Identification of type IV conjugative systems that are systematically excluded from metagenomic bins

Benjamin R. Joris
The University of Western
Tyler S. Browne
The University of Western
Thomas A. Hamilton
The University of Western
David R. Edgell
The University of Western
Gregory B. Gloor (ggloor@uwo.ca)
The University of Western

Research Article

Keywords:

Posted Date: March 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1428512/v1

Identification of type IV conjugative systems that are systematically excluded from metagenomic bins

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

07 March, 2022

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 related methods to identify conjugative elements in the human gut microbiome. Conjugative systems can effectively be identified from metagenomic assemblies either using a curated sets of profile hidden Markov models or by searching against large-scale databases, such as UniRef90. Both methods were successful at identifying type IV conjugative systems with profile hidden Markov models being faster, but less sensitive than alignment to the UniRef database. 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.

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.

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 *trans* where the conjugative machinery is on a different genetic element than the transferred element, and the remainder are mobilizable in *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 by conjugation, 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 catalyze the nicking reaction at the oriT DNA sequence [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 conjugative. 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 less than optimal 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. Second, we searched predicted protein sequences from those same assemblies versus UniRef90 [24] for proteins involved in conjugation to identify conjugative systems. Both methods were able to find conjugative systems in raw

metagenomic assemblies with pHMMs being computationally more efficient but less sensitive. Finally, we demonstrate that the majority of conjugative systems produced by a metagenomic assembly are not included in high-quality bins that are used as proxies for bacterial genomes in metagenomic analysis pipelines. 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* as follows (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].

Identification of conjugative systems using Profile hidden Markov models

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]. Instructions for installation of T4SS pHMMs into Anvi'o can be found in the online code repository (https://github.com/bjoris33/humanGutConj_Microbiome). 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]. Taxonomic prediction of the contigs was conducted with Kaiju version 1.7.2 utilizing the RefSeq non-redundant protein database [33]. MOB-suite verion 1.4.9.1 was utilized to characterize the incompatibility grouping of the conjugative system, if possible [34]. PlasFlow version 1.1.0 was used to classify whether the system was chromosomally integrated or a plasmid [22].

Identification of conjugative systems using protein alignments to the UniRef90 database

The contigs of the raw metagenomic assemblies had their ORFs predicted using Prodigal version 2.6.3 [32]. The predicted ORFs were then aligned to the UniRef90 database [24] using the `blastp` module of DIAMOND version 0.9.14 [35]. By using keywords such as "conjugal" or "mobilization", the protein alignments were search for contigs that contained annotations for both a relaxase and either a type IV secretion system protein or a type IV coupling protein. Through manual curation, type IV secretion system proteins and

coupling proteins often shared identical or very similar annotation entries in the UniRef90 database, so the decision was made not to distinguish between the two.

Binning of Assemblies

For each assembly, all 101 samples were mapped to the contigs using Bowtie2 [36]. 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/).

Results

Profile hidden Markov models and database alignment successfully identify conjugative elements from metagenomic assemblies

Fifty-one 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). Two separate methods we employed to search for contigs containing type IV conjugative proteins. For the pHMM method of identifying conjugative systems, 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.

The second method of identifying conjugative systems utilizes the UniRef90 database by aligning the predicted ORFs to it using DIAMOND [35]. The alignment results are searched using a keyword strategy for contigs that contain an alignment for a relaxase or mobilization protein and an alignment for a type IV secretion or type IV coupling protein. From the pre-term infant cohort assemblies 242 of 470500 contigs met the described criteria, and 4244 of the 15100646 contigs from the IBD cohort met the same criteria.

The two outlined methods represent potentially complimentary methods of tackling the same problem– identifying conjugative systems from a pool of metagenomic-assembled contigs. There is a large-degree of overlap between the two methods, however alignment to the UniRef90 database appears to be much more sensitive with only 280 of the 4486 identified conjugative systems also identified using the pHMM method (Figure 3). While it may be less sensitive, the pHMM method of identifying conjugative systems has a much smaller computationally footprint as it does not rely on aligning to a large protein database, but rather to

a small and specific set of profile hidden Markov models.

