Separation of Cohorts on the Basis of Bacterial Type IV Conjugation Systems Identified From Metagenomic Assemblies

Benjamin R Joris  
University of Western Ontario: Western University  
https://orcid.org/0000-0003-2786-6317

Tyler S Browne  
University of Western Ontario: Western University

Thomas A Hamilton  
University of Western Ontario: Western University

David R Edgell  
University of Western Ontario: Western University

Gregory B Gloor (✉ ggloor@uwo.ca)  
University of Western Ontario: Western University  
https://orcid.org/0000-0001-5803-3380

Research

Keywords: Bacterial Type IV, Conjugation, Metagenomic Assemblies

Posted Date: June 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-660934/v1

License: ☞ ☀ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Separation of cohorts on the basis of bacterial type IV conjugation systems identified from metagenomic assemblies

Benjamin R. Joris, Tyler S. Browne, Thomas A. Hamilton, David R. Edgell and Gregory B. Gloor

22 June, 2021

Author Affiliations: Department of Biochemistry, Schulich School of Medicine, The University of Western Ontario, London, Ontario, N6A 3K7, Canada

Corresponding author: Gregory B. Gloor, ggloor@uwo.ca
Abstract

Background

Conjugation enables the exchange of genetic elements throughout environments, including the human gut microbiome. Conjugative elements can carry and transfer clinically relevant metabolic pathways which makes precise identification of these systems in metagenomic samples clinically important.

Results

Here, we outline two distinct methods to identify conjugative elements in the human gut microbiome. We first show that conjugative elements exhibit strong population and age-level stratification. Furthermore, the taxonomic compositions of the conjugative elements differ from the composition of the metagenome assembled genomes, both in terms of the number of assembled elements and the relative abundances of the assembled systems. Finally, we demonstrate that the majority of assembled conjugative elements are not included within metagenomic bins, and that only a small proportion of the binned conjugative systems are included in “high-quality” metagenomic bins. Our findings highlight that conjugative systems differ between a North American inflammatory bowel disease cohort and a cohort of North American pre-term infants, but in a manner different than metagenome assembled genomes, revealing a potential use as an age-related biomarker. Additionally, conjugative systems can distinguish between other geographical-based cohorts.

Conclusions

Analysis of the human gut microbiome by shotgun metagenomic sequencing has revealed numerous connections to human health outcomes. Our findings emphasize the need to identify and analyze conjugative systems outside of standard metagenomic binning pipelines. We suggest that analysis of type IV conjugative systems should be added to the current metagenomic analysis approaches as they contain much information that could explain differences between cohorts beyond those we investigated.

Background

Bacteria can acquire exogenous DNA through horizontal gene transfer. Conjugation is a common mechanism of horizontal gene transfer that relies on direct cell-cell contact to unidirectionally transfer DNA from a bacterial donor to a recipient cell. In bacteria, integrative conjugative elements (ICEs) and conjugative plasmids are mobilizable through the actions of type IV secretion systems (T4SS). Approximately half of the known plasmids are mobilizable in \textit{trans} where the conjugative machinery is on a different genetic element than the transferred element, and the remainder are mobilizable in \textit{cis} because the conjugative machinery is present on the same genetic element [1]. ICEs encode their own T4SS, and can mobilize other elements [2]. Conjugative elements (CEs) often contain antibiotic resistance genes, but also can harbour useful biosynthetic and biodegradation genes [3]. Furthermore, conjugative systems can serve as vectors to introduce clustered
regularly interspaced short palindromic repeats (CRISPR) systems, metabolic pathways or novel functions into the gut microbiota [4–9]. Therefore, characterizing the full complement of conjugative systems in the human gut could expand the number of useable vectors for these applications. Precise identification of conjugative systems from metagenomic samples could also provide insights to their distribution in populations and their correlation with antibiotic exposure, age, and health status.

For a DNA sequence to be considered mobilizable, it must encode an origin of transfer (oriT) sequence that is recognized and nicked by a relaxase protein [1, 10]. Relaxase proteins contain a conserved histidine triad that coordinates a divalent metal ion, as well as tyrosine residues that bind the oriT DNA sequence and catalyzes the nicking reaction [11, 12]. In addition to a relaxase gene and an oriT sequence, a full complement of type IV secretion system and coupling proteins are required for a sequence to be mobilizable. In the well-studied Agrobacterium tumefaciens conjugative system, there are 12 proteins involved in the transfer of the DNA-relaxase complex from one bacterial cell to another [13, 14]. Homologs of the VirB4 ATPase that are essential for assembly of the conjugative system and DNA transfer are generally similar to the phylogeny of the bacteria harbouring them [15] and thus are useful for classifying conjugative systems [16]. The synteny of conjugative transfer genes is also highly conserved among conjugative systems [14]. Both the synteny and presence of highly-conserved genes involved in conjugation facilitates the classification of genetic elements as potentially conjugative if the sequences are annotated as belonging to the components of the T4SS [17] (Figure 1).

