BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 23 2010, pages 2977–2978
doi:10.1093/bioinformatics/btq536

Genome analysis

Advance Access publication October 19, 2010

SmashCommunity: a metagenomic annotation and analysis tool

Manimozhiyan Arumugam[1], Eoghan D. Harrington[2,3], Konrad U. Foerstner[1], Jeroen Raes[4] and Peer Bork[1,5,*]

[1]EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [2]Department of Microbiology & Immunology, [3]Department of Medicine, Stanford University, Stanford, CA 94305, USA, [4]VIB – Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium and [5]Max-Delbruck-Centre for Molecular Medicine, Berlin-Buch, Germany

Associate Editor: John Quackenbush

ABSTRACT

Summary: SmashCommunity is a stand-alone metagenomic annotation and analysis pipeline suitable for data from Sanger and 454 sequencing technologies. It supports state-of-the-art software for essential metagenomic tasks such as assembly and gene prediction. It provides tools to estimate the quantitative phylogenetic and functional compositions of metagenomes, to compare compositions of multiple metagenomes and to produce intuitive visual representations of such analyses.

Availability: SmashCommunity source code and documentation are available at http://www.bork.embl.de/software/smash

Contact: bork@embl.de

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

Received on May 13, 2010; revised on July 27, 2010; accepted on September 17, 2010

1 INTRODUCTION

Metagenomics allows the culture-free characterization of natural and host-associated microbial communities and provides an understanding of their structure, dynamics and functionality as well as the environmental factors that shape them (Handelsman *et al*., 2007). Although the volume of metagenomic data being deposited to public repositories is increasing exponentially (Singh *et al*., 2009) and more large-scale studies are underway (Peterson *et al*., 2009), there are still no standards for experimental and computational methods required to analyze such datasets (Raes *et al*., 2007), making it hard to compare results from these studies. Web-servers such as CAMERA (Seshadri *et al*., 2007), IMG/M (Markowitz *et al*., 2008) and MG-RAST (Meyer *et al*., 2008) host published metagenomic datasets and enable users to perform additional and comparative analysis on them. However, there is an imminent need for stand-alone computational pipelines that enable in-house analysis of new metagenomic datasets using standardized methods and comparison of datasets from different environments. MG-RAST and the MEGAN stand-alone tool (Mitra *et al*., 2009) perform functional and phylogenetic analyses of metagenomes. However, they do not estimate quantitative abundances as they simply count the reads mapping to known genes or species—a measure strongly affected by gene length and genome size. They also do not assemble metagenomic reads and are thus unable to identify operons and multidomain genes in low or medium complexity metagenomes. Finally, these tools do not have a modular, open source structure allowing users to plug in alternative tools for the various steps in metagenome analysis, nor do they provide the advantages of a locally installed, queryable database containing all raw analysis results (see Supplementary Table 2 for detailed comparison of metagenomic analysis tools). To address these issues we have developed SmashCommunity (Simple Metagenomics Analysis SHell for microbial communities) to annotate shotgun metagenomes with inbuilt tools for quantitative and comparative analyses.

2 DESIGN AND IMPLEMENTATION

SmashCommunity shares design principles and routines with SmashCell. (Harrington *et al*., 2010), a complementary framework for the analysis of high-throughput single cell-amplified microbial genomes. It is written in Perl with modular architecture, well-defined inter-modular interfaces and a locally installed database (Supplementary Figs S1 and S2). Each task in metagenomic analysis, such as sequence assembly or gene prediction, is implemented as a module that is a wrapper around a software program that implements this task. This design of independent modules with common interfaces supports multiple choices for each task and enables replacement of programs with better alternatives when available. SmashCommunity comes with built-in support for current state-of-the-art programs that are publicly available (see Supplementary Notes) and additional programs can easily be incorporated.

3 FEATURES

SmashCommunity can analyze sequences generated by Sanger and 454 sequencing technologies. It provides optimized parameter sets for Arachne (Jaffe *et al*., 2003) and Celera (Myers *et al*., 2000) for metagenome assembly, and GeneMark (Besemer and Borodovsky, 1999) and MetaGene (Noguchi *et al*., 2008) for predicting protein coding genes on metagenomes (see Supplementary Notes). Assembly or gene prediction performed outside of Smash-Community can also be loaded into the repository using ACE and GFF files.

SmashCommunity includes scripts for downstream analysis of datasets. They can generate intuitive tree-based visualizations of the results using the batch access API of the interactive Tree of Life (iTOL) web-tool (Letunic and Bork, 2007). For example, samples can be phylogenetically characterized (i) using best BLAST hits

*To whom correspondence should be addressed.