The majority of conjugative systems identified are omitted from metagenomic bins

Metagenomic assemblies from two distinct cohorts were binned using MetaBAT2 [38] to 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 4). 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.

For the conjugative systems established using alignments to the UniRef90 database, there is an even lower rate of binning (Figure 5). Of the 4486 conjugative systems, only 287 of them were binned–a rate of 6.4%. Again, a number of the bins that do form are low quality bins below 1mb in size that may be the collection of contigs that form a conjugative plasmid.

Discussion

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 is another factor that leads to 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.

We have outlined two methods for identifying contigs carrying potentially functional type IV conjugative systems from raw metagenomic assemblies. Using a curated set of pHMMs of the three main classes of type IV conjugative proteins (relaxases, secretion proteins, and coupling proteins), we were able to classify 364 total contigs as being potentially conjugative. In comparison, the method that utilized predicted protein alignments to the UniRef90 database found 4486 contigs that met the criteria, which indicates that it may be the more sensitive method. However, aligning all predicted open reading frames found in a metagenomic assembly to the full UniRef90 database is a computationally expensive task. Considering that many of

the conjugative systems identified by pHMMs are also found by the protein alignment method (280 of 364 contigs), using pHMMs may be appropriate in as a first pass method or in situations where computational resources are scarce.

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 binning rate of contigs carrying type IV conjugative systems identified by pHMMs and alignment to UniRef90 was considerably lower than the background binning rate of equivalently sized contigs (25.8% and 6.4% compared to 70.4%, respectively). Many of the binned conjugative systems were not within a bin that would pass the quality cutoff to be included in a curated MAG 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 could differ between cohorts and require special consideration to ensure that they are included in metagenomic 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, digestion, and host effector molecules. 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 4 and Figure 5), 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 UniRef90 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, expanding the curated set of pHMMs could increase the sensitivity of the method to detect conjugative systems and creating a curated set of conjugation proteins from the UniRef90, instead of exhaustively annotating with the full database, could improve the computational efficiency of the other method. Additionally, 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. For instance, long-read assembly permits the circularization of genomes and plasmids [40, 41] and the binning of plasmids to their cognate genomes using methylation data [42], 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.

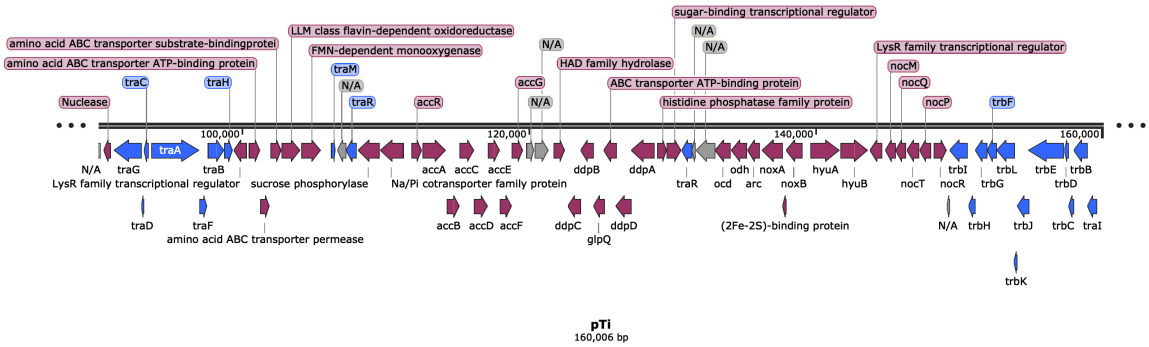

Figure 1: Example schematic of the gene organization of a bacterial conjugation system on the Agrobacterium tumefaciens pTi plasmid.

