

# MetaPrism: A versatile toolkit for joint taxa/gene analysis of metagenomic sequencing data

Jiwoong Kim<sup>1,†</sup>, Shuang Jiang<sup>2,†</sup>, Yiqing Wang<sup>2</sup>, Guanghua Xiao<sup>1,3,4</sup>, Yang Xie<sup>1,3,4</sup>, Dajiang J. Liu<sup>5</sup>, Qiwei Li<sup>6</sup>, Andrew Koh<sup>3,7,8</sup>, and Xiaowei Zhan<sup>1,3,9,\*</sup>

<sup>1</sup>Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA

<sup>2</sup>Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, USA

<sup>3</sup>Harold C. Simmons Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

<sup>4</sup>Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

<sup>5</sup>Department of Public Health Sciences, Pennsylvania State University, Hershey, PA, 17033, USA

<sup>6</sup>Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, USA

<sup>7</sup>Department of Microbiology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA

<sup>8</sup>Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

<sup>9</sup>Center for Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding author: xiaowei.zhan@utsouthwestern.edu.

## Abstract

In microbiome research, metagenomic sequencing generates enormous amounts of data. These data are typically classified into taxa for taxonomy analysis, or into genes for functional analysis. However, a joint analysis where the reads are classified into taxa-specific genes is often overlooked. To enable the analysis of this biologically meaningful feature, we developed a novel bioinformatic toolkit, MetaPrism, which can analyze sequence reads for a set of joint taxa/gene analyses to: 1) classify sequence reads and estimate the abundances for taxa-specific genes; 2) tabularize and visualize taxa-specific gene abundances; 3) compare the abundances between groups; and 4) build prediction models for clinical outcome. We illustrated these functions using a published microbiome metagenomics dataset from patients treated with immune checkpoint inhibitor therapy and showed the joint features can serve as potential biomarkers to predict therapeutic responses. MetaPrism is a toolkit for joint taxa and gene analysis. It offers biological insights on the taxa-specific genes on top of the taxa-alone or gene-alone analysis.

MetaPrism is open-source software and freely available at <https://github.com/jiwoongbio/MetaPrism>. The example script to reproduce the manuscript is also provided in the above code repository.

**Keywords:** metagenomics sequence analysis; joint analysis; microbiome biomarker

## Introduction

The human microbiome consists of ~39 trillion bacteria and influences host health (Sender et al. 2016). Recently, the use of metagenomic sequencing has become increasingly popular as a more unbiased approach to gut microbiome profiling as compared to 16S rRNA sequencing. A common approach to comparing differences in the gut microbiome between groups (cases and controls) is to identify significant differences in either taxa or microbial genes. Several popular bioinformatic tools have been developed for this purpose, including MetaPhlAn2 (Truong et al. 2015), Kraken (Wood and Salzberg 2014), HUMAnN2 (Franzosa et al. 2018), and FMAP (Kim et al. 2016b) (Table S1). However, these tools analyze either taxonomic abundances (taxonomic profiling) or gene abundances (function profiling) separately. As each microorganism carries its own genes, taxonomic and functional profiling results are not intrinsically independent. In fact, recent discoveries demonstrated that taxon-specific genes have a

causative role in disease progression and treatment responses. For example, Duan et al. found that a specific *Enterococcus faecalis* carrying the cytotoxin gene promotes alcoholic liver disease (Duan et al. 2019). Simms-Waldrup et al. found that the antibiotic resistance genes in the graft-versus-host-disease patients are enriched for *Klebsiella* (Simms-Waldrup et al. 2017). Therefore, a joint analysis, where taxonomy and functional features are analyzed together, could provide useful biological and clinical insights (Langille 2018). However, bioinformatics tools for joint analyses are lacking.

Our innovation in this manuscript is to define and utilize joint taxa/gene features via bioinformatics approach, with the goal of offering biologically interpretable findings. For example, our method characterizes the genes discovered for each species. This facilitates quantitative analysis of gene abundances in a species-specific manner, which is usually not readily available. Our approach is initiated from *de novo* assembled contigs, which are

Received: December 02, 2020. Accepted: January 28, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

both taxonomically and functionally annotated. Our simulations showed this method could accurately detect bacterial species and their carried genes. In a recent review article (Langille 2018), Langille prompted that understanding the gene contents at species level can offer better interpretation than using the taxon or gene content alone, and potentially improve outcome predictions. This confirmed that the joint feature is useful for general microbiome studies. Our tool provided these joint features as the first step for a wide range of downstream analysis tasks. For example, we demonstrated that the quantity of taxa-specific gene abundances is a potentially useful biomarker to predict the immunotherapy responses.

