

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#},

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*}

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

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

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

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

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

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

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

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

⁹Center 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

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 [5]. 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 [5]. The patient identifiers, phenotypes and their NCBI SRA Run information is as follows:

Sample Name	Outcome	SRA Run Information
P10	Response	SRR5930498
		SRR5930499
P14	Response	SRR5930500
		SRR5930499
P23	Response	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.

1.1 Derive patient-specific joint features

Retrieve sequence reads from NCBI SRA, we illustrate the analysis for P8. The analysis needs to be 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

1.2 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.

1.3 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 specify these joint features in a separate file. For example, we plotted the top features with variable importance values greater than 50% in **Supplementary Figure 2**.

1.4 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 **Supplementary Table 2** as well as online GitHub code repository.

Table 1: 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 [6] Kraken [7]	HUMAnN2 [8] FMAP [3]	MetaPrism

Table 2: 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 gene 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]

Reference

1. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator**. *Bioinformatics* 2012, **28**(4):593-594.
2. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA: **metaSPAdes: a new versatile metagenomic assembler**. *Genome research* 2017, **27**(5):824-834.
3. Kim J, Kim MS, Koh AY, Xie Y, Zhan X: **FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies**. *BMC Bioinformatics* 2016, **17**(1):420.
4. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nature methods* 2015, **12**(1):59-60.
5. Frankel AE, Coughlin LA, Kim J, Froehlich TW, Xie Y, Frenkel EP, Koh AY: **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):848-855.
6. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N: **MetaPhlan2 for enhanced metagenomic taxonomic profiling**. *Nature methods* 2015, **12**(10):902-903.
7. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome biology* 2014, **15**(3):R46.
8. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N *et al*: **Species-level functional profiling of metagenomes and metatranscriptomes**. *Nature methods* 2018, **15**(11):962-968.