877

18. Bi D, Xu Z, Harrison EM, Tai C, Wei Y, He X, et al. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. Nucleic Acids Res. 2012; 40: D621–6. https://doi.org/10.1093/nar/gkr846 PMID: 22009673

19. Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res. 2010; 38: D57–61. https://doi.org/10.1093/nar/gkp938 PMID: 19933762

20. Belda-Ferre P, Cabrera-Rubio R, Moya A, Mira A. Mining virulence genes using metagenomics. PLoS One. 2011; 6: e24975. https://doi.org/10.1371/journal.pone.0024975 PMID: 22039404

21. Trappe K, Marschall T, Renard BY. Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. Bioinformatics. 2016; 32: i595–i604. https://doi.org/10.1093/bioinformatics/btw423 PMID: 27587679

22. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015; 25: 1043–1055. https://doi.org/10.1101/gr.186072.114 PMID: 25977477

23. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486: 207–214. https://doi.org/10.1038/nature11234 PMID: 22699609

24. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17: 132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842

25. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. Oxford University Press; 2009; 25: 1754–1760.

26. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010; 11: 119. https://doi.org/10.1186/1471-2105-11-119 PMID: 20211023

27. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30: 1236–1240. https://doi.org/10.1093/bioinformatics/btu031 PMID: 24451626

28. Fouts DE. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 2006; 34: 5839–5851. https://doi.org/10.1093/nar/gkl732 PMID: 17062630

29. Abby SS, Cury J, Guglielmini J, Néron B, Touchon M, Rocha EPC. Identification of protein secretion systems in bacterial genomes. Sci Rep. 2016; 6: 23080. https://doi.org/10.1038/srep23080 PMID: 26979785

30. Cury J, Touchon M, Rocha EPC. Integrative and conjugative elements and their hosts: composition, distribution and organization. Nucleic Acids Res. 2017; 45: 8943–8956. https://doi.org/10.1093/nar/gkx607 PMID: 28911112

31. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5: R12. https://doi.org/10.1186/gb-2004-5-2-r12 PMID: 14759262

32. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

33. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25: 1972–1973. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

34. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

35. Revell LJ. Phytools: Phylogenetic tools or comparative biology (and other things). 2011.

36. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2010; 11: 431. https://doi.org/10.1186/1471-2105-11-431 PMID: 20718988

37. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2015; 9: 207–216. https://doi.org/10.1038/ismej.2014.106 PMID: 25003965

38. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics. 2004; 20: 1464–1465. https://doi.org/10.1093/bioinformatics/bth088 PMID: 14962934

39. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16: 284–287. https://doi.org/10.1089/omi.2011.0118 PMID: 22455463

40. Guglielmini J, Néron B, Abby SS, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. 2014; 42: 5715–5727. https://doi.org/10.1093/nar/gku194 PMID: 24623814

41. Guglielmini J, de la Cruz F, Rocha EPC. Evolution of conjugation and type IV secretion systems. Mol Biol Evol. 2013; 30: 315–331. https://doi.org/10.1093/molbev/mss221 PMID: 22977114

42. Rankin DJ, Rocha EPC, Brown SP. What traits are carried on mobile genetic elements, and why? Heredity. 2011; 106: 1–10. https://doi.org/10.1038/hdy.2010.24 PMID: 20332804

43. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Michael Cherry J, et al. Gene Ontology: tool for the unification of biology. Nat Genet. Nature Publishing Group; 2000; 25: 25–29. https://doi.org/10.1038/75556 PMID: 10802651

44. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42: D222–30. https://doi.org/10.1093/nar/gkt1223 PMID: 24288371

45. Connelly S, Bristol JA, Hubert S, Subramanian P, Hasan NA, Colwell RR, et al. SYN-004 (ribaxamase), an Oral Beta-Lactamase, Mitigates Antibiotic-Mediated Dysbiosis in a Porcine Gut Microbiome Model. J Appl Microbiol. 2017; https://doi.org/10.1111/jam.13432 PMID: 28245091

46. Volkova VV, Lu Z, Besser T, Gröhn YT. Modeling the infection dynamics of bacteriophages in enteric Escherichia coli: estimating the contribution of transduction to antimicrobial gene spread. Appl Environ Microbiol. 2014; 80: 4350–4362. https://doi.org/10.1128/AEM.00446-14 PMID: 24814786

47. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit M-A. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. ISME J. 2017; 11: 237–247. https://doi.org/10.1038/ismej.2016.90 PMID: 27326545

48. Shoemaker NB, Vlamakis H, Hayes K, Salyers AA. Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon. Appl Environ Microbiol. 2001; 67: 561–568. https://doi.org/10.1128/AEM.67.2.561-568.2001 PMID: 11157217

