

DePauw University

Scholarly and Creative Work from DePauw University

Computer Science Faculty publications

Computer Science

3-2020

CSMD: a computational subtraction-based microbiome discovery pipeline for species-level characterization of clinical metagenomic samples

Paul W. Bible

Marian University - Indianapolis, paulbible@depauw.edu

Yu Liu

Bin Zou

Qiaoxing Liang

Cong Dong

See next page for additional authors

Follow this and additional works at: https://scholarship.depauw.edu/compsci_facpubs



Part of the [Computer Sciences Commons](#)

Recommended Citation

Liu Y, Bible PW, Zou B, Liang Q, Dong C, Wen X, Li Y, Ge X, Li X, Deng X, Ma R. CSMD: A computational subtraction-based microbiome discovery pipeline for species-level characterization of clinical metagenomic samples. *Bioinformatics*. 2019 Oct 18. doi: 10.1093/bioinformatics/btz790



This Article is brought to you for free and open access by the Computer Science at Scholarly and Creative Work from DePauw University. It has been accepted for inclusion in Computer Science Faculty publications by an authorized administrator of Scholarly and Creative Work from DePauw University.

Authors

Paul W. Bible, Yu Liu, Bin Zou, Qiaoxing Liang, Cong Dong, and et al.

Data and text mining

CSMD: a computational subtraction-based microbiome discovery pipeline for species-level characterization of clinical metagenomic samples

Yu Liu^{1,2}, Paul W. Bible^{2,3}, Bin Zou ², Qiaoxing Liang², Cong Dong⁴, Xiaofeng Wen², Yan Li², Xiaofei Ge², Xifang Li², Xiuli Deng², Rong Ma², Shixin Guo², Juanran Liang², Tingting Chen², Wenliang Pan¹, Lixin Liu⁴, Wei Chen ^{5,6}, Xueqin Wang^{1,7,*} and Lai Wei^{2,*}

¹Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou 510275, China, ²State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China, ³College of Arts and Sciences, Marian University, Indianapolis, IN 46222, USA, ⁴College of Chemistry, Sun Yat-Sen University, Guangzhou 510275, China, ⁵Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA, ⁶Division of Pulmonary Medicine, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh, Pittsburgh, PA 15224, USA and ⁷Southern China Research Center of Statistical Science, Sun Yat-Sen University, Guangzhou 510275, China

*To whom correspondence should be addressed

Associate Editor: Jonathan Wren

Received on May 20, 2019; revised on September 22, 2019; editorial decision on October 13, 2019; accepted on October 16, 2019

Abstract

Motivation: Microbiome analyses of clinical samples with low microbial biomass are challenging because of the very small quantities of microbial DNA relative to the human host, ubiquitous contaminating DNA in sequencing experiments and the large and rapidly growing microbial reference databases.

Results: We present computational subtraction-based microbiome discovery (CSMD), a bioinformatics pipeline specifically developed to generate accurate species-level microbiome profiles for clinical samples with low microbial loads. CSMD applies strategies for the maximal elimination of host sequences with minimal loss of microbial signal and effectively detects microorganisms present in the sample with minimal false positives using a stepwise convergent solution. CSMD was benchmarked in a comparative evaluation with other classic tools on previously published well-characterized datasets. It showed higher sensitivity and specificity in host sequence removal and higher specificity in microbial identification, which led to more accurate abundance estimation. All these features are integrated into a free and easy-to-use tool. Additionally, CSMD applied to cell-free plasma DNA showed that microbial diversity within these samples is substantially broader than previously believed.

Availability and implementation: CSMD is freely available at <https://github.com/liuyu8721/csmd>.

Contact: weil9@mail.sysu.edu.cn

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Due to the decreasing costs and the ability to reach species-level or higher taxonomic resolution, metagenomic shotgun sequencing has become more popular in human microbiome studies, including the body sites with low microbial loads, such as stomach (Zhang *et al.*, 2015), ocular surface (Wen *et al.*, 2017) and blood (Kowarsky *et al.*, 2017). However, these low microbial biomass samples present distinct methodological challenges in microbiome profiling.

In clinical samples with low microbial biomass, DNA from host genomes will often dominate the study sample greatly diluting the signal of the actual microbiome (Walker *et al.*, 2018; Zhang *et al.*, 2015). A non-comprehensive removal of human sequences may confound the accurate detection of these microbial signals by misidentifying host DNA as novel microbial reads (Riley *et al.*, 2013). However, aggressive filtering of sequence data will further decrease the positive microbial signal. Therefore, microbiome studies interrogating human body sites with low microbial biomass must

carefully balance sequence identification with excluding the host's genomic signal in a computationally efficient manner.

Assignment-dependent taxonomic classification using some kind of reference data is the common strategy for microorganism identification and quantification in metagenomic study samples (McIntyre et al., 2017; Sczyrba et al., 2017). The reference data can be whole genome sequences [e.g. MEGAN (Buchfink et al., 2015; Huson et al., 2011)], marker genes [e.g. MetaPhlan2 (Truong et al., 2015) and mOTUs2 (Milanese et al., 2019)], compositional properties of genomes [e.g. CLARK (Ounit and Lonardi, 2016; Ounit et al., 2015) and Kraken (Wood and Salzberg, 2014)] or databases of known protein sequences [e.g. kaiju (Menzel et al., 2016)]. These methods attempt to assign every single read to its origin reference genome or the lowest common ancestor in the taxonomic tree. An individual short read has a small amount of information either causing false identification due to reads mismatch to related taxa or reads discarding at high-rank taxa due to lack of uniquely classified information. The limited amount of microbial reads in low biomass samples compound these issues increasing the challenge of accurate identification and estimation. Attempts have been made to mitigate these problems by using more data from the same kind of samples to enhance the microbial signal (Kowarsky et al., 2017; Pasolli et al., 2019) and utilizing reference genome read coverage models to improve individual species detection (Lindner and Renard, 2015; Zhang et al., 2015).