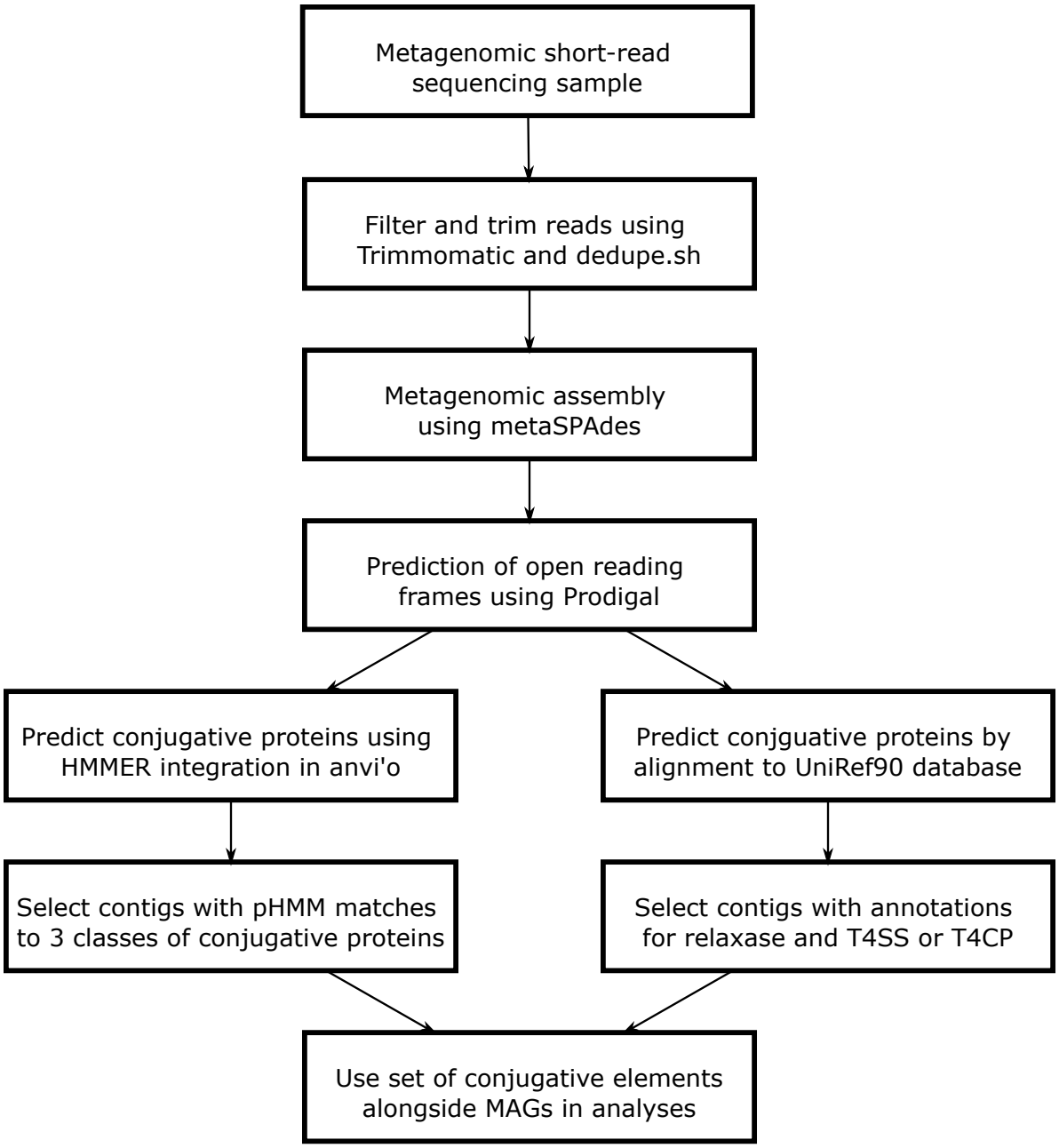

Figure 2: Overview of methods employed in this study. In the left panel is the workflow used to identify conjugative systems from previously assembled human gut bacterial genomes. The right panel outlines the workflow for the assembly of select North American samples and the use of pHMMs to identify the conjugative systems.

