Metagenomics: Focusing on the Haystack

5

Indu Khatri and Meenakshi Anurag

5.1 Introduction

The term "metagenomics" was first used to describe composite genomes of cultured soil microorganisms (Handelsman et al. 1998). In the context of the environmental studies, metagenomics, also known as "community genomics" or "ecogenomics" or "environmental genomics," is the study of composite genetic material in an environmental sample. There are a large number of microbes that are considered as uncultured in different environments, be it air, soil, water, mines, and animals, and are considered inaccessible for study with traditional approaches. Humans are constantly exposed to a large and diverse pool of microorganism, which can reside in, on, and around our bodies. These microbiotas and their genomes, collectively called as the microbiome, are being characterized by "metagenomics" approaches that integrate next-generation sequencing (NGS) technologies and bioinformatics analysis. The primary focus is on the assembly of 16S ribosomal RNA hypervariable region called as targeted sequencing or whole-genome shotgun DNA sequencing reads. Such studies have been possible because of advances made in the field of genomics and its constant growth in terms of sequencing technology. Apart from this, assembly algorithms and annotation pipelines have provided key opportunities to be exploited by the scientific community. Advances in single-cell genomics, transcriptomics, and metagenomics have revolutionized studies related to cancer genomics, gene expression, metabolic pathway studies, cellular analysis, environmental analysis, and many more areas. There has been tremendous growth in terms

I. Khatri
Leiden University Medical Center, Leiden University, Leiden, The Netherlands

M. Anurag (✉)
Lester & Sue Smith Breast Center & Department of Medicine, Baylor College of Medicine, Houston, TX, USA
e-mail: anurag@bcm.edu

A. Shanker (ed.), *Bioinformatics: Sequences, Structures, Phylogeny*,
https://doi.org/10.1007/978-981-13-1562-6_5

of sequencing, assembly, and annotation at the genomics level. However, for metagenomics there is a critical need to develop new technologies and in-depth analytical approaches. Here, we present a generalized methodology that can be used for sampling and analysis of metagenomics samples acquired from any environmental location.

5.2 Metagenomics: General Methodology

Metagenomics projects utilize various methodologies which depend on the aim, and a standard metagenomics analysis protocol is depicted in Fig. 5.1. The basic steps in metagenomics analysis including sampling, sequencing, metagenome assembly, binning, annotation of metagenomes, experimental procedures, statistical analysis, and data storage and sharing are discussed.

