Bacterial conjugation systems identified from metagenomic assemblies suggest structural differences between cohorts

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

30 November, 2020

Abstract

Background

Bacterial conjugation enables the exchange of genetic elements throughout environments, including the human gut microbiome. Conjugative elements can carry and transfer clinically relevant metabolic pathways such as antibiotic resistance and bile acid metabolism, which makes precise identification of these systems in metagenomic samples critically important.

Results

Here, we outline two distinct methods of identifying conjugative systems in the human gut microbiome. Using the abundances of these systems, we show that conjugative systems exhibit strong population and age-level stratification. Additionally, we find that overall abundance of conjugative systems is not an informative metric to use, regardless of the method of identifying the systems. Finally, we demonstrate that the majority of assembled conjugative systems are not included within metagenomic bins, and only a small proportion of the binned conjugative systems are included in "high-quality" metagenomic bins.

Conclusions

Our findings highlight that conjugative systems differ between a general North American and a cohort of North American pre-term infants, revealing a potential use as an age-related biomarker. Furthermore, conjugative systems can group other geographical-based cohorts. Our findings also underline the need to identify and analyze conjugative systems outside of standard metagenomic binning pipelines.

Background

Bacteria can acquire exogenous DNA by conjugation of integrative conjugative elements (ICEs) and conjugative plasmids. Roughly half of the known plasmids are mobilizable in trans (i.e. conjugative system on a different genetic element), and about half of those are also mobilizable in cis (i.e. conjugative system on the same genetic element) and are defined as conjugative [1]. Conjugative elements often contain antibiotic resistance genes, but also can harbour useful, biodegradative genes [2]. Furthermore, conjugative systems can serve as vectors to introduce metabolic pathways into the gut microbiome [3, 4], and characterizing the full complement of conjugative systems in the human gut could could expand the number of useable vectors for these applications. Precise identification of conjugative systems from metagenomic samples would 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 a relaxase protein that nick the DNA at the origin of transfer (oriT) sequence and unwind the DNA [1, 5]. Relaxase proteins contain a conserved histidine triad that coordinate a metal ion and tyrosine residues that are responsible for binding of the oriT sequence and for catalysis of the nicking reaction [6, 7]. While having a relaxase and an oriT allow for a genetic element to be mobilizable, a full compliment of type IV secretion system and coupling proteins are required for a sequence to be mobilizable in cis. The synteny is highly conserved among all conjugative systems [8]. There are 12 proteins involved in the formation of the complex that transfers the DNA-relaxase complex from one bacterial cell to another in the well-studied conjugative system identified in *Agrobacterium tumefaciens* [8, 9]. The VirB4 protein homologs are generally similar to the phylogeny of the bacteria harbouring them [10] and are useful to classify conjugative systems [11]. The many highly-conserved genes involved in conjugation allow classification of genetic elements as potentially conjugative if they contain multiple sequences annotated as belonging to the components of the type IV secretion system [12] (Figure 1).

Previous work has identified novel conjugative systems in the human and animal gut microbiomes, but the focus was mainly on ICEs [2, 13]. Identifying conjugative plasmids from a short-read metagenomic assembly is difficult for several reasons. First, it is nearly impossible to assemble circularized plasmids from short-read sequencing data [14]. Second, the contiguous DNA sequences (contigs) that compose metagenome-assembled genomes (MAGs) are binned together based on sequence composition and coverage. Binning with the 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; this is generally not the case because conjuagtive sytems are usually more AT rich than the cognate chromosome [1]. Alternate methods must be employed to assemble and identify conjugative plasmids from the metagenomic data because nearly 80% of the non-redundant set of genomes from the human-gut microbiome being from difficult-to-culture or unculturable species whose genomes had to be assembled metagenomically [15].

Here, we show that conjugative systems can be identified using two distinct methods (Figure 2). Firstly, we were able to identify conjugative systems directly from metagenomic assemblies of general North American and North American pre-term infant samples, identifying conjugative systems using profile HMMs (pHMMs) [12]. Second, using a separate approach, we searched for UniRef90 [16] annotations of proteins involved in conjugation to identify

conjugative systems from a human gut microbiome genome set, and with this approach differences between cohorts in the abundances of extracted systems can be recognized. The differences between cohorts found using the two methods were not identical, but both methods did illustrate that the cohorts were distinct from one another. This finding suggests some level of incompleteness or bias in the conjugative systems found in the non-redundant genome set. 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 genomic content of bacteria.

Methods

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

Samples belonging to a general North American (n=50) and a North American pre-term infant cohort (n=51) were assembled *de novo* (Supplementary Table 2). Reads from these samples were downloaded from the Sequence Read Archive using the SRA toolkit version 2.9.2, deduplicated with `dedupe.sh` [17], and trimmed with Trimmomatic version 0.36 [18] with options `LEADING:10 TRAILING:10`. Processed reads were assembled sample-by-sample using SPAdes version 3.14.0, option `--meta` [19]. The resultant assemblies were imported into Anvi'o version 6.0 [20] where the presence of T4SS, T4CP, and relaxase proteins were predicted using the `anvi-run-hmms` module, which integrates HMMER3 functionality [21]. 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 [22] to the UniRef90 database [16]. Before mapping, regions of the contigs where annotations for conjugative proteins were located, with no more than 20 ORFs between UniRef90 annotations for conjugative proteins. Extraction of the regions was to avoid an artificially high proportion of reads mapping in samples where the contig is present, but the ICE has not integrated (Supplementary Figure 1). Taxonomic prediction of the contigs was conducted with Kaiju version 1.7.2 utilizing the RefSeq non-redundant protein database [23]. The proportion of reads mapping to the conjugative systems was extracted from the Bowtie2 output, and the mapping data was visualized using Anvi'o. 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 used for clustering by hdbscan [24].

