Methods 79–80 (2015) 52–59

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Deciphering the human microbiome using next-generation sequencing data and bioinformatics approaches

Yihwan Kim [a], InSong Koh [a,b], Mina Rho [a,c,*]

[a] Department of Biomedical Informatics, Hanyang University, Seoul, Republic of Korea
[b] Department of Physiology, Hanyang University, Seoul, Republic of Korea
[c] Division of Computer Science and Engineering, Hanyang University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 9 June 2014
Received in revised form 6 October 2014
Accepted 13 October 2014
Available online 28 October 2014

Keywords:
Metagenomics
Human microbiome
Next generation sequencing
Bioinformatics

ABSTRACT

The human microbiome is one of the key factors affecting the host immune system and metabolic functions that are not encoded in the human genome. Culture-independent analysis of the human microbiome using metagenomics approach allows us to investigate the compositions and functions of the human microbiome. Computational methods analyze the microbial community by using specific marker genes or by using shotgun sequencing of the entire microbial community. Taxonomy profiling is conducted by using the reference sequences or by *de novo* clustering of the specific region of sequences. Functional profiling, which is mainly based on the sequence similarity, is more challenging since about half of ORFs predicted in the metagenomic data could not find homology with known protein families. This review examines computational methods that are valuable for the analysis of human microbiome, and highlights the results of several large-scale human microbiome studies. It is becoming increasingly evident that dysbiosis of the gut microbiome is strongly associated with the development of immune disorder and metabolic dysfunction.

1. Introduction

Next generation sequencing technology and efficient computational methods have made significant contributions to the genome studies of animals, plants, and microbes. In many environments, less than 1% of microbes can be cultivated [1]. This limitation makes the studies of microbes and their communities extremely challenging. Recent advances in the culture-independent analysis of the microbial communities using metagenomics approach have allowed an in-depth investigation of the structures and functions of human microbiome [2,3].

Human microbiome plays a crucial role in protecting the host against pathogenic microbes and providing metabolic functions that are not encoded by the human genome [4,5]. Recent studies have shown a significant association of the human gut microbiome with the development of obesity, type 2 diabetes, and inflammatory bowel disease [6]. For example, obese individuals show a significant increase of *Firmicutes* and decrease of *Bacteroidetes* in their gut microbiomes compared with lean individuals [7]. More interestingly, the proportion of *Bacteroidetes* increased when the subjects lost weight on low-calorie diet over a period of 52 weeks. Such dynamics of the human microbiome should be investigated to understand better the development or progress of many diseases.

Two large-scale international human microbiome projects, Metagenomics of the Human Intestinal Tract (MetaHit) [8,9] and Human Microbiome Project (HMP) [2,3], were initiated in late 2000s to study microorganisms in the human body and to develop computational methods that analyze sequenced metagenomes. MetaHit catalogued 3.3 M non-redundant genes from 124 gut microbiomes of healthy, overweight and obese adults, as well as inflammatory bowel disease patients. Based on this finding, it is estimated that more than 1000 bacterial species exist in the human gut microbiome [8,9]. This was one of the pioneering projects that performed *de novo* assembly on the human microbiome datasets of short reads. Genes were predicted from the metagenomic contigs to expand our understanding of diverse functions.

The HMP studied microbial communities from seven body sites (i.e. three different oral, gut, urogenital, nasal, and skin sites) in more than two hundred healthy adults [2,3]. Normal microbial flora of the different body sites have been investigated by using shotgun metagenomic sequences as well as marker genes. In addition to sharpening our understanding of the human microbiome, they also serve as valuable references for newly sequenced samples. *Firmicutes* and *Bacteroidetes* are the most represented

* Corresponding author at: Division of Computer Science and Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea.
E-mail address: minarho@hanyang.ac.kr (M. Rho).

http://dx.doi.org/10.1016/j.ymeth.2014.10.022
1046-2023/

bacterial phyla in the gut microbiome of healthy adults; *Actinobacteria*, *Firmicutes*, and *Proteobacteria* are major commensals in the skin. *Fusobacteria*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* also colonize the oral cavity.

In order to gain a better understanding of the microbial community structure, many computational methods have been developed. They use either specific marker genes such as 16S ribosomal RNAs, or the entire genetic materials of the community. With abundant sequencing data currently available, microbiome changes have been identified under different conditions such as dietary habits, human body sites, or physiological states [5,10]. Opportunities are also growing to better understand the dynamics of pathogenic microbes in the hosts, as well as their interactions in specific environments.

In this review are summarized computational methods that are currently available to analyze microbial communities using the next-generation sequencing data. Focus is placed on computational methods for two different approaches: composition profiling and function characterization. Brief description of several recent studies that are based on such computational methods is provided toward the end of this article.