The rapid growth of reference databases is another concern that confounds high-resolution taxonomic classification. With high database redundancy, it is in theory possible to identify organisms with higher accuracy but numerous highly similar sequences in the database complicate the read assignment to the correct source species in the sample. The inhomogeneous distribution and proliferation of reference genome taxa increase the challenge for taxonomic classification (Lindner and Renard, 2015). The limited or absent reference genomes in some taxa make classification imprecise or impossible, while over-represented reference genomes in other taxa may lead to very coarse or unreliable identifications. These demonstrate the necessity of a complete and high-quality reference genome database with low redundancy for microbial classification.

In order to tackle these problems, a computational subtraction-based microbiome discovery (CSMD) pipeline is presented to generate species-level characterization of microorganisms for low microbial biomass samples. CSMD functions in two critical steps: sensitive and specific computational subtraction of host-derived DNA followed by taxonomic profiling based on a mapping of metagenomic reads against a comprehensive, non-redundant and study-specific reference database. The first step amplifies the faint microbial signal in the sample through multiple stages of host subtraction. Next, a finely tuned and study-specific microbial database is generated and leveraged for rapid and accurate species-level classification and profiling for each sample. A stepwise process converges to a high-quality database of target species through a novel sampling and quality checking procedure. CSMD addresses the issues mentioned above in the following ways:

- i. All putative non-human data from the same study group are pooled together to maximize the microbial signal for the identification;
- ii. An initial and sufficiently redundant species-level microbial database that contains homogeneous reference genomes for each species is generated;
- iii. The species genomes selected for the reference database are gradually refined to maximize their sensitivity and accuracy within the study group samples;
- iv. The likelihood estimation of which species genomes are present in the sample can be measured and compared using agreement statistics estimated from modeling their genome coverage profile.

In this study, we demonstrate the advantages of CSMD over existing tools on simulated and artificial data, using performance

metrics under the framework of the Critical Assessment of Metagenome Interpretation (CAMI) (Meyer et al., 2019). Furthermore, we applied CSMD to cell-free DNA (cfDNA) plasma samples to test its performance on data characteristics of low microbial biomass clinical samples demonstrating the clinical relevance of CSMD in devising a new roadmap for the microbial community profiling.

2 Materials and methods

CSMD uses post-QC data as an input and is comprised of the following main steps: human-derived DNA removal, curation of a comprehensive and non-redundant database, and taxonomic profiling. Below, we describe each step along with the benchmark dataset. The complete workflow is visualized in Figure 1A and detailed in Supplementary Data S1.

2.1 Computational subtraction of human-derived data

CSMD performs human-derived reads subtraction through a series of successive detection and filtering of human assembled genomes and low complexity sequences (Figure 1B), as described by Kostic et al. (2011) and Zhang et al. (2015). CSMD makes some improvements to enhance the sensitivity and accelerate filtering. First, CSMD simplifies the procedure by forgoing alignment to large redundant databases such as the Ensembl human reference genomes. Second, two additional human genomes representing different races (specifically Asian and African) are introduced to provide a more complete human decontamination reference database. Finally, instead of using the computationally expensive program, BLAST, for extra human reads removal, Bowtie2 is employed with 'very-sensitive-local' algorithm to enormously speed up human-read identification.

2.2 Comprehensive and non-redundant reference database

After host read filtering, CSMD generates a comprehensive and non-redundant genome database including, as far as possible, any organisms that may emerge in the sample data. In order to maximize the signal from the samples, all putative non-human sample reads were pooled, and a painstaking microbiome discovery procedure was employed to determine their species of origin (Figure 1C). The following subsections describe this process.

2.2.1 Initial homogeneous reference genome database

Initially, CSMD compiles a list of microbial reference genomes that come from tens of thousands of species tabulated in the NCBI RefSeq archive (<ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq>). A representative genome for each species is designated according to the taxonomic information and when multiple genomes are available for a species, the complete overdraft genome is picked up or the genome with the latest sequenced date is selected. The large initial database represents a snapshot of all currently studied microorganisms and is constructed to provide both breadth and depth in identifying microbial signals. Considering restrictions of index size from the aligners and memory management, the database is divided into multiple segments to facilitate processing species identification step. All CSMD analysis used RefSeq bacteria database downloaded on November 1, 2018.