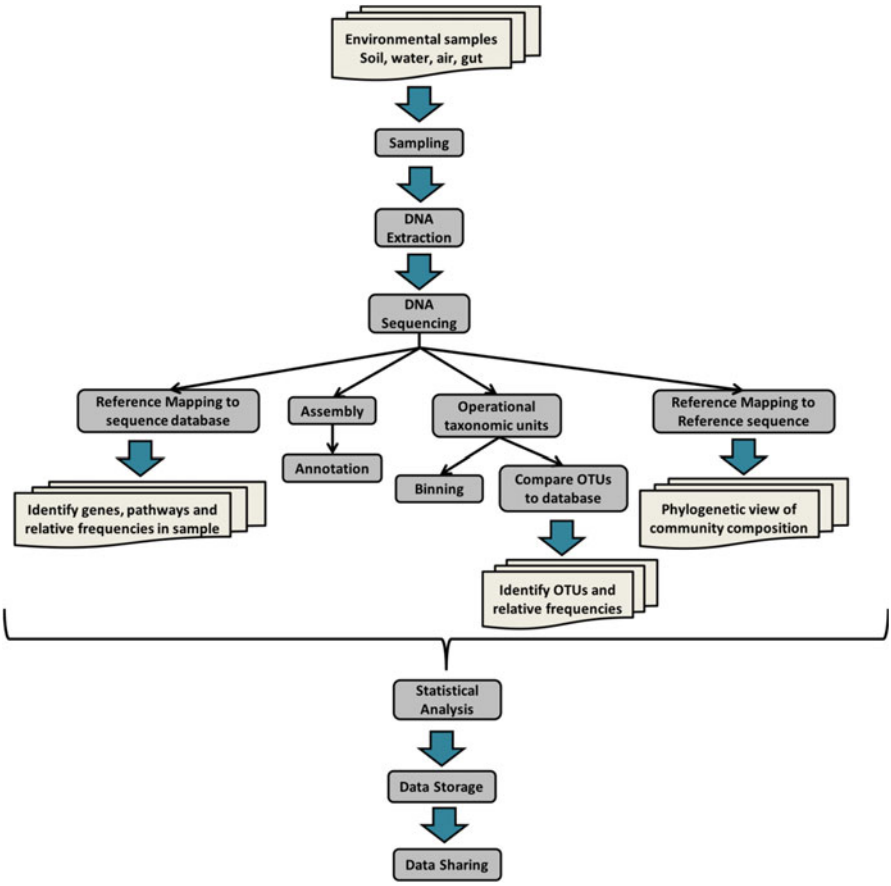

Fig. 5.1 Flow diagram of a typical metagenomic experiment

5.2.1 Sampling and DNA Extraction

The first and most crucial step is sample acquisition, which critically depends on the sample source. Collection of environmental samples from specific sites across various time points is analyzed in relative metagenomics studies which provide significant insight into both temporal and spatial characteristics of microflora. Another important step in a metagenomics data analysis is the processing of the samples efficiently and to ensure that DNA extracted from the sample represents all the cells present in the sample. In addition, special considerations should be given for sampling and DNA extraction which depends specifically on the sample source. For example, in a soil sample, physical separation and isolation of cells are important for maximizing DNA yield or avoid the co-extraction of enzymatic inhibitors which may interfere further in subsequent sample processing (Delmont et al. 2011). Samples from biopsies or groundwater often yield very small amounts of DNA (Singleton et al. 2011); therefore multiple displacement amplification can be performed (Lasken 2009) to amplify femtograms of DNA to micrograms.

Handling of metagenomics data with precision is a challenge for the scientific community due to large data volume leading to storage issues. Metagenomics data can be exploited for various purposes; therefore, strict and comprehensive guidelines are needed to make data publicly available with a proper format known as metadata. The metadata is known as "data about the data" that contains the when, where, and under what conditions the samples were collected. Metadata is as important as sequence data (Wooley et al. 2010), and minimum information about metagenome sequence (MIMS) contains standard formats that minimally describe the environmental and experimental data. The Genomic Standards Consortium (http://gensc.org/), an international group, has standardized the description, the exchange of genomes and metagenomes, and the rules for the associated metadata.

5.2.2 DNA Sequencing

Sequencing technologies revolutionized the genomics and metagenomics field with high-throughput sequencing. Big, dream projects consisting of sequencing genomes have become a relatively routine task owing to advances in NGS, multiplexing, reduced sequencing cost, and improved algorithms. Metagenomics samples are sequenced in the same manner; however, these samples contain both culturable and non-culturable organisms and also many such genera that have not been exploited yet by the field of genomics. The assignment of taxa to a larger percentage of the metagenome data is still a challenge. Currently, the majority of metagenomics analysis deals with sequencing of the 16S rRNA of the microbial community or a particular gene to trace the community composition which is not typical metagenomics and is referred as metagenetics or metabarcoding. In contrast, whole-genome sequencing is performed on metagenomics samples instead of sequencing a single gene. Of the various NGS sequencing technologies, 454/Roche and Illumina/Solexa have been used extensively for sequencing

metagenomics samples. 454/Roche generates longer reads facilitating the assignment of a read to a particular operational taxonomic unit (OTU) which is more reliable as compared to very short reads generated through Illumina high-throughput sequencing.

5.2.3 Assembly and Annotation

The majority of current assemblers have been designed to assemble single, clonal genomes, and their utility for assembling and resolution of large number of complex organisms has to be evaluated critically. Standard assembly methods and algorithms such as de novo assembly and reference mapping are employed in metagenomics data analysis; however, to tackle the significant variation in strain and at the species level, metagenomics assemblers have been designed with the "clonal assumption" that does not allow contig formation for some heterogeneous taxa. Out of various assemblers, *de Bruijn* graph-based assemblers like MetaVelvet (Namiki et al. 2012) and Meta-IDBA (Peng et al. 2011) deal explicitly with non-clonality in sequencing data and try to identify a subgraph that connects related genomes. The meta-assemblers are still in development, and their accuracy assessment is still a major goal of developers as no complete reference exists to which the interpretations can be compared. Assembly is more efficient for genome reconstruction when reference genomes of closely related species are available and in low complex samples (Luo et al. 2013; Teeling and Glockner 2012). However, low read coverage, high frequency of polymorphism, and repetitive regions can hamper the process (De Filippo et al. 2012).

The assembled contigs with minimal length of 30,000 bp or longer can be annotated through existing genome annotation pipelines, such as rapid annotation using subsystem technology (RAST; Aziz et al. 2008) or integrated microbial genomes (IMG; Markowitz et al. 2007, 2009). For the annotation of the entire communities, the standard genome annotation tools are less significant, and a two-step annotation is preferentially followed. First, genes or features of interest are identified, and second, functional assignments are performed by assigning gene functions and taxonomic neighbors (Thomas et al. 2012). FragGeneScan (Rho et al. 2010), MetaGeneMark (McHardy et al. 2007), MetaGeneAnnotator (Noguchi et al. 2008), and Orphelia (Hoff et al. 2009) are the metagenome annotation tools used for defining gene features, e.g., codon usage to find the coding regions. Also, nonprotein-coding genes such as tRNAs (Gardner et al. 2009; Lowe and Eddy 1997), signal peptides (Bendtsen et al. 2004), or clustered regularly interspaced short palindromic repeats (CRISPRs; Bland et al. 2007; Grissa et al. 2007) can be identified but might require long contiguous sequences and vast computational resources.