Previous work has identified novel CEs in the human and animal gut microbiomes, but the focus was mainly on ICEs and not on conjugative plasmids [3, 18, 19]. Identifying conjugative plasmids from a short-read metagenomic assembly is difficult for several reasons. The initial barrier is the difficulty in assembling circularized plasmids from short-read sequencing data [20]. A second barrier is that the contiguous DNA sequences (contigs) that compose metagenome-assembled genomes (MAGs) are binned together based on sequence composition and coverage. Binning of a plasmid with its cognate genome will not happen unless the contigs that compose the plasmid are maintained in the same copy-number and have the same sequence composition as the chromosome. These criteria are generally not met because conjugative systems are usually more AT rich than the cognate chromosome [1] and often do not have a unit copy number. Since nearly 80% of the non-redundant set of genomes from the human-gut microbiome are from difficult-to-culture species that are known only from MAGs [21], alternate methods must be employed to assemble and identify conjugative plasmids from the metagenomic sequencing data. Computational tools have recently been developed to identify plasmids from metagenomic assemblies [22], but would be rendered pointless if applied to already binned data that systematically excludes plasmids [23]. Methods that identify CEs prior to binning should be able to capture the full spectrum of ICEs and conjugative plasmids.

Here, we show that T4SS conjugative systems can be identified using two distinct methods (Figure 2). First, we used profile HMMs (pHMMs) to identify conjugative systems directly from metagenomic assemblies
of North American inflammatory bowel disease (IBD) and North American pre-term infant samples [17].

Second, we searched predicted protein sequences versus UniRef90 [24] for proteins involved in conjugation to identify conjugative systems from a human gut microbiome genome set. With this approach differences between additional cohorts in the relative abundances of extracted CEs could be recognized. While the differences between cohorts found using the two methods were not identical, both methods did illustrate that different age and population cohorts were distinct. As well, the taxonomic compositions of the assembled CEs and MAGs were different within cohorts. Finally, we demonstrate that the majority of conjugative systems produced by a metagenomic assembly are not included in high-quality bins that were used to compose human gut microbiome genome sets. Our findings provide a roadmap to integrate the analysis of conjugative systems alongside the chromosomal content of bacteria.

**Methods**

**Assembly and identification of conjugative systems in North American short-read data**

Samples belonging to a North American IBD (n=50) [25] and a North American pre-term infant cohort (n=51) [26] were assembled *de novo* (Supplemental Table 1). Reads from these samples were downloaded from the Sequence Read Archive using the SRA toolkit version 2.9.2, deduplicated with dedupe.sh [27], and trimmed with Trimmomatic version 0.36 [28] with options **LEADING**:10 **TRAILING**:10. Processed reads were assembled sample-by-sample using SPAdes version 3.14.0, option **--meta** [29]. The resultant assemblies were imported into Anvi’o version 6.0 [30] where the presence of T4SS, T4CP, and relaxase proteins were predicted using the anvi-run-hmms module, which integrates HMMER3 functionality [31]. Contigs that contained pHMM matches for all three classes of conjugative proteins were extracted and annotated by aligning open reading frames (ORFs) predicted with Prodigal version 2.6.3 [32] to the UniRef90 database [24]. Subregions of the contigs where annotations for conjugative proteins were present, with no more than 20 ORFs between successive UniRef90 annotations for conjugative proteins, were extracted. The processed read data were mapped to the extracted conjugative systems using Bowtie2 version 2.3.5 [33] with the settings **--no-unal --no-mixed --no-discordant**. Extraction of the subregions was to avoid an artificially high proportion of reads mapping in samples where the bacterium is present, but the ICE has not integrated in its chromosome (Supplemental Figure 1). Taxonomic prediction of the contigs was conducted with Kaiju version 1.7.2 utilizing the RefSeq non-redundant protein database [34]. MOB-suite verion 1.4.9.1 was utilized to characterize the incompatibility grouping of the conjugative system, if possible [35]. PlasFlow version 1.1.0 was used to classify whether the system was chromosomally integrated or a plasmid [22]. The proportion of reads mapping to the conjugative systems was extracted from the Bowtie2 output, and the mapping data was visualized using Anvi’o [30]. Raw counts of reads mapping to the extracted conjugative systems were transformed using a centered log-ratio. The principal component coordinates of the first 2 components were
Binning of Assemblies

