A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination

Mattock, J.; Watson, M.

Published in:
Nature Methods

DOI:
10.1038/s41592-023-01934-8

Print publication: 01/08/2023

Document Version
Peer reviewed version

Supplementary Information

Replication of results in human microbiome data

To ensure that our results are not limited to a single dataset, we replicate our findings in a human microbiome dataset which was included in a dataset of over 150,000 MAGs published by Pasolli *et* al and show that these genomes contain larger amounts of contamination than the published statistics. Rampelli *et al* studied the Hadza Hunter-Gatherer Gut Microbiome[1], and this study constitutes one of 46 used by Pasolli *et al* in their meta-assembly of microbial genomes from all human body sites[2]. The Rampelli *et al* dataset consists of 38 samples, and was assembled into genome bins by Pasolli *et al* using single-coverage binning, and filtered using CheckM cut-offs of completeness >50%, contamination <5%. The Rampelli *et al* datasets contributed 479 of Pasolli *et al*'s total number of reconstructed genomes, and 48 of the representative species-level genome bins.

Firstly, the Rampelli *et al* dataset was subject to the same pipeline as the rumen dataset, and we discovered similar results. Using CheckM cut-offs of completeness >50%, contamination <5%, single-coverage binning produced 471 genome bins, yet multi-coverage binning produced 641 genome bins, an increase of 36%.

Secondly, we used the multi-coverage data generated for the Rampelli *et al* dataset to examine the quality of the bins published by Pasolli *et al*. Of the 479 genome bins, we estimate that 259 have more than 5% contamination, and 131 have more than 10% contamination, by number of contigs. When looking at contamination by length, we estimate that 173 of the 479 genomes have greater than 5% contamination, and 74 have greater than 10% contamination. The CheckM cut-off for all genomes produced by Pasolli *et al* is 5%.

Figure S1 shows a genome bin, "RampelliS_2015__H17__bin_7", which is part of Pasolli *et al*'s representative species-level genome bin set, and which we show to be chimeric. This bin clearly contains two sets of contigs when assessed using multi-coverage data – one set, towards the bottom of the heatmap, dominated by coverage in a small number of samples; and a second set, towards the top of the heatmap, which show varying levels of coverage in most samples. We estimate that 36% of the contigs in this genome bin represent contamination, yet CheckM estimates the bin to be 92.6% complete and only 3.13% contaminated.

Rationale for using Pearson Correlation Coefficient

In statistical terms, metagenomic bins can be treated as clusters of points (contigs) in multi-dimensional space (coverage). The dispersion of points within the cluster can be examined by calculating pairwise similarities between each pair of points in the cluster. Points which have high levels of similarity with other points in the cluster are likely to belong in the cluster, and the inverse is true - points with high levels of dissimilarity with other points in the cluster are unlikely to belong in the cluster. Therefore, the overall cohesion of any cluster can be observed by examining the distribution of pairwise similarity measures. The cohesion of a metagenomic bin can be defined as how similar the coverage profiles are of contigs within those bins; and this can be calculated using a similarity measure such as the Pearson correlation coefficient, *r*, for all pairs of contigs within a bin.

Challenges in implementation of our approach

A challenge in our approach is the selection of cut-offs for correlation coefficients that signify contamination. In this paper we identify contaminant contigs as those whose Pearson correlation coefficient is < 0.9 with over 90% of the other contigs in the bin. These are relatively strict thresholds chosen to identify the most obvious contaminant contigs, and our estimates of contamination using these cut-offs are likely to be under-estimates. Researchers who wish to use multi-coverage data to

identify contamination may wish to change these thresholds after examining their own data, and after deciding how conservative they wish to be in identifying contaminant contigs.

Use of multi-coverage data to bin genomes or to identify contaminant contigs may risk discarding mobile genetic elements or members of the accessory genome, both of which by definition may be present in some genomes, but not others, of the same strain. However, it is already known that metagenomic binning techniques produce bins that are depleted in mobile genetic elements (MGEs)[3], and other core tools in the metagenomic binning process (such as CheckM) rely on single-copy core genes, which by definition exist on the core genome. It is therefore expected that bins created or QC-ed using multi-coverage data will represent only the core genome of strains or species. MGEs not only differ in coverage, but often also in sequence content, and therefore use of metagenomic binning is likely to preclude accurate binning of MGEs unless an additional technology is used, such as Hi-C[4].

Despite the clear benefits in using multi-coverage binning there remains one major limitation, the computational burden of performing this with many samples. For example, in an experiment of 400 microbiome samples, we would have 400 assembly jobs, which typically consume 8 cores each with 8Gb of RAM for approximately 12 hours - a total of 38,400 core-hours; for single-coverage binning, we would then have only 400 mapping jobs, which typically consume 4 cores each with 8Gb of RAM for 8 hours - a total of 12,800 core-hours. However, multi-coverage binning of 400 samples would require an incredible 160,000 mapping jobs - a total of 5,120,000 core-hours, a number which is out of reach for most researchers. One approach to limit the number of jobs could be to perform multi-coverage binning on subgroups of the samples, however it's likely that the optimal number of samples would vary per experiment, making identifying an ideal number of samples challenging.

