Vijay Tripathi · Pradeep Kumar
Pooja Tripathi · Amit Kishore *Editors*

Microbial Genomics in Sustainable Agroecosystems

Volume 1

Springer

Microbial Genomics in Sustainable Agroecosystems

Vijay Tripathi • Pradeep Kumar
Pooja Tripathi • Amit Kishore
Editors

Microbial Genomics in Sustainable Agroecosystems

Volume 1

Springer

Editors
Vijay Tripathi
Department of Molecular and Cellular Engineering, Jacob Institute of Biotechnology and Bioengineering
Sam Higginbottom University of Agriculture, Technology and Sciences
Prayagraj, Uttar Pradesh, India

Pooja Tripathi
Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering
Sam Higginbottom University of Agriculture, Technology and Sciences
Prayagraj, Uttar Pradesh, India

Pradeep Kumar
Department of Forestry
North Eastern Regional Institute of Science and Technology (Deemed To Be University-MHRD)
Itanagar, Arunachal Pradesh, India

Amit Kishore
Department of Botany
Kamla Nehru P.G. College
Raebareli, Uttar Pradesh, India

ISBN 978-981-13-8738-8 ISBN 978-981-13-8739-5 (eBook)
https://doi.org/10.1007/978-981-13-8739-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

Microorganisms are found ubiquitous in nature, and studying the microbiota of plant, soil, ocean, and human body using omics approaches represents the metagenomes by sequencing techniques. The exponential advancement of the next-generation sequencing-based techniques, i.e., the microbial diversity analysis, has reached a milestone in genomics. From microbial evolution to microbial diversity, host-pathogen interactions to disease-causing genetic variation, and microbes producing industrially important enzymes to plant growth-promoting microbes, genomics has provided deep insights into the microbial world.

The present book, *Microbial Genomics in Sustainable Agroecosystems Volume I*, primarily gives insights from the microbial taxonomy to pathogen detection along with their applications and updated information. Genomic techniques enable comparative analysis of multiple genomes and metagenomes of complex agroecosystem environment. This book packs comprehensive information on the present omics technologies employed to reveal the microbial communities present in the different agroecosystems for sustenance.

I am happy to note that the editors have framed a conceptual book and edited it with updated information and precise language. I am sure this book will unfold existing doubts and bottlenecks on microbial genomics. The book draft has 14 well-written chapters starting from bacterial pan-genomics, functional genomics, to CRISPR-Cas9 for pathogen detection.

Although the main emphasis of this book is microbial genomics insights in different agroecosystems, it contains vast information with easy and understanding language. I am confident that this book has tremendous potential to attract a wide group of researchers, students, academicians, and faculty work.

Dr. Pramod W. Ramteke
FNAAS, FAMSc, FNABS, FLS, FBRS

Professor and Head, Department of
Biological Sciences,
Dean, Postgraduate Studies
Sam Higginbottom University
of Agriculture, Technology and Sciences
Allahabad, Uttar Pradesh, India

Preface

Over the past several decades, high-throughput sequencing technologies have been used for the detection of different microbial species which are fundamentally important in understanding human infections, development of diagnostics and vaccines, biodefense studies, antimicrobial target identification, and drug designing. The rapid advancement of next-generation sequencing platforms has reduced the time and cost and provides the capability to produce several hundreds of prokaryotic genomes each year.

The current book, *Microbial Genomics in Sustainable Agroecosystems*, is a comprehensive textbook that highlights the role of microbial genomic techniques in various disciplines of science. Complementary chapters focus on the microbial genomic approaches to understand the agroecosystems, bioremediation of soil and water from organic pollutants, characterization of microbial communities, carbon management, and energy production. This highlights the latest developments of microbial genomic approaches in the understanding and analysis of different microbial genomes.

The overall objective of this book is to cover the importance of microbial genomics in agriculture, environment, and food technology. The book not only deciphers the importance of microbial genomics but also makes the reader and researchers in the present scenario of microbial genomics and metagenomics familiar with the existing techniques and tools. Chapter 1 discusses the role of functional genomics and systems biology in the bioremediation process of organic and inorganic wastes generated from various sources including agriculture. Chapter 2 describes the application and use of pan-genome in different areas. The pan-genome helps researchers to differentiate pathogenic isolates in a large population and understand their pathology. The pan-genome can also be applied as a reference to find genes or sequences that are significantly linked to the particular phenotypes such as invasiveness or antibiotic resistance, thus crucial for facilitating bacterial disease control, vaccine development, and drug design. Chapter 3 describes factors allowing the successful establishment of the phyllosphere microbial population and the method used for their role in plant fitness and health. Pathobiome study and the potential use of phyllosphere monitoring are important in sustainable agriculture practices (e.g. biocontrol agents, plant growth stimulators, and biofertilizers). Chapter 4 discusses the role of functional genomics to study the effect of stress on plants. Various approaches and tools of systems biology, required