2. Characterization of microbiome composition

The analysis of diverse microbiome has provided valuable opportunities to understand the relationships between microbiome and environmental conditions such as ages, dietary habits, different human body sites, and physiological states (Fig. 1). In order to characterize microbiome, taxonomic profiling has been carried out along with operational taxonomic unit (OTU) clustering. Here, 16S ribosomal RNAs have been used as a bacterial phylogenetic marker to determine the composition of microbiome and their dynamics under different environmental conditions.

Two different approaches have been taken in the development of computational methods to cluster the 16S ribosomal RNA sequences. One method is a homology-based approach using the 16S ribosomal RNA sequence database, such as GreenGenes [11], Ribosomal Database Project (RDP) [12], and Silva [13]. By using sequence alignment algorithms, sequencing reads from diverse bacterial 16S ribosomal RNAs are assigned to the closest species. This approach is limited to identifying known bacterial species, and therefore cannot find novel species. The other approach is clustering reads based on the sequence identity. The resulting clusters are searched to find the taxonomy assignment. Several methods are currently available for clustering typical DNA sequences, which include CD-HIT [14], UCLUST [15], and DNACLUST [18]. Many clustering methods use *k*-mer profiling to obtain rough estimates of sequence similarity between pairs of 16S RNA sequences as a pre-process step. If the distance of *k*-mer profiles between sequence pairs is close enough, those pairs could be the candidates of the sequence alignment for calculating accurate sequence similarities. This pre-process step can reduce the time for sequence alignments [15]. The essential idea behind the OTU clustering algorithms is that the similarity between the representative sequence (centroid) and any sequence in the cluster is higher than the desired threshold.

CD-HIT sorts the sequences to find representative sequences from the longest ones [14]. All sequences are compared with the representative sequences to find the correct cluster. If a sequence could not find similar representative sequences according to the threshold, it would become the representative sequences of a new cluster. UCLUST uses different heuristic approaches [15]. The representative sequences are sorted by the number of *k*-mers that are shared with the query sequence, which could reduce the number of sequence alignments that takes most of the time in clustering. Compared with the CD-HIT, UCLUST improved the running time by efficiently ordering pairwise alignments after *k*-mer comparison [15]. DNACLUST uses a similar greedy approach except for the way to find the member sequences of the cluster [18]. From the unclustered sequences that are sorted by a descending order of sequence length, the longest sequence is selected as the representative sequence of a new cluster. Subsequently, all the sequences that share higher sequence similarity than the threshold are assigned to a new cluster. CROP (Clustering 16S rRNA for OTU Prediction) splits sequences into different blocks, applies independent Bayesian clustering for each block, and merges all blocks to complete clustering into OTUs [16]. This method does not require the threshold for sequence similarity to cluster sequences, and therefore might be more robust against sequencing errors.

QIIME utilizes these two approaches as well as a combined approach [17]. Taxonomic profiling uses reference sequences from GreenGenes [16], RDP [12], or Silva [13] to assign taxonomy to the reads. The reads that are not mapped to any reference sequences are discarded. On the other hand, *de novo* approaches (OTU picking) use UCLUST and USEARCH [15] to cluster all reads based on the sequence similarity. The hybrid approach first applies taxonomy profiling, and then applies OTU picking for the rest of the reads that are not assigned to any taxonomy. As a pre-processing step, QIIME also provides useful functions for demultiplexing reads from different samples. For a more efficient analysis of microbiome diversity, it also provides functions for statistical analysis and visualization, such as OTU heatmap and clustering plots (Table 1).

Once the OTU clustering or reference-based taxonomy assignments are carried out, the microbial communities need to be compared to test whether environments are significantly different and to find which lineages contribute to distinguish the communities. UniFrac is one of the widely used web applications for community comparison with respect to phylogenetic information [18]. As an

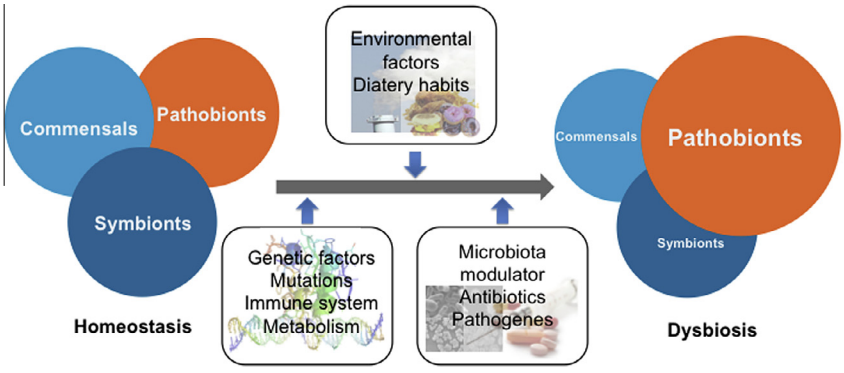

Fig. 1. Changes of the microbial community composition. Environmental and genetic factors, and microbial modulators affect the homeostasis of human microbiome, which is associated with the development of several diseases.