For each assembly, all 101 samples were mapped to the contigs using Bowtie2 [33]. The mapping files were sorted and indexed with SAMtools [37] and then the assemblies were binned using MetaBAT2 version 2.12.1 [38]. CheckM version 1.1.2 was used to assess the quality of the resultant bins [39]. High-quality bins were defined using the same cutoffs (>90% completion and <5% redundancy) as Almeida et al (2019) defined. Bins not passing that threshold were classified as “low-quality”. The previously identified contigs with conjugative systems were classified based on their presence in bins, and the types of bins they were present in. Results of this classification were visualized using SankeyMATIC (http://sankeymatic.com/).

Taxonomic composition comparison of assembled genetic elements

To compare the taxonomic composition Contig Annotation Tool (CAT) and Bin Annotation Tool (BAT) [40] were used to assign for conjugative elements and metagenome assembled genomes respectively. Reads were mapped to the all conjugative elements and all bins independently using Bowtie2 [33] with the settings --no-unal --no-mixed --no-discordant. The number of reads per sample mapping to each phylum was obtained by aggregating data output by the idxstats function of SAMtools [37].

Reference human gut metagenome set

A near-complete and non-redundant set of human gut microbiome genomes were downloaded from the European Bioinformatics Institute FTP site (ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/umgs_analyses/) [21]. These genomes were assembled from 13,133 metagenomic samples using SPAdes [29] and binned using MetaBAT2 [38]. The quality of binned genomes were assessed using CheckM [39]. High-quality genomes were defined as >90% completeness and <5% contamination and medium-quality genomes were defined as >50% completeness and <10% contamination, and these genomes were used to create the non-redundant set of genomes. The program dRep was used to cluster the genomes at 99% sequence identity [41] thereby dereplicating the genome bins, creating a set of 2505 genomes.

Identifying and quantifying conjugative systems in reference human gut metagenome set

ORFs were predicted in the genome by Prodigal version 2.6.3 [32]. The predicted protein sequences were then aligned to the UniRef90 database [24] using the Diamond protein aligner version 0.9.14 [42]. Contigs were extracted from the genomes if they contained annotations for a relaxase/mobilization protein and a type IV secretion/type IV coupling protein using a word-search strategy. MOB-suite version 1.4.9.1 was utilized to characterize the incompatibility grouping of the conjugative system, if possible [35]. PlasFlow version
1.1.0 was used to classify whether the system was predicted to be chromosomally integrated or located on
a plasmid [22]. Short-read data from 785 samples (Supplemental Table 2) [25, 26, 43–48] were downloaded
from the Sequence Read Archive using the SRA toolkit version 2.9.2. The reads were processed and aligned
to the regions of the contigs where type IV conjugative systems were located using the previously described
methods. The principal component coordinates of the first 3 components were used for clustering by hdbscan
[36].

**Results**

**Conjugative systems identified from assembly of short-read data distinguish North American cohorts**

51 samples from a pre-term infant cohort and 50 from a North American IBD cohort were assembled sample-
by-sample using metaSPAdes [29] to identify T4SS conjugative systems from a full pool of assembled contigs
(i.e. not binned) and to compare the relative abundances of these systems between cohorts. These two cohorts
were selected for their expected differences in microbiome composition and the sequencing quality available.
For these analyses, contigs with conjugative systems were defined by pHMM matches for a relaxase, a type
IV coupling protein, and a type IV secretion system, which offers a fast and precise method to annotate a
limited number of protein families. From the assembly of the pre-term infant cohort 96 of 470500 contigs met
the criteria, whereas 268 of 15100646 contigs from the IBD cohort did (Supplemental Table 3). Predicted
ORFs from contigs with conjugative systems were aligned to the UniRef90 database and the subregions
with conjugative systems were extracted. The short-read data from all 101 samples were mapped to all 391
subregions of conjugative proteins.