The functional annotations are provided as gene features via gene or protein mapping to existing nonredundant (NR) protein sequence database. The sequence that cannot be mapped to the known sequence space is termed as ORFans which represents the novel gene contents. ORFans could be erroneous coding sequence (CDS) calls or may be biochemically uncharacterized *bona fide* genes or have no

sequence but structural homology to the existing protein families or folds. The reference databases including Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al. 2004), eggnog (Muller et al. 2010), cluster of orthologous groups/eukaryotic orthologous groups (COG/KOG; Tatusov et al. 2003), PFAM (Finn et al. 2014), and TIGRFAM (Selengut et al. 2007) are used to provide functional context to metagenomics CDS. Three prominent systems Metagenome-RAST (MG-RAST; Glass et al. 2010), integrated microbial genomes and microbiomes (IMG/M; Markowitz et al. 2007), and CAMERA (Sun et al. 2011) perform quality control, feature prediction, and functional annotation through standardized protocols and also serve as large repositories of metagenomics datasets. These web servers have a graphical user-friendly interface that assists users to perform taxonomical and functional analysis of metagenomes, which, unfortunately, might be saturated and not customizable at times. Earlier it was reported that the standard metagenome annotation tools can only annotate 20–50% of the metagenomics sequences (Gilbert et al. 2010) and requires further refinement in the annotation algorithms, where sequence and structural homology can be taken into account altogether which is the major computational challenge.

Pathway reconstruction, one of the annotation goals, could be achieved reliably if there is robust functional annotation. To reconstruct a pathway, every gene should be in an apt metabolic context, missing enzymes should be filled in the pathways, and optimal metabolic states should be found. MinPath (Ye and Doak 2009) and MetaPath (Liu and Pop 2011) use KEGG (Kanehisa et al. 2004) and MetaCyc (Caspi et al. 2014) repositories for building networks. Most of the current platforms are not able to reconstruct variant metabolic pathways (de Crécy-Lagard 2014), since pathways and enzymes are not conserved among different environment and the inhabiting species. A web service implementation by KEGG, GhostKOALA (Kanehisa Laboratories www.kegg.jp/ghostkoala/), relates taxonomic origin of the metagenomes with their respective functional annotation, and the metabolic pathways from different taxa can be visualized in a composite map. Metabolic pathways can be constructed using gene-function interactions, synteny, and copy number of annotated genes and integrating them with the metabolic potential of metagenome consortium.

5.2.4 Taxonomic Classification and Binning

Binning, as name suggests, is to group the sequencing reads representing an individual genome or genomes of closely related organisms. The algorithms employed in grouping related sequences act either as supervised classifiers or unsupervised classifiers. Binning can be performed based on either sequence similarity/alignment or compositional features or both. Another strategy employed by tools is compositional binning that bins the genomes based on the property of conserved nucleotide composition that carry weak but detectable phylogenetic signals, e.g., GC content or particular K-mer (tetramer or hexamer) abundance distribution (Pride et al. 2003), or based on similarity-based binning where the unknown DNA fragments are binned

according to the known genes in the reference database. Compositional-based binning algorithms have been exploited in PhyloPythia (McHardy et al. 2007) and PCAHIER (Zheng and Wu 2010), whereas a similarity-based binning algorithm was employed in IMG/M (Markowitz et al. 2007), MG-RAST (Glass et al. 2010), MEtaGenome ANalyzer (MEGAN; Huson et al. 2016), CARMA (Krause et al. 2008), MetaPhyler (Liu et al. 2010), and many more. Some programs such as PhymmBL (Brady and Salzberg 2009) and MetaCluster (Leung et al. 2011) employ both compositional- and similarity-based algorithms. All these tools employ either an unsupervised or supervised approach to define the bins. The compositional-based binning is not reliable for short reads of approximately 100 bp length, but if reference data is available, then with supervised similarity-based method, the taxonomic assignment of the read can be made (McHardy et al. 2007). The bins obtained will be assigned taxonomy at the phylum level which is very high and results in chimeric bins composed of two or more genomes that belong to the same phylum. The similarity-based binning algorithm if improved to assignments at lower taxonomic levels may help in creating accurate bins for a specific organism at least to a species level. Such binned reads can be assembled to obtain partial genomes of yet-uncultured or unknown organisms. The binning of reads before assembling reduces the complexity of assembly efforts and computational requirements.

The metabolic potential of the metagenome can be deciphered after the microbial diversity is known. Whole-metagenome approach where whole DNA of the community is sequenced can be used to obtain the complete information of a microbial community. The choice of sequencing platform will influence the computational resources and selection of available software to process the sequencing results. These choices in turn will be reflected in taxonomic species/genus/family level classification. Novel microorganisms identified from the analysis can potentially establish new genes with novel functions.