2.2.2 Convergence of target microbiome

The convergent procedure of CSMD's target microbial community identification executes in three phases:

- i. Identify a list of alternatives of candidate species genomes through fast similarity search against the initial database;
- ii. Do species correction to adjust for possible misidentification and generate more reliable species candidates through BLAST analysis of their mapping reads;

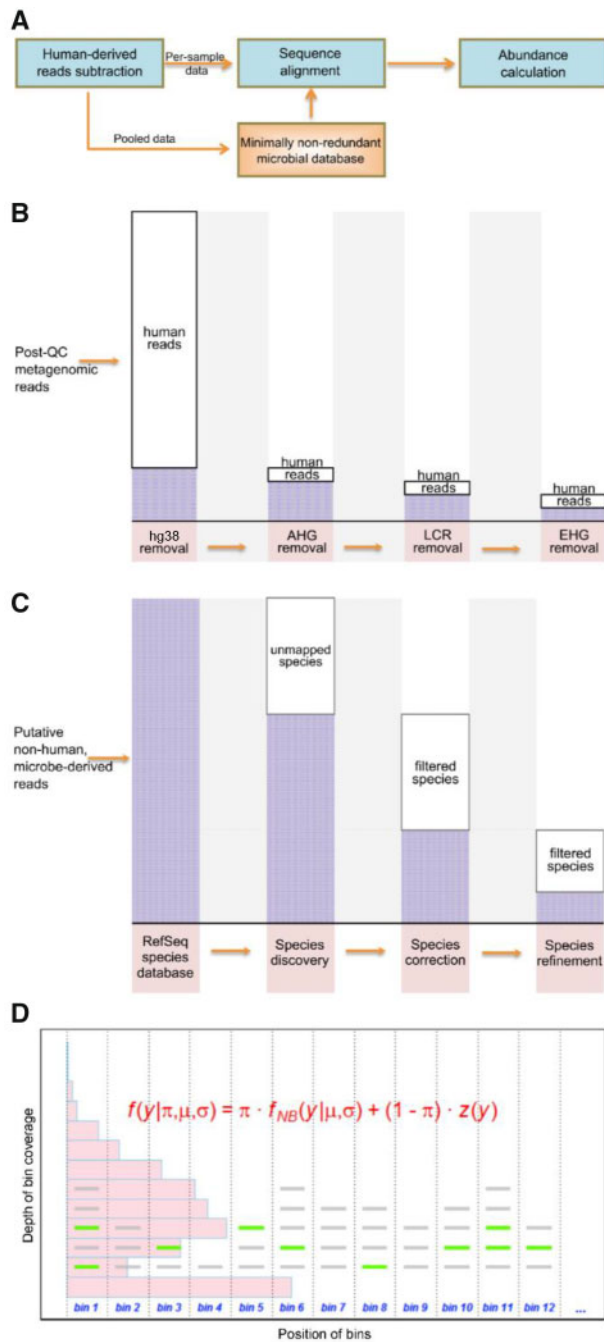


Fig. 1. The computational subtraction-based microbiome discovery pipeline for taxonomic profiling of low microbial biomass clinical samples. (A) The bioinformatic pipeline for microbial identification and profiling from low biomass clinical samples. In this pipeline, a more stringent filtering of host-derived reads is introduced with a more sensitive and accurate microbial profiling procedure. (B) The conceptual workflow to perform computational subtraction of human-derived reads. The size of the light blue bars represents the amount of remaining reads at the indicated step, and the size of the white bars represents the amount detected as human. AHG, reads from additional human genome; LCR, low complexity reads; EHG, reads from external human genome. (C) The conceptual workflow to generate the minimally non-redundant database. Starting with a complete RefSeq representative species database, candidate species genomes were discovered, corrected and refined. Finally, a comprehensive and minimally non-redundant microbial database was constructed for microbial taxa. (D) The coverage structure analysis to model the sequencing procedure for species refinement and evaluation. The number of sequencing reads mapped to a position in a source genome is expected to have a homogeneous coverage depth and modeled as a mixture of a negative binomial distribution and a zero distribution

iii. Perform species refinement through analyzing their genome coverage profile after re-aligning the metagenomic data to the species candidates.

We refer to these phases as species finding, species correction and species refinement, respectively.

CSMD first performs preliminary species identification through read alignment against the initial database. The process gives a long list of suspected species likely present in the sample. Due to the speed-sensitivity trade-off of fast mapping alignment tools, the problem of misidentification cannot be avoided. Misclassification of sample reads from a single species to multiple related species can confound classification efforts. Additionally, a suitable reference genome for a sample species member may not be present in the initial database. Thus, CSMD screens out species with significantly insufficient coverage (e.g. 25 pairs of 100 bp paired-end reads with average coverage $\geq 0.1\%$) and performs a species correction for possible misidentification using BLAST (Altschul *et al.*, 1990) for greater sensitivity and specificity. A random subsampling of reads mapped to each alternative species genome is analyzed using BLAST to reduce the processing time without losing specificity. The nt database, with RefSeq as a supplement for sequences using shotgun sequencing technique (that are excluded from nt), is applied to reassess the classification (see Supplementary Data S1.2.3). Each BLAST report is summarized by simply voting at two resolution levels: species and reference genome (Supplementary Table S1). The species with the vote number beyond the threshold, as described above, is selected as a candidate species and the reference genome with the highest vote for that species is included in the updated genome list. If there are more than one species passing this criterion, all representative genomes are preserved, or if there is no such a species, the alternative species are marked and discarded.

After species correction, a lot of alternative species may be corrected to the same candidate species. These candidate species will be merged and the representative genome with the highest vote number will be preserved. Then, a database with the updated representative genomes can be collected. This intermediary database provides more reliable species candidates for the identification of micro-organisms in the sample. Undergoing re-alignment against the updated database, we can further assess what species are likely to be present in the sample.