The patterns of conjugative system occurrence and relative abundance in the two cohorts are distinct (Figure
3). Furthermore, most identified conjugative contigs were private to the cohort from which they were
assembled. Conjugative systems belonging to the *Proteobacteria* phylum were only assembled from pre-term
infant samples and did not have any apparent occurrence in the IBD cohort. When the log-transformed
principal component analysis data were clustered with hdbscan [36], the two cohorts formed separate clusters
(Figure 4, Supplemental Figure 2). These findings show that North American pre-term infants have a
strikingly different array of conjugative systems than the North American IBD cohort.

**Taxonomic compositions of conjugative elements and metagenome assembled genomes are discordant**

The proportions of phyla assigned with CAT/BAT present in the assembled conjugative elements and binned
metagenome assembled genomes differ for those assembled from the preterm term infant samples (Figure 5a)
and the IBD samples (Figure 5b). The differences are generally similar in both cohorts; For both cohorts
there are a greater proportion of assembled Bacteroidetes CEs compared to Bacteroidetes MAGs and lower proportions of Actinobacteria and Firmicutes CEs than MAGs. Similarly, there are also more CEs in both cohorts that cannot have taxonomy successfully assigned, likely due to their fewer ORFs present for CAT to utilize as well as the possibility of some being phage [49]. In contrast, a greater proportion of CEs than MAGs were assembled from the Proteobacteria in the preterm infant samples, whereas in the IBD samples, there were no Proteobacteria conjugative elements assembled at all.

The number of assembled elements is not a perfect proxy for comparing to groups because there may be a CEs or MAGs that do not assemble, yet are present in the community, and thus do not provide a real indication of the relative abundances of the elements and how they compare between CEs and MAGs. As such, the reads from the two cohorts were mapped to all assembled CEs and binned MAGs to compare the composition of the communities based on the abundances. Taxonomic composition of reads mapping to CEs and MAGs is more similar than the proportion of assembled CEs and MAGs (Figure 5c and Figure 5d). However, there are still some notable deviations such as there being no reads mapping to Proteobacteria CEs from the IBD cohort, but there being successfully assembled and binned Proteobacteria genomes present. Furthermore, while the proportion of assembled Proteobacteria CEs was greater than Proteobacteria MAGs for the preterm infant cohort, a greater proportion of the reads aligned to the Proteobacteria MAGs than CEs. For the preterm infant cohort, an outsized proportion of reads mapped to taxonomically unassigned elements, which makes it difficult to make unambiguous assertions. However, for the data which has taxonomic assignments, it appears that the taxonomic composition of the CEs is different than the MAGs.

Mapping human gut microbiome data from cohorts to conjugative systems reveals distinct patterns

Next, we explored relative abundances of conjugative systems in a larger number of cohorts, without having to conduct computationally-expensive metagenomic assemblies. For this analysis, conjugative systems were identified from a set of 2505 bacterial genomes, which represent a non-redundant and near-complete picture of the human gut microbiome [21]. A total of 1598 contigs from 787 genomes that contain UniRef90 annotations for relaxase/mobilization and T4SS/T4CP proteins were identified (Supplemental Table 3). From these contigs, 3216 subregions where conjugative protein annotations were concentrated on the contig were extracted (Supplemental Figure 1), with 2413 being >1kb in size and used for visualization. Short-read human gut microbiome sequencing data from 785 samples, spread across 8 cohorts were aligned to the extracted subregions (Supplemental Table 1). With the conjugative systems identified from the human gut metagenome set, the two North American cohorts that were previously analyzed are still distinct, albeit in a different way (Figure 6). Only a very small number of the reads from the North American and European Infant cohorts mapped to conjugative systems. The only notable signal is in the Proteobacteria phylum for the North American pre-term infants, a finding consistent with what was found by de novo assembly. The
West African and South American cohorts also share similar characteristics as both have an overall lower apparent relative abundance of conjugative systems compared to the other non-infant cohorts, particularly in the Bacteroidetes phylum. The other four cohorts appear similar with regards to the presence and absence of the conjugative systems. The cohorts separated into three distinct clusters (Figure 7, Supplemental Figure 3), when the principal components of the centered log-ratio transformed data were clustered using hdbscan [36]. In this analysis infant cohorts were excluded because of their extreme sparsity. The majority of the West African and South American samples clustered together consistent with Figure 6. Not readily apparent from the cladogram was the East Asian cohort that clustered primarily on its own. The North American Indigenous, North American IBD, and Western European general samples largely clustered together. Like the conjugative systems identified from the de novo assemblies of short-read data, the relative abundances of conjugative systems identified from a human gut metagenome set separated cohorts into distinct groups.

