RBUD

Subjects: Microbiology
Contributor: zhi'kai xing

RBUD, read-based metagenomics profiling for unestablished database, is a new functional potential analysis approach for whole microbial genome shotgun sequencing. Based on whole metagenome shotgun sequencing data, it can be used to analysis microbial species and functions, especially for the study without relevant reference database. RBUD method is optimized by omitting the steps of contigs assembly and ORF prediction which improves the utilization of sequencing data and shortens the time of data analysis. In addition, RBUD method includes the steps of establishing databases of microorganisms from different sources to expand its application, which is a great help for small-sample research and can avoid the lack of reference database. By compared RBUD with the existing methods in practical applications, RBUD has great advantages in both species and functional analysis.

Keywords: read-based metagenomics profiling approach ; microbiome ; microorganism ; microbial community ; metagenome database ; microbial functional analysis ; MDGM database ; metagenome shotgun sequencing ; contains assembly ; ORF

1. Introduction

In recent years, with the improvement of high-throughput sequencing technology and the rapid development of microbial research methods, it has been possible to systematically analyze all microorganisms in samples, not just those that are amenable to cultivation. Previously, these methods were mainly applied to taxonomic studies of microorganism using phylogenetic information genes (such as ribosomal RNA) [1][2]. Moreover, these studies provide a new perspective for us to understand the essential role of microorganisms in human health, soil ecology, environmental remediation and many other fields [3][4][5]. However, due to the similarity of rRNA sequences and different functions of microorganisms in different environments, it is difficult to expand the understanding of their functions through taxonomic research [6].

Whole metagenome shotgun (WMS) sequencing data can guide researchers to focus on the whole microorganisms as a community, and classify the internal genes and protein coding functions by assembling these data into an annotated reference database [7]. The main processes of the current approaches are sequence alignment, assembly and subsequent annotation, which have high requirements for mapping rate, large sample size and computer resources [1][7]. At present, there are two kinds of prevailing strategies for the analysis of whole genome shotgun sequencing data, including assembly-based profiling and read-based profiling [8]. However, these approaches are greatly restricted. The former requires reads splicing, contigs assembly and prediction of open reading frame (ORF) before mapping the data to the reference database [9]. In all these steps, data utilization is reduced due to the loss of low coverage areas [8]. The latter requires a reference database of specific host species, such as Integrated Microbial Genomes and Microbiomes (IMG) and Metagenomics of the Human Intestinal Tract (MetaHIT) for humans in order to perform conventional analysis procedures [10][11]. There are also a small number of databases for other host species in metagenomics studies, including the chicken [12], pig [13] and mouse [14]. However, the construction of these databases requires a large number of sequencing data and sample sets, and the cost is very high [8]. This makes it difficult to succeed in the studies for small samples and uncommon host species. Therefore, the current methods are not sufficient to achieve the functional metagenomics studies for the lack of existing metagenomic databases and small sample size. Moreover, these methods cannot effectively improve the utilization of sequencing data.

2. Basic Workflow and Characteristics of Three Different Metagenomics Profiling

In this study, we employed two conventional methods (assembly-based metagenomics profiling for unestablished database (ABUD) and read-based metagenomics profiling for established database (RBED)) and our newly developed method (RBUD) to analyze the intestinal microbial metagenomes of avian colibacillosis chicken and T2D patients,

comparing the accuracy and effectiveness of the three methods [8]. The flow chart showed the difference of the three analyses (Figure 1). These three methods can be distinguished by whether there is an established reference gene catalog and whether the contigs are assembled.

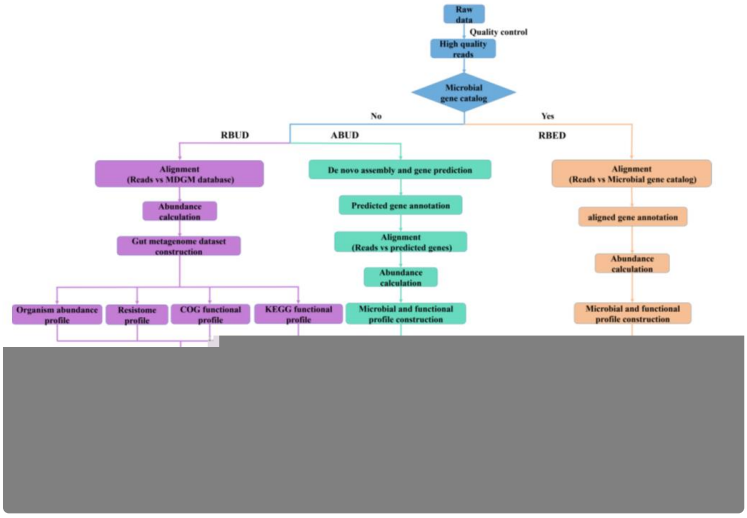

Figure 1. Flow chart of the read-based metagenomics profiling for unestablished database (RBUD) method, the assembly-based metagenomics profiling for unestablished database (ABUD) method and the read-based metagenomics profiling for established database (RBED) method to analyze metagenomic data in this study. The blue box represents the common steps of all methods. The purple box represents the steps of RBUD method. The light green box represents the steps of ABUD method, and the orange box represents the steps of RBED method.

