MICB 448A FINAL PROJECT
ERIC LEE

Full Length Research Paper

Consensus Bin Clustering for Metagenomic Time Series Datasets

Eric Lee, Ryan McLaughlin, Steven Hallam

Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada.

Submitted 13 December 2019.

The development of sophisticated algorithms and computational methods dedicated to resolving classification of metagenome-assembled genomes (MAGs) has blossomed from the increase of sequence throughput and capacity to assemble near-complete genomes from metagenomes. While current binning techniques identify core genomic functions accurately, they often miss flexible and accessory genes that are unassociated to core clusters. To address the necessity of a more complete representation of population genome diversity and function, we present a robust and reliable method to cluster MAGs into consensus bins. Our approach identifies pairwise MAG similarity through both average nucleotide identity (ANI) and min-wise independent permutations locality sensitive hashing methods (MinHash). The pairwise comparisons are then deduplicated, clouded, and joined to form consensus bins. Using this consensus binning technique, we have robustly and in a heuristically efficient way binned 1994 high quality MAGs from an anaerobic digester time-series dataset into ~200 consensus bins.

Key words: bioinformatics, cluster algorithm, metagenomics, binning, metagenomic-assembled genomes.

INTODUCTION

Metagenomics uses cultivation-independent sequencing and assembly to reconstruct the taxonomic identity and functional properties of microbial communities inhabiting natural and artificially engineered ecosystems[1]. Recently, the enhancement of high-throughput sequencing techniques has promoted the development of data driven methods for exploring and characterizing uncultured microbial taxa found across diverse environments[2]. As sequencing throughput increases, the capacity to assemble near-complete genomes from metagenomes increases as well[4]. This phenomenon has encouraged the development of sophisticated algorithmic approaches, dedicated to resolving specific classification challenges.

Binning is a commonly employed technique for extracting and assembling genomes from metagenomic datasets, resulting in the creation of metagenome-assembled genomes (MAGs)[13]. MAGs represent a collection of reads from multiple related donor genotypes that are assembled into contigs and sorted or binned into clusters of related sequences based on coverage, GC content and k-mer frequency distribution patterns[6]. In this light, MAGs can be thought of population genomes encompassing the shared coding potential of the closely related donor genotypes from a microbial community.

MAGs suffer from their own set of caveats which must be acknowledged[7]. Since they are created directly from metagenomes, MAGs are inherently from mixed sources. Closely related taxa, that have high levels of homology across their genomes or have similar abundances within the environment, may not be easily or confidently separated[14]. Microbial community structure such as diversity and evenness can also affect the ability of binning software to successfully build high quality bins[15].

While the use of MAGs has accelerated hypothesis generation related to phylogenetic relationships on the tree of life and motivated thinking about metabolic relationships within microbial communities, they are typically generated on a per sample basis or from merged assemblies across multiple independent samples[6, 17]. This can result in the formation of consensus bins enriched for core functions at the expense of more accessory or flexible genomic features[16,17,18]. Core genes are genes crucial to general biological processes, such as DNA repair, that all organisms need to survive[16]. Aside, accessory or flexible genes are groups of genes dedicated to specialized processes, such as metal-ion cycling, which are specific to a particular environment[17]. Accessory or flexible genes are not representatives across all environments, therefore they are of particular interest when studying niche ecosystem dynamics[19]. When binning co-assembled metagenomes, these genes might not be preserved since they may not be present in many samples, thus possibly resulting in important data to be lost. This is of particular concern in time-series context where metagenomic or meta-transcriptomic reads mapping back to MAGs provide a useful measure of abundance and activity, the previously described standard binning process to create MAGs can limit our capacity to identify accessory genomic features that contribute to niche adaptation and response along changing environmental and temporal gradients.

The testing dataset for this work set of MAGs was from a time-series collected from the anaerobic digester (AD) system at the Metro Vancouver Wastewater Treatment Plant located on Lulu Island British Columbia. The MAGs were created from 23 metagenomes collected over a 2 year time course, during which time the configuration parameters of the digesters and up/down stream systems were periodically modified. The goal of the overall study is to assist the wastewater treatment company with optimizing the methane production and biosolid degradation in the digestion system by assessing shifts in the microbial community during the time-series.

In order to properly assess the microbial community of the AD, MAGs needed to be created from the 23 metagenomes. However, as previously discussed, current standards for MAGs creation can lead to a loss of important niche functional information due to the practice of co-assembly of samples without regard to time of sampling. With this in mind, all AD metagenomes were binned separately to preserve the temporal aspect of the samples[3]. The goal of this project was to continue work on the time-resolved MAGs by establishing a highly robust and reliable method to cluster them into consensus bins, resulting in a more complete representation of population genome diversity and function across both a temporal gradient as well as the various system configurations.