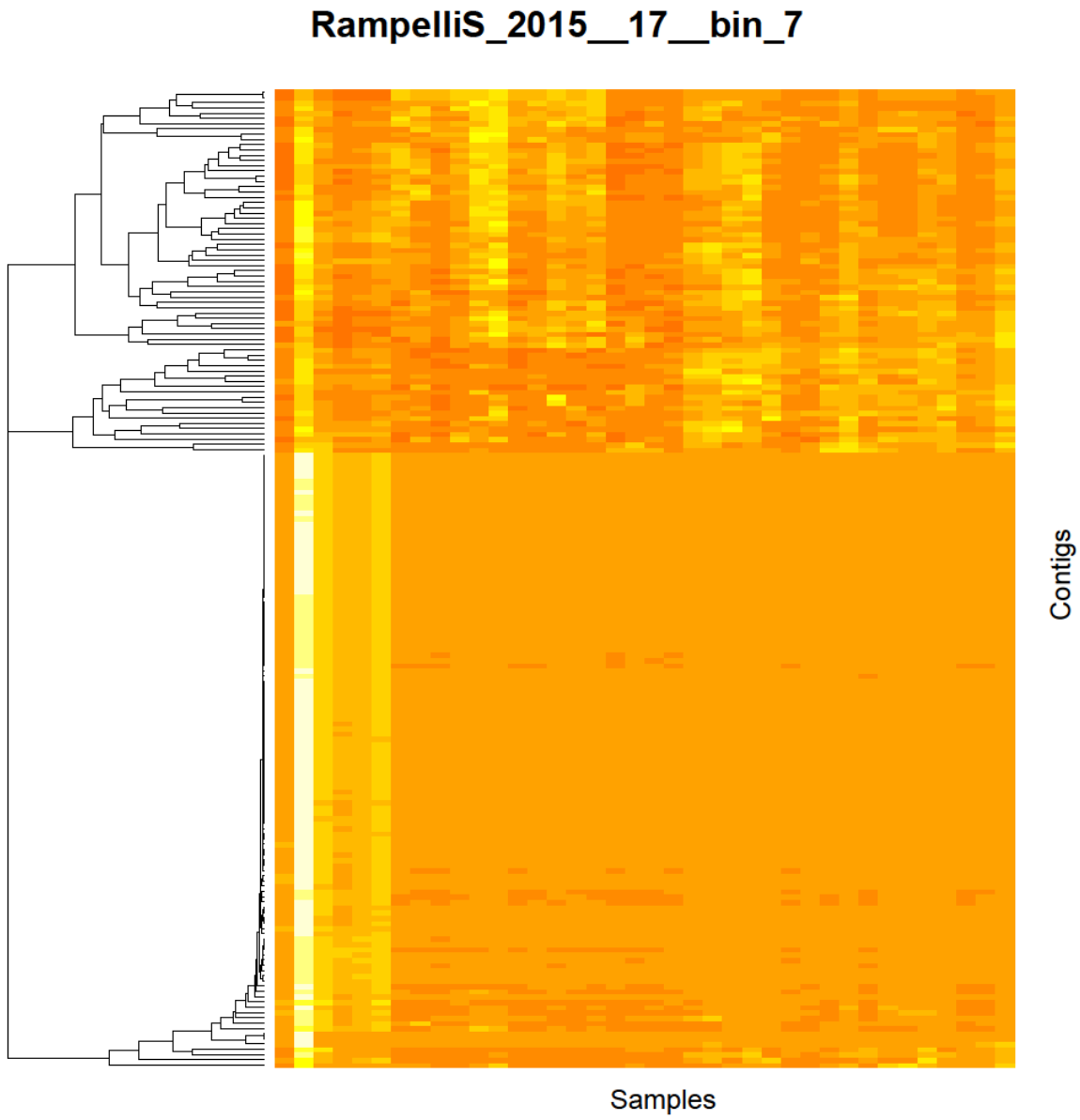

Figure S1 A heatmap of RampelliS_2015__H17__bin.7, a chimeric bin published by Pasolli *et* al. This bin clearly contains two sets of contigs when assessed using multi-coverage data – one set, towards the bottom of the heatmap, dominated by coverage in a small number of samples; and a second set, towards the top of the heatmap, which show varying levels of coverage amongst most samples

References

1. Rampelli, S. *et al.* Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).
2. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).
3. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
4. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 1–11 (2018).