The majority of conjugative systems identified by assembly are omitted from metagenomic bins

The assemblies were binned using MetaBAT2 [38], which was also used to bin the MAGs in the human gut genome set used in the prior analyses[21] to further explore how conjugative systems are distributed within common metagenomic analyses. Of the 364 assembled contigs containing pHMM matches to all three protein categories, 270 were not included in any metagenomic bins (Figure 8). For the 94 contigs included in metagenomic bins, 65 of those were found in high-quality bins (>90% completion and <5% redundancy). This is in stark contrast to the background binning rate of contigs; For contigs above 5kb in size the binning rate with MetaBAT2 [38] was 70.4% (116112/164843 contigs) and for contigs above 10kb the binning rate was 79.1% (57214/72300 contigs). Among the 29 contigs included in bins that do not meet the aforementioned threshold, 8 are within bins that are less than or equal to 1 megabase in size, potentially suggesting that fragments of a conjugative plasmid may have binned together.

Discussion

The relative abundances of conjugative systems identified from MAGs and isolate genomes from the human gut differ between cohorts in a similar manner to how relative abundances of human gut MAGs of different species are differential between cohorts [50]. The infant cohorts stood out the most from the other cohorts; the infant gut microbiome is composed largely of members of the genus Bifidobacteria and is recognized as being distinct to the microbiomes of adults over the first few years of life [51]. Furthermore, we observed that the CEs present in the gut microbiome of pre-term infants were distinct from other infants and adults. This could be because of exposure to antibiotics from birth and colonization by opportunistic Proteobacteria pathogens such as members of the genera Escherichia, Klebsiella, and Enterobacter [26]. As shown in Figure 6, the only conjugative systems that showed a signal for the samples belonging to the North American
pre-term infants were those belonging to the *Proteobacteria* phylum, and while there are the characteristic *Actinobacteria* MAGs present in the infant samples there were not any CEs of that phylum assembled and a lower relative abundance of *Actinobacteria* CEs than MAGs when mapping to those assembled from the IBD cohort (Figure 5). The degree of difference in relative abundances of conjugative systems between infants and adults suggests that conjugative systems could be a potential biomarker for age, or for the relative maturity of the infant microbiome.

The relative abundances and distributions of conjugative systems in the West African and South American cohorts were distinct from the other non-infant cohorts, which is similar to the findings in the relative abundances of bacterial species in these cohorts [50]. As well, the East Asian cohort clustered separately from the other non-infant cohorts. These findings suggest that conjugative systems might be useful biomarkers for other factors beyond just age and further focus on geographical or health-related differences may yet reveal additional separation between cohorts based on the relative abundances of conjugative systems.

We found that a reference bacterial genome set can be useful for identifying coarse differences in the conjugative systems between populations; however, this method may not capture the true diversity of conjugative between populations, because many conjugative systems may be omitted. To produce MAGs, contigs generated by metagenomic assembly are typically binned using a program such as MetaBAT2 [38]. Conjugative systems are often more AT rich than the parent genomes [1], which would result in the conjugative system and cognate genome not occurring in the same metagenomic bin because binning algorithms use GC content as a parameter for clustering. Additionally, plasmids are not necessarily maintained in a unit copy number within the cell, causing differential sequence coverage in comparison to the parent genome, which results in plasmids being excluded from MAGs. Therefore to capture a more complete image of the conjugative systems present in an environment, identification of the systems must take place before binning.

The assembled contigs were binned with MetaBAT2 [38] as a way of quantifying the effect of binning, which revealed that the vast majority of the assembled conjugative systems were not included in metagenomic bins and therefore would not be included in a MAG database, which confirms recent findings [23]. The background binning rate of contigs carrying type IV conjugative systems was considerably lower than the background binning rate of equivalently sized contigs (25.8% compared to 70.4%) Many of the binned conjugative systems were not within a bin that would pass the quality cutoff to be included in the genome set as well [21]. Interestingly, eight of the conjugative systems were binned into low-quality bins that were smaller than <1MB in size, which may suggest that the fragments of a conjugative plasmid could be binned together, which would increase the completeness of the conjugative system.
Conclusions