MATERIALS AND METHODS

Quality Assessment: CheckM

For each of the 4,965 MAGs in the time series dataset, the quality was analyzed using CheckM[8] (CheckM v1.0.18 released on August 11, 2019: https://github.com/Ecogenomics/CheckM). CheckM presents robust estimates of the completeness and contamination of each MAG by using collocated sets of genes that are ubiquitous and single-copy within a single phylogenetic lineage[8]. Completeness is measured by predicting the constructed genome's lineage and determining the presence of a lineage-specific genetic marker, whereas contamination is assessed based on the presence of multiple single-copy markers within a constructed genome. High-quality MAGs having the quality of being at least 90% complete with less than 5% contamination were retained in the dataset, where this is a standard of practice used in the field to obtain reliable results[6]. On the other hand, low-quality MAGs that demonstrated low completeness and substantial contamination were discarded. A master list of all classifications of MAGs along with the quality measures are necessary for the pipeline.

Similarity Assessment: Average Nucleotide Identity

One of the many effective approaches to identify similarity between MAGs in a pairwise fashion is referred to as average nucleotide identity (ANI). ANI is an alignment free method to find the mean nucleotide identity of orthologous gene pairs shared between two microbial genomes[9]. The open source software used to measure the time-series data was FastANI[9] (FastANI v1.2: https://github.com/ParBLiSS/FastANI). All MAGs in the time-series dataset were pairwise compared with ANI computed. The ANI similarity score outputted. Following the standard of practice, only highly similar pairs with a score above 95%, corresponding to a 70% DNA-DNA hybridization for species delineation, remained in the dataset to be pipelined[10].

Similarity Assessment: Min-wise Independent Permutations Locality Sensitive Hashing

Another effective approach to measure resemblance between MAGs is using the min-wise independent permutations locality sensitive hashing scheme (MinHash). MinHash is a data mining technique to eliminate duplicates as well as large-scale clustering using similarities of words[11]. The open source tool selected is called Sourmash[11] (Sourmash v2.3.0: https://github.com/dib-lab/sourmash). The signature for a k-mer of 31 was computed for all MAGs in the time series dataset. As a set, it is then pairwise computed to generate a Jaccard similarity coefficient to indicate the ratio of intersection between sets. The cutoff for high similarity is specified as retaining only the top n values, where n is the proportion of comparisons with an ANI value of at least 0.95 within the total data frame of pairwise computed comparisons .

Data Cleaning: ANI and MinHash Pairwise Similarity Output

The total amount of pairwise computed comparisons of the similarity assessment should have a size of the total amount of MAGs to the power of two. With the heuristic of consensus binning to be challenging under datasets with potentially millions of comparisons, data cleaning is performed using R scripts. The open source code are the files: pairwise2matrix.r for ANI cleaning and matrix2pairwise for MinHash cleaning available under the repository Consensus_Binning (Consensus_Binning v1.0: https://github.com/ericlee0920/Consensus_Binning). ANI and MinHash data cleaning were to perform cutoff for high similarity as well as deduplicate repetitive pairwise comparisons. Highly alike MAGs were then grouped together to form clouds, which are pre-consensus bins.

Consensus Binning: Joining ANI and MinHash Results

Consensus binning was performed by inner and outer joins based on the open source R script file: consensus_join.r created under the repository Consensus_Binning (same version) as mentioned previously. The pre-consensus bins formed in data cleaning were then annotated with a rough classification obtained from the output of CheckM, which is based on presence of a lineage-specific genetic marker in the MAG. Pairs of MAGs that have yield a high similarity yet failed to belong in the same lineage of classification were omitted. Inner join and outer join were both performed and outputted as csv files for reference and further pipelining to other workflows.

RESULTS

The dataset used in this paper is a time series dataset of 23 replicates from two anaerobic digesters. In total, 1994 MAGs were present in the dataset after applying CheckM quality assessment. These MAGs were pipelined.

FastANI and Sourmash Performance

A total of 48,440 pairwise comparisons were outputted from FastANI as the software having a policy of returning only comparisons of threshold of 70% ANI similarity[9]. 27,612 comparisons that were above 95% ANI similarity accounted for a factor of ~0.0069 of the total pairwise computations made. The software was heuristically slower than Sourmash with a runtime of roughly 3 hours.

