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

Editor summary:

This study shows when analyzing multi-sample metagenomic datasets, the multi-coverage binning approach outperforms the single-coverage binning alternative in generating bins with higher quality and less contamination.

Editor recognition statement:

Peer review information: Primary Handling editor: Lin Tang, in collaboration with the Nature Methods team.

Reviewer Recognition:

Nature Methods thanks Anders Andersson, C. Titus Brown and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Item	Present?	Filename Whole original file name including extension. i.e.: Smith_SI.pdf. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_Information_v3.pdf	Replication of results in human microbiome data, Rationale for using Pearson Correlation Coefficient, Challenges in implementation of our approach, Supplementary Figure 1
Reporting Summary	Yes	57714_2_attach_10_22535_v2.pdf	
Peer Review Information	Yes	*Peer Review file.pdf*	

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

Jennifer Mattock[1] and Mick Watson[2,3*]

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, Midlothian, UK.

[2] Centre for Digital Innovation, DSM Biotechnology Center, Delft 2613 AX, The Netherlands

[3] Scotland's Rural College, Peter Wilson Building, King's Buildings, Edinburgh, EH9 3JG, Scotland

* Corresponding author mick.watson@dsm.com

Abstract

Metagenomic binning has revolutionised the study of uncultured microorganisms. This study compares single- and multi-coverage binning on the same set of samples, and demonstrates that multi-coverage binning produces better results than single-coverage binning, and identifies contaminant contigsand chimeric bins that other approaches miss. Whilst resource expensive, multi-coverage binning is a superior approach and should always be performed over single-coverage binning.

Main

Metagenomic binning, the resolution of metagenomic sequence data into individual genomes, has been used to identify hundreds of thousands of genomes from microbiome samples[1–6]. These studies are enabled by software that groups together assembled contigs based on the assumption that contigs with similar sequence content and coverage profiles across multiple samples likely originate from the same genome[7,8]. However, calculating coverage from multiple samples represents a problem for large sample sizes, requiring an all-against-all comparison. It has therefore become routine for single-coverage binning to be performed for large datasets. Previous research has described multi-coverage binning in the context of co-assembly, finding that at least five samples are required for it to be worthwhile[3]; increasing the number of samples when performing multi-coverage binning decreased the contamination and increased the completeness of bins[7,9]. However co-assembly is sub-optimal as it allows the reconstruction of only one bin per species[3].

The aim of this paper is to compare single- and multi-coverage binning on the same dataset, to quantify the effect of the loss of coverage information on the quantity and quality of bins produced. We hypothesize that single-coverage binning will frequently bin together contigs that are co-abundant only in a single sample (Fig 1A), that these errors represent invisible contamination, and that they can be detected by using multi-coverage data.

Forty-two rumen microbiome samples were assembled and binned using two strategies, single-coverage and multi-coverage binning. All other parameters remained the same. The completeness and contamination results for all bins produced by both methods are shown in Figure 1B. Minimal difference is observed between the distribution of completeness scores in the single and multi-coverage bins, however, the single coverage bins have increased contamination: 22.5% (1273/5658) of the single coverage bins have a contamination score of 5 or greater versus 3.5% (293/8420) of the multi-coverage bins. This suggests that more contigs classed as contaminant DNA are incorporated using the single coverage approach.

The single coverage approach produced a total of 5658 bins across the 42 samples, whereas the multi-coverage approach produced 8420 (Figure 1C). A filtered set of bins was produced using completeness and contamination cut-offs that have previously been used in ruminants[6,10–12] (completeness ≥ 80% and contamination ≤ 10%). Using these cut-offs, the single coverage approach produced 931 filtered bins, compared to 1660 produced by the multi-coverage approach, an increase of 78%. This suggests that

the multi-coverage approach results in more bins of higher quality. The filtered bins were used for all downstream analysis.

