Beginner's guide to microbiome analysis

Bioinformatics guidelines and practical concepts for amplicon-based microbiome analysis.

PICHAHPUK UTHAIPAISANWONG
Systems Biology and Bioinformatics Research Group, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi

PANTAKAN PUENGRANG
Systems Biology and Bioinformatics Research Group, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi

CHALIDA RANGSIWUTISAK
Systems Biology and Bioinformatics Research Group, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi

PHOTCHANATHORN PROMBUN
Bioinformatics and Systems Biology Program, School of Bioresources and Technology and School of Information Technology, King Mongkut's University of Technology Thonburi

ATHISRI SITTHIPUNYA
Bioinformatics and Systems Biology Program, School of Bioresources and Technology and School of Information Technology, King Mongkut's University of Technology Thonburi

NATCHAPHON RAJUDOM
Bioinformatics and Systems Biology Program, School of Bioresources and Technology and School of Information Technology, King Mongkut's University of Technology Thonburi

KANTHIDA KUSONMANO*
Bioinformatics and Systems Biology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi

ABSTRACT

The advent of next-generation sequencing (NGS) allows to study living organisms by reading genetic materials in a high-throughput manner. The technology has opened up a field of microbial research in several areas such as medicine, agriculture, energy, and environment, to study a whole microbial community in an environment of interest without culturing. Bioinformatics analysis is a need in order to characterize and analyze microbiota in the studied samples. In this tutorial, we will give an overview of microbiome analysis based on high-throughput 16S rRNA genes sequencing, a commonly-used target sequence to classify bacteria and archaea. With biological and technology backgrounds, microbiome data from short-read sequencing platform will be elucidated followed by all important computational steps for microbiome analysis. The steps include data preprocessing, amplicon sequence variant analysis, taxonomy assignment, data normalization, and diversity analyses. Practical concepts and codes for the microbiome analysis will be demonstrated step by step providing a basic guideline for beginner.

CCS CONCEPTS

• **Applied Computing**; • **Life and medical sciences**; • **Bioinformatics**;

KEYWORDS

Microbiome analysis, Bioinformatics, 16S rRNA gene sequencing, Biological high-throughput data

ACM Reference Format:
PICHAHPUK UTHAIPAISANWONG, PANTAKAN PUENGRANG, CHALIDA RANGSIWUTISAK, PHOTCHANATHORN PROMBUN, ATHISRI SITTHIPUNYA, NATCHAPHON RAJUDOM, and KANTHIDA KUSONMANO*. 2020. Beginner's guide to microbiome analysis: Bioinformatics guidelines and practical concepts for amplicon-based microbiome analysis.. In *CSBio '20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics (CSBio2020), November 19–21, 2020, Bangkok, Thailand*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3429210.3429211

1 INTRODUCTION

The advent of next-generation sequencing (NGS) allows to study living organisms by reading genetic materials in a high-throughput manner. The technology has dramatically changed a field of microbial research as genetic materials of microorganisms could be sequenced directly from an environment without cultivation. Thus, we can study a whole microbial community in a sample of interest. Beforehand, only a few percent of microbes could be studied based on culturing or conventional method. This gives a vivid picture of microbes living together in a particular environment such as

CSBio2020, November 19–21, 2020, Bangkok, Thailand
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8823-8/20/11…$15.00
https://doi.org/10.1145/3429210.3429211

hosted by human, animals, plants or in a specific environment. For example, human gut microbiome has been known associated with human diseases such as diabetes, inflammatory bowel disease, obesity, asthma and rheumatoid arthritis [1]. The knowledge could be applied for improving human health and treatment. Not only human, but also animal and plant health are associated with microbiome. The understanding of microbial communities could be applied for several applications such as agriculture for improving productivities in farm animals [2] and economic crop plants [3]. Understanding animal and plant microbiome could provide alternative usages of beneficial microbes instead of antibiotics, chemical fertilizers or pesticides, respectively [4]. Insights information on microbiome in anaerobic wastewater treatment systems could be applied for industrial use to treat waste water and produce biogas as renewable energy [5].