for alteration in biological networks including gene regulatory, protein-protein, and metabolic network, are also discussed in detail. Chapter 5 elaborates the available genome in public databases and the role of next- and third-generation sequencing technology in the growth of genome and metagenome sequencing and its associated projects and their technology and functional characterization through bioinformatic analysis. Chapter 6 details the computational methods incorporated with sequencing techniques in the viral genome. This chapter mainly focuses on how the computational method incorporated with sequencing technique is made easy for microbial detection and characterization. Therefore, Chap. 7 discusses the importance of whole-genome sequencing techniques with great potential in the food safety and surveillance against foodborne pathogens and antimicrobial resistance. Chapter 8 comprises the understanding of taxonomic and functional microbiome profiles caused by the occurrence of microbial pathogens in the food production system. This can lead to valuable information to understand the negative effect of pathogens on crops. Chapter 9 describes the potential of microbes for energy production so that this may lead to limited usage of fossil fuels in the near future. The brief section of the role microbial world, ecosystem, and their relationship with the climatic change is also discussed in this chapter. Therefore, Chap. 10 summarizes the different methods to get the microbial diversity that may eventually enhance plant growth in sustainable agriculture. This chapter also highlights the molecular techniques for the identification and genotyping of microorganism, and these techniques developed for sequencing are a breakthrough for microbial systematics. Therefore, Chap. 12 is divided into two sections: the first section explains the molecular mechanism of the naturally occurring CRISPR-Cas9 systems, and the next section elaborates the applicative part of CRISPR-Cas9-based tools/systems in a range of microorganisms and parasites. Chapter 13 discusses the brown algae and its characteristics as bio-sorbent for the removal of heavy metals from industrial effluent. Chapter 14 discusses the use of microbial genomics approach in renewable energy production and global carbon management. The chapter addressed the global energy crises and suggests various omics technologies develop a better understanding to harness different renewable and carbon-neutral energy sources like lignocellulosic biomass, microalgae, and cyanobacteria to overcome this problem.

This book provides comprehensive knowledge about the use of microbial genomics approach in solving the various research questions related to agroecosystems. At the same time, it explains to the biologist some of the basics behind the next-generation sequencing techniques that are necessary for research in this field. The book will be mainly useful for the students, research scholars, academicians, and scientists who are studying and working in the field of genomics.

Prayagraj, Uttar Pradesh, India — Vijay Tripathi
Itanagar, Arunachal Pradesh, India — Pradeep Kumar
Prayagraj, Uttar Pradesh, India — Pooja Tripathi
Raebareli, Uttar Pradesh, India — Amit Kishore

Contents

Editors and Contributors

About the Editors

Vijay Tripathi is currently working as an Assistant Professor at the Department of Molecular and Cellular Engineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Prayagraj, India. He was previously awarded an ARO Post Doctoral Fellowship at the Department of Soil, Water, and Environmental Science, Agricultural Research Organization, Bet Dagan, Israel. He has also received two prestigious postdoctoral fellowships (Indo-Israel Government Fellowship and PBC Outstanding Post Doctoral Fellowship) and worked with Prof Edward Trifonov as a postdoctoral fellow at the Institute of Evolution, University of Haifa, Israel. Dr Tripathi began his research career as a doctoral student at the Center of Bioinformatics, University of Allahabad, India. During his doctoral thesis work he was also awarded a MUIR fellowship and visited the University of Cagliari, Italy.

Pradeep Kumar is currently working as an Assistant Professor at the Department of Forestry, North Eastern Regional Institute of Science and Technology (NERIST), Nirjuli, India. Before joining NERIST, he served as an International Research Professor/Assistant Professor at the Department of Biotechnology, Yeungnam University, South Korea. He was awarded a PBC Outstanding Post Doctoral Fellowship to work for more than three years as a Postdoctoral Researcher at the Department of Biotechnology Engineering, Ben Gurion University of the Negev, Israel. He is the recipient of many best paper presentations and the Narasimhan Award from the Indian Phytopathological Society, India. He has published four books and more than 50 research and review articles in peer-reviewed journals, as well as several book chapters.