Conjugative systems differ between cohorts and require special consideration to ensure that they are included in analyses. ICEs and plasmids can carry harmful systems, such as antimicrobial resistance, but also can act as vectors for bile salt metabolism and for detoxification modules [3]. These cargo genes are relevant for research relating to the gut microbiome’s role in pathogenicity as well as metabolism and digestion. Comprehensive identification and quantification of conjugative systems could allow for association of conjugative systems with different health outcomes. Because assembled plasmid-based conjugative systems are rarely included in metagenomic bins [23] (Figure 8), they need to be identified and analyzed outside of standard binning pipelines. At present, it is not possible to assemble complete plasmids from short-read metagenomic data [20], so it may helpful to identify bins containing conjugative systems in an attempt to cluster the fragments of plasmids present in an assembly together. Identifying type IV conjugative systems using pHMMs or annotations and using tools such as PlasFlow [22] to identify plasmids out of a full assembly in parallel with standard binning analyses will enhance research of the associations between the human gut microbiome and human health.

In the future, improvements in assembly and binning algorithms will continue to improve the recovery of low relative abundance conjugative elements and improve the completeness and accuracy of the assembled fragments. Additionally, long-read assembly permits the circularization of genomes and plasmids [52, 53] and the binning of plasmids to their cognate genomes using methylation data [54], which will reduce the ambiguity of the origins of conjugative systems (i.e. whether they are an ICE or independently circularized plasmid) and provide a more complete picture of the cargo they carry and the differences between cohorts.

Figure 1: Example schematic of the gene organization of a bacterial conjugation system on the Agrobacterium tumefaciens pTi plasmid.
Approach 1: Use metagenome database to search for conjugative systems

1. Download genomes
2. Predict open reading frames using Prodigal
3. Align open reading frames to UniRef90 database using Diamond
4. Extract contigs with UniRef90 annotations for conjugative proteins
5. Map short-read sequencing data to extracted contigs

Approach 2: Assemble raw reads and search assemblies for conjugative systems

1. Download short-read data from Sequence Read Archive
2. Sample-by-sample metagenomic assembly with metaSPAdes
3. Identification of conjugative genes using hidden Markov models
   - Extract contigs with HMM matches for conjugative proteins
   - Map short-read sequencing data to extracted contigs
4. Metagenomically bin contigs using MetaBat2
   - Extract bins with a contig containing HMM matches for conjugative proteins
   - Map short-read sequencing data to identified bins

Figure 2: Overview of methods employed in this study. In the left panel is the workflow used to identify conjugative systems from previously assembled human gut bacterial genomes. The right panel outlines the workflow for the assembly of select North American samples and the use of pHMMs to identify the conjugative systems.
Figure 3: Anvi’o cladogram of potentially conjugative contigs from 51 North American pre-term infants samples and 50 North American IBD samples. Inner rings of the phylogram represent individual samples, second-most outer ring being the phylum of conjugative system as predicted by Kaiju, and the outermost ring represents the cohort that the conjugative contig was assembled from. Each slice of the circle phylogram are individual conjugative contig identified by pHMMs of conjugative proteins. For the inner plot, intensity of the position represents the mean coverage of the contig for a given sample proportional to the other conjugative systems.
Figure 4: Clustering of the principal component coordinates of the CLR transformed relative abundances of the extracted conjugative regions from the assemblies of North American datasets. Coloured points represent membership to clusters with grey points not belonging to a cluster. Ellipses represent a 95 percent confidence interval using a multivariate t-distribution about the cluster.
Figure 5: (a) Taxonomic compositions of conjugative elements (CE) and metagenome-assembled genomes (MAG) assembled in from the North American pre-term infant samples. (b) Taxonomic compositions of conjugative elements (CE) and metagenome-assembled genomes (MAG) assembled in from the North American inflammatory bowel disease samples. (c) Taxonomic compositions of reads from North American pre-term infant samples mapping to conjugative elements (CE) and metagenome-assembled genomes (MAG) assembled from both cohorts. (d) Taxonomic compositions of reads from North American inflammatory bowel disease mapping to conjugative elements (CE) and metagenome-assembled genomes (MAG) assembled from both cohorts.
**Figure 6:** Anvi’o cladogram of potentially conjugative systems originating from 785 samples across 8 cohorts. Inner rings of the phylogram represent individual samples and the outermost ring being the phylum of conjugative system. Each slice of the circle phylogram are individual conjugative regions. For each point on the inner plot, the intensity of the black colouring represents the mean coverage of the system for a given sample proportional to the other conjugative systems.
Figure 7: Clustering of the principal component coordinates of the CLR transformed relative abundances of the extracted conjugative regions from the genome database. Coloured points represent membership to clusters with grey points not belonging to a cluster. Ellipses represent a 95 percent confidence interval using a multivariate t-distribution about the cluster.
Figure 8: Sankey diagram representing the flow of 364 contigs containing conjugative systems into bins generated by MetaBAT2 from assembled data.

