

Chapter 12

Methods for Microbiome Analysis

Kalibulla Syed Ibrahim and Nachimuthu Senthil Kumar

Abstract Metagenomics is gaining importance as an invaluable tool as it attempts to determine directly the whole collection of genes and analyze from microbes in a particular environment where they interact with each other by exchanging nutrients, metabolites, and signaling molecules. The development of affordable next-generation sequencers has led to democratization of sequencing, but their ever-growing throughput is making data analysis increasingly complex. This has introduced a plethora of challenges with respect to design of experiments, bioinformatics, and downstream processing. This chapter aims to provide an overview of the currently available methodologies and tools for performing every individual step of a typical metagenomic data set analysis and expected to serve as a useful resource for microbial ecologists and bioinformaticians.

Keywords 16S • Analysis pipeline • Bioinformatics • Genome annotation • Human microbiome • Metagenomics • Metatranscriptomics • Next-generation sequencing

12.1 Introduction

Microorganisms make up only 1 to 2% of the mass of the body of a healthy human, but they are suggested to outnumber human cells by 10 to 1 and to outnumber human genes by 100 to 1. The majority of microbes were identified to inhabit the gut and have profound influence on human well-being (Bäckhed et al. 2005). It has been recognized that microbes play major roles in maintaining health and causing illness, but relatively little is known about the role that microbial communities play in human health and disease (Cho and Blaser 2012; Lampe 2008). The knowledge about the human microbiome that we currently possess is from culture-based approaches using the 16S rRNA technology. However, it has to be noted around 20–60% of the microbiome associated with human is uncultivable (Peterson et al.

K.S. Ibrahim • N.S. Kumar (✉)

Department of Biotechnology, Mizoram University, Aizawl, Mizoram 796 004, India
e-mail: syedibrahim.k@gmail.com; nskmzu@gmail.com

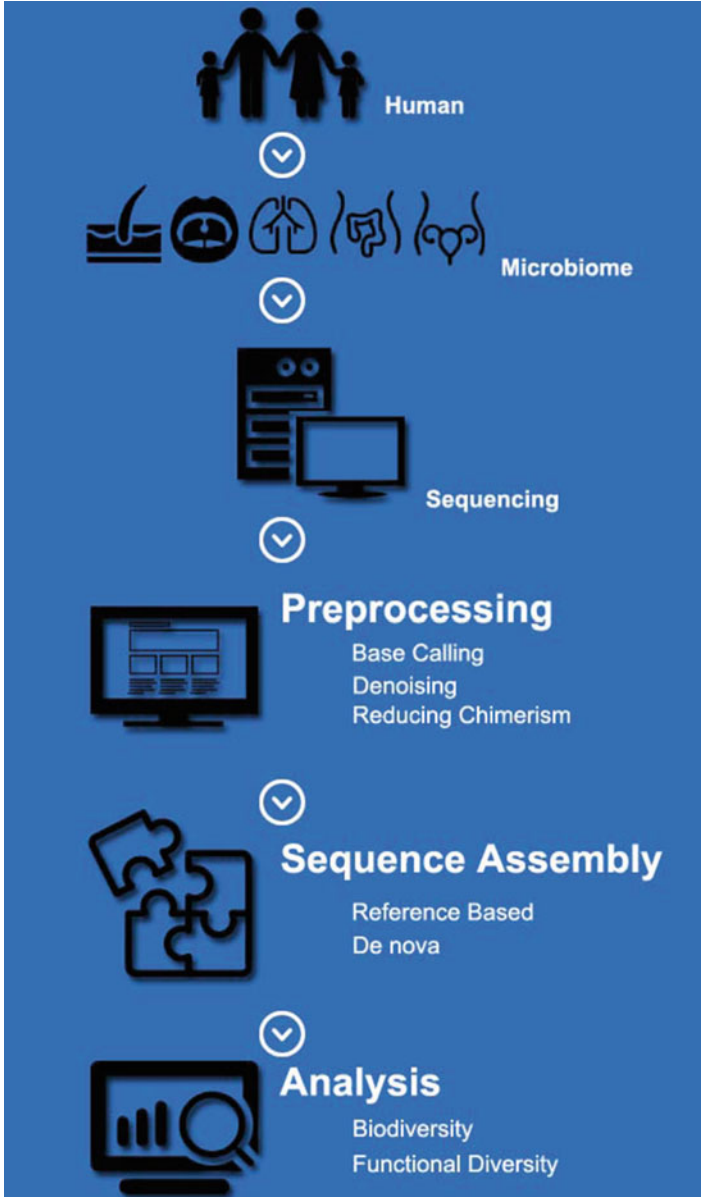


Fig. 12.1 Overall workflow of human microbiome analysis

2009). Projects such as Human Microbiome Project and MetaHIT (Qin et al. 2010) were launched with an intention to generate resources to enable a comprehensive characterization of the human microbiota and analysis of its role in human health

and disease. Figure 12.1 provides an overview of the methods involved in human microbiome analysis.

Metagenomics, the term coined by Handelsman et al. (1998), made it possible for direct genetic analysis of species that are refractory to culturing methods. Using metagenomics, several types of ecosystems including extreme environments and low-diversity environments have been studied so far (Oulas et al. 2015). Decoding the metagenome and its comprehensive genetic information can also be used to understand the functional properties of the microbial community besides studying population ecology. This has provided an infinite capacity for bioprospecting that allowed the discovery of novel compounds of biotechnological commercialization (Segata et al. 2011). Initially metagenomics was used mainly to identify novel biomolecules from environmental microbial assemblages (Chistoserdova 2010). But the advent of next-generation sequencing techniques at affordable costs has allowed for more comprehensive examination of microbial communities such as comparative community metagenomics, metatranscriptomics, and metaproteomics (Simon and Daniel 2010).

In order to disentangle complex ecosystem functions of the microbial communities and fulfill the promise of metagenomics, the comprehensive data sets derived from the next-generation sequencing technologies require intensive analyses (Scholz et al. 2011). This demand has created the need for more powerful tools and software that have unprecedented potential to shed light on ecosystem functions of microbial communities and evolutionary processes.

12.2 Sequence Processing

Compared to conventional Sanger sequencing, several next-generation sequencing platforms provide huge data at much lower recurring cost. Though these technologies include a number of methods like template preparation, sequencing and imaging, and data analysis in common, it is the unique combination of specific protocols that distinguishes one technology from another. Besides that, it also determines the type of data produced from each platform, posing challenges when comparing platforms based on data quality and cost. As these new sequencing technologies produce hundreds of megabases of data at affordable costs, metagenomics is within the reach of many laboratories. The metagenomic analysis workflow begins with sampling and metadata collection and then proceeds with DNA extraction, library construction, sequencing, read preprocessing, and assembly. Either for reads, contigs, or both, binning is applied. Community composition analysis is made using databases. Some details of the workflow will be different in different sequencing facilities.

One has to take greater care when processing sequences of metagenomic data sets than when processing genomic data sets because in the later there is no fixed end point and lacks many of the quality assurance procedures (Kunin et al. 2008).

12.2.1 Preprocessing

Preprocessing of sequence reads is a critical and largely overlooked aspect of metagenomic analysis. Preprocessing comprises the base calling of raw data coming off the sequencing machines, vector screening to remove cloning vector sequence, quality trimming to remove low-quality bases (as determined by base calling), and contaminant screening to remove verifiable sequence contaminants. Errors in each of these steps can have greater downstream consequences in metagenomes.

12.2.2 Sources of Bias and Error in 16S rRNA Gene Sequencing and Reducing Sequencing Error Rates

Irrespective of the technologies used, the scientist needs to understand the quality of their data and how to reduce errors that affect downstream analyses. Two main categories of errors that are commonly observed with 16S sequencing are due to misrepresentation of the relative abundances of microbial populations in a sample (bias) and misrepresentation of an actual sequence itself due to PCR amplification and sequencing (error) (Schloss et al. 2011). Misrepresentation of the relative abundances might be due to DNA extraction method (Miller et al. 1999), PCR primer and cycling conditions, 16S rRNA gene copy number, and the actual community composition in the original sample (Hansen et al. 1998). On the other hand, error due to misrepresentation of an actual sequence is due to PCR polymerases that typically have error rates of one substitution per 105–106 bases (Cline et al. 1996), risk of chimera formation (Haas et al. 2011), and errors introduced by sequencers (Margulies et al. 2005). Because of their relative rates, sequencing errors and chimeras are of the most concern (Schloss et al. 2011).