To facilitate joint analysis, we developed MetaPrism, a novel bioinformatics tool to (1) classify metagenomic sequence reads into both taxa and gene level, (2) normalize the taxa-specific gene abundances within samples, (3) tabularize or visualize these joint features, (4) perform comparative microbiome studies, and (5) build prediction models for clinical outcomes. Using simulated sequence reads, we validated that the performance of MetaPrism is accurate. We further applied the MetaPrism analysis to an immune checkpoint therapy and detected novel joint features as potential biomarkers.

MetaPrism is open-sourced and is available at <https://github.com/jiwoongbio/MetaPrism>. Given the advantages of joint analysis, MetaPrism is a useful tool for a wide range of microbiome-metagenomic sequence studies.

## Materials and Methods

### Analysis workflow

MetaPrism is a toolkit for joint analysis tasks. At its core, MetaPrism will infer the taxa and gene for each metagenome sequence read. One approach is to align each read to bacterial nucleotide reference genomes to obtain its taxonomy and align it to a protein database to obtain its gene functions. However, this approach is technically challenging: due to the short lengths of Illumina sequence reads and the high sequence similarities between bacteria genomes, alignment of short reads is not feasible. We thus developed a novel algorithm (Figure 1A) in an integrated toolkit (Figure 1B) to tackle this challenge.

First, we perform *de novo* assembly for each sample using metaSPAdes (Nurk et al. 2017) with all metagenomic sequence reads to obtain long contigs. These contigs are much longer than sequence reads, which allows for accurate taxonomical and functional profiling.

Second, we identify the taxonomy of these contigs. All the contigs are aligned to a large reference database of more than 4,000 bacterial genomes using centrifuge (Kim et al. 2016a). Ambiguous alignments will be filtered out from the subsequent analysis.

Third, we identify genes and their locations from the contigs. We detect the open reading frames from the contigs, translate the nucleotide bases to amino acids, and align them using DIAMOND (Buchfink et al. 2015) to a protein database. To comprehensively investigate all bacteria genes, either KEGG protein databases that include protein sequences from KEGG orthologue genes (Kanehisa et al. 2012) or KFU (KEGG orthology with UniProt protein sequences) (Kim et al. 2016b), can be utilized. By default, we require minimum coverage of 0.8 to ensure good protein alignments.

Lastly, we calculate and normalize gene abundance within-sample. We align metagenomic sequence reads to the contigs

using BWA (Li and Durbin), and count the number of aligned reads located in the genes of interest. We calculate the read depth normalized by contig length, and this quantity is denoted as mean depth to represent the gene abundances. Larger numbers often indicate higher gene abundance. Other abundance statistics, such as FPKM (Fragment Per Kilobase of transcript per Million reads) or depth per genome (normalized read depth per taxa genome length), are also provided.

