5 Advancements in Microbial Genome Sequencing and Microbial Community Characterization

Bhaskar Reddy

Abstract

The microorganism play an essential role in various metabolic activity associated with health, obesity, immune system, complex carbohydrate, nitrogen, sulfur, and xenobiotic metabolism etc. The identification of microorganism involved in such process is becoming possible with the sequencing of 16S rRNA amplicon and responsible gene through molecular cloning and then sequencing. The first-generation sequencing extensively facilitated the molecular characterization of microorganism and functional gene with expense of high cost with low throughput. The advent of next-generation sequencing technology enables the high-scale full-length 16S rRNA molecular characterization and genome sequencing with reduced time and cost with high yield. The present article describes available genomes in public database and the role of next- and third-generation sequencing technology contribution to the growth of genome and metagenome sequencing and its associated projects, their taxonomy, and functional characterization through bioinformatic analysis. This chapter also provides an overview on the metagenomic sequencing and functional characterization of three important ecological niches, viz., rumen, soil, and human gut. The massive advancement in high-throughput sequencing technology and bioinformatic analysis enabled robust genome and metagenome characterization in short time with reduced budget.

B. Reddy (✉)
Centre of Advanced Study in Botany, Institute of Science, Banaras Hindu University,
Varanasi, India
e-mail: 24breddy@gmail.com

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_5

5.1 Introduction

DNA sequencing is the key step in genomic studies and molecular characterizations. Sequencing techniques are widely applied, but not limited to fields such as molecular biology, biotechnology, genetics, genome sequencing, forensic sciences, archaeology, anthropology, and metagenomics. Two decades ago, the sequenced genome of the first bacterial genome *Haemophilus influenzae* Rd. was reported (Fleischmann et al. 1995). The extensive technological advancements in sequencing chemistry, significant growth of genomes, expressed sequence tags (ESTs), and metagenomes were observed (Sayers et al. 2018), because of tremendous throughput and drastic reduction in sequencing cost. The genome of *Eschericha Coli* were repprted to harbor nearly 5000 proteins oer genome. (Cook and Ussery 2013).

In order to analyze the sequenced genomes, bioinformatic-driven analysis facilitated the harvesting of functional signatures, comparison, and visualization. For such task fulfillment, various tools have been developed among that majority for second-generation sequencer. As traditional assembler and annotation pipelines are unable to handle such enormous data, the new method is continuously developing (Pop 2009; Ekblom and Wolf 2014). Also development of efficient computational algorithms coupled with high-performance computers (HPC) facilitated the robust genome, metatranscriptome, and metagenome analysis and raw read archival system with significantly reduced time (Leinonen et al. 2011; Keegan et al. 2016; Mitchell et al. 2018; Mukherjee et al. 2018).

5.1.1 Sequencing Projects

The extensive data generation and efficient computational resource development facilitated the finishing of various complete genomes and draft genomes. As shown in Fig. 5.1a, there was a remarkable growth of complete genomes from year 2010 to 2018, which increased from 506 to 2058 and permanent drafts from 718 to 15,098. The majority of bacterial genomes were obtained from medical sector (59%), followed by environment (7%) and agriculture (7%) projects (Fig. 5.1b). It is obvious that pathogens are greatly spreading with gain of resistance against antibiotics; medical sector-associated pathogen genome analysis could provide more insights of drug resistance and management (Dethlefsen et al. 2008). Table 5.1 shows domain-specific genome projects in which more than one lakh bacterial whole genome sequencing (WGS) projects and more than 60 K metagenome projects and nearly 1.5 K archaeal WGS were contributed/deposited in Genomes OnLine Database (GOLD (Mukherjee et al. 2018)). Further looking to archaea phyla level, majority of projects were associated with *Euryarchaeota* (58.46%) and *Crenarchaeota* (23.64%) (Table 5.2a), whereas among bacteria, majority of projects were associated with *Proteobacteria* (51.19%), *Firmicutes* (29.66), *Actinobacteria* (12.11), *Bacteroidetes* (2.67), and *Cyanobacteria* (0.97) (Table 5.2b).

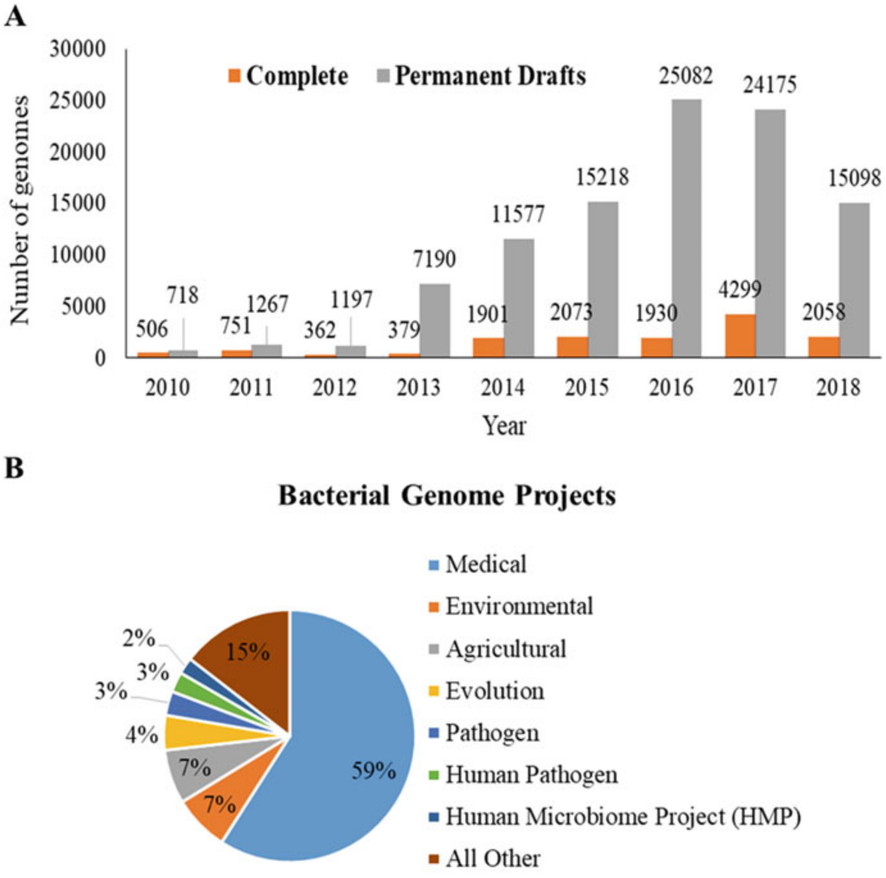

Fig. 5.1 The number of complete and permanent draft genomes (**a**) and projects' relevance to bacterial genome (**b**) in GOLD. Presented data accessed on December 26, 2018, from https://gold.jgi.doe.gov/

Table 5.1 Phylogenetic distribution of genome projects in GOLD

Domain	Total	% Domain
Archaea	16,120	7.62
Bacteria	135,101	63.84
Eukaryotic	51,481	24.33
Virus	8933	4.22

Presented data accessed on December 26, 2018, from GOLD (https://gold.jgi.doe.gov/)

It is also important to emphasize on the contribution of different ecological types in biosample and sequencing projects. It is observed that majority of projects were host-associated, followed by environment. Among the host-associated, majority were human, mammals, plants, arthropods, birds, and fungi. Among the