In contrast, Sourmash wrote all pairwise comparisons directly to file without filtering. The heuristic of using large scale k-mer clustering had a runtime of about 1.5 hours. The top factor of ~0.69% of the pairwise comparisons were extracted.

Data Cleaning and Consensus Binning Output

In the deduplication and clouding process, joins of the outputs from FastANI and Sourmash have further concentrated the dataset. Outer join created 193 groups with a total of 1937 comparisons in the resulting dataset. Inner join created 200 groups with a total of 1982 comparisons in the resulting dataset. Two separate files were outputted. The total runtime for this part were minimal to around 35 seconds.

Using the lineage classification in CheckM, disagreements with the major classification of each group were identified. In inner join, there were 56 cases and 46 cases in outer join. Comparisons with tags named, k_Bacteria, c__Deltaproteobacteria, c__Clostridia had the top three highest misplacement rates. Again, two separate files with disagreements removed were outputted as well.

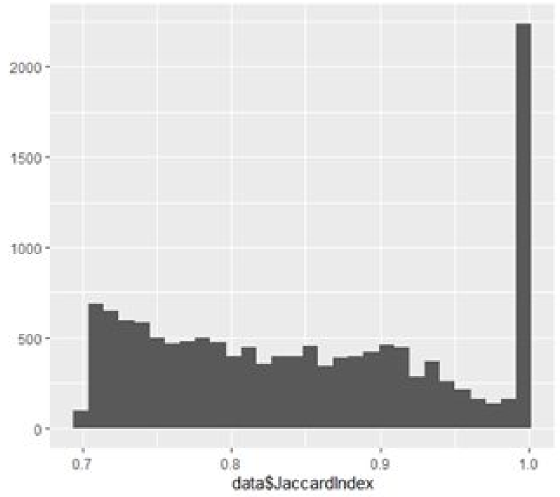
DISCUSSION

The objective of this implementation of the pipeline was to cluster AD MAGs from the Lulu Island Wastewater Treatment Plant time-series dataset into consensus bins. This software construction is foregrounded to serve as a critical segment of an extended pipeline that is meant to robustly and efficiently group MAGs, while preserving temporal or other sample-specific features, resulting in a higher resolution of flexible or accessory genomic feature sets.

Rationale for Combining ANI and MinHash

The choice here of combining results of ANI and MinHash methods is because of their highly positive correlation in output. It is now well known that Mash distances correlate well with ANI over multiple sketch and k-mer sizes, with very low disagreements in binning[12].

Prior to this implementation, the original brute-force method for clustering MAGs was using Sourmash as the sole method, as we can see in the Extended Document 1[3]. Problems arise from this method based on the non-determined Jaccard coefficient cut-off point. The Jaccard index cut-off point has been shown to vary between types of analyses in which no community agreed value has been established, thereby not robust. In order to tackle this issue, the design of this segment of the pipeline is constructed to value robust coverage and confidence of clustering over heuristic.



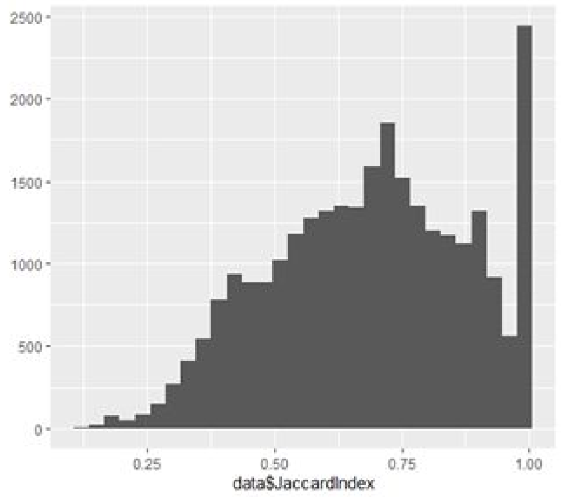


Figure 1-1. (left) Histogram Distribution of Pairwise Similarity Comparison with Arbitrary Cut-off at Jaccard Index = 0.70. X-axis represents Jaccard indexes. Y-axis represents the total count of comparisons in a specific Jaccard index. Bar at 1.0 are all the pairwise comparisons with the identical inputs. Total of 13,756 comparisons.

