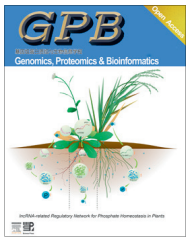
Genomics Proteomics Bioinformatics 13 (2015) 148–158

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



REVIEW

Metagenomic Surveys of Gut Microbiota

Rahul Shubhra Mandal [1,a], **Sudipto Saha** [2,*,b], **Santasabuj Das** [1,3,*,c]

[1] *Biomedical Informatics Centre, National Institute of Cholera and Enteric Diseases, Kolkata 700010, India*
[2] *Bioinformatics Centre, Bose Institute, Kolkata 700054, India*
[3] *Division of Clinical Medicine, National Institute of Cholera and Enteric Diseases, Kolkata 700010, India*

Received 8 July 2014; revised 10 February 2015; accepted 26 February 2015
Available online 13 July 2015

Handled by Fangqing Zhao

KEYWORDS

Disease;
Sequencing;
16S rRNA;
Operational taxonomic unit;
Microbial interaction network

Abstract Gut microbiota of higher vertebrates is host-specific. The number and diversity of the organisms residing within the gut ecosystem are defined by physiological and environmental factors, such as host genotype, habitat, and diet. Recently, culture-independent **sequencing** techniques have added a new dimension to the study of gut microbiota and the challenge to analyze the large volume of **sequencing** data is increasingly addressed by the development of novel computational tools and methods. Interestingly, gut microbiota maintains a constant relative abundance at **operational taxonomic unit** (OTU) levels and altered bacterial abundance has been associated with complex **diseases** such as symptomatic atherosclerosis, type 2 diabetes, obesity, and colorectal cancer. Therefore, the study of gut microbial population has emerged as an important field of research in order to ultimately achieve better health. In addition, there is a spontaneous, non-linear, and dynamic interaction among different bacterial species residing in the gut. Thus, predicting the influence of perturbed microbe–microbe interaction network on health can aid in developing novel therapeutics. Here, we summarize the population abundance of gut microbiota and its variation in different clinical states, computational tools available to analyze the pyrosequencing data, and gut microbe–microbe interaction networks.

Introduction

Metagenomics is the study of genetic material retrieved directly from environmental samples including the gut, soil, and water. Typically, human gut microbiota behaves like a multicellular organ, which consists of nearly 200 prevalent bacterial species and approximately 1000 uncommon species [1]. Several factors, such as diet and genetic background of the host and immune status, affect the composition of the microbiota [2,3]. It is also shown that early environmental exposure and the maternal inoculums have a large impact on gut microbiota in adulthood [4]. Gut microbiota complements the biology of an organism in ways that are mutually beneficial [5].

Gut microbiota can be studied using different approaches. For instance, descriptive metagenomics can reveal community structure and variation of the microbiome and microbial relative abundance is estimated based on different physiological

* Corresponding authors.
E-mail: ssaha4@jcbose.ac.in (Saha S), dasss@icmr.org.in (Das S).
[a] ORCID: 0000-0001-8939-3362.
[b] ORCID: 0000-0001-9433-8894.
[c] ORCID: 0000-0003-0745-469X.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.02.005

and environmental conditions [6,7]. On the other hand, functional metagenomics is the study of host–microbe and microbe–microbe interactions toward a predictive, dynamic ecosystem model. Such studies reflect connections between the identity of a microbe or a community and their respective functions in the environment (terms are defined in **Box 1**) [8,9]. However, a major challenge in the study of gut microbiota is the inability to culture most of the gut microbial species [10]. Several efforts have been previously made in this regard. Gordon et al. identified 86 culturable species in human colonic microbiota from three healthy adults (http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf). Gut ecosystems are currently being studied in the native state using 16S rRNA gene amplicon sequencing or whole genome sequencing (WGS) techniques [11]. 16S rRNA gene sequencing is widely used for phylogenetic reconstruction, nucleic acid-based detection, and quantification of microbial diversity. In contrast, WGS additionally explores the functions of the metagenome. The gut microbial community structure and function have been studied in different host species, including mouse [12], human [13], canine, [14], feline [14], cow [15], and yak [15]. Despite inter-species differences in community structure and function, gut microbiota frequently play a beneficial role in host metabolism and immunity across different species [16].

