Quantitative Biology 2016, 4(3): 192–206
DOI 10.1007/s40484-016-0078-x

REVIEW

An overview of major metagenomic studies on human microbiomes in health and disease

Hongfei Cui[1], Yingxue Li[1] and Xuegong Zhang[1,2,*]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, Center for Synthetic and Systems Biology/ Department of Automation, Tsinghua University, Beijing 100084, China
[2] School of Life Sciences, Tsinghua University, Beijing 100084, China
* Correspondence: zhangxg@tsinghua.edu.cn

Received January 30, 2016; Revised March 25, 2016; Accepted April 2, 2016

Many microbes are important symbiotes of human. They form specific microbiota communities, participate in various kinds of biological processes of their host and thus deeply affect human health status. Metagenomic sequencing has been widely used in human microbiota study due to its capacity of studying all genetic materials in an environment as a whole without any extra need of isolation or cultivation of microorganisms. Many efforts have been made by researchers in this area trying to dig out interesting knowledge from various metagenome data. In this review, we go through some prominent studies in the metagenomic area. We summarize them into three categories, constructing taxonomy and gene reference, characterization of microbiome distribution patterns, and detection of microbiome alternations associated with specific human phenotypes or diseases. Some available data resources are also provided. This review can serve as an entrance to this exciting and rapidly developing field for researchers interested in human microbiomes.

Keywords: metagenome; human microbiome; taxonomy and gene reference; distribution pattern; microbiome variation

INTRODUCTION

Microbes are invisible to our naked eyes but are major residents living on the earth. Any environment that can be imagined, such as water systems (oceans, lakes, rivers), soil, air, plants, animals including human, may harbor a set of microbes. They have significant influence on the environment or their host via complex interactions. And they play important roles on human health. They are small in size but often big in numbers. About 90% of cells in human body are of bacteria, archaea or some other forms of organisms, which are collectively known as microbiomes [1,2]. They live in various sites such as gut, skin, mouth or airway. Some of them are pathogens which can cause various diseases [3,4], but most of them are friendly commensals that participate our metabolic system and help maintain our health status [5,6]. They form specific microbiotas in their niches. Disorders of the microbiota may cause multiple types of diseases.

Researchers have long recognized that the study of human microbiota can improve our understanding of health. The traditional way studying microbiota is based on isolating and culturing each microbe strain separately, and studying their characters and functions. This kind of studies has given us much knowledge on microbes, but also has limitations. It has been reported that 99% of microbes cannot be isolated or cultured [7]. This implies that a vast number of microbes cannot be studied using the isolation approach. Another limitation is that microbes of a microbiota tend to live and function as members of a system rather than a group of isolated microbes [8]. In other words, microbes live as ecological communities. The traditional isolation-based techniques cannot reveal properties and functions of communities. New approaches are needed for understanding microbiomes.

In 1998, Handelsman *et al.* first used the term “metagenome”, defined as “the collective genomes of soil microflora”, in their research paper on soil microbes [9]. This concept started a new vision to study all microbes in an environment as a whole. With the help of DNA sequencing techniques, genetic materials from all

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

microbes of the environment can be extracted and sequenced. Taxa and function information can then be extracted with various bioinformatics methods and strategies. With the development of next generation sequencing (NGS) techniques, it is now very convenient to investigate microbiotas using metagenome sequencing. This technology has been widely used in the study of the microbiomes in many types of environments such as oceans, lakes, soil, air [10–13]. Because of the importance of microbiomes for human health, more and more researchers began to use metagenome sequencing to study human microbiotas [14,15].

Many metagenomic studies on human microbiomes have been published in recent years. They have covered many different aspects of human microbiomes and their impact on human health. Many data resources and knowledge have been accumulated rapidly from these studies. But they are scattered in multiple places and a systematic overview is lacking about what aspects of human microbiomes have been studied. In this review, we try to comb through major published metagenomic studies on human microbiomes. We organize the studies as three major categories: construction of taxonomic and/or gene references of human microbiomes, characterization of microbiome distribution patterns, and detection of microbiome alternations associated with specific human phenotypes or diseases (Figure 1). Available data sources from existing studies are also summarized according to these categories. We hope this review can serve as a portal for researchers to quickly grasp an overall picture of existing studies on human microbiomes and build a basic understanding on the major discoveries in this field, and also serve as a reference for major publically available data resources on human microbiomes.

CONSTRUCTING RESOURCES FOR TAXONOMIC AND GENE ANNOTATION OF HUMAN MICROBIOMES

Metagenomes are the mixture of sequences from the genomes of microbiomes in an environment. The taxonomic unit, gene and gene function are three basic elements in metagenomic studies. Just like in the area of

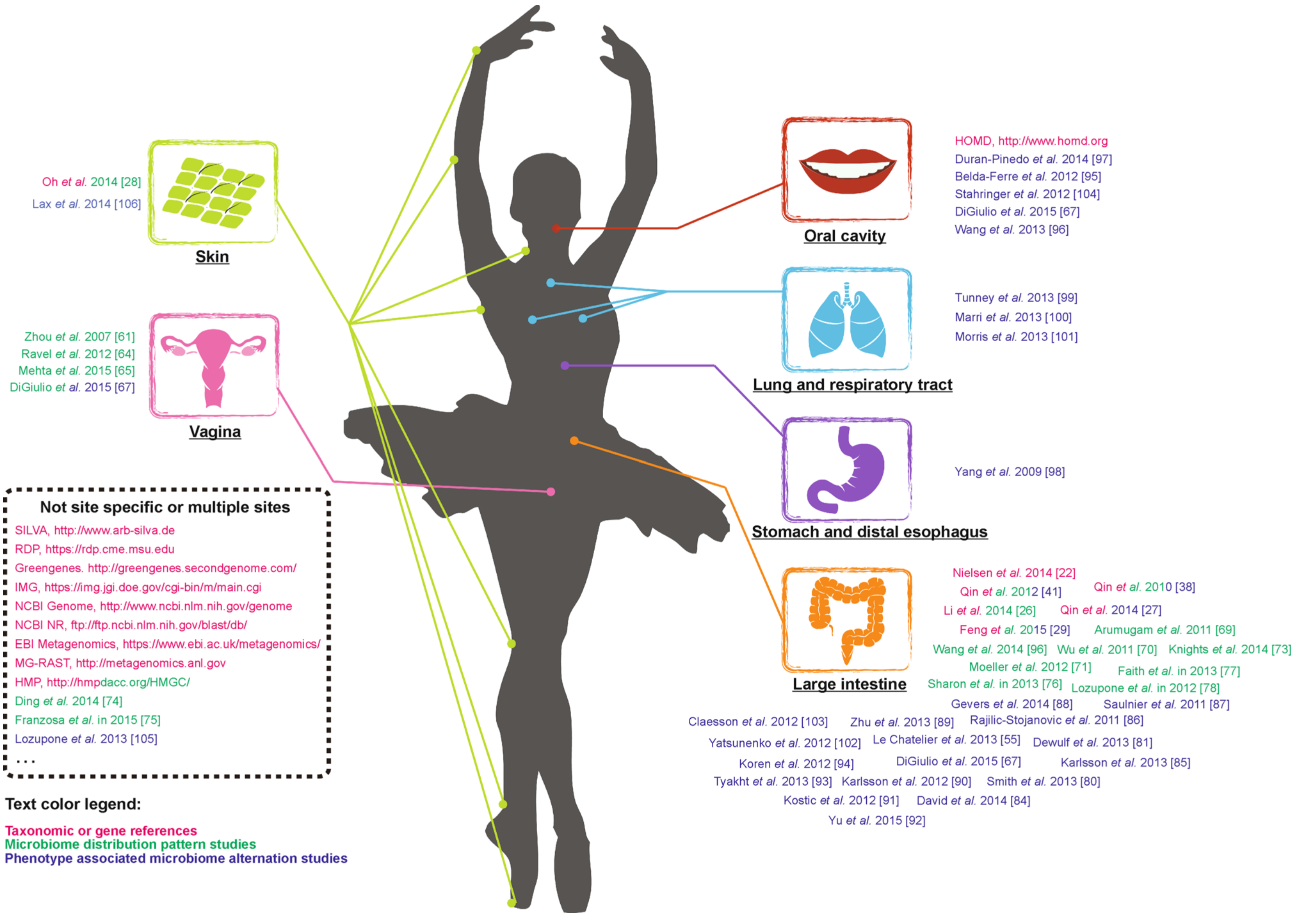
Oh et al. 2014 [28]
Lax et al. 2014 [106]
Skin
Zhou et al. 2007 [61]
Ravel et al. 2012 [64]
Mehta et al. 2015 [65]
DiGiulio et al. 2015 [67]
Vagina
Not site specific or multiple sites
SILVA, http://www.arb-silva.de
RDP, https://rdp.cme.msu.edu
Greengenes. http://greengenes.secondgenome.com/
IMG, https://img.jgi.doe.gov/cgi-bin/m/main.cgi
NCBI Genome, http://www.ncbi.nlm.nih.gov/genome
NCBI NR, ftp://ftp.ncbi.nlm.nih.gov/blast/db/
EBI Metagenomics, https://www.ebi.ac.uk/metagenomics/
MG-RAST, http://metagenomics.anl.gov
HMP, http://hmpdacc.org/HMGC/
Ding et al. 2014 [74]
Franzosa et al. in 2015 [75]
Lozupone et al. 2013 [105]
. . .
Text color legend:
Taxonomic or gene references
Microbiome distribution pattern studies
Phenotype associated microbiome alternation studies
Oral cavity
HOMD, http://www.homd.org
Duran-Pinedo et al. 2014 [97]
Belda-Ferre et al. 2012 [95]
Stahringer et al. 2012 [104]
DiGiulio et al. 2015 [67]
Wang et al. 2013 [96]
Lung and respiratory tract
Tunney et al. 2013 [99]
Marri et al. 2013 [100]
Morris et al. 2013 [101]
Stomach and distal esophagus
Yang et al. 2009 [98]
Large intestine
Nielsen et al. 2014 [22]
Qin et al. 2012 [41]
Qin et al. 2010 [38]
Li et al. 2014 [26]
Qin et al. 2014 [27]
Feng et al. 2015 [29]
Arumugam et al. 2011 [69]
Wang et al. 2014 [96]
Wu et al. 2011 [70]
Knights et al. 2014 [73]
Moeller et al. 2012 [71]
Faith et al. in 2013 [77]
Sharon et al. in 2013 [76]
Lozupone et al. in 2012 [78]
Gevers et al. 2014 [88]
Saulnier et al. 2011 [87]
Claesson et al. 2012 [103]
Zhu et al. 2013 [89]
Rajilic-Stojanovic et al. 2011 [86]
Yatsunenko et al. 2012 [102]
Le Chatelier et al. 2013 [55]
Dewulf et al. 2013 [81]
Koren et al. 2012 [94]
DiGiulio et al. 2015 [67]
Karlsson et al. 2013 [85]
Tyakht et al. 2013 [93]
Karlsson et al. 2012 [90]
Smith et al. 2013 [80]
Kostic et al. 2012 [91]
David et al. 2014 [84]
Yu et al. 2015 [92]