49. Devlin AS, Fischbach MA. A biosynthetic pathway for a prominent class of microbiota-derived bile acids. Nat Chem Biol. 2015; 11: 685–690. https://doi.org/10.1038/nchembio.1864 PMID: 26192599

50. Begley M, Gahan CGM, Hill C. The interaction between bacteria and bile. FEMS Microbiol Rev. 2005; 29: 625–651. https://doi.org/10.1016/j.femsre.2004.09.003 PMID: 16102595

51. Jones BV, Begley M, Hill C, Gahan CGM, Marchesi JR. Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proc Natl Acad Sci U S A. 2008; 105: 13580–13585. https://doi.org/10.1073/pnas.0804437105 PMID: 18757757

52. Johansson MEV, Phillipson M, Petersson J, Velcich A, Holm L, Hansson GC. The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. Proceedings of the National Academy of Sciences. 2008; 105: 15064–15069.

53. Johansson MEV, Larsson JMH, Hansson GC. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. Proc Natl Acad Sci U S A. 2011; 108 Suppl 1: 4659–4665.

54. Tailford LE, Crost EH, Kavanaugh D, Juge N. Mucin glycan foraging in the human gut microbiome. Front Genet. 2015; 6: 81. https://doi.org/10.3389/fgene.2015.00081 PMID: 25852737

55. Li H, Limenitakis JP, Fuhrer T, Geuking MB, Lawson MA, Wyss M, et al. The outer mucus layer hosts a distinct intestinal microbial niche. Nat Commun. 2015; 6: 8292. https://doi.org/10.1038/ncomms9292 PMID: 26392213

56. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014; 42: D490–5. https://doi.org/10.1093/nar/gkt1178 PMID: 24270786

57. Koropatkin NM, Cameron EA, Martens EC. How glycan metabolism shapes the human gut microbiota. Nat Rev Microbiol. 2012; 10: 323–335. https://doi.org/10.1038/nrmicro2746 PMID: 22491358

58. Ravcheev DA, Godzik A, Osterman AL, Rodionov DA. Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks. BMC Genomics. 2013; 14: 873. https://doi.org/10.1186/1471-2164-14-873 PMID: 24330590

59. Comstock LE, Coyne MJ, Tzianabos AO, Pantosti A, Onderdonk AB, Kasper DL. Analysis of a capsular polysaccharide biosynthesis locus of Bacteroides fragilis. Infect Immun. 1999; 67: 3525–3532. PMID: 10377135

60. Zitomersky NL, Coyne MJ, Comstock LE. Longitudinal analysis of the prevalence, maintenance, and IgA response to species of the order Bacteroidales in the human gut. Infect Immun. 2011; 79: 2012–2020. https://doi.org/10.1128/IAI.01348-10 PMID: 21402766

61. Wu M, McNulty NP, Rodionov DA, Khoroshkin MS, Griffin NW, Cheng J, et al. Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut Bacteroides. Science. 2015; 350: aac5992. https://doi.org/10.1126/science.aac5992 PMID: 26430127

62. Patrick S, Blakely GW, Houston S, Moore J, Abratt VR, Bertalan M, et al. Twenty-eight divergent polysaccharide loci specifying within- and amongst-strain capsule diversity in three strains of Bacteroides fragilis. Microbiology. 2010; 156: 3255–3269. https://doi.org/10.1099/mic.0.042978-0 PMID: 20829291

63. Coyne MJ, Chatzidaki-Livanis M, Paoletti LC, Comstock LE. Role of glycan synthesis in colonization of the mammalian gut by the bacterial symbiont Bacteroides fragilis. Proceedings of the National Academy of Sciences. 2008; 105: 13099–13104.

64. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, et al. Evolution of symbiotic bacteria in the distal human intestine. PLoS Biol. 2007; 5: e156. https://doi.org/10.1371/journal.pbio.0050156 PMID: 17579514

65. Albenberg L, Esipova TV, Judge CP, Bittinger K, Chen J, Laughlin A, et al. Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota. Gastroenterology. 2014; 147: 1055–63.e8. https://doi.org/10.1053/j.gastro.2014.07.020 PMID: 25046162

66. Duncan SH, Hold GL, Harmsen HJM, Stewart CS, Flint HJ. Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. Int J Syst Evol Microbiol. 2002; 52: 2141–2146. https://doi.org/10.1099/00207713-52-6-2141 PMID: 12508881

67. Browne HP, Forster SC, Anonye BO, Kumar N, Anne Neville B, Stares MD, et al. Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation. Nature. Nature Research; 2016; 533: 543–546.

