MetaBAT: Metagenome Binning based on Abundance and Tetranucleotide frequency

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

1Department of Energy Joint Genome Institute, Walnut Creek, CA 94598 USA
2Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA
*Email Address: ddkang@lbl.gov

March 2014

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

MetaBAT: Metagenome Binning based on Abundance and Tetranucleotide frequency

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

[1] Department of Energy Joint Genome Institute, Walnut Creek, CA
[2] Genomics Division, Lawrence Berkeley National Lab, Berkeley, CA

Abstract

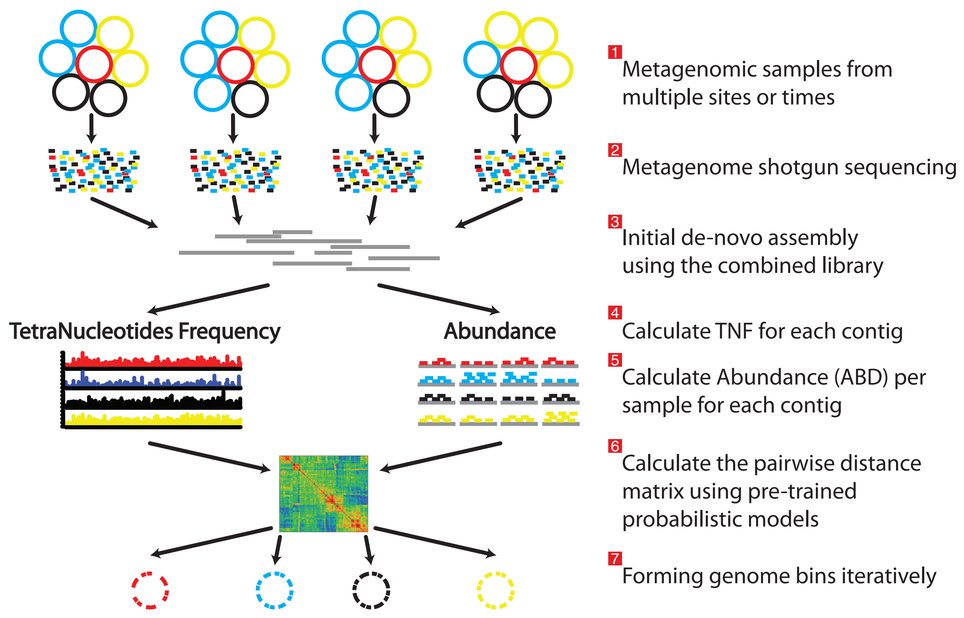
Grouping large fragments assembled from shotgun metagenomic sequences to deconvolute complex microbial communities, or metagenome binning, enables the study of individual organisms and their interactions. Here we developed an automated metagenome binning software, called MetaBAT, that integrates empirical probabilistic distances of genome abundance and tetranucleotide frequency. On synthetic datasets MetaBAT on average achieves 98% precision and 90% recall at the strain level with 281 near complete unique genomes. Applying MetaBAT to a human gut microbiome data set we recovered 176 genome bins with 90% precision and 78% recall. Further analyses suggest MetaBAT is able to recover genome fragments missed in reference genomes up to 19%, while 53 genome bins are novel. In summary, we believe MetaBAT is a powerful tool to facilitate comprehensive understanding of complex microbial communities.

Introduction

Direct assembly of full-length genomes from complex microbial communities is impractical. Therefore, an intermediate step grouping contigs from the same organism, called metagenome binning, has been adopted [1, 2]. Draft genomes formed by metagenome binning, although lacking contiguity, approximates reference genomes as it can contain the full set of genes of a species.

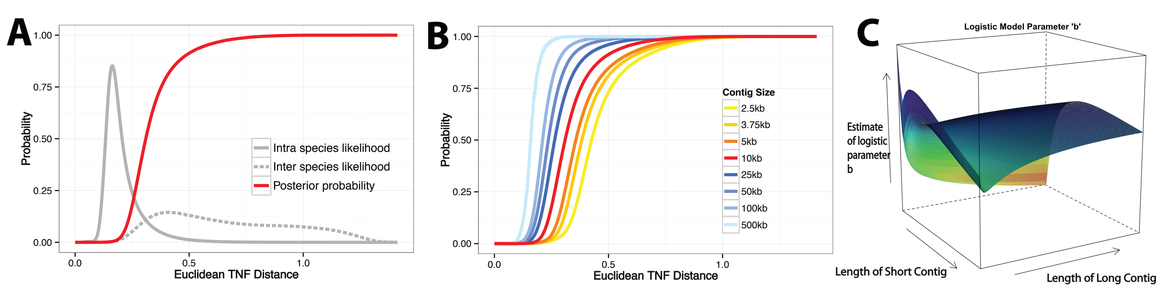
Two classes of metagenome binning approaches have been developed recently. The supervised binning approach requires known genomes, which does not work well on environmental samples. To overcome this limit, the unsupervised approach relies on either oligonucleotide composition biases or tspecies abundance, or both, to bin metagenomic reads or contigs. Most of the methods based on this approach depend on data visualization techniques to manually select a few species and determine their boundaries; therefore they are not suitable for binning large and complex metagenome data sets comprehensively.