Sequencing errors can be reduced by the following ways: removing sequence associated with low-quality scores, removing ambiguous base calls, removing mismatches to the PCR primer, or removing sequences that were shorter or longer than expected. Besides these, using denoising and removing sequences that cannot be taxonomically classified are also followed. But the later generally reduce the number of spurious OTUs and phylotypes and do not minimize the actual error rate. Laehnemann et al. (2015) has reported an extensive survey of the errors that are generated during sequencing by the commonly used high-throughput sequencing platforms.

12.2.3 *Base Calling and Quality Trimming*

Base calling involves identifying DNA bases from the readout of a sequencing machine. Popular base caller widely used is Phred (Ewing et al. 1998). The quality score, q , assigned to a base is related to the estimated probability, p , of erroneously calling the base by the following formula: $q = -10 \times \log^{10}(p)$. Thus, a Phred quality score of 20 corresponds to an error probability of 1%. Paracel's TraceTuner (www.paracel.com) and ABI's KB (www.appliedbiosystems.com) are the other two frequently used base callers, which behave very similar to Phred by converting raw data into accuracy probability base calls. Since metagenomic assemblies have lower coverage than genomes, errors are more likely to propagate to the consensus. Some post-processing pipelines ignore base quality scores associated with reads and contigs, and few take positional sequence depth into account as a weighting factor for consensus reliability. Because of this, for an average user, low-quality data will be indistinguishable from the rest of the data set. When poor-quality read that inadvertently passed through to gene prediction it may pass into public repositories. Hence, quality trimming is highly recommended.

12.2.4 *Denoising*

Denoising is a computationally intensive process that removes problematic reads and increases the accuracy of the taxonomic analysis. This is critically important for 16S metagenomic data analysis as it may give rise to erroneous OTUs, and it is sequencing platform-specific too. Illumina require less denoising than others. Though generally a considerable number of sequences is lost, it usually results in high-quality sequences (Gaspar and Thomas 2013) at certain level of stringency (Bakker et al. 2012). Notable software packages that are commonly used to correct amplicon pyrosequencing errors include Denoiser (Reeder and Knight 2010), AmpliconNoise (Quince et al. 2011), Acacia (Bragg et al. 2012), DRISSEE (duplicate read inferred sequencing error estimation) (Keegan et al. 2012), JATAC (Balzer et al. 2013), and CorQ (Iyer et al. 2013). Denoiser uses frequency-based heuristics rather than statistical modeling to cluster reads and makes more accurate assessments of alpha diversity when combined with chimera-checking methods. AmpliconNoise is highly effective but is computationally intensive and applies an approximate likelihood using empirically derived error distributions to remove pyrosequencing noise from reads. These two tools do not modify individual reads; rather they both select an "error-free" read to represent reads in a given cluster. Acacia, on the other hand, is an error-correction tool, reduces the number and complexity of alignments, and uses a quicker but less sensitive statistical approach to distinguish between error and genuine sequence differences. DRISSEE assess sequencing quality and provides positional error estimates that can be used to inform read trimming within a sample. JATAC algorithm identifies duplicate reads

based on the flowgram that has been shown to be superior for noise removal in metagenomics amplicon data and also allows for a more effective removal of artificial duplicates. CorQ corrects homopolymer and non-homopolymer insertion and deletion (indel) errors by utilizing inherent base quality in a sequence-specific context.

12.2.5 Reducing Chimerism

Chimeras are fusion products that are formed between multiple parent sequences. These are falsely interpreted as novel organisms. These are not sequencing errors as they are not derived from a single reference sequence to which it can be mapped. Few commonly used programs for combating chimerism are Bellerophon, Pintail (Ashelford et al. 2005), ChimeraSlayer (Haas et al. 2011), Perseus (Quince et al. 2011), and Uchime (Edgar et al. 2011). The two algorithms most widely used for 16S chimera detection are Pintail and Bellerophon. The former is used by the databases like the RDP (Cole et al. 2009) and SILVA (Pruesse et al. 2007) and the latter is used by the GreenGenes 16S rRNA sequence collection (DeSantis et al. 2006). Pintail is generally visualized as 16S anomaly detection tool rather than a chimera detection tool. But interestingly most anomalies detected by Pintail were chimeras (Ashelford et al. 2005). Perseus, unlike Pintail and Bellerophon, does not use a reference database, but does require a training set of sequences similar to the sequences for characterization. Uchime outperformed ChimeraSlayer, especially in cases where the chimera has more than two parents and its performance was comparable to that of Perseus.

12.3 Sequence Assembly

The shotgun sequencing generates sequences for multiple small fragments separately which are then combined into a reconstruction of the original genome using computer programs called genome assemblers. These programs assemble shorter reads first into contigs, and these are then oriented into scaffolds that provide a more compact and concise view of the sequenced community. New challenges for the assembly process are posed by recent advances in genome sequencing technologies in terms of volume of data generated, length of the fragments, and new types of sequencing errors especially in metagenomics (Pop 2009). Earlier metagenomic data assemblies used tools that were originally designed for conventional whole-genome shotgun sequence (WGS) projects with minor parameter modifications (Wooley and Ye 2009). But recent ones have evolved as more robust specifically in handling samples containing multiple genomes. The assembly process can be approached either as reference-based assembly or as de novo assembly.

12.3.1 Reference-Based Assembly

In reference-based assembly, contigs are created by mapping on one or more reference genomes that belong to a particular species or genus, or sequences from closely related organism would have already been deposited in online data repositories and databases. Reference-based assembly tools are not computationally intensive and can perform well when metagenomic samples are derived from the areas that are extensively studied. Tools like GS Reference Mapper (Roche), MIRA 4 (Chevreux et al. 2004) or AMOS, and MetaAMOS (Treangen et al. 2013) are commonly used in metagenomics applications. The assemblies can be visualized using tools such as Tablet (Milne et al. 2009), EagleView (Huang and Marth 2008), and MapView (Bao et al. 2009). Gaps in the query genome(s) of the resulting assembly indicate that the assembly is incomplete or that the reference genomes used are too distantly related to the community under investigation.

12.3.2 De Novo Assembly

On the other hand, de novo assembly is a computationally expensive process requiring hundreds of gigabytes of memory and has long execution times, which assembles the contigs based on the de Bruijn graphs without any reference genome (Miller et al. 2010). Though tools such as EULER (Pevzner et al. 2001), FragmentGluer (Pevzner et al. 2004), Velvet (Zerbino and Birney 2008), SOAP (Li et al. 2008), ABySS (Simpson et al. 2009), and ALLPATHS (Maccallum et al. 2009) were built for assembling a single genome, even today they are used for metagenomics applications. EULER and ALLPATHS attempt to correct errors in reads prior to assembly, while Velvet and FragmentGluer deal with errors by editing the graphs. These often underperform when used for metagenome assemblies due to problems coming from variation between similar subspecies and genomic sequence similarity between different species. Besides that, difference in abundance for species in a sample was also affected by different sequencing depths for individual species. Tools like Genova (Laserson et al. 2011), MAP (Lai et al. 2012), MetaVelvet (Namiki et al. 2012), MetaVelvet-SL (Afiahayati and Sakakibara 2014), and Meta-IDBA (Peng et al. 2011) managed to create more accurate assemblies especially from data sets containing a mixture of multiple genomes by making use of k-mer frequencies to detect kinks in the de Bruijn graph. Using k-mer thresholds, they decompose the graph into subgraphs and further assemble contigs and scaffolds based on the decomposed subgraphs. The IDBA-UD algorithm (Peng et al. 2012) additionally address the issue of metagenomic sequencing technologies with uneven sequencing depths by making use of multiple depth-relative k-mer thresholds in order to remove erroneous k-mers in both low-depth and high-depth regions.

12.4 Analyzing Community Biodiversity

12.4.1 *The Marker Gene*

Microbial community fundamentally is a collection of individual cells, with distinct genomic DNA. In order to describe the community, it is impractical to fully sequence every genome in every cell. Hence, microbial ecology has defined a number of unique tags to distinct genomes called molecular markers. A marker is a small segment of DNA sequence that identifies the genome that contains it, eliminating the need to sequence the entire genome. Despite its numerous varieties, there are some which are desirable for properties for a good marker like it should be present in every member of a population and discriminate individuals with distinct genomes and, ideally, should differ proportionally to the evolutionary distance between distinct genomes.