Figure 1. Major body sites of published human microbiota studies.

linguistics, dictionary plays a fundamental role. Constructing a complete list of all possible microbe taxa, genes and their functions in the environment is a basis for all microbiome studies. Many metagenomic studies begin with the identification of taxa, genes or annotated functions from metagenome sequences using existing references or forming their own *de novo* references. There are a number of studies aiming at building comprehensive references of microbiomes based on single projects or the integration of data from multiple projects. Table 1 illustrates some representative studies and websites of reference databases that are used frequently in metagenomic studies.

There are two major types of DNA sequences that are used to construct taxonomic reference: ribosomal RNA (rRNA) genes and whole microbial genomes. At the earlier days of metagenomic study, rRNA genes, especially 16S rRNA genes for bacteria and archaea and 18S rRNA genes for eukaryotic microbes, were most widely used. The 16S rRNA is an RNA gene with length of about 1,500 base pairs. It is regarded as an "evolutionary clock" [32] of microbes due to its highly conserved characters between different species of bacteria and archaea. One can identify the phylogenetic position of a microbe by comparing the sequence similarity of its16S rRNA gene with annotated rRNA gene references. Several well-organized rRNA gene databases have been constructed. SILVA [16], Greengenes [17] and RDP [18] are the three most famous ones. The newest SILVA version (Version 123, updated in July 23, 2015, http://www.arb-silva.de/documentation/release-123/) provides 1,756,783 small subunit (SSU, 16S for prokaryotes and 18S for eukaryota) rRNA genes, containing 1,575,088 bacteria sequences, 60,993 archaea sequences and 120,702 eukaryota sequences. SILVA also provides extensive large subunit (LSU) rRNA genes. The latest Greengenes version provided in its official website (release 13_5 updated in May 2013, http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) contains rRNA of 1,242,330 bacteria sequences and 20,656 archaea sequences. But some software such as QIIME and Mothur use a later version (release 13_8, ftp://greengenes.microbio.me/greengenes_release/), which provides a minor update of OTU structure based on release 13_5. The latest RDP version (Release 11.4, updated on May 26, 2015, https://rdp.cme.msu.edu/index.jsp) contains 3,224,600 16S rRNAs and 108,901 fungal 28S rRNAs. The RDP database also provides a genome browser containing rRNA copy number for 1,960 bacteria, 131 archaea and 25 fungi organisms. Each database has its own characteristics. Researchers can choose an appropriate one according to their research interests. For example, SILVA and Greengenes are often used as alignment references. SILVA has advantages in eukaryotic taxonomy profiling, while RDP is more famous for its classifier. Moreover, there is also a very useful reference named rrnDB (the Ribosomal RNA Operon Copy Number Database [33], https://rrndb.umms.med.umich.edu/) that contains systematically organized 16S rRNA copy number information for 2,865 bacteria and 160 archaea. The RDP classifier uses this database to make the copy-number adjustment.

Besides rRNA genes, the whole genome sequences (WGS) of microbes also serve as taxonomy reference. Researchers can identify the taxonomy information in their data set by directly mapping the sequencing reads to a whole genome reference using alignment tools such as BWA [34], Bowtie2 [35] and SOAP2 [36]. The most traditional database is the NCBI Genome database (http://www.ncbi.nlm.nih.gov/genome). It is updated frequently, and contains genome sequences of 8,075 bacteria, 521 archaea, 1,675 eukaryota (including non-microbes), 5,428 viruses and 48 viroids as of March 21, 2016. Projects such as HMP (Human Microbiome Project) [25,37], MetaHIT [38] have used the genomes from NCBI as their references. Software such as Kraken [39] for taxonomy profiling and NeSSM [40] for simulating metagenome sequence procedure also use NCBI genomes as their fundamental database. The IMG (Integrated Microbial Genomes) [20] is also a famous database, which provides both reference genomes and sequences of public metagenome studies. Its newest version (updated on March 7, 2016, https://img.jgi.doe.gov/cgi-bin/m/main.cgi) has genome sequences of 34,186 bacteria, 675 archaea, 220 eukaryota, 1,193 plasmids and 3,905 viruses and 1,192 assembled genome fragments. It also has 4,842 metagenome samples from 244 projects (updated on March 1, 2016). More and more researches are paying attentions on the IMG database [26,41–43]. The taxonomy profiling tool MetaPhlan uses the marker genes from IMG database to do taxa reconstruction [44]. There are also some site-specific databases. The Human Oral Microbiome Database (HOMD, http://www.homd.org [21]) is a good example, which included microbes from human mouth, and re-organized the taxa with their own Human Oral Taxon (HOT) number. Up to March 21, 2016, it contains 731 taxa with 406 of them have at least one annotated genomes (1,530 annotated genomes in total). It is widely used by studies targeting on oral microbiota [45–47].

It should be noticed that the majority of genomes provided by the above databases are based on the isolation and culture of the sequenced organisms, but the diversity of a microbial community extends far beyond the cultured organisms. With higher throughput of sequence platforms and deeper sequencing of metagenome data, some studies try to assemble new genomes from metagenome short reads in a *de novo* manner [22]. Although this kind of

Table 1. Major studies that provided metagenomic databases or references.

