HoCoRT: Host contamination removal tool

Ignas Rumbavicius[1], Trine B. Rounge[1,2,3,*], Torbjørn Rognes[1,4,*]

[1] Centre for Bioinformatics, Department of Informatics, University of Oslo, PO Box 1080 Blindern, 0316 Oslo, Norway
[2] Centre for Bioinformatics, Department of Pharmacy, University of Oslo, PO Box 1068 Blindern, 0316 Oslo, Norway
[3] Cancer Registry of Norway, PO Box 5313 Majorstuen, 0304 Oslo, Norway
[4] Department of Microbiology, Oslo University Hospital, PO Box 4950 Nydalen, 0424 Oslo, Norway
* Corresponding authors

Abstract

Motivation: Microbiome sequencing data obtained from a host environment will usually be contaminated with sequences from the host organism. Host sequences should be removed before further analysis to avoid biases, to reduce downstream computational load, or for privacy reasons in case of a human host. The available tools designed specifically to perform host contamination sequence removal are either outdated, not maintained or complicated to use. We have therefore created HoCoRT, a fast and user friendly tool that implements several methods for optimised host sequence removal.

Results: HoCoRT is an open-source command-line tool for host contamination removal. It is designed to be simple to install and use, including a one-step option for genome indexing. HoCoRT uses well-known mapping, classification and alignment methods for the classification. The underlying classification method and its parameters are selectable by the user, but default and recommended choices are provided for data sets with short and long reads. To decontaminate an artificial human gut microbiome with HoCoRT, the best combination of speed and accuracy was found with Bowtie2 in end-to-end mode for Illumina reads, while Minimap2 performed best for Nanopore reads. HoCoRT was found to be much faster and slightly more accurate than DeconSeq.

Availability: HoCoRT is distributed as a Bioconda package. Source code and documentation are available at https://github.com/ignasrum/hocort for Linux and macOS under the MIT licence.

Contacts: t.b.rounge@farmasi.uio.no and torognes@ifi.uio.no

Supplementary material: Supplementary material is available in a separate document.

Introduction

Sequencing the genomes of microbial communities in an environment of a host organism have opened new avenues for research on host - microbe interactions. A number of analysis steps must be carried out after metagenomic sequencing to obtain a good picture of the composition of the microbiome. The huge amounts of data often involved require efficient processing, and it is therefore important to avoid unnecessary computations. Some sequences of host origin are inevitably included in the obtained data, and when the host is human, privacy becomes an issue. Sequences of non-microbial origin could also bias the results of downstream analyses. These host sequences should hence be removed at an early stage (Barthi & Grimm, 2021).

We suspect that decontamination is handled in an ad-hoc manner in many studies by running an alignment tool against a host genome database. This makes this essential analysis step more complicated than necessary and could result in inferior performance. There is a lack of best practices for this procedure, and comparison across studies is more difficult without standardisation.

We could identify only two dedicated tools specifically designed for removal of contaminating sequences, namely DeconSeq and GenCoF. DeconSeq (Schmieder & Edwards, 2011) is a command-line tool for identification and removal of sequence contamination from genomic and metagenomic datasets. DeconSeq integrates its underlying classifier, a modified version of BWA-SW (Li & Durbin, 2009), directly into its source code, making modifications difficult. DeconSeq only supports single-end Illumina reads, and the code has not been updated since 2013. GenCoF (Czajkowski et al, 2019) is a graphical user interface to rapidly remove human genome contaminants from metagenomic datasets. It is limited to short reads. The GUI makes it unsuited for scripting. Extending GenCoF is also difficult as it also has its underlying classifier Bowtie2 (Langmead & Salzberg, 2012) directly integrated into its source code. Both tools clearly have severe limitations making them less suitable for most large-scale modern datasets. We therefore decided to develop the new tool HoCoRT to fill this need.

Implementation

HoCoRT is an open-source command-line based tool written in Python 3. It is designed to be easy to use and can be easily installed as a package with Bioconda (Grüning et al., 2018), or using the accompanying Docker container. HoCoRT has a modular pipeline design and uses well established classification, mapping and alignment tools to perform the classification of the sequences into host and microbial sequences. Currently available pipeline modules include BBMap (https://sourceforge.net/projects/bbmap/), Bowtie2 (Langmead & Salzberg, 2012), BWA-MEM2 (Vasimuddin et al, 2019), HISAT2 (Kim et al., 2019), Kraken2 (Wood et al., 2019), and Minimap2 (Li, 2018). In addition, HoCoRT depends on samtools (Li et al., 2009). The configuration of the options to the pipelines can be specified by the user, but recommended defaults are provided. New modules can be written to extend HoCoRT to use other tools. Modules may pipe data through different tools after each other. HoCoRT also provides an extensive Python library with an API that could be used as a backend by other tools. HoCoRT reads and writes optionally compressed FASTQ files. The building of

database index files for the underlying classification tools is also handled by HoCoRT. The tool is documented with built-in help functions and error messages.

