SmashCommunity: a metagenomic annotation and analysis tool

Manimozhiyan Arumugam¹, Eoghan D. Harrington²,³, Konrad U. Foerstner¹, Jeroen Raes⁴ and Peer Bork¹,⁵,*

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

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., 2004). 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 (Patterson 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 (Seshadrhi 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 SIHell for microbial communities) to annotate shotgun metagenomes with built-in 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 SmashCommunity 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 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 Youssel 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 Oddysseus excellence programme (J.R.).

Conflict of Interest: none declared.

REFERENCES


