# CSMD tutorial (Version 0.05)

Computation Subtraction-based Microbiome Discovery (CSMD) is a computational pipeline for high-resolution profiling of low abundance microbiome in clinical samples using whole genome shotgun sequencing.

CSMD is developed at the Wei's Lab at Zhongshan Ophthalmic Center

Lai Wei, Ph.D.
State Key Laboratory of Ophthalmology
Zhongshan Ophthalmic Center, Sun Yat-sen University
54 South Xianlie Road
Guangzhou 510060, China

Developers:
Yu Liu, Qiuzhuang Lian

For support queries, please contact us at liuyu@gzzoc.com

## 1. Pre-requisite software

Table 1. List of pre-requisite software and the available information

| No. | Software | Version | Availability |
|---|---|---|---|
| 1 | BWA | 0.7.15 | http://bio-bwa.sourceforge.net/ |
| 2 | SAMtools | 1.4 | http://www.htslib.org/ |
| 3 | bedtools | 2.26.0 | https://bedtools.readthedocs.io/en/latest/ |
| 4 | seqdk | 1.2 | https://github.com/lh3/seqtk |
| 5 | RepeatMasker | 4.0.7 | http://repeatmasker.org/ |
| 6 | PathoScope 2.0 | 0.02 | https://sourceforge.net/projects/pathoscope/ |
| 7 | R | 3.3.3 | https://www.r-project.org/ |
| 7.1 | countreg | 0.2-0 | https://r-forge.r-project.org/projects/countreg/ |
| 8 | Bowtie2 | 2.2.1 | https://sourceforge.net/projects/bowtie-bio/files/bowtie2/ |
| 9 | fastq-tools | 0.8 | https://github.com/dcjones/fastq-tools |
| 10 | blast | 2.6.0+ | https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download |

## 2. Database availability

Table 2. List of pre-download databases and the available information

| No. | Databases | Availability |
|---|---|---|
| 1 | Human reference genome (hg38) | https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38 |
| 2 | Three assembled human genomes available on NCBI: HuRef, YH and BGIAF. | ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/002/125/GCA_000002125.2_HuRef<br>ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/004/845/GCA_000004845.2_YH_2.0<br>ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/005/465/GCA_000005465.1_BGIAF |
| 3 | Repbase | https://www.girinst.org/ |
| 4 | Ensembl Homo sapiens cDNA database | ftp://ftp.ensembl.org/pub/current_fasta/homo_sapiens/cdna/ |
| 5 | NCBI Homo sapiens RNA database | ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/ |
| 6 | NCBI BLAST human genome database | ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/ |
| 7 | NCBI non-redundant nucleotide sequences (nt) | ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/ |
| 8 | NCBI RefSeq Bacteria database | ftp://ftp.ncbi.hlm.nih.gov/genomes/refseq/bacteria |
| 9 | Taxonomy files | https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz |

The manual also provides a step-by-step guide to download all the databases in Section 6.

## 3. Computational pipeline

**Step 3.1: A four-phase human-derived sequence subtraction.**

- Phase I: Subtract reads from standard human reference genome (hg38)
- Phase II: Subtract reads from three additional assembled human genomes (AHG), including HuRef, YH and BGIAF
- Phase III: Subtract low complexity reads (LCR)

- Phase IV: Subtract reads from three extra human sequence databases (EHG), including Ensembl Homo sapiens cDNA database, NCBI Homo sapiens RNA database, and NCBI BLAST human genome database

**Step 3.2: A four-phase microbiome discovery procedure.**

- Phase I: Generate an initial redundant reference library that may include all the species genomes of interest.
- Phase II: Identify a list of alternatives of candidate genomes through fast similarity search against the initial library
- Phase III: Screen out genomes with significantly insufficient coverage and do the species correction for possibly misidentified genomes using BLAST analysis.
- Phase IV: Perform genome refinement through analyzing their coverage structure.

**Step 3.3: Per-sample microbiome profiling.**

## 4. Installation

### 4.1 Download

Download the code from https://gitlab.com/qiuzhuang/csmd.

You could issue the following command to extract the files: "unzip csmd-master.zip".