Distinct from single-read-based taxonomic classification in the first two phases, CSMD identifies micro-organisms using their genome coverage profile to further improve false detection and get a higher precision (Lindner and Renard, 2015). In this phase, a candidate genome with length L_g is first divided into non-overlapping and equal-sized bins. The number of sequencing reads mapped to a bin in a source genome, denoted by y , is expected to have a homogeneous coverage depth and modeled as a zero-inflated negative binomial distribution, that is, a mixture of a negative binomial distribution f_{NB} and a zero distribution z (Figure 1D):

$$f(y|\pi, \mu, \sigma) = \pi \cdot f_{NB}(y|\mu, \sigma) + (1 - \pi) \cdot z(y),$$

where π , μ and σ are the distribution parameters. Agreement statistics coming from this procedure, Genome-Dataset Validity score (Lindner *et al.*, 2013), π in the formula, denoted by GDV_{bin} , and bin divergence (Sampson *et al.*, 2011), σ in the formula, denoted by σ_{bin} , are estimated to distinguish the correctly identified microorganisms from the false positives. The GDV_{bin} measures the similarity between the candidate reference genome and the source genome, and takes values from 0 (no similarity) to 1 (complete similarity). The σ_{bin} measures the reliability referring to the extent to which the source genome can be represented by the candidate genome, and larger values imply more reliability. Thus, the larger these two statistics are, the more likely the candidate genome is to be present in the sample. In this study, we used $GDV_{bin} > 20\%$ and $\sigma_{bin} > 0.5$ to determine the positive bacteria identification. A detailed mathematical description of the modeling process can be found in Supplementary Data S2.

2.3 Taxonomic profiling

Once the comprehensive and non-redundant genome database is built, CSMD aligns per-sample non-human data against it to generate the microbiome characterization for each sample. Alignment efficiency is improved by depleting nearly all host data as well as searching a much smaller non-redundant microbial genome database. The process of mapping to the new non-redundant reference additionally recovers previously ambiguously mapped reads leading to improved accuracy. CSMD further enhances taxonomic sensitivity by running Bowtie2 with the ‘very-sensitive-local’ algorithm. Finally, microbial abundance estimates including their taxonomic information (e.g. species, genus, family, etc.) are reported.

2.4 Benchmarking

CSMD, specifically adapted for microbiome characterization of clinical samples with low microbial loads, was designed to: (i) efficiently subtract human-derived sequences but maximally preserve microbe-derived sequences and (ii) provide sensitive and accurate detection of microbial signal and minimize the false positive identification. We benchmarked CSMD along two groups of other published bioinformatics tools: one included BWA (Li and Durbin, 2009), Hisat2 (Kim et al., 2015), Bowtie2 (Langmead et al., 2009) and GATK PathSeq (Walker et al., 2018) for human sequence identification and filtering; and another included MEGAN, Kraken, CLARK, CLARK-S, MetaPhlan2 and mOTUs2 for microbial detection. These tools were selected because they are all well-studied and commonly applied for metagenomic sequence processing pipelines.

The human sequence subtraction process was benchmarked by assessing the final read composition in the putative non-human data. And the microbial profiles were compared using precision-recall plots for species identification and L1 distance for species quantification. Precision is calculated as $TP/(TP+FP)$ and recall as $TP/(TP+FN)$, where TP are true positives, FP are false positives and FN denotes false negatives. The L1 distance is defined as the absolute difference between estimated relative abundances and relative abundances simulated as ground truth.

Three sets of data were used in the benchmarking. Artificial data containing six simulated datasets were generated using DWGSIM (<https://github.com/nh13/DWGSIM>). Each dataset is combined with reads generated from the human genome and several bacterial genomes commonly found to be commensal with the human body (Supplementary Table S2). Twenty million 100 bp paired-end reads were first generated from hg38 with the same parameters as Li and Durbin (2009). The simulated bacteria reads were derived from 22 bacterial genomes. To simulate distinct evolutionary distances, six different substitutional mutation rates of 0%, 1%, 3%, 5%, 10% and 20%, were introduced for each of these genomes. Under each mutation rate, 4400 paired-end reads of length 100 bp were produced for the 22 bacteria, each with 200 reads, and then pooled with the 20 million simulated human reads to comprise of a simulated metagenomic dataset. Artificial datasets from the CAMI study were included for a more comprehensive performance evaluation of the profiling process, which contained one low complexity, two medium complexity and five high complexity datasets. In addition, the biological samples were acquired and sequenced as described previously (McIntyre et al., 2017). Three human-spike-in samples were included and served as negative controls for the detection of experimental contaminants which can significantly confound microbiome analyses of low biomass clinical samples.

2.5 Cell-free plasma DNA samples

A recent study held by Stanford University used nucleic acids in plasma to investigate the microbial diversity within the human body (Kowarsky et al., 2017). In this study, the cohort of 32 pregnant women with 120 plasma samples throughout their pregnancies were included. The cell-free DNA was collected and sequenced on the Illumina platform. All fastq data were pre-processed using Trimmomatic (Bolger et al., 2014), and the quality was confirmed by FASTQC (Andrews, 2010).

3 Results

3.1 Performance on human-derived sequence subtraction

After the whole computational subtractive process, almost all human-derived reads (>99.9999%) were correctly removed, and only very few reads (25 for each, ~0.53%) from bacteria-derived genomes were filtered. The bacteria-derived filtered reads were mainly removed during the low complexity read filtering phase (Figure 2A and Supplementary Table S3). After filtering, the microbial signal emerged from the remaining reads and became dominant in the sample data.

Conversely, commonly used one-stage human filtering processes, including BWA, Hisat2 and Bowtie2, failed to detect a substantial proportion of human reads (Figure 2B). The newer multi-stage filtering process, GATK PathSeq, filtered more host sequences but simultaneously depleted the microbial signal. Consequently, all these tools except CSMD resulted in the microbial signal constituting a minority of the remaining sequences (Figure 2C).

