BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 16 2011, pages 2298–2299
doi:10.1093/bioinformatics/btr376

Genome analysis

Advance Access publication June 22, 2011

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

Chien-Hao Su[1,2,3,†], Ming-Tsung Hsu[1,†], Tse-Yi Wang[1], Sufeng Chiang[1], Jen-Hao Cheng[1], Francis C. Weng[4], Cheng-Yan Kao[3], Daryi Wang[4,*], Huai-Kuang Tsai[1,2,*]

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

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: MetaABC is a metagenomic 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; hktsai@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 (Monzoorul 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; Beszteri *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 FASTA 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).

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

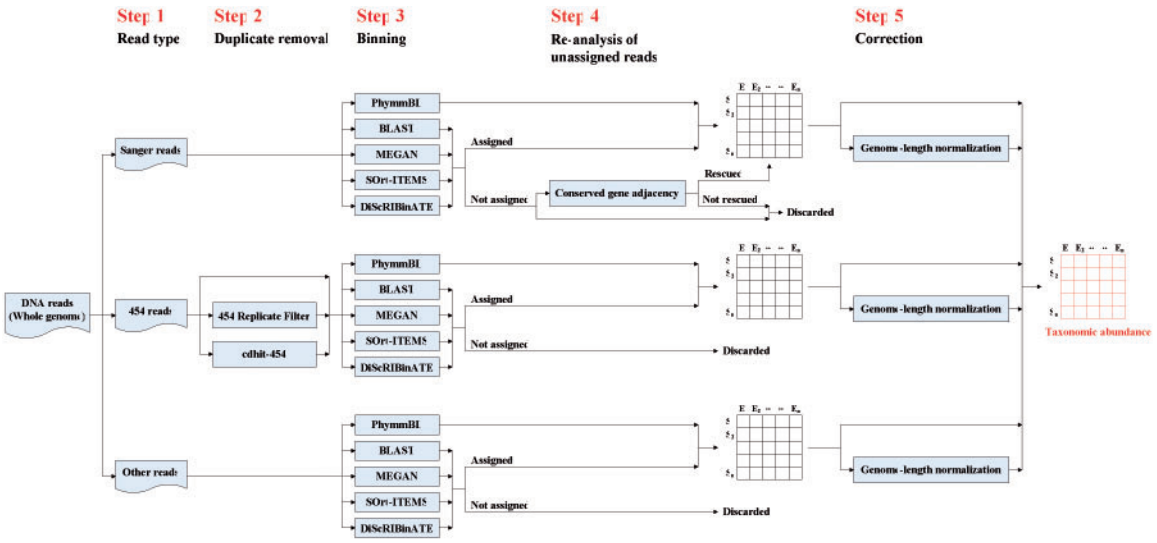

Fig. 1. Flowchart of MetaABC.

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.

Funding: Institute of Information Science; Academia Sinica; National Science Council of Taiwan under grants (NSC 99-2627-B-001-005-MY3 to D.W. and NSC99-2627-B-001-003 to H.-K.T.).

Conflict of Interest: none declared.

REFERENCES

Angly,F.E. *et al.* (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol.*, **5**, e1000593.

Beszteri,B. *et al.* (2010) Average genome size: a potential source of bias in comparative metagenomics. *ISME J.*, **4**, 1075–1077.

Brady,A. and Salzberg,S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods*, **6**, 673–676.

Ghosh,T.S. *et al.* (2010) DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics*, **11** (Suppl. 7), S14.

Gomez-Alvarez,V. *et al.* (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J.*, **3**, 1314–1317.

Handelsman,J. *et al.* (2007) *The New Science of Metagenomics: Revealing the Secrets of our Microbial Planet*. The National Academies Press, Washington, DC.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

McHardy,A.C. *et al.* (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.

Monzoorul Haque,M. *et al.* (2009) SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, **25**, 1722–1730.

Niu,B. *et al.* (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, **11**, 187.

Simon,C. and Daniel,R. (2011) Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.*, **77**, 1153–1161.

Weng,F.C. *et al.* (2010) Reanalyze unassigned reads in Sanger based metagenomic data using conserved gene adjacency. *BMC Bioinformatics*, **11**, 565.