By far the most ubiquitous and significant (Lane et al. 1985) is the small or 16S ribosomal RNA subunit gene (Tringe and Hugenholtz 2008) as the preferred target marker gene for bacteria and archaea. But in case of fungi and eukaryotes, the preferred marker genes are the internal transcribed spacer (ITS) and 18S rRNA gene, respectively (Oulas et al. 2015). The gold standard (Nilakanta et al. 2014) for the 16S data analysis is QIIME (Caporaso et al. 2010). Yet another popular tool is Mothur (Schloss et al. 2009) which provides the user with a variety of choices by incorporating software such as DOTUR (Schloss and Handelsman 2005), SONS (Schloss and Handelsman 2006a), Treeclimber (Schloss and Handelsman 2006b), and many more algorithms. Other tools include SILVAngs (Quast et al. 2012) and MEGAN (Huson et al. 2007). These marker gene analyses generally involve searching a reference database to find the closest match to an OTU from which a taxonomic lineage is inferred. Some widely utilized databases for 16S rRNA gene analysis include GreenGenes (DeSantis et al. 2006) and Ribosomal Database Project (Cole et al. 2007; Cole et al. 2009). Besides 16S, SILVA (Pruesse et al. 2007) also supports analysis of 18S in case of fungi and eukaryotes. Unite (Koljal et al. 2013) can be used for analyzing ITS.

Unfortunately, not much databases are available for analyzing extremely diverse protists and viruses for which considerably less sequence information is available compared to bacteria. Humans are not only reported to carry viral particles consisting mainly of bacteriophages (Haynes and Rohwer 2011) but also a substantial number of eukaryotic viruses (Virgin et al. 2009). Like bacterial microbiota, viromes show similar patterns in different stages of human (Caporaso et al. 2011; Koenig et al. 2010), but the effects of these patterns in the human virome are mostly not understood, although certain bacteriophages in other animals are beneficial to the host (Oliver et al. 2009). The lack of a universal gene that is present in all virus makes amplicon-based studies difficult for characterizing the virome in its totality.

12.5 Analyzing Functional Diversity

This generally involves identifying protein coding sequences from the metagenomic reads and comparing the coding sequence to a database (for which some functional information is identified) to infer the function based on its similarity to sequences in the database. Besides picturing the functional composition of the community (Looft et al. 2012) or functions that associate with specific environmental or host-physiological variables (Morgan et al. 2012), they may also reveal the presence of novel genes (Nacke et al. 2011) or provide insight into the ecological conditions associated with those genes for which the function is currently unknown (Buttigieg et al. 2013). Functional annotation of metagenome involves two non-mutually exclusive steps: gene prediction and gene annotation.

12.5.1 Gene Prediction

This can be done on assembled or unassembled metagenomic sequences. Metagenomic reads/contigs are scanned for identifying protein coding genes (CDSs), as well as CRISPR repeats, noncoding RNAs, and tRNA. Predicting CDSs from metagenomic reads is a fundamental step for annotation. Gene prediction for metagenomic sequences can be performed in three ways: first, by mapping the metagenomic reads or contigs to a database of gene sequences; second, based on protein family classification; and, third, by de novo gene prediction.

Mapping the metagenomic reads or contigs to a database of gene sequences is a straightforward method of identifying coding sequences in a metagenome. This method of gene prediction can simultaneously provide functional annotation, if functional annotation of the gene is available. It comes under high-throughput gene prediction procedure as the mapping algorithms assess rapidly whether a genomic fragment is nearly identical to a database sequence or not. This method is generally useful for cataloging the specific genes present in the metagenome but not appropriate from predicting novel or highly divergent genes due to underrepresentation of genomes in sequence databases.

The second method is the most frequently used gene prediction procedure where each metagenomic read is translated into all six possible protein coding frames and each of the resulting peptides is compared to a database of protein sequences. Tools like transeq (Rice et al. 2000), USEARCH (Edgar 2010), RAPsearch (Zhao et al. 2011), and lastp (Kielbasa et al. 2011) translate reads prior to conducting protein sequence alignment. On the other hand, algorithms like blastx (Altschul et al. 1997), USEARCH with the ublast option, or lastx (Kielbasa et al. 2011) translate nucleic acid sequences on the fly. As this also relies on database, it can reveal only diverged homologues of known proteins and not useful for identifying novel types of proteins. Common functional databases includes SMART (Schultz et al. 1998), SEED (Overbeek et al. 2005), NCBI nr (Pruitt et al. 2011), the KEGG Orthology

(Kanehisa and Goto 2000), COGs (Tatusov et al. 1997), MetaCyc (Caspi et al. 2012), eggNOGs (Powell et al. 2011), and PFAM (Punta et al. 2011). Integrated pipelines with integrated functional annotation like MG-RAST (Meyer et al. 2008), MEtaGenome ANalyzer (MEGAN) (Huson et al. 2007), and HUMAnN (Abubucker et al. 2012) are also available to automate these tasks.

Contrary to the above two methods, de novo gene prediction does not rely on a reference database for identifying sequence similarity. Rather, gene prediction systems are trained by evaluating various properties of microbial genes like length of the gene, codon usage, GC bias, etc. Hence this method can potentially identify novel genes, but it is difficult to determine if the predicted gene is real or spurious. Tools like MetaGene (Noguchi et al. 2006), MetaGeneAnnotator (Meyer et al. 2008), Glimmer-MG (Kelley et al. 2011), MetaGeneMark (Zhu et al. 2010), FragGeneScan (Rho et al. 2010), Orphelia (Hoff et al. 2009), and MetaGun (Liu et al. 2013) can be used for de novo gene prediction. Yok and Rosen (2011) recommended that gene prediction in metagenomes can be improved when multiple methods are applied to the same data like following a consensus approach. Though time-consuming, this method tends to be more discriminating than 6-frame translation while annotating (Trimble et al. 2012).

RNA genes (tRNA and rRNA) can be predicted using tools like tRNAscan (Lowe and Eddy 1997). Predictions of tRNA predictions are quite reliable, but not the rRNA genes. Other types of noncoding RNA (ncRNA) genes can be detected by comparison to covariance models (Griffiths-Jones et al. 2005) and sequence-structure motifs (Macke et al. 2001). These methods are computationally intensive and take long time for metagenomic data sets. Predicting ncRNAs are usually excluded from downstream analyses because of the complexity due to lack conservation and reliable “ab initio” methods even for isolated genomes.

Errors in gene prediction mainly occur due to chimeric assemblies or frameshifts (Mavromatis et al. 2007). Hence, the quality of the gene prediction normally relies on the quality of read preprocessing and assembly. Though gene prediction can be performed with both assembled reads (contigs) and unassembled reads, it is advised to perform gene calling on both reads and contigs. It was observed that gene prediction methods used on accurately assembled sequences predicted more than 90% when compared to predictions made on unassembled reads which exhibited lower accuracy (~70%) (Mavromatis et al. 2007).

12.5.2 Functional Annotation

Functional annotation of metagenomic data sets are made by comparing predicted genes to existing, previously annotated sequences or by context annotation. Metagenomic data will have complications when predicted proteins are short and lack homologues. Databases that are used for comparing protein sequences include alignment of profiles from the protein families in TIGRFAMs (Selengut et al. 2007), PFAM (Finn et al. 2008), COGs (Tatusov et al. 1997), and RPS-BLAST

(Markowitz et al. 2006). PFAMs allow the identification and annotation of protein domains. TIGRFAM database include models for both domain and full-length proteins. Though COGs also allow the annotation of the full-length proteins, it is not frequently updated like PFAMs and TIGRFAMs. It is also recommended not to assign protein function solely based on BLAST results as there is a potential for error propagation through databases (Kyrpides and Ouzounis 1999). Context-based annotation methods include genomic neighborhood (Overbeek et al. 1999), gene fusion (Marcotte et al. 1999b), phylogenetic profiles (Pellegrini et al. 1999), and coexpression (Marcotte et al. 1999a). It was observed that neighborhood analysis was performed on metagenomic data, which, combined with homology searches, inferred specific functions for 76% of the metagenomic data sets (83% when nonspecific functions are considered) (Harrington et al. 2007) and is expected to be used in predicting protein function in metagenomic data in the future.