Study/Document	Reference name	Reference type	Body site	Data provided	Website
Quast *et al*. 2013 [16]	SILVA	rRNA	Not specific	1,756,783 small subunit and 96,642 large subunit	http://www.arb-silva.de/documentation/release-123/
DeSantis *et al*. 2006 [17]	Greengenes	rRNA	Not specific	1,262,986 16S rRNAs	http://greengenes.secondgenome.com/
Cole *et al*. 2014 [18]	RDP	rRNA	Not specific	3,224,600 16S rRNAs and 108,901 fungal 28S rRNAs	https://rdp.cme.msu.edu/index.jsp
NCBI Resource Coordinators, 2016 [19]	NCBI genome	Genome	Not specific	8,075 bacteria, 521 archaea, 1,675 eukaryota, 5,428 viruses and 48 viroids	http://www.ncbi.nlm.nih.gov/genome
Markowitz *et al*. 2014 [20]	IMG (Genome)	Genome	Not specific	34,186 bacteria, 675 archaea, 220 eukaryota, 1,193 plasmids and 3,905 viruses and 1,192 assembled genome fragments	https://img.jgi.doe.gov/cgi-bin/m/main.cgi
Chen *et al*. 2010 [21]	HOMD	Genome	Oral cavity	1,530 annotated genomes from 406 Human Oral Taxon	http://www.homd.org
Nielsen *et al*. 2014 [22]	–	Genome	Gut	238	EBI accession PRJEB674 to PRJEB1046
NCBI Resource Coordinators, 2016 [19]	NCBI nr	Gene	Not specific	83,785,854	ftp://ftp.ncbi.nlm.nih.gov/blast/db/
Kanehisa *et al*. 2016 [23]	KEGG	Gene	Not specific	17,956,002 genes related to 19,214 KEGG Ortholog groups (KOs)	http://www.kegg.jp/
The UniProt Consortium 2016 [24]	UniProt	Gene	Not specific	550,740 Swiss-Prot sequences and 63,039,659 TrEMBL sequences	http://www.uniprot.org/
The HMP Consortium, 2012 [25]	HMP	Gene	15 body sites	15,006,602 in Total	http://hmpdacc.org/HMGC/
Li *et al*. 2014 [26]	IGC	Gene	Gut	9,879,896	http://meta.genomics.cn/metagene/meta/home
Qin *et al*. 2014 [27]	Liver cirrhosis catalogue	Gene	Gut	2,688,468	Genes not given. Raw sequences in http://www.ebi.ac.uk/ena/data/view/ERP005860
Oh *et al*. 2014 [28]	–	Gene	Skin	5,922,920	Genes not given. Raw sequences in http://www.ncbi.nlm.nih.gov/bioproject/?term = 46333
Feng *et al*. 2015 [29]	–	Gene	Gut	~3.5 million	Genes not given. Raw sequences in http://www.ebi.ac.uk/ena/data/view/ERP008729
Meyer *et al*. 2008 [30]	MG-RAST	Metagenome	Not specific	33,809 samples from 1,159 metagenomic projects	http://metagenomics.anl.gov/
Hunter *et al*. 2014 [31]	EBI metagenomics	Metagenome	Not specific	7,085 metagenome samples, 884 metatranscriptome samples, 18,820 amplication samples and 69 assemblies from 182 projects	https://www.ebi.ac.uk/metagenomics/
Markowitz *et al*. 2014 [20]	IMG (Metagenome)	Metagenome	Not specific	4,842 metagenome samples from 244 projects	https://img.jgi.doe.gov/cgi-bin/m/main.cgi

assembled genomes are not as validated as genomes from traditional ways, they provide means to uncover the unknown taxa.

The usage of whole metagenome sequencing goes far beyond taxonomy identification and profiling. An important part is that we can identify the gene composition of a microbial community and figure out the functions the microbes play in their habitat. Moreover, the genes, as functional units, which are more conserved comparing with species, are more deeply understood by researchers. Large amounts of metagenomic gene reference resources have been generated by various projects.

Gene references can be built from high throughput WGS data. Sequencing reads are first assembled using tools such as SOAPdenovo2 [48], MetaVelvet [49] and Meta-IDBA [50]. Then ORFs (open reading frames) can be predicted from the assembled sequences using tools such as MetaGeneMark [51] and Glimmer [52]. When constructing a final gene reference catalogue, there is often an extra step removing the redundancies inside the catalogue. Tools like CD-HIT [53] and UCLUST [54] are alternative. In the HMP project, 681 samples from multiple body sites were sequenced as WGS samples for gene catalogue construction. They constructed a gene catalogue with about 15 M non-redundant genes in total. The project tremendously extended our horizon about human microbiota, and provided a systematic and comprehensive reference for future studies. In another project, Oh *et al.* collected 263 specimens from 18 body skin sites of 15 healthy adults to build a "multi-kingdom skin microbial catalogue" with about 5.9 M genes [28]. Human gut is the most frequently studied body site in metagenomics and many gut microbial gene references have been built by different studies. The MetaHIT (Metagenomics of the Human Intestinal Tract) project is one of them. The MetaHIT Consortium collected fecal samples from 124 European (Denmark or Spain) adult volunteers as the first cohort, and constructed a catalogue with about 3.3 M non-redundant genes in the year 2010 [38]. They further expanded the dataset to 760 European (Denmark or Spain) adults [22,55], which resulted in a more complete gut gene catalogue with about 8.1 M genes. The largest gut gene catalog for now was built by Li *et al.* in 2014 [26] by integrating the catalogue from samples of MetaHIT, 368 Chinese adults from Qin *et al.* [41] and the 139 American adults from HMP [25]. They named this integrated gene catalogue as 3CGC (three cohorts non-redundant gene catalog), and further merged it with a reference genome-based gut related gene catalogue SPGC (sequenced prokaryotic gene catalog), resulting in an integrated gene catalogue (IGC) of gut microbiota with about 9.9 M genes [26]. Qin *et al.* recruited a cohort of 181 Chinese samples with 98 liver cirrhosis patients and 83 healthy individuals, and constructed a 2.7 M gut gene catalogue [27]. Feng *et al.* collected and sequenced 156 fecal samples and constructed a non-redundant catalogue with about 3.5M genes [29]. The samples contain healthy controls and patients of advanced adenoma or carcinoma.

The gene references built from WGS studies include both known genes and new genes. For known genes, there are databases providing sequence information and functional annotations. The NCBI nr database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) [19] provides "all non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF excluding environmental samples from WGS projects". It is the default database of NCBI online BLAST tool. Most WGS-based gene catalogues contain genes that can be mapped to the NCBI nr database and also genes that cannot be mapped. The newest version of NCBI nr database updated on March 20, 2016 contains 83,785,854 gene sequences. The KEGG (Kyoto Encyclopedia of Genes and Genomes [23], http://www.kegg.jp/) database is a widely used database for gene annotation, which contains 17,956,002 genes related to 19,214 KEGG Ortholog groups (KOs) in its latest update in Jan. 2016. The KOs can be further linked to KEGG pathway information, which is very important for understanding the biological processes. UniProt is another widely used database (http://www.uniprot.org/) [24]. Its newest version (release 2016_03 updated on March 16, 2016) contains 550,740 manually annotated (Swiss-Prot) sequences and 63,039,659 computationally annotated (TrEMBL) sequences. Other databases such as the COG (Clusters of Orthologous Groups [56]), SEED [57], hidden Markov models based Pfam [58] and TIGRFAMs [59] are also for functional annotations. As databases often have cross-references to each other, especially the NCBI entry can be linked by most of databases, tools have been developed to parse NCBI mapping results to other annotation databases. MEGAN (MEta Genome Analyzer, http://ab.inf.uni-tuebingen.de/software/megan5/ [60]) is a good example, which can report KEGG, SEED and COG annotations just from an NCBI mapping file.

