MetaABC—an integrated metagenomics platform for data adjustment, binning and clustering

Chien-Hao Su1,2,3,†, Ming-Tsung Hsu1,†, Tse-Yi Wang1, Sufeng Chiang1, Jen-Hao Cheng1, Francis C. Weng3, Cheng-Yan Kao3, Daryi Wang4,5, Huai-Kuang Tsai1,2,∗

1Institute of Information Science, 2Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, 3Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106 and 4Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan

ABSTRACT
Summary: MetaABC is a metagenomics platform that integrates several binning tools coupled with methods for removing artifacts, analyzing unassigned reads and controlling sampling biases. It allows users to arrive at a better interpretation via series of distinct combinations of analysis tools. After execution, MetaABC provides outputs in various visual formats such as tables, pie and bar charts as well as clustering result diagrams.

Availability: MetaABC source code and documentation are available at http://bits2.iis.sinica.edu.tw/MetaABC/
Contact: dywang@gate.sinica.edu.tw; hkttsai@iis.sinica.edu.tw
Supplementary information: Supplementary data are available at Bioinformatics online.

Received on February 21, 2011; revised on May 3, 2011; accepted on June 16, 2011

1 INTRODUCTION
Metagenomics is a field that involves sampling, sequencing and analyzing the genetic material of unculturable microorganisms in microbial communities while maintaining their physiological conditions in habitats (Handelsman et al., 2007). Large amounts of metagenomic datasets have been accumulated for analysis in these years through the rapid advancement of sequencing technology. One of the main focuses in metagenomics is the comparative study of metagenomes based on microbial compositions and diversities that are obtained by binning. Binning assigns sequences to phylogenetic groups according to their taxonomic origins (Simon and Daniel, 2011). Recently, many binning tools, such as MEGAN (Huson et al., 2007), PhymmBL (Brady and Salzberg, 2009), SOrt-ITEMS (Monzourul Haque et al., 2009) and DiScRIBinATE (Ghosh et al., 2010), were developed to achieve higher binning accuracy. As the above binning methods become widely used, some data adjustment methods have lately been proposed to improve metagenomic data analysis. For example, it is known that 454 pyrosequencing produces artificial duplicated reads (Gomez-Alvarez et al., 2009), and Niu et al. (2010) indicated that removing these duplicates reduced 5–23% of artifacts in 10 metagenomic datasets. Further, when analyzing Sanger sequencing data, Weng et al. (2010) suggested that reanalyzing unassigned reads utilizing conserved neighboring gene adjacency can improve taxonomic assignment. Moreover, latest studies (Angly et al., 2009; Breslauer et al., 2010) showed that genome length normalization could help in reducing sampling biases in estimating taxon and gene abundances. However, current binning tools do not incorporate these data adjustment methods while assigning reads to their respective taxa and producing abundance profiles. Hence, it is essential to integrate these adjustments and develop a more comprehensive binning platform.

The aim of this work is to develop a single platform, MetaABC, that integrates several binning methods, coupled with data filters and normalization techniques for improving the taxonomic assignment in metagenomic analysis. In addition, MetaABC presents a user-friendly interface, provides outputs in several visualizations which are downloadable in printable figure-ready formats and implements a hierarchical clustering program for comparative analysis of metagenomes. Further, MetaABC is capable of handling data produced by both Sanger and next-generation sequencing.

MetaABC also provides a stand-alone version of the software to deal with large datasets.

2 METHODS
MetaABC is an integrated metagenomics platform for data adjustment, binning and clustering (Fig. 1). MetaABC incorporates two means for removing artifacts, five tools for taxonomic binning, an approach to re-analyze unassigned reads using conserved gene adjacency and an option to control sampling biases via genome length normalization. Also, MetaABC includes a hierarchical clustering program for metagenomic comparative analysis (see Supplementary Material for more details of these components).

3 USAGE OF METAABC
3.1 Input
MetaABC accepts data in two different forms: (i) sequences in SFF, FASTQ and PASTA formats; (ii) a three-column, tab-delimited abundance table. The sequence file is used for taxonomic binning, while the abundance table is for clustering. (More details of the inputs are provided in Supplementary Material).
3.2 Data presentation and visualization
Upon uploading the sequence file, the user must select a series of tools and execute them. Then, MetaABC provides the total and assigned sequence number of each file. After executing all the selected tools, the final results of MetaABC output different formats, including tables, pie charts and bar charts of abundance profiles (see Supplementary Material for two case studies in details).

3.3 Statistical analysis of run time
MetaABC integrates different steps for better estimation of the taxonomic assignment, including duplicates removal, taxonomic binning, reanalysis of unassigned reads and sampling biases control. In two case studies, we found that the most time-consuming step of MetaABC is either the taxonomic binning or the reanalysis of unassigned reads (see Supplementary Material for details).

ACKNOWLEDGEMENTS
We wish to thank Ting-Wei Hsu for helping the construction of clustering figures and Krishna B.S. Swamy for his valuable suggestions and comments.

Conflict of Interest: none declared.

REFERENCES