12.6 Metatranscriptomic Analysis

Metatranscriptome sequencing has been recently employed to identify RNA-based regulation and expression in human microbiome (Markowitz et al. 2008). Accessing metatranscriptome of the microbiome through metatranscriptomic shotgun sequencing (RNAseq) has led to the discovery and characterization of new genes from uncultivated microorganisms under different conditions. Few investigations (Bikel et al. 2015; Franzosa et al. 2014; Gosalbes et al. 2011; Jorth et al. 2014; Knudsen et al. 2016) have been performed on metatranscriptomics combined with metagenomics. Several technical issues affecting large-scale application of metatranscriptomics are discussed by Bikel et al. (2015). Though metagenomic and metatranscriptomic data provide extensive information about microbiota diversity, gene content, and their potential functions, it is very difficult to say whether DNA comes from viable cells or whether the predicted genes are expressed at all and, if so, under what conditions and to what extent (Gosalbes et al. 2011).

The bioinformatics pipeline for analyzing the data obtained from a metatranscriptomic experiment is similar to the one used in metagenomics. Basically this is also divided in two strategies: mapping sequence reads to reference genomes or pathways to identify the taxonomical classification of active microorganism and the functionality of their expressed genes and de novo assembly of new transcriptomes. For de novo assembly, there are several programs like SOAPdenovo (Li et al. 2009), ABySS (Birol et al. 2009), and Velvet-Oases (Schulz et al. 2012) that have been reported to be successfully applied to the metatranscriptome assembly (Ghaffari et al. 2014; Ness et al. 2011; Schulz et al. 2012; Shi et al. 2011). A program specially developed for de novo transcriptome assembly from short-read RNAseq data, Trinity (Haas et al. 2013), is one of the most used bioinformatics tools to assemble de novo transcriptomes of different species. It is a very efficient and sensitive in recovering full-length transcripts and isoforms (Ghaffari et al. 2014; Luria et al. 2014).

Metatranscriptome analyses involves stepwise approach for detecting the different RNA types, such as rRNAs, mRNAs, and other noncoding RNAs, facilitating the researchers to study them individually. The reads can be then compared against the small subunit rRNA reference database (SSUrd), and later, the remaining unassigned reads can be analyzed with the large subunit rRNA reference database (LSUrd)—the databases compiled from SILVA (Pruesse et al. 2007) or RDP II (Cole et al. 2009). The non-rRNA representation can be then identified from subtracting the LSU rRNA and SSU rRNA reads from the total reads obtained. The non-rRNAs are finally carried forward for functional analyses.

The functional diversity of the microbiome can be predicted by annotating metatranscriptomic sequences with known functions. cDNA sequences with no significant homology with any of the rRNA databases can be searched against the NCBI nr protein database using BLASTX (Altschul et al. 1997). The sequence reads that contain protein coding genes are identified, and their sequences are compared to the coding sequences of protein databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG), protein family annotations (Pfam), gene ontologies (GO), and clusters of orthologous groups (COG). Thus, the function of the query sequence is assigned based on its homology to sequences functionally annotated in all the above mentioned databases.

Pipelines for combined metatranscriptomics with metagenomics include INFERNAL, a powerful tool for predicting small RNA in the metagenomic data (Nawrocki and Eddy 2013). HUMAnN is another automated pipeline, an offline platform, to determine the presence/absence and abundance of microbial pathways and gene families in a community directly from metagenomic sequence. This is done by converting sequence reads into coverage and abundance and finally summarizes the gene families and pathways in a microbial community (Abubucker et al. 2012). Other offline platforms used to analyze metagenomic data include MEGAN (Huson et al. 2007), IMG/M server (Markowitz et al. 2008), RAST (MG-RAST) (Meyer et al. 2008), and JCVI Metagenomics Reports (METAREP) (Goll et al. 2010).

12.7 Statistical Analysis in Metagenomics

Statistical analysis plays critical role in analyzing and interpreting metagenomic data. Even simple metagenomic analysis like estimate of species diversity seems not so straightforward and obviously needs statistical attention due to the artifacts created during the sequencing (discussed earlier).

Often critical statistical analysis precedes with normalization (i.e., normalization to a reference sample), a step that reduces the systematic variance and improves the overall performance for downstream statistical analysis. These include methods like centering, autoscaling, pareto scaling, range scaling, vast scaling, log transformation, and power transformation. Appropriate selection of data pretreatment methods and its significance have been by van den Berg et al. (2006).

Robust data processing algorithms for wide range of analysis are mostly created using repositories available from the open-source R-project (<http://www.R-project.org>) and the R-based bioconductor project (<https://www.bioconductor.org/>). These are widely considered to be the most complete collection of up-to-date statistical and machine learning algorithms (Xia et al. 2009). Common statistical analysis includes missing value estimation, diversity analysis, and univariate and multivariate analysis like directions of variance, cluster analysis, etc.

Missing value exclusion, missing value replacement, and missing value imputation can be identified by probabilistic PCA (PPCA), Bayesian PCA (BPCA), and singular value decomposition imputation (SVDImpute) (Stacklies et al. 2007; Steinfath et al. 2008). Univariate analysis includes three commonly used methods—fold-change analysis, *t*-tests, and volcano plots. The *t*-test attempts to determine whether the means of two groups are distinct. With *t*-value, *P*-value can be calculated which can be used to determine whether the distinction is statistically significant or not. The volcano plots compare the size of the fold change to the statistical significance level (Xia et al. 2009). Directions of maximum variance can be determined by principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). PCA is an unsupervised method aiming to find the directions of maximum variance in a data set (X) without referring to the class labels (Y), and PLS-DA is a supervised method that uses multiple linear regression technique to find the direction of maximum covariance between a data set (X) and the class membership (Y). In both PCA and PLS-DA, the original variables are summarized into much fewer variables using their weighted averages called scores. Diversity analysis can be performed by estimating the alpha diversity, which provides a summary statistic of a single population, or beta diversity, which gives organismal composition between populations. Chao1 (Chao 1984), abundance-based coverage estimator (ACE) (Chao et al. 1993), and Jackknife (Heltshe and Forrester 1983) measure alpha diversity, species richness, and evenness (species distribution) expected within a single population. These results in collector's or rarefaction curves (Colwell and Coddington 1994). Alpha diversity is often quantified by the Shannon Index (Shannon 1948) or by Simpson Index (Simpson 1949). Beta diversity can be measured by simple taxa overlap or quantified by the Bray-Curtis dissimilarity (Bray and Curtis 1957) or UniFrac (Lozupone and Knight 2005). Two major approaches of clustering analysis include Hierarchical clustering and partitional clustering. Hierarchical, which is also called as agglomerative clustering, begins with each sample considered as separate cluster and then proceeds to combine them until all samples belong to one cluster. The result of hierarchical clustering is usually presented as a dendrogram or as a heat map, which displays the actual data values using color gradients. Clustering methods include average linkage, complete linkage, single linkage, and Ward's linkage. A dissimilarity measure includes Euclidean distance, Pearson's correlation, and Spearman's rank correlation. On the other hand, partitional clustering attempts to directly decompose the data set into a user-specified number of disjoint clusters. This uses methods like k-means clustering and self-organizing map (SOM). k-Means clustering create k clusters such that the sum of squares from points to

the assigned cluster centers' is minimized. SOM is an unsupervised neural network based around the concept of a grid of interconnected nodes, each of which contains a model.

Demands for new statistical methods to support emerging trends in metagenomics applications have resulted in more efficient implementations and better data visualization to lodge the tremendous increase in data analysis workloads. Web-based server with its user-friendly interface, comprehensive data processing options, wide array of statistical methods, and extensive data visualization and analysis support are playing key role. Servers like GEPAS (Herrero et al. 2003) and CARMAweb (Rainer et al. 2006), MG-RAST (Meyer et al. 2008), MEGAN (Huson et al. 2007), QIIME (Caporaso et al. 2010), Mothur (Schloss et al. 2009), and MetaboAnalyst (Xia et al. 2015) are few worth mentioning. Table 12.1 summarizes some the commonly used tools in microbiome analysis and their internet resources.

12.8 Analysis of Human Microbiome

Since birth, continuous exposure to microbial challenges has shaped the human microbiome and whose perturbation affects both human health and disease (Segal and Blaser 2014). In recent years, the knowledge about composition, distribution, and variation of bacteria in the human body has dramatically increased. Besides external factors like air, food, and environment, routine activity, habit, and physiology create selective pressure of each organism. In order to understand the influence of human microbiome, several studies have assessed the microbial compositions in different locations like stool, nasal, skin, vaginal, and oral of health and unhealthy individuals (Kraal et al. 2014). Thus, determining the extent of the variability of the human microbiome is therefore crucial for understanding the microbiology, genetics, and ecology of the microbiome. Besides that, it is useful for practical issues in designing experiments and interpretation of clinical studies (Zhou et al. 2014).