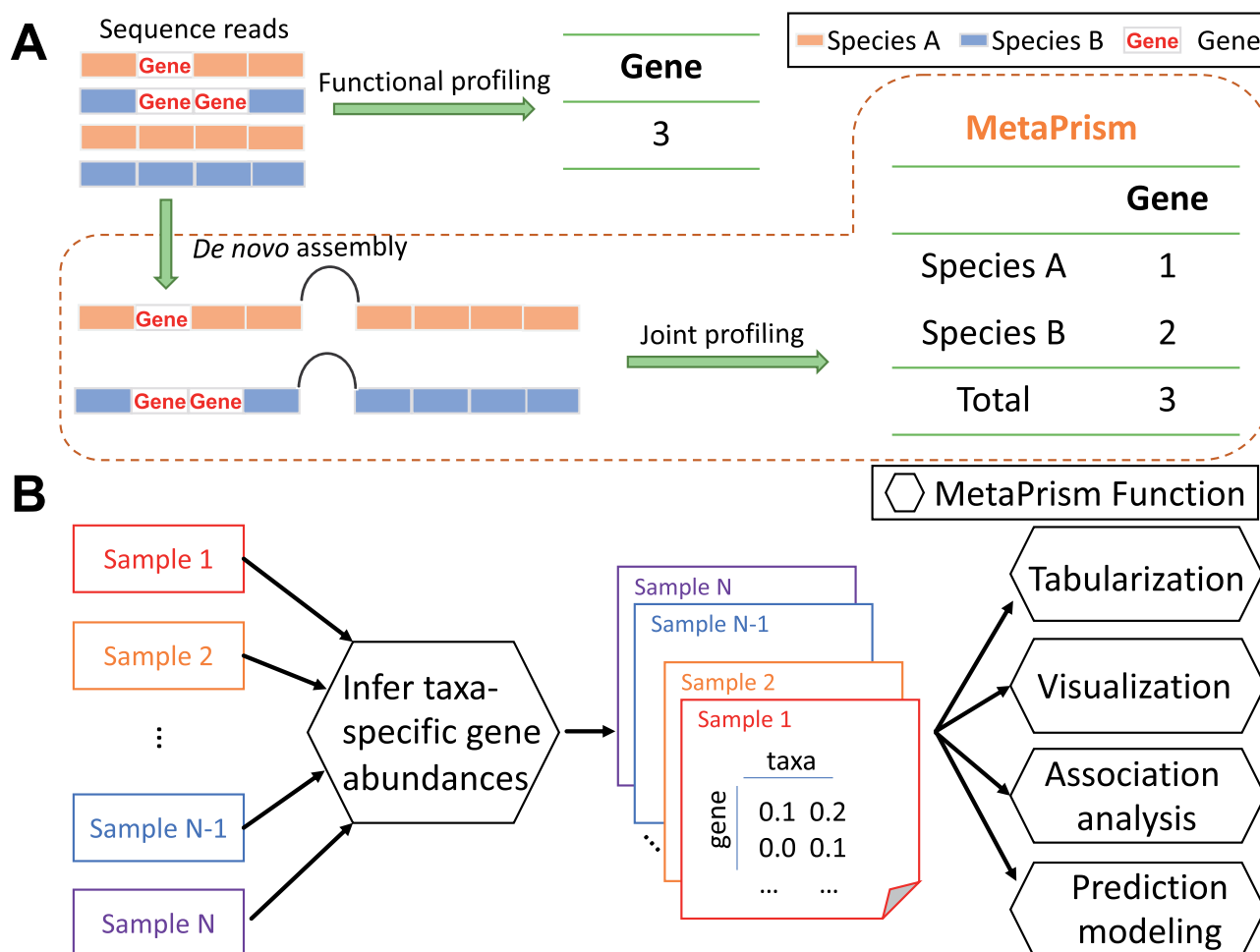
Through the above steps, we obtain the joint feature where the gene abundances are associated with taxonomy information. These features can be viewed as novel microbiome measurements, which provide more information than taxonomic abundances or gene abundances alone. To utilize these features, MetaPrism provides four downstream analysis modules (Figure 1B): (1) tabularization module allows to export the joint features such as the mean depth of genes per contig at the genus level; (2) visualization module allows to visualize the abundances of the joint features in an HTML webpage; (3) differential abundances analysis modules can calculate the fold change of the gene abundances for two groups of samples along with statistical significances in terms of p-values; and (4) prediction module can construct a random forest model or extreme gradient boosting model to detect joint abundance features as potential biomarkers (Breiman 2001; Chen and Guestrin 2016). In all, these modules provide a common set of functions for the typical analysis of joint features. Meanwhile, users are in full control to utilize the exported tabular data for their customized analysis. A list of available functions, command line, and major customization options in MetaPrism are listed in Table S2.

### Simulation setting

To assess the accuracy of the joint features estimated by MetaPrism, we conducted a simulation study using simulated sequence reads from a collection of bacteria species. First, we selected all 118 bacterial species with complete reference genomes where the latest genome collected from June 8, 2018 (Table S3). Then, we downloaded their sequences from NCBI FTP (<http://ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria>). These sequences include 229 contigs including both bacterial chromosomes and plasmids. Their lengths range from 1,308 bp to 10,236,715 bp (mean length is 1,969,971 bp). Finally, we simulated shotgun metagenomic sequencing reads and generated at 10X coverage to resemble typical read lengths from the Illumina using ART (Huang et al. 2012). Specifically, we set read length to be 100 base pair and the mean and standard deviation of the fragment size to be 200 bp and 50 bp, respectively.