3.2 Performance on microbiome discovery

The whole host-filtered sample datasets were pooled as an input to CSMD’s three-phase microbiome discovery procedure. All 22 species of origin, each with a representative strain genome, were correctly identified to be present in the simulated data (Figure 3A).

The convergent procedure of target microbial communities can be found in Figure 3B. Fast similarity analysis in the finding phase identified 273 species in the simulated data, which included 21 correct ones (Table 1). This indicates that very high false positives arise in the species-level identification using a highly redundant reference genome database. After screening, 36 species were kept and 1/3 were the close relatives of the positive control species. These false classifications were corrected in the following phase by the BLAST analysis. Some rare exceptions remained as BLAST could not distinguish some reads from *Bacteroides vulgatus*/*Enterococcus faecalis* or *B.dorei*/E. sp. 7L76 because of their high sequence similarity.

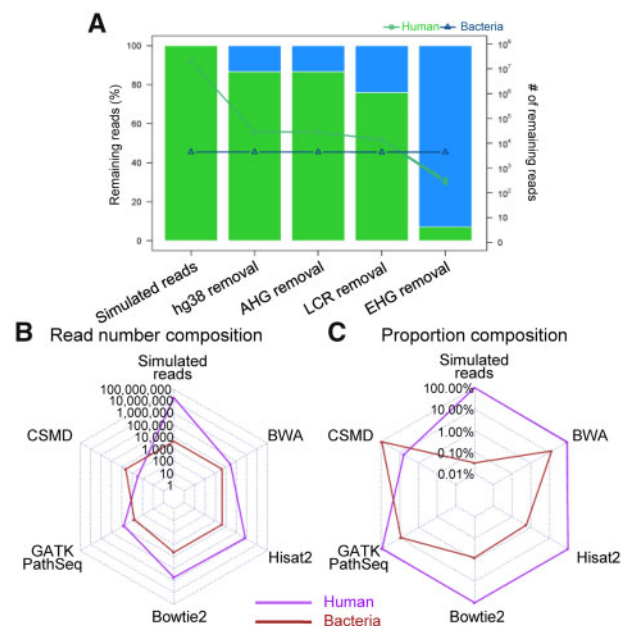


Fig. 2. Pipeline performance on simulated data for human-derived reads subtraction. (A) The bars represent the composition of human (green) and bacteria (blue) after the indicated phase in the CSMD computational subtraction pipeline. And the lines with the same color represent the number of remaining reads with $\log_{10}(x+1)$ scale. (B) Relative performance of different tools for human reads identification with different metrics. Read numbers of false negatives (purple) and true negatives (red) are shown. (C) The relative composition of human reads (purple) and bacteria reads (red) after human reads filtering using different tools. (Color version of this figure is available at *Bioinformatics* online.)

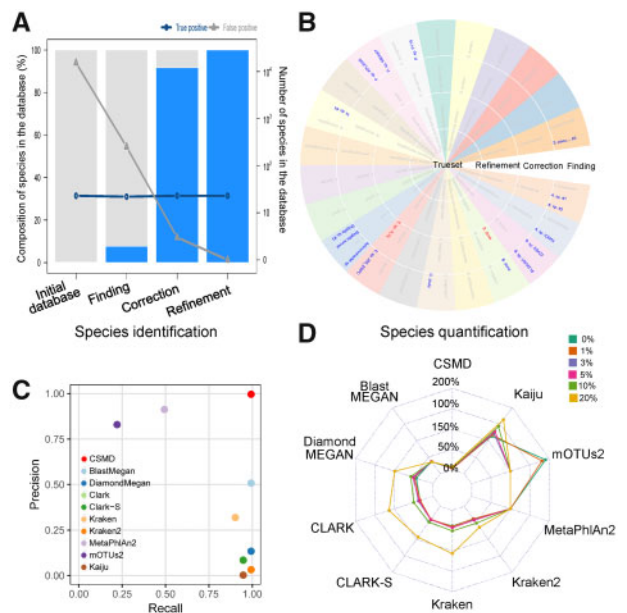


Fig. 3. Pipeline performance on simulated data for microbiome discovery at species resolution. (A) The bars show the composition of true positive (blue) and false positive (grey) species included in the library after the indicated step in the CSMD microbiome discovery pipeline. The lines with the same color represent the number of correctly identified bacteria-derived reads with $\log_{10}(x+1)$ scale. (B) Microbiome discovery detail. Two hundred and seventy-three species were identified by fast similarity analysis in the finding phase according to the CSMD pipeline and 36 species were preserved after screening out genomes with significantly insufficient coverage. After species correction and refinement, species discovered by the CSMD converged to the true set. True positive and false positive detection are colored with blue and gray texts. (C) Precision-recall plot, where each data point corresponds to 1 of 10 different tools and all the inputs are the same, that is, from pooled hg38 removal reads. (D) The L1 distances between observed and expected abundances are used to measure the consistency of different pipelines across simulated datasets. (Color version of this figure is available at *Bioinformatics* online.)