Table 5.2a Phylogenetic distribution of archaea at phyla level associated projects in GOLD

Phyla	Total phyla	% Phyla
Euryarchaeota	947	58.46
Crenarchaeota	383	23.64
Thaumarchaeota	215	13.27
Unclassified	29	1.79
Candidatus Parvarchaeota	13	0.80
Nanoarchaeota	12	0.74
Candidatus Woesearchaeota	10	0.62
Candidatus Aenigmarchaeota	4	0.25
Candidatus Diapherotrites	3	0.19
Candidatus Bathyarchaeota	2	0.12
Candidatus Korarchaeota	1	0.06
Candidatus Micrarchaeota	1	0.06

Presented data accessed on December 26, 2018, from GOLD (https://gold.jgi.doe.gov/)

Table 5.2b Phylogenetic distribution of bacteria at phyla level associated projects in GOLD

Phyla	Total phyla	% Phyla
Proteobacteria	69,154	51.19
Firmicutes	40,075	29.66
Actinobacteria	16,362	12.11
Bacteroidetes	3608	2.67
Cyanobacteria	1313	0.97
Spirochaetes	873	0.65
Tenericutes	558	0.41
Unclassified	454	0.34
Chlamydiae	446	0.33
Fusobacteria	260	0.19
Chloroflexi	244	0.18
Verrucomicrobia	187	0.14
Thermotogae	175	0.13
Deinococcus-Thermus	134	0.10
Planctomycetes	130	0.10
Fibrobacteres	112	0.08
Candidatus Parcubacteria	88	0.07
Acidobacteria	80	0.06
Candidatus Microgenomates	64	0.05
Deferribacteres	62	0.05
Nitrospirae	46	0.03
Chlorobi	33	0.02
Nitrospinae	31	0.02
Aquificae	36	0.03
Others	576	0.43

Presented data accessed on December 26, 2018, from GOLD (https://gold.jgi.doe.gov/)

Table 5.3 The number of sequencing projects associated biosample from different ecosystem hosts submitted to GOLD

Host-associated (28015)	Total	Environmental (26803)		Engineered (5127)	
Algae	86	Air	104	Bioreactor	219
Animal	79	Aquatic	19,074	Bioremediation	93
Annelida	99	Terrestrial	7623	Biotransformation	31
Arthropoda	915	Unclassified	2	Built environment	1869
Birds	783			Food production	443
Cnidaria	157			Industrial production	81
Echinodermata	39			Lab enrichment	331
Endosymbionts	2			Lab synthesis	12
Fish	30			Modeled	354
Fungi	691			Paper	18
Human	17,336			Solid waste	185
Insecta	43			Unclassified	18
Invertebrates	94			Wastewater	1473
Mammals	3987				
Microbial	102				
Mollusca	65				
Plants	3414				
Porifera	35				
Protists	4				
Reptilia	36				
Tunicates	10				
Unclassified	8				

Presented data accessed on December 26, 2018, from GOLD (https://gold.jgi.doe.gov/)

environmental ecosystem, aquatic and terrestrial were in majority, and among the engineered ecosystem built environment, wastewater, food production modeled, and lab enrichment were in majority (Table 5.3). Looking in details, 111 different ecosystem types contributed to enormous biosamples. Among these, the digestive system, marine, freshwater, soil, and thermal springs were in majority, while tooth, solar panel, microbial solubilization of coal, and hair were the least (Table 5.4).

The Genomes Online Database (GOLD) contains 340,849 total organisms; among those 300,052 were bacteria and 3093 were archaea. The MG-RAST v4.03 system listed 362,238 metagenomes with 1329 billion sequences constituted 183.08 Tbp (Tera base pair). This shows the high demand of next-generation sequencing (NGS) in various ecosystem biosamples for their whole genome sequencing (WGS) and metagenomics. Microbial genomes available in Ensembl genome browser consist of 61 phyla, 1600 genera, and 9800 species. Interestingly, among the available sequenced genomes, *Proteobacteria* accounted the major fraction (Mukherjee et al. 2018). Additionally, the advancements in sequencing of uncultivable microbial genomes and reconstruction of genomes from metagenomes through second and third generation contribute in the enlargement of database repositories.

Table 5.4 The number of sequencing projects associated biosample from different ecosystem types submitted to GOLD

Ecosystem type	Total	Ecosystem type	Total	Ecosystem type	Total
Digestive system	19,373	Bacteria	89	Integument	10
Marine	8740	House	78	Landfill	10
Fresh water	6186	Engineered product	78	Sponge	10
Soil	5933	Indoor air	69	Lymphatic system	9
Thermal springs	1498	Hospital	55	Ascidians	9
City	1464	Fermented beverages	53	Nodule	8
Skin	1381	Symbiotic fungal gardens and galleries	45	Oil reservoir	8
Non-marine saline and alkaline	1343	Aquaculture	45	Milk	8
Roots	1085	Green algae	43	Sclerotium	8
Phyllosphere	880	Bone	43	Oil refinery	7
Plant litter	640	Hydrocarbon	40	Eye	7
Activated sludge	637	Simulated communities (DNA mixture)	39	Agricultural field	7
Mycelium	619	Asteroidea	36	Fermented vegetables	6
Gastrointestinal tract	511	Outdoor air	34	Beetle	6
Circulatory system	496	Lichen	30	Cave	6
Respiratory system	454	Wood	25	Volcanic	5
Rhizosphere	433	Rock-dwelling (subaerial biofilms)	24	Aerobic	5
Peat	424	Rock-dwelling (endoliths)	23	Spacecraft assembly cleanrooms	5
Rhizoplane	377	Ant dump	21	Dinoflagellates	4
Phylloplane	363	Red algae	21	Shell	4
Sediment	346	Continuous culture	20	Ctenophora	4
Anaerobic digestor	300	Mixed alcohol bioreactor	20	Tooth	4
Deep subsurface	291	Nervous system	20	Solar panel	3
Simulated communities (microbial mixture)	290	Thiocyanate	18	Biochar	3
Dairy products	277	Currency notes	18	Metal	3
Industrial wastewater	235	Larvae	15	Microbial solubilization of coal	3
Water treatment plant	231	Terephthalate	14	Brown algae	2
Nutrient removal	224	Canal	14	Tailings pond	2
Geologic	192	Fruiting body	13	Whole body	2

(continued)

Table 5.4 (continued)

Ecosystem type	Total	Ecosystem type	Total	Ecosystem type	Total
Defined media	174	Tetrachloroethylene and derivatives	13	Fungi	2
Composting	160	Intracellular endosymbionts	12	Breviatea	1
Tissue	152	Simulated communities (sequence read mixture)	12	Microbial enhanced oil recovery	1
Leaf	133	Seeds	12	Asphalt lakes	1
Anaerobic	126	Peat moss	12	Swine wastewater	1
Reproductive system	104	Mosquito	12	Hair	1
Meat products	103	Endosphere	11	Nematoda	1
Cnidaria	89	Solid animal waste	11	Persistent organic pollutants (POP)	1

Presented data accessed on December 26, 2018, from GOLD (https://gold.jgi.doe.gov/)

5.2 Genome Characteristics

The sequenced genomes deposited in public databases, such as NCBI, GOLD, ENA, DDBJ, and Ensembl, offer to study the functional features and contribution to the ecosystem (Leinonen et al. 2011). Also, there is a significant variation in gene content and genome size in species to species. Moreover, a species and strain display very streamlined and homogenous in terms of genetic variations observed in transposable elements and resistance genes (e.g., *Mycobacterium tuberculosis*) (Land et al. 2014). Comparisons made within genes and between genes of different organisms provide a distinct type of closeness, leading to the development of genes common to most genomes (core genes) and total genes (pan genes) set. This provides a reasonable knowledge of species closeness and molecular evolution. The wide range of *E. coli* genome analysis revealed that pan-genomes are increasing than core gene sets, and letter various pan and core genomes have been determined (Land et al. 2014).