Pooja Tripathi is currently working as an Assistant Professor at the Department of Computational Biology & Bioinformatics, JIBB at Sam Higginbottom University of Agriculture, Technology and Sciences, Prayagraj, India. She completed her PhD in Bioinformatics at the University of Allahabad, Prayagraj. She was also awarded the prestigious PBC Outstanding Post Doctoral Fellowship from the Ministry of Higher Education, Israeli Government, to pursue her postdoctoral research at the Department of Plant and Environmental Science, Weizmann Institute of Science, Rehovot, Israel. She also received an ARO Post Doctoral Fellowship and joined the

Department of Plant Pathology and Weed Research, Agricultural Research Organization, Bet Dagan, Israel as a postdoctoral fellow.

Amit Kishore is graduated from V.B. Singh Purvanchal University, Jaunpur, India, then after postgraduate and doctoral degree in Botany from Banaras Hindu University, Varanasi, India. He gained his post-doctoral research experience in host-pathogen interaction area from Agricultural Research Organization (ARO), Israel. Currently, Dr. Amit Kishore is working as an Assistant Professor (Botany Department) at Kamla Nehru Post Graduate College, Raebareli, India. He has been credited by several by fellowships and awards like CSIR- NET JRF, CSIR SRF, Best poster presentation and ARO post-doctoral fellowship. Dr. Singh is life member of the Association of Microbiologists of India (AMI) and Biotech Research Society, India (BRSI).

Contributors

Gulab Chand Arya Department of Vegetable and Field Crops, Institute of Plant Sciences, Agricultural Research Organization, Volcani Center, Rishon LeZion, Israel

Shiv Bharadwaj Department of Biotechnology, College of Life and Applied Sciences, Yeungnam University, Gyeongbuk-do, Republic of Korea

Yashpal Bhardwaj Laboratory of Molecular Ecology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Gargi Bhattacharjee School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

Shatabisha Bhattacharjee School of Basic Sciences, Indian Institute of Technology (IIT), Mandi, Himachal Pradesh, India

Kadiya Calderón Departamento de Investigaciones Científicas y Tecnológicas de la Universidad de Sonora, Universidad de Sonora, Hermosillo, Sonora, Mexico

Vivek Dhar Dwivedi Center for Bioinformatics, Pathfinder Research and Training Foundation, Greater Noida, Uttar Pradesh, India

Estefanía Garibay-Valdez Centro de Investigación en Alimentación y Desarrollo, A.C. Coordinación de Tecnología de Alimentos de Origen Animal, Hermosillo, Sonora, Mexico

Nisarg Gohil School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

Arye Harel Department of Vegetable and Field Crops, Institute of Plant Sciences, Agricultural Research Organization, Volcani Center, Rishon LeZion, Israel

Arash Iranzadeh Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Aditi Jangid School of Basic Sciences, Indian Institute of Technology (IIT), Mandi, Himachal Pradesh, India

Khushal Khambhati School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

Nikhil Kirtipal Department of Biotechnology, Modern Institute of Technology, Rishikesh, Uttarakhand, India

Asunción Lago-Lestón Centro de Investigación Científica y de Educción Superior de Ensenada, Ensenada, BC, Mexico

Ashutosh Mani Department of Biotechnology, Motilal Nehru National Institute of Technology, Allahabad, India

Luis Rafael Martínez-Córdova Departamento de Investigaciones Científicas y Tecnológicas de la Universidad de Sonora, Universidad de Sonora, Hermosillo, Sonora, Mexico

Marcel Martínez-Porchas Centro de Investigación en Alimentación y Desarrollo, A.C. Coordinación de Tecnología de Alimentos de Origen Animal, Hermosillo, Sonora, Mexico

Vishal Mishra School of Biochemical Engineering, IIT (BHU) Varanasi, Varanasi, Uttar Pradesh, India

Nicola Jane Mulder Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Neha Laboratory of Molecular Ecology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Happy Panchasara School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

Shreya Patel School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

Mayank Pokhriyal Viral Research and Diagnostic Laboratory (VRDL), Government Medical College, Haldwani, Uttarakhand, India

Tulika Prakash School of Basic Sciences, Indian Institute of Technology (IIT), Mandi, Himachal Pradesh, India

Barkha Ratta Division of Biochemistry, Indian Veterinary Research Institute, Bareilly, Uttar Pradesh, India

Bhaskar Reddy Centre of Advanced Study in Botany, Institute of Science, Banaras Hindu University, Varanasi, India

Abhishek Singh Laboratory of Molecular Ecology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Jyoti Singh School of Biochemical Engineering, IIT (BHU) Varanasi, Varanasi, Uttar Pradesh, India