MetaBAT Pipeline



Briefly, MetaBAT works as the following. For each pair of contigs, it first calculates two probabilities of pairwise distances from genome signatures and abundances, and it integrates all pairwise probabilities into a composite distance matrix. It then employs a modified k-medoid clustering algorithm to iteratively cluster the contigs into genome bins, each of which corresponds to a single genome.

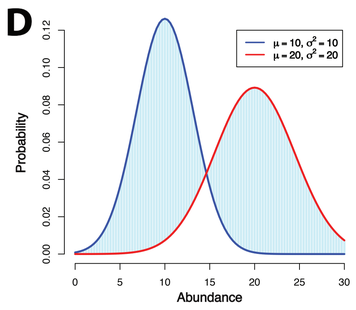
Probablistic Modeling of TNF and ABD



A. Converting Euclidean TNF distance to empirical probability by statistical modeling. The likelihood of inter- and intra-species distance using unique, complete genomes from NCBI, and posterior probability distribution of inter-species distance for a pair of genomic fragments of fixed size (10kb).

B. Dynamics of posterior probabilities depend on contig sizes. Better inter-species separation is achieved with larger fragment sizes. And each line can be fit to a logistic curve, F(*d*) = 1/(1+exp(-(*b*+*c***d*))), where parameters b and c are functions of contig sizes.

C. Parameter estimation of a dynamic logistic curve based on two different contig lengths. We modeled the posterior inter-species probability as a logistic curve that changes depending on two different contig lengths.



D. Probabilistic abundance distance between two contigs. The shaded area represents the probability of two contigs originating from different genomes calculated as follows:

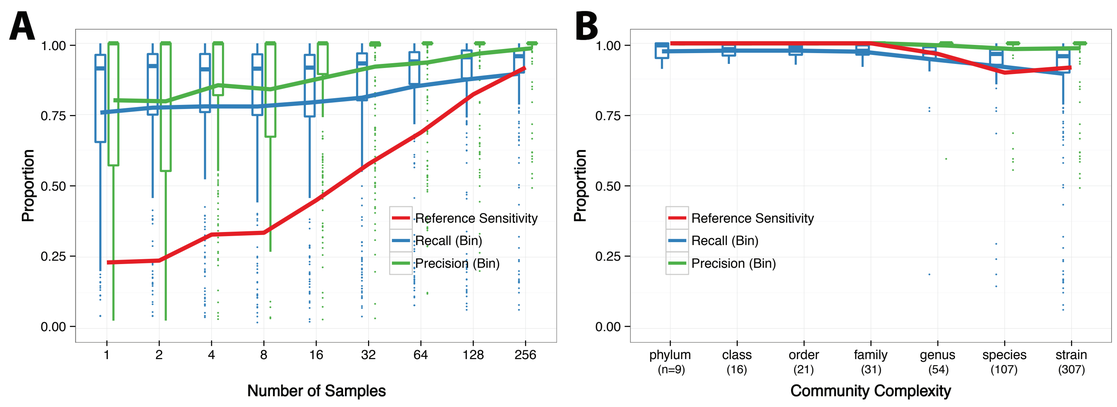
$$P(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \quad = \quad \frac{1}{2} \sum_{i=0}^{\infty} |\phi_{\mu_1,\sigma_1^2}(i) - \phi_{\mu_2,\sigma_2^2}(i)|$$

The combined probability of multiple abundance distances are calculated as weighted geometric mean of each probabilities:

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

Performance on synthetic metagenome data

Using 307 draft genomes from MetaHIT (Metagenomics of the Human Intestinal Tract) data, we tested the effect of the number of samples and community complexity to MetaBAT performance.