### Data analysis for the microbiome in an immune checkpoint therapy

Immune checkpoint therapy is a revolutionary cancer treatment regime. Researchers realize that the gut microbiome plays an indispensable role in modulating the immune system and boost the therapy efficacy (Frankel et al. 2017). We demonstrated a joint analysis using MetaPrism to build a therapy-response prediction model. We collected stool samples of 12 melanoma patients before anti-PD1 (pembrolizumab) therapy and performed metagenomic sequencing (Frankel et al. 2017). Six patients responded to the therapy and six did not. We performed quality-control procedures on the metagenomic sequence reads. That included the removal of human contamination as previously described (Frankel et al. 2017).



**Figure 1. A schematic illustration of the algorithm and the functions in MetaPrism.** A) Illustration of the MetaPrism algorithm to infer taxa-specific gene abundances. While function profiling infers that three reads are mapped to a gene, it cannot provide further taxonomic information. Through joint profiling, MetaPrism can utilize *de novo* assembled contigs to estimate taxa-specific features: two gene copies are from species A and one copy is from species B; B) An overview of the joint analysis workflow in MetaPrism. The hexagon shapes represent implemented functions in MetaPrism.

## Results

### Joint features inferred by MetaPrism are accurate in simulation

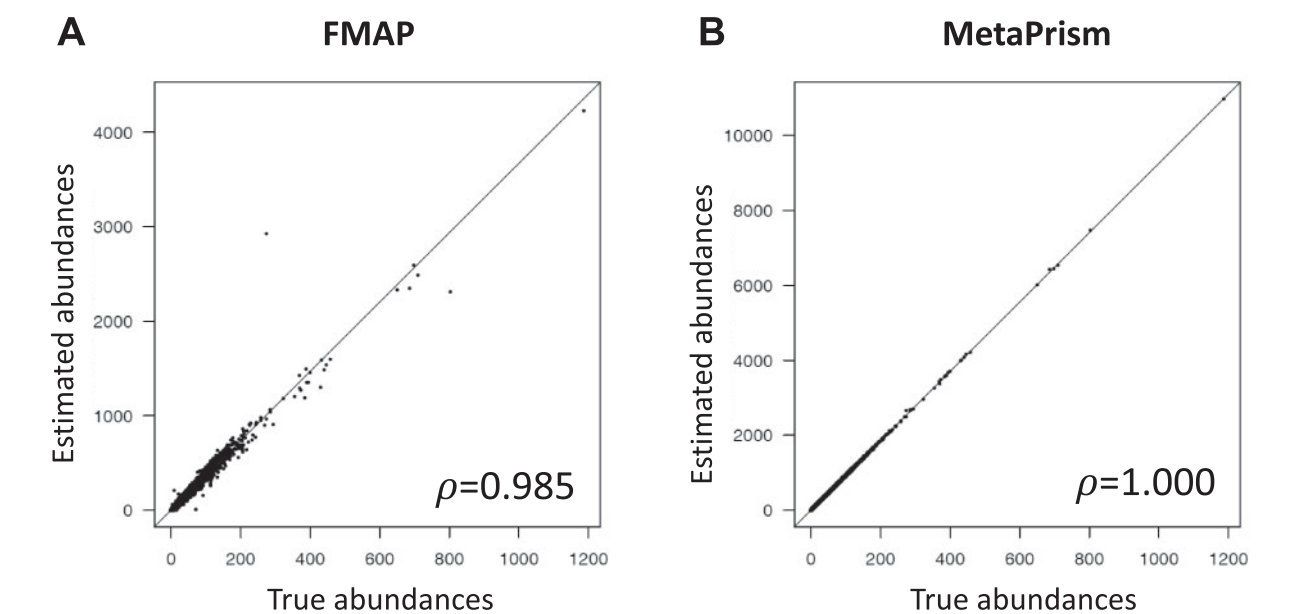
We evaluate the gene abundances calculated by MetaPrism and other methods to the true abundances. We determined the true abundances by the multiplication of sequence depth and the depth of KEGG ortholog (KO) genes in the reference genomes. Notably, there could be more than one copy of KO genes in one contig; thus, the true abundance of KO genes can vary from 0 to 1,200. This is also verified by aligning the gene sequences to the KEGG protein database using DIAMOND (Buchfink et al. 2015).