Looking to inside of sequenced genomes showed that 2671 complete/finished genomes consist of 88% of average protein coding region in bacteria, available in GenBank, and it ranges between 40% and 97% (Land et al. 2014). Meanwhile bacteria generally consist of 5 Mb genome size which encodes near about 5000 proteins. Among the sequenced genomes available in GenBank, the largest genome is *Sorangium cellulosum* strain So0157–2 with a size of 14,782,125 bp and contains 11,021 genes (Han et al. 2013), and the smallest bacterial genome is *Candidatus Nasuia deltocephalinicola* strain NAS-ALF; the genome consists of 112,091 bp in length and encodes137 proteins (Bennett and Moran 2013). The microorganism such as *Kineococcus radiotolerans* SRS30216, *Sorangium cellulosum* So0157–2, and *Rhodococcus aetherivorans* strain IcdP1 consists of (%GC) 74.4, 72.1, and 70.6,

Table 5.5 List of microorganism with genome size, %GC, gene content, and accession number

Organism	Length	Mb	% GC	No. of genes	RefSeq Accession
Escherichia coli UTI89	5,065,741	5.06	50.6	5363	NC_007946.1
Paeniclostridium sordellii strain AM370	3,550,458	3.55	27.9	3484	NZ_CP014150.1
Paenibacillus durus strain DSM 1735	6,038,347	6.03	50.8	5427	NZ_CP009288.1
Paenibacillus lautus strain E7593–69	7,128,120	7.12	51.2	6434	NZ_CP032412.1
Pseudomonas aeruginosa PAO1	6,264,404	6.24	66.6	5697	NC_002516.2
Pseudomonas putida KT2440	6,181,873	6.18	62.4	5389	NC_002947.4
Mycobacterium tuberculosis H37Rv	4,411,532	4.41	65.6	4008	NC_000962.3
Arcobacter butzleri RM4018	2,341,251	2.34	27	2332	NC_009850.1
Bacillus cereus ATCC 14579	5,411,809	5.41	35.3	5473	NC_004722.1
Rhodococcus hoagii 103S	5,043,170	5.04	68.8	4649	NC_014659.1
Rhodococcus aetherivorans strain IcdP1	5,922,748	5.92	70.6	5388	NZ_CP011341.1
Rhodococcus erythropolis PR4	6,516,310	6.51	62.3	6092	NC_012490.1
Candidatus Sulcia muelleri PSPU	285,352	0.285	20.9	296	NZ_AP013293.1
Kineococcus radiotolerans SRS30216	4,761,183	4.76	74.4	4536	NC_009664.2
Sorangium cellulosum So0157–2	14,782,125	14.78	72.1	11,021	NC_021658.1
Candidatus Tremblaya princeps	138,410	0.138	61.8	168	LN999011.1
Candidatus Nasuia deltocephalinicola strain NAS-ALF	112,091	0.11	17.1	165	NC_021919.1

The data presented in the above table is retrieved from NCBI Genome (https://www.ncbi.nlm.nih.gov/genome/) directory database

respectively, whereas *Candidatus Sulcia muelleri* PSPU and *Candidatus Nasuia deltocephalinicola* strain NAS-ALF consist of (%GC) 20.9 and 17.1, respectively (Table 5.5). Further, biochemical processes are the primary mechanism for driving biological processes that occur in different species of a living organism. Using genome sequencing various key metabolic pathways could be efficiently identified (Francke et al. 2005). Using such technique, the species-specific association between phenotypes and genotypes by network reconstruction of metabolic pathway can be performed, as it is applied widely for genome-scale metabolic model (Thiele and Palsson 2010).

The bacterial genome average protein coding density (PCD) is 87% with a usual range of 85–80% (McCutcheon and Moran 2011), but in some bacterial genomes, the protein coding density is less than 40%. Among these several are obligate pathogens and symbionts or consist of pseudogenes. As an example in an insect cosymbiont *Serratia symbiotica* str. Cinara cedri, the PCD is 38% and it comprises at least 58 pseudogenes (Lamelas et al. 2011). Similarly, the symbiotic cyanobacteria *Nostoc azollae* 0708 residing with fresh water fern consist of 52% PCD, which is the lowest of any other cyanobacteria (Ran et al. 2010). Although cyanobacteria *Trichodesmium erythraeum* IMS101 with 63% PCD harbor 12% of pseudogenes without the influence of environment, these cyanobacteria are free-living, nitrogen-fixing, bloom-causing, filamentous, and colony-forming and thrive in tropical and subtropical oceans with suitability to known reasons for undergoing a genome reduction (Pfreundt et al. 2014).

5.3 First-Generation DNA Sequencing

The DNA sequencing technology in the market was automated capillary sequencer also called chain termination sequencing or Sanger sequencing. In this sequencing chemistry, DNA is randomly fragmented, cloned into plasmid, and transformed to generally *E. coli*. The cloned fragment is amplified using flanking PCR primer. Each PCR round is terminated using incorporation of fluorescently labelled dideoxyribonucleotide (ddNTP). The resultant terminated fragments are then separated in electrophoretic capillary containing polymer gel, following exposing capillary to excite the fluorescently labelled dye by argon laser, and then emitted spectrum is recorded in a form of chromatogram using charge-coupled device camera. This gives read length of 800 to 1000bp with base call accuracy of 99.99%. However, its technology with very low output and high production cost limits the application (Swerdlow and Gesteland 1990).

5.3.1 Next-Generation Sequencing

In year 2005, massive parallel high-throughput sequencing technologies arrived among the scientific community also referred as next-generation sequencing, which delivers the tremendous output with high coverage and eventually becomes one of the essential tools for microbial genomics (Cao et al. 2017). The revolution of NGS over Sanger sequencing can be presented as (1) construction of multiplexed sequencing library, (2) clonal amplification of libraries, (3) immobilization of amplified libraries on solid substrate, and (4) chip-based sequencing. Depending on the variation in methodology used to immobilize DNA on a solid substrate and detection, the following technologies were mostly utilized in scientific community: (1) pyrosequencing, (2) sequencing by reversible termination, and (3) semiconductor sequencing.