As more and more metagenomic data are being generated, researchers need comprehensive platforms for storing and arranging public metagenomic data from multiple projects. This kind of resources can help us find data of interest, download them and study them, which can promote data collaboration. There are already several well-organized data sources of this kind. The MG-RAST (Metagenome RAST, with "RAST" for "Rapid Annotation using Subsystem Technology", http://metagenomics.anl.gov/) website is a representative. It was first published in 2008 [30]. It has three main features: storing metagenomic data, processing user-uploaded data, and comparing multiple samples. Now it contains 1,159 metagenomic projects with 33,809 samples (by March 21, 2016). For each sample, it provides a brief summary for each data processing step such as quality control, function distribution, taxonomy distribution. Users can also upload their own data and use MG-RAST to process them. Simple visualization of comparison such as bar plot, heat map, and PCoA, can be done for multiple samples. This provide great convenience for metagenomic researchers, especially for those who are new to this field. The EBI metagenomics portal (https://www.ebi.ac.uk/metagenomics/) is a recently developed and rapidly growing resource for metagenomic data storage and processing [31]. It also provides features for storing, processing and comparing metagenomic samples, and has a better organization for metagenomic samples and projects. It now contains 8,136 metagenome samples, 884 metatranscriptome samples, 18,884 amplicons and 69 assemblies from 201 projects (by March 24, 2016). Many new studies publish their data in the MG-RAST and EBI database. The IMG database also archives metagenomic samples and their project information. Other resources of metagenomics data include iMicrobe (http://imicrobe.us/) and Metagenome.jp (http://mg.bio.titech.ac.jp/mg), etc.

The dictionary of metagenomes and their genes is still far away from complete. While generating more metagenomic data, researchers should also make efforts to improve the technology and computational methods to construct these references.

CHARACTERIZATION OF MICROBIOME DISTRIBUTION PATTERNS

Microbes interact with each other and with their habitat in a complicated manner. Characterizing the microbiome distribution patterns of different microbial communities is an important step to understand this mechanism. Many studies have focused on investigating the general characters of specific microbial communities across people.

Some early studies toward this direction described the compositional members of different microbial communities and how they distributed among populations [26,28,38]. One of the main aims of the HMP project [25] is to characterize the specialization of microbial communities in different body sites of healthy human via analyzing both inter- and intra- individual diversity.

As more and more studies indicate the large diversity within and among individuals, many researchers try to summarize the distribution patterns in a discrete view. In human vagina microbiome studies, clustering methods to find potential groups in microbial abundance profiles have been widely used. In 2007, Zhou *et al*. used the word "community type" to refer to the identified groups by categorizing the 16S rRNA T-RFLP (terminal restriction fragment length polymorphism) data [61]. The "community type" was also called "community state types (CSTs)" in some later studies. Some studies further investigated the associations between vagina CSTs and vaginal health in groups of particular ages [62,63], with the influence of HIV treatment [64,65], in pregnancy and birth [66–68], and so on.

In human gut microbiome studies, the concept of "enterotypes" was first proposed by Arumugam *et al.* in 2011 [69], similar with the concept of CST in human vagina microbiota studies. Arumugam *et al.* found three clusters in human gut microbiome phylogenetic structures using partitioning around medoid (PAM) method and named each cluster as an "enterotype". Each enterotype is enriched by one specific genus (*Bacteroides*, *Prevotella* or *Ruminococcus*, respectively) and has its own genera co-occurrence network. The three-enterotypes phenomenon was found in different human gut metagenome datasets, which indicated that the structure of our gut microbiota may behave in a similar manner as the A/B/O blood types. Some studies also confirmed the three enterotypes in their own datasets [41] and many studies found enterotypes with slight differences comparing to the original ones [29,70,71] such as the enriched genus and the number of the enterotypes. There is no clear and consistent understanding of how different enterotypes influence human yet. The works of Qin *et al*. [41] and Arumugam *et al.* [69] found no significant relationship between enterotypes and host phenotypes (such as age, gender or T2D diseases), but Wu *et al*. [70] found a strong association between enterotypes and long-term diet. Moeller *et al.* [71] even found that enterotypes can be switched across a long period of time over a year. Moreover, the rational of the enterotype hypothesis was also doubted by some studies, which suggested a gradient manner in changes of human microbial communities instead of summarizing the variations in a few discrete types [72,73].

In 2014, Ding *et al*. [74] extended the concept of enterotypes in gut or CSTs in vagina to the characterization of the structure of microbial communities across different body sites. Using a Dirichlet multinomial mixture (DMM) model, "microbial community types" were detected in different body sites. They observed that individual phenotypes can be related with these community types and community types also have strong correlations between oral and gut sites. They also studied the stability of community types at oral, gut and vagina sites.

Different from the cluster-based studies trying to group the microbial diversity into limited or consecutive classes, some studies were designed to investigate the discriminant power of microbiome composition. An example is the "metagenomic code" concept proposed by Franzosa *et al*. in 2015 [75]. The work was asking whether the microbial compositional variation is sufficient to mark personality. They built body-site-specific metagenomic codes based on the first visit of HMP individuals and test their performance on samples of the second visit. This study indicated that the strain-level variation in stool site microbiomes has the highest potential to identify individuals.

The above studies can all be viewed as studies on sampled static portraits of human microbiome distribution in a human population. Dynamics of microbiome distribution patterns have also been studied by some researchers. Sharon *et al.* collected 11 fecal samples from a premature infant during the first month of life and tracked variations at species and strain levels in microbial communities [76]. Faith *et al.* studied the long-term dynamics and stability of human gut microbiota using 16S rRNA sequencing on fecal samples of 37 healthy adults in the course of 296 weeks [77]. Lozupone *et al.* investigated the factors that influence the stability of gut microbial community and how they behave once being interrupted [78]. David *et al.* studied the daily dynamics of two individuals in their gut and salivary microbiota in one year and found that the microbial communities are general stable, but "can be quickly and profoundly altered by common human actions and experiences" [79]. Table 2 gives a summary of the studies reviewed in this section.

DETECTIONS OF MICROBIOME VARIATIONS ASSOCIATED WITH SPECIFIC HUMAN PHENOTYPES OR DISEASES

As the "second genome" of human beings, microbiomes have deep influences on individuals' health. Based on the knowledge of the general characters of microbial communities, digging out the major shifts or variations in the microbiome community associated with phenotypic changes, such as host habitats, age and especially human diseases, is a major goal of the metagenomic research community. Table 3 lists some representative studies on this direction.

Nutrition-related diseases are among the earliest research targets of metagenomic studies. Many microbiome studies on obesity, diabetes, kwashiorkor and diet have been reported. For example, in 2013, Le Chatelier *et al*. collected fecal samples from 292 individuals including 169 obese samples. They investigated gene richness in the non-obese and obese groups, and found several bacterial species that have strong relations with low or high gene richness and obesity [55]. The kwashiorkor study by Smith *et al*. collected samples of 317 Malawian twin pairs to study the gut microbiota differences between healthy children and kwashiorkor children under a diet treatment [80]. David *et al*. studied the effect of diet on human gut

Table 2. Major studies on mining the composition and structure of human microbiomes.

Study	Body site	Aim of study
Qin *et al*. 2010 [38]	Gut	General survey
The HMP Consortium, 2012 [25]	Multiple sites	General survey
Li *et al*. 2014 [26]	Gut	General survey
Oh *et al*. 2014 [28]	Skin	General survey
Zhou *et al*. 2007 [61]	Vagina	Community type
Ravel *et al*. 2012 [64]	Vagina	CST
Mehta *et al*. 2015 [65]	Vagina	CST
DiGiulio *et al*. 2015 [67]	Vagina	CST
Arumugam *et al*. 2011 [69]	Gut	Enterotype
Wu *et al*. 2011 [70]	Gut	Enterotype
Qin *et al*. 2012 [41]	Gut	Enterotype
Moeller *et al*. 2012 [71]	Gut	Enterotype
Knights *et al*. 2014 [73]	Gut	Rethinking of enterotype
Feng *et al*. 2015 [29]	Gut	Enterotype and microbioal community type
Ding *et al*. 2014 [74]	Multiple sites	Microbioal community type
Franzosa *et al*. 2015 [75]	Multiple sites	Metagenomic code
David *et al*. 2012 [79]	Gut and saliva	Dynamics
Lozupone *et al*. 2012 [78]	Gut	Dynamics
Sharon *et al.* 2013 [76]	Gut	Dynamics
Faith *et al*. 2013 [77]	Gut	Dynamics

Table 3. Some of phenotype studies in recent years.