Results

The classification speed and accuracy of HoCoRT with several different underlying methods and settings was investigated with synthetic datasets. They were created with HiSeq, MiSeq and Nanopore reads containing a mix of 1% human host sequences and 99% microbial sequences, similar to the human gut microbiome (figure 1 with additional details in supplementary table S1) and 50% human host and 50% microbial sequences, similar to oral microbiomes (supplementary figure S1 and table S2). Overall, Bowtie2 in end-to-end mode is recommended for Illumina reads (both HiSeq and MiSeq) as it was both fast and accurate in all cases. Minimap2 is recommended for Nanopore reads due to its combination of high accuracy and speed across all Nanopore tests, but Kraken2 also works well, especially when the amount of host contamination is high, as seen in supplementary figure S1.

The performance of HoCoRTwas also compared to DeconSeq using similar datasets, but with single-ended HiSeq and MiSeq reads. The HoCoRT Bowtie2 pipeline was much faster than DeconSeq at aligning both HiSeq (34X) and MiSeq (49X) reads, and also slightly more accurate (Supplementary table S3).

Additional results and a more detailed description of HoCoRT can be found in the master thesis of Rumbavicius (2022).

Conclusion

A dedicated, flexible, extendable, and modular tool for host sequence contamination removal is now freely available for all users. It should simplify the decontamination step in microbiome data analysis and provide solid performance.

References

Bharti, R., & Grimm, D. G. (2021) Current challenges and best-practice protocols for microbiome analysis. Briefings in Bioinformatics, 22(1), 178–193. https://doi.org/10.1093/bib/bbz155

Czajkowski, M. D., Vance, D. P., Frese, S. A., & Casaburi, G. (2019) GenCoF: A graphical user interface to rapidly remove human genome contaminants from metagenomic datasets. Bioinformatics, 35(13), 2318–2319. https://doi.org/10.1093/bioinformatics/bty963

Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., & Bongcam-Rudloff, E. (2019) Simulating illumina metagenomic data with InSilicoSeq. Bioinformatics, 35(3), 521–522. https://doi.org/10.1093/bioinformatics/bty630

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J. & The Bioconda Team (2018) Bioconda: Sustainable and comprehensive software distribution for the life sciences. Nature Methods, 15(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Koster, J., & Rahmann, S. (2012) Snakemake–a scalable bioinformatics workflow engine. Bioinformatics, 28(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Langmead, B., & Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Rumbavicius, I. (2022) Tool to remove specific organisms from microbiome sequencing data - Host Contamination Removal Tool (HoCoRT). Master thesis, Department of Informatics, University of Oslo, Norway. http://urn.nb.no/URN:NBN:no-98212

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S.T., Karsch-Mizrachi, I. (2022) GenBank. Nucleic Acids Res, 50(D1), D161-D164. https://doi.org/10.1093/nar/gkab1135

Schmieder, R., & Edwards, R. (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLOS ONE, 6(3), e17288. https://doi.org/10.1371/journal.pone.0017288

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019) Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Wood, D. E., Lu, J., & Langmead, B. (2019) Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Yang, C., Chu, J., Warren, R. L., & Birol, I. (2017) NanoSim: Nanopore sequence read simulator based on statistical characterization. GigaScience, 6(4). https://doi.org/10.1093/gigascience/gix010

