MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities

Dongwan D Kang, Jeff Froula, Rob Egan, Zhong Wang

Grouping large genomic fragments assembled from shotgun metagenomic sequences to deconvolute complex microbial communities, or metagenome binning, enables the study of individual organisms and their interactions. Because of the complex nature of these communities, existing metagenome binning methods often miss a large number of microbial species. In addition, most of the tools are not scalable to large datasets. Here we introduce automated software, called MetaBAT that integrates empirical probabilistic distances of genome abundance and tetranucleotide frequency for accurate metagenome binning. MetaBAT outperforms alternative methods in accuracy and computational efficiency on both synthetic and real metagenome datasets. It automatically forms hundreds of high quality genome bins on a very large assembly consisting millions of contigs in a matter of hours on a single node. MetaBAT is open source software and available at https://bitbucket.org/berkeleylab/metabat .

MetaBAT, An Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities

Dongwan D. Kang[1, 2], Jeff Froula[1, 2], Rob Egan[1, 2], and Zhong Wang[1, 2, 3]*

[1] Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA
[2] Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3] School of Natural Sciences, University of California at Merced, Merced, CA, 95343, USA;

* To whom correspondence should be addressed. Tel: +1 925 296 5795; Email: zhongwang@lbl.gov

ABSTRACT

Grouping large genomic fragments assembled from shotgun metagenomic sequences to deconvolute complex microbial communities, or metagenome binning, enables the study of individual organisms and their interactions. Because of the complex nature of these communities, existing metagenome binning methods often miss a large number of microbial species. In addition, most of the tools are not scalable to large datasets. Here we introduce automated software, called MetaBAT that integrates empirical probabilistic distances of genome abundance and tetranucleotide frequency (TNF) for accurate metagenome binning. MetaBAT outperforms alternative methods in accuracy and computational efficiency on both synthetic and real metagenome datasets. It automatically forms hundreds of high quality genome bins on a very large assembly consisting millions of contigs in a matter of hours on a single node. MetaBAT is open source software available at https://bitbucket.org/berkeleylab/metabat.

INTRODUCTION

High throughput metagenome shotgun sequencing is a powerful tool to study microbial communities directly taken from their environment, thereby avoiding the requirement for cultivation or the biases that may arise from it. Assembling short metagenome shotgun reads into larger genomic fragments (contigs) by short read assemblers (Pevzner & Tang 2001; Pevzner et al. 2001) often fail to produce full-length genomes. Predicting draft genomes from assembled metagenomic contigs by metagenome binning provides a substitute for full-length genomes (Mande et al. 2012; Mavromatis et al. 2007). Despite their fragmented nature, these draft genomes are often derived from individual species (or population genomes representing

consensus sequences of different strains, Imelfort et al. 2014), and they approximate full genomes as they can contain a near full set of genes.

Two metagenome binning approaches have been developed (reviewed in (Mande et al. 2012)). The supervised binning approach uses known genomes as references and relies on either sequence homology or sequence composition similarity for binning (Krause et al. 2008; Wu & Eisen 2008). This approach does not work well on environmental samples where many microbes do not have closely related species with known genomes. In contrast, the unsupervised approach relies on either discriminative sequence composition (Teeling et al. 2004b; Yang et al. 2010) or species (or genomic fragments) co-abundance (Cotillard et al. 2013; Le Chatelier et al. 2013; Nielsen et al. 2014; Qin et al. 2012; Wu & Ye 2011) or both (Albertsen et al. 2013; Alneberg et al. 2014; Imelfort et al. 2014; Sharon et al. 2013; Wrighton et al. 2012; Wu et al. 2014) for binning. Recent studies have shown that species co-abundance feature can be very effective to deconvolute complex communities if there are many samples available (Albertsen et al. 2013; Alneberg et al. 2014; Cotillard et al. 2013; Imelfort et al. 2014; Karlsson et al. 2013; Le Chatelier et al. 2013; Nielsen et al. 2014; Sharon et al. 2013). A few recent methods, particularly CONCOCT (Alneberg et al. 2014) and GroopM (Imelfort et al. 2014), are also fully automated binning procedures.

Many of the above tools do not scale well to large metagenomic datasets. In this study, we developed MetaBAT (Metagenome Binning with Abundance and Tetra-nucleotide frequencies), as an efficient, fully automated software tool that is capable of binning millions of contigs from thousands of samples. By using a novel statistical framework to combine tetra-nucleotide frequency (TNF) and contig abundance probabilities, we demonstrated that MetaBAT produces high quality genome bins.

MATERIALS AND METHODS

An overview of MetaBAT software and its probabilistic models

As a pre-requisite for binning, the user must create BAM files (Li et al. 2009) by aligning the reads of each sample separately to the assembled metagenome (Figure 1. steps from 1 to 3). MetaBAT takes an assembly file (fasta format, required) and sorted bam files (one per sample, optional) as inputs. For each pair of contigs in a metagenome assembly, MetaBAT calculates

their probabilistic distance based on tetranucleotide frequency (TNF) and abundance (i.e. mean base coverage), then the two distances are integrated into one composite distance. All the pairwise distances form a matrix, which then is supplied to a modified k-medoid clustering algorithm to bin contigs iteratively and exhaustively into genome bins (Figure 1).

We use tetranucleotide frequency as sequence composition signatures as it has been previously shown that different microbial genomes have distinct TNF biases (Mrazek 2009; Pride et al. 2003; Saeed et al. 2012; Teeling et al. 2004a). To empirically derive a distance to discriminate TNFs of different genomes, we calculated the likelihood of inter- and intra-species Euclidean distance by using 1,414 unique, complete genome references from NCBI (Figure 2a). This empirically derived distance is termed tetranucleotide frequency probability distance (TDP).

To evaluate the effect of contig sizes on inter-species distance, we obtained posterior probability distributions of inter-species distance with several fixed sizes and observed better inter-species separation as contig size increases (Figure 2b). As contigs in real metagenome assemblies have various sizes, we then modeled TDP between contigs of different sizes by fitting a logistic function to reflect the dynamic nature of non-linear relationship between Euclidean TNF distance and TDP in different contig sizes. The results (Figure 2, c and d) suggested that the values of two parameters of the model, b and c, are unstable if the size of either contig is very small (< 2kb) and one should be cautious to allow smaller contigs to be binned.