### 4.2 Configuration

Make sure the pre-requisite software listed in Table 1 has been installed and available in the runtime environment. They should be first configured in 'config.sh', for example:

```
export PATH=/public/software/bwa/v0.7.15/bin:$PATH
```

Change all *.sh and csmd file to be executable:

```
chmod +x *.sh; chmod +x csmd
```

Add csmd into your PATH available:

```
export CSMD_HOME=/public/users/liuyu/csmd_test/code/csmd-master
export PATH=$PATH:${CSMD_HOME}
```

### 4.3 Databases

CSMD will work with a series of libraries listed in Table 2, including human-related genomes or sequences (30G) and all RefSeq bacteria genomes (150G, as of November 2018). The build process will then require approximately 500GB of additional disk space and 200GB of RAM. These genomes or sequences can be found in DBPATH/hg38/SEQ, DBPATH/AHG/SEQ, DBPATH/EHG/SEQ and DBPATH/RefSeq/bacteria/SEQ, respectively. And the indexed files will be saved in DBPATH/hg38, DBPATH/AHG, DBPATH/EHG and DBPATH/RefSeq/bacteria, respectively.

```
csmd --download-library LIBNAME --db DBPATH
NOTE: --download-library LIBNAME: Permissible LIBNAME includes “hg38”, “AHG”, “EHG”, “nt”, or “RefSeqBac”.
      --db DBPATH: the store path for the download.
csmd --build-library LIBNAME --db DBPATH
NOTE: --build-library LIBNAME: Permissible LIBNAME includes “hg38”, “AHG”, “EHG”, “nt”, or “RefSeqBac”. This command generates csmd needed indexed databases, BWA index for hg38 and AHG databases, Bowtie2 index for EHG and RefSeq representative species databases, and blast index for nt and representative genomes databases. To obtain RepBase, go to http://www.girinst.org.
       --db DBPATH: the store path for the databases, as described above.
```

### 4.4 Taxonomy files

Taxonomy files include the taxonomic lineage of taxa, information on type strains and material, and host information. This command will download the taxonomy files from NCBI and re-build it according to csmd running. These files can be found in DBPATH/taxonomy/taxdump/ and the re-build file will be saved in DBPATH/taxonomy/taxtree.txt. The files taxdb.tar.gz (ftp://ftp.ncbi.nlm.nih.gov/blast/db/taxdb.tar.gz) for blast and GenBank2RefSeq.txt in the CSMD download page for GenBank acceccion number translation should also be saved in DBPATH/taxonomy.

```
csmd --download-taxonomy --db DBPATH
csmd --build-taxonomy --db DBPATH
NOTE: --download-taxonomy: download taxdump.tar.gz from NCBI.
       --build-taxonomy: re-organize the taxonomy tree.
```

## 5. Running CSMD

### 5.1 CS step

This CS procedure performs sensitive and specific computational subtraction of human DNA from the clinical samples. Post-QC paired-end fastq data are assumed as the initial input and indexed human-related genomes or sequences are assumed to be available.

```
1.   csmd --cs hg38Removal --ref REFNAME --thread NUMBER --r1 SAMPLE.R1.fastq --r2 SAMPLE.R2.fastq --output SAMPLE.hg38removal.fastq
2.   csmd --cs ahgRemoval --ref REFNAME --thread NUMBER --input SAMPLE.hg38removal.fastq --output SAMPLE.hg38removal.AHGremoval.fastq
3.   csmd --cs lcrRemoval --ref REFNAME --thread NUMBER --input SAMPLE.hg38removal.AHGremoval.fastq --output SAMPLE.hg38removal.AHGremoval.LCRremoval.fasta
4.   csmd --cs ehgRemoval --ref REFNAME --thread NUMBER \
     --input SAMPLE.hg38removal.AHGremoval.LCRremoval.fasta --output SAMPLE.hg38Removal.AHGRemoval.LCRRemoval.EHGRemoval.fasta
NOTE: --cs: cs step detail name. Permissible step name includes “hg38Removal”, “ahgRemoval”, “lcrRemoval”, or “ehgRemoval”.
      --ref: the reference name with the path used for human-derived reads alignment and removal.
      --thread: number of threads (CPUs) to use.
      --r1, --r2, --input: the input file with the path. For hg38Removal, the input should be paired-end reads using the parameters “--r1” and “--r2”, but
for others, using the parameter “--input". In the phase of hg38Removal, paired-end reads will be combined into single-end data after finishing hg38 reads
```

alignment and removal. The format of the input files is illustrated as the examples.