With NGS technologies, two main approaches could be conducted for microbial community studies which are amplicon-based sequencing and shotgun metagenomic sequencing. The amplicon-based sequencing provides a microbial community composition (taxonomic and abundance estimation), while metagenome allows to insight investigation of both microbial community composition and functional features [6]. For amplicon-based microbiome analysis, several marker genes can be used to represent distinct microbial populations such as 16S ribosomal RNA (rRNA) genes of bacteria and archaea, 18S rRNA genes of eukaryotic species, and nuclear ribosomal internal transcribed spacer (ITS) regions of fungi. Among these markers, 16S rRNA gene is the most popular marker. It presents in almost all bacteria and archaea and contains conserved function with hypervariable regions for distinguishing different species [7-9]. Metagenomic data provide more information as all genetic materials are sequenced including functional genes. The data can be used to characterize microbial and functional diversities including antimicrobial resistance genes (AMR) or resistome profiles [10, 11]. Bioinformatics analysis is required to analyze the data leading proper and informative results [12-14]. In this tutorial, we will focus on the amplicon-based microbiome analysis.

Amplicon-based microbiome analysis workflow could be divided into four main steps, which are data assessment and preprocessing, operational taxonomic units (OTUs) clustering [15] or amplicon/exact sequence variants (ASVs) [16, 17] method, taxonomic assignment and diversity analysis. The first step is to assess sequences quality. Adapter sequences, DNA barcodes, primers and low-quality reads from sequencing steps will be removed to obtain only target regions for the analysis. Also, some noise sequences such as chimeric sequences, non-targeted amplicons, etc will be discarded. After deriving high-quality sequences, OTUs or ASVs will be identified. The methods group a unit of microorganism by sequence similarity usually 97% similarity at a taxonomic genus level or get down to the level of single-nucleotide differences for OTU and ASV, respectively. In order to perform taxonomic classification, the derived OTUs or ASVs will be compared against database [18]. The microbiome profiles can then be normalized to obtain comparable abundance scales between samples [19]. From the identified microbial profiles, diversity analyses could be performed. Alpha diversity analysis provides the diversity estimation within samples, whereas the beta diversity analysis provides the different microbial composition between samples [20]. Visualizations for diversity analyses can be performed, for instance, rarefaction curves, Principal coordinates analysis (PCoA) and heatmap.

In this tutorial, an overview and concepts of microbiome analysis based on high-throughput 16S rRNA genes sequencing will be explained along with all steps of the analysis via practical hands-on.

2 TUTORIAL OUTLINE

This is a one-day tutorial. Basic concepts, backgrounds and practical hands-on for microbiome analysis will be explained in the following orders.

- Introduction and background. The importance of microbiome study and how NGS technologies facilitate the study will be introduced.
- Approaches for microbiome analysis. An overview of microbiome analysis approaches utilizing NGS technologies and how the approaches facilitate the studies microbiota including types of information we could extracted from the data will be explained.
- Amplicon-based microbiome analysis. This tutorial will be focused on the amplicon-based microbiome analysis using 16S rRNA genes. The principles for the analysis and methods will be elucidated including existing tools and databases. We will start from explaining raw data of the NGS sequencing technology. Practical hands-on will be demonstrated as the following steps.
  - ○ Data quality assessment and data preprocessing
  - ○ Amplicon sequence variant analysis
  - ○ Taxonomy assignment
  - ○ Data normalization
  - ○ Diversity analysis. This is a downstream analysis allows us to extract information from the microbiome data. The tutorial will include both alpha (diversity within a sample) and beta (diversity between sample) diversity analyses. Commonly-used visualizations and interpretation will be elucidated e.g. rarefaction curves, PCoA and heatmap.
- Conclusion and discussion. All content will be concluded and discussed.

3 DATASET AND MATERIALS

The published microbiome data on the study of microbial communities from anaerobic sludge sources of wastewater treatment plants [5] will be used in this tutorial. It is short-read sequencing data targeting V3-V4 hypervariable region of 16S rRNA genes. To provide efficient time and less computational power through the tutorial session, the data were sub-sampled to reduce the size of the dataset. The raw sequencing data including codes for analysis and visualization are provided at https://github.com/BioM-SBI/microbiome_tutorial_csbio20. To follow the practical hands-on, participants are required to install mothur [21], an open source software for microbiome analysis, and RStudio on their computers with recommended minimum 8 GB of RAM and 100 GB of free hard drive space.

4 AUDIENCE

Undergraduates, graduates, researchers, lectures or private sectors who are interested in studying or analyzing microbiome data. Backgrounds on microbiome study, coding, or basic command line will be a benefit.