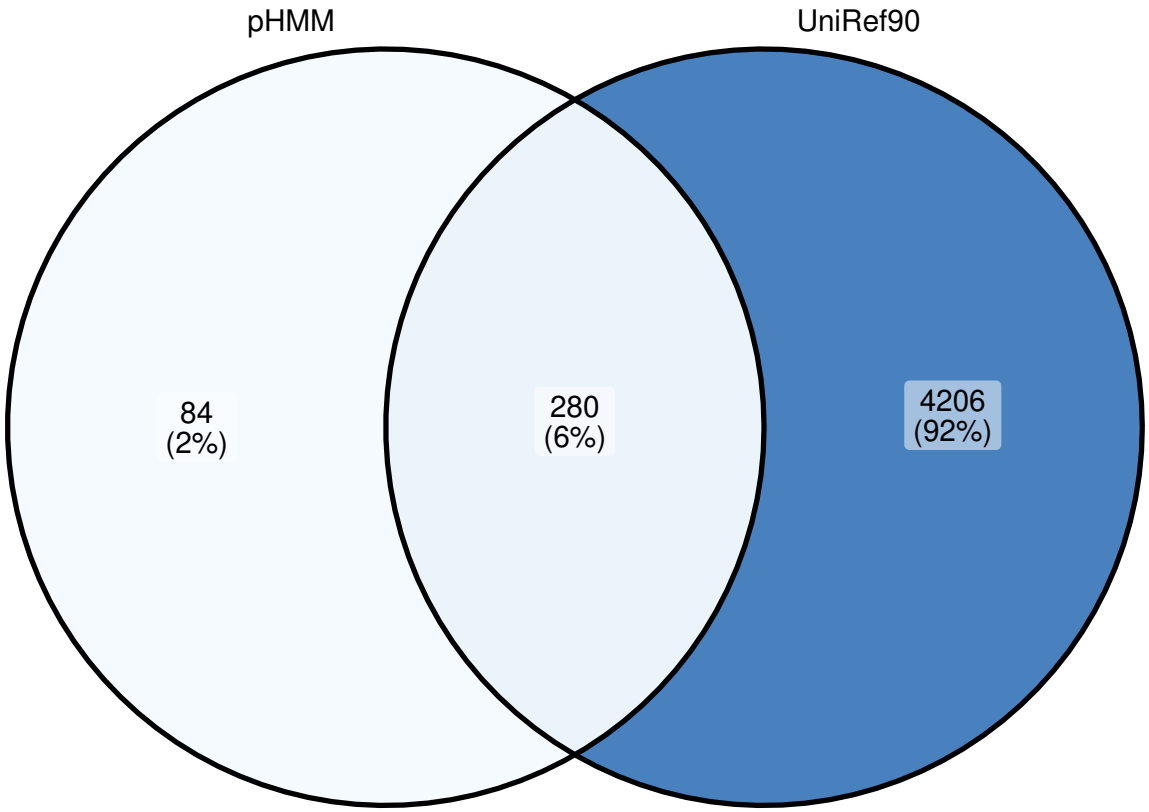

Figure 3: Venn diagram illustrating the overlap of the two methods of identifying type IV conjuagtive systems from the 101 assembled metagenomic samples.