Figures

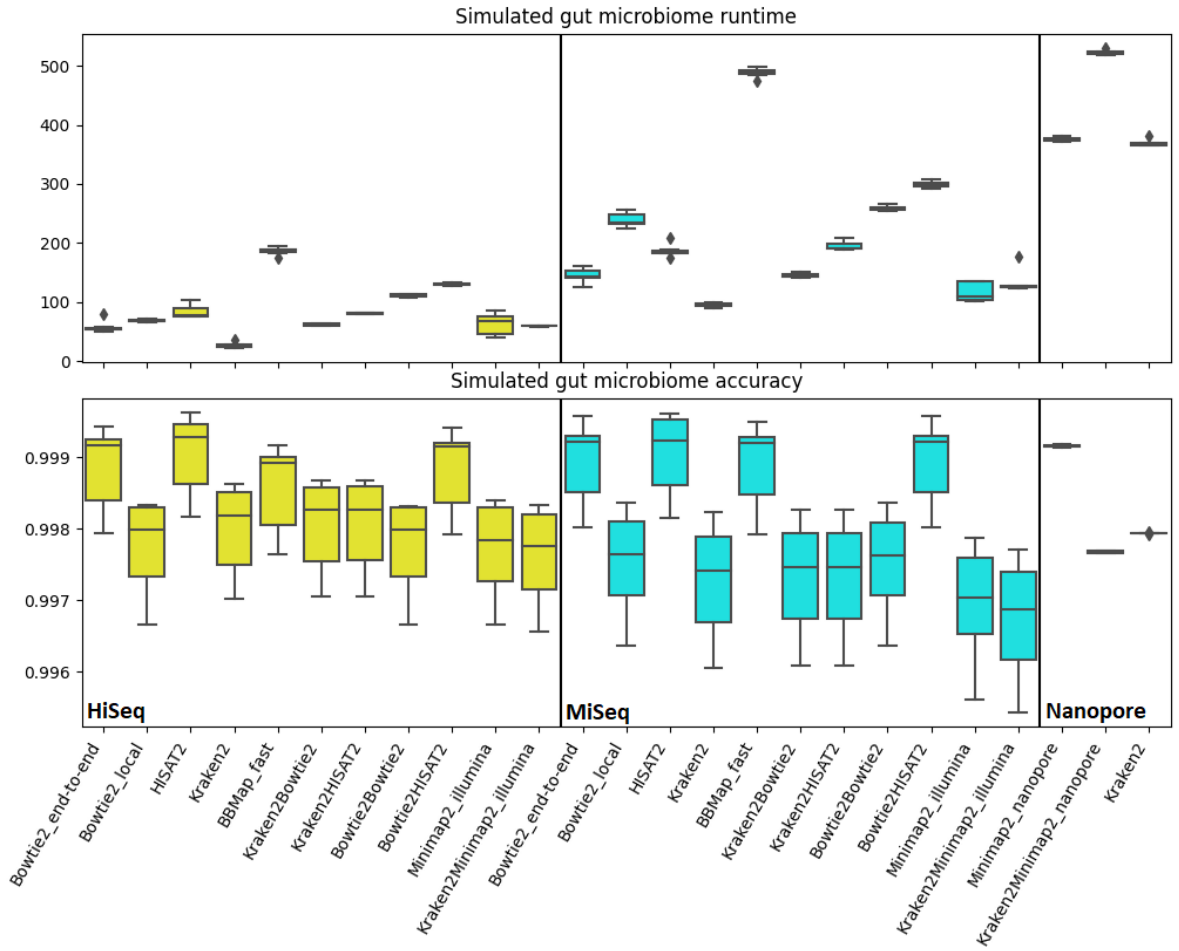

Fig. 1. HoCoRT performance on simulated gut microbiome datasets. Boxplots of HoCoRT runtime in seconds (top) and classification accuracy (bottom) using several different classification modules and parameters on HiSeq (yellow, left), MiSeq (cyan, middle) and Nanopore data (red, right). The datasets contain 1% human reads and 99% microbial reads. The human reads are generated from the Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13), whereas the microbial reads are generated from a mix of bacterial, fungal and viral sequences pseudo-randomly extracted from NCBI GenBank (Sayers et al., 2022). Each boxplot represents 7 different datasets, each consisting of 5 million reads randomly generated using InSilicoSeq (Gourlé et al., 2019) (HiSeq 125bp and MiSeq 300bp, paired-end) and NanoSim (Yang et al., 2017) (Nanopore, average 2159bp, range 54-98320bp, single-end). For Illumina data the following 15 pipelines were examined: Bowtie2 version 2.4.5 in end-to-end and local mode, both with and without the “un_conc” option, HISAT2 version 2.2.1, Kraken2 version 2.1.2, BBMap version 38.96 in default and fast mode, BWA-MEM2 version 2.2.1, Kraken2 followed by Bowtie2 in end-to-end mode, Kraken2 followed by HISAT2, Bowtie2 in end-to-end mode followed by local mode, Bowtie2

in end-to-end mode followed by HISAT2, Minimap2 version 2.24, and Kraken2 followed by Minimap2. For Nanopore data the following 3 pipelines were examined: Minimap2, Kraken2 followed by Minimap2, and Kraken2. When the “un_conc” option is given, Bowtie2 requires both reads in a pair to map concordantly to the genome. Performance analysis was carried out using a Snakemake pipeline (Koster & Rahmann, 2012) on a laptop PC with a AMD Ryzen 7 1700X 8 core 3.4 GHz CPU, 64GB RAM and 4TB HDD running Linux. No quality filtering or other pre-processing was performed. Supplementary table S1 shows additional results, including those for Bowtie2 with the “un_conc” option, BBMap in default mode and BWA-MEM2, which were excluded from this boxplot due to outliers.