The taxonomies produced by either binning method were compared. Variation was observed in the proportion of bins belonging to each taxa at each rank. A greater proportion of the multi-coverage bins were archaea (4.3%) than in the single coverage bins (3.1%). In both approaches the predominant phyla was *Bacteroidota* with a slight variation in the *Firmicutes/Bacteroidota* ratio, 1.28 in multi-coverage bins vs 1.05 in multi-coverage bins. One Phylum, *Patescibacteria*; two Classes, *Endomicrobia* and *Saccharimonadia*, three Orders, nine Families, 35 Genera and 96 Species were found exclusively in the multi-coverage bins. Just two Genera and 11 Species were found exclusively in the single coverage bins. This suggests that single coverage binning may overlook taxa that can be recovered using multi-coverage binning, perhaps due to the increased coverage data available with multi-coverage binning enabling the splitting of contigs by coverage at a greater resolution.

Dereplication of the bins was performed at the species and strain level to determine the overlap between single- and multi- coverage bins. In the single coverage bins, 460 species and 573 strains were identified; this increased in the multi-coverage bins to 682 species and 943 strains. When all bins were dereplicated together 700 species were found, 240 of which were unique to multi-coverage bins and 18 to single coverage bins. At the strain level 969 strains were present, 398 only found using the multi-coverage method and 23 the single coverage method. This illustrates how including coverage information from multiple samples can help recover species and strains that would otherwise be missed.

We used the distribution of observed values of r to assess the quality differences between bins and to detect contaminant contigs (see rationale in Supplementary Information). The mean pairwise correlation coefficient for each bin was significantly higher in multi-coverage bins (p-value < $2x10^{-16}$); 89% (1480/1660) of multi-coverage bins had a mean pairwise correlation coefficient greater than 0.9 compared to 44% (406/931) of single-coverage bins (Figure 1D). Furthermore, the distribution is clearly skewed towards 1 for multi-coverage bins, whereas the distribution for single-coverage bins was flatter with a tail stretching down into lower values of r. This suggests that the single-coverage bins are more dispersed and contain many more pairs of contigs that are dissimilar to one another than the multi-coverage bins: contigs with low levels of similarity with the rest of the bin are likely to be contaminants.

Examining the minimum value for r within each bin allows us to identify the pair of contigs with the least similar coverage profiles. The minimum value of r observed in each bin was significantly lower in the single coverage bins (p-value < $2x10^{-16}$); 73% (684/931) of single coverage bins contained at least one pair of contigs which were negatively correlated, versus only 10% (157/1660) of multi-coverage bins (Figure 1E). This is consistent with the hypothesis that single-coverage bins contain higher numbers of contaminant contigs. The single coverage bins had significantly (p-value < $2x10^{-16}$) more bins with higher proportions of coverage coefficients of less than 0.5 than the multi-coverage bins (Figure 1F). Just 8.3% (77/931) single coverage bins had no coverage coefficients less than 0.5 compared to 39.3% (653/1660) of multi-coverage bins.

To identify the most contaminated bins, all filtered bins were ranked by their mean pairwise r. Ninety-eight of the hundred lowest ranked bins were produced using single-coverage binning. The lowest ranked bin, single_ERR2027909.44, has a mean pairwise r of just 0.25. The coverage profiles of the contigs can be seen in Figure 2A. Contigs were predicted as contamination if they had an $r \leq 0.9$ with more than 90% of the contigs in the bin. Using this approach, 303 of the 529 contigs (57%) in single_ERR2027909.44 represent contamination, the equivalent of 949kb of contaminant sequence from a total of 3.12Mb (30%). The non-contaminant contigs are all highly correlated with one another and can be seen in the top half of the heatmap in Figure 2A, with their coverage profiles plotted in Figure 2B. In contrast, the contigs predicted to be contaminants show dissimilar coverage profiles, with no discernible pattern in the multi-coverage data (Fig 2A and 2C). Therefore, multi-coverage data for this single-coverage bin suggests that this is a highly contaminated bin with hundreds of contigs that do not belong together. However, CheckM estimates this bin to be 93.04% complete, with just 8.06% contamination[13]. A taxonomic method for identifying chimeric bins, GUNC[14], does not detect any

chimerism in this bin, estimating a clade separation score of just 0.16. It is therefore clear that multi-coverage data can identify potential contamination and chimerism that current methods miss.

