Long read- and DNA methylation-based binning of metagenomic contigs and single molecules

Background

Whole-metagenome shotgun sequencing is a comprehensive approach for characterizing the population structure and genetic architecture of complex microbial communities. However, significant challenges arise in the analysis of metagenomic sequences, often stemming from the presence of bacterial species and strains with high sequence similarity, complex DNA repeats, and widely varying relative abundance. Short-read metagenomic assemblies usually result in many thousands of short-to-medium length contigs that subsequently must be annotated using existing reference sequences or segregated by taxa through the process of binning.

Supervised binning methods require existing references to train classification algorithms, while unsupervised (reference-free) methods do not rely on any training data and therefore have the potential to identify novel species. Most reference-free binning methods attempt to cluster contigs from a metagenomic assembly, either using sequence composition alone or using coverage covariance statistics[1]. Sequence composition-based approaches, however, are limited by the fragmented nature of the *de novo* assemblies and often fail to distinguish between genomes with high sequence homology. Coverage covariance-based methods can provide additional power to separate similar genomes, but require the sequencing of many related samples, which is often not feasible, especially in clinical settings when in-depth analysis of few or a single microbiome is desired.

Single-molecule, real-time (SMRT) sequencing has the potential to address many of these challenges, but its applicability in metagenomics has not been extensively explored. Here, we present multiple novel methods for binning metagenomic sequences that not only leverage the long read lengths of SMRT sequencing, but also, for the first time, utilize the DNA methylation signatures inferred from these reads to resolve assembled contigs and single reads into clusters representing distinct biological entities. The diversity of methylation systems found in the bacterial world[2,3], often observed between closely related species and strains, suggests that DNA methylation profiles can be leveraged as an epigenetic feature to separate taxa with high sequence homology. Using single-molecule methylation detection[4] and metagenomic sequencing data from several synthetic microbiome communities and infant gut microbiota, we demonstrate that the proposed methods can segregate both assembled contigs and low abundance reads at a high resolution.

Results

The proposed contig- and read-binning framework relies on the use of two types of features: sequence composition as assessed by k-mer frequencies and DNA methylation profiles (Figure 1a). To evaluate the power of these binning methods, we build upon dimensionality reduction methods that have been shown to provide relatively effective segregation of metagenomic contigs using the t-distributed stochastic neighbor embedding (t-SNE) algorithm.

While previous methods used sequence composition and t-SNE to bin assembled contigs, we incorporate the DNA methylation profiles and further extend the application to unaligned SMRT reads. The inclusion of DNA methylation profiles for binning purposes significantly increases the ability to segregate contigs from closely related species and strains in an infant microbiome sample. By extending the binning methods to unaligned SMRT reads (Figure 1b), we can remove the requirement for a successful *de novo* assembly and thus highlight taxon-specific clusters for very low-abundance members of a metagenomic mixture that would otherwise fail to generate contigs. Finally, we can improve the quality of multi-strain metagenomic assemblies by first binning the reads based on DNA methylation profiles. The binned reads are then assembled separately, generating strain-specific assemblies without the contig fragmentation and chimerism that occurs when multiple strains are assembled together.

Conclusions

The methods discussed here are empowered by emerging sequencing technologies that are not subject to the limitations of short read lengths or biases in amplification and sequencing of GC-rich regions, allowing a more accurate and comprehensive reconstruction of microbial genomes in a metagenomic sample. Furthermore, this work represents the first time that DNA methylation

signatures have been used to improve metagenomic binning and assembly, an advance that facilitates the separation of high-homology species and strains, both at the level of assembled contigs and raw reads.

In addition to helping examine the taxonomic diversity in metagenomic sequencing data, the method will also be useful for associating plasmids to their bacterial hosts, as the plasmids inherit the same methylation signature as their host despite possible differences in sequence composition. This is an important step in understanding the full genomic potential of a certain species in a sample. In addition, bacteriophages that are often responsible for the transmission of antibiotic resistance carry methylation signatures indicative of their most recent host organism. Therefore, methylation profiles can be used to track the transmission of bacteriophages and associated antibiotic resistance elements among species and strains.