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/) [15]. These genomes were assembled from 13,133 metagenomic samples using SPAdes [19] and binned using MetaBAT2 [25]. The quality of binned genomes were assessed using CheckM [26]. 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. To dereplicate the genome bins, dRep was to used to cluster the genomes at 99% sequence identity [27], creating a set of 2505 genomes.

Identifying conjugative systems in reference human gut metagenome set

ORFs were predicted in the genome by Prodigal version 2.6.3 [22]. The predicted protein sequences were then aligned to the UniRef90 database [16] using the Diamond protein aligner version 0.9.14 [28]. Using a word-search strategy contigs were extracted from the genomes if they contained annotations for a relaxase/mobilization protein and a type IV secretion/type IV coupling protein.

Mapping of short-read data from various cohorts to conjugative systems

Short-read data from 785 samples (Supplementary Table 1) were downloaded from the Sequence Read Archive using the SRA toolkit version 2.9.2. The downloaded reads were deduplicated with `dedupe.sh` [17] and trimmed with Trimmomatic version 0.36 [18] using options `LEADING:10 TRAILING:10`. As previously explained, regions with conjugative systems were extracted from the identified contigs to avoid artificially high average mapping coverage (Supplementary Figure 1). The processed read data were mapped to the extracted conjugative systems using Bowtie2 version 2.3.5 [29] with the settings `--no-unal --no-mixed --no-discordant`. Proportions of reads mapping to conjugative systems were extracted from Bowtie2 output. Alignment files were sorted and indexed using SAMtools version 1.10 [30] and imported into Anvi'o version 6.0 for visualization [20]. Raw counts of reads mapping to the extracted conjugative systems were transformed using a centered log-ratio. The principal component coordinates of the first 3 components were used for clustering by hdbscan [24].

Binning of Assemblies

For each assembly, all 101 samples were mapped to the contigs using Bowtie2 [29]. The mapping files were sorted and indexed with SAMtools [30] and then the assemblies were binned using MetaBAT2 version 2.12.1 [25]. CheckM version 1.1.2 was used to assess the quality of the resultant bins [26]. 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

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

Fifty-one samples from a pre-term infant cohort and fifty from a general North American cohort were assembled sample-by-sample using metaSPAdes [19] to identify conjuagtive systems from a full pool of assembled contigs (i.e. not binned) and to compare the abundances of these systems between cohorts. 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 only 96 contigs met the criteria, whereas 268 contigs from the general cohort did. 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 systems in the two cohorts are distinct (Figure 3). Conjugative systems belonging to the *Proteobacteria* phylum were only assembled from pre-term infant samples and did not have any apparent occurrence in the general population. Furthermore, most identified conjugative contigs were private to the cohort they were assembled from. When the log-transformed principal component analysis data were clustered with hdbscan [24], the two cohorts formed separate clusters (Figure 4, Supplementary Figure 2). These findings show that North American pre-term infants have a different array of conjugative systems than a general North American cohort.

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

To explore abundances of conjugative systems in a greater number of cohorts, without having to conduct computationally-expensive metagenomic assemblies, 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 [15]. A total of 1598 contigs from 787 genomes that contain UniRef90 annotations for relaxase/mobilization and T4SS/T4CP proteins were identified. From these contigs, 3216 subregions where conjugative protein annotations were concentrated on the contig were extracted (Supplementary 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 (Supplementary 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 much different way (Figure 5). Virtually none 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; both have an overall lower apparent abundance

of conjugative systems compared to the other non-infant cohorts, particularly in the *Bacteroidetes* phylum. The other four cohorts are appear similar with regards to the presence and absence of the conjugative systems. When the principal components of the centered log-ratio transformed data, excluding infants due to their sparsity, were clustered using hdbscan [24], the cohorts separated into three distinct clusters (Figure 6, Supplementary Figure 3). As suggested by Figure 5, the West African and South American samples almost exclusively clustered together. Not readily apparent from the cladogram was the East Asian cohort clustering primarily on its own. The North American Indigenous, North American general, and Western European general clustered together as well. Like the conjugative systems identified from the *de novo* assemblies of short-read data, the abundances of conjugative systems identified from a human gut metagenome set separated cohorts into distinct groups.

Proportions of reads mapping to conjugative systems is inconsistent between methods

For the cohorts that were examined using both methods of conjugative system identification, the proportions of total reads mapping to the identified conjugative systems were not equal between methods (Figures 7, 8). From the conjugative systems identified from the general North American cohort assembly of the short reads the mean proportion of reads mapping was 0.00619 (95% CI [0.00219,0.0110]), whereas the mean proportion of reads mapping to the conjugative systems identified from the genome set was (0.0269 95% CI [0.0125,0.0464]). This suggests that the assemblies of the short read sequences were not able to successfully capture the full diversity of conjugative systems found in an average North American individual. Inversely, for the pre-term North American infants, the mean proportion of total reads mapping to conjugative systems identified from the assemblies was 0.00333 (95% CI [0.0000182,0.0239]), whereas from the reference genome set the mean proportion was (0.00271 95% CI [0.000245,0.0121]). In terms of the composition of the reference gut metagenome set, it is probable that the bulk of the bacterial genomes are sourced from deeply sequenced cohorts, like a general North American population, rather than a niche cohort like pre-term infants. This results in a more precise abundance estimate of conjugative systems in the general North American cohort than in the pre-term infant cohort. These findings together suggest that the total proportion of reads mapping to conjugative systems is not particularly informative and these data should be treated as compositions with relative abundances of features being compared between groups.

