Pipeline

Prior to using GMHI-webtool, users must profile their stool metagenomes (.fastq). The following is the pipeline used to preprocess and profile the stool metagenomes used to compute GMHI. Users should follow this pipeline closely to ensure GMHI is computed correctly. For brevity, this pipeline assumes that paired end metagenomes are used.

0. Setup

```
# make sure the paired end metagenome files are available
ls
.
├── in1.fastq
└── in2.fastq

# Install/update softwares
"bbmap (repair.sh)": "38.93"
"fastqc": "0.11.9"
"bowtie2": "2.4.4"
"samtools": "0.1.19"
"bedtools": "2.30.0"
"trimmomatic": "0.39"
"metaphlan2.py": "2.x.x"

# set this var to directory containing human reference genome (use
GRCh38/hg38)
$HUMAN_REFERENCE_GENOME

# set this var to desired number of parallel processes
$N_JOBS

# set this var to directory containing metaphlan2 clade markers
$CLADE_MARKERS

# Prepare adapter sequence file
echo ">PrefixPE/1
TACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PrefixPE/2
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT" > TruSeq3-PE.fa
```

1. Repair fastq files using bbmap

```
repair.sh in1=in1.fastq in2=in2.fastq out1=repaired1.fastq
out2=repaired2.fastq outs=garbage
```

2. Quality check and identification of overrepresented sequences

```
fastqc repaired1.fastq
fastqc repaired2.fastq
unzip repaired1_fastq.zip
unzip repaired2_fastq.zip
```

3. Extract overrepresented sequences (probable adapter sequences) from FastQC outputs

```
for f in repaired1_fastqc/fastqc_data.txt; do
    echo $f `grep -A100 ">>Overrepresented sequences" $f | \
    grep -m1 -B100 ">>END_MODULE" | \
    grep -P "Adapter|PCR" | awk '{print ">overrepresented_sequences" "_"
++c "/1" $1}'` | \
    awk '{gsub(/\/1/,"/1\n")}1' | \
    awk '{gsub(/>/,"\n>")}1' | \
    awk '{gsub(/fastqc_data.txt/,"")}1' | \
    awk 'NF > 0';
done > adapter1.txt

for f in repaired2_fastqc/fastqc_data.txt; do
    echo $f `grep -A100 ">>Overrepresented sequences" $f | \
    grep -m1 -B100 ">>END_MODULE" | \
    grep -P "Adapter|PCR" | awk '{print ">overrepresented_sequences" "_"
++c "/1" $1}'` | \
    awk '{gsub(/\/1/,"/1\n")}1' | \
    awk '{gsub(/>/,"\n>")}1' | \
    awk '{gsub(/fastqc_data.txt/,"")}1' | \
    awk 'NF > 0';
done > adapter2.txt
```

4. Remove human contaminants

```
bowtie2 -p $N_JOBS -x $HUMAN_REFERENCE_GENOME -1 repaired1.fastq -2
repaired2.fastq -S mapped.sam

samtools view -bS mapped.sam > mapped.bam

samtools view -b -f 12 -F 256 mapped.bam > human.bam

samtools sort -n human.bam human_sorted -@ $N_JOBS

bedtools bamtofastq -i human_sorted.bam -fq human1.fastq -fq2 human2.fastq
```

5. Remove adapter sequences and low quality reads

```
cat adapter1.txt adapter2.txt TruSeq3-PE.fa > adapters.txt
```

```
trimmomatic PE -threads $N_JOBS human1.fastq human2.fastq -baseout
QC.fastq.gz \
ILLUMINACLIP:adapters.txt:2:30:10:2:keepBothReads LEADING:3 TRAILING:3
MINLEN:60
```

6. Profile the metagenome

```
metaphlan2.py QC_1P.fastq.gz,QC_2P.fastq.gz --bowtie2db $CLADE_MARKERS \
--bowtie2out --index mpa_v20_m200 --nproc $N_JOBS --input_type fastq -o
profiled_metagenome.txt
```

After running the pipeline, users can upload the taxonomic profile "profiled_metagenome.txt" to GMHI-webtool.

Users can use MetaPhlAn2 to merge multiple taxonomic profiles

```
ls
.
├── profiled_metagenome1.txt
├── profiled_metagenome2.txt
└── profiled_metagenome3.txt

merge_metaphlan_tables.py profiled_metagenome*.txt >
merged_abundance_table.txt
```

And upload the merged file "merged_abundance_table.txt" to GMHI-webtool.