Study demonstrating the feasibility of using the composition of the gut microbiome to detect the presence of precancerous and cancerous lesions (Zackular et al. 2014), ethnic relation to significant differences in the vaginal microbiome (Fettweis et al. 2014), and discovery closely related oligotypes, differing sometimes by as little as a single nucleotide, showing dramatic different distributions among oral sites and among individuals (Eren et al. 2014), a less robustly interrogated placental microbiome by Aagaard et al. (2014), altered interactions between intestinal microbes, and the mucosal immune system resulting in inflammatory bowel disease (IBD) (Kostic et al. 2014) have taken us to the next level of understanding the human microbiome. Other studies like understanding of the etiology and pathogenesis of reflux disorders and esophageal adenocarcinoma (Yang et al. 2014) and altered microbiome on pulmonary responses (Segal and Blaser 2014) will be definitely be critical and open door for future investigations.

Table 12.1 Selected tools and their resources for microbiome analysis

Software	Brief description	URLs
<i>Preprocessing</i>		
FASTX-Toolkit	A collection of command line tools for short-read FASTA/FASTQ files preprocessing	hannonlab.cshl.edu/fastx_toolkit
FastQC	A quality-controlled tool for high-throughput sequence data	www.bioinformatics.babraham.ac.uk/projects/fastqc
SolexaQA	Calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data	http://solexaqa.sourceforge.net/
Lucy 2	Raw DNA sequence trimming and visualization tool based on the command-line tool Lucy1	http://www.complex.iastate.edu/download/Lucy2/index.html
CutAdapt	Removal of adapter sequences from high-throughput sequencing data	https://code.google.com/p/cutadapt/
NGS QC Toolkit	Perl-based stand-alone program package for the quality control (QC)	www.nipgr.res.in/ngsqctoolkit.html
Trimmomatic	Employed in trimming tasks for illumina paired-end and single ended data	http://www.usadellab.org/cms/?page=trimmomatic
ngsShoRT	Commonly used preprocessing algorithms in PERL	research.bioinformatics.udel.edu/genomics/ngsShoRT/
QC-Chain	A fast, accurate, and holistic NGS data quality-controlled method	http://www.computationalbioenergy.org/qc-chain.html
Meta-QC-Chain	A tool that combines multiple QC functions like identifying potential errors, quality trimming filters for poor sequencing quality bases and reads, and contamination screening that identifies higher eukaryotic species, which are considered as contamination for metagenomic data	http://computationalbioenergy.org/meta-qc-chain.html
PathoQC	A streamlined toolkit for preprocessing next-generation sequencing data	http://sourceforge.net/projects/PathoScope/
PRINSEQ	Provides summary statistics of FASTA (and QUAL) or FASTQ files	http://prinseq.sourceforge.net/
<i>Denosing</i>		
AmpliconNoise	A collection of programs for the removal of noise from 454 sequenced PCR amplicons	https://code.google.com/p/ampliconnoise/

(continued)

Table 12.1 (continued)

Software	Brief description	URLs
DADA	Algorithm for fast and accurate removal errors from PCR-amplified sequence data	http://sites.google.com/site/dadadenoiser
Acacia	Error corrector for pyrosequenced amplicon reads	http://sourceforge.net/projects/acaciaerrorcorr
<i>Chimera detection</i>		
UCHIME	Detects very low-divergent chimeras with a reference database	http://drive5.com/usearch/manual/uchime_algo.html
ChimeraSlayer	A chimeric sequence detection utility, compatible with near-full-length Sanger sequences and shorter 454-FLX sequences	http://microbiomeutil.sourceforge.net/
DECIPHER	Chimeric sequence detection utility developed using the R statistical programming language	http://decipher.cee.wisc.edu
<i>Reference-based assembly</i>		
Newbler (Roche)	Assembling sequence data generated by the 454 GS-series of pyrosequencing platforms sold by 454 Life Sciences, a Roche Diagnostics company	http://swes.cals.arizona.edu/maier_lab/kartchner/documentation/index.php/home/docs/newbler
MIRA 4	A multi-pass DNA sequence data assembler/mapper for whole-genome and EST/RNASeq projects	http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html
AMOS	A consortium committed to the development of open-source whole-genome assembly software	http://amos.sourceforge.net/wiki/index.php/AMOS
MetAMOS	An integrated assembly and analysis pipeline for metagenomic data	http://www.cbcb.umd.edu/software/metamos
Bowtie 2	Ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
<i>De novo assembly</i>		
EULER	A suite of programs for correcting errors in short reads (454 and Illumina) and assembling them	http://euler-assembler.ucsd.edu/
Velvet	de Bruijn graph-based single-genome assembler for short reads	https://www.ebi.ac.uk/~zerbino/velvet/
SOAPdenovo	The program is specially designed to assemble Illumina GA short reads for the human-sized genomes	http://soap.genomics.org.cn/soapdenovo.html
Abyss	A de novo, parallel, paired-end sequence assembler that is designed for short reads	http://www.bcgsc.ca/platform/bioinfo/software/abyss
MetaVelvet	Modified and extended de Bruijn graph-based single-genome	http://metavelvet.dna.bio.keio.ac.jp/

(continued)

Table 12.1 (continued)

Software	Brief description	URLs
	assembler, Velvet, for de novo metagenomic assembly	
MetaVelvet-SL	An extended Velvet assembler for detecting chimeric nodes by using supervised machine learning	metavelvet.dna.bio.keio.ac.jp/
Meta-IDBA	An iterative de Bruijn graph de novo short-read assembler specially designed for de novo metagenomic assembly	http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba/
Genovo	A tool for de novo metagenomic assembly and handle reads with length > 1000	http://cs.stanford.edu/group/genovo/
Trinity	Assembles transcript sequences from Illumina RNAseq data	https://github.com/trinityrnaseq/trinityrnaseq/wiki
<i>Binning tools</i>		
TETRA	To calculate how well tetranucleotide usage patterns in DNA sequences correlate	http://www.megx.net/tetra/index.html
PhylopythiaS	Taxonomic assignment of metagenome sequences among from three different models	http://phylopythias.cs.uni-duesseldorf.de/index.php?phase=wait
TACOA	Predicting the taxonomic origin of genomic fragments from metagenomic data sets by combining the advantages of the k-NN approach with a smoothing kernel function	http://www.cebitec.uni-bielefeld.de/index.php/2-uncategorised/99-tacoa?highlight=WyJ0YWVnYSJd
ESOM	A suite of programs to perform data mining tasks like clustering, visualization, and classification	http://databionic-esom.sourceforge.net/
ClaMS	A sequence composition-based classifier for metagenomic sequences	http://clams.jgi-psf.org/
MetaPhyler	Taxonomic classifier for metagenomic shotgun reads	http://metaphyler.cbcb.umd.edu/
Sort-ITEMS	A similarity-based binning method	http://metagenomics.atc.tcs.com/binning/Sort-ITEMS/
PhymmBL	Hybrid classifier tool which combines analysis from both Phymm and BLAST and produces even higher accuracy	http://www.cbcb.umd.edu/software/phymm/
MetaCluster	Binning and annotating short paired-end reads	http://i.cs.hku.hk/~alse/MetaCluster/
<i>OTU clustering</i>		
UCLUST	An algorithm that divides a set of sequences into clusters	http://www.drive5.com/usearch

(continued)

Table 12.1 (continued)

