Hostile: accurate host decontamination of microbial sequences

Bede Constantinides [1,2]* and Derrick W Crook [1,2,3]

[1]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[2]The National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, UK
[3]The National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford, Oxford, UK

*To whom correspondence should be addressed.

Abstract

Summary: Microbial sequences generated from clinical samples are often contaminated with human host sequences that must be removed for ethical and legal reasons. Care must be taken to excise host sequences without inadvertently removing target microbial sequences to the detriment of downstream analyses such as variant calling and *de novo* assembly. To facilitate accurate host decontamination of both short and long sequencing reads, we developed Hostile, a tool capable of rapid host read removal using laptop specification hardware. We demonstrate that our approach removes at least 99.868% of real human reads and retains at least 99.997% of simulated bacterial reads. Use of a masked reference genome further increases bacterial read retention (>=99.997%) with negligible (<0.001%) reduction in human read removal performance. Compared with an existing tool, Hostile removed up to 11x more human reads and up to 11x fewer microbial reads while taking less time for typical workloads.

Availability and implementation: Hostile is implemented as an MIT licensed Python package available from https://github.com/bede/hostile

1 Introduction

Microbial specimens are often contaminated with host sequences. Since experimental host genome depletion protocols are imperfect, host DNA often reaches the sequencing instrument. Where the specimen host is a human, it is important that host sequences are subsequently deleted in order to protect anonymity. The widespread human contamination of publicly deposited microbial sequence data (Bush *et al.*, 2020) is therefore regrettable and raises regulatory concerns, particularly in light of the rapid growth of metagenomic diagnostics. Furthermore, unwanted host sequences waste computing resources and may adversely affect downstream analyses such as variant calling and *de novo* assembly. Host decontamination is therefore the first step performed in many microbial genomic analyses. Existing approaches employ one of three strategies: *i)* exclusive retention of reads aligning to a target microbial genome (Hunt *et al.*, 2022), *ii)* subtractive removal of reads aligning to a host genome, and *iii)* subtractive removal after metagenomic read classification (Kim *et al.*, 2016; Wood *et al.*, 2019). Where the target microbe is known *a priori*, the first strategy (exclusive retention) may be most suitable: for SARS-CoV-2 it is both more accurate and computationally efficient than subtractive removal (Hunt *et al.*, 2022). However, the second and third strategies (subtractive removal) are generalisable, and thus necessary for analysis of microbes that are unknown *a priori*, mixtures, or novel.

In this article, we describe a simple tool implementing subtractive removal of contaminant human genome sequences, together with rigorous evaluation of its performance against real human genomes and simulated bacterial reads representing the 985 complete bacterial assemblies in the FDA-ARGOS dataset (Sichtig *et al.*, 2019). We also report performance using simulated reads for 140 complete mycobacterial genomes. These results provide evidence of the accuracy of the approach in terms of both its ability to remove human host reads (sensitivity), and to retain microbial reads (specificity).

2 Materials and Methods

Hostile is implemented as a Python package providing a command line interface and Python API. The decontamination process involves a series of streaming operations on optionally gzip-compressed input FASTQ files: *i)* alignment to a human genome (Minimap2 or Bowtie2), *ii)* counting distinct reads (Samtools), *iii)* discarding mapped reads (and their mate reads if applicable; Samtools), *iv)* counting remaining reads (Samtools), *v)* Replacing read names with incrementing integers, and *vi)* writing gzip-compressed FASTQ files (Samtools) (Li, 2018; Langmead and Salzberg, 2012; Danecek *et al.*, 2021). These operations are streaming to reduce execution time and disk IO. Bowtie2 is the default aligner for short reads due to its relatively compact (<4GB) memory footprint, while Minimap2 is the default aligner for long reads, requiring approximately 12GB of RAM using the map-ont preset for ONT reads. Hostile outputs summary statistics in JSON format including the total number of reads before and after decontamination.

A custom human reference genome was built from the current telomere-to-telomere human genome assembly (Nurk *et al.*, 2022) and human leukocyte antigen (HLA) sequences. Human Illumina (Byrska-Bishop *et al.*, 2022) and ONT (Jain *et al.*, 2018) reads from the well-characterised NA12878 sample were downsampled using BBTools

Table 1. Evaluation of Hostile and the Human Read Removal Tool (HRRT) on real human and simulated bacterial and mycobacterial reads

			Reads retained (%)			Execution time (s)		
Dataset	Samples	Total reads	Hostile	Hostile (masked)	HRRT	Hostile	Hostile (masked)	HRRT
Human Illumina (real)	1	199,510,296	0.1317%	0.1320%	1.4687%	1,299	1,215	1,056
Human ONT (real)	1	2,498,111	0.0380%	0.0381%	0.0373%	2,744	2,747	448
Bacteria Illumina	985	273,511,602	99.9999%	99.9999%	99.9988%	3,526	3,529	937
Bacteria ONT	985	8,230,970	99.9892%	99.9970%	99.9890%	1,375	1,370	5,559
Mycobacteria Illumina	140	51,360,128	100.0000%	100.0000%	99.9999%	680	691	838
Mycobacteria ONT	140	1,544,982	99.9995%	100.0000%	99.9986%	312	309	808

Note: Percentages of retained reads represent the sum of reads from all samples in the dataset

(Bushnell, 2014) to a target depth of 10. Refer to the Supplementary Text for detailed information about test data and masked reference genome construction. Illumina reads were simulated with DWGSIM (Homer, 2010) while ONT reads were simulated with PBSIM2 (Ono *et al.*, 2021).