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

The assemblies were binned using MetaBAT2 [25], which was also used to bin the MAGs in the human gut genome set used in the prior analyses[15] 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 9). For the 94 contigs included in metagenomic bins, 65 of those were found in high-quality bins (>90% completion and <5% redundancy). Among the 29 contigs included in bins that

do not meet the aforementioned threshold, 8 are within bins that are greater than or equal to 1 megabase in size, potentially suggesting that fragments of a conjugative plasmid may have binned together.

Discussion

The abundances of conjugative systems identified from MAGs and isolate genomes from the human gut differ between cohorts similarly to how abundances of human gut MAGs of different species are differential between cohorts [31]. 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 [32]. Furthermore, we observed that 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* [33]. As shown in Figure 2a, 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. The degree of difference in abundances of conjugative systems between infants and adults suggests that conjugative systems could be a potential biomarker for age.

The 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 abundances of bacterial species in these cohorts [31]. 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 abundances of conjugative systems.

Comparing the findings of Figure 7 and Figure 8, it is clear that the total percentage of reads mapping to conjugative systems is not an effective metric; the differences between the abundances in conjugative systems that were present from the genome set between the general North American and North American pre-term infants did not persist when examining the abundances of the assembled systems. Comparing Figure 2a and Figure 3a, the conjugative systems assembled in pre-term infants are almost entirely missing in the genome set. It is clear that there is a degree of incompleteness in terms of infant conjugative systems in a database consisting of primarily MAGs assembled from non-infant datasets. However, for the non-infant data, the genome set appeared to capture a larger percentage of the conjugative systems. Neither method of identifying and quantifying conjugative systems is perfect; the reference bacterial genome set may be incomplete for less commonly studied cohorts, but the sequencing depth of a single sample may not be sufficient to assemble the less abundant conjugative systems in an environment.

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 [25]. 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. 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 [25] 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. 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 [15]. 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 [2]. These cargo proteins 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 conjugative systems are rarely included in metagenomic bins, 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 [14], so it may helpful to identify bins containing conjugative systems in an attempt to cluster the fragments of plasmids present in an assembly together.

In the future, improvements in assembly and binning algorithms will continue to improve the recovery of low abundance conjugative elements and improve the completeness and accuracy of the assembled fragments. Additionally, long-read assembly permits the circularization of genomes and plasmids [34, 35], 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.

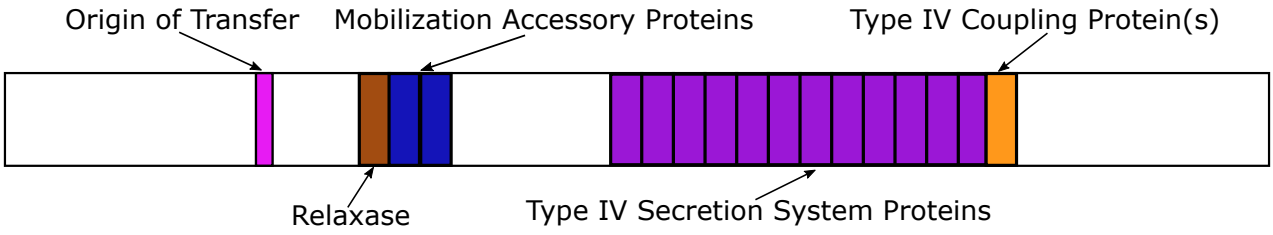

Figure 1: Example schematic of the gene organization of a bacterial conjugation system.