Study	Main phenotype	Body site	Number of individuals	Main features	Main data accession
		Single phenotype			
Smith *et al*. 2013 [80]	Kwashiorkor	Gut	317 pairs	Taxa, functions	EBI accessions ERP001928 and ERP001911
Le Chatelier *et al*. 2013 [55]	Obesity	Gut	292	Genes	EBI accession ERP003612
Dewulf *et al*. 2013 [81]	Obesity	Gut	30	Taxa	
Cotillard *et al*. 2013 [82]	Diet, Obesity	Gut	49	Genes	EBI accession ERP003699
Adler *et al*. 2013 [83]	Diet of ancient	Ancient dental calculus	34	Taxa	EBI or SRA accession ERP002107
David *et al*. 2014 [84]	Diet	Gut	10	Taxa, functions	GEO accession GSE46761 and MG-RAST accession 6248
Qin *et al*. 2012 [41]	Diabetes	Gut	345	Taxa(MLGs), functions	SRA accession SRA045646 and SRA050230
Karlsson *et al*. 2013 [85]	Diabetes	Gut	145	Taxa(MGCs), functions	EBI or SRA accession ERP002469
Rajilic-Stojanovic *et al*. 2011 [86]	IBS	Gut	108	Taxa	
Saulnier *et al*. 2011 [87]	IBS	Gut	44	Taxa	SRA accession SRP002457
Qin *et al*. 2010 [38]	IBD	Gut	124	Taxa, functions	EBI accession ERA000116, http://www.bork.embl.de/~arumugam/Qin_et_al_2010/
Gevers *et al*. 2014 [88]	IBD	Gut	668	Taxa	BioProjects accessions PRJNA237362 and PRJNA205152
Zhu *et al*. 2013 [89]	Fatty liver	Gut	63	Taxa	MG-RAST accession 1195
Qin *et al*. 2014 [27]	Cirrhosis	Gut	181	Taxa(MGS), functions	EBI accession ERP005860
Karlsson *et al*. 2012 [90]	Atherosclerosis	Gut	27	Taxa, functions	SRA accession SRA059451
Kostic *et al*. 2012 [91]	Cancer	Colorectal carcinoma tumors and adjacent non-affected tissues	18+190	Taxa	SRA accession SRP000383
Yu *et al*. 2015 [92]	Cancer	Gut	168+40	Taxa(MLGs), functions	EBI accession PRJEB10878
Feng *et al*. 2015 [29]	Cancer	Gut	156	Taxa(MLGs), functions	EBI accession ERP008729
Tyakht *et al*. 2013 [93]	Urban/countryside	Gut	96	Taxa, functions	SRA accession SRA059011, http://www.metagenome.ru/files/rus_met/
Koren *et al*. 2012 [94]	Pregnant	Gut	91	Taxa	
DiGiulio *et al*. 2015 [67]	Pregnant	Vagina, gut, saliva, and tooth/gum	49	Taxa	SRA accession SRP288562 (Not available now)

(*Continued*)

Study	Main phenotype	Body site	Number of individuals	Main features	Main data accession
Belda-Ferre *et al*. 2012 [95]	Oral diseases	Supragingival plaque	25	Taxa, functions	MG-RAST accessions 4447192.3, 4447102.3, 4447103.3, 4447101.3, 4447943.3, 4447903.3, 4447971.3 and 4447970.3
Wang *et al*. 2013 [96]	Oral diseases	Dental surface and plaque	16	Taxa, functions	
Duran-Pinedo *et al*. 2014 [97]	Oral diseases	Subgingival plaque	13	Taxa, functions	ftp://ftp.homd.org/publication_data/20130522/
Yang *et al*. 2009 [98]	Distal esophagus diseases	Biopsy samples of the distal esophagus	34	Taxa	GEO accessions DQ537536-DQ537935 and DQ632752-DQ639751
Tunney *et al*. 2013 [99]	Bronchiectasis	Sputum	40	Taxa	EBI or SRA accession ERP002060
Marri *et al*. 2013 [100]	Asthma	Sputum	20	Taxa	
Morris *et al*. 2013 [101]	Smoke	Upper and lower respiratory tract	45	Taxa	
Multiple phenotypes					
Yatsunenko *et al*. 2012 [102]	Country, age	Gut	531	Taxa, genes	MG-RAST accession ‘qiime:850’ and ‘qiime:621’
Claesson *et al*. 2012 [103]	Diet, health	Gut	178	Taxa, genes	MG-RAST accession 154
Stahringer *et al*. 2012 [104]	Host gene, age	Saliva	107	Taxa	EBI or SRA accession ERP001346
Lozupone *et al*. 2013 [105]	Body sites, studies	Multiple sites	Public data	Taxa	Public data
Lax *et al*. 2014 [106]	Family, indoor environment, time	Skin	15+3+3+1	Taxa	EBI accession ERP005806

microbiome by recruiting ten volunteers participating in a 15-day diet plan to observe the influence on gut microbiota by plant-based or animal-based diet [84]. Results show notable effects brought by diet changes, overcoming the effect by individual genetic difference. Qin *et al.* studied the association of gut microbiome with type-2 diabetes (T2D) [41]. Bowl diseases such as IBD (inflammatory bowel disease) and IBS (irritable bowel disease) were also wildly studied [38,86–88]. Microbiome research on cancer also takes a large part of metagenomic studies [29,91,92]. Many other interesting host phenotypes, such as smoking [101], living in urban/ countryside [93], and pregnancy [67,94], have also been studied. Most of these works studied gut microbiomes using stool samples. There have been studies based on dental samples such as sub- or super- gingival plaques [95–97], or even ancient dental calculus [83]. Sputum samples are often used in respiratory disease studies [99,100]. Studies using sample from other sites such as saliva [67,104], vaginal [67], and some more complex forms of samples, such as biopsy samples [98] or samples from respiratory tract [101] have also been reported.

Some studies aimed at multiple phenotypes. They studied not only relationships between microbiota communities and phenotypes, but also the cross effects of different phenotypes on the community, and which factor is more influential. Yatsumenko *et al.* investigated the microbiome alternations across host age and geography [102]. They collected fecal samples of 531 individuals including mono- or dizygotic twins. The individuals came from 151 families across three countries (Malawi, Venezuela and United States). They used their phylogenetic profiles and function profiles of 110 of them to show the influence on gut microbiota by age, geography and human gene content. Though the structure of gut microbiota and functions vary among the three populations, they share similar functional changes along with the growth of age. Lax *et al.* collected the skin and home-surface samples of 7 families for over 6 weeks [106], and found that each family has their own microbiota pattern, and this pattern can even affect the microbiota of their living environment rapidly.

DISCUSSION

Human microbiome study is rapidly becoming a hot research topic nowadays, displaying huge potentials in forming a systems biology understanding of human health that includes microbiomes as parts of the system. We reviewed major representative studies in recent years under three categories. Studies for constructing taxonomic or gene references have created many basic resources for further metagenomic research. Studies for characterization of the microbiome distribution patterns are toward building understandings on the general properties of microbiome composition. And studies for associating microbiome variations with specific phenotypes or diseases are opening the door for uncovering the interaction between microbes and their hosts in health and disease.

During the writing of this review, new publications kept coming in all aspects of microbiome study. Kuleshov *et al.* updated the understanding of human microbiome complexity [107] by applying a quite deep sequencing using Illumina TruSeq synthetic long reads sequencing technique on a human gut microbiome. They assembled long contigs from the data and observed vast number of intra-species variation. Forslund *et al.* focused on the gut microbiome alternations with treatment on type-2 diabetes [108]. They studied how metformin treatment affect gut microbiota community using 784 human gut metagenome samples of T2D or non-T2D individuals, and revealed its therapeutic effects on T2D in a metagenomic view. Liu *et al.* found that the fecal miRNAs generated by human or murine are also very important for gut microbial composition [109]. They showed that the phenotype "host miRNA structure" provides a new angle to explain how host interact with gut microbes and shape the gut microbial structure.