Table 1. Generation of non-redundant database for each test data

Study group	Initial database	Finding phase	Correction phase	Refinement phase
Simulated data	15 639	273	24	22
Human-spike-in samples	15 639	3051	182	74
Plasma cfDNA	15 639	9104	1566	550

The coverage structure analysis showed deeper and more uniform coverage validating their presence in the sample (Supplementary Figure S1). In addition, BLAST using nt database could not detect *Actinomyces odontolyticus* in the simulated data because no such a complete sequence exists in nt. Instead, its closest relative, *A.meyeri*, was identified with a very low mapping rate. However, this species was correctly detected with a very high mapping rate when using the RefSeq database because it contained its whole-genome shotgun sequencing assembly. Finally, all the species converged to the simulated true species and over 2/3 were successfully predicted with strain-level resolution (Supplementary Table S2).

We compared results with a set of nine commonly used programs for microbiome detection to evaluate false detection. The usage parameters for each tool can be found in Supplementary Table S4. As shown in Figure 3C, low precision which represents numerous false positives plague metagenomic analyses for all commonly used tools except marker gene methods (MetaPhlan2 and mOTUs2) which struggle from low recall rate. This result highlights their limitations for microbiome identification in low biomass clinical samples. In order to know whether it is

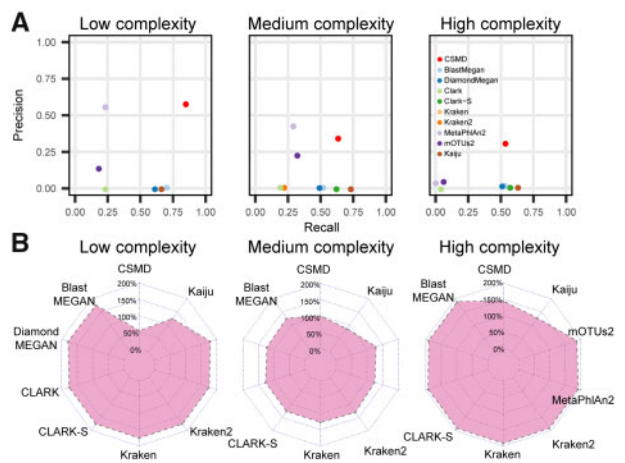


Fig. 4. Pipeline performance on CAMI data for microbiome discovery at species resolution. For the medium and high complexity datasets, plotted values are the average of two or five samples, respectively. (A) Precision-recall plot by complexity. (B) L1 distance by complexity

the better depletion of human genomes that is responsible for the better species identification, we included comparison of species identification using data from different phases in CSMD’s host sequence filtering process. The results showed that more meticulous screening against human genomes can highly improve identification precision for all kinds of tools except marker gene methods (MetaPhlan2 and mOTUs2) (Supplementary Figure S2 and Supplementary Table S5). A closer investigation of these approaches showed that the total nucleotide length mapped to the marker gene catalog was not sufficient to allow species-specific alignment (Milanese et al., 2019; Segata et al., 2012).

3.3 Improvement in abundance estimation

CSMD’s comprehensive and non-redundant refined database leads to a significant increase in the proportion of correctly identified bacteria-derived reads on the simulated data (Supplementary Figure S3A). CSMD correctly identified greater than 98% of all low mutation rate (<5%) bacteria-derived reads and over 30% of the extreme 20% substitutional mutation rate reads (Supplementary Figure S3A). No residual simulated human reads were misclassified as bacterial sequences. The improved mapping rate naturally leads to improvements in abundance estimation. The accuracy of species abundance estimation progressively improved, and high consistency was observed as measured by the L1 distance (Supplementary Figure S3B).

According to the L1 distance metric, the estimated relative abundances from all commonly used tools showed large deviations from the expected values (Figure 3D). Specifically, the marker gene methods (MetaPhlan2 and mOTUs2) had the largest deviation of abundance. These deviations likely result from falsely classified host reads and poor classification rates at species level (Supplementary Figure S4). The design of CSMD systematically addresses these limitations to improve microbial detection and abundance quantification. And as shown in Supplementary Figure S2 and Supplementary Table S5, species quantification of the profilers also benefits from the better removal of host sequence.

3.4 Performance on CAMI challenge data

Even though CSMD was specifically designed for low microbial biomass sample profiling, CSMD outperformed almost all commonly used tools in terms of precision for species identification with comparable recall at any complexity challenge on CAMI data (Figure 4A and Supplementary Table S6). And the L1 distance also showed that CSMD had better abundance estimation (Figure 4B and Supplementary Table S6). As shown in Supplementary Figure S5 and Supplementary Table S6, CSMD achieved an even better performance at other higher taxonomic levels.

3.5 Contamination in the sequencing experiment

Sequence-based microbiome analyses for low biomass clinical samples can be significantly confounded by the contamination. We used the CSMD pipeline to detect the experimental contaminants from the human-spike-in samples.

Even though hg38 removal phase identified and filtered most of the human reads (~94%), extra human-derived reads removal phases identified a lot of reads from human or low complexity sequences (~4%) in the human-spike-in samples, which may complicate and bias the microbial profiling of other tools. At last, less than 1% of reads were left after human sequence subtraction (Supplementary Table S7).

Like with the simulated dataset, CSMD identified or rectified numerous species unlikely to be present in the human-spike-in samples (Table 1). Finally, 74 species were identified as positive contaminants in these samples (Supplementary Table S8). Among of them, six species were found to be predominant in the bacterial composition with the sum of the abundance beyond 50% of classified reads. These are *Mesorhizobium* sp. UASWS1009, *Bradyrhizobium erythrophlei*, *Escherichia coli*, *Acinetobacter gandensis*, *Cloacibacterium normanense* and *Acidovorax* sp. JS42. Their coverage profiles are shown in Supplementary Figure S6A, which indicates their presence in the samples. The species, *Enterobacter cloacae*, which was identified by ‘high-precision pair of taxonomic classification tools’ in McIntyre et al. (2017), seemed to have an insufficient and inconsistent coverage across the genome and failed to pass the coverage evaluation (Supplementary Figure S6B). Through profiling each control sample using the results from microbiome discovery, similar bacterial compositions were observed (Supplementary Figure S7). Thus, CSMD can be employed to detect experimental contaminants that can easily integrate into the filtering pipeline.