Using the above cut-off to detect contaminant contigs in all filtered bins, we predict that 428 of the 931 single coverage bins (46%) contain more than 10% contamination in terms of the number of contigs, and 151 (16%) contain more than 10% contamination by sequence length. The worst single-coverage bin by contig (single_ERR2027912.135) contains 58.4% contamination, and the worst single-coverage bin by sequence length (single_ERR2027901.59) contains 43.6% contamination. In contrast, 177 out of 1660 multi-coverage bins (10.6%) contain more than 10% contamination by contig, and only 39 out of 1660 (2%) contain more than 10% contamination by sequence length. However, despite the superior performance of multi-coverage binning, this technique also produces some contaminated bins - the worst multi-coverage bin (multi_ERR2027898.229) contains 75% contamination by contig and 44% contamination by sequence length.

To ensure that our results were not limited to a single dataset, we replicate our findings in a human microbiome dataset (Supplementary Information) and show that these genomes contain larger amounts of contamination than the published statistics[15]. By measuring the pairwise Pearson correlation coefficient between each pair of contigs in each bin, using multi-coverage data, the results demonstrate that single-coverage bins contain large amounts of hidden contamination that are not detected by existing techniques, and that multi-coverage binning performs much better (though not perfectly) when assessed using this method.

There are challenges in the implementation of our approach, such as the computational burden, selection of appropriate cutoffs and the potential loss of mobile genetic elements – these are discussed further in the Supplementary Information. However, our results demonstrate that, wherever possible, multi-coverage data should be used for metagenomic binning; and in all cases, significant effort must be devoted to quality control and filtering of metagenomic bins that go beyond existing methods such as CheckM and GUNC, as both single-copy-core-gene and taxonomic methods miss hidden contamination that statistical methods do not.

Acknowledgements

The Roslin Institute forms part of, and is supported by, the Royal (Dick) School of Veterinary Studies, University of Edinburgh. This project was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; BB/S006680/1, BB/R015023/1, BB/V018450/1), including institute strategic program grant BBS/E/D/30002276.

Contributions

JM and MW carried out all analyses and wrote the paper.

Competing interests

Mick Watson is an employee of DSM, and the remaining authors declare no competing interests

Figure Legends

Figure 1 a comparison of single- and multi- coverage metagenomic binning. **A** hypothetical example using simulated data demonstrating our hypothesis that contigs from two different genomes may only be co-abundant in a single-sample and therefore may be mistakenly binned together in single-coverage binning **B** a comparison of completeness and contamination statistics for single-coverage bins (top row)

and multi-coverage bins (bottom row). **C** The number of bins produced by single and multi-coverage binning. **D** violin plot of the mean pairwise inter-contig correlations for single (n=931) and multi-coverage bins (n=1660). The boxplot centre represents the median, the box the 25th and 75th percentiles and the whiskers 1.5x the interquartile range. **E** violin plot of the minimum pairwise inter-contig correlations for single (n=931) and multi-coverage bins (n=1660). The boxplot centre represents the median, the box the 25th and 75th percentiles and the whiskers 1.5x the interquartile range. **F** scatter plot of the percentage of pairwise inter-contig correlations below r=0.5 for single and multi-coverage bins.

Figure 2 The worst performing single-coverage bin according to the mean pairwise correlation coefficient. **A** heatmap of the worst single coverage bin, showing the coverage of every contig (rows) in each sample (columns). Orange represents low coverage and yellow high coverage. Contaminant contigs, identified as having an $r \leq 0.9$ with more than 90% of the contigs in the bin, are indicated to the right of the heatmap. **B** Line plot showing the coverage profile of the core (non-contaminant) contigs. **C** Line plot showing the coverage profile of the contaminant contigs

References

1. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
2. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
3. 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).
4. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
5. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
6. 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).
7. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
8. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
9. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
10. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 1–11 (2018).
11. Glendinning, L., Genç, B., Wallace, R. J. & Watson, M. Metagenomic analysis of the cow, sheep, reindeer and red deer rumen. *Sci. Rep.* **11**, 1990 (2021).
12. Wilkinson, T. *et al.* 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. *Genome Biol.* **21**, 229 (2020).
13. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
14. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
15. Rampelli, S. *et al.* Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).