Vijai Singh School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India
Present address: Department of Biosciences, School of Sciences, Indrashil University, Kadi, Gujarat, India

Francisco Vargas-Albores Centro de Investigación en Alimentación y Desarrollo, A.C. Coordinación de Tecnología de Alimentos de Origen Animal, Hermosillo, Sonora, Mexico

Brijesh S. Yadav Center of Bioengineering Research, University of Information Science and Technology (UIST), St. Paul, Ohrid, Republic of North Macedonia

Birendra Singh Yadav Department of Biotechnology, Motilal Nehru National Institute of Technology, Allahabad, India

Priyanka Yadav School of Biochemical Engineering, IIT (BHU) Varanasi, Varanasi, Uttar Pradesh, India

Suman Yadav Laboratory of Molecular Ecology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi, Uttar Pradesh, India

1 Functional Genomics and System Biology Approach in Bioremediation of Soil and Water from Organic and Inorganic Pollutants

Suman Yadav, Yashpal Bhardwaj, Neha, and Abhishek Singh

Abstract

In the modern world, where environmental pollution is a great concern, a cost-effective and environmental-friendly way is the necessity of the time. In the recent past, bioremediation emerges as a promising tool to deal with this problem. Several types of microbes and their bioremediation strategies are reported till date. The system biology and functional genomic approaches in bioremediation of wastes are the need of the hour. However, biological activities are complex and need a better understanding of interactions and networks at molecular, cellular, community, and ecosystem level, which can be achieved by system biology approach. The knowledge of these interactions is helpful in understanding the strategies adopted by the microbes at a polluted site. Several molecular and bioinformatic tools such as genomics, transcriptomics, proteomics, metabolomics, etc. are used to gather the knowledge of cellular interaction at different levels on a polluted site. Similarly, functional genomics also use the knowledge of genomics and transcriptomics to describe the gene functions and interactions in bioremediation process and helpful in phylogenetic identification of microbes involved in bioremediation. In this chapter, we discuss the role of functional genomics and system biology in bioremediation process of organic and inorganic wastes generated from various sources including agriculture.

Authors Suman Yadav and Abhishek Singh have equally contributed to this chapter.

S. Yadav · Y. Bhardwaj · Neha · A. Singh (✉)
Laboratory of Molecular Ecology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi, Uttar Pradesh, India
e-mail: abhishek.pharma08@gmail.com

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_1

1.1 Introduction

Any unwanted substance coming into the environment through various processes is referred to as a "contaminant." Deleterious effects or damages on the environment, caused by the contaminants, lead to "pollution" (Megharaj et al. 2011). Relative to the pre-industrialization era, abundant use of chemical substances such as petroleum oil, hydrocarbons [e.g., aliphatic, aromatic, and polycyclic aromatic hydrocarbons (PAHs); BTEX (benzene, toluene, ethylbenzene, and xylenes); chlorinated hydrocarbons like polychlorinated biphenyls (PCBs), trichloroethylene (TCE), and perchloroethylene; nitroaromatic compounds; and organophosphate compounds], solvents, pesticides, and heavy metals is the main cause of environmental pollution in the current age of industrialization. Due to their mobile and water-soluble nature, many organic and inorganic chemicals such as pesticides, fertilizers, and heavy metals are present in drainage and runoffs from the contaminated sites. Regular cycling via volatilization and condensation of many organic chemicals such as volatile organic carbons (VOCs) are consequently found in rain, fog and snow (Dubus et al. 2000). When these pollutants migrated from contaminated site to non-contaminated place as vapors and leachate through the soil or as dust, further contributes to the pollution of natural ecosystems (Lopez-Errasquin and Vazquez 2003). Soil and water resources are polluted severely with various organic and inorganic pollutants which leads to deterioration of soil and water quality. The pollutants are mainly polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), explosives, metals, metalloids, and radionuclides reported in soils (Zhu and Shaw 2000; Kumar et al. 2012; Testiati et al. 2013; Vane et al. 2014). These pollutants from soil are reaching into underground and surface water by leaching and runoff, respectively, thus polluting water bodies. The presence of these pollutants in soil and water causes hazardous effect due to their toxic nature which consequently hampers the human health and environment.

Nowadays, pesticides are more commonly used in agricultural ecosystem and public health program worldwide. In several cases, the environmental effects of these chemical substances overshadow the profits that human beings are getting. This necessitates the remediation of these chemicals after their deliberated use. Bioremediation process is an effective and economical way to curb out such environmental problems by enhancing the natural degradation processes.