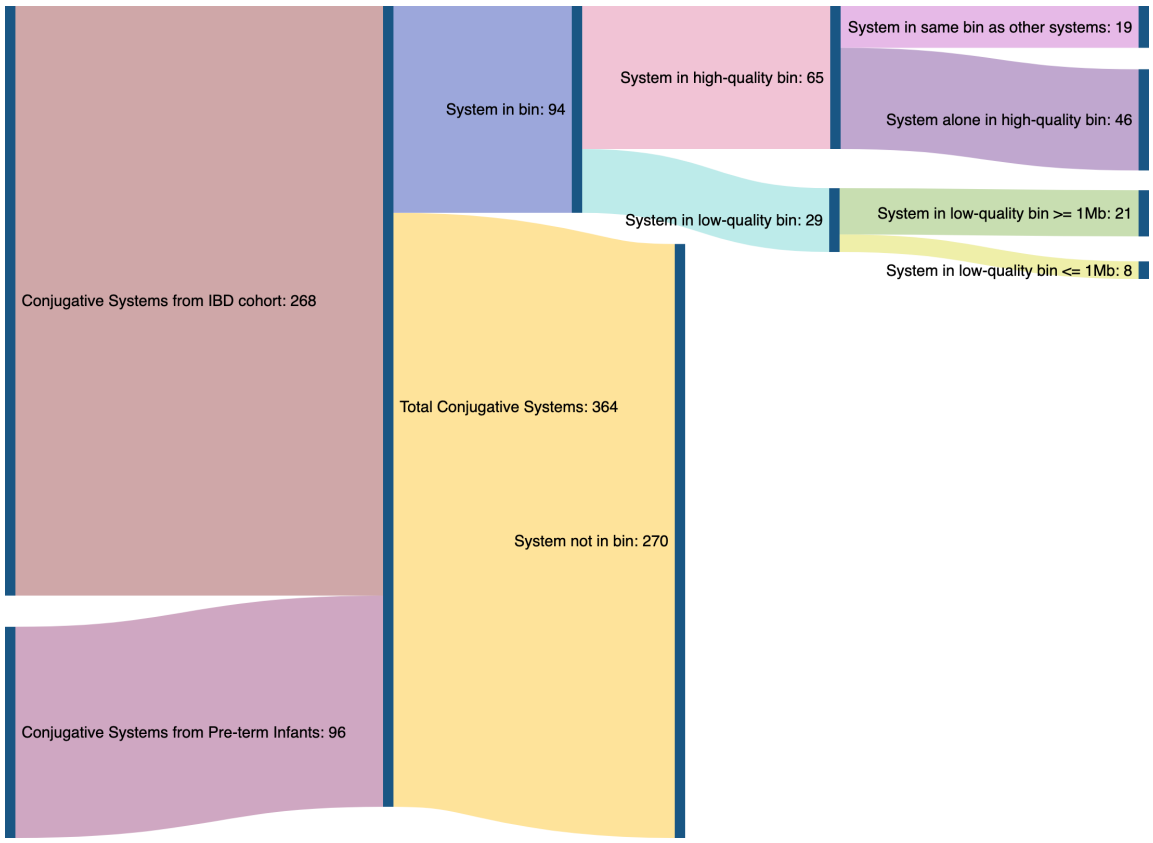

Figure 4: Sankey diagram representing the flow of 364 contigs containing conjugative systems identified using pHMMs into bins generated by MetaBAT2 from assembled data.