Large numbers of metagenomic sequence datasets have been generated, thanks to the advances in WGS and 16S rRNA pyrosequencing techniques [17]. These datasets are available in different repositories including the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra), the Data Analysis and Coordination Center (DACC) under the Human Microbiome Project (HMP) (http://hmpdacc.org) supported by the National Institutes of Health (NIH), metagenomic data resource from the European Bioinformatics Institute (EBI) (https://www.ebi.ac.uk/metagenomics/) and the UniProt Metagenomic and Environmental Sequences (UniMES) database (http://www.uniprot.org/help/unimes). All these sequence archives also provide different tools for the analysis of metagenomic sequences. Starting with the first-generation Sanger (*e.g.*, Applied Biosystems) platforms to the second-generation 454 Life Sciences Roche (*e.g.*, GS FLX Titanium) and Illumina (*e.g.*, GA II, MiSeq, and HiSeq) platforms and finally, the recently developed Ion Torrent Personal Genome Machines (PGM) and Single-Molecule Real-Time (SMRT) third generation sequencing techniques introduced by Pacific Bioscience have evolved according to the need for generating cost-effective and faster metagenomic sequencing techniques. The Roche-454 Titanium platform generates consistently longer reads compared to the latest PGM platform. Whereas the MiSeq platform from Illumina produces consistently higher sequence coverage in both depth and breadth, the Ion Torrent is unique for its speed of sequencing. However, the short read length, higher complexity, and inherent incompleteness make metagenomic sequences difficult to assemble and annotate [18]. The sequences obtained from metagenomic studies are fragmented (lies between 20 and 700 base pairs) and incomplete, because of the limitations in the available sequencing techniques. Each genomic fragment is sequenced from a single species, but within a sample there are many different species, and for most of them, a full genome is absent. It becomes impossible to determine the species of origin of a particular sequence. Moreover, the volume of sequence data acquired by environmental sequencing is several orders of magnitude higher than that acquired by sequencing of a single genome [19].

Box 1 Glossary

Microbiome: the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space.
Metagenome: all the genetic material present in an environmental sample, consisting of the genomes of many individual organisms.
Metagenomic sequencing: the high-throughput sequencing of metagenome using next-generation sequencing technology.
Metagenomics: the study of genetic material or the variation of species recovered directly from environmental samples.
Descriptive metagenomics: estimation of microbial relative abundance based on different physiological and environmental conditions to reveal community structure and variation of the microbiome.
Functional metagenomics: the study of host–microbe and microbe–microbe interactions toward a predictive dynamic ecosystem model to reflect a connection between the identity of a microbe or a community.

It is well established that gut microbes constantly interact among themselves and with the host tissues. Different types of interactions are present, but most are of commensal nature. The composition of the microbial community varies significantly between and within the host species. For example, there is similarity of the microbiota between humans and mice at the super kingdom level, but significant difference exists at the phylum level [20]. In this review, we focus on different gut microbial communities residing within various host species, different software used for metagenomic data analysis, clinical importance of metagenomic studies, and importance of the microbial network toward predicting ecosystem structure and relationship among different species.

Gut microbiota studied in mammals

The gut microbial composition of only a few host species has been investigated with respect to diet, genetic potential, and disease conditions (**Table 1**). It was reported that human gut microbial communities were transplanted into gnotobiotic animal models, such as germ-free C57BL/6J mice, to examine the effects of diet on the human gut microbiome [3,21]. Diet plays a vital role in determining the composition of the resident gut microbes [3]. Turnbaugh et al. found that the human gut microbiome is shared among family members, who have similar microbiota even if they live at different locations [4]. In a study, Tap et al. identified 66 dominant and prevalent operational taxonomic units (OTUs) from human fecal samples, which included members of the genera *Faecalibacterium*, *Ruminococcus*, *Eubacterium*, *Dorea*, *Bacteroides*, *Alistipes*, and *Bifidobacterium* [22]. Another study in mice showed that host genetics along with diet is important in shaping the gut microbiota [23]. Using 16S rRNA sequencing, common microbes that belong to the *Cytophaga-Flavobacterium-Bacteroides* (CFB) phylum had been identified in the intestines of mice, rats, and humans [24]. Diversity in the fecal bacterial