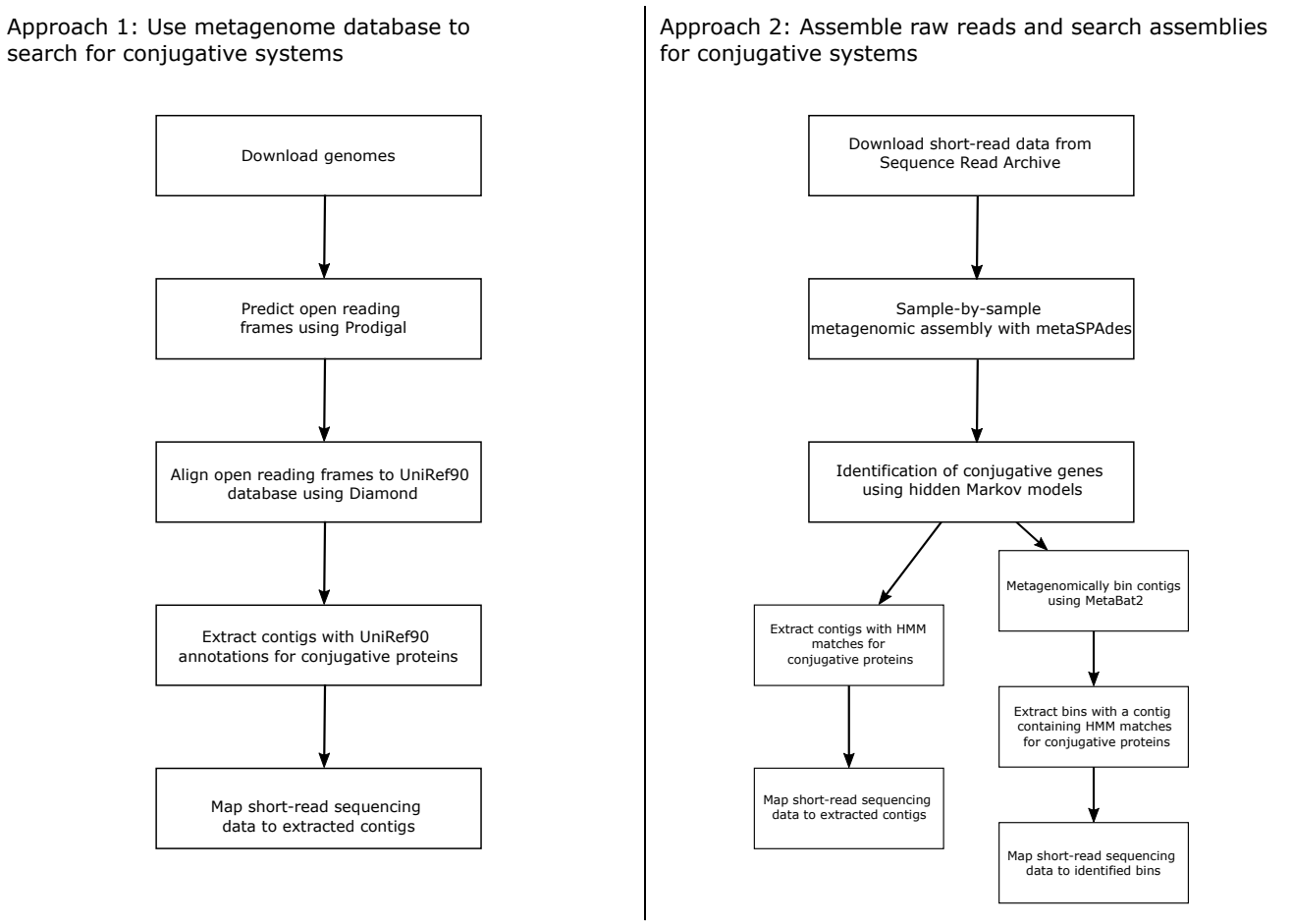

Figure 2: Overview of methods employed in this study. In the left panel is the workflow used to identify conjugative systems from previouly assembled human gut bacterial genomes. The right panel outlines the workflow for the assembly of select North American samples and the use of pHMMs to identify the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

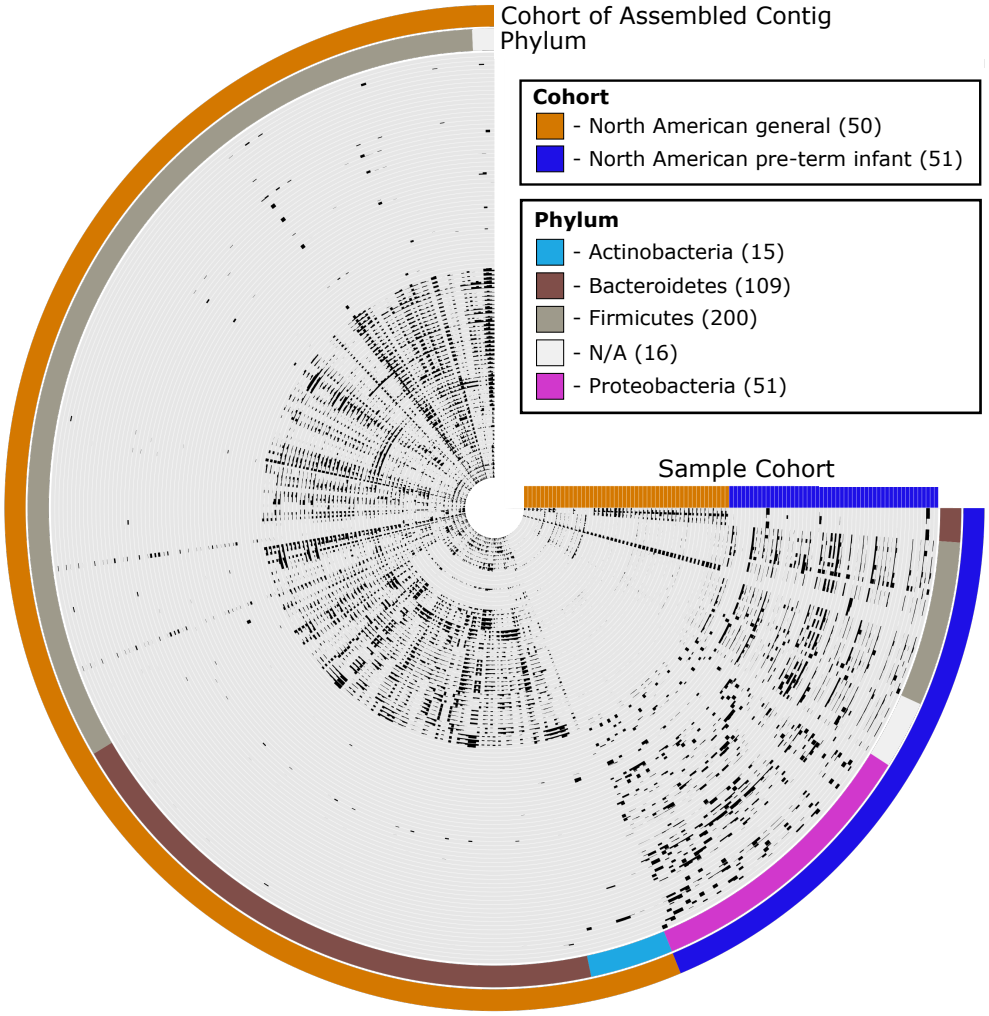

Figure 3: Anvi'o cladogram of potentially conjugative contigs from 51 North American pre-term infants samples and 50 general North American samples. Inner rings of the phylogram represent individual samples, second-most outer ring being the phylum of conjugative system as predicted by Kaiju, and the 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.