We use the simulated reads totally  $4.2 \times 10^9$  nucleotide bases. We ran two programs: MetaPrism and FMAP. The FMAP software used translation alignment (BLASTX) and our previous benchmarks showed it can report gene abundances accurately (Kim et al. 2016b). Another popular approach is HUMAnN2. However, our simulation showed that its performance to report KO gene abundances is not accurate (Supplementary: Simulation results using HUMAnN2). In Figure 2, we visualized the true abundances (X-axis) and the estimated abundances (Y-axis) for FMAP and MetaPrism using scatterplots. The correlation coefficient ( $\rho = 1.000$ ) from MetaPrism is higher than that from FMAP ( $\rho = 0.985$ ). In brief, this simulation mimics a metagenomic sequence data from known species. We inferred the gene

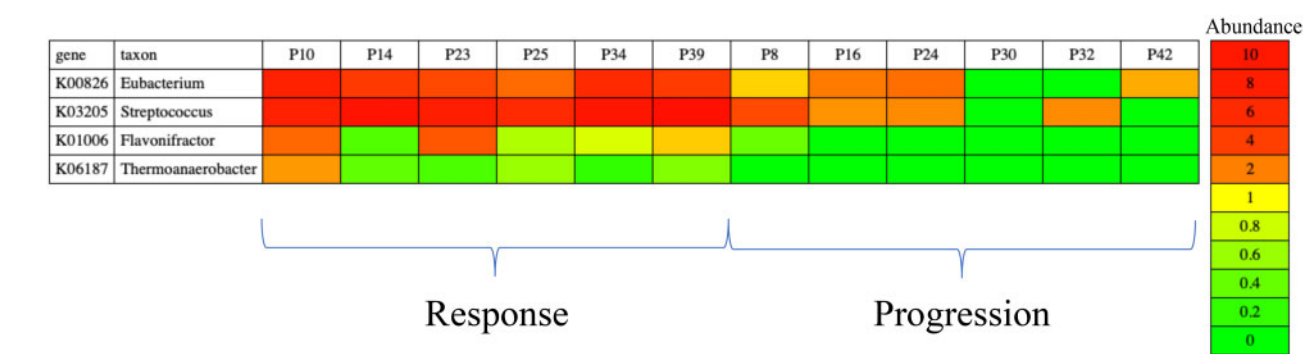
abundances using FMAP (Kim, et al., 2016b) and MetaPrism, and the benchmark showed that gene abundances inferred by MetaPrism were accurate and achieved the highest correlation between inferred abundances and true abundances (Figure 2).

### Joint features can be potential biomarkers in immune checkpoint therapy

We used MetaPrism on the remaining sequence reads (detailed data retrieval, analysis steps, and command lines were available in Supplementary: Discover species-specific biomarker in an immune checkpoint therapy study). On average, each sample has 1.2 billion reads. We profiled sequence in MetaPrism and there are on average 24,532 joint features consisting of 2,058 taxa and 3,432 KO genes per sample. Next, we used MetaPrism to normalize read counts for each sample by reporting the mean depth per assembled contig. As demonstrated in previous simulation, the inferred abundance represents the gene counts of specific taxa. These taxa-specific gene abundances were ranked using a random forest model with 500 trees and leave-one-out cross-validation. This prediction model reached 69% accuracy to predict the immunotherapy responses. It was higher than the accuracy using taxa features alone (54%), gene features alone (62%), or just random guess (50%). The prediction accuracy based on the proposed joint features achieved a 7% lead compared to the second-



**Figure 2. Comparison of gene abundances reported by FMAP and MetaPrism.** We used simulations to compare the estimated gene abundances using FMAP and MetaPrism. The Pearson correlation coefficients between true abundances and the software-estimated abundances were listed on the bottom right.



