MetaPrism: A Toolkit for Joint Taxa/Gene Analysis of Metagenomic Sequencing Data

Jiwoong Kim (Jiwoong.Kim@UTSouthwestern.edu) [1#],

Shuang Jiang (Shuangj@mail.smu.edu) [2#],

Yiqing Wang (lucy@mail.smu.edu) [2],

Guanghua Xiao (Guanghua.Xiao@UTSouthwestern.edu) [1,3,4],

Yang Xie (Yang.Xie@UTSouthwestern.edu) [1,3,4],

Dajiang J. Liu (dxl46@psu.edu) [5],

Qiwei Li (Qiwei.Li@utdallas.edu) [6],

Andrew Koh (Andrew.Koh@UTSouthwestern.edu) [3,7,8],

Xiaowei Zhan (Xiaowei.Zhan@UTSouthwestern.edu) [1,3,9*]

[1]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, 75390

[2]Department of Statistical Science, Southern Methodist University, Dallas, TX 75275

[3]Harold C. Simmons Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas, 75390

[4]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas, 75390

[5]Department of Public Health Sciences, Pennsylvania State University, Hershey, Pennsylvania, 17033

[6]Department of Mathematical Science, The University of Texas at Dallas, Dallas, Texas, 75080

[7]Department of Microbiology, University of Texas Southwestern Medical Center, Dallas, Texas, 75390

[8]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas, 75390

[9]Center for Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, Texas, 75390

*To whom correspondence should be addressed

[#]These authors contributed equally to this work

1. Simulation results using HUMAnN2

HUMAnN2 is a popular program as part of the biobakery software suits. It can derive abundances of taxa-specific genes from metagenomic sequence reads, but it does not offer downstream analysis tools such as comparative analysis and prediction modeling. We evaluated the performance of HUMAnN2 using simulation. First, we used ART [1] to simulate short sequence reads at the depth of 10 using 118 bacterial genomes (details in **Results: Joint features inferred by MetaPrism are accurate in simulation**). Then we downloaded HUMAnN2 (version 2.8.1) from biobakery GitHub repository (https://github.com/biobakery/humann) as well as the required short sequence read aligner, bowtie (version 2.3.4.3). We applied HUMAnN2 to the simulated reads and obtain taxa-specific UniRef90 genes abundances. Next, we followed the HUMAnN2 documentation to regroup the abundances from UniRef90 genes to KEGG ortholog groups using the utility program, humann2_regroup_table, provided in the HUMAnN2 distribution. Finally, we compared the inferred taxa-specific gene abundances reported by HUMAnN2 to the true gene abundances (**Figure S1**). The joint features reported by HUMAnN2 is similar to those by MetaPrism. The abundances provided by HUMAnN2 are highly correlated with the true abundances (correlation coefficient = 0.827, p-value < 0.0001), however, the reported abundances are much larger than the true abundances.

2. Discover species-specific biomarker in an immune checkpoint therapy study

This section illustrated a MetaPrism-based workflow to derive joint features as a potential biomarker to predict immune checkpoint therapy responses. Microbiome metagenomics sequencing data was generated for a previously published study [2]. We focused on a subset of patients treated with anti-PD1 (pembrolizumab) therapy. 6 melanoma patients respond to this

therapy and the rest 6 patients did not respond according to the RECIST standard [2]. The patient identifiers, phenotypes and their NCBI SRA Run information is as follows:

Sample Name	Outcome	SRA Run Information
P10	Response	SRR5930498
P14	Response	SRR5930499 SRR5930500
P23	Response	SRR5930499 SRR5930500
P25	Response	SRR5930527
P34	Response	SRR5930533
P39	Response	SRR5930511
P8	Progressive	SRR5930497
P16	Progressive	SRR5930501
P24	Progressive	SRR5930526
P30	Progressive	SRR5930521
P32	Progressive	SRR5930522
P42	Progressive	SRR5930510

We are interested in microbiome-based biomarkers to predict the therapy responses. Below we listed the command line to reproduce our analysis results.

1) Derive patient-specific joint features

Retrieve sequence reads from NCBI SRA, we illustrate the analysis for the sample, P8. The analysis can be similarly applied to all samples.

```
fastq-dump --gzip --split-files SRR5930497
```

De novo assemble metagenome sequence reads

(this step can take large memory and time)

```
spades.py -1 SRR5930497_1.fastq.gz -2 SRR5930497_2.fastq.gz -o
P8.SPAdes –meta
```

Gene annotation and abundance quantification

```
perl MetaPrism_gene.pl -p 8 P8.gene P8.SPAdes/scaffolds.fasta
SRR5930497_1.fastq.gz,SRR5930497_2.fastq.gz
```

Taxon annotation

```
# Obtain Centrifuge - https://ccb.jhu.edu/software/centrifuge/
# Obtain Centrifuge references
#
ftp://ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data/p_compressed_2
018_4_15.tar.gz
perl MetaPrism_taxon_centrifuge.pl -p 8
P8.gene.region.abundance.txt P8.SPAdes/scaffolds.fasta
centrifuge/data/p_compressed > P8.gene_taxon.region.abundance.txt
```

Repeat the above procedure for all patients' sequence reads

2) Build a prediction model

We first prepare a group file with two columns: the first column is sample name and the second column is either response or progressive.