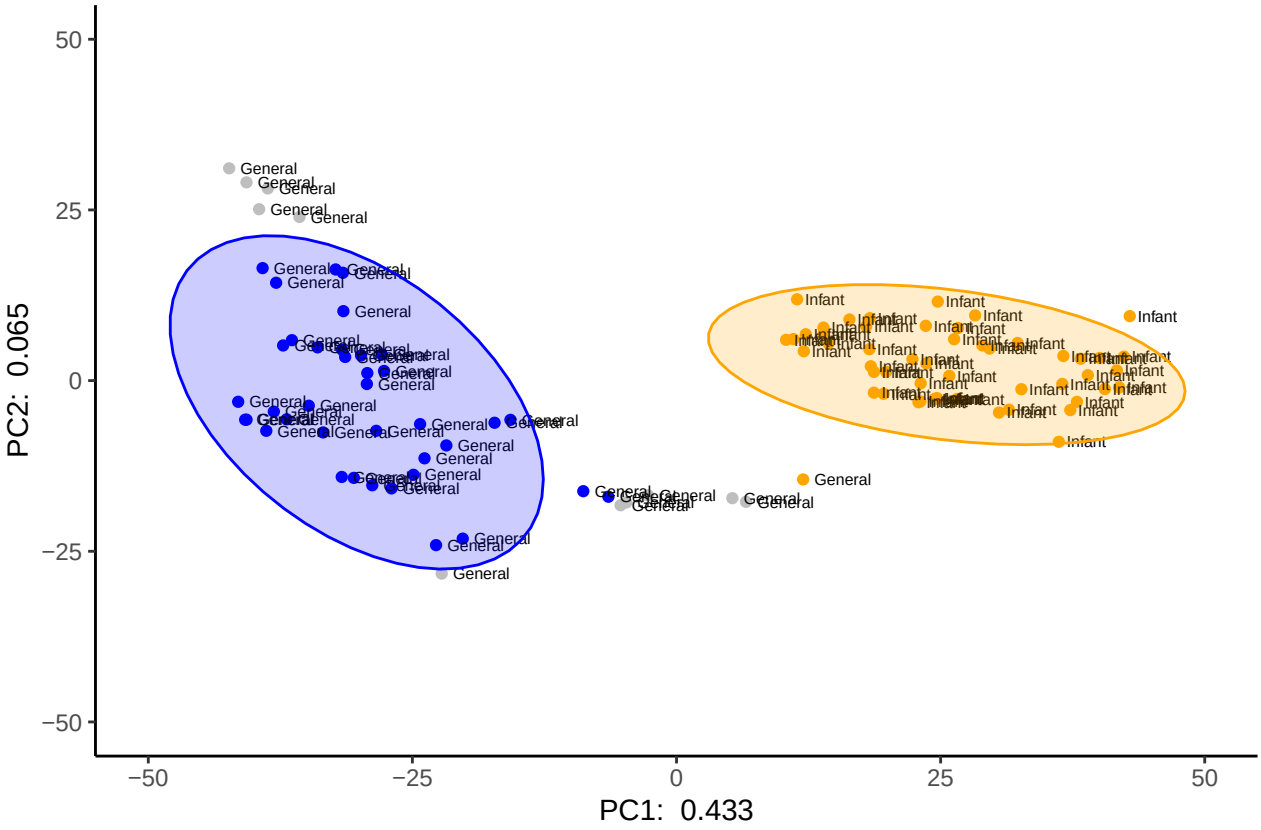

Figure 4: Clustering of the principal component coordinates of the CLR transformed abundances of the extracted conjugative regions from the assemblies of North American datasets. Ellipses represent a 95 percent CI using a multivariate t-distribution.