--output: the output file with the path. The format of the output files is illustrated as the examples.

### 5.2 MD step

This MD procedure develops a comprehensive and minimally non-redundant reference database using pooled data from the study samples. To overcome the limitations of microbial identification from samples with low microbial biomass and to maximize detection power, all putatively non-human reads in the study group are combined as the input in this step. It includes three key sub-steps to make the microbiome finding as accurate as possible: species finding, species correction and species refinement.

```
1. csmd --md finding --ref REFNAME --thread NUMBER --input pooled_nonHuman.fasta --outdir OUTDIR
2. csmd --md correction --sam csmd_finding.sam --report csmd_finding_report.tsv \
   --cutoff 25 --thread NUMBER --nt NTNAME --refseq RSNAME --taxdir TAXDIR --outdir OUTDIR
3. csmd --md refinement --seqdir SEQPATH --seqlist DBLIST --thread NUMBER --input pooled_nonHuman.fasta --outdir OUTDIR
```

### 5.3 Profile step

This procedure provides accurate taxonomy classification for each sample based on a mapping of metagenomic reads against the comprehensive and minimally non-redundant reference database generated in the MD step.

```
1  csmd --pf dbsetup --seqdir SEQPATH --seqlist DBUPDATE --outdir OUTDIR
2  csmd --pf profile --ref REFNAME --thread NUMBER \
   --input SAMPLE.hg38Removal.AHGRemoval.LCRRemoval.EHGRemoval.fasta --sam SAMPLE.csmd.sam --report SAMPLE.csmd.profile.report
```

NOTE: --pf: profile step detail name. Permissible step name includes “dbsetdup” for the indexed CSMD database and “profile” for single sample microbiome profiling.

--seqdir, --seqlist: see MD step.

--sam: reads alignment detail in SAM format.

```
--report: pathoscope style report in tsv format.
```

# 6. Database download

## 6.1 Human reference genomes

(1) hg38 (6.5G with sequence and index files)

```
wget -c ftp://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/* -P DBPATH
zcat DBPATH/*.fa.gz > DBPATH/hg38_ucsc.fasta
```

(2) AHG (16G)

HuRef

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/002/125/GCA_000002125.2_HuRef/GCA_000002125.2_HuRef_genomic.fna.gz -P DBPATH
```

YH

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/004/845/GCA_000004845.2_YH_2.0/GCA_000004845.2_YH_2.0_genomic.fna.gz -P DBPATH
```

BGIAF

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/005/465/GCA_000005465.1_BGIAF/GCA_000005465.1_BGIAF_genomic.fna.gz -P DBPATH
```

Merge the sequences:

```
zcat DBPATH/*.fna.gz > DBPATH/ahg.fasta
```

(3) Repbase (0.2G)

Go to https://www.girinst.org/ and download the Repbase according to the website tutorial.

(4) EHG (150G)

Ensembl Homo sapiens cDNA database

```
wget -c ftp://ftp.ensembl.org/pub/current_fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz -P DBPATH
```

NCBI Homo sapiens RNA database

```
wget -c ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/rna.fa.gz -P DBPATH
```

NCBI BLAST human genome database

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/blast/db/human_genomic.*.tar.gz -P DBPATH
for item in `ls DBPATH/human_genomic.*.tar.gz`;
do
    tar xzvf ${item}
done
blastdbcmd -entry all -db DBPATH/human_genomic -out DBPATH/NCBIBlastHumanSequence.fasta
rm DBPATH/human_genomic.??.n*
rm DBPATH/human_genomic.nal
```

