Bioinformatics, 35(13), 2019, 2318–2319
doi: 10.1093/bioinformatics/bty963
Advance Access Publication Date: 26 November 2018
Applications Note

Sequence analysis

GenCoF: a graphical user interface to rapidly remove human genome contaminants from metagenomic datasets

Matthew D. Czajkowski[1], Daniel P. Vance[1], Steven A. Frese[1,2] and Giorgio Casaburi[1,*]

[1]Evolve BioSystems, Inc., Davis, CA 95618, USA and [2]Department of Food Science and Technology, University of Nebraska, Lincoln, NE 68583, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on August 20, 2018; revised on November 7, 2018; editorial decision on November 21, 2018; accepted on November 23, 2018

Abstract

Summary: The removal of human genomic reads from shotgun metagenomic sequencing is a critical step in protecting subject privacy. Freely available tools addressing this issue require advanced programing knowledge or are limited by analytical time and data load due to their server-based nature. Here, we compared the most cited tools for host-DNA removal using synthetic and real metagenomic datasets. Then, we integrated the most efficient pipeline in a graphical user interface to make these tools available without command line use. This interface, GenCoF, rapidly removes human genome contaminants from metagenomic datasets. Additionally, the tool offers quality-filtering, data reduction and interactive modification of any parameter in order to customize the analysis. GenCoF offers both quality and host-associated filtering in a non-commercial, freely available tool in a local, interactive and easy-to-use interface.

Availability and implementation: GenCoF is freely available (under a GPL license) for Mac OS and Linux at https://github.com/MattCzajkowski/GenCoF.

Contact: gcasaburi@evolvebiosystems.com

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

1 Introduction

Improvements in sequencing technologies within the past decade have reduced the cost of shotgun metagenomic sequencing tremendously (Kunin *et al.*, 2008). Consequently, greater access and lower cost has resulted in rapid growth of metagenomics datasets from humans and other systems (Pagani *et al.*, 2011). After sequencing, the first step of analysis is the removal of host-derived reads (Kunin *et al.*, 2008). Residual human DNA in publicly deposited data sets is a major privacy concern and a roadblock to proper study privacy protections (Gurwitz *et al.*, 2009).

Although several tools are currently available to perform this task, the majorities require advanced programing knowledge or have limitations in terms of analysis time or computational requirements. In this study, we compared the most cited tools for host-removal filtering in metagenomics studies. Then, selected tools were integrated into the most efficient pipeline in a graphical user interface (GUI) for broad use. We present Genomic Contamination Filter (GenCoF), a GUI to rapidly remove human genome contaminants from metagenomic datasets. The application is available as an executable in Unix and Mac OS environments and does not require command line interaction. In addition to host-derived contaminant removal, GenCoF offers the possibility to quality-filter sequencing reads and split datasets into smaller sizes to increase manageability of data.

GenCoF allows for parameter customization of the analysis based on user needs. Further, a step-by-step tutorial of installation and sequence filtering is available. GenCoF is the first tool offering an interactive and easy-to-use interface to filter metagenomic sequencing reads for both quality and host-associated components.

The application is freely available under GPL license at https://github.com/MattCzajkowski/GenCoF.

2 Features and methods

The software included in GenCoF has been coded with Python v3.6.1 and implements freely available packages (see Supplementary Methods).

2.1 Tool description

GenCoF has bundled several programs, including Sickle (Joshi and Fass, 2018), Prinseq (Schmieder and Edwards, 2011a), Bowtie2 (Langmead and Salzberg, 2012) and Fastq and Fasta Splitter from the FASTX-Toolkit. Before read decontamination, samples can be quality-filtered. Read trimming in GenCoF uses either Sickle or Prinseq. Users can decide whether to use multi-threading options and build custom reference databases for sequences removal. Bowtie2 is then employed to perform the decontamination step with optional custom parameterization. Lastly, GenCoF offers the option to concatenate output files if they were initially split.

2.2 Program performance methods

The programs Deconseq v0.4.3 (Schmieder and Edwards, 2011b), Bowtie2 v2.3.4, BBMap (specifically BBSplit) v37.80 (Bushnell, 2014) and BMTagger v1.1.0 (Rotmistrovsky and Agarwala, 2011) were compared for their speed, accuracy and size of reference files created (see Supplementary Results). Four synthetic datasets were created containing an average of 96 166 185 reads from viral, fungal, bacterial, archaeal and human genomes. Datasets had ∼30, 50, 70 or 100% of human reads, respectively. All tools were compared against the human genome (vGRCh38.p7, NCBI accession GCF_000001405.33) as reference. From the synthetic dataset test, the highest performing parameters were chosen for the individual tools, which were finally tested against a published metagenomic dataset (Casaburi *et al.*, 2018). BLASTn (Altschul *et al.*, 1990). was used as baseline using an *E*-value cut-off of $\leq 10^{-10}$ to determine whether the reads reported as positive hits by the programs were true positives (Haas *et al.*, 2011; Turnbaugh *et al.*, 2009). Only reads with BLASTn-positive hits were considered true human contaminants.