Figure 1-2. (right) Histogram Distribution of Pairwise Similarity Comparison with Cut-off of the Top 0.069% of all Jaccard Indexes. X-axis represents Jaccard indexes. Y-axis represents the total count of comparisons in a specific Jaccard index. Bar at 1.0 are all the pairwise comparisons with the identical inputs. 95% ANI similarity accounted for the top ~0.069% of the total pairwise computations made. Total of 27,612 comparisons.

The proposed method here is through a histogram approach. The percentage of comparisons that reached the cut-off of 95% similarity in FastANI is recorded, which is about 0.069 percent. As to benchmark, an arbitrary cut-off of Jaccard index 0.7, as seen in figure 1-1, was used. Comparing to figure 1-2 which uses a histogram approach, we can see that arbitrary cut-offs in Sourmash can lead to data missing. In fact, one of the pairwise comparisons with lowest ANI score at 95.0543% correlated to 0.156 in Jaccard index which is the lowest value in the Sourmash output. Therefore, using a random 70% set commonality subset of the dataset causes almost half of potentially useful data to be lost.

Despite high correlation of output between ANI and MinHash methods for clustering as mentioned before, we cannot suppose a high ANI value will necessarily mean that the set similarity will be high as well. Genome-wise ANI (Equation 1) outputs the proportional of the sum of the percent identities for each bidirectional best hit (BBH) against the whole length of the genes. By calculating the bidirectional best genes, it helps identify the pairs of genes in two different genomes that are more similar to each other than either is to any other gene in the other genome. Whereas, calculating the Jaccard index, we treat each MAG as a set and find the similarities between sets.

$$\text{gANI}= \sum_{\text{BBH}} (\text{Percent Identity} * \text{Alignment Length}) / \text{Lengths of BBH genes} \qquad (1)$$

$$\text{Jaccard}(X,Y)= |X \cap Y| / |X \cup Y| \qquad (2)$$

Using both methods ensures that we not only obtain the mean nucleotide identity of orthologous gene pairs shared between two genomes but also the set similarity. Based on the observation that correlation is high between MinHash and ANI outputs, we can safely use the histogram approach to obtain the comparisons with the best similarity for confident consensus binning. The drawback for using combining two methods, however, is the heuristic. Sourmash takes approximately half of the time compared to FastANI. This greatly slows down the whole consensus binning process.

Joins and Implications

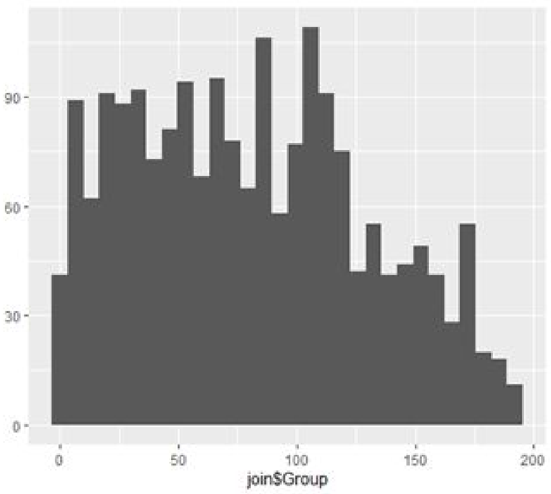
The selection of inner join and outer join depends on how the pipeline will proceed. For the sake of convenience, both joins with and without CheckM lineage disagreements removed are outputted as four separate files. It is important to mention the benefits and drawbacks of using each file.

Inner join is a join that returns all rows from participating tables where the key record of a table is equal to the key records of other table(s). It ensures that only comparisons which passes the selectivity test of both ANI and MinHash will be clouded into a group, yet this method entirely ignores potential and possible flaws from each of the two similarity calculation algorithms. The inner join method increases false negatives of the result while causing coupling problems in software. This can potentially lead to the formation of good quality consensus bins, but may accidentally remove those which belong in a consensus bin.

On the contrary, outer join return results by combining rows from two or more tables, and saving only one copy of those in common. It merges comparisons passing either ANI or MinHash selectivity tests. This method clouds all potentially similar genes into consensus bins but may result in groups having unexpected false positive comparisons. The outer method potentially increases false positives of the results while decoupling problems in software. The resulting output can have large consensus bins with a need of sorting out false positives.

To address the need of sorting, a feature of removing lineage disagreements using the lineage classification in CheckM was implemented. We see that we have a disagreement rate of approximately 2.8% in inner join and 2.3% in outer join. While not examined in this project, the disagreements in these joins need manual consideration of whether or not it belongs in a consensus bin. A challenge is that we may not have high confidence in using lineage removal since the CheckM lineage classification only provides up to 19 tags with very general classification such as k_Bacteria.