Merge or split the sequences:

```
### Split the fasta file in order that each one has the size less than 6G.
grep '^>' DBPATH/NCBIBlastHumanSequence.fasta | sed 's/.//' > DBPATH/header.list
sed -n '1,128p' DBPATH/header.list > DBPATH/ehg.split.list00
sed -n '129,558p' DBPATH/header.list > DBPATH/ehg.split.list01
```

```
sed -n '559,846p' DBPATH/header.list > DBPATH/ehg.split.list02
sed -n '847,1294p' DBPATH/header.list > DBPATH/ehg.split.list03
sed -n '1295,1683p' DBPATH/header.list > DBPATH/ehg.split.list04
sed -n '1684,1835p' DBPATH/header.list > DBPATH/ehg.split.list05
sed -n '1836,1998p' DBPATH/header.list > DBPATH/ehg.split.list06
sed -n '1999,2171p' DBPATH/header.list > DBPATH/ehg.split.list07
sed -n '2172,2421p' DBPATH/header.list > DBPATH/ehg.split.list08
sed -n '2422,3472p' DBPATH/header.list > DBPATH/ehg.split.list09
sed -n '3473,3505p' DBPATH/header.list > DBPATH/ehg.split.list10
for item in `ls DBPATH/ehg.split.list*`
do
    no=$(echo ${item} | awk -F "." '{print $NF}' | cut -c5-6)
    seqtk subseq -l 80 DBPATH/NCBIBlastHumanSequence.fasta ${item} > DBPATH/ehg.${no}.fasta
done
zcat DBPATH/Homo_sapiens.GRCh38.cdna.all.fa.gz DBPATH/rna.fa.gz > DBPATH/ehg.11.fasta
rm DBPATH/human_genomic.*
gzip NCBIBlastHumanSequence.fasta
```

### 6.2 The nt database

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz -P DBPATH
```

Or user can directly download the BLAST index database:

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt*.tar.gz -P DBPATH
for item in `ls DBPATH/nt*.tar.gz`;
do
     tar xzvf ${item}
```

```
done
```

### 6.3 NCBI RefSeq Bacteria database

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt -P DBPATH
awk '{FS="\t"} !/^#/ {print $20} ' DBPATH/assembly_summary.txt | \
   sed -r 's|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/.+/)(GCF_.+)|\1\2/\2_genomic.fna.gz|' > DBPATH/genomic_file.txt
for item in $(cat DBPATH/genomic_file.txt); do wget -c ${item}; done
```

### 6.4 Taxonomy files

```
wget -c ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz -P DBPATH
tar -xvf DBPATH/taxdump.tar.gz
```

# 7. Help document

## 7.1 Introduction

```
csmd -h/--help
```

```
Program: csmd (A computational pipeline for high-resolution profiling of low abundance microbiome in clinical samples using whole genome shotgun sequencing)
Version: 0.1

Usage:   csmd <command> [options]

Commands:

  Database preparation
  --download-library    Download csmd needed databases
  --build-library       Index csmd needed databases
  --download-taxonomy   Download taxonomy files from NCBI
  --build-taxonomy      Re-organize the taxonomy tree

  Computational pipeline
  --cs                  Perform a sensitive and specific human-derived sequence subtraction
  --md                  Provide a painstaking microbiome discovery procedure
  --pf                  Perform an accurate species-level classification
```

## 7.2 Database preparation

```
csmd --download-library -h/--help
```

```
About: This command download csmd needed databases
Usage: csmd --download-library LIBNAME --db DBPATH

Options:

   --download-library LIBNAME
          Permissible LIBNAME includes “hg38”, “AHG”, “EHG”, “nt”, or “RefSeqBac”.
   --db DBPATH
          The save path for the download.
```

```
csmd --build-library -h/--help
```

```
About: This command generates csmd needed indexed databases, BWA index for hg38 and AHG databases, Bowtie2 index for EHG and RefSeq representative species databases, and blast index for nt and representative genomes databases. To obtain RepBase, go to http://www.girinst.org.
Usage: csmd --build-library LIBNAME --db DBPATH

Options:

   --build-library LIBNAME
          Permissible LIBNAME includes “hg38”, “AHG”, “EHG”, “nt”, or “RefSeqBac”.
   --db DBPATH
          The save path for the indexed database.
```

```
csmd --download-taxonomy -h/--help
```

```
About: This command downloads taxonomy files from NCBI
```

```
Usage: csmd --download-taxonomy --db DBPATH

Options:

  --download-taxonomy
        No argument is needed.
  --db DBPATH
        The save path for the taxonomy files.
```

`csmd --build-taxonomy -h/--help`

```
About: This command re-organizes taxonomy files
Usage: csmd --build-taxonomy --db DBPATH

Options:

  --build-taxonomy
        No argument is needed.
  --db DBPATH
        The save path for the re-organized taxonomy
```

## 7.3 Computational pipeline

`csmd --cs -h/--help`