We evaluated Hostile version 0.0.2 performance alongside the Human Read Removal Tool (HRRT; also known as Human Scrubber; `https://github.com/ncbi/sra-human-scrubber`) version 2.1.0. Testing was performed using an Ubuntu 22.04 AMD64 virtual machine.

3 Results

Full benchmarking results are shown in Supplementary Table S1, summarised in Table 1 and described here. Refer to the Supplementary Text for detailed information about test data preparation.

Accuracy of human read removal: human read removal accuracy was evaluated using real Illumina and Nanopore reads. Refer to the Supplementary Text for detailed information about test data. For Illumina data, Hostile retained 0.132% of human reads while HRRT retained 1.47% of human reads (11-fold more). For ONT data, Hostile and HRRT retained similar percentages of human reads – 0.038% and 0.037% respectively. Use of a reference genome masked against bacterial sequences negligibly increased Hostile's retention of human reads from 0.131738% to 0.131979% (Illumina) and from 0.038029% to 0.038069% (ONT).

Accuracy of microbial read retention: accuracy of bacterial read retention was evaluated using Illumina and ONT sequences simulated from reference-grade complete bacterial assemblies in the FDA-ARGOS dataset. For simulated Illumina data, Hostile retained 99.99989% of reads while HRRT retained 99.99875%, corresponding to HRRT removing 11 times as many bacterial reads as Hostile. Hostile's bacterial read retention increased to 99.99994% through the use of a reference genome masked against bacterial sequences. For simulated ONT data, Hostile and HRRT retained similar percentages of bacterial reads – 99.98918% and 99.98901% respectively. Use of a masked reference genome reduced the number of bacterial ONT reads removed by Hostile by 72% (from 891 to 251).

For mycobacterial reads, 140 complete assemblies were simulated in the same fashion. For simulated mycobacterial Illumina data, Hostile retained 99.99998% of reads while HRRT retained 99.999868%, corresponding to HRRT removing 9 times more mycobacterial reads than Hostile. For simulated mycobacterial ONT data, Hostile retained 99.99948% of reads while HRRT retained 99.99864%. Use of a masked reference with Hostile resulted in perfect (100%) retention of both Illumina and ONT reads.

Performance and execution time: Execution time was measured as the median wall clock time required to process gzip-compressed FASTQ input and create gzip-compressed decontaminated FASTQ output with 8 threads. See Table S1 for the commands used. Neither tool was faster for all datasets tested. HRRT was consistently faster at decontaminating human reads, taking 1,056s and 448s to decontaminate real Illumina and ONT reads respectively, while Hostile took 1,215s and 2,743s. For simulated bacterial Illumina reads, HRRT was faster, taking 937s vs. 3,507s for Hostile. However, for ONT reads, Hostile was faster, taking 1,369s vs. 5,559s for HRRT. For mycobacterial reads, Hostile decontaminated both Illumina and ONT reads faster than HRRT (687s and 309s for Hostile vs. 838s and 808s for HRRT). Hostile's overall per-base throughput of 8-30Mbp/s during testing was more consistent than HRRT's 7-46Mbp/s throughput. HRRT's creation of uncompressed temporary FASTQ files may explain its varied performance characteristics for host-light and host-heavy input data.

4 Discussion

In any diagnostic or experiment where microbial genomes might be contaminated with human genomes, host decontamination is necessary both to safeguard patient anonymity and to avoid encumbering downstream analyses with redundant and potentially detrimental off-target sequences. For downstream analysis it is also of critical importance that microbial sequences are not inadvertently removed, leading to false variant calls and broken *de novo* assemblies. Where target microbes are unknown *a priori*, mixed or sufficiently novel, a subtractive human read removal approach is required, involving non-trivial computation using gigabytes of RAM. Hostile uses one of two complementary seed-and-extend aligners to accurately excise human reads. Bowtie2 is well-suited for decontaminating short reads due to its small memory footprint, fast index loading, and memory-mapped index support, while Minimap2 offers excellent long (and short) read performance in return for a larger index that is considerably slower to load. Compared to an existing approach, Hostile is more sensitive in terms of removing human reads, and specific in retaining diverse bacterial reads, and we have shown that its specificity increases further by use of a masked reference genome. Unlike some existing tools, Hostile streams compressed fastq input to compressed fastq output without creating intermediate files. Hostile's RAM requirements are increasingly met by consumer laptops, creating scope for client-side host decontamination using what we hope will be broadly useful software.

Funding

BC and DC are funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR200915)

References

Bush, S. J. *et al.* (2020). Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microbial Genomics*, **6**(7).

Bushnell, B. (2014). Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Byrska-Bishop, M. *et al.* (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, **185**(18), 3426–3440.e19.

Danecek, P. *et al.* (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, **10**(2), giab008.

Homer, N. (2010). DWGSIM: Whole Genome Simulator for Next-Generation Sequencing.

Hunt, M. *et al.* (2022). ReadItAndKeep: rapid decontamination of SARS-CoV-2 sequencing reads. *Bioinformatics*, **38**(12), 3291–3293.

Jain, M. *et al.* (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, **36**(4), 338–345.

Kim, D. *et al.* (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, **26**(12), 1721–1729.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4), 357–359.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

Nurk, S. *et al.* (2022). The complete sequence of a human genome. *Science*, **376**(6588), 44–53.

Ono, Y. *et al.* (2021). PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics*, **37**(5), 589–595.

Sichtig, H. *et al.* (2019). FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. *Nature Communications*, **10**(1), 3313.

Wood, D. E. *et al.* (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, **20**(1), 257.