5.3.1.1 Pyrosequencing

The first commercially launched next-generation sequencer was 454 GS20 pyrosequencing machine (Margulies et al. 2005). This technology is based on sequencing by synthesis and inorganic pyrophosphate-light emission detection chemistry. In this technology, initially DNA molecule is sheared using frequent site cutter restriction enzyme or fragmented through sonicator (nebulization). The sheared/fragmented DNA is end repaired and then subjected to oligonucleotide adapters and barcode ligation for multiplexing, a process called library preparation. The prepared library is then clonally amplified on beads (28 μm bead) with supplement of dNTPs, polymerase, and primer in an oil-water emulsion mixture, a process called emulsion PCR. The clonally amplified libraries were recovered, enriched, hybridized with sequencing primer, and loaded on picotiter plate for sequencing in the machine. The oil-water mixture acts as a microreactor for clonal amplification of sample on beads. During the sequencing, clonally amplified DNA fragments polymerized by the addition of nucleotides into daughter strands by sequencing polymerase result in the release of inorganic pyrophosphate (PPi). This released PPi combines with APS to form the ATP by sulfurylase, and then ATP combines with luciferin by luciferase resulting in the emission of oxyluciferin and light. This released light is captured by CCD camera in image format and then converted to nucleotides through image processing. The subsequent/iterative flow of sequencing cycles generates the average mean read length of 400–500 nucleotides (Margulies et al. 2005). More details are shown in Table 5.6. While producing the tremendous output, this technology is prone to sequencing of homopolymer repeats (Goodwin et al. 2016). Applying this technique, the first sequenced genome was bacterium *Myxococcus xanthus*, a soil inhabitant (Vos and Velicer 2006). Using such technology, a study of buffalo rumen microbial diversity associated with high roughage diet (Pitta et al. 2014b; Singh et al. 2015a) and fresh water (Dinsdale et al. 2008) has been performed.

5.3.1.2 Sequencing by Reversible Termination

The sequencing by reversible termination technology was implemented in Illumina Genome Analyzer (SOLEXA) marketed in the year 2006 (Fedurco et al. 2006). In this method, the sample preparation involves the random fragmentation, followed by the ligation of oligonucleotide adaptors and indexes, called sequencing libraries. The prepared libraries were amplified through bridge amplification (Adessi et al. 2000; Fedurco et al. 2006). The PCR forward and reverse primers complementary with adapters are hybridized on glass surface, amplified using modified DNA polymerase, a process called cluster generation. It is then followed by annealing of sequencing primer with adapters and followed by sequencing. In this sequencing chemistry, a modified DNA polymerase and different fluorophore-labelled nucleotides at 3′ are used. In each cycle, incorporation of single nucleotide followed to cleavage of fluorescent reporter which is the corresponding to the incorporated base and recorded by camera (Ju et al. 2006). The advancements in this technology permitted the 300$*$2 paired-end sequencing with a total average read length of 600 nucleotides (Table 5.6) (Goodwin et al. 2016). The limitation of this technology is high error rate of transition (Ts) to transversion (Tv) SNPs and Ts/Tv ratio.

Table 5.6 List of NGS machines with their chemistry, throughput, and runtime

Platform	Sequencing by	Detection	Read length (bp)	Throughput	Reads	Runtime
SOLiD 5500 Wildfire	Ligation	Fluorescence di-base probes	50 (SE)	160 Gb	~700M	6 day
SOLiD 5500xl	Ligation	Fluorescence of di-base probes	50 (SE)	160 Gb	~1.4B	10 day
454 GS Junior	Synthesis	Pyrophosphate	600(SE)	50 Mb	~0.1M	10 h
454 GS Junior+	Synthesis	Pyrophosphate	1,000(SE)	70 Mb	~0.1M	18 h
454 GS FLX Titanium XLR70	Synthesis	Pyrophosphate	600(SE)	600 Mb	~1M	10 h
454 GS FLX Titanium XL+	Synthesis	Pyrophosphate	1,000(SE)	750 Mb	~1M	23 h
Ion PGM 314	Synthesis	Proton	400 (SE)	60–150 Mb	1 M	3.7 h
Ion PGM 316	Synthesis	Proton	400 (SE)	500 Mb–1 Gb	2-3 M	5 h
Ion PGM 318	Synthesis	Proton	400 (SE)	0.5–2 Gb	4–6 M	8 h
Ion proton	Synthesis	Proton	Up to 200 (SE)	10 Gb	60–80 M	2–4 h
Ion S5 530	Synthesis	Proton	400 (SE)	5–8 Gb	15–25M	4 h
Ion S5 540	Synthesis	Proton	200 (SE)	10–15 Gb	60–80 M	2.5 h
Pacific BioSciences RS II	Synthesis	Fluorescence, phospholinked	~20Kb	400 Mb–1 Gb	~55,000	4 h
Pacific Biosciences Sequel	Synthesis	Fluorescence, phospholinked	8–12Kb	3.5–7Gb	~350,000	0.5–6 h
Oxford Nanopore MinION	Nanopore	Nanopores	200Kb	1.5Gb	>100,000	Up to 48 h
Illumina MiSeq v2	Synthesis	Reversible termination	250 (PE)	7.5–8.5 Gb	24–30M (PE)	39 h
Illumina MiSeq v3	Synthesis	Reversible termination	300 (PE)	13.2–15 Gb	44–50M (PE)	21–56 h
Illumina NextSeq 500/550 Mid	Synthesis	Reversible termination	150 (PE)	100–120 Gb	800M (PE)	29 h
Illumina HiSeq2500 v3	Synthesis	Reversible termination	100 (PE)	270–300 Gb	3 B (PE)	11 day
Illumina HiSeq2500 v4	Synthesis	Reversible termination	125 (PE)	450–500 Gb	4 B (PE)	6 day

(continued)

Table 5.6 (continued)

Platform	Sequencing by	Detection	Read length (bp)	Throughput	Reads	Runtime
Illumina HiSeq3000/4000	Synthesis	Reversible termination	150 (PE)	650–750 Gb	2.5 B (PE)	1–3.5 day
Illumina HiSeq X	Synthesis	Reversible termination	150 (PE)	800–900Gb per flow cell	2.6–3B (PE)	<3 day

Partially adapted from Goodwin et al. (2016). SE= single end, PE= pair end, Gb= giga base, M= million, B= billion, h= hours.

5.3.1.3 Semiconductor Sequencing

This sequencing technology is based on the detection of proton (H^+) released after the incorporation of nucleotide in a complementary strand. This released proton ion triggers an ion-sensitive field-effect transistor (ISFET) ion sensor as a signal, and generated signal is translated into the corresponding nucleotide through signal processing by Torrent Suite. The device on which sample is loaded consists of millions of microwells on a semiconductor chip in which sequencing occurs (Pennisi 2010). This technology library preparation is similar to pyrosequencing. The difference in library amplification through emulsion PCR, recovery and enrichment wherein pyrosequencing is time consuming, laborious while semiconductor (Ion Torrent) takes less time and labor.

5.4 Single-Molecule Real-Time (SMRT) Sequencing

The third-generation sequencer involves direct DNA sequencing without utilizing the PCR amplification step, as amplification introduces a bias in read content and presence of high GC content affects depth and coverage. The major advantage of this technique is the longer read length with an average of 5–10 Kb. With this technology, the first commercially launched chemistry was single-molecule real-time (SMRT®) by Pacific Biosciences (Eid et al. 2009). In this chemistry, sample library preparation involves the incorporation of DNA molecule to be circularized by ligating the adapter to both the ends of template. The prepared circular library is placed into SMRT® cell comprising 150,000 zeptoliter wells. Each well of SMRT cell contains single immobilized DNA polymerase (modified) at the bottom. The DNA polymerase binds with adapter sequence and then initiates the template replication. The incorporation of complementary four different fluorescently labelled nucleotides into reaction well. As the labelled base gets incorporated enzymatically, a light signal is generated and identified as the corresponding nucleotide (Eid et al. 2009). The general data output of PacBio RS II machine is 0.5–1 billion bases per SMRT® cell with very high error rate (typically 10–15%) (Goodwin et al. 2016). More details are presented in Table 5.6.