2.1. Read-based Metagenomics Profiling for Unestablished Database (RBUD)

RBUD was developed in this study to analyze metagenome data without assembly steps. The first important step for RBUD was to establish a relevant database, especially for the rare types of samples. In this study, we built a metagenome database (MDGM) based on the data of microorganisms from National Center for Biotechnology Information (NCBI) [15]. It contains microbial species from different hosts, different environmental sources and different sampling parts of the same host, which are competent for most metagenomic studies. To fulfill the construction of MDGM, there were several sub-steps that needed to be done as follows: firstly, whole microbial genome data (5133 bacteria, 9548 viruses and 243 fungi) and corresponding species and their taxonomic annotation information were downloaded from the NCBI database to construct the microbial species dataset (on 3 December 2015) [15]. Secondly, the CDS(coding sequence) region sequences and corresponding annotations of these sequences including gene ID, protein ID, location in chromosome, Cluster of Orthologous Group of Proteins (COG) function and protein product were collected to build the functional dataset. Most of the functional annotations can be obtained according to gene ID or protein ID of the NCBI database [15]. When some databases do not support information retrieval by NCBI gene ID or protein ID, the sequence of the CDS region was aligned with the nucleic acid or protein sequence in these databases through BLASTN or BLASTP search (e-value < $1x10^{-5}$) to obtain functional annotation [16]. The databases that we have used for functional annotation in this study were eggNOG [17], Kyoto Encyclopedia of Genes and Genomes (KEGG) [18], Antibiotic Resistance Gene Database (ARDB) [19], Carbohydrate-Active enZymes Database (CAZy) [20], The Comprehensive Antibiotic Resistance Database (CARD) [21], Universal Protein (UniProt) [22] and Metabolic Pathways From all Domains of Life (MetaCyc) [23].

The second step of RBUD was to single out the high-quality reads. In this step, the raw data were obtained by high-throughput sequencing or from published data. Then, quality control was carried out to remove low-quality reads and host DNA contamination.

The third step of RBUD was to establish microbial species profiling and functional profiling of our testing samples. In this step, we firstly aligned the high-quality reads with MDGM to calculate the abundance of all microbiome species. Then, we started to analyze the difference of microbial composition among sample groups and to calculate the microbial diversity. Meanwhile, the high-quality reads were also aligned with the CDS region sequences of MDGM to calculate the abundance of all genes and obtain their functional annotations. The genes were clustered by their functions to get multiple functional orthologues containing different genes. Then, redundant genes with the same abundance and in the same functional orthologue were removed. The functional abundance was the sum of the non-redundant gene abundance in the same functional orthologue. Subsequently, the differential genes were identified, and functional analyses were performed.

The codes used in the RBUD method and for constructing MDGM database have been successfully uploaded to https://github.com/DMsiast/RBUD.git, which is open to all users. Researchers can download the full processing code from the website. For the purpose of studying bacteria, viruses and fungi separately, we provided three individual databases that can save time and improve accuracy.

2.2. Assembly-based Metagenomics Profiling for Unestablished Database (ABUD)

ABUD is a kind of commonly used method to analyze microbial shotgun genome data, which can be applied without an existing reference database [8]. RAST [24], Megan4 [25], MOCat2 [26], Carnelian [27] and IMG4 [28] belong to ABUD method. The basic principal and the main workflows of these tools are similar. First, low quality reads and host DNAs are removed from raw data. Then, the ORFs are obtained after contigs assembly and gene prediction by MetaGeneMark software. The predicted genes are annotated via aligning ORFs with a universal database (e.g., MDGM), and a non-redundant reference gene catalogue is built. After that, high quality sequencing reads are aligned with the above gene catalogue to calculate gene abundance capable of building microbial and functional profiles. Finally, the characteristics of microflora can be acquired through the comparative analysis of different sample groups. However, for the ABUD method, the utilization rate of sequencing reads is reduced during data processing and the loss of biological information is serious. Although increasing sample size and sequencing depth can solve this problem in a certain degree, more computing resources and economic investments are required.