3.6 Cell-free plasma DNA data

Having established the pipeline with high performance in simulated samples and contaminants finding in control samples, we next show the application in profiling of microbiome contents in low biomass clinical samples. A recent study held by Stanford University used cell-free DNA from the plasma to investigate the microbial diversity within the human body (Kowarsky et al., 2017). A sub-cohort with 32 pregnant women was involved and cell-free DNA from their 120 plasma samples throughout their pregnancies was collected and sequenced on the Illumina platform.

About 75 million reads were obtained for each cfDNA plasma sample, of which 92% of reads passed quality control. Aside from that 86% of all were identified as hg38 reads, additional 5% human or low complexity reads were detected and removed. Finally, an average of 0.71% of input reads, with a total of 62 million, remained as candidate non-human, quality-filtered Illumina microbial sequence reads (Supplementary Table S7).

A summary of this bacteriome database generation is detailed in Table 1. Finally, the cfDNA Bacteria Reference Database includes 550 assembly sequences that are likely to be of importance in studies of the cfDNA-derived microbiome each for a species representative (Supplementary Table S9).

Using the CSMD generated comprehensive and minimally non-redundant database, the bacterial taxa of each cfDNA sample were characterized. Surprisingly, no species were found in 16 (13.3%) of the 120 samples. Among the rest of 104 (86.7%) samples, there was a large variability in the number of identified species ranging from 1 to 227 with a median of 14. To determine those bacteria likely to show meaningful distribution in the human body, we limited our analysis to those bacteria which were identified in more than 12 (>10%) samples (Supplementary Figure S8).

4 Discussion

Various clinical samples contain small quantities of microbial biomass relative to that of the human host. Microbiome analyses of these samples are challenging due to factors including host contamination and inefficient detection procedures (Minich et al., 2018;

Salter et al., 2014). Overabundance of human-derived reads leads traditional analyses to confound true microbial community data or host sequences that may map to microbial genomes due to insensitive mapping tools. Many studies that attempt to use traditional methods to evaluate these low-biomass microbiome samples fail to adequately address concerns about false positive signals and host content removal.

In contrast to methods based on microbiome-rich analysis, CSMD was developed to efficiently identify and quantify microbial content in low biomass clinical samples (Figure 1). CSMD accurately and efficiently subtracts host-derived data to enhance the true microbial signals (Figure 2). Subsequently, it integrates the entire microbial signal from all study samples to generate a comprehensive and non-redundant reference database (Figure 3) that greatly improves the sensitivity and precision of taxonomic classification for each tested sample (Supplementary Figure S3). To accurately and comprehensively identify the microbial species within a study, a convergent process of species-level target microbiome refinement is designed. Starting from a highly redundant broad coverage species warehouse, CSMD narrows down its detection from a long list of alternatives to a group of much more reliable candidate species reference genomes. This stepwise process ensures that the sensitivity and precision of the database gradually improve after the three phases of species discovery, correction, and refinement. The major benefits of this approach are as follows: (i) The mapping process sees better sequence information utilization as related but irrelevant reference sequences are discarded from the database. (ii) CSMD allows for better taxonomic resolution through a statistical modeling of genome coverage profiles as well as the use of much more accurate (as opposed to fast) alignment tools thanks to a reduction in reference database size. (iii) The decreased redundancy and increased specificity increase the number of mapped and uniquely mapped reads.

In microbiome analysis of low microbial biomass clinical samples, the host is often the major source of contamination overwhelming the microbial signal or confounding it with homologous regions in the host genome. Therefore, the accurate removal of human-derived data becomes essential. Based on previous work from Kostic et al. (2011) and Zhang et al. (2015), we devised an efficient computational subtraction method for human data. Compared with commonly used programs, CSMD demonstrated higher sensitivity in the detection of human reads. Compared to GATK PathSeq, CSMD provides an advantage in specificity which plays a critical role in the analysis of low microbial load samples (Figure 2). The species identification and quantification also benefit from the better depletion of human sequence for all tested programs except marker gene approaches which showed poor performance for low-biomass data because of insufficient usage of data (Supplementary Figure S2 and Supplementary Table S5).

Efficient analysis of low biomass clinical microbiome data suffers from difficulties that arise from a rapidly growing and poorly curated pool of microbial reference genomes. The large and unwieldy database becomes a new barrier in high-resolution microbiome discovery (Dadi et al., 2017). CSMD utilizes all the replicate samples under study and an initial high-redundancy database to generate a comprehensive and non-redundant reference for the study data. The scale of initial database is highly correlated with taxonomic breadth and depth (Tessler et al., 2017). As such, CSMD begins with a sufficiently large database with species resolution satisfying the need for both breadth and depth in the investigation of microbial diversity in low biomass samples. The convergent process of database refinement drives the discovery of the true microbiome present in the sample. The coverage profile modeling of each candidate species genome supports the refinement procedure and contributes to the detection and classification of redundant, and true positive communities within the study dataset (Lindner, et al., 2013; Sampson, et al., 2011). CSMD automatically integrates coverage modeling to improve performance in three ways. First, false positives that are often present as outliers in coverage depth profiles are identified and removed. Second, genomes with inconsistent coverage can be determined as low confidence species present in the sample. Third, a species present in the sample can be validated based on the coverage depth profile. All of these avoid mistakenly discarding species based on low abundance (Shin et al., 2016; Zhang et al., 2017).