**Figure 3. Heatmap of joint features for predicting immune checkpoint therapy response.** We used MetaPrism\_heatmap.pl to visualize four joint features (taxa-specific gene abundances, with variable importance values greater than 50%) in the immune checkpoint therapy study. The colors from red to green represent the increased gene abundances, the mean depth normalized by the contig lengths. P10, P14, P23, P25, P34, and P39 are patients who respond to the therapy; P8, P16, P24, P30, P32, and P42 are patients having progressive outcomes. K00826, branched-chain amino acid aminotransferase; K03205, type IV secretion system protein VirD4; K01006, pyruvate, orthophosphate dikinase; K06187, recombination protein RecR.

best model where gene features were used. Furthermore, it detected four joint features with variable importance greater than 50%. We examined the abundances these abundances with red to green colors representing the depth values (Figure 3). We observed these joint features are more abundant in the responder group suggesting that they may improve the treatment efficacy. Among them, the most important feature is the K00826 gene (branched-chain amino acid aminotransferase, BCAT1) from the genus *Eubacterium* (Table 1). The average abundance of this joint feature in the response group is three-fold higher compared to that in the progression group (response = 4.70 vs progression = 1.17). Interestingly, BCAT1, as an important enzyme in branched-chain amino acid, is associated with glycolysis and oxygen consumption (Kelly and Pearce 2020). These biological procedures determine the cancer growth (Bertout et al. 2008; Yttersian Sletta et al. 2017), and they may be interfered by the high activity of BCAT of *Eubacterium*, the top abundant taxon in this dataset (20.9%). Although alteration of BCAA metabolism from the

bacterial contributes to creating a tumor-favoring metabolic condition in the host remains a hypothesis, further mechanistic studies may investigate the K00826 genes from *Eubacterium* as a biomarker for cancer immune checkpoint therapy.

In terms of computation, all the above analyses can be accomplished on a standard computation cluster (e.g., 128 GB memory with 2 GB hard drive space per sample).

### Discussion

We present a novel bioinformatics tool, MetaPrism. It implements functions to quantify the joint features (both taxonomic and functional) from metagenomic sequence reads, as well as other functions for downstream data analyses including comparative studies and prediction modeling. We demonstrate that the joint features can provide novel insights to understand the microbial role in a cancer immunotherapy study.



**Table 1. Prediction models and performances for taxonomical analysis, functional analysis, and joint analysis.** We tabularized the details of prediction models used in three types of analyses and their prediction performances.

	Taxonomic profiling	Functional profiling	Joint profiling
<b>Model</b>	Random forest	Random forest	Random forest
<b>Number of trees</b>	500	500	500
<b>Number of features</b>	1,048	5,227	62,086
<b>Top features (if variable importance &gt; 50%)<sup>#</sup></b>			
<b>1<sup>st</sup> feature</b>	Chondromyces (100)	K07705 (100)	K00826 <i>Eubacterium</i> (100)
<b>2<sup>nd</sup> feature</b>	Roseateles (65)	–	K03205 <i>Streptococcus</i> (89)
<b>3<sup>rd</sup> feature</b>	–	–	K01006 <i>Flavonifractor</i> (81)
<b>4<sup>th</sup> feature</b>	–	–	K06187 <i>Thermoanaerobacter</i> (74)
<b>Accuracy<sup>*</sup></b>	53.8%	61.5%	69.2%

<sup>#</sup> : The variable importance values are listed in parentheses.

<sup>\*</sup> : Prediction accuracy was evaluated using leave-one-out cross-validations.

MetaPrism is flexible and can be customized. For example, we can prepare a specific gene database to investigate taxa-specific antibiotic resistance genes (ARGs). We have used reference protein databases with ARGs, such as ARDB (Liu and Pop 2009) or CARD (McArthur et al. 2013). In a graft-versus-host disease (GVHD) study, we used MetaPrism with the ARDB to infer taxa-specific ARGs for joint resistome profiling. Then we correlated patients' resistome to the outcome of GVHD. We found increased abundances of antibiotic-resistance genes (e.g., *mdtG*, *AcrA*, *AcrB*, and *TolC*) in *Klebsiella* and *E. coli* in the GVHD patients compared with the abundances in non-GVHD patients. This finding may hint optimal antibiotic prescription for better management of GVHD.

MetaPrism characterizes the joint features based on the contigs that are *de novo* assembled from metagenomic sequence reads. This is a distinct feature compared with other software. For example, HUMAnN2 used a tiered search strategy that relied on a curated reference database for organism-specific genes (Franzosa et al. 2018). However, many bacterial genes are shared across organisms and can be missed by the organism-specific gene database. Thus, we designed the MetaPrism to reduce the dependency on curated reference databases. The tradeoff for this decision is that MetaPrism requires more computational resources for the *de novo* assembling step.