It can be observed from figure 2-1 and 2-2, that both joins yield approximately the same shape. The group number has increased, and some groups have increased comparison count.



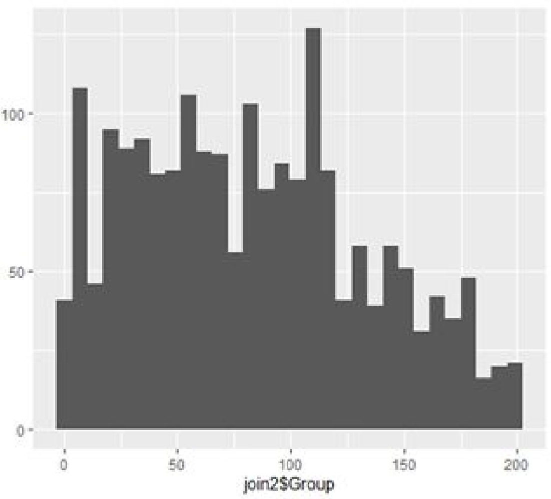


Figure 2-1. (left) Histogram Distribution of an Inner Join of Outputs from FastANI and Sourmash. X-axis represents group numbers from 1 to 193. Y-axis represents the total count of comparisons in a specific group. Total of 1937 comparisons.

Figure 2-2. (right) Histogram Distribution of an Outer Join of Outputs from FastANI and Sourmash. X-axis represents group numbers from 1 to 200. Y-axis represents the total count of comparisons in a specific group. Total of 1982 comparisons.

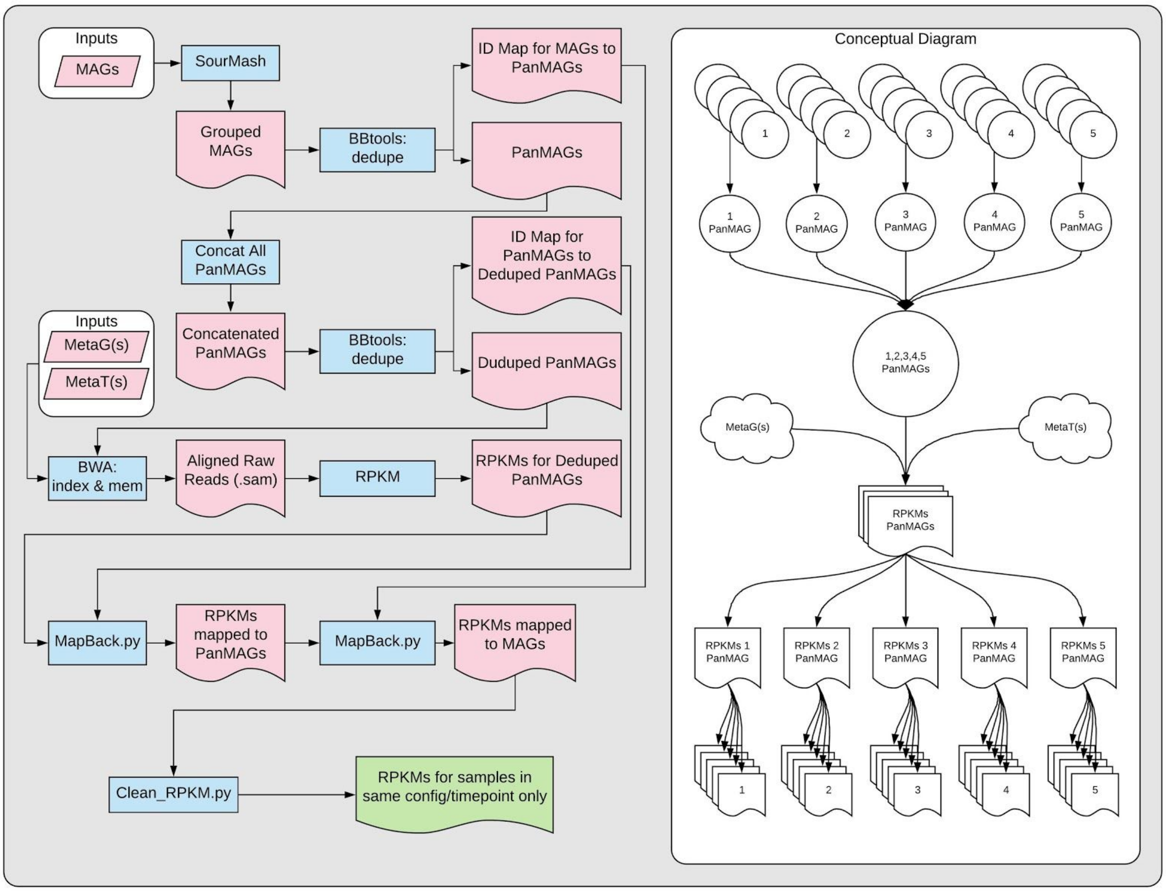
Concluding, through the combination of ANI and MinHash methods, this project provides an implementation of a highly robust and reliable method to cluster MAGs overlapping in a time series dataset into consensus bins, resulting in a more complete representation of population genome diversity and function. This method will be implemented as a component within a larger framework which serves to preserve temporal patterns within MAGs datasets to allow for better resolution of the functional variation between populations through time.