5 TUTORIAL TEAM

Dr. Kanthida Kusonmano (KK) and team are members of Systems Biology and Bioinformatics Research group at King Mongkut's University of Technology Thouburi (KMUTT), Thailand. KK is a lecturer in Bioinformatics and Systems Biology Program, School of Bioresources and Technology, KMUTT. The team members are postdoctoral researcher (Dr. Pichahpuk Uthaipaisanwong), research assistants (Pantakan Puengrang and Chalida Rangsiwutisak), Ph.D. Students (Photchanathorn Prombun and Athisri Sitthipunya) and M.Sc. students (Natchaphon Rajudom) under her supervision. They are currently working on bioinformatics and microbiome researches involved in several projects, for example, microbiome and metagenomics in anaerobic wastewater treatment systems for efficient biogas production, dynamisms of gut microbiota, functional profiles and resistome in pigs and chicken towards the improvement of animals health without using antibiotics as growth promoters, and microbiota analysis of coffee plants in order to characterize plant growth promoting microbes for replacing the use of chemical fertilizers. The team have been developing computational pipelines for microbiome analysis to answer specific biological questions by applying bioinformatics, statistics and machine learning methods. They have experiences as a part of organizers in several national and international microbiome and metagenomics workshops.

ACKNOWLEDGMENTS

We acknowledge the supports from the National Science and Technology Development Agency (NSTDA) grant#P-18-50463, and King Mongkut's University of Technology (KMUTT). We also acknowledge King Mongkut's University of Technology for funding a postdoctoral fellowship to Pichahpuk Uthaipaisanwong.

REFERENCES

[1] Bahrndorff, S., *et al.*, The Microbiome of Animals: Implications for Conservation Biology. Int J Genomics, 2016. 2016: p. 5304028.

[2] Markowiak, P. and K. Slizewska, The role of probiotics, prebiotics and synbiotics in animal nutrition. Gut Pathog, 2018. 10: p. 21.

[3] Berg, G., *et al.*, The plant microbiome and its importance for plant and human health. Front Microbiol, 2014. 5: p. 491.

[4] Compant, S., *et al.*, A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. J Adv Res, 2019. 19: p. 29-37.

[5] Puengrang, P., *et al.*, Diverse Microbial Community Profiles of Propionate-Degrading Cultures Derived from Different Sludge Sources of Anaerobic Wastewater Treatment Plants. Microorganisms, 2020. 8(2).

[6] Hamady, M. and R. Knight, Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res, 2009. 19(7): p. 1141-52.

[7] Janda, J.M. and S.L. Abbott, 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol, 2007. 45(9): p. 2761-4.

[8] Wang, Y. and P.Y. Qian, Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One, 2009. 4(10): p. e7401.

[9] Rangsiwutisak, C., S. Cheevadhanarak, and K. Kusonmano. Hypervariable regions assessment for 16S rRNA gene-based metagenomic analysis in activated sludge and anaerobic digestion. in The 30th Annual Meeting of the Thai Society for Biotechnology and International Conference. 2018. Bangkok, Thailand.

[10] Lanza, V.F., *et al.*, In-depth resistome analysis by targeted metagenomics. Microbiome, 2018. 6(1): p. 11.

[11] Dos Santos, D.F., *et al.*, Functional Metagenomics as a Tool for Identification of New Antibiotic Resistance Genes from Natural Environments. Microb Ecol, 2017. 73(2): p. 479-491.

[12] van den Bogert, B., *et al.*, On the Role of Bioinformatics and Data Science in Industrial Microbiome Applications. Front Genet, 2019. 10: p. 721.

[13] Uthaipaisanwong, P., *et al.* Evaluation of Shotgun Metagenomic Classification Performance via Microbial Richness and Diversity. in The 9th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2018). 2018. Bangkok, Thailand.

[14] Laeman, K., *et al.* SBI Metagenomics: An integrated database and visualization platform for metagenomic studies. in The 30th Annual Meeting of the Thai Society for Biotechnology and International Conference. 2018. Bangkok, Thailand.

[15] Schloss, P.D. and S.L. Westcott, Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol, 2011. 77(10): p. 3219-26.

[16] Callahan, B.J., P.J. McMurdie, and S.P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J, 2017. 11(12): p. 2639-2643.

[17] Bolyen, E., *et al.*, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol, 2019. 37(8): p. 852-857.

[18] Quast, C., *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res, 2013. 41(Database issue): p. D590-6.

[19] Weiss, S., *et al.*, Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome, 2017. 5(1): p. 27.

[20] Jost, L., Partitioning diversity into independent alpha and beta components. Ecology, 2007. 88(10): p. 2427-39.

[21] Schloss, P.D., *et al.*, Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol, 2009. 75(23): p. 7537-41.