Bioinformatics, 2019, 1–4
doi: 10.1093/bioinformatics/btz414
Advance Access Publication Date: 22 May 2019
Applications Note

Genome expression

MetaQUBIC: a computational pipeline for gene-level functional profiling of metagenome and metatranscriptome

Anjun Ma[1,†], Minxuan Sun[2,†], Adam McDermaid[1,3], Bingqiang Liu[4] and Qin Ma[1,*]

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA, [2]Department of Computer Science, [3]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57006, USA and [4]School of Mathematics, Shandong University, Jinan 250100, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on December 5, 2018; revised on April 15, 2019; editorial decision on May 11, 2019; accepted on May 14, 2019

Abstract

Motivation: Metagenomic and metatranscriptomic analyses can provide an abundance of information related to microbial communities. However, straightforward analysis of this data does not provide optimal results, with a required integration of data types being needed to thoroughly investigate these microbiomes and their environmental interactions.
Results: Here, we present MetaQUBIC, an integrated biclustering-based computational pipeline for gene module detection that integrates both metagenomic and metatranscriptomic data. Additionally, we used this pipeline to investigate 735 paired DNA and RNA human gut microbiome samples, resulting in a comprehensive hybrid gene expression matrix of 2.3 million cross-species genes in the 735 human fecal samples and 155 functional enriched gene modules. We believe both the MetaQUBIC pipeline and the generated comprehensive human gut hybrid expression matrix will facilitate further investigations into multiple levels of microbiome studies.
Availability and implementation: The package is freely available at https://github.com/OSU-BMBL/metaqubic.
Contact: qin.ma@osumc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The metagenomic and metatranscriptomic analyses can provide genetic composition and expression information of microbial communities, respectively, collected from environmental samples (Franzosa *et al.*, 2018). In human, the gut microbiome has been discovered to be associated with inflammatory bowel disease, diabetes, obesity and cancer, among other diseases (Huttenhower *et al.*, 2012). Currently, much emphasis has been placed on 16S rRNA analysis and diversity analysis at species or strain level, yet, fewer studies have been carried out for gene functions and regulatory pathways due to the high heterogeneity of data (Niu *et al.*, 2017). While it is necessary to understand the underlying species composition for a microbiome, that alone is not sufficient to gain a comprehensive understanding of this environment. Hence, integrating metatranscriptomic analyses—which provide insight into the entire gene expression of the microbiome—along with metagenomic information, i.e. species/strain composition, provides a more robust method for understanding the mechanisms of a microbial community.

Integrating these levels of information requires paired DNA and RNA data that represent the samples at simultaneous time points and identical conditions. Often of the primary interest is the detection of modules from the gene expressions. Module detection groups

or partitions sets of samples or genes (or both) to generate relatively homogenous sets, in which within-set genes/samples are distinctly more similar than those external to the set. Biclustering is a powerful data mining technique that allows clustering of rows (genes) and columns (samples), simultaneously, in any quantitative matrix-format dataset. Thus, this method is highly useful for predicting condition-specific gene modules for co-expression gene analysis. However, currently, no biclustering tool has been developed specifically for metatranscriptomic data which generally contains genes 10- to 100-fold larger and sparser than the bulk or single-cell RNA-Seq data.

On this basis, we developed MetaQUBIC, an efficient and compatible biclustering-based module detection pipeline for metagenomic and metatranscriptomic data. MetaQUBIC takes paired DNA and RNA fastq files as input and generates conditional-specific gene modules. To showcase the effectiveness of MetaQUBIC, we found 735 paired DNA and RNA data from all the 2598 samples under BioProject accession PRJNA354235 (Abu-Ali *et al.*, 2018), and aligned them to the IGC database (Li *et al.*, 2014), resulting in 200 biclusters. The biclusters were evaluated by Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment test and showed the best performance compared with other four biclustering tools: FABIA (Hochreiter *et al.*, 2010), ISA (Ihmels *et al.*, 2004), Bimax (Prelić *et al.*, 2006) and Plaid (Lazzeroni and Owen, 2002). Additionally, the biclustering method in MetaQUBIC is considerably efficient considering the high computational complexity of extracting gene modules from the extremely large filtered hybrid gene expression matrix (hGEM) (2 365 438 genes ×735 samples). Moreover, such hGEM can benefit users by saving the time of alignment processes for the 735 datasets. They can either directly combine our hGEM with their new human gut microbial data to increase the analysis pool, or independently use the hGEM for other further analyses, such as differential expression.