Going forward, it would be interesting to develop a modified implementation of the current FastANI and Sourmash implementation to decrease the runtime as the throughput-limiting step is FastANI. Another meaningful work would be to provide a better lineage classification software to do lineage removal in the joining step. This will increase our confidence in consensus binning.

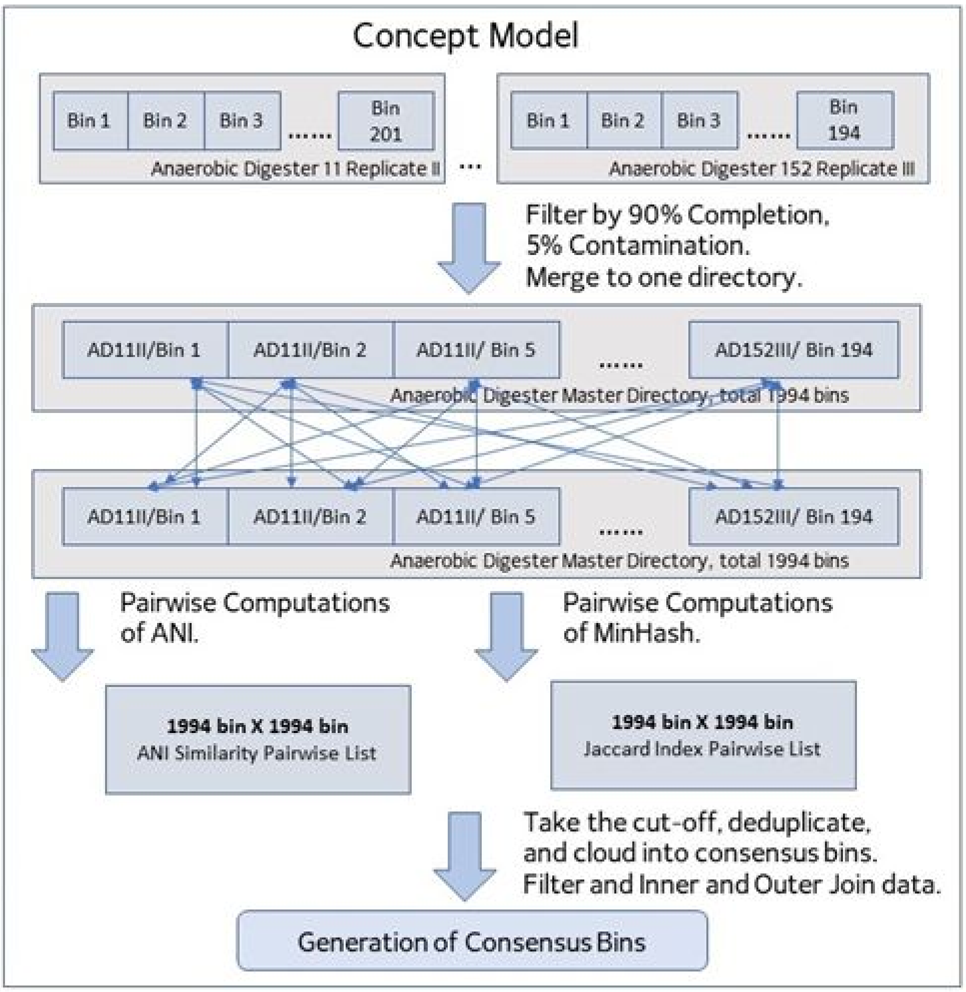
CONFLICTS OF INTEREST

There are no conflicts over this paper.