Software	Brief description	URLs
CD-HIT-OTU	Fast and accurate in identifying true OTUs and produces much fewer spurious OTUs	http://weizhong-lab.ucsd.edu/cd-hit-otu
TBC	Algorithm for defining operational taxonomic units (OTUs) without multiple sequence alignment	http://www.ezbiocloud.net/sw/tbc
<i>16S databases</i>		
RDP	A database that provides quality-controlled, aligned, and annotated bacterial and archaeal 16S rRNA sequences, fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community	http://rdp.cme.msu.edu/index.jsp
SILVA	A comprehensive online resource for quality-checked and aligned ribosomal RNA sequence data	http://www.arb-silva.de
GreenGenes	A collection of tools for choosing phylogenetically specific probes, interpreting microarray results, and aligning/annotating novel sequences	http://greengenes.lbl.gov
EzTaxon	A database that covers uncultured species often found in microbial ecological studies	http://www.ezbiocloud.net/eztaxon
<i>ITS database</i>		
UNITE	A platform for sequence-borne identification of ectomycorrhizal asco- and basidiomycetes	http://unite.ut.ee
<i>Sub-cellular localization</i>		
CoBaltDB	Predicting prokaryotic protein localizations	http://www.umr6026.univ-rennes1.fr/english/home/research/basic/software/cobalten
PSLpred	To predict the subcellular location for Gram-negative bacteria proteins	http://www.imtech.res.in/raghava/pslpred/
CELLO	Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions	http://cello.life.nctu.edu.tw/
PSORT-B	To predict the subcellular location for Gram-positive or Gram-negative bacterial proteins	
<i>Functional annotation databases</i>		
BLAST nr	Basic Local Alignment Search Tool against nonredundant database	http://blast.ncbi.nlm.nih.gov/Blast.cgi
SWISSPROT	Manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB)	http://www.uniprot.org/

(continued)

Table 12.1 (continued)

Software	Brief description	URLs
KEGG	Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/
SEED	A resource that provide consistent and accurate genome annotations across thousands of genomes	http://pubseed.theseed.org/
EggNOG	A database of orthologous groups and functional annotation	http://eggnogdb.embl.de/#/app/home
COG/KOG	EuKaryotic Orthologous Groups (KOG) is a eukaryote-specific version of the Clusters of Orthologous Groups (COG) tool for identifying ortholog and paralog proteins	http://genome.jgi.doe.gov/Tutorial/tutorial/kog.html
PFAM	Collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)	http://pfam.xfam.org/
TIGRFAMs	A resource consisting of curated multiple sequence alignments, hidden Markov models (HMMs) for protein sequence classification, and associated information designed to support automated annotation of (mostly prokaryotic) proteins	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi
MetaBioMe	A web resource to find novel homologues for known commercially useful enzymes (CUEs) in metagenomic data sets and completed bacterial genomes	http://metasystems.riken.jp/metabiome/
TSdb	The transporter substrate database (TSdb)—a central repository of formatted substrate information of transporters as well as their annotation	http://tsdb.cbi.pku.edu.cn/
TCDB	Functional and phylogenetic classification of membrane transport proteins	http://www.tcdb.org/
CAZy	A specialist database dedicated to the display and analysis of genomic, structural, and biochemical information on carbohydrate-active enzymes (CAZymes)	http://www.cazy.org/
dbCAN	A database for carbohydrate-active enzymes	http://csbl.bmb.uga.edu/dbCAN/
<i>Annotation of metagenomics sequences</i>		
MetaGeneMark	For gene prediction in metagenomes	http://exon.gatech.edu/meta_gmhmp.cgi
MetaGeneAnnotator	A gene-finding program for prokaryote and phage	http://metagene.nig.ac.jp/

(continued)

Table 12.1 (continued)

Software	Brief description	URLs
Prodigal	A gene-finding program developed at Oak Ridge National Laboratory and the University of Tennessee	http://prodigal.ornl.gov/
Orphelia	A metagenomic ORF finding tool for the environmental DNA sequences with unknown phylogenetic origin	http://orphelia.gobics.de/
FragGeneScan	Software for predicting prokaryotic genes in incomplete assemblies or complete genomes	http://sourceforge.net/projects/fraggenescan/
PILER-CR	Software for finding CRISPR repeats	http://www.drive5.com/pilercr/
tRNAscan-SE	A web server for predicting tRNAs	http://lowelab.ucsc.edu/tRNAscan-SE/
WebMGA	A web server for rapid metagenomic data analysis using fast and effective algorithms	http://weizhong-lab.ucsd.edu/metagenomic-analysis/
METAREP	An open-source tool to view, query, browse, and compare metagenomics annotation profiles from short reads or assemblies	http://jcvl.org/metarep/
STAMP	A software package for analyzing taxonomic or metabolic profiles	http://kiwi.cs.dal.ca/Software/STAMP
CoMet	A web server for fast comparative functional profiling of metagenomes	http://comet.gobics.de/
RAMMCAP	Analysis and comparison of very large metagenomes with fast clustering and functional annotation	http://weizhong-lab.ucsd.edu/rammcap/cgi-bin/rammcap.cgi
<i>Analytical pipelines for 16S</i>		
CARMA	Software pipeline for characterizing the taxonomic composition and genetic diversity of short-read metagenomes	http://www.cebitec.uni-bielefeld.de/index.php/2-uncategorised/47-carma?highlight=WyJjYXJtYSJd
IMG/M	Integrated Microbial Genomes with Microbiome	http://img.jgi.doe.gov/m/doc/background.html
MG-RAST	An automated analysis platform for metagenomes	http://metagenomics.anl.gov/
Mothur	An open-source software for microbial ecology community analysis	http://www.mothur.org
QIIME	An open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data	http://qiime.org
SILVAngs	A data analysis service for ribosomal RNA gene (rDNA) amplicon	https://www.arb-silva.de/ngs/

(continued)

Table 12.1 (continued)

Software	Brief description	URLs
	reads from high-throughput sequencing	
MEGAN	Tool for studying the taxonomic content from short-read metagenomes	http://ab.inf.uni-tuebingen.de/software/megan5/
WATERS	From 16S rDNA contigs to biological interpretation and analysis	http://code.google.com/p/waters16s
RDPipeline	For processing large rRNA sequence libraries (single-strand and paired-end reads) obtained through high-throughput sequencing technology	https://pyro.cme.msu.edu
VAMPS	A collection of tools for visualization and analyze data for microbial population structures and distributions	http://vamps.mbl.edu
Genboree	A web-based platform for multi-omic research and data analysis using the latest bioinformatics tools	http://genboree.org
SnoWMan	Pipeline for analysis of microbiome data	https://snowman.genome.tugraz.at/snowman

12.9 Conclusion

Human microbiota includes microorganisms living on the surface and inside the body. They are important for the host's health. These are highly dynamic and can be influenced by a number of factors such as age, diet, and physiology. Studies have shown that most of the human adult microbiota lives in the gut and follows specific microbial signatures but with high intraindividual variability over time. Any alterations of the human gut microbiome can play a role in disease development. Thus, exploring microbiome could make themselves as potent target for diagnostic and therapeutic applications. Since early microbial studies were based on the direct cultivation and isolation of microbes, clinical applications posed several limitations especially growth conditions. Studies have shown that not all microbes are currently uncultivable. Methods to study cultivable organisms are also not suitable for the study of entire microbiome. Metagenomics helped in the direct genetic analysis of genomes contained within an environmental sample without the need for cultivating. Metagenomic studies using NGS-based methods can be approached by amplifying 16S rRNA genes using specific primers or through whole-genome shotgun sequencing. 16S sequences identified can be used to describe their community relative abundance and/or their phylogenetic relationships by clustering into operational taxonomic units (OTUs) using databases of previously annotated sequences. In whole-genome shotgun sequencing approach, where random primers

are used for amplifying all microbial genes, the relative abundances of genes and pathways can be determined by comparing the sequences to functional databases.

Next-generation sequencing (NGS) technologies not only increased the throughput of bases sequenced/run but also reduced sequencing costs. This had a major impact on the field of metagenomics where a specific microbiome can be qualitatively and quantitatively characterized in depth without the selection bias and constraints associated with cultivation methods. Continuous advancements in sequencing technologies have not only allowed address more complex habitats but also have imposed growing demands on bioinformatic data post-processing. Analyzing the huge amount of data by these technologies has become the bottleneck especially in case of larger metagenome projects. From assembly to analysis, bioinformatic post-processing requires dedicated data integration pipelines, some of which have yet to be developed.

Acknowledgments The authors wish to acknowledge the Department of Biotechnology (DBT), Govt. of India, New Delhi for the financial support in the form of State Biotech Hub (BT/04/NE/2009) and Bioinformatics Infrastructure Facility (BT/B1/12/060/2012 (NERBIF-MUA)). KSI acknowledge the financial assistance provided by DST-SERB, New Delhi through young scientist scheme (YSS/2014/000657).