Then we use the MetaPrism prediction script. The default prediction model is random forest with 500 trees and all features will be randomly sampled as candidate split.

```
perl MetaPrism_prediction.pl -F gene_taxon
group.mel.response.vs.progressive P10=gene_taxon.P10
P14=gene_taxon.P14 P23=gene_taxon.P23 P25=gene_taxon.P25
P34=gene_taxon.P34 P39=gene_taxon.P39 P1=gene_taxon.P1
P3=gene_taxon.P3 P6=gene_taxon.P6 P11=gene_taxon.P11
P43=gene_taxon.P43 P8=gene_taxon.P8 P16=gene_taxon.P16
P24=gene_taxon.P24 P30=gene_taxon.P30 P32=gene_taxon.P32
P42=gene_taxon.P42 P26=gene_taxon.P26 P38=gene_taxon.P38
P47=gene_taxon.P47
```

The `-F` option specifies the join feature (gene_taxon) was used. Alternatively, taxa or gene feature can be specified for traditional taxa-based or gene-based analysis.

3) Visualize joint features across all samples

To visualize the joint features across all samples, we used MetaPrism_heatmap.pl script.

```
perl MetaPrism_heatmap.pl P10=gene_taxon.P10 P14=gene_taxon.P14
P23=gene_taxon.P23 P25=gene_taxon.P25 P34=gene_taxon.P34
```

```
P39=gene_taxon.P39 P8=gene_taxon.P8 P16=gene_taxon.P16
P24=gene_taxon.P24 P30=gene_taxon.P30 P32=gene_taxon.P32
P42=gene_taxon.P42 > group.mel.response.vs.progressive.html
```

By default, this script visualized all joint features (all gene and taxa combinations detected in any samples). If a subset of joint features is needed, we can use `-g` option and specific these joint features in a separate file. For example, we plotted the top features with variable importance values greater than 50% in **Figure 3**.

4) Other functions

MetaPrism offer tabularization function to export the joint features in a tabular format. That allows users to utilize for methodology development or further customizations. They can integrate MetaPrism into their existing analysis pipeline. For example, to export data, an example command line is:

```
perl MetaPrism_table.pl P10=gene_taxon.P10 P14=gene_taxon.P14
P23=gene_taxon.P23 P25=gene_taxon.P25 P34=gene_taxon.P34
P39=gene_taxon.P39 P8=gene_taxon.P8 P16=gene_taxon.P16
P24=gene_taxon.P24 P30=gene_taxon.P30 P32=gene_taxon.P32
P42=gene_taxon.P42 > gene_taxon.tsv
```

We provide a full list of MetaPrism function in **Table S2** as well as an online GitHub code repository.

Figure S1: Comparison of gene abundances reported by HUMAnN2. Based on the simulations of 118 bacterial genomes at depth of 10, we compare the estimated gene abundances by HUMAnN2 and the true gene abundance. The Pearson correlation coefficient between true abundances and the HUMAnN2 estimated abundances is 0.827.