2 Methods and results

In 2014, Li *et al.* released the IGC comprised of 9 879 896 genes and 1257 samples to create a cohort that is 3-fold larger than cohorts used for other gene catalogs (Li *et al.*, 2014). The IGC database includes close-to-complete sets of genes for most human gut microbes, thus can be prominently considered as a pan-metagenome—which we define as an entire gene set of all strains and species of various species existing in the same environment—and used for gut microbial gene mapping. The 735 paired metagenome and metatranscriptome datasets from human fecal samples in the Health Professionals Follow-Up Study were provided by Abu-Ali *et al.* (2018). The general MetaQUBIC pipeline is composed by three main steps: (i) gene mapping and catalog construction; (ii) hGEM generation and filtering and (iii) biclustering and enrichment analyses (Fig. 1A–C).

The paired metagenomic and metatranscriptomic fastq files of 735 samples were first aligned to the IGC pan-metagenome reference using Bowtie2 (Langmead and Salzberg, 2012). Samtools (Li *et al.*, 2009b) was used to convert file format and Bedtools (Quinlan and Hall, 2010) was used to generate a matrix of read counts and coverage over each gene in both DNA and RNA (Fig. 1A). The two mapping results from the same sample were merged to create a sample catalog file (Fig. 1D) containing gene names and nine more columns recording mapping information, including the calculated RD_{RPK} normalized expression values (Supplementary Method S1). A full list of tools integrated into MetaQUBIC can be found in Supplementary Table S1. MetaQUBIC also supports self-provided

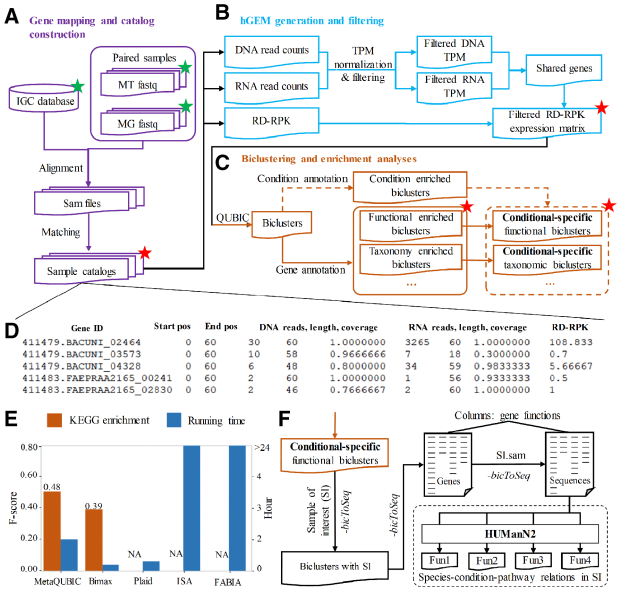

Fig. 1. MetaQUBIC workflow and tool comparison. The workflow consists of three main steps: (**A**) step 1: gene mapping and catalog construction, (**B**) step 2: hGEM generation and filtering and (**C**) step 3: biclustering and enrichment analysis. Green stars indicate user's input and red stars indicate important outputs. Dash lines indicate condition annotation required steps. (**D**) Example of sample catalog. (**E**) Tool comparison of KEGG functional enrichment analysis and running time for analyzing the human gut hGEM. (**F**) Extension application of MetaQUBIC to HUMannN2. Abbreviations: metatranscriptome (MT), metagenome (MG)

reference databases by editing the pathway in the configuration file to broaden its application to metatranscriptomic researchers focusing on environments other than human gut (Supplementary Tutorial). Unlike traditional metatranscriptomic analysis tools that require a command line for every single sample, MetaQUBIC allows for users to provide the location of a root folder containing all sequencing samples to automatically perform gene mapping of each sample, greatly expediting and simplifying the analysis pipeline. Additionally, MetaQUBIC can even perform parallel mapping using multi-cores. A full description of parameters' options for MetaQUBIC can be found in Supplementary Table S2 and Supplementary Tutorial.