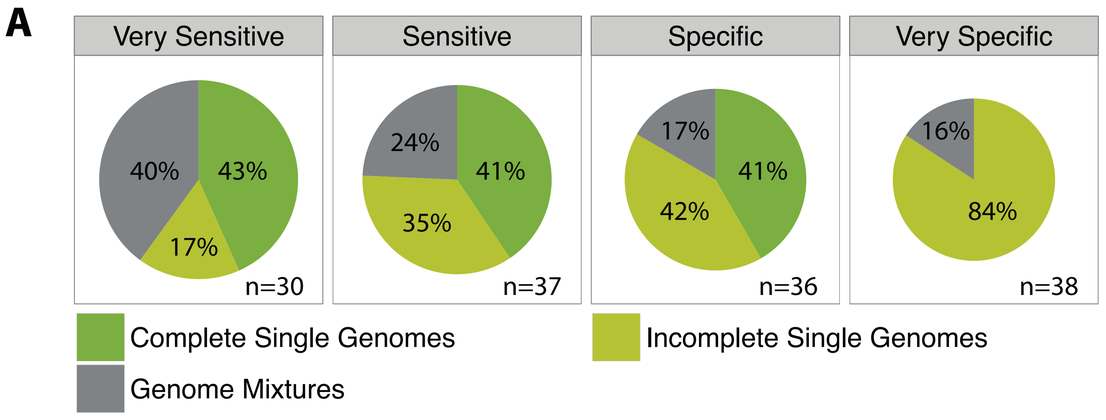


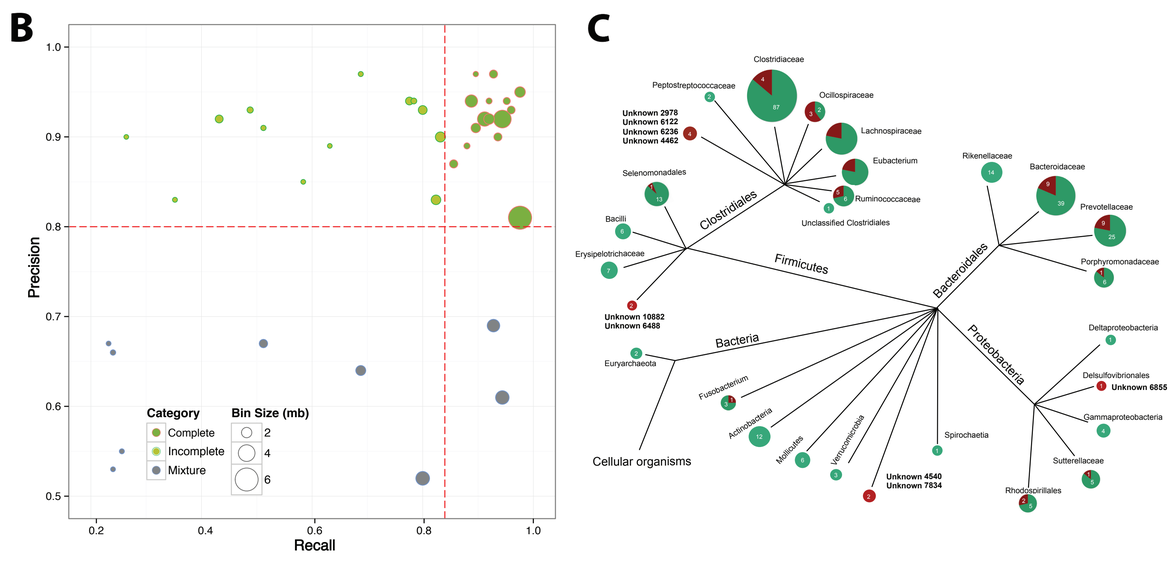
A. Both the precision and recall of each bin found by MetaBAT increase as the number of samples increases. Reference sensitivity result suggests that more genome bins are formed in great quality, as more samples are included in the analysis.

B. MetaBAT is robust at various phylogeny resolutions. Genome bins from MetaBAT contain no contamination from other families, and very few contigs from other genus. At species level it achieves on average 92% recall and 98% precision on this data set. Interestingly, it achieves almost the same performance in strain level with 90% recall and 98% precision with recovery of 281 genomes out of 307.

Recovering hundreds of genomes from MetaHIT

To validate the performance of MetaBAT on real metagenomic data sets, we applied MetaBAT to a MetaHIT dataset with 262 human gut microbiome samples [3]. We assembled the entire dataset using Ray Meta Assembler [4] and selected 60,619 contigs longer than 2.5kb for binning. To systematically explore the trade-offs between precision and recall rates, we ran MetaBAT under 4 pre-specified parameter settings: very sensitive, sensitive, specific, and very specific. The purpose of the parameter is to decide tightness of genome bins and their boundaries: bin sizes would be largest in 'very sensitive' setting but would be least homogeneous, in contrast 'very specific' would identify the smallest size of bins but highly homogeneous.





A. Comparative results from 4 pre-specified parameter settings in MetaBAT. Each panel represents the result from different settings of sensitivity and specificity. 4 evaluation categories were defined as follows: Complete Single Genomes, which are composed of bins having 80% or greater recall and precision; Incomplete Single Genomes, having less than 80% recall with 80% or greater precision; Genome Mixtures, having less than 80% precision; Novel Genomes, which has no match with any known genomes in NCBI databases.

B. Distribution of each bin in terms of recall and precision using the result from 'Sensitive' setting. Each color category corresponds to the same category in panel A and the sizes of circle corresponds to bin sizes. The upper right corner represents 'Complete Single Genome.'

C. A phylogenetic tree from our bins. Unknown bins (red) were placed on the tree by MEGAN [5]. 32 of the 53 novel bins could be placed within a genus. 12 into a family, 4 into an order, 2 into a phyla, and 3 only at the kingdom level.

References

[1] Mande, S.S., M.H. Mohammed, and T.S. Ghosh, Classification of metagenomic sequences: methods and challenges. Brief Bioinform, 2012. 13(6): p. 669-81.
[2] Mavromatis, K., et al., Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods, 2007. 4(6): p. 495-500.
[3] Qin, J.J., et al., A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010. 464(7285): p. 59-U70.
[4] Boisvert, S., et al., Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol, 2012. 13(12): p. R122
[5] Huson, Daniel H., et al. "Integrative analysis of environmental sequences using MEGAN4." Genome research 21.9 (2011): 1552-1560.