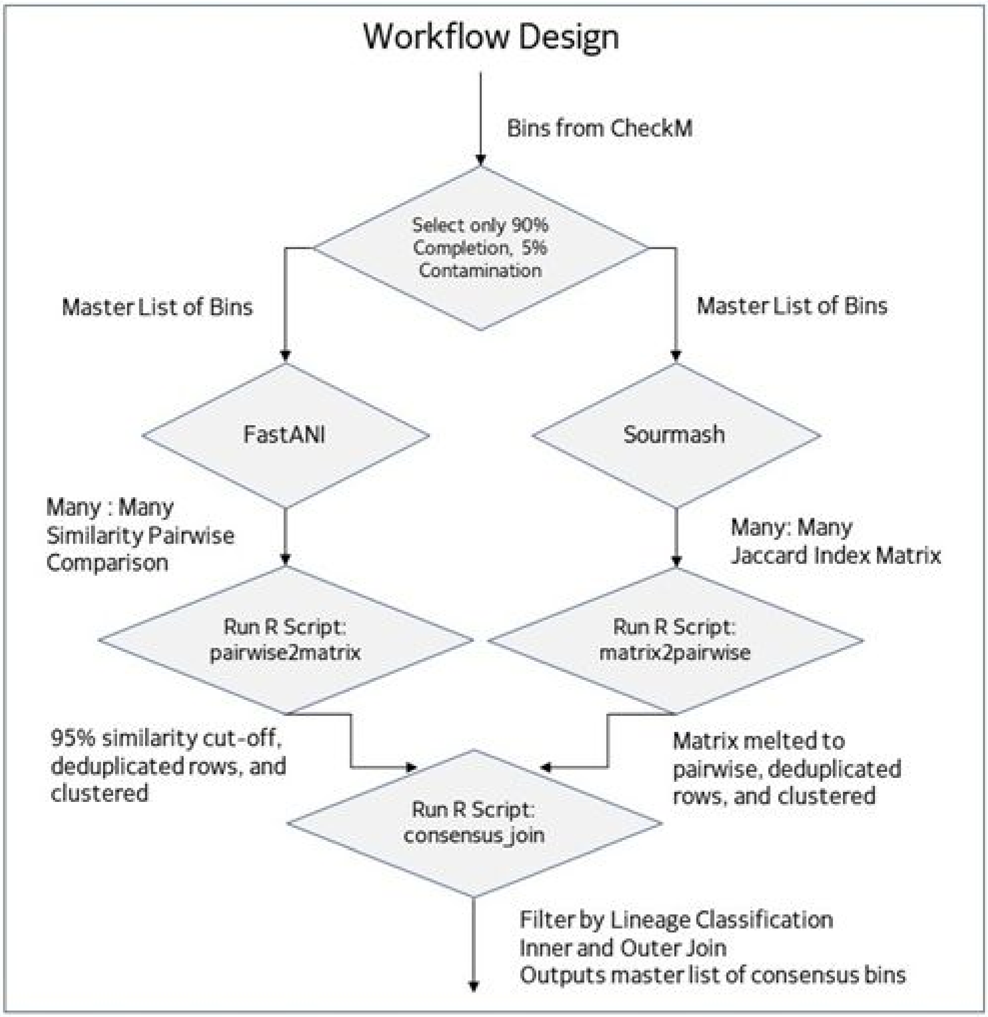
APPENDIX



Extended Document 1. Workflow Diagram.[3] This workflow diagram on the left outlines the processing steps for creating sample-resolved population genomes. The right image shows the conceptual diagram to aid in understanding the problem space.


Concept Model
Bin 1
Bin 2
Bin 3
......
Bin 201
Anaerobic Digester 11 Replicate II
...
Bin 1
Bin 2
Bin 3
......
Bin 194
Anaerobic Digester 152 Replicate III
Filter by 90% Completion,
5% Contamination.
Merge to one directory.
AD11II/Bin 1
AD11II/Bin 2
AD11II/ Bin 5
......
AD152III/ Bin 194
Anaerobic Digester Master Directory, total 1994 bins
AD11II/Bin 1
AD11II/Bin 2
AD11II/ Bin 5
......
AD152III/ Bin 194
Anaerobic Digester Master Directory, total 1994 bins
Pairwise Computations of ANI.
Pairwise Computations of MinHash.
1994 bin X 1994 bin
ANI Similarity Pairwise List
1994 bin X 1994 bin
Jaccard Index Pairwise List
Take the cut-off, deduplicate, and cloud into consensus bins. Filter and Inner and Outer Join data.
Generation of Consensus Bins

Extended Document 2. Concept Model. This concept model echoes the theoretical processing ideas generated explained in the previous methods section. Arrows represent the pipelining of data. Description with the arrow represents the actions that should be performed.