References

- Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. *Sci Transl Med.* 2014;6(237):237–65.
- Abubucker S, Segata N, Goll J, Schubert A, Izard J, Cantarel B, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley S, Methe B, Schloss P, Gevers D, Mitreva M, Huttenhower C. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol.* 2012;8:e1002358.
- Afiahayati SK, Sakakibara Y. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Res.* 2014;22(1):69–77.
- Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25–9.
- Ashelford K, Chuzhanova N, Fry J, Jones A, Weightman A. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol.* 2005;71:7724–36.
- Bäckhed F, Ley R, Sonnenburg J, Peterson D, Gordon J. Host-bacterial mutualism in the human intestine. *Science.* 2005;307(5717):1915–20.
- Bakker M, Tu Z, Bradeen J, Kinkel L. Implications of pyrosequencing error correction for biological data interpretation. *PLoS One.* 2012;7(8):e44357.
- Balzer S, Malde K, Grohme M, Jonassen I. Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics.* 2013;29(7):830–6.
- Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S. MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics.* 2009;25(12):1554–5.

- Bikel S, Valdez-Lara A, Cornejo-Granados F, Rico K, Canizales-Quinteros S, Soberón X, Del Pozo-Yauner L, Ochoa-Leyva A. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput Struct Biotechnol J*. 2015;13:390–401.
- Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ. De novo transcriptome assembly with ABySS. *Bioinformatics*. 2009;25(21):2872–7.
- Bragg L, Stone G, Imelfort M, Hugenholtz P, Tyson G. Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat Methods*. 2012;9(5):425–6.
- Bray J, Curtis J. An ordination of upland forest communities of southern Wisconsin. *Ecol Monogr*. 1957;27:325–49.
- Buttigieg P, Hankeln W, Kostadinov I, Kottmann R, Yilmaz P, Duhaime M, Glöckner F. Ecogenomic perspectives on domains of unknown function: correlation-based exploration of marine metagenomes. *PLoS One*. 2013;8(3):e50869.
- Caporaso J, Kuczynski J, Stombaugh J, Bittinger K, Bushman F, Costello E, Fierer N, Peña A, Goodrich J, Gordon J, Huttley G, Kelley S, Knights D, Koenig J, Ley R, Lozupone C, McDonald D, Muegge B, Pirrung M, Reeder J, Sevinsky J, Turnbaugh P, Walters W, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–6.
- Caporaso J, Lauber C, Costello E, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon J, Knight R. Moving pictures of the human microbiome. *Genome Biol*. 2011;12(5):R50.
- Caspi R, Altman T, Dreher K, Fulcher C, Subhraveti P, Keseler I, Kothari A, Krummenacker M, Latendresse M, Mueller L, Ong Q, Paley S, Pujar A, Shearer A, Travers M, Weerasinghe D, Zhang P, Karp P. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2012;40:D742–53.
- Chao A. Nonparametric estimation of the number of classes in a population. *Scand J Stat*. 1984;11:265–70.
- Chao A, Ma M-C, Yang M. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*. 1993;80:93–201.
- Chevreur B, Pfisterer T, Drescher B, Driesel A, Müller W, Wetter T, Suhai S. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res*. 2004;14(6):1147–59.
- Chistoserdova L. Recent progress and new challenges in metagenomics for biotechnology. *Biotechnol Lett*. 2010;32:1351–9.
- Cho I, Blaser M. The Human Microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012;13(4):260–70.
- Cline J, Braman J, Hogrefe H. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res*. 1996;24:3546–51.
- Cole J, Chai B, Farris R, Wang Q, Kulam-Syed-Mohideen A, McGarrell D, Bandela A, Cardenas E, Garrity G, Tiedje J. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*. 2007;35(Database issue):D169–72.
- Cole J, Wang Q, Cardenas E, Fish J, Chai B, Farris R, Kulam-Syed-Mohideen A, McGarrell D, Marsh T, Garrity G, Tiedje J. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res*. 2009;37:D141–5.
- Colwell R, Coddington J. Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B*. 1994;345:101–18.
- DeSantis T, Hugenholtz P, Larsen N, Rojas M, Brodie E, Keller K, Huber T, Dalevi D, Hu P, Andersen G. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*. 2006;72(7):5069–72.
- Edgar R. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460–1.

- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*. 2011;27:2194–200.
- Eren AM, Borisy GG, Huse SM, Mark Welch JL. Oligotyping analysis of the human oral microbiome. *Proc Natl Acad Sci U S A*. 2014;111(28):E2875–84.
- Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8(3):175–85.
- Fettweis JM, Brooks JP, Serrano MG, Sheth NU, Girerd PH, Edwards DJ, Strauss 3rd JF, Jefferson KK, Buck GA. Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology*. 2014;160(Pt 10):2272–82.
- Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. *Nucleic Acids Res*. 2008;36 (Database issue):D281–8.
- Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. *Nucleic Acids Res*. 2013;42:D222–30.
- Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A*. 2014;111 (22):E2329–38.
- Gaspar JM, Thomas WK. Assessing the consequences of denoising marker-based metagenomic data. *PLoS One*. 2013;8(3):e60458.
- Ghaffari N, Sanchez-Flores A, Doan R, Garcia-Orozco KD, Chen PL, Ochoa-Leyva A, Lopez-Zavala AA, Carrasco JS, Hong C, Briebe LG, Rudiño-Piñera E, Blood PD, Sawyer JE, Johnson CD, Dindot SV, Sotelo-Mundo RR, Criscitiello MF. Novel transcriptome assembly and improved annotation of the whiteleg shrimp (*Litopenaeus vannamei*), a dominant crustacean in global seafood mariculture. *Sci Rep*. 2014;4:7081.
- Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methé BA, Yooseph S. METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. *Bioinformatics*. 2010;26(20):2631–2.
- Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, Pérez-Cobas AE, Latorre A, Moya A. Metatranscriptomic Approach to Analyze the Functional Human Gut Microbiota. *PLoS One*. 2011;6(3):e17447.
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. 2005;33(Database issue):D121–4.
- Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methe B, DeSantis TZ, Petrosino JF, Knight R, Birren BW. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*. 2011;21(3):494–504.
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8(8):1494–512.
- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*. 1998;5 (10):R245–9.
- Hansen M, Tolker-Nielsen T, Givskov M, Molin S. Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. *FEMS Microbiol Ecol*. 1998;26:141–9.
- Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci U S A*. 2007;104(35):13913–8.
- Haynes M, Rohwer F. *Metagenomics of the Human Body* Springer. New: York; 2011.

- Heltshel J, Forrester N. Estimating species richness using the jackknife procedure. *Biometrics*. 1983;39:1–11.
- Herrero J, Al-Shahrour F, Diaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J, Dopazo J. GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*. 2003;31(13):3461–7.
- Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res*. 2009;37(Web Server issue):W101–5.
- Huang W, Marth G. EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res*. 2008;18(9):1538–43.
- Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res*. 2007;17(3):377–86.
- Iyer S, Bouzek H, Deng W, Larsen B, Casey E, Mullins JI. Quality score based identification and correction of pyrosequencing errors. *PLoS One*. 2013;8(9):e73015.
- Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the human oral microbiome during health and disease. *MBio*. 2014;5(2):e01012–4.
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27–30.
- Keegan KP, Trimble WL, Wilkening J, Wilke A, Harrison T, D'Souza M, Meyer F. A platform-independent method for detecting errors in metagenomic sequencing data: DRISSEE. *PLoS Comput Biol*. 2012;8(6):e1002541.
- Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res*. 2011;40(1):e9.
- Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Res*. 2011;21(3):487–93.
- Knudsen BS, Kim HL, Erho N, Shin H, Alshalhafa M, Lam LL, Tenggara I, Chadwick K, Van Der Kwast T, Fleshner N, Davicioni E, Carroll PR, Cooperberg MR, Chan JM, Simko JP. Application of a clinical whole-transcriptome assay for staging and prognosis of prostate cancer diagnosed in needle core biopsy specimens. *J Mol Diagn*. 2016; pii: S1525–1578(16) 00051–9. doi:[10.1016/j.jmoldx.2015.12.006](https://doi.org/10.1016/j.jmoldx.2015.12.006).
- Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A*. 2010;108(Suppl 1):4578–85.
- Koljal U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AF, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Duenas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lucking R, Martin MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, Peterson M, Poldmaa K, Saag L, Saar I, Schussler A, Scott JA, Senes C, Smith ME, Suija A, Taylor DL, Telleria MT, Weiss M, Larsson KH. Towards a unified paradigm for sequence-based identification of fungi. *Mol Ecol*. 2013;22(21):5271–7.
- Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*. 2014;146(6):1489–99.
- Kraal L, Abubucker S, Kota K, Fischbach MA, Mitreva M. The prevalence of species and strains in the human microbiome: a resource for experimental efforts. *PLoS One*. 2014;9(5):e97279.
- Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev*. 2008;72(4):557–78. , Table of Contents
- Kyrpides NC, Ouzounis CA. Whole-genome sequence annotation: 'going wrong with confidence'. *Mol Microbiol*. 1999;32(4):886–7.
- Laehnemann D, Borkhardt A, McHardy AC (2015) Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform*
- Lai B, Ding R, Li Y, Duan L, Zhu H. A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics*. 2012;28(11):1455–62.