Three hGEM (DNA read counts, RNA read counts and RD_{RPK} normalized values) are generated by extracting and merging the corresponding columns from all sample catalogs. The DNA and RNA hGEM are then normalized by transcripts per kilobase million (TPM) (Supplementary Method S1), and genes are preserved if they demonstrate sufficient expression, in our case TPM > 1 in at least two samples. Genes remaining in both filtered DNA and RNA TPM-normalized hGEM are used to filter the RD_{RPK} hGEM, resulting in a filtered expression matrix with 2.3 million genes ×735 samples (Fig. 1B). Sample catalogs generated from users' new metatranscriptomic data regarding to the human gut microbial research can be easily integrated with the 735 provided catalogs to readily increase the sample pool (which could save a large amount of time for gene mapping for 735 datasets). Details for integration of user data can be found in Supplementary Tutorial.

To detect modules from the expression matrix, our in-house QUBIC biclustering tool was implanted to predict biclusters consisting of functional genes sharing similar expression patterns among a group of samples (Li *et al.*, 2009a; Zhang *et al.*, 2016) (Fig. 1C).

Considering the large number of genes, we further modified the QUBIC algorithm by increasing the seed generation threshold, which tremendously decreases the running time without losing accuracy (Supplementary Method S2). Two hundred biclusters were obtained from the filtered RD_{RPK} expression matrix using default parameters.

The functional enrichment analysis of the identified biclusters was performed by the SciPy package (Jones *et al.*, 2014) using the KEGG functional annotations provided in the IGC database (Fig. 1C). The most enriched pathway was chosen for each bicluster and its corresponding *P*-value was assigned to this bicluster. Then the selected *P*-values were adjusted by False Discovery Rate (Supplementary Method S3), giving rise to biclusters that are significantly enrich with specific annotated pathways (adjusted *P*-value <0.05). The MetaQUBIC pipeline allows users to perform other enrichment tests, as long as the annotation file is provided. Additionally, with sample information provided, such as diseases and treatments, a condition enrichment analysis using the same hypergeometric test can be performed to link each significant bicluster to the conditions and become a condition-specific bicluster (as such information was not provided by Abu-Ali *et al.*, we omitted the step here) (Fig. 1C).

3 Performance and comparison

The complete MetaQUBIC code with explanations and examples is shown in Supplementary Table S3. Run time of the biclustering step in MetaQUBIC was compared with the other four high-performing biclustering tools: FABIA (Hochreiter *et al.*, 2010), ISA (Ihmels *et al.*, 2004), Bimax (Prelić *et al.*, 2006) and Plaid (Lazzeroni and Owen, 2002), on XSEDE (Towns *et al.*, 2014) clusters with memory usage set to 512 GB. All tools were tested and compared on the basis of the filtered RD_{RPK} hGEM with 2 365 438 genes (Fig. 1E; Supplementary Tables S3 and S4). The result showed that MetaQUBIC spent 2 h to identify 200 biclusters with genes spanning from 146 to 374 (average 207). Among the 200 biclusters, 155 were significantly enriched with at least one KEGG functional category (32 categories, annotated by IGC database) with evaluation scores of 0.48 (*F*1-score), 0.77 (precision), 0.34 (recall) and 0.1325 (purity) (Supplementary Method S3). On the contrary, Bimax identified 200 biclusters in 8 min with genes spanning from 932 to 31 435 (average 4011). All biclusters identified by Bimax are significantly enriched to at least one pathway, however, its application power is limited by the relative low evaluation scores of 0.39 (*F*1-score), 1 (precision), 0.25 (recall) and 0.0336 (average purity). Additionally, based on our previous research, Bimax, though quick, showed significantly high variations in biclustering when using different settings or different datasets (Xie *et al.*, 2018). Plaid took 16 min to complete the running, while no bicluster was produced from the large hGEM due to its size and complex structure. ISA and FABIA were terminated after 24 h without the program completion. As a result, we evaluated the biclustering function in MetaQUBIC to be considerably accurate, efficient and controllable for module detection from the metagenome and metatranscriptome data.