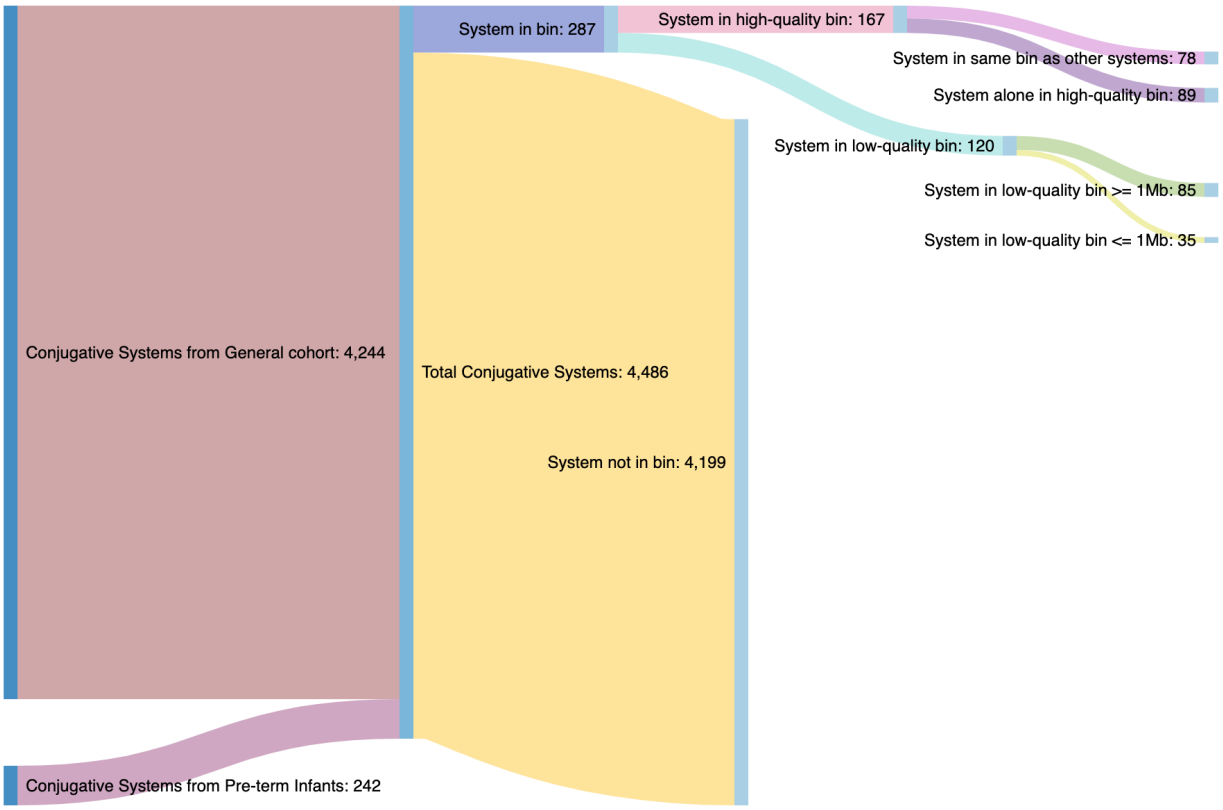

Figure 5: Sankey diagram representing the flow of 4486 contigs containing conjugative systems identified using predicted protein alignments to the UniRef90 database into bins generated by MetaBAT2 from assembled data.

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

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

34. 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.

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

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

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. 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.

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

42. 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.

SRR3131692	pre-term infant
SRR3131694	pre-term infant
SRR3131696	pre-term infant
SRR3131698	pre-term infant
SRR3131700	pre-term infant
SRR3131702	pre-term infant
SRR3131704	pre-term infant
SRR3131706	pre-term infant
SRR3131708	pre-term infant
SRR3131710	pre-term infant
SRR3131712	pre-term infant
SRR3131714	pre-term infant
SRR3131716	pre-term infant
SRR3131718	pre-term infant
SRR3131720	pre-term infant
SRR3131722	pre-term infant
SRR3131724	pre-term infant
SRR3131726	pre-term infant
SRR3131728	pre-term infant
SRR3131730	pre-term infant
SRR3131732	pre-term infant
SRR3131734	pre-term infant
SRR3131736	pre-term infant
SRR3131738	pre-term infant
SRR3131740	pre-term infant
SRR3131742	pre-term infant
SRR3131744	pre-term infant
SRR3131746	pre-term infant
SRR3131748	pre-term infant
SRR3131750	pre-term infant
SRR3131752	pre-term infant
SRR3131754	pre-term infant
SRR3131756	pre-term infant
SRR3131758	pre-term infant
SRR3131760	pre-term infant
SRR3131762	pre-term infant
SRR3131764	pre-term infant
SRR3131766	pre-term infant
SRR3131768	pre-term infant
SRR3131770	pre-term infant
SRR3131772	pre-term infant
SRR3131774	pre-term infant
SRR3131776	pre-term infant
SRR3131778	pre-term infant
SRR3131780	pre-term infant
SRR3131782	pre-term infant
SRR3131784	pre-term infant
SRR3131786	pre-term infant
SRR3131788	pre-term infant
SRR3131790	pre-term infant
SRR3131812	pre-term infant
SRR5650110	general
SRR5650111	general
SRR5650112	general

SRR5650113	general
SRR5650114	general
SRR5650115	general
SRR5650116	general
SRR5650117	general
SRR5650118	general
SRR5650119	general
SRR5650120	general
SRR5650121	general
SRR5650122	general
SRR5650123	general
SRR5650124	general
SRR5650125	general
SRR5650126	general
SRR5650127	general
SRR5650128	general
SRR5650129	general
SRR5650130	general
SRR5650131	general
SRR5650132	general
SRR5650133	general
SRR5650134	general
SRR5650135	general
SRR5650136	general
SRR5650137	general
SRR5650138	general
SRR5650139	general
SRR5650140	general
SRR5650141	general
SRR5650142	general
SRR5650143	general
SRR5650144	general
SRR5650145	general
SRR5650146	general
SRR5650147	general
SRR5650148	general
SRR5650149	general
SRR5650150	general
SRR5650151	general
SRR5650152	general
SRR5650153	general
SRR5650154	general
SRR5650155	general
SRR5650156	general
SRR5650157	general
SRR5650158	general
SRR5650159	general