5.5 Oxford Nanopore

Another third-generation sequencer is MinIon commercialized by Oxford Nanopore Technology in 2014. In this technology, DNA/RNA is passes through a nanopore through electrophoresis, involves utilization of electrolytic solutions with constant electric field. As the DNA/RNA passes through nanopore, alteration in current occurs, and the resultant magnitude is recorded. MinIon library preparation consists of DNA fragmentation and end repaired, and then poly A tail is added to 3’OH end. In this two different adapter, a hair pin adapter and Y adapter (shape based). With the help of motor protein, sequencing templated dsDNA is unzipped at Y adapter and passes the ssDNA through nanopore. It is followed through base calling of ssDNA and hundred to thousand base pair read length is obtained, with an accuracy of 88% (Laszlo et al. 2014). More details are presented in Table 5.6. This technology delivers long reads, low cost, and small size with real-time nature of sequencing and invites attention in genomics and microbial community study (Judge et al. 2015).

5.5.1 Microbial Genome Sequencing and Bioinformatic Analysis

On the publication of first bacterial genome *Haemophilus influenza* (Fleischmann et al. 1995), the revolution in genomics data grew with tremendous improvements in sequencing mechanism such as application of paired-end sequencing and mate-pair sequencing (Pop 2009; Forde and O'Toole 2013; Cao et al. 2017). The publication of the first complete genome has led to the efforts to scientific community for the sequencing of larger genomes of *E. coli* (Blattner et al. 1997), *Bacillus subtilis* (Kunst et al. 1997), and eukaryotic genomes of *Saccharomyces cerevisiae* (Goffeau 1998), *Arabidopsis thaliana* (Arabidopsis Genome 2000), and ultimately the human genome (Venter et al. 2001). The advancement in genome sequencing has led to the development of various bioinformatic tools for de novo genome assembly and annotation. The most frequently used tools for genome assembly, majority of them, are command-line interface and available only for Ubuntu (free and open source) operating system. Among those, CLC-Bio, SOAP denovo2 (Luo et al. 2012), Velvet (Zerbino and Birney 2008), IDBA-UD (Peng et al. 2012), and SPAdes (Bankevich et al. 2012) are widely used. These tools detail algorithm and input data type, and dependencies are given in Table 5.7. With the development of computational tools for reference-based gene finder, the BLAST+ (Camacho et al. 2009), InterProScan (Quevillon et al. 2005), DIAMOND (Buchfink et al. 2015), and Blast2GO (Conesa et al. 2005) were highly used, while the ab initio gene prediction-based tools such as GeneMarkS (Besemer et al. 2001), GLIMMER (Delcher et al. 1999), AUGUSTUS (Stanke and Morgenstern 2005), and ORF Finder (Stothard 2000) were highly used. More details of each tool are presented in Table 5.8.

5.6 Application of NGS in Microbiome Study

5.6.1 16S rRNA Gene-Based Community Analysis

Various bacteria are un-cultivable in laboratory conditions, either they are unknown or suitable media compositions are unknown. Therefore to comprehensively study microbial composition and diversity, metagenomics was extensively applied. Metagenomics is described as a culture-independent approach to investigate the genetic diversity, community composition, and their interaction in their habitat (Handelsman 2004). The initial metagenomic study involves the microbial diversity using 16S rRNA gene-targeted amplicon sequencing (Schloss and Handelsman

Table 5.7 List of widely used tools for the microbial genome assembly

Assembler	Algorithm	Assembly method	Standard input	Read length	Pairedend	Output format	Availa bility	References
CLC-Bio	De Bruijn graph	Denovo and reference	Fasta, fastq	Arbitrary	Yes	Fasta, sam, bam	Licence	–
SeqMan Ngen	Patented	Denovo and reference	Fasta, fastq	Arbitrary	Yes	Fasta, sam, bam	Licence	–
SOAP denovo2	De Bruijn graphs	Denovo	Fastq	Arbitrary	Yes	Fasta	Open source	Luo et al. (2012)
MaSuRCA	Hybrid de (Bruijn graph +overlap-based)	Denovo	Fastq	Arbitrary	Yes	Fasta	Open source	Zimin et al. (2013)
Velvet	De Bruijn graphs	Denovo	Fastq	Arbitrary	Yes	Fasta	Open source	Zerbino and Birney (2008)
Meta-Velvet	De Bruijn graph	Denovo	Fastq	Arbitrary	Yes	Fasta	Open source	Namiki et al. (2012)
IDBA-UD	De Bruijn graph	Denovo	Fastq	Arbitrary	Yes	Fasta	Open source	Peng et al. (2012)
Meta-IBDA	De Bruijn graph	Denovo	Fastq	Arbitrary	Yes	Fasta	Open source	Peng et al. (2011)
CAP3	Overlap Layout Consensus	Denovo	Fasta	Arbitrary	No	Fasta	Open source	Huang and Madan (1999)
SPAdes	De Bruijn Graphs	Denovo	Fastq	Arbitrary	Yes	Fasta	Open source	Peng et al. (2012)

Table 5.8 List of tools used for gene identification and prediction in genomes and metagenomes

Gene	Input	Single/ Paired end	Output format	Availability	Suitability	References
Reference based						
BLAST+	Fasta, fastq	Both	.txt, sam, .xml	Open source	Genome, Metagenome	Camacho et al. (2009)
InterProScan	Fasta	Single	.txt, .xml	Open source	Genome, Metagenome	Quevillon et al. (2005)
DIAMOND	Fasta, fastq	Both	.txt, .sam, .xml, standard	Open source	Genome, Metagenome	Buchfink et al. (2015)
Usearch	Fasta, fastq	Both	standard	Open source	Metagenome	Edgar (2010)
RAPSearch	Fasta, fastq	Both	standard	Open source	Genome Metagenome	Ye et al. (2011)
PALADIN	Fasta, fastq	Both	standard	Open source	Metagenome	Westbrook et al. (2017)
Blast2GO	Fasta, fastq	Single	.txt, .xml	License	Genome	Conesa et al. (2005)
Ab-initio gene prediction						
Meta-GeneMark	Fasta, fatsq	Single	.txt	Open source	Metagenome	Zhu et al. (2010)
GLIMMER	Fasta	Single	.txt	Open source	Genome	Delcher et al. (1999)
GLIMMER -MG	Fasta, fatsq	Single	.txt	Open source	Metagenome	Kelley et al. (2012)
AUGUSTUS	Fasta	Single	.txt	Open source	Genome	Stanke and Morgenstern (2005)
FragGeneScan	Fasta, fastq	Single, paired	.txt	Open source	Metagenome	Rho et al. (2010)

GeneMark	Fasta	Single	.txt	Open source	Genome	Besemer et al. (2001)
ORF Finder	Fasta	Single	.txt	Open source	Genome	Stothard (2000)
Prodigal	Fasta	Single	.txt	Open source	Genome	Hyatt et al. (2010)

2005; Xu 2006) and later followed by whole metagenome shotgun sequencing (Reddy et al. 2014; Singh et al. 2014a) using NGS platforms.

