Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Sequence analysis

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

Jiwoong Kim[1#], Shuang Jiang[2#], Guanghua Xiao[1,3,4], Yang Xie[1,3,4], Andrew Koh[3,5,6], Xiaowei Zhan[1,3,7*]

[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, [4]Department of Bioinformatics, [5]Department of Microbiology, [6]Department of Pediatrics, [7]Center for Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, Texas, 75390, USA.

*To whom correspondence should be addressed;

#These authors contributed equally to this work;

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: 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: 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 illustrate 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 freely available at https://github.com/jiwoongbio/MetaPrism.

Contact: Xiaowei.Zhan@UTSouthwestern.edu

1 Introduction

The human microbiome consists of ~39 trillion bacteria and influences host health. 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 (**Supplementary Table 1**). 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. Therefore, joint analysis, where taxonomy and functional features are analyzed together, could provide useful biological and clinical insights (Langille, 2018). However, software tools for joint analyses are comparatively lacking.

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-abundances within samples, (3) tabularize or visualize these joint features, (4) perform comparative microbiome studies, and (5) build prediction models for clinical outcomes. 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 microbiome sequence studies.

2 Methods

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 technical 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**) 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. As these contigs are much longer than sequence reads, that 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 subsequent analysis.

Third, we identify genes and their locations from the contigs. We detect the open reading frames from the contigs, translated the nucleotide bases to amino acids, and aligned 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., 2016a), can be utilized. By default, we required a 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, the gene abundances are associated with taxonomy information. To assess the accuracy of these estimations, we conducted a simulation study (detailed in **Supplementary 1**). In brief, we simulated metagenomic sequence reads from known species and inferred the gene abundances using FMAP (Kim, et al., 2016b) and MetaPrism. This benchmark showed that gene abundances inferred by MetaPrism were accurate and achieved the highest correlation between inferred abundances and true abundances (**Supplementary Figure 1**).

Based on these joint features, MetaPrism provided the following downstream joint analysis functions (demonstrated in **Figure 1B**): 1) tabularize the abundances of these features (MetaPrism_table.pl); 2) visualize the features in heatmaps (MetaPrism_heatmap.pl); 3) compare the taxa-specific genes abundances across different experimental conditions such as case-control studies (MetaPrism_compare.pl); 4) indicate which features may serve as potential biomarkers in a prediction model (MetaPrism_predict.pl). A list of available functions, command line and major customization options in MetaPrism are listed in **Supplementary Table 2**.

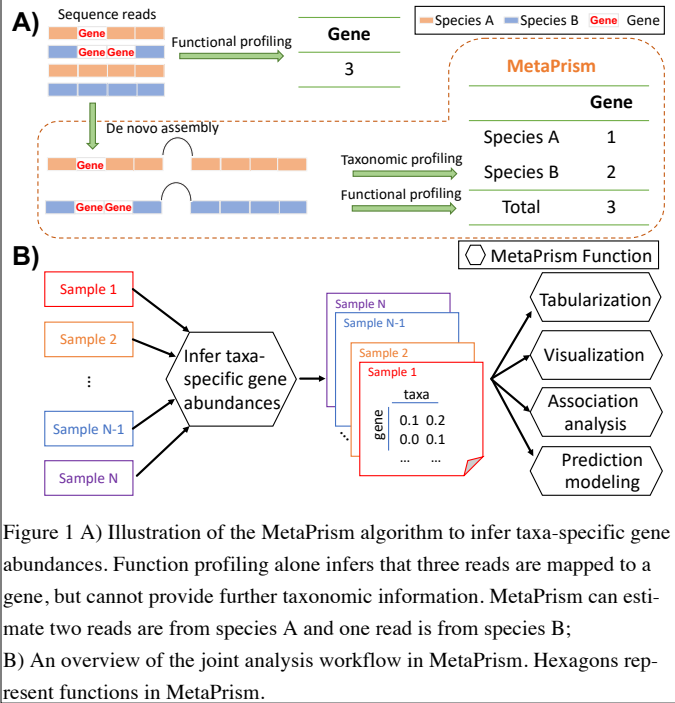

Figure 1 A) Illustration of the MetaPrism algorithm to infer taxa-specific gene abundances. Function profiling alone infers that three reads are mapped to a gene, but cannot provide further taxonomic information. MetaPrism can estimate two reads are from species A and one read is from species B;
B) An overview of the joint analysis workflow in MetaPrism. Hexagons represent functions in MetaPrism.

3 Application

The gut microbiome plays an important role in modulating immune checkpoint therapy (Frankel, et al., 2017). Here 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 (pembrolizomab) therapy , and performed metagenomic sequencing. 6 patients responded to the therapy; 6 did not.

Starting from metagenome sequence reads, we performed quality control, including removal human contamination. Then all remaining sequence reads were processed in MetaPrism (**Supplementary 2**). On average, MetaPrism inferred the taxonomy and gene features for 1.2 billion reads per sample. Next, MetaPrism normalized the reads within samples by reporting the mean depth per assembled contig. The taxa-specific gene abundances were ranked using a random forest model with leave-one-out cross validation. This prediction model reached 69% accuracy, which is higher than the accuracy using taxa features alone (54%) or gene features alone (62%). Furthermore, it detected four joint features with variable importance greater than 50%. MetaPrism visualized these abundances with red to green colors representing the depth values (**Supplementary Figure 2**). The most important feature is the K00826 gene (branched-chain amino acid aminotransferase) from the genus Eubacterium (**Supplementary Table 3**). This is a novel joint feature that may serve as a potential biomarker for cancer therapy.

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

4 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. We demonstrate that the joint features can provide novel insights to understand the microbial role in a cancer immunotherapy study. It is noteworthy that our tool is flexible and can be customized. For example, to study species-specific antibiotic resistance genes (ARGs), a reference protein database with ARGs, such as ARDB (Liu and Pop, 2009) or CARD (McArthur, et al.), can be used. MetaPrism can infer taxa-specific ARGs, thus enabling joint resistome profiling. In all, MetaPrism is free software to facilitate joint analyses and is suitable for general microbiome studies.

Funding

This work has been supported by the following grants: NIH R01 [R01GM115473 (YX), R01GM126479 (XZ)]; Cancer Center: [P30CA142543 (YX, XZ)]; Specialized Programs of Research Excellence [P50CA070907 (YX, XZ)].
Conflict of Interest: None declared.

References

Buchfink, B., Xie, C. and Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12(1):59-60.

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):848-855.

Kanehisa, M., *et al.* KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;40(Database issue):D109-114.

Kim, D., *et al.* Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016a;26(12):1721-1729.

Kim, J., *et al.* FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics* 2016b;17(1):420.

Langille, M.G.I. Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. *mSystems* 2018;3(2).

Li, H. and Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*;26(5):589-595.

Liu, B. and Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res* 2009;37(Database issue):D443-447.

McArthur, A.G., *et al.* The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013;57(7):3348-3357.

Nurk, S., *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;27(5):824-834.