```
About: A four-phase human-derived sequence subtraction
Usage:
     csmd --cs hg38Removal --ref REFNAME --thread NUMBER --r1 SAMPLE.R1.fastq --r2
SAMPLE.R2.fastq --output SAMPLE.hg38removal.fastq
     csmd --cs ahgRemoval --ref REFNAME --thread NUMBER --input SAMPLE.hg38removal.fastq --
output SAMPLE.hg38removal.AHGremoval.fastq
     csmd --cs lcrRemoval --ref REFNAME --thread NUMBER --input
SAMPLE.hg38removal.AHGremoval.fastq --output SAMPLE.hg38removal.AHGremoval.LCRremoval.fasta
     csmd --cs ehgRemoval --ref REFNAME --thread NUMBER --input
SAMPLE.hg38removal.AHGremoval.LCRremoval.fasta                                    --output
SAMPLE.hg38Removal.AHGRemoval.LCRRemoval.EHGRemoval.fasta

Options:

     --cs STEPNAME
          The cs step detail name. Permissible STEPNAME includes “hg38Removal”, “ahgRemoval”,
“lcrRemoval”, or “ehgRemoval”.
   --ref REFNAME
          The reference name with the path used for human-derived reads alignment and removal.
   --thread NUMBER
          The number of threads (CPUs) to use.
   --r1, --r2, --input
          The input file with the path. For hg38Removal, the input should be paired-end reads
using the parameters “--r1” and “--r2”, but for others, using the parameter “--input". In the
phase of hg38Removal, paired-end reads will be combined into single-end data after finishing
hg38 reads alignment and removal. The format of the input files is illustrated as the examples.
   --output
          The output file with the path. The format of the output files is illustrated as the
examples.
```

`csmd --md -h/--help`

```
About: A three-phase microbiome discovery procedure
```

```
Usage:
     csmd --md finding --ref REFNAME --input pooled_nonHuman.fasta --outdir OUTPATH
     csmd --md correction --sam csmd_finding.sam --report csmd_finding_report.tsv --cutoff 25
--thread NUMBER --nt NTNAME --refseq RSNAME --output DBLIST
     csmd --md refinement --seqdir SEQPATH --seqlist DBLIST --input pooled_nonHuman.fasta --
outdir OUTPATH


Options:

     --md STEPNAME
          The md step detail name. Permissible STEPNAME includes “finding”, “correction”, or
“refinement”.
   --ref REFNAME
          The reference name with the path used for the initial alignment to the redundant
microbial database containing all the organisms or the nearest neighbors likely to be present
in the sample.
     --thread NUMBER
          The number of threads (CPUs) to use.
     --nt NTNAME, --refseq RSNAME
          The nt or RefSeq reference name with the path used for blast analysis.
     --cutoff NUMBER
          The minimum number of mapping reads used for single microbial species discovery.
     --sam FILENAME, --report FILENAME
          The input or output sam file in SAM format and report file in pathoscope tsv format.
The path should be provided.
     --outdir OUTPATH
          The output directory and the default output file(s) will be generated. NOTE: If --
outdir is provided in discovery step, --sam and --report are not used.
     --output DBLIST, --seqlist DBLIST
          The updated microbial genome list which are likely present in the samples.
     --seqdir SEQPATH
          The RefSeq microbial genome directory containing all needed updated microbial genome
list which are likely present in the samples.
```

```
csmd --pf -h/--help
```

```
About: Per-sample microbiome profiling
Usage:
     csmd --pf dbsetup --seqdir SEQPATH --seqlist DBUPDATE --outdir OUTPATH
     csmd --pf profile --ref REFNAME --thread NUMBER \

Options:

     --input FILENAME
     --pf STEPNAME
          The profile step detail name. Permissible STEPNAME includes “dbsetdup” for the
indexed CSMD database and “profile” for single sample microbiome profiling.
   --seqdir SEQPATH
          The RefSeq microbial genome directory containing all needed updated microbial genome
list which are likely present in the samples.
     --seqlist DBLIST
          The updated microbial genome list which are likely present in the samples.
   --sam FILENAME
          Reads alignment detail in SAM format.
   --report FILENAME
          The pathoscope style report in tsv format
```