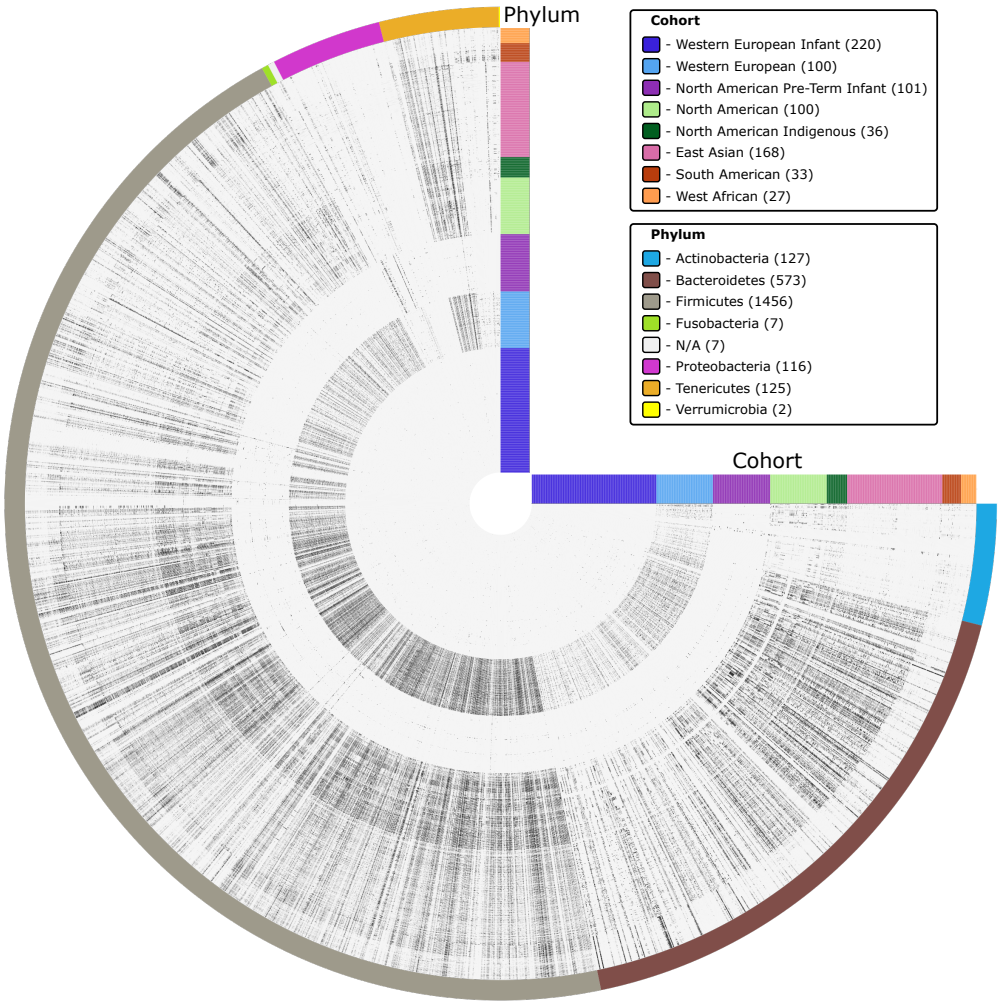

Figure 5: Anvi'o cladogram of potentially conjugative systems originating from 785 samples across 8 cohorts. Inner rings of the phylogram represent individual samples and the outermost ring being the phylum of conjugative system. Each slice of the circle phylogram are individual conjugative regions. For each point on the inner plot, the intensity of the black colouring represents the mean coverage of the system for a given sample.

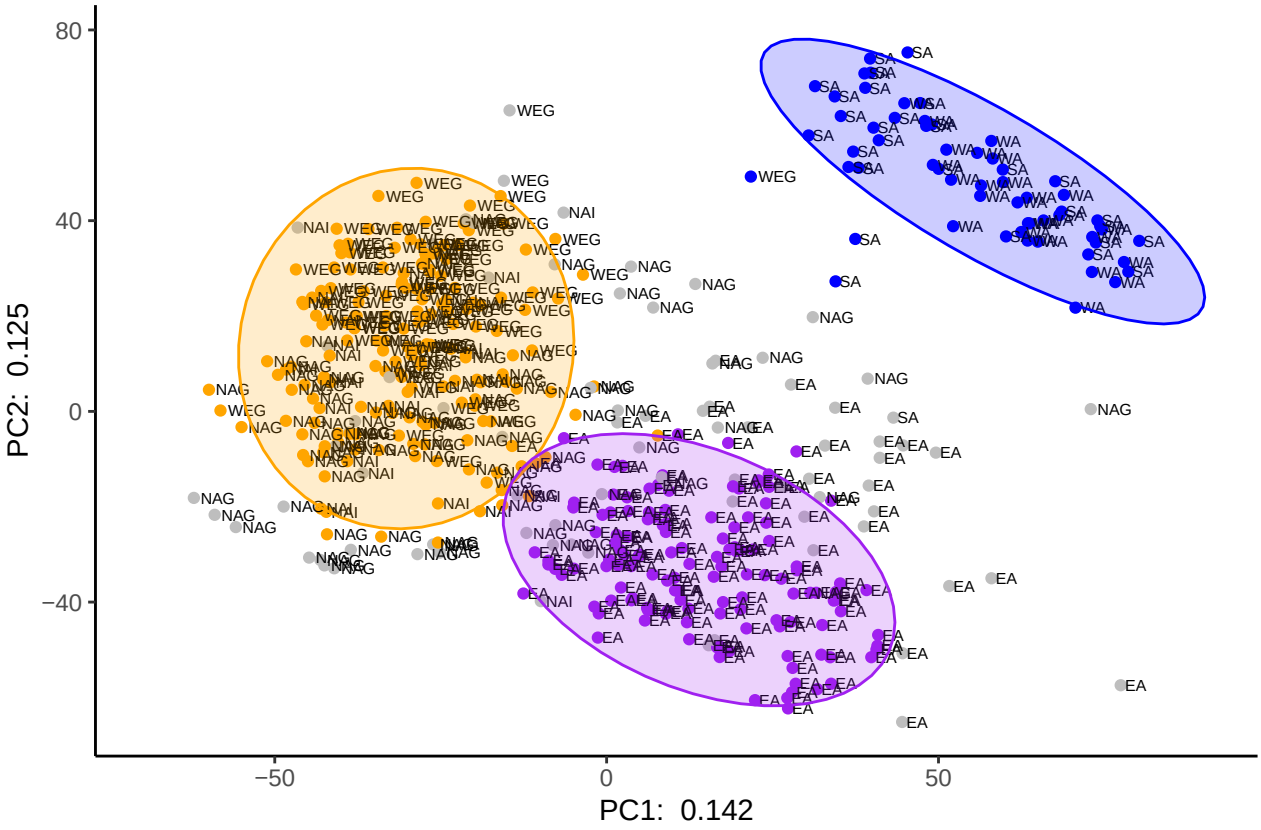

Figure 6: Clustering of the principal component coordinates of the CLR transformed abundances of the extracted conjugative regions from the genome database. Ellipses represent a 95 percent CI using a multivariate t-distribution.