Taxonomic annotation can be made better by using more than one phylogenetic marker. Metagenome shotgun sequencing allows for the identification of single copy marker genes among various databases. Parallel-META (Su et al. 2014) can be used to extract ribosomal marker genes from metagenomics sequences to conduct taxonomic annotations. Single copy marker genes can be extracted using MOCAT (Kultima et al. 2012) that uses the RefMG database (Ciccarelli et al. 2006), a collection of 40 single copy universal marker genes, and "a pipeline for AutoMated PHylogenOmic infeRence" (AMPHORA; Wu and Eisen 2008), a database with 31 single copy marker genes. This pipeline, distinct from identification of marker genes, performs multiple sequence alignment, distance calculations, and clustering. The reference genomes were used to perform taxonomic annotation at a species-level resolution.

5.2.5 Statistical Analysis

The metagenomics data consists of large number of species, corresponding genes, and their functions as compared to the number of samples analyzed. Thus, multiple hypotheses are to be formed, tested, and implemented for comprehensive presentation

of data. Various multivariate statistical visualization programs such as Metastats (White et al. 2009) and R packages, viz., ShotgunFunctionalizeR (Kristiansson et al. 2009), have been built to statistically analyze the metagenome data.

5.2.6 Data Storage and Sharing

Genome research has always been connected to sharing raw data, the final assemblies and annotations; however, to store metagenomics data, database management and storage system are required. All the data is stored at the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and other metagenomics repositories. The digital form of data storage is generally preferred, and despite the decreasing cost of generating NGS data, storage costs may not decline (Weymann et al. 2017); therefore, acquiring data storage in a cost-effective manner is also important.

The microbial systems can be very dynamic at different time points, e.g., as in the human gut; therefore, temporal sampling has substantial impact on data analysis, interpretations, and results (Thomas et al. 2012). Due to the magnitude of variation in small-scale experiments (Prosser 2010), a sufficient number of replicates are needed. Samples should be collected from the same habitat and should be processed in a similar fashion. The experimental plan and interpretations, if done carefully, facilitate dataset integration into new or existing theories (Burke et al. 2011). The critical aim of metagenomics projects is to relate functional and phylogenetic information to the biological, chemical, and physical characteristics of that environment and ultimately achieve retrospective correlation analysis.

5.3 Species Diversity

The diversity of species in an environmental sample is a critical question where the vast majority of marker genes have been used to classify metagenomics reads. Species-specific gene markers such as 16S/18S ribosomal DNA (rDNA) sequences have been used to estimate the species diversity and coverage in most of the analyses. rDNA as a marker gene has limitations including horizontal transfers within microbes (Schouls et al. 2003) and the presence of multiple copies of the marker gene (DeSantis et al. 2006). Other housekeeping genes such as *rpoB* (Walsh et al. 2004) are strong candidates, and also *amoA*, *pmoA*, *nirS*, *nirK*, *nosZ*, and *pufM* (Case et al. 2007) have been exploited in different contexts as molecular markers.

Quantifying species diversity is not trivial due to the incorporation of species richness, evenness of species, or differential abundance (Simpson 1949). In comparison of two communities, if both the communities have the same number of species but their abundance varies, then the community with the shortest difference with "assumed even abundance" will be considered as more diverse.

The diversity indices of the species are measured as α-diversity, β-diversity, and γ-diversity in ecology and microbial ecology. The α-diversity is defined as the biodiversity in a defined habitat (i.e., a smaller ecosystem), whereas β-diversity compares species diversity between habitats (or between two ecosystems).The γ-diversity is considered as the total biodiversity over a large region containing several ecosystems (Wooley et al. 2010). Rarefraction curves are used to estimate the coverage obtained from sampling which tells whether the species in a particular habitat has been exhaustively sampled or not. All these indices are calculated in metagenomics data analysis by employing various software and tools including EstimateS (Colwell et al. 2004), Quantitative Insights Into Microbial Ecology (QIIME; Caporaso et al. 2010), and Kraken (Davis et al. 2013). Another method to calculate species diversity is through the use of statistical estimators, in particular nonparametric estimators. Simpson's index (Simpson 1949) is based on the probability of the same species taken randomly from the community and is used to assign two independent subjects. The Shannon–Wiener index H' (Shannon 1948) is an entropy measurement and is directly proportional to the number of species in the sample. These methods are used for heterogeneity measurements and differ primarily in calculating the taxa abundance to measure the final richness estimation (Escobar-Zepeda et al. 2015). Simpson and Shannon–Wiener indices prioritize more-frequent and rare species, respectively, in the sample (Krebs 2014).

The use of diversity indices which quantify and compare microbial diversity among samples is a better approach as compared to ones based on molecular markers. The species diversity analysis should be done carefully as it can be uninformative. The biases related to sampling should be reduced considering the criteria for species or OTU definition.

5.4 Comparative Metagenomics

The comparison between two or more metagenomes facilitates the understanding of genomic differences and how they are affected by the abiotic environment. Various sequence-based traits such as GC content (Yooseph et al. 2007), microbial genome size (Raes et al. 2007), taxonomy (von Mering et al. 2007), and functional content (Turnbaugh et al. 2006) have been compared to gather biological insights through comparison between two or more metagenomes. Statistical analysis is a necessity to analyze several metagenomics datasets, and principal component analysis (PCA) and nonmetric multidimensional scaling (NM-MDS) have been used to visualize the metagenomics data analysis and reveal major factors that affect the data most (Brulc et al. 2009).