Although contigs originating from the same genome are expected to have similar sequence coverage, i.e. genome abundance, the coverage of contigs can vary significantly within a library due to biases originated from the current sequencing technology (Benjamini & Speed 2012; Harismendy et al. 2009; Nakamura et al. 2011; Ross et al. 2013). As illustrated in Figure 2e, the observed coverage variance derived from data consisting of isolate genome sequencing projects (total 99 from IMG Database (Markowitz et al. 2012), henceforth referred as the IMG dataset) significantly deviate from the theoretical Poisson distribution, consistent with the notion that both the variance and the mean should be modelled (Clark et al. 2013). For computational convenience, we chose the normal distribution as an approximation since it fits the observation much better (Figure 2e). To compute the abundance distance of two contigs in one sample, we use the area not shared by their inferred normal distributions with given coverage mean and variance (Figure 2f). A geometric mean of the distances for all samples is used for the final abundance distance probability (ADP) of two contigs. In addition, we applied a progressive

weighting mechanism to adjust the relative strength of the information from abundance distance, meaning that we put more weight on abundance distance when it was calculated from many samples (see below).

We then integrate TDP and ADP of each contig pair as the following:

$$P(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \begin{cases} max\,(TDP, ADP), & \text{if } TDP > 0.05 \\ ADP \cdot w + TDP \cdot (1-w), & \text{otherwise} \end{cases}$$

, where w = min [log(n+1) / log(m+1), α]. n, m, and α represent the number of samples, a large number (100 as the default), and the maximum weight of ADP (0.9 as the default), respectively. For instance, in the default setting, the weight would be about 0.5 when the number of samples is 10 and TDP is less than 0.05. The resulting distance matrix is used for binning (see below).

Tetranucleotide Frequency Probability Distance (TDP)

To establish empirical probabilities of intra- and inter- species for tetranucleotide frequency distance, we downloaded 1,414 unique, completed bacterial genomes from the NCBI database and shredded them into fragments ranging from 2.5kb to 500kb. Next, we obtained 1 billion random contig pairs from within or between genomes. The empirical posterior probability that two contigs are from different genomes is given as the following:

$$P(T|D) = \frac{P(T)P(D|T)}{P(T)P(D|T) + P(R)P(D|R)}$$

, where T or R represent cases where two contigs are from different (inter) or the same (intra) species, respectively. D is the Euclidean TNF distance between two contigs. The same uninformative priors of T and R were chosen. In reality, P(T) is expected to be much bigger than P(R), thus we set P(T) = 10* P(R) as the default implementation to adjust the possible under-sampling issue in inter species distance.

The TDP of contig pairs with different sizes is approximated using logistic regression:

$$P(D_{ij}; b_{ij}, c_{ij}) = \frac{1}{(1 + e^{-(b_{ij} + c_{ij} * D_{ij})})}$$

, where D_{ij} represents a Euclidean TNF distance between contig *i* and *j*. b and c, the two parameters for the logistic regression, are estimated from the empirical data.

Abundance Distance Probability (ADP)

The probabilistic abundance distance was calculated as follows: Suppose two contigs have the mean coverage of μ_1 and μ_2 and the variances of σ_1^2 and σ_2^2, then we defined the abundance distance as the non-shared area of two normal distribution of $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$:

$$P(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \quad = \quad \frac{1}{2} \int |\phi_{\mu_1,\sigma_1^2} - \phi_{\mu_2,\sigma_2^2}|$$

, where ϕ represents a normal distribution having two parameters μ and σ^2. Numerically this can be simplified using cumulative distribution functions as follows assuming σ_2^2 is greater than or equal to σ_1^2:

$$P(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \begin{cases} \Phi_{\mu_1,\sigma_1^2}(k_0) - \Phi_{\mu_2,\sigma_2^2}(k_0), & \text{if } \sigma_1^2 = \sigma_2^2 \\ \Phi_{\mu_1,\sigma_1^2}(k_2) - \Phi_{\mu_1,\sigma_1^2}(k_1) + \Phi_{\mu_2,\sigma_2^2}(k_1) - \Phi_{\mu_2,\sigma_2^2}(k_2), & \text{otherwise} \end{cases}$$

, where Φ represents a cumulative normal distribution, and

$$k_0 = \frac{\mu_1 + \mu_2}{2},$$

$$k_1^* = \frac{\sqrt{\sigma_1^2 \cdot \sigma_2^2 \cdot ((\mu_1 - \mu_2)^2 - 2 \cdot (\sigma_1^2 - \sigma_2^2) \cdot \log(\sigma_2/\sigma_1))} - \mu_1 \cdot \sigma_2^2 + \mu_2 \cdot \sigma_1^2}{\sigma_1^2 - \sigma_2^2},$$

$$k_2^* = \frac{\sqrt{\sigma_1^2 \cdot \sigma_2^2 \cdot ((\mu_1 - \mu_2)^2 - 2 \cdot (\sigma_1^2 - \sigma_2^2) \cdot \log(\sigma_2/\sigma_1))} + \mu_1 \cdot \sigma_2^2 - \mu_2 \cdot \sigma_1^2}{\sigma_1^2 - \sigma_2^2},$$

$$k_1 = min(k_1^*, k_2^*) \; and \; k_2 = max(k_1^*, k_2^*)$$

To combine multiple abundance probabilities across different samples, we calculated the geometric mean of probabilities:

$$P_{ij} = \sqrt[\sum_n \mathbb{1}\{\mu_{in} > c \text{ OR } \mu_{jn} > c\}]{\prod_n P_{ijn}(\mu_{in}, \sigma_{in}^2, \mu_{jn}, \sigma_{jn}^2) * \mathbb{1}\{\mu_{in} > c \text{ OR } \mu_{jn} > c\}}$$

, where P_{ijn} represents the probability calculated from two abundances μ_{in} and μ_{jn}, and c represents a cut-off for reasonable minimum abundance for a contig.