4 Conclusion and discussion

Comprehensive understanding of microbiomes and their relation to hosts and environments is integral to understanding numerous biological settings. Investigations into these microbial communities rely on metagenomic and metatranscriptomic data, which should be used simultaneously to provide a more thorough undertaking. Especially when combined with comprehensive phenotype information for samples, MetaQUBIC can comprehensively provide functional insight to microbial communities in the form of gene modules. Furthermore, MetaQUBIC can be enhanced to facilitate compatibility with other high-performance methods, such as the case with HUMannN2 (Franzosa *et al.*, 2018) (Fig. 1F, Supplementary Method S4).

The flexibility of MetaQUBIC allows different implementations for analysis, including functionalities that allow researchers to add their human gut microbial data to our existing dataset for further investigation of a more robust set of meta-data using MetaQUBIC. Additionally, users can modify the current MetaQUBIC pipeline to integrate their own unique paired DNA and RNA data representing an environment of their choice, when paired with an appropriate database or self-provided inter-species reference. With the imminent need for large amount sequencing data processes, it is foreseeable that the mapping data, both input and output, can be extremely huge (e.g. nanopore can produce TBs of data per sample). MetaQUBIC takes the advantages in automatically performing the alignment steps over all samples in parallel to free users from laborious and error-prone work, and most importantly, summarizes tons of data into a single hGEM matrix reserving both genetic and transcriptomic information. The generated large-scale hGEM can also be used by researchers for a variety of purposes, beyond that of what was explored in this study, such as using MetaMap to screen existing human RNA-seq datasets for the presence of microbial and viral reads by re-inspecting the non-human-mapping read fraction (Simon *et al.*, 2018).

Acknowledgements

The authors would like to thank Prof. Yu Zhang from the Colleges of Computer Science and Technology at Jilin University, China, for his insightful discussions and valuable comments on the QUBIC algorithm.

Funding

The project described was supported by Award Number Grant [UL1TR002733] from the National Center for Advancing Translational Sciences. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation [grant number ACI-1548562]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center For Advancing Translational Sciences or the National Institutes of Health.

Conflict of Interest: none declared.

References

Abu-Ali,G.S. *et al.* (2018) Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat. Microbiol.*, **3**, 356–366.

Franzosa,E.A. *et al.* (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods*, **15**, 962–968.

Hochreiter,S. *et al.* (2010) FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, **26**, 1520–1527.

Huttenhower,C. *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207.

Ihmels,J. *et al.* (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Jones,E. *et al.* (2014) {SciPy}: open source scientific tools for {Python} 2001, http://www.scipy.org/ (23 May 2019, date last accessed).

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Lazzeroni,L. and Owen,A. (2002) Plaid models for gene expression data. *Stat. Sin.*, **12**, 61–86.

Li,G. *et al*. (2009a) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.*, **37**, e101.

Li,H. *et al*. (2009b) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,J. *et al*. (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, **32**, 834–841.

Niu,S.-Y. *et al*. (2017) Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief. Bioinform.*, **1**, 15.

Prelić,A. *et al*. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Simon,L.M. *et al*. (2018) MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. *Gigascience*, **7**, giy070.

Towns,J. *et al*. (2014) XSEDE: accelerating scientific discovery. *Comput. Sci. Engineer.*, **16**, 62–74.

Xie,J. *et al*. (2018) QUBIC2: a novel biclustering algorithm for large-scale bulk RNA-sequencing and single-cell RNA-sequencing data analysis. *bioRxiv*, 409961.

Zhang,Y. *et al*. (2016) QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics*, **33**, 450–452.