- Lampe JW. The Human Microbiome Project: getting to the guts of the matter in cancer epidemiology. *Cancer Epidemiol Biomark Prev.* 2008;17(10):2523–4.
- Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A.* 1985;82(20):6955–9.
- Laserson J, Jojic V, Koller D. Genovo: de novo assembly for metagenomes. *J Comput Biol.* 2011;18(3):429–43.
- Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics.* 2008;24(5):713–4.
- Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics.* 2009;25(15):1966–7.
- Liu Y, Guo J, Hu G, Zhu H. Gene prediction in metagenomic fragments based on the SVM algorithm. *BMC Bioinf.* 2013;14(Suppl 5):S12.
- Looft T, Johnson TA, Allen HK, Bayles DO, Alt DP, Stedtfeld RD, Sul WJ, Stedtfeld TM, Chai B, Cole JR, Hashsham SA, Tiedje JM, Stanton TB. In-feed antibiotic effects on the swine intestinal microbiome. *Proc Natl Acad Sci U S A.* 2012;109(5):1691–6.
- Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997;25(5):955–64.
- Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol.* 2005;71(12):8228–35.
- Luria N, Sela N, Yaari M, Feygenberg O, Kobiler I, Lers A, Prusky D. De-novo assembly of mango fruit peel transcriptome reveals mechanisms of mango response to hot water treatment. *BMC Genomics.* 2014;15:957.
- Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, McKernan K, Ranade S, Shea TP, Williams L, Young S, Nusbaum C, Jaffe DB. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol.* 2009;10(10):R103.
- Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* 2001;29(22):4724–35.
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science.* 1999a;285(5428):751–3.
- Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature.* 1999b;402(6757):83–6.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005;437(7057):376–80.
- Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P, Kyrpides NC. An experimental metagenome data management and analysis system. *Bioinformatics.* 2006;22(14):e359–67.
- Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 2008;36:D534–8.
- Mavromatis K, Ivanova N, Barry K, Shapiro H, Goldsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, Kyrpides NC. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods.* 2007;4(6):495–500.

- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinf.* 2008;9:386.
- Miller DN, Bryant JE, Madsen EL, Ghiorse WC. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl Environ Microbiol.* 1999;65(11):4715–24.
- Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics.* 2010;95(6):315–27.
- Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. Tablet – next generation sequence assembly visualization. *Bioinformatics.* 2009;26(3):401–2.
- Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 2012;13(9):R79.
- Nacke H, Engelhaupt M, Brady S, Fischer C, Tautz J, Daniel R. Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. *Biotechnol Lett.* 2011;34(4):663–75.
- Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 2012;40(20):e155.
- Nawrocki EP, Eddy SR. Computational identification of functional RNA homologs in metagenomic data. *RNA Biol.* 2013;10(7):1170–9.
- Ness RW, Siol M, Barrett SC. De novo sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. *BMC Genomics.* 2011;12:298. [936]
- Nilakanta H, Drews KL, Firrell S, Foulkes MA, Jablonski KA. A review of software for analyzing molecular sequences. *BMC Res Note.* 2014;7:830.
- Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 2006;34(19):5623–30.
- Oliver KM, Degan PH, Hunter MS, Moran NA. Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science.* 2009;325(5943):992–4.
- Oulas A, Pavlodi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos I. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinf Biol Insight.* 2015;9:75–88.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A.* 1999;96(6):2896–901.
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goessmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 2005;33(17):5691–702.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A.* 1999;96(8):4285–8.
- Peng Y, Leung HC, Yiu SM, Chin FY. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics.* 2011;27(13):i94–101.
- Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012;28(11):1420–8.
- Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C,

- Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. The NIH Human Microbiome Project. *Genome Res.* 2009;19(12):2317–23.
- Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A.* 2001;98(17):9748–53.
- Pevzner PA, Tang H, Tesler G. De novo repeat classification and fragment assembly. *Genome Res.* 2004;14(9):1786–96.
- Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinform.* 2009;10(4):354–66.
- Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* 2011;40(Database issue):D284–9.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 2007;35(21):7188–96.
- Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 2011;40(Database issue):D130–5.
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. *Nucleic Acids Res.* 2011;40(Database issue):D290–301.
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Jian M, Zhou Y, Li Y, Zhang X, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464(7285):59–65.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2012;41(Database issue):D590–6.
- Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC Bioinf.* 2011;12:38.
- Rainer J, Sanchez-Cabo F, Stocker G, Sturn A, Trajanoski Z. CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res.* 2006;34(Web Server issue):W498–503.
- Reeder J, Knight R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods.* 2010;7(9):668–9.
- Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010;38(20):e191.
- Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 2000;16(6):276–7.
- Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I. De novo assembly and analysis of RNA-seq data. *Nat Methods.* 2010;7(11):909–12.
- Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol.* 2005;71(3):1501–6.
- Schloss PD, Handelsman J. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol.* 2006a;72(10):6773–9.

- Schloss PD, Handelsman J. Introducing TreeClimber, a test to compare microbial community structures. *Appl Environ Microbiol.* 2006b;72(4):2379–84.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* 2009;75(23):7537–41.
- Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One.* 2011;6(12):e27310.
- Scholz MB, Lo CC, Chain PS. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol.* 2011;23(1):9–15.
- Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A.* 1998;95(11):5857–64.
- Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics.* 2012;28(8):1086–92.
- Segal LN, Blaser MJ. A brave new world: the lung microbiota in an era of change. *Ann Am Thorac Soc.* 2014;11(Suppl 1):S21–7.
- Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. *Genome Biol.* 2011;12(6):R60.
- Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.* 2007;35(Database issue):D260–4.
- Shannon C. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27:379–423. , 623–656
- Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T, Wan XC. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics.* 2011;12:131.
- Simon C, Daniel R. Metagenomic analyses: past and future trends. *Appl Environ Microbiol.* 2010;77(4):1153–61.
- Simpson E. Measurement of diversity. *Nature.* 1949;163:688.
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 2009;19(6):1117–23.
- Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods – a bioconductor package providing PCA methods for incomplete data. *Bioinformatics.* 2007;23(9):1164–7.
- Steinfath M, Groth D, Lisek J, Selbig J. Metabolite profile analysis: from raw data to regression and classification. *Physiol Plant.* 2008;132(2):150–61.
- Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science.* 1997;278(5338):631–7.
- Thomas T, Gilbert J, Meyer F. Metagenomics – a guide from sampling to data analysis. *Microb Info Exp.* 2012;2(1):3.
- Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaia I, Ondov B, Darling AE, Phillippy AM, Pop M. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 2013;14(1):R2.
- Trimble WL, Keegan KP, D’Souza M, Wilke A, Wilkening J, Gilbert J, Meyer F. Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinf.* 2012;13:183.
- Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol.* 2008;11(5):442–6.
- van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics.* 2006;7:142.
- Virgin HW, Wherry EJ, Ahmed R. Redefining chronic viral infection. *Cell.* 2009;138(1):30–50.

- Wooley JC, Ye Y. Metagenomics: facts and artifacts, and computational challenges. *J Comput Sci Technol.* 2009;25(1):71–81.
- Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.* 2009;37(Web Server issue):W652–60.
- Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0 – making metabolomics more meaningful. *Nucleic Acids Res.* 2015;43(W1):W251–7.
- Yang L, Chaudhary N, Baghdadi J, Pei Z. Microbiome in reflux disorders and esophageal adenocarcinoma. *Cancer J.* 2014;20(3):207–10.
- Yok NG, Rosen GL. Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinf.* 2011;12:20.
- Zackular JP, Rogers MA, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev Res (Phila).* 2014;7(11):1112–21.
- Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821–9.
- Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics.* 2011;28(1):125–6.
- Zhou Y, Mihindukulasuriya KA, Gao H, La Rosa PS, Wylie KM, Martin JC, Kota K, Shannon WD, Mitreva M, Sodergren E, Weinstock GM. Exploration of bacterial community classes in major human habitats. *Genome Biol.* 2014;15(5):R66.
- Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 2010;38(12):e132.