Besides the studies reviewed in the three categories, there are many other interesting topics that do not belong to those categories. For example, all studies discussed above are based on the taxonomy, gene and gene function features obtained from metagenome sequencing data. Actually, there is another major type of feature called *k*-mer or *k*-tuple features which are the frequency of nucleotide words of length-*k* in the metagenome sequences. They can calculated more efficiently without depending on existing references, and can be associated with microbiome genotypes as well as host phenotypes (e. g., [110–112]). This is a direction with high potential that has not been put on sufficient attention. Most of studies we reviewed in this paper are studies on prokaryotic and eukaryotic microbes. Human viromes are also an important emerging area in metagenomics. Recent studies in this area have reported associations between viromes and human health [113–118], providing deeper understanding of the symbiotic mechanisms of human and the ubiquitous virus. Application topics such as the prebiotic concept [81], probiotics [119,120], microbiota transplant [121] or microbiota-targeted therapies [122], also have not been discussed, nor those microbiome studies based on animal models [123,124].

Although the door to the microbiome world has been opened and many recent work have reported a variety of interesting and promising views in this field, we are still at the starting phase of metagenome research. This review focused on the scientific questions of existing studies, but did not cover much of the bioinformatics methods and tools that are crucial to enable those studies [125]. With the rapid development of next generation sequence techniques, the accumulation of data far exceeds our ability in handling, analyzing and understanding the data. For a typical single metagenomic sample, we can have more than 50 M sequencing reads corresponding to about 20 GB raw data file. A small-scale project with 100~200 sample can easily generate more than sequence files at TB level. The efficiency of storing and accessing such data can already be a challenge for a biological lab, let alone all the assembly, mapping, annotation and analysis tasks. For almost all metagenome studies, one needs to extract features (such as the taxonomy, gene and function profiles) from data, based on which downstream analysis aiming at knowledge discovery is conducted. However, methods for all these tasks are far from perfect yet. In fact, for many important steps, several methods have been developed but different methods may produce different results. Take the taxonomy profiling step as an example, one can choose alignment tools such as BWA [34], Bowtie2 [35] or SOAP2 [36] and profiling methods such as MEGAN [60] or self-designed rules. One can also use one-stop methods such as MetaPhlAn [44] or Kraken [39]. Both the numbers of species and their abundances estimated with those approaches can be quite different on the same data [125]. Other steps such as metagenome assembly and gene prediction also have similar problem. Solution or improvement on this issue needs deeper understanding of the nature of metagenome data and the development of better bioinformatics methods. A solid theoretical foundation of microbiome analysis is still to be established. For example, in studies on microbiome associations with host phenotypes or diseases, many features can be picked but their discriminating power may need more stringent investigation. The recent debate of Bajaj *et al*. and Qin *et al*. [126] on the cirrhosis research of Qin *et al*. [27] is a reflection of this need. Basic study on the heritability and evolution of microbiomes is still largely open [8]. There remain a lot of places in shadow awaiting for us to put light on. Solving these questions will bring us brand new sights in reading the book of metagenomes, which require the active participation and collaboration of scientists from a wide range of disciplines from biology to computer science and statistics. Global cooperation of microbiome research has been advocated and an International Microbiome Initiative (IMI) was proposed [127]. We can foresee a very bright future of metagenomic studies that will eventually lead us to the systematic understanding of the complex system of a human being and its numerous "tiny friends", the microbiome.

ACKNOWLEDGEMENTS

This work has been funded in part by the National Basic Research Program of China (No. 2012CB316504) and NSFC (No. 61561146396).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Hongfei Cui, Yingxue Li and Xuegong Zhang declare that they have no conflict of interests.

All data were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with Helsinki Declaration of 1975, as revised in 2000 (5).

REFERENCES

1. Savage, D. C. (1977) Microbial ecology of the gastrointestinal tract. Annu. Rev. Microbiol., 31, 107–133
2. Lundberg, J. O., Weitzberg, E., Cole, J. A. and Benjamin, N. (2004) Nitrate, bacteria and human health. Nat. Rev. Microbiol., 2, 593–602
3. Relman, D. A. (2011) Microbial genomics and infectious diseases. N. Engl. J. Med., 365, 347–357
4. Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z., Quick, J., Weir, J. C., Quince, C., Smith, G. P., Betley, J. R., *et al.* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. JAMA, 309, 1502–1510
5. Kamada, N., Chen, G. Y., Inohara, N. and Núñez, G. (2013) Control of pathogens and pathobionts by the gut microbiota. Nat. Immunol., 14, 685–690.
6. Gallo, R. L. and Hooper, L. V. (2012) Epithelial antimicrobial defence of the skin and intestine. Nat. Rev. Immunol., 12, 503–516
7. Schloss, P. D. and Handelsman, J. (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol., 6, 229
8. van Opstal, E. J. and Bordenstein, S. R. (2015) MICROBIOME. Rethinking heritability of the microbiome. Science, 349, 1172–1173
9. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol., 5, R245–R249
10. Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., *et al.* (2015) Ocean plankton. Structure and function of the global ocean microbiome. Science, 348, 1261359
11. Debroas, D., Humbert, J. F., Enault, F., Bronner, G., Faubladier, M. and Cornillot, E. (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget—France). Environ. Microbiol., 11, 2412–2424
12. Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., Loiacono, K. A., Lynch, B. A., MacNeil, I. A., Minor, C., *et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol., 66, 2541–2547
13. Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z. J., Badaloni, C., Beelen, R., Caracciolo, B., de Faire, U., Erbel, R., Eriksen, K. T., *et al.* (2014) Long term exposure to ambient air pollution and incidence

of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. BMJ, 348, f7412

14. Walker, A. (2010) A glut from the gut: metagenomics takes a giant step foward. Nat. Rev. Microbiol., 8, 315
15. Lepage, P., Leclerc, M. C., Joossens, M., Mondot, S., Blottière, H. M., Raes, J., Ehrlich, D. and Doré, J. (2013) A metagenomic insight into our gut's microbiome. Gut, 62, 146–158.
16. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res., 41, D590–D596
17. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol., 72, 5069–5072
18. Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R. and Tiedje, J. M. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res., 42, D633–D642
19. NCBI Resource Coordinators. (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 44, D7–D19
20. Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., Huntemann, M., *et al.* (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res., 42, D560–D567
21. Chen, T., Yu, W. H., Izard, J., Baranova, O. V., Lakshmanan, A. and Dewhirst, F. E. (2010) The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford), 2010, baq013
22. Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol., 32, 822–828
23. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res., 44, D457–D462
24. UniProt Consortium. (2015) UniProt: a hub for protein information. Nucleic Acids Res., 43, D204–D212
25. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., *et al.* (2012) Structure, function and diversity of the healthy human microbiome. Nature, 486, 207–214
26. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., *et al.* (2014) An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol., 32, 834–841
27. Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., *et al.* (2014) Alterations of the human gut microbiome in liver cirrhosis. Nature, 513, 59–64
28. Oh, J., Byrd, A. L., Deming, C., Conlan, S., Kong, H. H., Segre, J. A., Segre, J. A., and the NISC Comparative Sequencing Program. (2014) Biogeography and individuality shape function in the human skin metagenome. Nature, 514, 59–64
29. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z.,*et al.*(2015) Gut microbiome development along the colorectal adenoma-carcinoma sequence. Nat. Commun., 6, 6528
30. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9, 386
31. Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., *et al.* (2014) EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res., 42, D600–D606
32. Woese, C. R. (1987) Bacterial evolution. Microbiol. Rev., 51, 221–271
33. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. and Schmidt, T. M. (2015) rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res., 43, D593–D598
34. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760
35. Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359
36. Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K. and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25, 1966–1967
37. Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., Gevers, D., Petrosino, J. F., Abubucker, S., Badger, J. H., *et al.* (2012) A framework for human microbiome research. Nature, 486, 215–221
38. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464, 59–65
39. Wood, D. E. and Salzberg, S. L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol., 15, R46
40. Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L. and Wei, C. (2013) NeSSM: a next-generation sequencing simulator for metagenomics. PLoS One, 8, e75448
41. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490, 55–60
42. Baker, B. J., Sheik, C. S., Taylor, C. A., Jain, S., Bhasi, A., Cavalcoli, J. D. and Dick, G. J. (2013) Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. ISME J., 7, 1962–1973
43. Inskeep, W. P., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Rusch, D. B., Tringe, S. G., Macur, R. E., Jennings, R., Boyd, E. S., Spear, J. R., *et al.* (2013) Phylogenetic and functional analysis of metagenome sequence from high-temperature archaeal habitats demonstrate linkages between metabolic potential and geochemistry. Front. Microbiol., 4, 95
44. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods, 9, 811–814
45. Jagtap, P., McGowan, T., Bandhakavi, S., Tu, Z. J., Seymour, S., Griffin, T. J. and Rudney, J. D. (2012) Deep metaproteomic analysis of