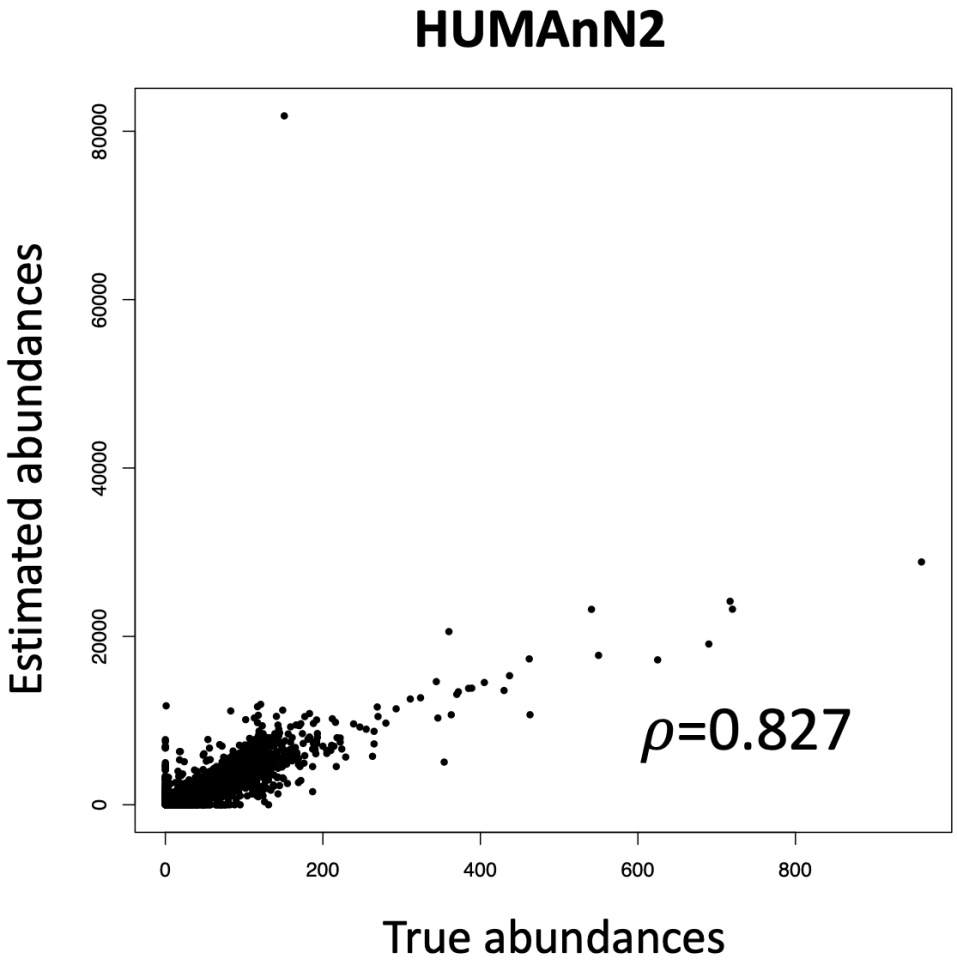

Table S1: Comparison of different metagenomic sequence analysis. We illustrate the goal and representative software to perform taxonomic profiling, functional profiling, and joint profiling.

	Taxonomic profiling	**Functional profiling**	**Joint profiling**
Analysis goal	Classify sequence reads into taxa	Classify sequence reads into genes	Classify sequence reads into both taxa and genes
Example software	MetaPhlAn2 [3] Kraken [4]	HUMAnN2 [5] FMAP [6]	MetaPrism

Table S2: MetaPrism functions. MetaPrism is a toolkit for a wide range of joint analysis. We provide a list of MetaPrism functions and useful command options. That allows in-depth customizations for different research routines. We also provide the latest documentation along the source codes at https://github.com/jiwoongbio/MetaPrism .

1. Prepare database (MetaPrism_gene_prepare.pl)
Options:
-r redownload data
-m FILE executable file path of mapping program, "diamond" or "usearch" [diamond]
-k download prebuilt KEGG files
-a download ARDB database
-b download beta-lactamase database

2. Infer gene abundances (MetaPrism_gene.pl)
Options:
-B input indexed sorted BAM file instead of FASTQ file
-p INT number of threads [1]

3. Infer taxonomy of contigs (MetaPrism_taxon_centrifuge.pl)
Options:
-p INT number of threads [1]

4. Tabularize joint features (MetaPrism_table.pl)

Options:

-A STR abundance column [meanDepth/genome]

-R STR taxon rank [genus]

-F STR feature type, "gene_taxon", "gene", "gene_average", "taxon", "taxon_average"

5. Visualize joint features (MetaPrism_heatmap.pl)

Options:

-A STR abundance column [meanDepth/genome]

-R STR taxon rank [genus]

-F STR feature type, "gene_taxon", "gene", "gene_average", "taxon", "taxon_average"

-g FILE feature file

6. Differential abundance analysis (MetaPrism_comparison.pl)

Options:

-A STR abundance column [meanDepth/genome]

-R STR taxon rank [genus]

-F STR feature type, "gene_taxon", "gene", "gene_average", "taxon", "taxon_average"

-g FILE feature file

7. Prediction model (MetaPrism_prediction.pl)

Options:

-A STR abundance column [meanDepth/genome]

-R STR taxon rank [genus]

-F STR feature type, "gene_taxon", "gene", "gene_average", "taxon", "taxon_average"

-t STR train method [rf]

-c STR train control method [LOOCV]

-s INT seed [1]

References

1. Huang, W., et al., *ART: a next-generation sequencing read simulator.* Bioinformatics, 2012. **28**(4): p. 593-4.
2. Frankel, A.E., et al., *Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients.* Neoplasia, 2017. **19**(10): p. 848-855.
3. Truong, D.T., et al., *MetaPhlAn2 for enhanced metagenomic taxonomic profiling.* Nat Methods, 2015. **12**(10): p. 902-3.
4. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biol, 2014. **15**(3): p. R46.
5. Franzosa, E.A., et al., *Species-level functional profiling of metagenomes and metatranscriptomes.* Nat Methods, 2018. **15**(11): p. 962-968.
6. Kim, J., et al., *FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies.* BMC Bioinformatics, 2016. **17**(1): p. 420.