Supplemental Table 1: SRA accession numbers and cohorts for samples used in de novo assembly workflow.

Supplemental Table 2: SRA accession numbers and cohorts for samples used in reference human gut metagenome set workflows.

Supplemental Table 3: Summary data for all predicted conjugative systems.

Declarations

Ethics approval and consent to participate

Not applicable.
Consent for publication

Not applicable.

Availability of data and materials

All code needed to reproduce the results are available on Github, https://github.com/bjoris33/humanGutConj_Microbiome

Competing interests

The authors declare that they have no competing interests.

Funding

Supported by CIHR Project Grant (PJT-159708) to D.R.E. and G.B.G. T.A.H. was supported by an NSERC PGs-D scholarship. B.R.J was supported by the Schulich School of Medicine Dean’s Research Scholarship.

Authors’ Contributions

BRJ designed the experiments, analyzed and interpreted the data, and wrote the manuscript. TSB analyzed the data. TAH interpreted the data. DRE designed the experiments, interpreted the data, and provided funding. GBG designed the experiments, interpreted the data, edited the manuscript, and provided funding.

Acknowledgements

We thank Daniel Giguere for his input on the analyses and figures.

References

1. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, Cruz F de la. Mobility of plasmids. Microbiol Mol Biol Rev. 2010;74:434–52.
2. Daccord A, Ceccarelli D, Burrus V. Integrating conjugative elements of the sxt/r391 family trigger the excision and drive the mobilization of a new class of vibrio genomic islands. Mol Microbiol. 2010;78:576–88.
3. Jiang X, Hall AB, Xavier RJ, Alm EJ. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PLoS One. 2019;14:e0223680.
4. Neil K, Allard N, Grenier F, Burrus V, Rodrigue S. Highly efficient gene transfer in the mouse gut microbiota is enabled by the incl2 conjugative plasmid tp114. Commun Biol. 2020;3:523.
5. Hamilton TA, Pellegrino GM, Therrien JA, Ham DT, Bartlett PC, Karas BJ, et al. Efficient inter-species conjugative transfer of a crispr nuclease for targeted bacterial killing. Nat Commun. 2019;10:4544.

6. Peters JM, Koo B-M, Patino R, Heussler GE, Hearne CC, Qu J, et al. Enabling genetic analysis of diverse bacteria with mobile-crispri. Nat Microbiol. 2019;4:244–50.

7. Citorik RJ, Mimeo M, Lu TK. Sequence-specific antimicrobials using efficiently delivered rna-guided nucleases. Nat Biotechnol. 2014;32:1141–5.

8. Bikard D, Euler CW, Jiang W, Nussenzweig PM, Goldberg GW, Duportet X, et al. Exploiting crispr-cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol. 2014;32:1146–50.

9. Gomaa AA, Klumpe HE, Luo ML, Selle K, Barrangou R, Beisel CL. Programmable removal of bacterial strains by use of genome-targeting crispr-cas systems. mBio. 2014;5:e00928–13.

10. Francia MV, Varsaki A, Gargillán-Barcia MP, Latorre A, Drainas C, Cruz F de la. A classification scheme for mobilization regions of bacterial plasmids. FEMS Microbiol Rev. 2004;28:79–100.

11. Nash RP, Habibi S, Cheng Y, Lujan SA, Redinbo MR. The mechanism and control of DNA transfer by the conjugative relaxase of resistance plasmid pCU1. Nucleic Acids Res. 2010;38:5929–43.

12. Becker EC, Meyer RJ. Recognition of oriT for DNA processing at termination of a round of conjugal transfer. J Mol Biol. 2000;300:1067–77.

13. Fronzes R, Christie PJ, Waksman G. The structural biology of type IV secretion systems. Nat Rev Microbiol. 2009;7:703–14.

14. Cabezón E, Ripoll-Rozada J, Peña A, Cruz F de la, Arechaga I. Towards an integrated model of bacterial conjugation. FEMS Microbiol Rev. 2015;39:81–95.

15. Bhatty M, Laverde Gomez JA, Christie PJ. The expanding bacterial type IV secretion lexicon. Res Microbiol. 164:620–39.