human salivary supernatant. Proteomics, 12, 992–1001

46. Liu, B., Faller, L. L., Klitgord, N., Mazumdar, V., Ghodsi, M., Sommer, D. D., Gibbons, T. R., Treangen, T. J., Chang, Y. C., Li, S., *et al.* (2012) Deep sequencing of the oral microbiome reveals signatures of periodontal disease. PLoS One, 7, e37919
47. Warinner, C., Rodrigues, J. F., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R. Y., Fiddyment, S., *et al.* (2014) Pathogens and host immunity in the ancient human oral cavity. Nat. Genet., 46, 336–344
48. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 1, 18
49. Namiki, T., Hachiya, T., Tanaka, H. and Sakakibara, Y. (2012) MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. Nucleic Acids Res., 40, e155
50. Peng, Y., Leung, H. C., Yiu, S. M. and Chin, F. Y. (2011) Meta-IDBA: a *de novo* assembler for metagenomic data. Bioinformatics, 27, i94–i101
51. Zhu, W., Lomsadze, A. and Borodovsky, M. (2010) *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res., 38, e132
52. Delcher, A. L., Harmon, D., Kasif, S., White, O. and Salzberg, S. L. (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res., 27, 4636–4641
53. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28, 3150–3152
54. Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26, 2460–2461
55. Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J. M., Kennedy, S., *et al.* (2013) Richness of human gut microbiome correlates with metabolic markers. Nature, 500, 541–546
56. Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) A genomic perspective on protein families. Science, 278, 631–637
57. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res., 33, 5691–5702
58. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., *et al.* (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res., 44, D279–D285
59. Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., Richter, A. R. and White, O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res., 35, D260–D264
60. Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007) MEGAN analysis of metagenomic data. Genome Res., 17, 377–386
61. Zhou, X., Brown, C. J., Abdo, Z., Davis, C. C., Hansmann, M. A., Joyce, P., Foster, J. A. and Forney, L. J. (2007) Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. ISME J., 1, 121–133
62. Brotman, R. M., Bradford, L. L., Conrad, M., Gajer, P., Ault, K., Peralta, L., Forney, L. J., Carlton, J. M., Abdo, Z. and Ravel, J. (2012) Association between *Trichomonas vaginalis* and vaginal bacterial community composition among reproductive-age women. Sex. Transm. Dis., 39, 807–812
63. Brotman, R. M., Shardell, M. D., Gajer, P., Fadrosh, D., Chang, K., Silver, M. I., Viscidi, R. P., Burke, A. E., Ravel, J. and Gravitt, P. E. (2014) Association between the vaginal microbiota, menopause status, and signs of vulvovaginal atrophy. Menopause, 21, 450–458
64. Ravel, J., Gajer, P., Fu, L., Mauck, C. K., Koenig, S. S., Sakamoto, J., Motsinger-Reif, A. A., Doncel, G. F. and Zeichner, S. L. (2012) Twice-daily application of HIV microbicides alter the vaginal microbiota. MBio, 3, e00370-12
65. Mehta, S. D., Donovan, B., Weber, K. M., Cohen, M., Ravel, J., Gajer, P., Gilbert, D., Burgad, D. and Spear, G. T. (2015) The vaginal microbiota over an 8- to 10-year period in a cohort of HIV-infected and HIV-uninfected women. PLoS One, 10, e0116894
66. Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., Galuppi, M., Lamont, R. F., Chaemsaithong, P., Miranda, J., *et al.* (2014) The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome, 2, 4
67. DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S., Wong, R. J., Shaw, G., *et al.* (2015) Temporal and spatial variation of the human microbiota during pregnancy. Proc. Natl. Acad. Sci. USA, 112, 11060–11065
68. Huang, Y. E., Wang, Y., He, Y., Ji, Y., Wang, L. P., Sheng, H. F., Zhang, M., Huang, Q. T., Zhang, D. J., Wu, J. J., *et al.* (2015) Homogeneity of the vaginal microbiome at the cervix, posterior fornix, and vaginal canal in pregnant Chinese women. Microb. Ecol., 69, 407–414
69. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., *et al.* (2011) Enterotypes of the human gut microbiome. Nature, 473, 174–180
70. Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., *et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science, 334, 105–108
71. Moeller, A. H., Degnan, P. H., Pusey, A. E., Wilson, M. L., Hahn, B. H. and Ochman, H. (2012) Chimpanzees and humans harbour compositionally similar gut enterotypes. Nat. Commun., 3, 1179
72. Jeffery, I. B., Claesson, M. J., O'Toole, P. W. and Shanahan, F. (2012) Categorization of the gut microbiota: enterotypes or gradients? Nat. Rev. Microbiol., 10, 591–592
73. Knights, D., Ward, T. L., McKinlay, C. E., Miller, H., Gonzalez, A., McDonald, D. and Knight, R. (2014) Rethinking “enterotypes”. Cell Host Microbe, 16, 433–437
74. Ding, T. and Schloss, P. D. (2014) Dynamics and associations of microbial community types across the human body. Nature, 509, 357–360
75. Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J. and Huttenhower, C. (2015) Identifying personal microbiomes using metagenomic codes. Proc. Natl. Acad. Sci. USA, 112, E2930–E2938
76. Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A. and Banfield, J. F. (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res., 23, 111–120
77. Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf,

H., Goodman, A. L., Clemente, J. C., Knight, R., Heath, A. C., Leibel, R. L., *et al.* (2013) The long-term stability of the human gut microbiota. Science, 341, 1237439

78. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. and Knight, R. (2012) Diversity, stability and resilience of the human gut microbiota. Nature, 489, 220–230
79. David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., Erdman, S. E. and Alm, E. J. (2014) Host lifestyle affects human microbiota on daily timescales. Genome Biol., 15, R89
80. Smith, M. I., Yatsunenko, T., Manary, M. J., Trehan, I., Mkakosya, R., Cheng, J., Kau, A. L., Rich, S. S., Concannon, P., Mychaleckyj, J. C., *et al.* (2013) Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science, 339, 548–554
81. Dewulf, E. M., Cani, P. D., Claus, S. P., Fuentes, S., Puylaert, P. G., Neyrinck, A. M., Bindels, L. B., de Vos, W. M., Gibson, G. R., Thissen, J. P., *et al.* (2013) Insight into the prebiotic concept: lessons from an exploratory, double blind intervention study with inulin-type fructans in obese women. Gut, 62, 1112–1121
82. Cotillard, A., Kennedy, S. P., Kong, L. C., Prifti, E., Pons, N., Le Chatelier, E., Almeida, M., Quinquis, B., Levenez, F., Galleron, N., *et al.* (2013) Dietary intervention impact on gut microbial gene richness. Nature, 500, 585–588
83. Adler, C.J., Dobney, K., Weyrich, L.S., Kaidonis, J., Walker, A.W., Haak, W., Bradshaw, C.J., Townsend, G., Soltysiak, A., Alt, K.W. *et al.* (2013) Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet, 45, 450–455, 455e451
84. David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., *et al.* (2014) Diet rapidly and reproducibly alters the human gut microbiome. Nature, 505, 559–563
85. Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J. and Bäckhed, F. (2013) Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature, 498, 99–103
86. Rajilić-Stojanović, M., Biagi, E., Heilig, H. G., Kajander, K., Kekkonen, R. A., Tims, S. and de Vos, W. M. (2011) Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology, 141, 1792–1801
87. Saulnier, D. M., Riehle, K., Mistretta, T. A., Diaz, M. A., Mandal, D., Raza, S., Weidler, E. M., Qin, X., Coarfa, C., Milosavljevic, A., *et al.* (2011) Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology, 141, 1782–1791
88. Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., *et al.* (2014) The treatment-naive microbiome in new-onset Crohn’s disease. Cell Host Microbe, 15, 382–392
89. Zhu, L., Baker, S. S., Gill, C., Liu, W., Alkhouri, R., Baker, R. D. and Gill, S. R. (2013) Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH. Hepatology, 57, 601–609
90. Karlsson, F. H., Fåk, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., Bäckhed, F. and Nielsen, J. (2012) Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat. Commun., 3, 1245
91. Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., *et al.* (2012) Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. Genome Res., 22, 292–298
92. Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q. Y., Qin, Y., Tang, L., Zhao, H., Stenvang, J., Li, Y., *et al.* (2015) Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut, gutjnl-2015-309800
93. Tyakht, A. V., Kostryukova, E. S., Popenko, A. S., Belenikin, M. S., Pavlenko, A. V., Larin, A. K., Karpova, I. Y., Selezneva, O. V., Semashko, T. A., Ospanova, E. A., *et al.* (2013) Human gut microbiota community structures in urban and rural populations in Russia. Nat. Commun., 4, 2469
94. Koren, O., Goodrich, J. K., Cullender, T. C., Spor, A., Laitinen, K., Bäckhed, H. K., Gonzalez, A., Werner, J. J., Angenent, L. T., Knight, R., *et al.* (2012) Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell, 150, 470–480
95. Belda-Ferre, P., Alcaraz, L. D., Cabrera-Rubio, R., Romero, H., Simón-Soro, A., Pignatelli, M. and Mira, A. (2012) The oral metagenome in health and disease. ISME J., 6, 46–56
96. Wang, J., Qi, J., Zhao, H., He, S., Zhang, Y., Wei, S. and Zhao, F. (2013) Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. Sci. Rep., 3, 1843
97. Duran-Pinedo, A. E., Chen, T., Teles, R., Starr, J. R., Wang, X., Krishnan, K. and Frias-Lopez, J. (2014) Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. ISME J., 8, 1659–1672
98. Yang, L., Lu, X., Nossa, C. W., Francois, F., Peek, R. M. and Pei, Z. (2009) Inflammation and intestinal metaplasia of the distal esophagus are associated with alterations in the microbiome. Gastroenterology, 137, 588–59
99. Tunney, M. M., Einarsson, G. G., Wei, L., Drain, M., Klem, E. R., Cardwell, C., Ennis, M., Boucher, R. C., Wolfgang, M. C. and Elborn, J. S. (2013) Lung microbiota and bacterial abundance in patients with bronchiectasis when clinically stable and during exacerbation. Am. J. Respir. Crit. Care Med., 187, 1118–1126
100. Marri, P.R., Stern, D.A., Wright, A.L., Billheimer, D. and Martinez, F. D. (2013) Asthma-associated differences in microbial composition of induced sputum. J. Allergy. Clin. Immunol., 131, 346–352. e3
101. Morris, A., Beck, J. M., Schloss, P. D., Campbell, T. B., Crothers, K., Curtis, J. L., Flores, S. C., Fontenot, A. P., Ghedin, E., Huang, L., *et al.* (2013) Comparison of the respiratory microbiome in healthy nonsmokers and smokers. Am. J. Respir. Crit. Care Med., 187, 1067–1075
102. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., *et al.* (2012) Human gut microbiome viewed across age and geography. Nature, 486, 222–227.
103. Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O’Connor, E. M., Cusack, S., Harris, H. M., Coakley, M., Lakshminarayanan, B., O’Sullivan, O., *et al.* (2012) Gut microbiota composition correlates with diet and health in the elderly. Nature, 488, 178–184
104. Stahringer, S. S., Clemente, J. C., Corley, R. P., Hewitt, J., Knights, D., Walters, W. A., Knight, R. and Krauter, K. S. (2012) Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. Genome Res., 22, 2146–2152
105. Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J. K., Gordon, J. I. and

Knight, R. (2013) Meta-analyses of studies of the human microbiota. Genome Res., 23, 1704–1714

106. Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., Gibbons, S. M., Larsen, P., Shogan, B. D., Weiss, S., *et al.* (2014) Longitudinal analysis of microbial interaction between humans and the indoor environment. Science, 345, 1048–1052

107. Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S. and Snyder, M. (2016) Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat. Biotechnol., 34, 64–69

108. Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, S., Gudmundsdottir, V., Krogh Pedersen, H., *et al.* (2015) Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature, 528, 262–266

109. Liu, S., da Cunha, A. P., Rezende, R. M., Cialic, R., Wei, Z., Bry, L., Comstock, L. E., Gandhi, R. and Weiner, H. L. (2016) The host shapes the gut microbiota via fecal microRNA. Cell Host Microbe., 19, 32–43

110. Cui, H. and Zhang, X. (2013) Alignment-free supervised classification of metagenomes by recursive SVM. BMC Genomics, 14, 641

111. Jiang, B., Song, K., Ren, J., Deng, M., Sun, F. and Zhang, X. (2012) Comparison of metagenomic samples using sequence signatures. BMC Genomics, 13, 730

112. Wang, Y., Liu, L., Chen, L., Chen, T. and Sun, F. (2014) Comparison of metatranscriptomic samples based on k-tuple frequencies. PLoS One, 9, e84348

113. Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F. and Gordon, J. I. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature, 466, 334–338

114. Kuss, S. K., Best, G. T., Etheredge, C. A., Pruijssers, A. J., Frierson, J. M., Hooper, L. V., Dermody, T. S. and Pfeiffer, J. K. (2011) Intestinal microbiota promote enteric virus replication and systemic pathogenesis. Science, 334, 249–252

115. Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res., 21, 1616–1625

116. Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D. and Bushman, F. D. (2013) Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. USA, 110, 12450–12455

117. Edlund, A., Santiago-Rodriguez, T.M., Boehm, T.K. and Pride, D.T. (2015) Bacteriophage and their potential roles in the human oral cavity. 2015, 27423

118. Wang, J., Gao, Y. and Zhao, F. (2015) Phage-bacteria interaction network in human oral microbiome. Environ. Microbiol, 10.1111/1462-2920.12923

119. Bisanz, J. E. and Reid, G. (2011) Unraveling how probiotic yogurt works. Sci. Transl. Med., 3, 106ps41

120. Ghishan, F. K. and Kiela, P. R. (2011) From probiotics to therapeutics: another step forward? J. Clin. Invest., 121, 2149–2152

121. Borody, T. J. and Khoruts, A. (2012) Fecal microbiota transplantation and emerging applications. Nat. Rev. Gastroenterol. Hepatol., 9, 88–96

122. Lemon, K. P., Armitage, G. C., Relman, D. A. and Fischbach, M. A. (2012) Microbiota-targeted therapies: an ecological perspective. Sci. Transl. Med., 4, 137rv5

123. Sonnenburg, E. D., Smits, S. A., Tikhonov, M., Higginbottom, S. K., Wingreen, N. S. and Sonnenburg, J. L. (2016) Diet-induced extinctions in the gut microbiota compound over generations. Nature, 529, 212–215

124. Chevalier, C., Stojanović, O., Colin, D. J., Suarez-Zamorano, N., Tarallo, V., Veyrat-Durebex, C., Rigo, D., Fabbiano, S., Stevanović, A., Hagemann, S., *et al.* (2015) Gut microbiota orchestrates energy homeostasis during cold. Cell, 163, 1360–1374

125. Zhang, X., Liu, S., Cui, H. and Chen, T.Reading the underlying information from massive metagenome sequencing data. *To be published.*

126. Bajaj, J. S., Betrapally, N. S. and Gillevet, P. M. (2015) Decompensated cirrhosis and microbiome interpretation. Nature, 525, E1–E2

127. Dubilier, N., McFall-Ngai, M. and Zhao, L. (2015) Microbiology: Create a global microbiome effort. Nature, 526, 631–634