As early studies realized (Chen *et al.*, 2010; Fettweis *et al.*, 2012; Griffen *et al.*, 2011), the well-curated specific databases speed up the species-level classifications and provide much higher sensitivity and accuracy for microbiome profiles derived from the study samples (Supplementary Figure S3). These improvements likely stem from reductions in ambiguous mapping that can plague large redundant databases. Better identification and phylogenetic classification of sequences substantially improve our ability to separate health-associated communities from those associated with diseases.

Some limitations of the CSMD pipeline have been recognized despite tackling some important challenges in microbiome analysis for low microbial biomass clinical samples. First, species present in the samples with insufficient sequencing are identified and filtered in coverage analysis. This is an unavoidable issue because species identification relies on the coverage information from the whole genome requiring a suitable signal. More efficient sequencing such as KatharoSeq (Minich *et al.*, 2018) has been proposed. Second, some species identified by CSMD should be further validated experimentally using qPCR or culture. Another potential limitation may be the coverage bin width used to detect species. Using a 5k bp window worked empirically for our controls, but other setting may lead to some true positive species to be missed based on uneven coverage and will be tested in future practice. Lastly, CSMD simply finds the species, the investigator is responsible for the interpretation of their relevance to health with consideration of other conditions such as study design and decontamination procedure.

In conclusion, CSMD is a viable, powerful and freely available pipeline program for high-resolution microbiome profiling of low microbial biomass clinical samples using metagenomic shotgun sequencing. It can maximally eliminate the interference from the host with minimal loss of microbial signal and effectively detect microorganisms present in the sample with minimal false positives.

Acknowledgements

The authors thank Qiuzhuang Lian for the help in the development of CSMD software.

Funding

This work was supported by the National Basic Research Program of China [2015CB964601 to L.W.]; the National Natural Science Foundation of China [81570828 to L.W., 11771462 to X.Q.W.]; and the International Science and Technology Cooperation Program of Guangdong [2016B050502007 to X.Q.W.].

Conflict of Interest: none declared.

References

- Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Andrews,S. (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
- Chen,T. *et al.* (2010) The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database Oxford*, **2010**. doi: 10.1093/database/baq013.
- Dadi,T.H. *et al.* (2017) SLIMM: species level identification of microorganisms from metagenomes. *PeerJ*, **5**, e3138.
- Fettweis,J.M. *et al.* (2012) Species-level classification of the vaginal microbiome. *BMC Genomics*, **13 Suppl 8**, S17.
- Griffen,A.L. *et al.* (2011) CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLoS One*, **6**, e19051.
- Huson,D.H. *et al.* (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, **21**, 1552–1560.
- Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–U121.
- Kostic,A.D. *et al.* (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.*, **29**, 393–396.
- Kowarsky,M. *et al.* (2017) Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci. USA*, **114**, 9623–9628.
- Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
- Li,H., and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
- Lindner,M.S. *et al.* (2013) Analyzing genome coverage profiles with applications to quality control in metagenomics. *Bioinformatics*, **29**, 1260–1267.
- Lindner,M.S., and Renard,B.Y. (2015) Metagenomic profiling of known and unknown microbes with microbeGPS. *PLoS One*, **10**, e0117711.
- McIntyre,A.B.R. *et al.* (2017) Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.*, **18**, 182.
- Menzel,P. *et al.* (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.*, **7**.
- Meyer,F. *et al.* (2019) Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.*, **20**, 51.
- Milanesi,A. *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.*, **10**, 1014.
- Minich,J.J. *et al.* (2018) KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. *mSystems*, **3**, e00218-17.
- Ounit,R., and Lonardi,S. (2016) Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics*, **32**, 3823–3825.
- Ounit,R. *et al.* (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, **16**, 236.
- Pasolli,E. *et al.* (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, **176**, 649–662.e620.
- Riley,D.R. *et al.* (2013) Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. *PLoS Comput. Biol.*, **9**, e1003107.
- Salter,S.J. *et al.* (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.*, **12**, 87.
- Sampson,J. *et al.* (2011) Efficient study design for next generation sequencing. *Genet. Epidemiol.*, **35**, 269–277.
- Szyrba,A. *et al.* (2017) Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.
- Segata,N. *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811.
- Shin,H. *et al.* (2016) Changes in the eye microbiota associated with contact lens wearing. *mBio*, **7**, e00198.
- Tessler,M. *et al.* (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci. Rep.-UK*, **7**, 6589.
- Truong,D.T. *et al.* (2015) MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.
- Walker,M.A. *et al.* (2018) GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, **34**, 4287–4289.
- Wen,X. *et al.* (2017) The influence of age and sex on ocular surface microbiota in healthy adults. *Invest. Ophthalm. Vis. Sci.*, **58**, 6030–6037.
- Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
- Zhang,C. *et al.* (2015) Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome Biol.*, **16**, 265.
- Zhang,H. *et al.* (2017) Conjunctival microbiome changes associated with soft contact lens and orthokeratology lens wearing. *Invest. Ophthalm. Vis. Sci.*, **58**, 128–136.