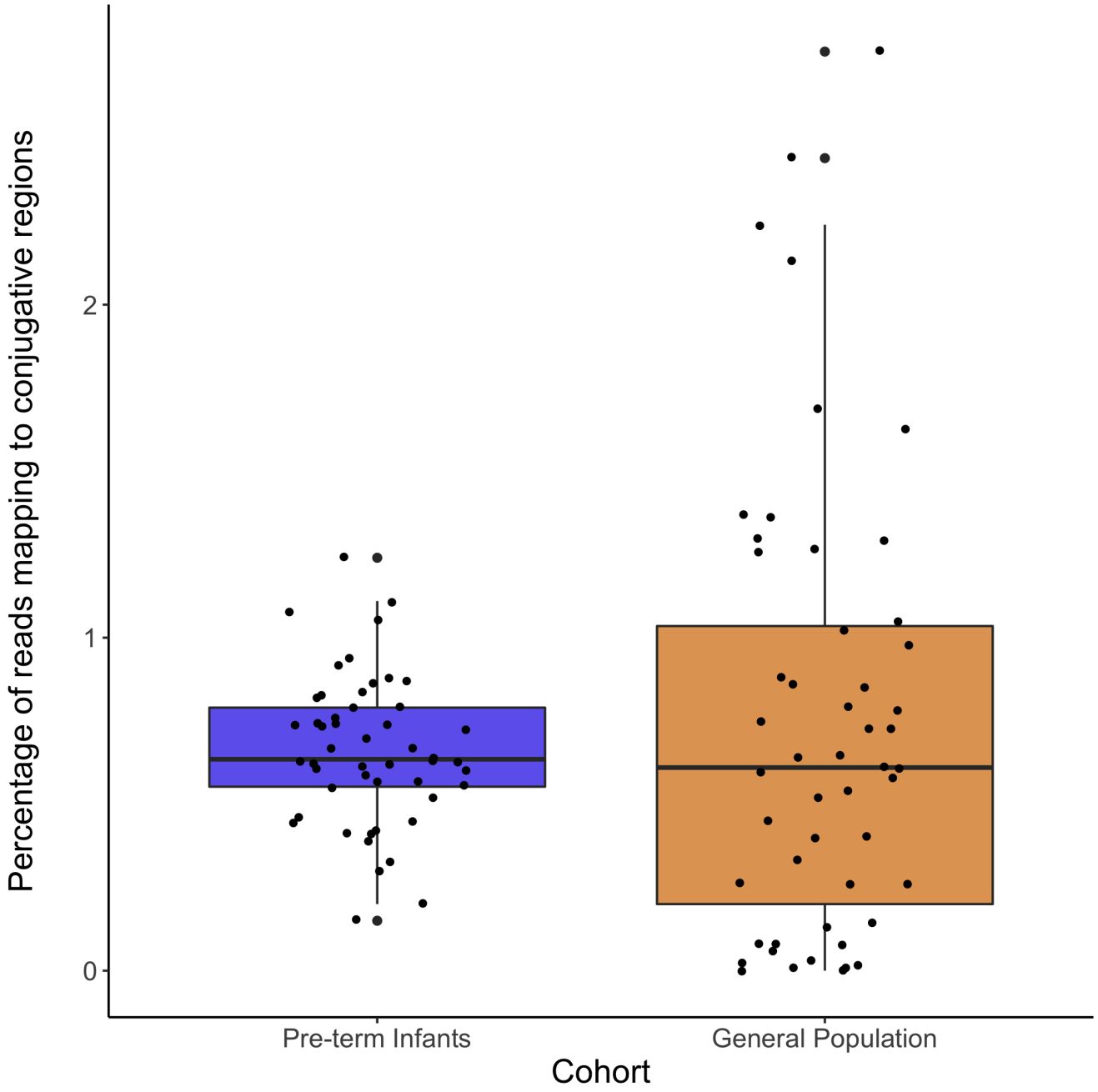

Figure 7: Plot of the percentage of total reads mapping to conjugative regions extracted from assemblies of North American datasets, separated by cohort.

