OXFORD
UNIVERSITY PRESS | Bioinformatics

Binning_refiner: improving genome bins through the combination of different binning programs

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2016-1806
Category:	Applications Note
Date Submitted by the Author:	10-Nov-2016
Complete List of Authors:	Song, Weizhi; University of New South Wales, Centre for Marine Bio-Innovation Thomas, Torsten; UNSW, School of Biotechnology and Biomolecular Sciences
Keywords:	Bioinformatics, Metagenomics, Microbiology

Bioinformatics, xxxx, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Genome analysis

Binning_refiner: improving genome bins through the combination of different binning programs

Wei-Zhi Song and Torsten Thomas*

Centre for Marine Bio-Innovation, University of New South Wales, Sydney, NSW 2052, Australia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Microbial genomes have recently been reconstructed from metagenomic dataset using binning approaches. Inconsistent binning results are however often observed between different binning programs, likely due to the different algorithm or statistical models used. We present Binning_refiner, a pipeline that merges the results of different binning programs. Our results demonstrated that this pipeline can significantly reduce the contamination level of genome bins and increase the contamination-free total genome size. Binning_refiner is thus a useful tool to improve the quality of genome bins derived from metagenomic data.

Availability: Binning_refiner is implemented in Python3 and is freely available at: https://github.com/songweizhi/Binning_refiner.

Contact: songwz03@gmail.com, t.thomas@unsw.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput shotgun sequencing provides a powerful way to study the "unexplored" and uncultured diversity of microbial communities (Eloe-Fadrosh *et al*., 2016). A number of programs have been recently developed to reconstruct complete or partial microbial genomes from metagenomics shotgun sequences, a process called genome binning (Alneberg *et al*., 2014; Imelfort *et al*., 2014; Kang *et al*., 2015; Lin and Liao, 2016). These programs cluster contigs assembled from metagenomics sequences based on compositional properties (e.g. GC content, tetra-nucleotide frequency) or sequence coverage profiles across multiple samples (or combinations thereof). Inconsistent results have however often been observed between binning programs, which is likely due to the differences in algorithm or statistical models employed (Kang *et al*., 2015; Lin and Liao, 2016). Furthermore, it is important that all sequences of a bin are specific to a given organism (i.e. bins are free of contamination from any other organism) as this would otherwise lead to erroneous functional and metabolic inferences (Albertsen *et al*., 2013; Rinke *et al*., 2013). To assess this, a number of algorithms have been established that measure the level of contamination and genome completeness using sets of marker sequences (Parks *et al*., 2015; Sangwan *et al*., 2016; Seah and Gruber-Vodicka, 2015).

Here we have developed a pipeline called Binning_refiner that reconciles the outputs of different binning programs with the aim to improve the quality of genomic bins, in particular with respect to contamination levels.

2 Methods

The main steps of Binning_refiner are summarized in **Fig. 1A**. The output bins of two binning programs are used as inputs for Binning_refiner. In the first step, the bin names are added to contig names. The bins from each binning program are then combined into one file and a pairwise blastN (Madden, 2013) is performed between the two combined files. Blast results are then filtered to keep only full-length matches. Each set of shared contigs between the two sets of bins are treated as a refined bin (see **Supplementary Figure S1**) and exported in multi-fasta format, if their total length is longer than the defined cutoff (e.g. 0.5 Mbp). CheckM (Parks *et al*., 2015) is used after Binning_refiner to assess the quality (contamination and completeness) of input and refined bins.

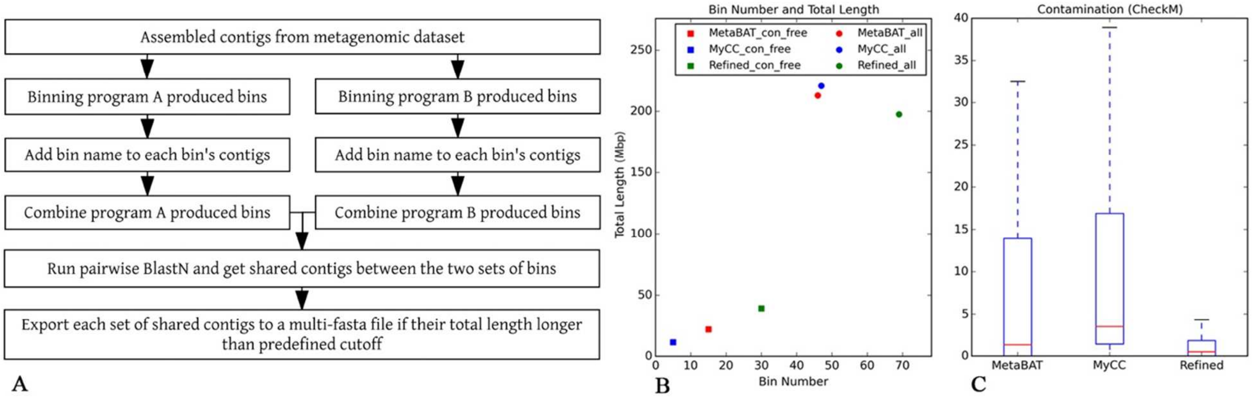

Fig. 1. (A) Binning refiner workflow. (B-C) Qualities of bins from real metagenomic dataset. "con free" refers to "contamination-free bins", while "all" refers to all bins produced.