The 16S rRNA gene consists of hypervariable regions of V1 to V9, with some conserveness between species to species, thus utilized as a molecular tool for bacterial characterization (Kolbert and Persing 1999). The high-throughput 16S rRNA amplicon sequencing analysis of habitats such as the gut (Claesson et al. 2012), oral cavity (Crielaard et al. 2011), and buffalo rumen (Pitta et al. 2014a) microbiota has been characterized. The taxonomic composition estimation using 16S rRNA depends on sampling site and varies organism to organism. As an instance, buffalo rumen (Patel et al. 2014; Singh et al. 2015a) and human digestive tract prevalent with *Bacteroidetes* and *Fermicutes* bacterial phyla with remarkable difference at phyla level (Human Microbiome Project Consortium 2012b). The 16S rRNA-based taxon abundance has been correlated with diet and health in human (Claesson et al. 2012; Conlon and Bird 2014). In summary, 16S rRNA-based study provides information for microbial abundance, diversity, and variation to diet alteration, effect of disease condition, and contribution in the ecosystem.

5.6.2 Whole Community Shotgun Metagenomics

The functional contribution of microorganism in various habitats is identifiable by performing the whole metagenome sequencing, and their annotation determines the functional genes (Singh et al. 2014b). The whole metagenome study revealed that the prevailing organism in the environment is correlated with genome size, GC content, horizontal gene transfer and optimum growth temperature (Popa et al. 2011; Wu et al. 2014), and antibiotic and metal ion resistance genes (Reddy and Dubey 2018). Metagenomic investigations also identified that microbes which thrive in soil generally have higher GC content with larger genome size compared to aquatic environment (Wu et al. 2014).

5.6.3 Metagenomics of Rumen

The animal rumen is anaerobic in nature and prevailing microbes are generally anaerobes, and thus these microbes are very difficult to culture in laboratory conditions and determination of molecular diversity. With the massive advancements of microbial community study using targeted 16S rRNA amplicon high-throughput sequencing, it becomes possible to explore the deeper insights of rumen microbiome diversity efficiently. Using such technique, various researchers applied the targeted 16S rRNA amplicon sequencing to characterize the adaptation of microbial community in response to experimental conditions. As an example, V3–V5 targeted amplicon in pre-ruminant calves results in the identification of 15 different phyla. Among these phyla, *Bacteroidetes* constituted 78% at the 42-day-old age and also in agreement that *Bacteroidetes* is one of the abundant phyla in ruminants (Li et al. 2012b). The wild ruminant *Tragelaphus strepsiceros*’s first metagenomic report showed that *Firmicutes* is dominant with 39% contribution of the total microbiota, followed by ~22% unassigned bacteria and then occurrence of *Bacteroidetes* (~18%) (Dube et al. 2015). The rumen microbiome adaptation to 50–100% forage diet investigation with respect to liquid and solid fraction, using V1 to V9 targeted amplicon study, indicated that *Bacteroidetes* were dominant in liquid fraction while *Fermicutes* were dominant in solid fractions (Pitta et al. 2014b). However, amplicon sequencing analysis provides insights of microbial community structure but is unable to explore the microbiota functional role in defined ecological niche. Therefore, application of whole metagenome sequencing removes such limitation and provides the functional role of microbes in the given niche. Using such technique, various studies had shown that various genes were involved in carbohydrate metabolism, protein metabolism, hydrolase activity, transferase and oxidoreductase activity, DNA and RNA metabolic process, butyrate and propionate metabolism (Patel et al. 2014), and methanogenesis and acetogenesis (Singh et al. 2015b). Functional annotation of whole metagenome data of Mehasani buffalo breed revealed that various environmental gene tags (EGTs) were involved in virulence disease and defense, stress response, and phages and prophages. The virulence disease and defense deeper study revealed that majority of EGTs were associated with resistance to antibiotic and toxic compounds (RATC). Similarly, stress response and phages and prophages extensive study revealed that heat shock, oxidative stress, and phages-prophages and pathogenicity islands were in majority (Reddy et al. 2014). Similarly, functional annotation of whole metagenome data of Jafarabadi buffalo revealed that various EGTs were significantly varied with a variation of feeding diet in liquid and solid fraction. In such study, EGTs such as carbohydrate, nitrogen, protein, DNA, sulfur, amino acid and derivative etc. EGTs exclusively associated with carbohydrate metabolism and protein metabolism such as monosaccharides, polysaccharides, di- and oligosaccharides, amino sugars and protein biosynthesis, protein degradation, and protein folding respectively, were also detected (Nathani et al. 2015). The most widely used tools for 16S rRNA amplicon classification are Quantitative Insights Into Microbial Ecology (QIIME (Kuczynski et al. 2011)), Mothur (Schloss et al. 2009), Ribosomal Database Project (Wang et al. 2007) etc., while for functional classification, Metagenomics Rapid Annotation using Subsystem Technology (MG-RAST (Keegan et al. 2016)), MEtaGenome ANalyzer (MEGAN (Huson et al. 2007)), and EBI-Metagenomics (Mitchell et al. 2018) have been frequently used. In overall, it gives the functional mechanism mediated by microbes in response to experimental conditions and invites the attention for developing catalogue of functional genes of aerobic and anaerobic microbes.

5.7 Metagenomics of Soil

Soil is the main site of food production and peculiar to support life functionality. Soil plays an essential role for plant growth, cycling of carbon, and other nutrients which are mainly mediated by soil microbiota. The first report on soil microbial community using DNA-based study revealed that soil microbiota composition is enormously diversified (Torsvik et al. 1996). The microbial community diversity of soil is mainly driven by soil properties and minimum by temperature and elevation (Xue et al. 2018). It is estimated that more than 10 K bacterial species are present in one gram of soil, with strongly correlated complex network (Nesme et al. 2016). The advancements in microbial genomics facilitated the soil microbiome study at various levels such as genus and species with abundance estimation (Nannipieri 2014), including the functional gene content and actively involved genes. Additionally, it is reported that microorganism displayed increased activity in soil hot spots such as mycosphere, rhizosphere, drilosphere, and detritusphere. The soil rhizosphere consists of surrounding complex microorganism and influenced by plant root, and these microbes play a vital role in plant growth and health promotion. For example, microorganisms beneficial to plants are symbiotic nitrogen-fixing rhizobia, the phosphate-solubilizing bacteria, and pathogen defeating such as *Pseudomonads* and *Bacilli* (Berendsen et al. 2012).

The one of highly studied genes of soil microbiota characteristic is *nif* various types. Among those, *nifH* was extensively targeted with different PCR primers for identification of N-fixing bacteria through molecular approach (Widmer et al. 1999; Zani et al. 2000), which is time-consuming with limited microorganism identification. The high-throughput sequencing analysis offers new horizons of diversity and composition estimation of soil microbiota across various soil niches without cultivation (Thompson et al. 2017). The deep metagenome study explored the microbial community functional capacity for carbon cycling (Howe et al. 2016) and correlation among community's functional genes (Hartman et al. 2017). There are some examples of big soil microbiome projects such as Earth Microbiome Project (EM) (Gilbert et al. 2014)), Brazilian Microbiome Project (Pylro et al. 2014), TerraGenome (Vogel et al. 2009), China Soil Microbiome Initiative (http://english.issas.cas.cn/), MicroBlitz (http://www.microblitz.com.au/), and EcoFINDERS (http://ecofinders.dmu.dk/), which characterized the soil microbiota community structure and functional diversity.