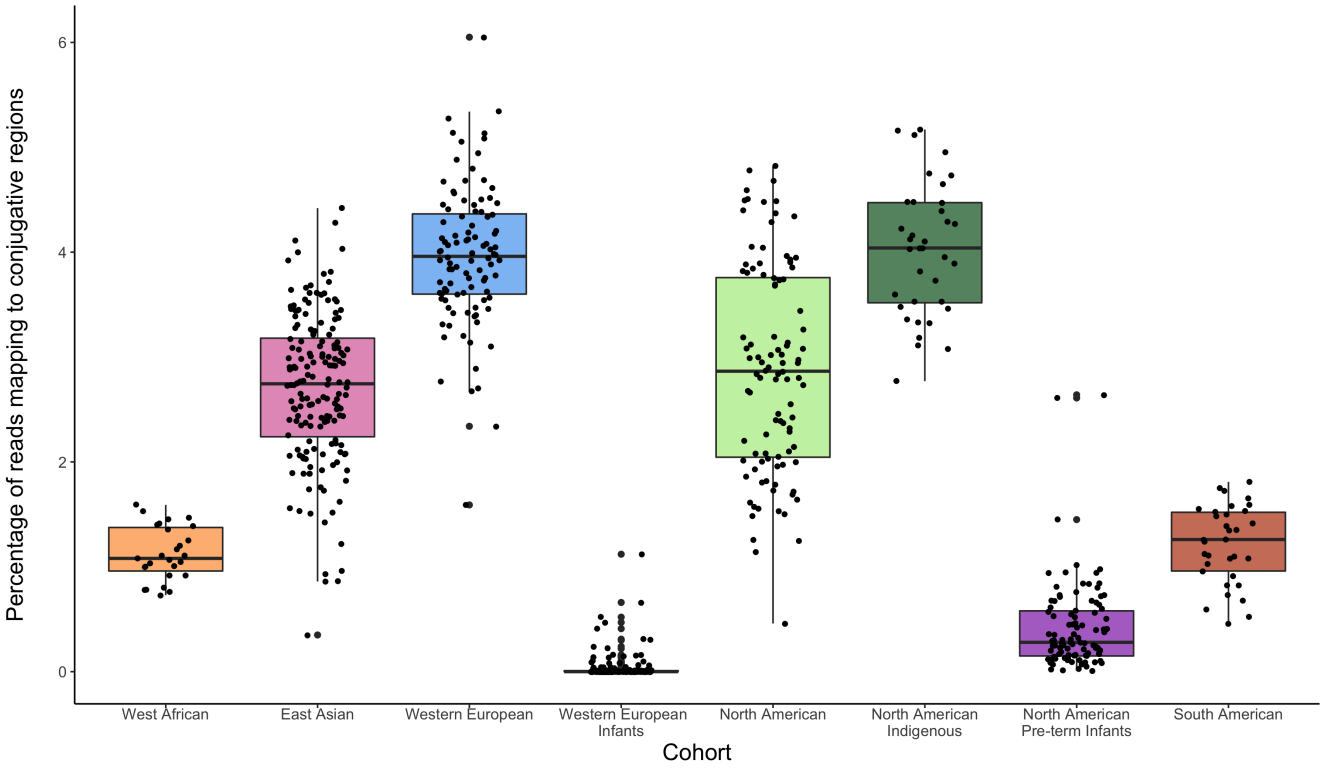

Figure 8: Plot of the percentage of total reads mapping to conjugative regions extracted from genome database, separated by cohort.

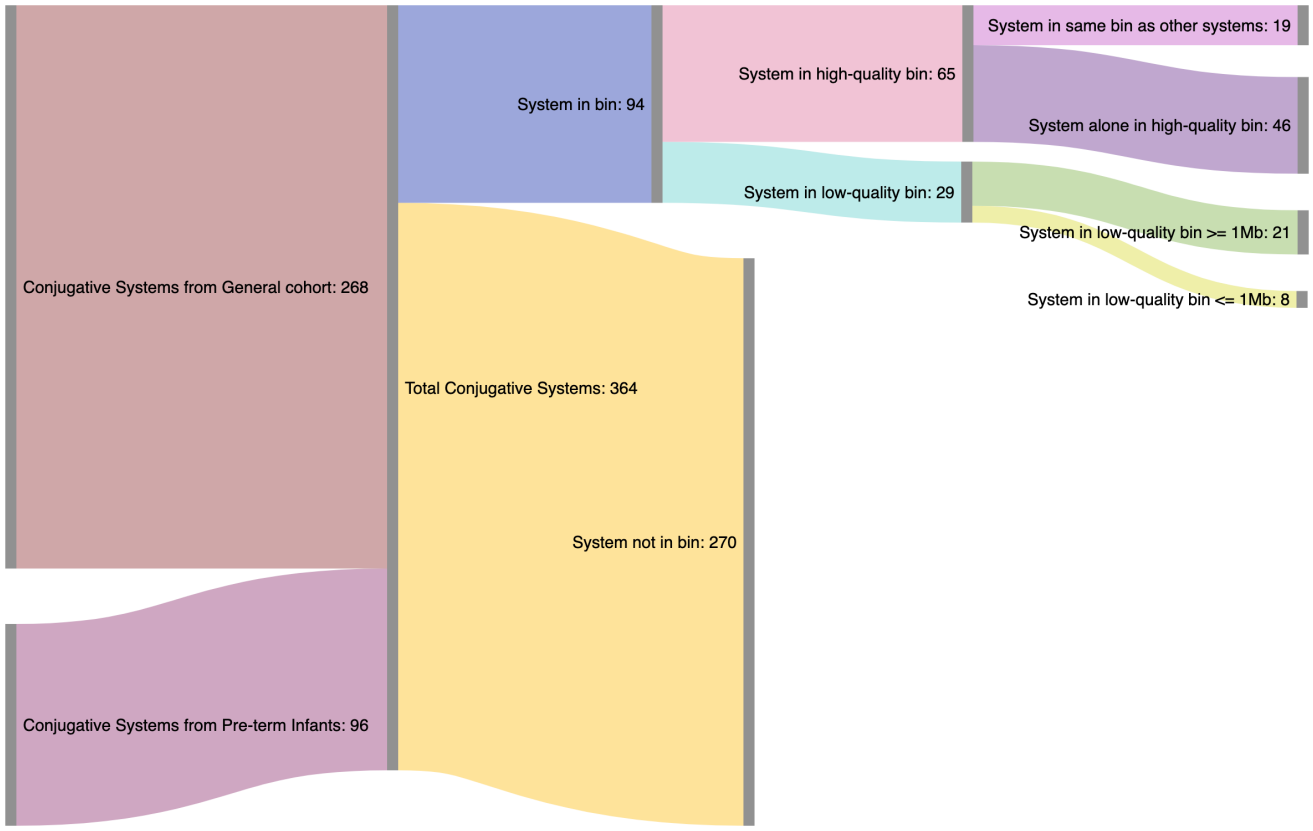

Figure 9: Sankey diagram representing the flow of 364 contigs containing conjugative systems into bins generated by MetaBAT2 from assembled data.

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/gut_conjugation

Competing interests

The authors declare that they have no competing interests.

Funding

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. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

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

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

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

12. 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.

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

14. 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.

15. 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.

16. 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.

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

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

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

20. 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.

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

22. 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.

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

24. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. Journal of Open Source Software. 2017;2:205. doi:10.21105/joss.00205.

25. 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.

26. 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.

27. 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.

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

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

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.