5.5 Challenges in Metagenomics Analysis

Sequencing of a complex environmental community for metagenomics analysis often represents only a minute fraction of the vast number of culturable and unculturable microorganisms actually present (Desai et al. 2012). To obtain just onefold coverage of the entire community in a gram of soil requires hundreds of millions of reads without guarantee that every member of that community was sequenced. The unknown community composition and relative abundance of microorganisms limits our ability to calculate the coverage robustly. Even perfect 16S amplicon-based characterization of microbial species fails to distinguish between different strains (Desai et al. 2012). Furthermore, no tools are available that determine the availability of sufficient coverage to interpret data of a certain depth for a community. The low coverage data represents randomly subsampled genomic content of the community. Despite complete coverage with millions invested, the analysis of metagenomics data requires tools and protocol development comparable to genomic analysis. Moreover, if the approaches led to the identification of new microbial community members and discovery of new molecules, problems associated with cloning biases, sampling biases, misidentification of "decorating enzymes" and incorrect promoter sites in genomes, and dispersion of genes involved in secondary metabolite production (Escobar-Zepeda et al. 2015) should be considered.

Similarly, human metagenomic experiments and analysis also have associated limitations and pitfalls as they are sensitive to the environment including any particular condition or intervention (Kim et al. 2017). Various factors including diet, drugs, age, geography, and sex have all been reported to influence function and composition of the human microbiome (Blaser et al. 2013;Dave et al. 2012; Lozupone et al. 2012). Another challenge is the longitudinal stability. Unlike gut, the microbiome of other sites, like the human vagina, can vary in short periods without always indicating dysbiosis (Williams and Lin 1971). In animal experiments, the prime limitation is the cage effect, which is best studied in mice kept in the same cage and can share the same microbiome because of coprophagia (Campbell et al. 2012). When it comes to handling and analyzing samples, issues pertaining to low microbial biomass, environmental contamination, and presence of negative/positive control samples should be addressed. The major informatics challenges associated with human metagenome analysis, similar to other metagenomes, are the large volume and bulkiness of the data and the heterogeneous microbial community. One additional challenge has been the rapid identification of host sequences contaminating metagenomics datasets, which is time- and memory-extensive process and hence needs to be revisited. There have been efforts to overcome these challenges with tools like CS-SCORE (Haque et al. 2015); however, algorithm improvement is needed.

5.6 Applications of Metagenomics

5.6.1 Correlations Between Environmental Data and Metadata

Metagenomics studies aid in investigating genomic potential of the bacterial community and how it is affected by and is affecting its habitat. The correlation between sequence data, environment, and environmental attributes or their correlation among themselves reveals new biological insights. For example, a bivariate metagenome study in obese vs lean mouse reveals that obese individuals are enriched in carbohydrate-active enzymes (Turnbaugh et al. 2006). Multivariate correlation analysis in a nutrient poor ocean habitat revealed covariation in amino acid transport and cofactor synthesis molecules (Gianoulis et al. 2009).

5.6.2 Investigating Symbiosis

Symbiotic relationships occur when two or more organisms are symbionts which represent a small-scale metagenomics and can be analyzed in a similar fashion. The organisms in symbiotic relations are few, and their distance to each other phylogenetically eases the binning of the reads in separate bins and can be assembled separately. Wu and collegues (2006) exploited a similar method to bin the ESS data from bacterial symbionts living in the glassy-winged sharpshooter and inferred that one member of a symbiont synthesizes amino acids for the host insect, while the other produces cofactors and vitamins (Wu et al. 2006).

5.6.3 Gene Family Enrichment

The immense amount of genetic material has led to the possibility of associating new gene families with new members of existing gene families. The small bacterial eukaryotic protein kinase-like (ELK) gene family was enriched severalfold through the Global Ocean Sampling (GOS) metagenomics project (Wooley et al. 2010).

5.6.4 Human Microbiome

Symbiotic microbes have coevolved with humans for millions of years and play a critical role in health of the host. The focus of human microbiome research has been on the bacteria residing in the gut, which represents the most abundant and diverse part of the human microbiome (Consortium 2012). Colonization of these bacteria commences at birth, and the method of delivery (i.e., vaginal or cesarean section) influences the basal community (Dominguez-Bello et al. 2010). Early-life events, such as mode of delivery (Fig. 5.2 – adapted from Rutayisire et al. 2016), dietary transitions or restrictions (Bergstrom et al. 2014; Rutayisire et al. 2016), and antibiotic use (Cho et al. 2012), shape the dynamic microbiome of infants. This gradually