As metagenome assemblies contain many small contigs, whether or not to include them is a dilemma; including small contigs will likely improve the genome completeness, at a cost in genome quality because their larger abundance variations make it harder to bin them correctly. We tried to empirically determine a reasonable contig size cut-off by plotting the ratio of mean and variance from the IMG single genome dataset. Although most genome variances are much larger than their means, their ratio becomes stabilized after contig size increases to 2.5kb

(Supplementary Figure S1). Therefore we used 2.5kb or larger contigs for the initial binning. Smaller contigs can be recruited after binning based on their correlation to the bins (Imelfort et al. 2014).

Iterative Binning

We modified the k-medoid clustering algorithm (Kaufman & Rousseeuw 1987) to eliminate the need to input a value for *k* and to reduce search space for efficient binning. Specifically, the binning algorithm works as the following:

1. Find a seed contig (e.g. having the greatest coverage), and set it as the initial medoid.
2. Recruit all other contigs within a cutoff distance (i.e. parameters p1 and p2) to the seed.
3. Find a new medoid out of all member contigs.
4. Repeat 2-3 until there are no further updates to the medoid. These contigs form a bin.
5. For the rest of the contigs, repeat 1-4 to form more bins until no contigs are left.
6. Keep large bins (e.g. >200kb), and dissolve all other bins into free contigs.
7. (optional) For dataset with at least 10 samples, recruit additional free contigs to each bin based on their abundance correlation.

RESULTS

Binning performance on "error-free" metagenomic assemblies

A metagenomic dataset (Accession #: ERP000108) from the MetaHIT consortium (henceforth referred to as the MetaHIT dataset) (Qin et al. 2010) was chosen to benchmark MetaBAT because it contains a large number of samples and the community contains many species with reference genomes. To derive a reference genome set, we selected 290 known genomes from NCBI that are present in MetaHIT at >5X mean coverage (Supplementary Table S1). These reference genomes were then shredded into contigs of random sizes (> 2.5kb) following an exponential distribution modeled to mimic real metagenome assemblies. The abundance of each contig in every sample was also obtained using real data. These "error-free" metagenome contigs, their abundance information, along with their parental reference genomes (as "true answers"), were used in the following analysis to benchmark binning performance. For a full description of the experiment, refer to MetaBAT wiki page:

https://bitbucket.org/berkeleylab/metabat/wiki/Home

For comparison, we ran several alternative binning tools on the same dataset described above. These software include Canopy (Nielsen et al. 2014), CONCOCT v.0.4.0 (Alneberg et al. 2014), GroopM v.0.3.0 (Imelfort et al. 2014), and MaxBin v.1.4.1 (Wu et al. 2014). Among them, CONCOCT, GroopM, and MaxBin are also fully automated binning tools. An optional manual step in GroopM for improving the quality of bins was excluded. Since MaxBin does not consider multiple samples, we combined multiple samples into one.

We used >90% precision (lack of contamination) and >30% recall (completeness) as the minimum criteria for a bin to be considered "good" which basically means the bin should be composed of one or more strains of a single species (for results of other thresholds, refer to Supplementary Figure S3 and S4). Formulas for this calculation are described in the Supplementary Material. Among all binning tools, MetaBAT binned the greatest number of genomes at almost every recall threshold (Figure 3A). CONCOCT is the only tool that produces more genome bins with over 80% or 90% completeness than MetaBAT, but MetaBAT produces many more bins at 70% completeness threshold. Interestingly, we found these tools complement each other in forming genome bins (Figure 3B, GroopM was omitted since it detects only a few genomes). Among the unique 133 genome bins collectively formed by all tools, MetaBAT binned the most number of genomes (111, 83.5%), with 23 bins (17.2%) that were not found by any other tool.

MetaBAT runs very efficiently in computation, as the entire binning process only took 14 minutes and 4GB of RAM (Table 1). Multiple simulations produced almost identical performance results and thus were not shown.

Binning performance on real metagenomic assemblies

We next tested the performance of MetaBAT on real metagenomic assemblies. Using the same raw sequence data from the above MetaHIT dataset, we pooled the sequences from all samples and then assembled them using Ray Meta assembler (Boisvert et al. 2012). Because real metagenomic assemblies often contain many small contigs, we lowered the minimum contig size requirement from 2.5kb to 1.5kb to include more contigs into our binning experiment. As a result, 118,025 contigs were used for binning. We then ran the above 5 binning tools with their default settings on this dataset. In contrast to the previous simulation experiment, in this experiment we

do not have a reference genome for every genome bin; we instead used CheckM (Parks et al. 2014) to calculate the approximate recall (percent of expected single-copy-genes that are binned) and precision (the absence of genes from different genomes) rates. A full description of this experiment is available of MetaBAT wiki page: https://bitbucket.org/berkeleylab/metabat/wiki/Benchmark_MetaHIT.

Similar to the previous "error-free" experiment, MetaBAT again identified the greatest number of unique genome bins having >90% precision (Figure 4A). In this experiment with real metagenomic contigs the superior completeness we saw in CONCOCT during the "error-free" experiment was lost. Moreover, the number of genome bins formed by MetaBAT was consistently greater than the others at every completeness threshold. Similarly different tools produced complementary binning results as before (Figure 4B). MetaBAT's contribution appears to be more pronounced this time. It missed 17 bins formed by all other tools combined, but recovered 31 bins that no other tools produced. MetaBAT alone recovered 90.2% (133/144) of genome bins from all tools. These results suggest MetaBAT is very robust against a real metagenome assembly. Consistent with the simulation experiment, MetaBAT is computationally very efficient and requires only 4 minutes to complete this experiment (Table 2).