2.3. Read-based Metagenomics Profiling for Established Database (RBED)

The RBED method can be implemented through external sequence data sources (such as open reference genomes) without reads assembly [8]. Since assembly is a slow, resource intensive and lossy process, reads directly mapping to the existing database is the core concept for RBED. MG-RAST [29], ShotMap [30], COGNIGER [31] and HUMAnN2 [32] belong to RBED method, which have similar procedures with different reference databases. For the RBED method, the data pretreatment of RBED is consistent with that of ABUD. Then, the high-quality reads are aligned with the reference gene catalog, which has been built in the existing database to calculate the relative abundance of these genes. Retrieving the gene annotation information in the gene catalog, the bacterial species and functional profiles are established. Finally, through the comparative analysis of different sample groups, the microflora characteristics can be obtained.

The RBED approach can mitigate the assembly problems, speed up computation and analyze the low abundance microorganisms that cannot be assembled. Nowadays, many reference genomes are rapidly increasing [33][34], and each reference genome can be used for the analysis of a certain sample type, such as the human gut [35]. However, the lack of representative reference genome hampers the analysis of some more diverse environments, such as soil and oceans. Thus, the application scope of the RBED method based on the existing reference database is limited by the source of samples.

References

1. Scholz, M.B.; Lo, C.C.; Chain, P.S. Next generation sequencing and bioinformatic bottlenecks: The current state of met agenomic data analysis. Opin. Biotechnol. 2012, 23, 9–15.
2. Fukuda, K.; Ogawa, M.; Taniguchi, H.; Saito, M. Molecular approaches to studying microbial communities: Targeting th e 16s ribosomal rna gene. UOEH 2016, 38, 223–232.
3. Cho, I.; Blaser, M.J. The human microbiome: At the interface of health and disease. Rev. Genet. 2012, 13, 260–270.
4. Hultman, J.; Waldrop, M.P.; Mackelprang, R.; David, M.M.; McFarland, J.; Blazewicz, S.J.; Harden, J.; Turetsky, M.R.; McGuire, A.D.; Shah, M.B.; et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature 2015, 521, 208–212.
5. Fierer, N.; Ladau, J.; Clemente, J.C.; Leff, J.W.; Owens, S.M.; Pollard, K.S.; Knight, R.; Gilbert, J.A.; McCulley, R.L. Re constructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the united states. Science 20 13, 342, 621–624.
6. Sentausa, E.; Fournier, P.E. Advantages and limitations of genomics in prokaryotic taxonomy. Microbiol. Infect. Off. Pub l. Eur. Soc. Clin. Microbiol. Infect. Dis. 2013, 19, 790–795.
7. Nagarajan, N.; Pop, M. Sequence assembly demystified. Rev. Genet. 2013, 14, 157–167.
8. Quince, C.; Walker, A.W.; Simpson, J.T.; Loman, N.J.; Segata, N. Shotgun metagenomics, from sampling to analysis. B iotechnol. 2017, 35, 833–844.
9. Gilbert, J.A.; Dupont, C.L. Microbial metagenomics: Beyond the genome. Rev. Mar. Sci. 2011, 3, 347–371.

10. Chen, I.A.; Markowitz, V.M.; Chu, K.; Palaniappan, K.; Szeto, E.; Pillay, M.; Ratner, A.; Huang, J.; Andersen, E.; Hunte mann, M.; et al. Img/m: Integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. 20 17, 45, D507–D516.

11. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K.S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010, 464, 59–65.

12. Huang, P.; Zhang, Y.; Xiao, K.; Jiang, F.; Wang, H.; Tang, D.; Liu, D.; Liu, B.; Liu, Y.; He, X., et al. The chicken gut meta genome and the modulatory effects of plant-derived benzylisoquinoline alkaloids. Microbiome 2018, 6, 211.

13. Xiao, L.; Estellé, P.; Ramayo-Caldas, Y.; Xia, Z.; Feng, Q.; Liang, S.; Pedersen, A.; Kjeldsen, N.J.; Liu, C.; Maguin, E.; e t al. A reference gene catalogue of the pig gut microbiome. Nature Microbiol. 2016, Epub ahead of print.

14. Xiao, L.; Feng, Q.; Liang, S.; Sonne, S.B.; Xia, Z.; Qiu, X.; Li, X.; Long, H.; Zhang, J.; Zhang, D., et al. A catalog of the mouse gut metagenome. Biotechnol. 2015, 33, 1103–1108.

15. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Ostell, J. Database resources of the national center for biotechn ology information. Nucleic Acids Res. 2018, 46, D8–D13.

