Supplementary Information

for

GMHI-webtool: a user-friendly tool for assessing health through metagenomic gut microbiome profiling

Chang *et al.*

1 Webtool Information

1.1 Implementation

Python and the Numpy [HMvdW+20] library were used to pre-compute GMHI, α-diversities, and taxonomic distributions of the pooled dataset. The scikit-learn [PVG+11] library was used to perform Principal Component Analysis (PCA) of the pooled dataset. Access the data here.

Regarding the front-end, GMHI-webtool is a client-side JavaScript application. The D3.js [Bos12] library was used for plotting within the web browser. The Math.js [dJM18] library is used to project the input sample onto the first two principal components of the pooled dataset. α-diversities of the input sample, along with text parsing and validation, are implemented using JavaScript functions.

1.2 User Input

Users need to first upload (**Fig. 1A**) the taxonomy profile (see **Section 2**). If the file has multiple samples, users can select the sample for analysis (**Fig. 1B**). Users may choose to compare the input sample with healthy samples, nonhealthy samples, or all samples in the pooled gut microbiome dataset (**Fig. 1C**). This selection pertains to the first two plots only. After making the proper selections, users can press the "display results" button (**Fig. 1D**) to compute and visualize the analyses described.

1.3 User Interaction and Exports

Users can change plot options (e.g., index, taxonomic level) by using the tabs above plots (**Fig. 2A**). Additionally, users can hover their mouse over legend text (**Fig. 2B**) to highlight information.

Users can export index data by clicking the "export as CSV" button (**Fig. 1E**), and export plots by clicking links below them (**Fig. 2C**).

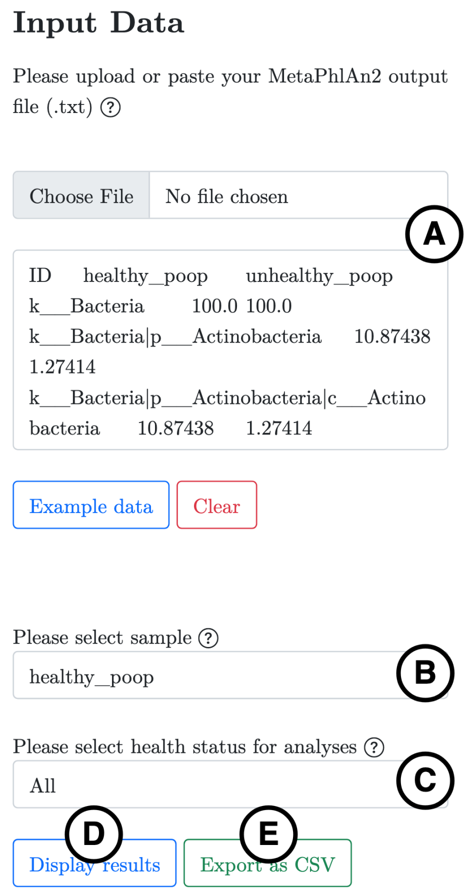

Figure 1: User Input Panel

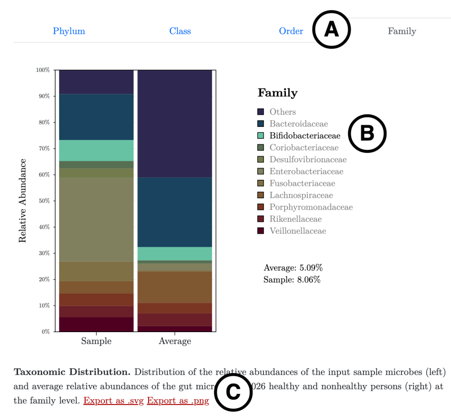

Figure 2: Taxonomic Distribution Plot

2 Pipeline

Prior to using GMHI-webtool, users need to profile their metagenome .fastq files using MetaPhlAn2 [TFT+15]. 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.

2.1 Setup

Install/update softwares:

- bbmap (repair.sh) v38.93 [Bus]
- fastqc v0.11.9 [Source]
- bowtie2 v2.4.4 [LS12]
- samtools v0.1.19 [LHW+09]
- bedtools v2.30.0 [QH10]
- trimmomatic v0.39 [BLU14]
- MetaPhlAn2 v2.x.x [TFT+15]

```
# make sure the paired end metagenome files are available
ls
```

```
.
├── in1.fastq
└── in2.fastq
```

```
# set this var to directory containing human reference genome (use GRCh38/hg38)
# example:
$HUMAN_REFERENCE_GENOME=/users/mynamejeff/human_genomes/GRCh38_noalt_as/

# set this var to desired number of parallel processes
# example:
$N_JOBS=16

# set this var to directory containing metaphlan2 clade markers
# example:
$CLADE_MARKERS=/users/mynamejeff/metaphlan2_data/clade_markers

# Prepare adapter sequence file
```

```
echo ">PrefixPE/1
TACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PrefixPE/2
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT" > TruSeq3-PE.fa
```

2.2 Repair fastq files using bbmap

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

2.3 Quality check and identification of overrepresented sequences

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

2.4 Extract 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
```

2.5 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
```

2.6 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
```

2.7 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 also 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.

3 α-diversities

GMHI-webtool computes a number of α-diversities from stool metagenome samples. Let p_i be the relative abundance of the ith species in the sample. For consistency with the original GMHI work, only the 313 species considered during the computation of GMHI were considered [GKB+20]. Let $S = 313$ be the maximum number of species in a single sample. Let $c = 0.00001$ be the presence threshold.

3.1 Richness

Richness R is the number of species with relative abundance greater than the presence threshold.

$$R = |\{\, i \mid p_i > c \,\}|$$

3.2 Shannon Diversity

Shannon Diversity H' is derived from Shannon entropy, and is a measure of the uncertainty associated with predicting the species of any microbe in the sample.

$$H' = -\sum_{\forall i [p_i > 0]} p_i ln(p_i)$$

3.3 Evenness

Evenness E is a measure of close in number (or relative abundance) different species are.

$$E = \frac{H'}{ln(S)}$$

3.4 Simpson Diversity

Simpson diversity is equivalent to the probability that two randomly selected microbes are of the same species [SIM49].

$$\lambda = \sum_{\forall i [p_i > 0]} ln(p_i)$$

3.5 Inverse Simpson

Inverse Simpson diversity is the reciprocal of Simpson diversity.

$$I = \frac{1}{\lambda}$$

References

[BLU14] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, April 2014.

[Bos12] Mike Bostock. D3.js - data-driven documents, 2012.

[Bus] Brian Bushnell. Bbmap: A fast, accurate, splice-aware aligner.

[dJM18] Jos de Jong and Eric Mansfield. Math.js: An advanced mathematics library for JavaScript. *Computing in Science &amp Engineering*, 20(1):20–32, January 2018.

[GKB+20] Vinod K. Gupta, Minsuk Kim, Utpal Bakshi, Kevin Y. Cunningham, John M. Davis, Konstantinos N. Lazaridis, Heidi Nelson, Nicholas Chia, and Jaeyun Sung. A predictive index for health status using species-level gut microbiome profiling. *Nature Communications*, 11(1), sep 2020.

[HMvdW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[LHW+09] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin and. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, June 2009.

[LS12] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, mar 2012.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[QH10] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, January 2010.

[SIM49] E. H. SIMPSON. Measurement of diversity. *Nature*, 163(4148):688–688, April 1949.

[TFT+15] Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, September 2015.