The sensitivity and effectiveness of the approaches described here, and interest in long read metagenomic sequencing generally, will only increase as third generation sequencing technologies continue to mature, generating larger yields and longer reads. These methods will serve as a framework for the development of further approaches that take advantage of the unique features of existing and emerging third generation sequencing technologies for characterization of metagenomic communities.

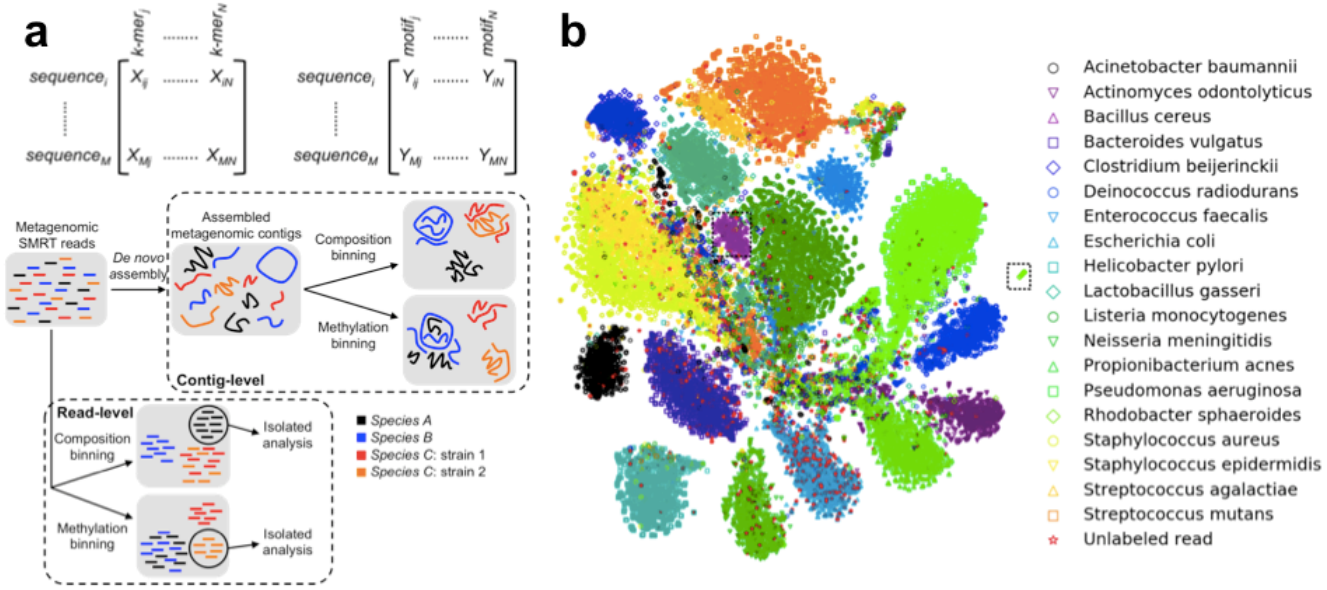

Figure 1: (a) Schematic of metagenomic SMRT binning approaches, where composition-based binning helps separate divergent species, while methylation-based binning helps untangle highly similar genomes. The two binning approaches can also be combined to leverage the strengths of each. (b) Composition-based binning of unaligned reads (length>15kb) from a Human Microbiome mock community containing 20 members. Clusters of reads belonging to two low-abundance species, *Bacillus cereus* and *Rhodobacter sphaeroides*, are highlighted.

1. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** (2014).
2. Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat. Biotechnol.* **30,** 1232–9 (2012).
3. Blow, M. J. *et al.* The Epigenomic Landscape of Prokaryotes. *PLOS Genet.* **12,** e1005854 (2016).
4. Beaulaurier, J. *et al.* Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat. Commun.* **6,** 7438 (2015).