In all, MetaPrism is free and useful software to facilitate joint analyses and it is suitable for general microbiome studies. Researchers can expect MetaPrism to quantify species-specific gene abundances and use these interpretable features in association studies and prediction tasks.

## Data Availability

The metagenomic shotgun sequence dataset in the immune checkpoint therapy is available from the NCBI BioProject PRJNA397906. The treatment responses for the 12 patients as well as the analysis codes were available in the **Supplementary: Discover species-specific biomarker in an immune checkpoint therapy study**. The source codes of MetaPrism software are available at: <https://github.com/jiwoongbio/MetaPrism>. That resource contains the software requirements, usage example, and documentations for all MetaPrism components (e.g., download bacterial database, quantify species-specific gene abundances, build association models and prediction models, tabularize results, and visualize results in heatmap plots). Supplemental material is available at figshare DOI: <https://doi.org/10.25387/g3.13944521>.

## Acknowledgements

We thank Jessie Norris for her comments on the manuscript.

## Funding

This work has been supported by the following grants: National Institutes of Health R01 [R01GM115473 (YX), R01GM126479 (DJL, XZ), R56HG011035 (DJL, XZ)]; Cancer Center: [P30CA142543 (YX, XZ)]; Specialized Programs of Research Excellence [P50CA070907 (YX, XZ)].

**Conflicts of Interest:** The authors declare that they have no competing interests.

## Authors' contributions

JK, SJ, and XZ conceived of the project and wrote the first draft of the manuscript. GX, YX, AY, and XZ coordinated and oversaw the study. QL and DL provided critical inputs for the study. JK and SJ developed the software and associated databases. JK, SJ, and YW perform statistical analysis. All authors contributed to the review of the manuscript before submission for publication. All authors read and approved the final manuscript.

## References

- Bertout JA, Patel SA, Simon MC. 2008. The impact of o2 availability on human cancer. *Nat Rev Cancer*. 8:967–975.
- Breiman L. 2001. Random forests. *Machine Learning*. 45:5–32.
- Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using diamond. *Nat Methods*. 12:59–60.
- Chen T, Carlos G. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*; ACM.
- Duan Y, Llorente C, Lang S, Brandl K, Chu H, et al. 2019. Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. *Nature*. 575:505–511.
- Frankel AE, Coughlin LA, Kim J, Froehlich TW, Xie Y, et al. 2017. Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients. *Neoplasia*. 19:848–855.
- Franzosa EA, McIver LJ, Rahnnavard G, Thompson LR, Schirmer M, et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. 15:962–968.
- Huang W, Li L, Myers JR, Marth GT. 2012. Art: a next-generation sequencing read simulator. *Bioinformatics*. 28:593–594.

- Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*. 40:D109–114.
- Kelly B, Pearce EL. 2020. Amino assets: how amino acids support immunity. *Cell Metab*. 32:154–175.
- Kim D, Song L, Breitwieser FP, Salzberg SL. 2016a. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res*. 26:1721–1729.
- Kim J, Kim MS, Koh AY, Xie Y, Zhan X. 2016b. Fmap: functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics*. 17:420.
- Langille MGI. 2018. Exploring linkages between taxonomic and functional profiles of the human microbiome. *mSystems*. e00163-17.3.
- Liu B, Pop M. 2009. Ardb-antibiotic resistance genes database. *Nucleic Acids Research*. 37:D443–447.
- McArthur AG, Wagglechner N, Nizam F, Yan A, Azad MA, et al. 2013. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother*. 57:3348–3357.
- Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. Metaspades: a new versatile metagenomic assembler. *Genome Res*. 27: 824–834.
- Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*. 14: e1002533.
- Simms-Waldrup TR, Sunkersett G, Coughlin LA, Savani MR, Arana C, et al. 2017. Antibiotic-induced depletion of anti-inflammatory clostridia is associated with the development of graft-versus-host disease in pediatric stem cell transplantation patients. *Biol Blood Marrow Transplant*. 23:820–829.
- Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, et al. 2015. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nat Methods*. 12:902–903.
- Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 15: R46.
- Yttersian Sletta K, Tveitaras MK, Lu N, Engelsen AST, Reed RK, et al. 2017. Oxygen-dependent regulation of tumor growth and metastasis in human breast cancer xenografts. *PLoS One*. 12:e0183254.

Communicating editor: M. S. Sachs