To test the performance of MetaBAT on large-scale metagenomic data sets, we used a dataset containing 1,704 (with replicates) human gut microbiome samples (Accession #: ERP002061) (Nielsen et al. 2014). The entire dataset was assembled using Ray Meta assembler (Boisvert et al. 2012) and Megahit (Li et al. 2015) resulting in 1,058,952 contigs (>1kb) that were then used for binning. MetaBAT took less than 2 hours to generate 1,634 genome bins (>200kb) using a single node with 16 CPU cores (32 hyper-threads), and the peak memory consumption was at 17G. In comparison, Canopy took 18 hours using 36G memory in the same setting. The other binning tools--CONCOCT, GroopM, and MaxBin--all failed to generate any genome bins for this data set likely due to their inability to scale. For accuracy evaluation, we used CheckM (Parks et al. 2014) and identified 610 high quality bins (>90% precision and >50% completeness) among the bins predicted by MetaBAT, which is 35% more than the published CAG bins (Nielsen et al. 2014) and 11% more than Canopy bins using our assembly. For details to use MetaBAT on this large dataset please refer to: https://bitbucket.org/berkeleylab/metabat/wiki/Example_Large_Data.

Post-binning processing to further improve quality

Assembly from pooled samples in the above experiment raises the possibility that similar genomes (e.g., different strains) present in different samples are assembled into chimeric contigs. This level of contamination may not be tolerated in some applications. Based on the assumption that a single sample will contain fewer strains of the same species than all pooled samples, we implemented an optional post-binning process to reduce the strain-level contamination. Briefly, we first selected a single sample with the most reads mapped to a specific bin, and then assembled these reads into a new set of contigs. If the new contigs produces better CheckM results, we subsequently replaced the old contigs in this bin with the new ones.

This post-processing step significantly improved both completeness and precision (lack of contamination) for 61% (992 out of 1,634) of the genome bins. Overall there were 571 bins with >95% precision and >50% completeness, compared with 375 without post-processing. This improvement was more obvious when we increased the precision threshold to >99%, as the number high quality bins increases from 46 to 186.

By incorporating additional sequencing data and other post-binning optimizations, Nielson et al. generated 373 high quality draft genomes ("MGS genomes") (Nielsen et al. 2014). We therefore used these MGS draft genomes as reference for additional quality assessment of the MetaBAT genome bins after post-processing. As shown in Figure 5A, 31 MGS draft genomes were not well represented by MetaBAT bins, but MetaBAT recovered 55 additional genome bins not reported by MGS draft genomes. For those overlapping bins, most MetaBAT bins closely approximate the MGS draft genomes in accuracy—94% precision and 82% completeness (Figure 5B).

DISCUSSION

In conclusion, we developed an efficient and fully automated metagenome binning tool, MetaBAT, and evaluated its capability to reconstruct genomes using both synthetic and real world metagenome datasets. Applying MetaBAT to a large-scale complex human microbiome data recovers hundreds of high quality genome bins including many missed by alternative tools. An optional post-processing step improves the overall binning quality.

The algorithm implemented in MetaBAT is different from existing methods in several ways. For example, MetaBAT uses both sequence composition and co-abundance as features, while

Canopy (Nielsen et al. 2014) only uses co-abundance. Among the methods that use sequence composition feature, CONCOCT and GroopM both rely on principal component analysis (PCA) to discriminate contigs belong to different genomes. In contrast, MetaBAT and Maxbin calculate posterior probabilities by modeling intra- and inter-genome TNF distances. While Maxbin uses a universal model to calculate TNF distance probability for all contig sizes, MetaBAT uses a dynamic model weighing on different contig sizes.

One of the noticeable improvements in MetaBAT over other automated tools is its computational efficiency. In addition to the low memory requirement and fast computing speed, if one runs several rounds of binning to fine-tune parameters on a large dataset (by default MetaBAT does little parameter optimization), MetaBAT can be even faster as it saves intermediate calculations. For example, binning with ~1M contigs and ~1K samples for a second time only takes a few minutes.

There are a couple of considerations to keep in mind before applying MetaBAT. One important consideration is the minimum number of samples required for a reasonable binning performance. Although MetaBAT can run with only one sample or even in TNF-only mode for binning, as shown in Supplementary Figure S6, our general advice is that more samples achieve better binning results. The greater the abundance variation among samples of a target species, the more likely MetaBAT will produce a good genome bin for this species. A second consideration is the quality of the metagenome assembly. We do not expect binning to work well with poor metagenome assemblies, e.g. assemblies including many small contigs less than 1kb since the distance metrics computed for small contigs will not be very reliable.

ACKNOWLEDGEMENT

The authors thank Drs. Matt Blow, Rex Malmstrom and Tanja Woyke for their stimulating discussions and critical comments.

REFERENCES

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, and Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 31:533-538.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, and Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nat Methods* 11:1144-1146.

Benjamini Y, and Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 40:e72.

Boisvert S, Raymond F, Godzaridis E, Laviolette F, and Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* 13:R122.

Clark SC, Egan R, Frazier PI, and Wang Z. 2013. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* 29:435-443.

Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, Almeida M, Quinquis B, Levenez F, Galleron N, Gougis S, Rizkalla S, Batto JM, Renault P, consortium ANRM, Dore J, Zucker JD, Clement K, and Ehrlich SD. 2013. Dietary intervention impact on gut microbial gene richness. *Nature* 500:585-588.

Daniel R. 2005. The metagenomics of soil. *Nat Rev Microbiol* 3:470-478.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, and Frazer KA. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 10:R32.

Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, and Tyson GW. 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2:e603.

Kaufman L, and Rousseeuw P. 1987. *Clustering by means of medoids*: North-Holland.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, and Stoye J. 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* 36:2230-2239.

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jorgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clement K, Dore J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker JD, Raes J, Hansen T, Meta HITc, Bork P, Wang J, Ehrlich SD, and Pedersen O. 2013. Richness of human gut microbiome correlates with metabolic markers. *Nature* 500:541-546.

Li D, Liu CM, Luo R, Sadakane K, and Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.

Mande SS, Mohammed MH, and Ghosh TS. 2012. Classification of metagenomic sequences: methods and challenges. *Brief Bioinform* 13:669-681.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, and Kyrpides NC. 2012. IMG: the

Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 40:D115-122.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, and Kyrpides NC. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* 4:495-500.

Mrazek J. 2009. Phylogenetic Signals in DNA Composition: Limitations and Prospects. *Molecular Biology and Evolution* 26:1163-1169.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, and Kanaya S. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 39:e90.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto JM, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Dore J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Leonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sorensen S, Tap J, Tims S, Ussery DW, Yamada T, Meta HITC, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, and Ehrlich SD. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol*.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, and Tyson GW. 2014. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. PeerJ PrePrints.