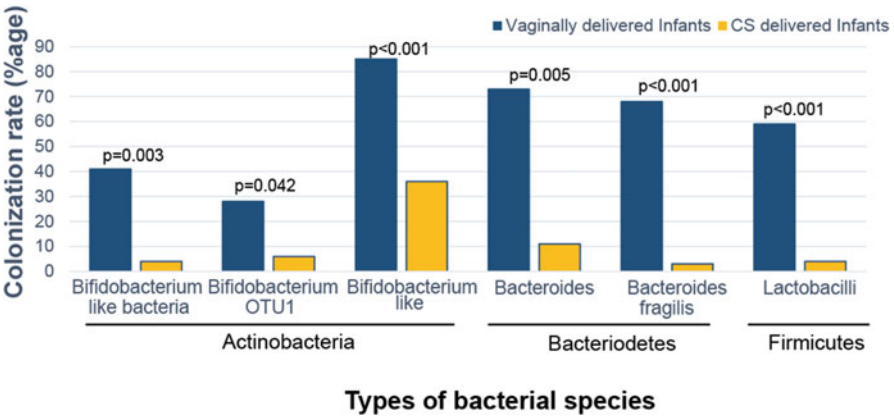

Fig. 5.2 Microbiota colonization pattern significantly associated with the mode of delivery during the first 7 days after birth. Bacterial species with quantified colonization rate has been shown. (Adapted from Rutayisire et al. 2016)

stabilizes with age and leads to adult gut microbiota, which is highly resilient to minor perturbations. This longitudinal stability, collectively with vast interpersonal diversity of the microbiome, allows identification of ~80% individuals by their distinct "microbial fingerprint" (Franzosa et al. 2015). The human microbiota communities contribute to various host biological processes, thus deeply influencing human health. Global initiatives have been taken to understand the healthy microbiome and its composition.

5.6.5 Metagenomics in Diseases

Recent findings have emphasized the effect of gut microbiome in human health and therapeutic response (Scarpellini et al. 2015). The gut microbiome, primarily, is composed of viruses and fungi and has been shown to be modulated in diet-associated insulin resistance in type 2 diabetic patients using a metagenome-wide association analysis (Qin et al. 2012). Gut microbiota has been established as a metformin action site, and metformin–microbiota interactions have been studied to show that altered gut microbiota mediates some of metformin's antidiabetic effects (Wu et al. 2017). The Human Pan-Microbe Communities (HPMC) database (http://www.hpmcd.org/) is an excellent source of highly curated, searchable, metagenomic resource focusing on facilitating the investigation of human gastrointestinal microbiota (Forster et al. 2016).

Historically, cancer has been associated with different forms of microorganisms. The metagenomics era has revolutionized microbiome profiling which helps to boost a number of studies exploring microbial linkage to cancer. Several studies on microbes and cancers have shown distinct associations between various viruses and different types of cancers. Human papilloma virus (HPV) causes the majority of

cervical, anal, and oropharyngeal cancer (Chaturvedi et al. 2011; Daling et al. 2004; Gillison et al. 2008; Winer et al. 2006). Similarly, Epstein–Barr virus has been found to be responsible for nasopharyngeal carcinoma, Hodgkin's, Burkitt's lymphoma, etc. (Anagnostopoulos et al. 1989; Henle and Henle 1976; Leung et al. 2014).

5.6.6 Clinical Implications

In translating the role of microbiomes into clinical applications, Danino et al. (2015) engineered a probiotic *E. coli* to harbor specific gene circuits that produce signals allowing detection of tumor in urine, in case of liver metastases. This concept was based on the fact that metastasis leads to translocation of the probiotic *E. coli* to the liver. Metagenomics has also allowed physicians to probe complex phenotypes such as microbial dysbiosis with intestinal disorders (Antharam et al. 2013) and disruptions of the skin microbiome that may be associated with skin disorders (Weyrich et al. 2015). Recently, different bacterial profiles in the breast were observed between healthy women and breast cancer patients. Interestingly, higher abundances of DNA damage causing bacteria were detected in breast cancer patients, along with decrease in some lactic acid bacteria, known for their beneficial health effects (Urbaniak et al. 2016). Such studies raise important questions regarding the role of the mammary microbiome in risk assessment to develop breast cancer.

Metagenomics analytics is changing rapidly with evolutions of tools and analysis procedures in terms of scalability, sensitivity, and performance. The field allows us to discover new genes, proteins, and the genomes of non-cultivable organisms with better accuracy and less time as compared to classical microbiology or molecular methods. However, no standard tool or method is available that can answer all our questions in metagenomics. The lack of standards reduces reproducibility and is still a case by case study. The major problem associated with metagenomics study is also data management as most institutes lack computational infrastructure to deal with long-term storage of raw, intermediate data, and final analyzed datasets.

Comparison between different biomes and different environmental locations will provide insight into the microflora distribution and help understand the environment around us.

All the advances in the field of human metagenomics add up to the profound impact that the microbiome and their metagenomics have on human health in providing new diagnostic and therapeutic opportunities. However, existing therapeutic approaches for modulating microbiomes in the clinic remain relatively underdeveloped. More studies focused on metagenomics of different organs need to be performed, comparing the tissues from healthy versus affected individuals. Further exploration of additive, subtractive, or modulatory strategies affecting the human microbiota and its clinical implementation could potentially be the next big milestone in the field of translational and applied microbiology. The near future challenge is in the accurate manipulation and analysis of the vast amounts of data and to develop approaches to interpret data in a more integrative way that will reflect the

biodiversity present in our world. The development of more bioinformatics tools for metagenomics analysis is necessary, but the expertise of scientific community to manipulate such tools and interpret their results is a critical parameter for successful metagenomics studies.