16. Guglielmini J, Cruz F de la, Rocha EPC. Evolution of conjugation and type IV secretion systems. Mol Biol Evol. 2013;30:315–31.

17. Guglielmini J, Quintais L, Gargillán-Barcia MP, Cruz F de la, Rocha EPC. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet. 2011;7:e1002222.

18. Shterzer N, Mizrahi I. The animal gut as a melting pot for horizontal gene transfer. Can J Microbiol. 2015;61:603–5.

19. Kaufman JH, Terrizzano I, Nayar G, Seabolt E, Agarwal A, Slizovskiy IB, et al. Integrative and conjugative elements (ice) and associated cargo genes within and across hundreds of bacterial genera. bioRxiv. 2020. doi:10.1101/2020.04.07.030320.
20. Arredondo-Alonso S, Willems RJ, Schaik W van, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genom. 2017;3:e000128.

21. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568:499–504.

22. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: Predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. 2018;46:e35.

23. Maguire F, Jia B, Gray KL, Lau WYV, Beiko RG, Brinkman FSL. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. Microb Genom. 2020;6.

24. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31:926–32.

25. Hall AB, Yassour M, Sank J, Garner A, Jiang X, Arthur T, et al. A novel ruminococcus gnavus clade enriched in inflammatory bowel disease patients. Genome Med. 2017;9:103.

26. Gibson MK, Wang B, Ahmadi S, Burnham C-AD, Tarr PI, Warner BB, et al. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nat Microbiol. 2016;1:16024.

27. Bushnell B, Rood J, Singer E. BBMerge - accurate paired shotgun read merging via overlap. PLoS One. 2017;12:e0185056.

28. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics. 2014;30:2114–20.

29. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 2017;27:824–34.

30. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi’o: An advanced analysis and visualization platform for ’omics data. PeerJ. 2015;3:e1319.

31. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.

32. 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.

33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

34. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257.

35. Robertson J, Bessonov K, Schonfeld J, Nash JHE. Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. Microb Genom. 2020;6.
36. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. Journal of Open Source Software. 2017;2:205. doi:10.21105/joss.00205.

37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.

38. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.

39. 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;25:1043–55.

40. Meijenfeldt FAB von, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with cat and bat. Genome Biol. 2019;20:217.

41. Olm MR, Brown CT, Brooks B, Banfield JF. DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.

42. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.

43. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498:99–103.

44. Pehrsson EC, Tsukayama P, Patel S, Mejia-Bautista M, Sosa-Soto G, Navarrete KM, et al. Interconnected microbiomes and resistomes in low-income human habitats. Nature. 2016;533:212–6.

45. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, et al. Metagenome sequencing of the hadza hunter-gatherer gut microbiota. Curr Biol. 2015;25:1682–93.

46. Sankaranarayanan K, Ozga AT, Warinner C, Tito RY, Obregon-Tito AJ, Xu J, et al. Gut microbiome diversity among cheyenne and arapaho individuals from western Oklahoma. Curr Biol. 2015;25:3161–9.

47. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490:55–60.

48. Zhao G, Vatanen T, Droit L, Park A, Kostic AD, Poon TW, et al. Intestinal virome changes precede autoimmunity in type i diabetes-susceptible children. Proc Natl Acad Sci U S A. 2017;114:E6166–75.

49. Karaolis DK, Somara S, Maneval DR, Johnson JA, Kaper JB. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. Nature. 1999;399:375–9.

50. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell. 2019;176:649–662.e20.
51. Milani C, Duranti S, Bottacini F, Casey E, Turroni F, Mahony J, et al. The first microbial colonizers of the human gut: Composition, activities, and health implications of the infant gut microbiota. Microbiol Mol Biol Rev. 2017;81.

52. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nature Biotechnology. 2020. doi:10.1038/s41587-020-0422-6.

53. Giguere DJ, Bahcheli AT, Joris BR, Paulsen JM, Gieg LM, Flatley MW, et al. Complete and validated genomes from a metagenome. bioRxiv. 2020. doi:10.1101/2020.04.08.032540.

54. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang X-S, Davis-Richardson A, et al. Metagenomic binning and association of plasmids with bacterial host genomes using dna methylation. Nat Biotechnol. 2018;36:61–9.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementalfigures.pdf
- supplementaltable1.txt
- supplementaltable2.txt
- supplementaltable3.txt