Pevzner PA, and Tang H. 2001. Fragment assembly with double-barreled data. *Bioinformatics* 17 Suppl 1:S225-233.

Pevzner PA, Tang H, and Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* 98:9748-9753.

Pride DT, Meinersmann RJ, Wassenaar TM, and Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 13:145-158.

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto JM, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, and Wang J. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490:55-60.

Qin JJ, Li RQ, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li JH, Xu JM, Li SC, Li DF, Cao JJ, Wang B, Liang HQ, Zheng HS, Xie YL, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu HM, Yu C, Li ST, Jian M, Zhou Y, Li YR, Zhang XQ, Li SG, Qin N,

Yang HM, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J, and Consortium M. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59-U70.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, and Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biol* 14:R51.

Saeed I, Tang SL, and Halgamuge SK. 2012. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res* 40:e34.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, and Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 23:111-120.

Teeling H, Meyerdierks A, Bauer M, Amann R, and Glockner FO. 2004a. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* 6:938-947.

Teeling H, Waldmann J, Lombardot T, Bauer M, and Glockner FO. 2004b. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5:163.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, and Banfield JF. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337:1661-1665.

Wu M, and Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 9:R151.

Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, and Singer SW. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2:26.

Wu YW, and Ye Y. 2011. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comp Biol* 18:523-534.

Yang B, Peng Y, Leung HC, Yiu SM, Chen JC, and Chin FY. 2010. Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics* 11 Suppl 2:S5.

FIGURES LEGENDS

Figure 1. Overview of the MetaBAT pipeline. Three preprocessing steps before MetaBAT are applied: 1) A typical metagenome experiment may contain many spatial or time-series samples, each consisting of many different genomes (different color circles). 2) Each sample is sequenced by next-generation sequencing technology to form a sequencing library with many short reads. 3) The libraries are combined for de novo assembly. After assembly, the reads from each sample are aligned to form separate BAM files. MetaBAT then automatically performs the remaining

steps: 4) For each contig pair, a tetranucleotide frequency distance probability (TDP) is calculated. 5) For each contig pair, an abundance distance probability (ADP) across all the samples is calculated. 6) The TDP and ADP of each contig pair are then combined, and the resulting distance for all pairs form a distance matrix. 7) Each bin will be formed iteratively and exhaustively from the distance matrix.

Figure 2. Probabilistic modeling of TNF and Abundance distances. **a-d**) TNF distance modeling. **a**) Empirical probabilities of intra- (solid gray line) or inter- (dotted gray line) species Euclidean TNF distance are estimated from sequenced genomes. The posterior probability of two contigs originated from different genomes given a TNF distance is shown as a red solid line. All probabilities are calculated using a fixed contig size of 10kb. **b**) Different posterior inter-species probabilities for two equal-size contigs under various contig sizes. **c, d**) The estimation of parameters for a logistic curve with two contigs of different sizes. x and y axis represent the lengths of short and long contig, respectively, and z axis represents the estimates of each parameter *b* or *c* in a logistic curve, TDP = 1/(1+exp(-(*b*+*c**TNF))), where TNF and TDP represents the Euclidean TNF distance and probabilistic TNF distance, respectively. **e-f**) Abundance distance modeling. **e**) The relationship between mean and variance of base depths (coverage) which were shown in x and y axis, respectively. Each dot represents this relationship in each genome, which calculated by median of mean and variance of the coverage. Theoretical Poisson model was shown as blue line and normal model was shown as red line. **f**) Probabilistic abundance distance between two contigs. The shaded area represents the abundance distance between two contigs in a given library.

Figure 3. Binning performance on synthetic metagenomic assemblies. A) The number of genomes (X-axis) identified by each binning method (Y-axis) in different recall (completeness) threshold and the same >90% precision threshold. B) A Venn diagram of identified genomes by top 4 binning methods.

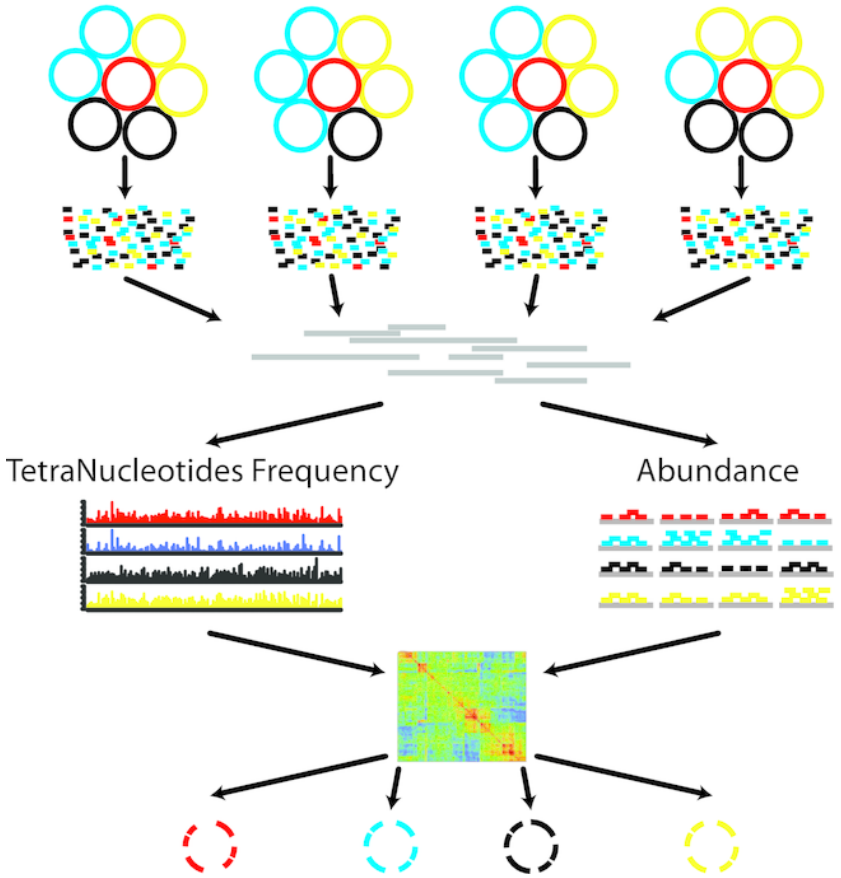
Figure 4. Binning performance on real metagenomic assemblies. A) The number of genomes (X-axis) identified by each binning method (Y-axis) in different recall (completeness) threshold and the same >90% precision threshold. B) A Venn diagram of identified genomes by top 4 binning methods.