5.8 Metagenomics of Human Gut

Initially, the NGS-based 16S rRNA targeted amplicon sequencing provided the fast and cost-effective information of bacteria present in human gut (Qin et al. 2010). The MetaHIT consortium firstly performed the metagenomic study of the human microbiome of 124 Spanish and Danish subject stool samples. They showed that 1150 bacterial species were common in gut and a total of 3.3 million genes. However, 294,000 genes from 75 organisms were common in more than half samples (Qin et al. 2010). Sequenced data functional annotation revealed that various genes and pathways are involved in complex sugar metabolism, cell adhesion, vitamin, xenobiotic, and halogenated aromatic compound metabolism. On the other hand, the human microbiome project (HMP) was the largest for human host-associated microbiota characterization and reported that 3500 and 35,000 species-level operational taxonomic units (OTUs) in humans (Human Microbiome Project, 2012b). The GIT, oral cavity, and stool were the highly diversified, covering over 1000 OTUs from near about 150 genera. HMP data also showed that oral and GIT are more diversified than the back side of the elbow and ear. The diversity index of vaginal microbiome was the lowest with dominance of *Lactobacillus* (Huse et al. 2012) and becomes less diverse during pregnancy (Aagaard et al. 2012). Looking to the involvement of microbes for a functional role, stool dominated with complex carbohydrate degradation genes, whereas gut dominated with low abundance of hydrogen sulfide production and methionine degradation. The oral microbiota harbored genes for simple sugar metabolism and mostly for dextran, whereas vaginal microbiota harbored genes for glycogen and peptidoglycan degradation (Morgan et al. 2013).

Interestingly, high gut microbial community diversity is an essential feature of health. Aging and Crohn's disease are associated with bacterial diversity. The alteration of gut microbial community is well known to offer the progression of obesity, diabetes, and irritable bowel disease (Dicksved et al. 2008). The pathobionts are generally found in normal microbiota, while with certain alteration in homeostasis of the host, they increase the disparity by promoting the inflammation and production of bacteriocin and sometimes improving pathogenicity of other pathogens (Cho and Blaser 2012). It is established that the adult's microbiota is steady; however, broad-spectrum antibiotics kill the majority of commensal gut microbiota (Yassour et al. 2016). An experiment of ciprofloxacin 5-day course causes the reduction of gut bacterial diversity and quantified 30% species abundance (Dethlefsen et al. 2008). As antibiotic usually equally targets commensal microbes which are involved in metabolism and immunity, its removal potentially triggers malfunctioned metabolism and immune system. This offers development of susceptible environment for intestinal pathogens and homeostasis disparities.

The some examples of big project are the HMP (http://www.hmpdacc.org), MetaHIT (http://www.metahit.eu), and Global Ocean Survey (http://www.jcvi.org/cms/research/projects/gos/) applied such technique to explore the microbial diversity and functional genes, allowed our understanding of microbe contribution to sampled ecosystems. The National Institute of Health (NIH) sponsored HMP (http://www.hmpdacc.org) developed the 16S rRNA and whole metagenome data of large populations with comprehensive details of microbial communities at different bodies (Human Microbiome Project Consortium 2012a). This project developed an extensive reference of normal individuals and comparable with diseased individual microbiota (Human Microbiome Project Consortium 2012b; Li et al. 2012a).

5.9 Conclusion

The advent of high-throughput sequencing technology robustly enhanced the data generation, which allowed the massive whole genome sequencing, metagenomics, and their characterization. The taxonomic and functional analysis coupled with bioinformatic tools facilitated the development of microbial community and function genes catalogue. Among the published whole genomes, phyla such as *Proteobacteria*, *Firmicutes*, *Actinobacteria, Bacteroidetes*, and *Cyanobacteria* constitute nearly 96% of total phyla. The medical sector has contributed in the majority of genome projects as pathogens are greatly spreading with gain of resistance against antibiotics and host-associated ecosystem as a majority for biosamples. The metagenomic sequencing is a widely used tool for taxonomy and functional annotation and provided the identification of various novel genes from different ecological niches. This study shed light on available whole genomes and metagenomes and further provides the base for advanced application of next-generation sequencing and functional annotation.

References

Aagaard K, Riehle K, Ma J, Segata N, Mistretta TA, Coarfa C, Raza S, Rosenbaum S et al (2012) A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. PLoS One 7:e36466

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28(20):E87

Arabidopsis Genome I (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

Bennett GM, Moran NA (2013) Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. Genome Biol Evol 5:1675–1688

Berendsen RL, Pieterse CM, Bakker PA (2012) The rhizosphere microbiome and plant health. Trends Plant Sci 17:478–486

Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD et al (1997) The complete genome sequence of Escherichia coli K-12. Science 277:1453–1462

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Cao Y, Fanning S, Proos S, Jordan K, Srikumar S (2017) A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. Front Microbiol 8:1829

Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13:260–270

Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, Harris HM, Coakley M et al (2012) Gut microbiota composition correlates with diet and health in the elderly. Nature 488:178–184

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Conlon MA, Bird AR (2014) The impact of diet and lifestyle on gut microbiota and human health. Nutrients 7:17–44

Cook H, Ussery DW (2013) Sigma factors in a thousand E. coli genomes. Environ Microbiol 15:3121–3129

Crielaard W, Zaura E, Schuller AA, Huse SM, Montijn RC, Keijser BJ (2011) Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. BMC Med Genet 4:22

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641

Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol 6:e280

Dicksved J, Halfvarson J, Rosenquist M, Jarnerot G, Tysk C, Apajalahti J, Engstrand L, Jansson JK (2008) Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J 2:716–727

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C et al (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632

Dube AN, Moyo F, Dhlamini Z (2015) Metagenome sequencing of the greater kudu (Tragelaphus strepsiceros) rumen microbiome. Genome Announc 3

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Ekblom R, Wolf JB (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34(3):e22

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512

Forde BM, O'Toole PW (2013) Next-generation sequencing technologies and their impact on microbial genomics. Brief Funct Genomics 12:440–453

Francke C, Siezen RJ, Teusink B (2005) Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol 13:550–558

Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. BMC Biol 12:69

Goffeau A (1998) The yeast genome. Pathol Biol (Paris) 46:96–97

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351

Han K, Li ZF, Peng R, Zhu LP, Zhou T, Wang LG, Li SG, Zhang XB et al (2013) Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. Sci Rep 3:2101

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685

Hartman WH, Ye R, Horwath WR, Tringe SG (2017) A genomic perspective on stoichiometric regulation of soil carbon cycling. ISME J 11:2652–2665

Howe A, Yang F, Williams RJ, Meyer F, Hofmockel KS (2016) Identification of the core set of carbon-associated genes in a bioenergy grassland soil. PLoS One 11:e0166578

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Human Microbiome Project Consortium (2012a) A framework for human microbiome research. Nature 486:215–221

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

Huse SM, Ye Y, Zhou Y, Fodor AA (2012) A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One 7:e34242

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17:377–386

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119

Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, Li X, Marma MS et al (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci U S A 103:19635–19640

Judge K, Harris SR, Reuter S, Parkhill J, Peacock SJ (2015) Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. J Antimicrob Chemother 70:2775–2778

Keegan KP, Glass EM, Meyer F (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol 1399:207–233

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 40:e9

Kolbert CP, Persing DH (1999) Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. Curr Opin Microbiol 2:299–305

Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R (2011) Using QIIME to analyze 16S rRNA gene sequences from microbial communitie*s*. Curr Protoc Bioinformatics Chapter 10:Unit 10 17

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P et al (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390:249–256