Methods

Binning methods

The Illumina sequence data of 42 rumen microbiome samples from a previous study were downloaded from the European Nucleotide Archive repository, project accession PRJEB21624 [10]. Trim Galore (0.6.2) was used to trim the Illumina adaptors and poor quality base-calls; the trimmed reads were then single-sample assembled with MEGAHIT (v1.1.3) using options --k-list 27,47,67,87, --kmin-1pass, -m 0.95, --min-contig-len 1000 and -t 8 [16,17]. The reads were mapped against their own assembly and also against each of the other assemblies with BWA MEM (v0.7.17) [18]. Samtools (v1.9) was used to convert the output to BAM files and the jgi_summariza_bam_contig_depths script from Metabat2 (v2.15) to calculate the coverage for each of the bam files [8,19]. These coverage files were used as input for Metabat2 to perform single coverage binning with the option --minContig 1500. The coverage files for each sample were combined and processed with Metabat2 and --minContig 1500 to produce multi-coverage bins.

Bin quality assessment

CheckM (v.1.0.7), with the options lineage_wf, -x fa and --tab_table, was used to calculate the completeness and contamination of all bins. Bins with contamination ≤ 10% and completeness ≥ 80% were kept for downstream analysis. GUNC (v1.0.4) was run with options --contig_taxonomy_output --detailed_output --db_file gunc_db_progenomes2.1.dmnd. Complete rRNAs were searched for using Barrnap (v0.9) with the arc, bac, euk and mito options (https://github.com/tseemann/barrnap).

Single and multi-coverage bin comparisons

The bins were assigned taxonomies using the GTDB-Tk (v1.4.0) classify_wf option [20]. Phylophlan (v3.0.60), using the phylophlan database (downloaded automatically by the tool) and options --diversity high and --min_num_markers 40, was then used to create a phylogeny with 460 rumen microbiota genomes from the Hungate 1000 project [21,22]. The resulting phylogeny was rooted at the Bacteria/Archaea branch with FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and visualised with iTOL [23]. Annotations for the bins were improved using the phylogeny, for example if a bin was surrounded by Hungate genomes of a different family then the bin's family was corrected to match theirs.

The filtered single and multi-coverage bins, both together and separately, were dereplicated at the species (95%) and strain level (99%) using dRep (v3.2.0) with the options -p 4, -comp 80, -con 10, -nc 0.6 and -sa 0.95 or -sa 0.99 [24].

Bin cohesion

To explore how cohesive the single coverage bins were relative to the multi-coverage bins, the coverage for each of the contigs in the single coverage bins was pulled from the multi-coverage files. The correlation between all of the contigs coverage within each bin was calculated, with comparisons against self and duplicate pairwise comparisons removed, using Pearson's correlation within R [25]. Plots were drawn with the ggplot2 package (v3.3.5), ggsci (v2.9), ggplotify (v.0.1.0), cowplot (v.1.1.1), patchwork (v.1.1.1), gridGraphics (v.0.5-1) and dplyr (v1.0.7) [26–32]. Significance calculations were performed with significance defined as $p < 0.05$ using the Mann-Whitney U test and R (v4.1.0).

Code availability

Code for producing single- and multi- coverage assemblies and bins is available at: https://github.com/WatsonLab/single_and_multiple_binning

Data availability

Raw rumen FASTQ datasets are available under BioProject accession PRJEB21624. Raw human FASTQ datasets are available under BioProject accession PRJNA278393. Bins from Rampelli *et al*, assembled by Pasolli *et al*, are available from http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html. Metagenome assemblies of the Rampelli *et al* data, assembled by Pasolli *et al*, are available from https://www.dropbox.com/s/5qqtbyuufmgycp6/RampelliS_2015.tar.bz2. Finally, our analysis of the rumen and human datasets and bins can be downloaded from DOI: 10.6084/m9.figshare.19733509

Methods-only references

16. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo. (2021) doi:10.5281/zenodo.5127899.
17. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
18. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).
19. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
20. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
21. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
22. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).
23. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
24. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
25. R Core Team. R: A Language and Environment for Statistical Computing. (2021).
26. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
27. Xiao, N. ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'. (2018).
28. Yu, G. ggplotify: Convert Plot to 'grob' or 'ggplot' Object. (2021).
29. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2020).
30. Pedersen, T. L. patchwork: The Composer of Plots. (2020).
31. Murrell, P. & Wen, Z. gridGraphics: Redraw Base Graphics Using 'grid' Graphics. (2020).
32. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A Grammar of Data Manipulation. (2021).