Figure 5. A comparison between MetaBAT bins after post-processing and MGS draft genomes from Nielsen et al. A) A Venn diagram of identified genome bins by MetaBAT and MGS draft genomes. B) A scatterplot of completeness and precision for MetaBAT genome bins using MGS draft genomes as reference. X-axis represents shared proportion of bases in terms of MetaBAT bins (i.e. precision), and y-axis represents shared proportion of bases in terms of MGS genomes (i.e. completeness). Each circle represents a unique MetaBAT bin having a corresponding MGS genome (342 bins in total), and the size of the circle corresponds to the bin size.

1

Figure 1. Overview of the MetaBAT pipeline.

Three preprocessing steps before MetaBAT is applied: 1) A typical metagenome experiment may contain many spatial or time-series samples, each consisting of many different genomes (different color circles). 2) Each sample is sequenced by next-generation sequencing technology to form a sequencing library with many short reads. 3) The libraries may be combined before de novo assembly. After assembly, the reads from each sample must be aligned in separate BAM files. MetaBAT then automatically performs the remaining steps: 4) For each contig pair, a tetranucleotide frequency distance probability (TDP) is calculated from a distribution modelled from 1,414 reference genomes. 5) For each contig pair, an abundance distance probability (ADP) across all the samples is calculated. 6) The TDP and ADP of each contig pair are then combined, and the resulting distance for all pairs form a distance matrix. 7) Each bin will be formed iteratively and exhaustively from the distance matrix.



Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

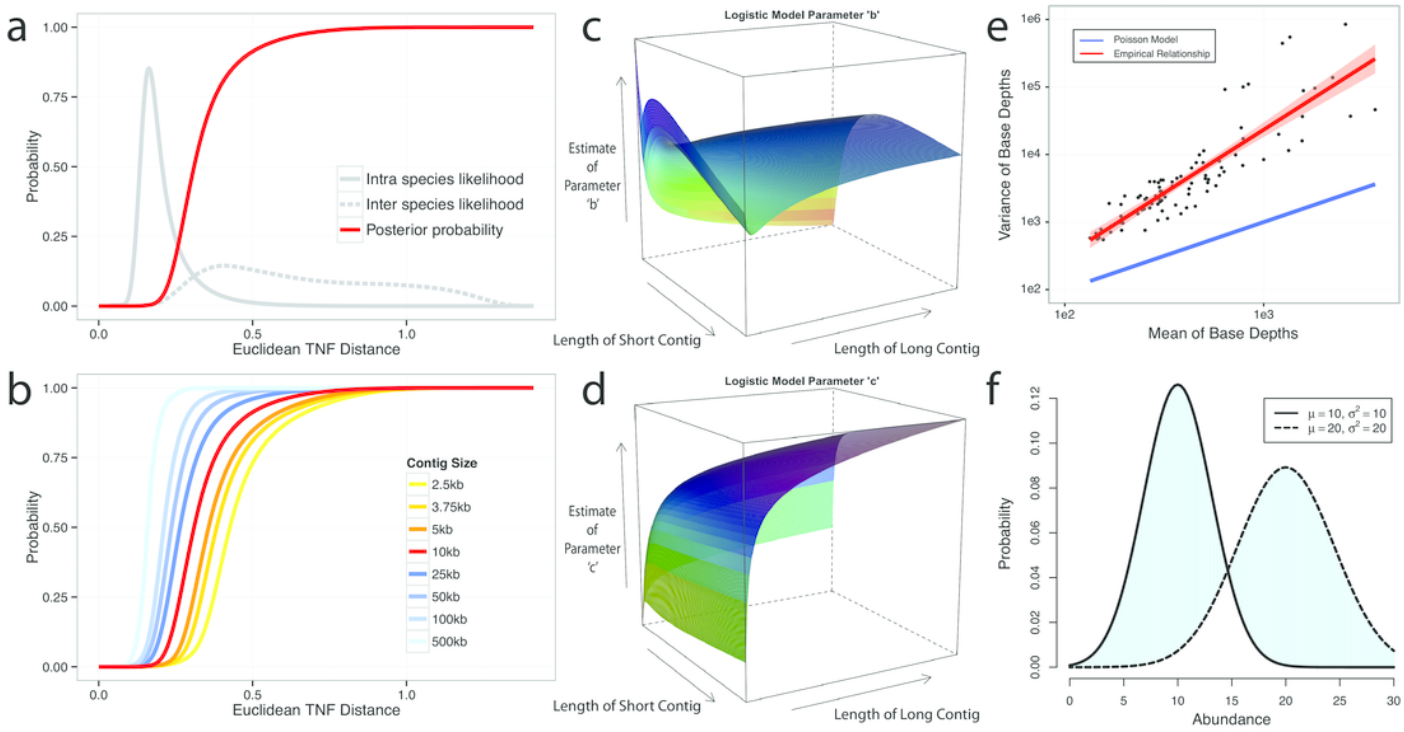
MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

2

Figure 2. Probabilistic modeling of TNF and Abundance distances.