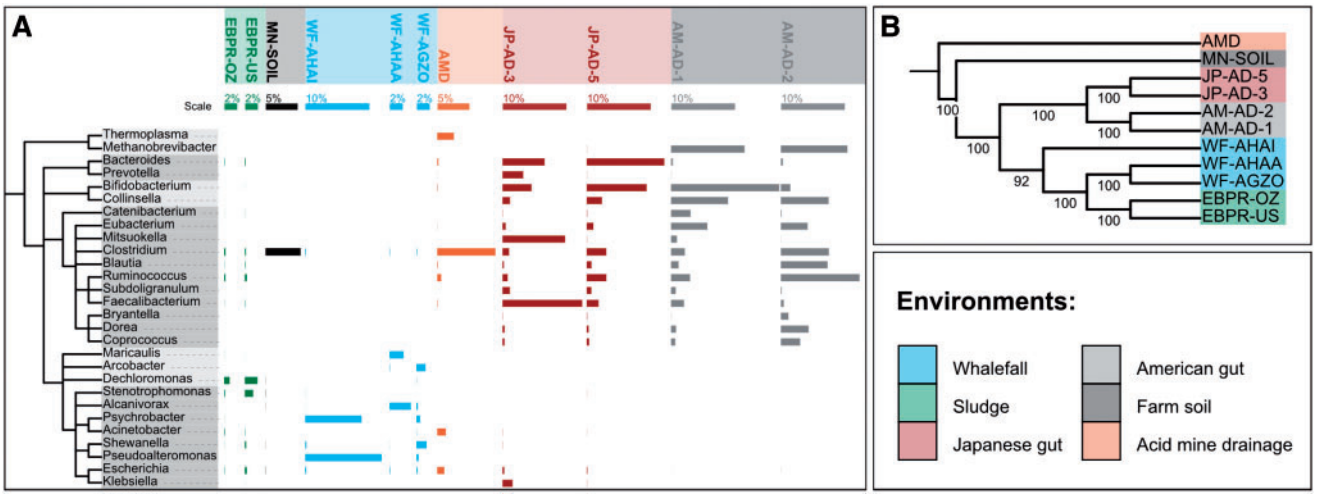

Fig. 1. Visual representation of metagenomic analysis using SmashCommunity. (**A**) Phylogenetic (genus) profiles of 11 different samples (see Supplementary Table 1) using reference genome mapping of reads. Different phyla appear in different environments and similar environments sometimes have different dominant genera. (**B**) Hierarchical clustering of samples using phylogenetic profiles reveals clustering based on habitats.

to microbial reference genomes above taxonomic rank specific sequence similarity thresholds, or (ii) by identifying reads containing 16S rRNA sequences (Huang *et al*., 2009) and classifying them (Wang *et al*., 2007). Quantitative phylogenetic profiles (relative abundances) are then calculated more accurately by counting the reads and correcting for genome size or 16S rRNA gene copy number variation. These profiles could be uploaded to the iTOL website, browsed online, downloaded and automatically post-processed to generate useful visual representations (Fig. 1A). Protein-coding genes can be annotated using BLAST-based homology to orthologous groups from eggNOG (Muller *et al*., 2010) and KEGG pathway (Kanehisa *et al*., 2008) databases and functional profiles are estimated using read abundance after normalizing for gene length. SmashCommunity can also compare multiple metagenomes using these profiles, cluster them based on a relative entropy-based distance measure suitable for comparing such quantitative profiles, perform bootstrap analysis of the clustering and generate visual representation of the clustering results (Fig. 1B). Several of the analysis tasks in SmashCommunity can be performed on data from SmashCell and vice versa. Documentation for the full set of features is available on SmashCommunity website.

ACKNOWLEDGEMENTS

We thank Daniel Mende, Aino Jaervelin and Youssef Darzi for testing the pipeline, Ivica Letunic for extending iTOL functionalities, Alison Waller and Julien Tap for comments and suggestions on the article and members of the Bork group for helpful discussions.

Funding: European Community's MetaHIT, grant agreement HEALTH-F4-2007-201052. FWO Oddyseus excellence programme (J.R.).

Conflict of Interest: none declared.

REFERENCES

Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, **27**, 3911–3920.

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

Harrington,E.D. *et al.* (2010) SmashCell: a software framework for the analysis of single-cell amplified genome sequences. *Bioinformatics*, **26**, 2979–2980.

Huang,Y. *et al.* (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics*, **25**, 1338–1340.

Jaffe,D.B. *et al.* (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.*, **13**, 91–96.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

Markowitz,V.M. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.*, **36**, D534–D538.

Meyer,F. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Mitra,S. *et al*. (2009) Visual and statistical comparison of metagenomes. *Bioinformatics*, **25**, 1849–1855.

Muller,J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.

Myers,E.W. *et al.* (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.

Noguchi,H. *et al.* (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.*, **15**, 387–396.

Peterson,J. *et al.* (2009) The NIH Human Microbiome Project. *Genome Res.*, **19**, 2317–2323.

Raes,J. *et al.* (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr. Opin. Microbiol.*, **10**, 490–498.

Seshadri,R. *et al.* (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.*, **5**, e75.

Singh,A.H. *et al.* (2009) Discovering functional novelty in metagenomes: examples from light-mediated processes. *J. Bacteriol.*, **191**, 32–41.

Wang,Q. *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.