Workflow Design
Bins from CheckM
Select only 90% Completion, 5% Contamination
Master List of Bins
Master List of Bins
FastANI
Sourmash
Many : Many Similarity Pairwise Comparison
Many: Many Jaccard Index Matrix
Run R Script: pairwise2matrix
Run R Script: matrix2pairwise
95% similarity cut-off, deduplicated rows, and clustered
Matrix melted to pairwise, deduplicated rows, and clustered
Run R Script: consensus_join
Filter by Lineage Classification
Inner and Outer Join
Outputs master list of consensus bins

Extended Document 3. Workflow Design. This workflow design further draws the computational processing explained in a UML fashion. Each trapezoid represented a program. Arrows represent the pipelining of data from source to destination. Description with the arrow represents the actions performed at the source.

REFERENCES

1. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The Next Culture-Independent Game Changer. *Front Microbiol*. 2017 Jul 4;8:1069. doi: 10.3389/fmicb.2017.01069. PMID: 28725217; PMCID: PMC5495826.
2. Diaz PI, Dupuy AK, Abusleme L, Reese B, Obergfell C, Choquette L, Dongari-Bagtzoglou A, Peterson DE, Terzi E, Strausbaugh LD. Using high throughput sequencing to explore the biodiversity in oral bacterial communities. *Mol Oral Microbiol*. 2012 Jun;27(3):182-201. doi: 10.1111/j.2041-1014.2012.00642.x. Epub 2012 Mar 3. PMID: 22520388; PMCID: PMC3789374.
3. McLaughlin R. Workflow Diagram. 2019. Image.
4. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*. 2015 Mar;15(2):141-61. doi: 10.1007/s10142-015-0433-4. Epub 2015 Feb 27. PMID: 25722247; PMCID: PMC4361730.
5. Spits, C., Le Caignec, C., De Rycke, M. *et al.* Whole-genome multiple displacement amplification from single cells. *Nat Protoc* 1, 1965–1970 (2006) doi:10.1038/nprot.2006.326
6. Parks, D.H., Rinke, C., Chuvochina, M. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2017;2, 1533–1542. doi:10.1038/s41564-017-0012-7
7. Alneberg, J., Karlsson, C.M.G., Divne, A. *et al.* Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* 6, 173 (2018) doi:10.1186/s40168-018-0550-0
8. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2014. Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25: 1043-1055.
9. Jain, C., Rodriguez-R, L.M., Phillippy, A.M. et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9, 5114 doi:10.1038/s41467-018-07641-9
10. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol.* 2007 Jan;57(Pt 1):81-91. doi: 10.1099/ijs.0.64483-0. PubMed PMID: 17220447.
11. Pierce NT, Irber L, Reiter T *et al.* Large-scale sequence comparisons with sourmash [version 1; peer review: 2 approved]. *F1000Research* 2019;8:1006 (https://doi.org/10.12688/f1000research.19675.1)
12. Ondov, B.D., Treangen, T.J., Melsted, P. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 17, 132 (2016) doi:10.1186/s13059-016-0997-x
13. Fernando Meyer, Peter Hofmann, Peter Belmann, Ruben Garrido-Oter, Adrian Fritz, Alexander Sczyrba, Alice C McHardy, AMBER: Assessment of Metagenome BinnERs, *GigaScience*, Volume 7, Issue 6, June 2018, giy069, https://doi.org/10.1093/gigascience/giy069

14. Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications. *Brief Bioinform*. 2012;13:711–727.
15. Bozzuto C, Blanckenhorn W. Taxonomic resolution and treatment effects – alone and combined – can mask significant biodiversity reductions. *International Journal of Biodiversity Science, Ecosystem Services & Management* 2016: 86-99. https://doi.org/10.1080/21513732.2016.1260638.
16. Lan R, Reeves PR. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol*. 2000;8(9):396–401.
17. van Passel MW, Marri PR, Ochman H. The emergence and fate of horizontally acquired genes in Escherichia *coli*. *PLoS Comput Biol*. 2008;4(4):e1000059.
18. Hall James P. J., Brockhurst Michael A. and Harrison Ellie. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Phil Trans. R. Soc. B.* 2017:372 https://doi.org/10.1098/rstb.2016.0424
19. Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microb Genom*. 2015 Nov 5;1(5):e000038. doi: 10.1099/mgen.0.000038. PMID: 28348822; PMCID: PMC5320679.