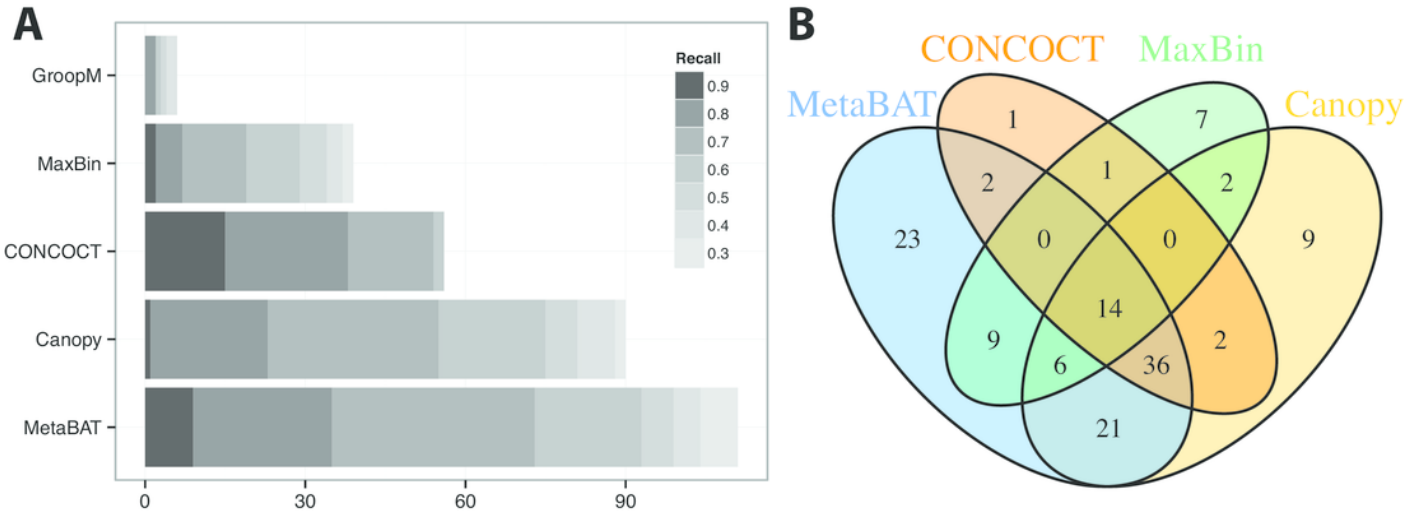
a-d) TNF distance modeling. **a**) Empirical probabilities of intra- (solid gray line) or inter- (dotted gray line) species Euclidean TNF distance are estimated from sequenced genomes. The posterior probability of two contigs originated from different genomes given a TNF distance is shown as a red solid line. All probabilities are calculated using a fixed contig size of 10kb. **b**) Different posterior inter-species probabilities for two equal-size contigs under various contig sizes. **c, d**) The estimation of parameters for a logistic curve with two contigs of different sizes. x and y axis represent the lengths of short and long contig, respectively, and z axis represents the estimates of each parameter *b* or *c* in a logistic curve, TDP = 1/(1+exp(-(*b*+*c**TNF))), where TNF and TDP represents the Euclidean TNF distance and probabilistic TNF distance, respectively. **e-f**) Abundance distance modeling. **e**) The relationship between mean and variance of base depths (coverage) which were shown in x and y axis, respectively. Each dot represents this relationship in each genome, which calculated by median of mean and variance of the coverage. Theoretical Poisson model was shown as blue line and normal model was shown as red line. **f**) Probabilistic abundance distance between two contigs. The shaded area represents the abundance distance between two contigs in a given library.


a
Probability
Euclidean TNF Distance
Intra species likelihood
Inter species likelihood
Posterior probability
b
Probability
Euclidean TNF Distance
Contig Size
2.5kb
3.75kb
5kb
10kb
25kb
50kb
100kb
500kb
c
Logistic Model Parameter 'b'
Estimate of Parameter 'b'
Length of Short Contig
Length of Long Contig
d
Logistic Model Parameter 'c'
Estimate of Parameter 'c'
Length of Short Contig
Length of Long Contig
e
Poisson Model
Empirical Relationship
Variance of Base Depths
Mean of Base Depths
f
Probability
Abundance

3

Figure 3. Binning performance on synthetic metagenomic assemblies.

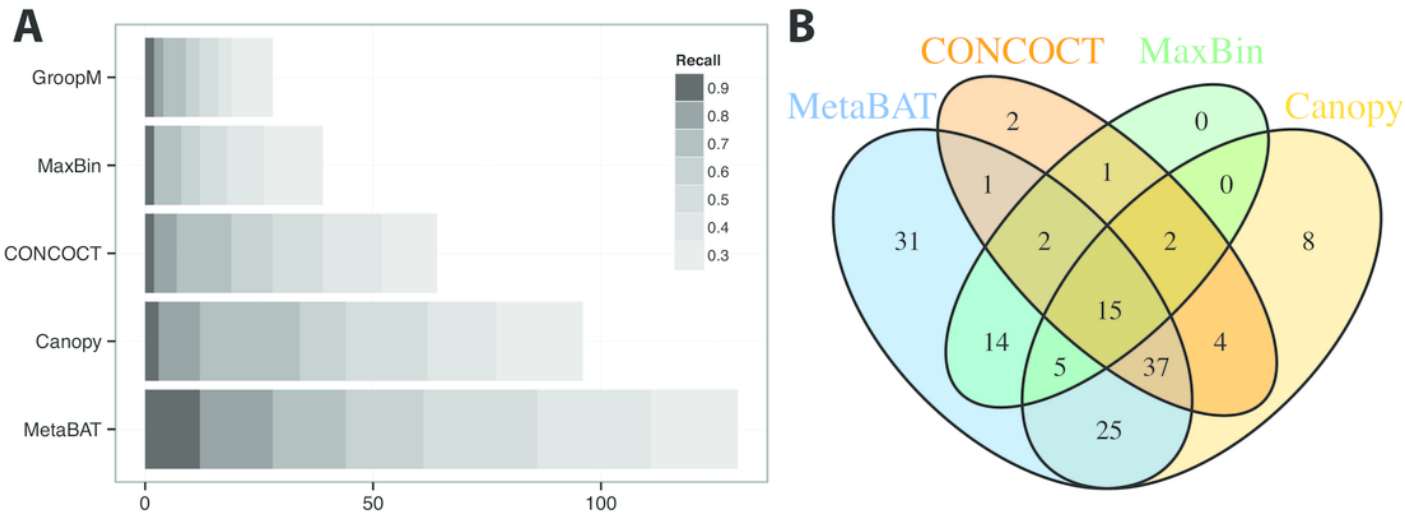
A) The number of genomes (X-axis) identified by each binning method (Y-axis) in different recall (completeness) threshold and >90% precision, which calculates the lack of contamination. B) Venn diagram of identified genomes by top 4 binning methods.