68. Berendsen EM, Boekhorst J, Kuipers OP, Wells-Bennik MHJ. A mobile genetic element profoundly increases heat resistance of bacterial spores. ISME J. 2016; 10: 2633–2642. https://doi.org/10.1038/ismej.2016.59 PMID: 27105070

69. Sommer MOA, Church GM, Dantas G. The human microbiome harbors a diverse reservoir of antibiotic resistance genes. Virulence. 2010; 1: 299–303. https://doi.org/10.4161/viru.1.4.12010 PMID: 21178459

70. Whittle G, Hund BD, Shoemaker NB, Salyers AA. Characterization of the 13-kilobase ermF region of the Bacteroides conjugative transposon CTnDOT. Appl Environ Microbiol. 2001; 67: 3488–3495. https://doi.org/10.1128/AEM.67.8.3488-3495.2001 PMID: 11472924

71. Garnier F, Taourit S, Glaser P, Courvalin P, Galimand M. Characterization of transposon Tn1549, conferring VanB-type resistance in Enterococcus spp. Microbiology. 2000; 146 (Pt 6): 1481–1489.

72. Napolitano MG, Almagro-Moreno S, Boyd EF. Dichotomy in the evolution of pathogenicity island and bacteriophage encoded integrases from pathogenic Escherichia coli strains. Infect Genet Evol. 2011; 11: 423–436. https://doi.org/10.1016/j.meegid.2010.12.003 PMID: 21147268

73. Van Houdt R, Leplae R, Lima-Mendez G, Mergeay M, Toussaint A. Towards a more accurate annotation of tyrosine-based site-specific recombinases in bacterial genomes. Mob DNA. 2012; 3: 6. https://doi.org/10.1186/1759-8753-3-6 PMID: 22502997

74. Whittle G, Hamburger N, Shoemaker NB, Salyers AA. A bacteroides conjugative transposon, CTnERL, can transfer a portion of itself by conjugation without excising from the chromosome. J Bacteriol. 2006; 188: 1169–1174. https://doi.org/10.1128/JB.188.3.1169-1174.2006 PMID: 16428422

75. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, et al. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic Acids Res. 2019; 47: D660–D665. https://doi.org/10.1093/nar/gky1123 PMID: 30407568

76. Rizzatti G, Lopetuso LR, Gibiino G, Binda C, Gasbarrini A. Proteobacteria: A Common Factor in Human Diseases [Internet]. BioMed Research International. 2017. pp. 1–7. https://doi.org/10.1155/2017/9351507 PMID: 29230419

77. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015; 12: 902–903. https://doi.org/10.1038/nmeth.3589 PMID: 26418763

78. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15: R46. https://doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807

79. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015; 16: 236. https://doi.org/10.1186/s12864-015-1419-2 PMID: 25879410

80. Ahn T-H, Chai J, Pan C. Sigma: strain-level inference of genomes from metagenomic analysis for bio-surveillance. Bioinformatics. 2015; 31: 170–177. https://doi.org/10.1093/bioinformatics/btu641 PMID: 25266224

81. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015; 33: 1045–1052. https://doi.org/10.1038/nbt.3319 PMID: 26344404

82. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. Cell. 2015; 160: 583–594. https://doi.org/10.1016/j.cell.2014.12.038 PMID: 25640238

83. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016; 26: 1612–1625. https://doi.org/10.1101/gr.201863.115 PMID: 27803195

84. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. Institute for Systems Biology http://repeatmasker org. 2015;

85. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015; 6: 11. https://doi.org/10.1186/s13100-015-0041-9 PMID: 26045719

86. Antibiotic / Antimicrobial Resistance | CDC [Internet]. [cited 14 Jun 2017]. Available: https://www.cdc.gov/drugresistance/index.html

87. Caballero S, Kim S, Carter RA, Leiner IM, Sušac B, Miller L, et al. Cooperating Commensals Restore Colonization Resistance to Vancomycin-Resistant Enterococcus faecium. Cell Host Microbe. 2017; 21: 592–602.e4. https://doi.org/10.1016/j.chom.2017.04.002 PMID: 28494240

88. Mimee M, Tucker AC, Voigt CA, Lu TK. Programming a Human Commensal Bacterium, Bacteroides thetaiotaomicron, to Sense and Respond to Stimuli in the Murine Gut Microbiota. Cell Syst. 2015; 1: 62–71. https://doi.org/10.1016/j.cels.2015.06.001 PMID: 26918244

89. Liu Y-J, Zhang J, Cui G-Z, Cui Q. Current progress of targetron technology: development, improvement and application in metabolic engineering. Biotechnol J. 2015; 10: 855–865. https://doi.org/10.1002/biot.201400716 PMID: 25735546