3 Results

Synthetic analysis showed that BMTagger had the best CPU/hour (Fig. 1B). However, BMTagger is limited by a single-CPU usage and needed the largest reference (∼32 GB). Conversely, Bowtie2 presented similar overall error rates (Fig. 1A), could run with multiple CPUs, and only required creation of a ∼3 GB reference file. In comparison to the other programs, BBSplit returned a higher rate of incorrectly mapped human reads (Fig. 1A), but outperformed all the other applications in correctly assigning microbial reads (Fig. 1C). Further, while BBSplit only generated a 3 GB reference file, it required java runtime environment to run and returned a high CPU/hour across all tested parameters (Fig. 1B). Finally, Deconseq returned very high error rate (Fig. 1C) and CPU/hour (Fig. 1B), and was limited to a single CPU.

However, Deconseq only used a 4 GB reference database and only required Mac or Linux OS. From the published metagenomic dataset, Bowtie2 reported the lowest error rate, followed by Deconseq, BBSplit and BMTagger, respectively (Fig. 1D).

Bowtie2 outperformed the other tools in terms of accuracy, while BBSplit had the worst accuracy in correctly assigning reads. Overall, Bowtie2 was chosen for GenCoF because of its low error rate, limited pre-requisites and low CPU/hour. It also performed comparably on both real and synthetic datasets. Although it mapped the least number of reads from the published reads, it only differed by ∼3000 reads of the 1.5 million tested (0.2%).

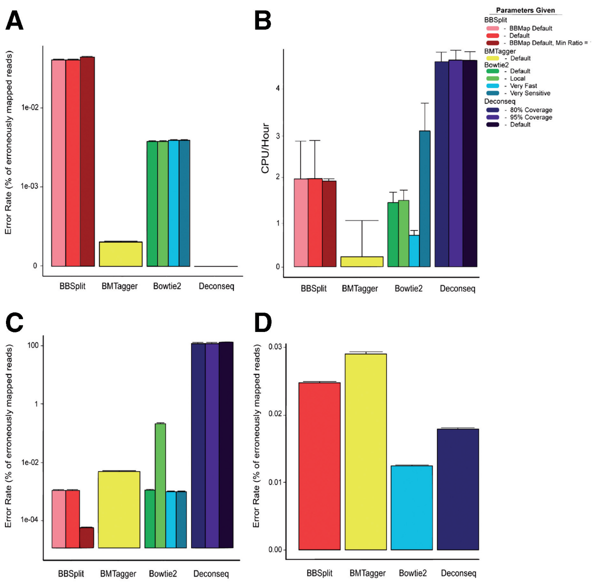

Fig. 1. Analysis of program performance. (**A**) Average error rate of synthetic reads wrongly assigned as non-human. (**B**) Average CPU/Hour. (**C**) Average error rate of synthetic reads wrongly assigned as human. (**D**) Average error rate of real dataset. Error bars represent standard error

Funding

This work was supported by Evolve Biosystems.

Conflict of Interest: Authors are employees of Evolve Biosystems. The software presented here is completely free and available under GPL v3 license.

References

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Bushnell,B. (2014) Bbmap: a fast, accurate, splice-aware aligner.

Casaburi,G. *et al*. (2018) Colonization of breastfed infants by Bifidobacterium longum subsp. Infanti EVC001 reduces virulence gene abundance. *Hum. Microbiome J*., **9**, 7–10.

Gurwitz,D. *et al*. (2009) Children and population biobanks. *Science*, **325**, 818–819.

Haas,B.J. *et al*. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*., **21**, 494–504.

Joshi,N. and Fass,J. (2015) Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files.

Kunin,V. *et al*. (2008) A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev*., **72**, 557–578.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie2. *Nat. Methods*, **9**, 357–359.

Pagani,J. *et al*. (2011) The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*., **40**, 571–579.

Rotmistrovsky,K. and Agarwala,R. (2011) *BMTagger: Best Match Tagger for Removing Human Reads from Metagenomics Datasets*. ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/.

Turnbaugh,P.J. *et al*. (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480.

Schmieder,R. and Edwards,R. (2011a) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schmieder,R. and Edwards,R. (2011b) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*, **6**, e17288.