4

Figure 4. Binning performance on real metagenomic assemblies.

A) The number of genomes (X-axis) identified by each binning method (Y-axis) in different recall (completeness) threshold and >90% precision, which calculates the lack of contamination. B) Venn diagram of identified genomes by top 4 binning methods.



5

Figure 5. Comparison between MetaBAT bins after post-processing and MGS draft genomes from Nielsen et al.

A) Venn diagram of identified genome bins by MetaBAT having >90% precision and >30% completeness calculated by CheckM and one-to-one corresponding genomes in MGS draft genomes. B) Scatterplot of completeness and precision for MetaBAT genome bins when considered MGS draft genomes as the gold standard. X-axis represents shared proportion of bases in terms of MetaBAT bins (i.e. precision), and y-axis represents shared proportion of bases in terms of MGS genomes (i.e. completeness). Each circle represents a unique MetaBAT bins having uniquely corresponding MGS genomes (342 bins in total), and the size of it corresponds to bin size.

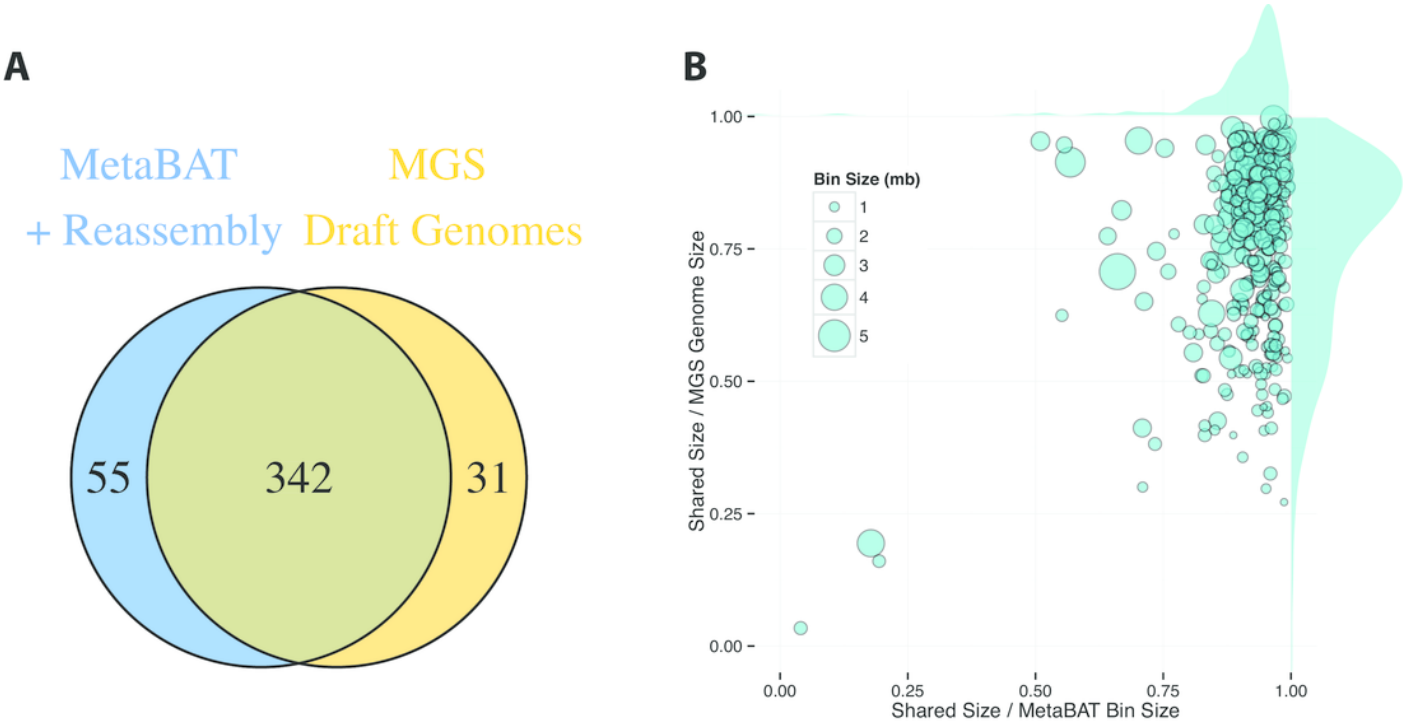


Table 1(on next page)

Table 1

A summary of the binning performance on synthetic metagenomic assembly.

Table 1. A summary of the binning performance on synthetic metagenomic assembly.

	MetaBAT	Canopy	CONCOCT	MaxBin	GroopM**
Number of Bins Identified (>200kb)	340	230	235	260	445
Number of Genomes Detected (Precision > .9 & Recall > .3)	111	90	56	39	6
Wall Time (16 cores; 32 hyper-threads)	00:13:55	00:21:01*	104:58:01	20:51:19	45:29:39
Peak Memory Usage (for binning step)	3.9G	1.82G*	9.55G	7.7G	38G

*Canopy only use abundance table as input, so it should have taken more time and memory to read and write sequence data like the others

**Manual steps were not used

Table 2(on next page)

Table 2

A summary of the binning performance on real metagenomic assembly.

Table 2. A summary of the binning performance on real metagenomic assembly.

	MetaBAT	Canopy	CONCOCT	MaxBin	GroopM**
Number of Bins Identified (>200kb)	234	223	260	168	335
Number of Genomes Detected (Precision > .9 & Recall > .3)	130	96	64	39	28
Wall Time (16 cores; 32 hyper-threads)	00:03:36	00:02:31*	82:19:53	06:49:39	12:19:12
Peak Memory Usage (for binning step)	3.0G	1.6G*	7G	5.8G	6.3G

*Canopy only use abundance table as input, so it should have taken more time and memory to read and write sequence data like the others

**Manual steps were not used