3 Results and Discussion

Binning_refiner's performance was first assessed on the MBARC-26 mock dataset, which consists of shotgun sequences for a defined mixture of 23 bacterial and 3 archaeal strains with publicly available complete genomes (Singer *et al*., 2016). Sequence assembly using idba_ud (Peng *et al*., 2012) generated 92.4 Mbp of contiguous sequences ⩾ 5 Kbp, and of which 52.3 Mbp were assigned to 24 complete genomes with an identity cutoff of 99% (contigs assigned to more than one genomes were excluded). These 52.3 Mbp of contigs with clear genome assignments were used as input for the binning programs MetaBAT (Kang *et al*., 2015) and MyCC (Lin and Liao, 2016) using default parameters. Both MetaBAT and MyCC resulted in 23 bins ⩾ 0.5 Mbp. Binning_refiner produced 26 bins ⩾ 0.5 Mbp (**Supplementary Figure S1**). Given the availability of reference genome for this mock dataset, precision (defined as how pure a bin is) and recall (defined as how compete a bin is) were calculated with Evaluate.py from the MyCC package (Lin and Liao, 2016). The precision was increased from 87% (MetaBAT) and 91% (MyCC) to 96% (Binning_refiner), while the recall was decreased from 91% (MetaBAT) and 94% (MyCC) to 88% (Binning_refiner).

Binning_refiner was also tested on metagenomic sequence data for three replicate bacterial communities from the surface of the marine alga *Caulerpa filiformis* (Roth-Schulze *et al*., 2016). The paired-end sequence data was generated on an Illumina HiSeq2000 platform and contains a total of 97.94 Gbp. Sequence assembly with idba_ud generated 273 Mbp of contiguous sequences ⩾ 5 Kbp. MetaBAT and MyCC gave 45 and 47 bins ⩾ 0.5 Mbp, respectively, while Binning_refiner produced 69 refined bins ⩾ 0.5 Mbp (**Supplementary Figure S2**). Total length of MetaBAT, MyCC and Binning_refiner bins are 211.3 Mbp, 220.6 Mbp and 197.5 Mbp, respectively, showing only a small loss of overall sequence data through the refinement (**Fig. 1B**). However, the contamination level of refined bins was significantly reduced when compared to the bins from MetaBAT ($p = 0.026$) and MyCC ($p = 0.004$) (**Fig. 1C**). The total length of the refined contamination-free bins is 39.2 Mbp, which is 1.8 times higher than that of the MetaBAT bins (22.0 Mbp) and 3.4 times higher than that of the MyCC bins (11.5 Mbp) (**Fig. 1B**).

Our results demonstrate that Binning_refiner can significantly reduce contamination levels and increase the precisions of genome bins, which resulted in an improved total length of contamination-free bins.

Funding

This research was funded by the Australian Research Council. Wei-Zhi Song was funded by the China Scholarship Council (201508200019).

Conflict of Interest: none declared.

References

Albertsen, M. *et al.* (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. biotechnol*, **31**, 533-538.

Alneberg, J. *et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nat. methods*, **11**, 1144-1146.

Eloe-Fadrosh, E.A. *et al.* (2016) Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.*, **1**, 15032.

Imelfort, M. *et al.* (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, **2**, e603.

Kang, D.D. *et al.* (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.

Lin, H.H. and Liao, Y.C. (2016) Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. rep.*, **6**.

Madden, T. (2013) The BLAST sequence analysis tool.

Parks, D.H. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome res.*, **25**, 1043-1055.

Peng, Y. *et al.* (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420-1428.

Rinke, C. *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431-437.

Roth-Schulze, A.J. *et al.* (2016) Partitioning of functional and taxonomic diversity in surface-associated microbial communities. *Environ. microbiol.*

Sangwan, N. *et al.* (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, **4**, 1.

Seah, B.K. and Gruber-Vodicka, H.R. (2015) gbtools: Interactive Visualization of Metagenome Bins in R. *Front. microbiol.*, **6**.

Singer, E. *et al.* (2016) Next generation sequencing data of a defined microbial mock community. *Sci. Data*, **3**.