References

Anagnostopoulos I, Herbst H, Niedobitek G, Stein H (1989) Demonstration of monoclonal EBV genomes in Hodgkin's disease and Ki-1-positive anaplastic large cell lymphoma by combined Southern blot and in situ hybridization. Blood 74:810–816

Antharam VC, Li EC, Ishmael A, Sharma A, Mai V et al (2013) Intestinal dysbiosis and depletion of butyrogenic bacteria in Clostridium difficile infection and nosocomial diarrhea. J Clin Microbiol 51:2884–2892

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Bergstrom A, Skov TH, Bahl MI, Roager HM, Christensen LB et al (2014) Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of Danish infants. Appl Environ Microbiol 80:2889–2900

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K et al (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinf 8:209

Blaser M, Bork P, Fraser C, Knight R, Wang J (2013) The microbiome explored: recent insights and future challenges. Nat Rev Microbiol 11:213–217

Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods 6:673–676

Brulc JM, Antonopoulos DA, Miller MEB, Wilson MK, Yannarell AC et al (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc Natl Acad Sci U S A 106:1948–1953

Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T (2011) Bacterial community assembly based on functional genes rather than species. Proc Natl Acad Sci 108:14288–14293

Campbell JH, Foster CM, Vishnivetskaya T, Campbell AG, Yang ZK et al (2012) Host genetic and environmental effects on mouse intestinal microbiota. ISME J 6:2033–2044

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336

Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 73:278–288

Caspi R, Altman T, Billington R, Dreher K, Foerster H et al (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 42:D459–D471

Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W et al (2011) Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J Clin Oncol 29:4294–4301

Cho I, Yamanishi S, Cox L, Methe BA, Zavadil J et al (2012) Antibiotics in early life alter the murine colonic microbiome and adiposity. Nature 488:621–626

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

Colwell RK, Mao CX, Chang J (2004) Interpolating, Extrapolating, and comparing incidence-based species accumulation curves. Ecology 85:2717–2727

Consortium THMP (2012) Structure, function and diversity of the healthy human microbiome. Nature 486:207–214

Daling JR, Madeleine MM, Johnson LG, Schwartz SM, Shera KA et al (2004) Human papillomavirus, smoking, and sexual practices in the etiology of anal cancer. Cancer 101:270–280

Danino T, Prindle A, Kwong GA, Skalak M, Li H et al (2015) Programmable probiotics for detection of cancer in urine. Sci Transl Med 7:289ra284

Dave M, Higgins PD, Middha S, Rioux KP (2012) The human gut microbiome: current knowledge, challenges, and future directions. Transl Res: J Lab Clin Med 160:246–257

Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ (2013) Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Methods 63:41–49

de Crécy-Lagard V (2014) Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. Comput Struct Biotechnol J 10:41–50

De Filippo C, Ramazzotti M, Fontana P, Cavalieri D (2012) Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Brief Bioinform 13:696–710

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011) Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods 86:397–400

Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F (2012) From genomics to metagenomics. Curr Opin Biotechnol 23:72–76

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072

Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G et al (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A 107:11971–11975

Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A (2015) The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front Genet 6:348

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Forster SC, Browne HP, Kumar N, Hunt M, Denise H et al (2016) HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic Acids Res 44:D604–D609

Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP et al (2015) Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112:E2930–E2938

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL et al (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136–D140

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO et al (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A 106:1374–1379

Gilbert JA, Field D, Swift P, Thomas S, Cummings D et al (2010) The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. PLoS ONE 5:e15545

Gillison ML, Chaturvedi AK, Lowy DR (2008) HPV prophylactic vaccines and the potential prevention of noncervical cancers in both men and women. Cancer 113:3036–3046

Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc 2010: pdb.prot5368

Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35:W52–W57

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245–R249

Haque MM, Bose T, Dutta A, Reddy CV, Mande SS (2015) CS-SCORE: rapid identification and removal of human genome contaminants from metagenomic datasets. Genomics 106:116–121
Henle G, Henle W (1976) Epstein-Barr virus-specific IgA serum antibodies as an outstanding feature of nasopharyngeal carcinoma. Int J Cancer 17:1–7
Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37:W101–W105
Huson DH, Beier S, Flade I, Górska A, El-Hadidi M et al (2016) MEGAN community edition – interactive exploration and analysis of large-scale microbiome sequencing data. PLOS Comput Biol 12:e1004957
Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:277D–280D
Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C et al (2017) Optimizing methods and dodging pitfalls in microbiome research. Microbiome 5:52
Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW et al (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 36:2230–2239
Krebs C (2014) Species diversity measures. In: Ecological methodology. Addison-Wesley Educational Publishers, Inc, Boston
Kristiansson E, Hugenholtz P, Dalevi D (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics 25:2737–2738
Kultima JR, Sunagawa S, Li J, Chen W, Chen H et al (2012) MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS ONE 7:e47656
Lasken RS (2009) Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochem Soc Trans 37:450–453
Leung HCM, Yiu SM, Yang B, Peng Y, Wang Y et al (2011) A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics 27:1489–1495
Leung SF, Chan KC, Ma BB, Hui EP, Mo F et al (2014) Plasma Epstein-Barr viral DNA load at midpoint of radiotherapy course predicts outcome in advanced-stage nasopharyngeal carcinoma. Ann Oncol 25:1204–1208
Liu B, Pop M (2011) MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. BMC Proc 5:S9
Liu B, Gibbons T, Ghodsi M, Pop M (2010) MetaPhyler: taxonomic profiling for metagenomic sequences. In: 2010 I.E. international conference on Bioinformatics and Biomedicine (BIBM). IEEE, Hong Kong, pp 95–100
Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964
Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R (2012) Diversity, stability and resilience of the human gut microbiota. Nature 489:220–230
Luo C, Rodriguez-R LM, Konstantinidis KT (2013) A user's guide to quantitative and comparative analysis of metagenomic datasets. Methods Enzymol 531:525–547
Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K et al (2007) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36:D534–D538
Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu K, Kyrpides NC (2009) IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 25:2271–2278
McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods 4:63–72
Muller J, Szklarczyk D, Julien P, Letunic I, Roth A et al (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res 38:D190–D195
Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40:e155–e155

Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15:387–396

Peng Y, Leung HCM, Yiu SM, Chin FYL (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27:i94–i101

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res 13:145–158

Prosser JI (2010) Replicate or lie. Environ Microbiol 12:1806–1810

Qin J, Li Y, Cai Z, Li S, Zhu J et al (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490:55–60

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8:R10

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38:e191–e191

Rutayisire E, Huang K, Liu Y, Tao F (2016) The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life: a systematic review. BMC Gastroenterol 16:86

Scarpellini E, Ianiro G, Attili F, Bassanelli C, De Santis A, Gasbarrini A (2015) The human gut microbiota and virome: Potential therapeutic implications. Dig Liver Dis 47:1007–1012

Schouls LM, Schot CS, Jacobs JA (2003) Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol 185:7241–7246

Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M et al (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res 35:D260–D264

Shannon CE (1948) A mathematical theory of communication, Part I. Bell Syst Tech J 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Simpson EH (1949) Measurement of diversity. Nature 163:688

Singleton DR, Richardson SD, Aitken MD (2011) Pyrosequence analysis of bacterial communities in aerobic bioreactors treating polycyclic aromatic hydrocarbon-contaminated soil. Biodegradation 22:1061–1073

Su X, Pan W, Song B, Xu J, Ning K (2014) Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. PLoS ONE 9:e89323

Sun S, Chen J, Li W, Altintas I, Lin A et al (2011) Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Res 39:D546–D551

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform 4:41

Teeling H, Glockner FO (2012) Current opportunities and challenges in microbial metagenome analysis – a bioinformatic perspective. Brief Bioinform 13:728–742

Thomas T, Gilbert J, Meyer F (2012) Metagenomics – a guide from sampling to data analysis. Microb Inf Exp 2:3

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1131

Urbaniak C, Gloor GB, Brackstone M, Scott L, Tangney M, Reid G (2016) The Microbiota of Breast Tissue and Its Association with Breast Cancer. Appl Environ Microbiol 82:5039–5048

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T et al (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315:1126–1130

Walsh DA, Bapteste E, Kamekura M, Doolittle WF (2004) Evolution of the RNA polymerase B′ subunit gene (rpoB′) in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. Mol Biol Evol 21:2340–2351

Weymann D, Laskin J, Roscoe R, Schrader KA, Chia S, Yip S, Cheung WY, Gelmon KA, Karsan A, Renouf DJ, Marra M, Regier DA (2017) The cost and cost trajectory of whole-

genome analysis guiding treatment of patients with advanced cancers. Mol Genet Genomic Med 5:251–260
Weyrich LS, Dixit S, Farrer AG, Cooper AJ, Cooper AJ (2015) The skin microbiome: associations between altered microbial communities and disease. Aust J Dermatol 56:268–274
White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5:e1000352
Williams HR, Lin TY (1971) Methyl- 14 C-glycinated hemoglobin as a substrate for proteases. Biochim Biophys Acta 250:603–607
Winer RL, Hughes JP, Feng Q, O'Reilly S, Kiviat NB et al (2006) Condom use and the risk of genital human papillomavirus infection in young women. N Engl J Med 354:2645–2654
Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6: e1000667
Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9:R151
Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL et al (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol 4:e188
Wu H, Esteve E, Tremaroli V, Khan MT, Caesar R et al (2017) Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. Nat Med 23:850–858
Ye Y, Doak TG (2009) A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput Biol 5:e1000465
Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ et al (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16
Zheng H, Wu H (2010) Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. J Bioinform Comput Biol 8:995–1011