Lamelas A, Gosalbes MJ, Manzano-Marin A, Pereto J, Moya A, Latorre A (2011) Serratia symbiotica from the aphid Cinara cedri: a missing link from facultative to obligate insect endosymbiont. PLoS Genet 7:e1002357

Land ML, Hyatt D, Jun S-R, Kora GH, Hauser LJ, Lukjancenko O, Ussery DW (2014) Quality scores for 32,000 genomes. Stand Genomic Sci 9:20

Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, Craig JM, Langford KW et al (2014) Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol 32:829–833

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. Nucleic Acids Res 39:D19–D21

Li E, Hamm CM, Gulati AS, Sartor RB, Chen H, Wu X, Zhang T, Rohlf FJ et al (2012a) Inflammatory bowel diseases phenotype, C. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition. PLoS One 7:e26284

Li RW, Connor EE, Li C, Baldwin Vi RL, Sparks ME (2012b) Characterization of the rumen microbiota of pre-ruminant calves using metagenomic tools. Environ Microbiol 14:129–139

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26

Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S et al (2018) EBI metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res 46:D726–D735

Morgan XC, Segata N, Huttenhower C (2013) Biodiversity and functional genomics in the human microbiome. Trends Genet 29:51–58

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Katta HY, Mojica A, Chen IA, Kyrpides NC et al (2018) Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res **47**:D649–D659

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

Nannipieri P (2014) Soil as a biological system and omics approaches. EQA – Int J Environ Qual 13:61

Nathani NM, Patel AK, Mootapally CS, Reddy B, Shah SV, Lunagaria PM, Kothari RK, Joshi CG (2015) Effect of roughage on rumen microbiota composition in the efficient feed converter and sturdy Indian Jaffrabadi buffalo (Bubalus bubalis). BMC Genomics 16:1116

Nesme J, Achouak W, Agathos SN, Bailey M, Baldrian P, Brunel D, Frostegard A, Heulin T et al (2016) Back to the future of soil metagenomics. Front Microbiol 7:73

Patel V, Patel AK, Parmar NR, Patel AB, Reddy B, Joshi CG (2014) Characterization of the rumen microbiome of Indian Kankrej cattle (Bos indicus) adapted to different forage diet. Appl Microbiol Biotechnol 98:9749–9761

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

Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428

Pennisi E (2010) Genomics. Semiconductors inspire new sequencing technologies. Science 327:1190

Pfreundt U, Kopf M, Belkin N, Berman-Frank I, Hess WR (2014) The primary transcriptome of the marine diazotroph Trichodesmium erythraeum IMS101. Sci Rep 4:6187

Pitta DW, Kumar S, Veiccharelli B, Parmar N, Reddy B, Joshi CG (2014a) Bacterial diversity associated with feeding dry forage at different dietary concentrations in the rumen contents of Mehshana buffalo (Bubalus bubalis) using 16S pyrotags. Anaerobe 25:31–41

Pitta DW, Parmar N, Patel AK, Indugu N, Kumar S, Prajapathi KB, Patel AB, Reddy B et al (2014b) Bacterial diversity dynamics associated with different diets and different primer pairs in the rumen of Kankrej cattle. PLoS One 9:e111710

Pop M (2009) Genome assembly reborn: recent computational challenges. Brief Bioinform 10:354–366

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609

Pylro VS, Roesch LF, Ortega JM, do Amaral AM, Totola MR, Hirsch PR, Rosado AS, Goes-Neto A et al (2014) Brazilian microbiome project: revealing the unexplored microbial diversity--challenges and prospects. Microb Ecol 67:237–241

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120

Ran L, Larsson J, Vigil-Stenman T, Nylander JA, Ininbergs K, Zheng WW, Lapidus A, Lowry S et al (2010) Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS One 5:e11486

Reddy B, Dubey SK (2018) River Ganges water as reservoir of microbes with antibiotic and metal ion resistance genes: high throughput metagenomic approach. Environ Pollut 246:443–451

Reddy B, Singh KM, Patel AK, Antony A, Panchasara HJ, Joshi CG (2014) Insights into resistome and stress responses genes in Bubalus bubalis rumen through metagenomic analysis. Mol Biol Rep 41:6405–6417

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

Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N et al (2018) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res

Schloss PD, Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol 6:229

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541

Singh KM, Reddy B, Patel AK, Panchasara H, Parmar N, Patel AB, Shah TM, Bhatt VD et al (2014a) Metagenomic analysis of buffalo rumen microbiome: effect of roughage diet on dormancy and sporulation genes. Meta Gene 2:252–268

Singh KM, Reddy B, Patel D, Patel AK, Parmar N, Patel A, Patel JB, Joshi CG (2014b) High potential source for biomass degradation enzyme discovery and environmental aspects revealed through metagenomics of Indian buffalo rumen. Biomed Res Int 2014:267189

Singh KM, Jisha TK, Reddy B, Parmar N, Patel A, Patel AK, Joshi CG (2015a) Microbial profiles of liquid and solid fraction associated biomaterial in buffalo rumen fed green and dry roughage diets by tagged 16S rRNA gene pyrosequencing. Mol Biol Rep 42:95–103

Singh KM, Patel AK, Shah RK, Reddy B, Joshi CG (2015b) Potential functional gene diversity involved in methanogenesis and methanogenic community structure in Indian buffalo (Bubalus bubalis) rumen. J Appl Genet 56:411–426

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465–W467

Stothard P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. BioTechniques 28(1102):1104

Swerdlow H, Gesteland R (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res 18:1415–1419

Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A et al (2017) A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463

Torsvik V, Sørheim R, Goksøyr J (1996) Total bacterial diversity in soil and sediment communities—a review. J Ind Microbiol 17:170–178

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M et al (2001) The sequence of the human genome. Science 291:1304–1351

Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD, Bailey MJ, Nalin R et al (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Microbiol 7:252

Vos M, Velicer GJ (2006) Genetic population structure of the soil bacterium Myxococcus xanthus at the centimeter scale. Appl Environ Microbiol 72:3615–3625

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267

Westbrook A, Ramsdell J, Schuelke T, Normington L, Bergeron RD, Thomas WK, MacManes MD (2017) PALADIN: protein alignment for functional profiling whole metagenome shotgun data. Bioinformatics 33:1473–1478

Widmer F, Shaffer BT, Porteous LA, Seidler RJ (1999) Analysis of nifH gene pool complexity in soil and litter at a Douglas fir forest site in the Oregon cascade mountain range. Appl Environ Microbiol 65:374–380

Wu H, Fang Y, Yu J, Zhang Z (2014) The quest for a unified view of bacterial land colonization. ISME J 8:1358–1369

Xu J (2006) Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Mol Ecol 15:1713–1731

Xue PP, Carrillo Y, Pino V, Minasny B, McBratney AB (2018) Soil properties drive microbial community structure in a large scale transect in south eastern Australia. Sci Rep 8:11725

Yassour M, Vatanen T, Siljander H, Hamalainen AM, Harkonen T, Ryhanen SJ, Franzosa EA, Vlamakis H et al (2016) Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med, 8:343ra381

Ye Y, Choi JH, Tang H (2011) RAPSearch: a fast protein similarity search tool for short reads. BMC Bioinformatics 12:159

Zani S, Mellon MT, Collier JL, Zehr JP (2000) Expression of nifH genes in natural microbial assemblages in Lake George, New York, detected by reverse transcriptase PCR. Appl Environ Microbiol 66:3119–3124

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38:e132

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677