16. Lopez, R.; Silventoinen, V.; Robinson, S.; Kibria, A.; Gish, W. Wu-blast2 server at the european bioinformatics institute. Nucleic Acids Res. 2003, 31, 3795–3798.

17. Jensen, L.J.; Julien, P.; Kuhn, M.; von Mering, C.; Muller, J.; Doerks, T.; Bork, P. Eggnog: Automated construction and a nnotation of orthologous groups of genes. Nucleic Acids Res. 2008, 36, D250–D254.

18. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The kegg resource for deciphering the genome. Nucleic Acids Res. 2004, 32, D277–D280.

19. Liu, B.; Pop, M. Ardb--antibiotic resistance genes database. Nucleic Acids Res. 2009, 37, D443–D447.

20. Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P.M.; Henrissat, B. The carbohydrate-active enzymes databa se (cazy) in 2013. Nucleic Acids Res. 2014, 42, D490–D495.

21. Jia, B.; Raphenya, A.R.; Alcock, B.; Waglechner, N.; Guo, P.; Tsang, K.K.; Lago, B.A.; Dave, B.M.; Pereira, S.; Sharma, A.N.; et al. Card 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucl eic Acids Res. 2017, 45, D566–D573.

22. Renaux, A. Uniprot: The universal protein knowledgebase (vol 45, pg d158, 2017). Nucleic Acids Res. 2018, 46, 2699.

23. Caspi, R.; Billington, R.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Midford, P.E.; On g, Q.; Ong, W.K.; et al. The metacyc database of metabolic pathways and enzymes. Nucleic Acids Res. 2018, 46, D633 –D639.

24. Aziz, R.K.; Bartels, D.; Best, A.A.; DeJongh, M.; Disz, T.; Edwards, R.A.; Formsma, K.; Gerdes, S.; Glass, E.M.; Kubal, M.; et al. The rast server: Rapid annotations using subsystems technology. BMC Genom. 2008, 9, 75.

25. Huson, D.H.; Mitra, S.; Ruscheweyh, H.J.; Weber, N.; Schuster, S.C. Integrative analysis of environmental sequences u sing megan4. Genome Res. 2011, 21, 1552–1560.

26. Kultima, J.R.; Coelho, L.P.; Forslund, K.; Huerta-Cepas, J.; Li, S.S.; Driessen, M.; Voigt, A.Y.; Zeller, G.; Sunagawa, S.; Bork, P. Mocat2: A metagenomic assembly, annotation and profiling framework. Bioinformatics 2016, 32, 2520–2523.

27. Nazeen, S.; Yu, Y.W.; Berger, B. Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. Genome Biol. 2020, 21, 47.

28. Markowitz, V.M.; Chen, I.M.; Palaniappan, K.; Chu, K.; Szeto, E.; Pillay, M.; Ratner, A.; Huang, J.; Woyke, T.; Hunteman n, M.; et al. Img 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res. 2014, 42, D560–D567.

29. Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.M.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilk e, A.; et al. The metagenomics rast server-A public resource for the automatic phylogenetic and functional analysis of m etagenomes. BMC Bioinform. 2008, 9, 386.

30. Nayfach, S.; Bradley, P.H.; Wyman, S.K.; Laurent, T.J.; Williams, A.; Eisen, J.A.; Pollard, K.S.; Sharpton, T.J. Automate d and accurate estimation of gene family abundance from shotgun metagenomes. PLoS Comput. Biol. 2015, 11, e1004 573.

31. Bose, T.; Haque, M.M.; Reddy, C.; Mande, S.S. Cognizer: A framework for functional annotation of metagenomic datas ets. PLoS ONE 2015, 10, e0142102.

32. Franzosa, E.A.; McIver, L.J.; Rahnavard, G.; Thompson, L.R.; Schirmer, M.; Weingart, G.; Lipson, K.S.; Knight, R.; Cap oraso, J.G.; Segata, N.; et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods 2018, 15, 962–968.

33. Stewart, E.J. Growing unculturable bacteria. J. Bacteriol. 2012, 194, 4151–4160.
34. Rinke, C.; Schwientek, P.; Sczyrba, A.; Ivanova, N.N.; Anderson, I.J.; Cheng, J.F.; Darling, A.; Malfatti, S.; Swan, B.K.; Gies, E.A.; et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 2013, 499, 431–437.
35. Nelson, K.E.; Weinstock, G.M.; Highlander, S.K.; Worley, K.C.; Creasy, H.H.; Wortman, J.R.; Rusch, D.B.; Mitreva, M.; Sodergren, E.; Chinwalla, A.T.; et al. A catalog of reference genomes from the human microbiome. Science 2010, 328, 994–999.