Bioremediation uses plants and/or microorganisms, such as bacteria, protozoa, and fungi to degrade contaminants into a less toxic or nontoxic compounds (Pierzynski et al. 1994; US-EPA 1996). In the modern era, various tools and techniques such as genomics, transcriptomics, proteomics, metabolomics, phenomics, and lipidomics are used to examine the system biology in different environments. It is used to unravel, optimize, propose, and scrutinize the function and survival tactics for bioremediation in the desired ecosystem (de Lorenzo 2008; Chakraborty et al. 2012). The functional genomic approach is useful in identification of novel microbes and their strategies adopted for bioremediation and to uncover the biological function of a gene and their products in a cell. The present-day innovations in metagenomics and whole genome sequencing create a whole new

possibility for identification and function of novel genes and their regulatory elements involved in biodegradation of various types of contaminants from various types of cultivable and non-cultivable microbes from their respective environments (Golyshin et al. 2003; Zhao and Poh 2008).

This chapter highlights the system biology approaches with emphasis on functional genomics to effective bioremediation.

1.2 Organic and Inorganic Pollutants in Soil and Water

Soil is the "the biogeochemical engine of Earth's life support system" (Robinson et al. 2012). It provides us food, fodder, fiber and fuel. In addition to these, many agricultural and forestry goods, soils deliver ecosystem services that we cannot get in markets. These life-giving functions include recycling of carbon and essential nutrients of all living materials, filtering and storage of water, and regulation of the atmosphere and biological control of pests (Robinson et al. 2012). Soil organic matter is composed of the biological and chemical degradation of dead plant and animal remains in soil. After several alterations and cross-linkages, the products that are formed show no resemblance to the plants and animals and are called as natural organic matter or humus, which is defined as the total organic compounds, exclusive of undecayed plant and animal tissues, their "partial decomposition" products and the soil biomass (Stevenson 1994). In recent years the input of the heavy metal has been increased in the nature due to anthropogenic activities (Shi et al. 2007) like traffic, industry, waste disposal, and agricultural practices (Weber and Karczewska 2004). It is reported that the presence of hydrocarbons and metals in the soils adversely affects the seed germination, plant growth (Smith et al. 2006; Ahmad and Khan 2012), soil microbial population and activities (Guo et al. 2012; Alrumman et al. 2015), and the ability of plants and microorganisms to absorb water and nutrients from the soil (Nie et al. 2009). Pesticides have also been reported as toxic to humans (Anderson and Meade 2014). Likewise, they adversely affect other forms of life including arthropods (Desneux et al. 2007), plants (Siddiqui and Ahmed 2006), soil microbes (Ahemad and Khan 2012), etc. In soils the heavy metal content depends on the geological parent material, soil-forming processes, and anthropogenic activities. The agricultural practices, like use of superphosphate fertilizers, instigate the soil contamination by cadmium, while calcium nitrate can contain significant amounts of nickel. Certain fungicides containing copper and zinc increase the presence of these elements in the upper soil horizons (Lopez-Mosquera et al. 2005). In rural areas population is relatively small; thus most fertilizers, pesticides, and eroded soil are released as pollutants and reach the water bodies through runoff after rain and flood (Letchinger 2000), leading to eutrophication in freshwater bodies.

Phosphate plays a major role in eutrophication and promotes cyanobacterial growth when present in higher concentration, which ultimately decreases the dissolved oxygen concentration in water (Werner 2002). These cyanobacterial blooms release some harmful toxins that get accumulated in the food chain (Schmidt

et al. 2013). Nitrogen-rich fertilizer causes dissolved oxygen deficiency in rivers, lakes, and coastal zones which has hazardous effects on oceanic fauna. These fertilizers have high water solubility and increased runoff and leaching rate which results in groundwater pollution (Rosen and Horgan 2009; NOFA 2004; Singh et al. 2006). Sandy soil favors leaching (EFP 2015), thus promoting pollutant flow from the soil. Selenium (Se) is a heavy metal that is naturally found in soil, but irrigation practices favor its accumulation in the soil which ultimately reaches to water reservoirs and is very toxic for animals and humans (Ganje 1966).

1.3 Mechanism of Bioremediation

It is a process mediated by the microorganism to degrade the contaminants or pollutants into less toxic compounds. The bioremediation process includes (1) microorganisms or plants, (2) a potentially biodegradable contaminant, and (3) a bioreactor in which the process can take place. The microbes in the bioreactor use carbon of the contaminants as a source of energy and, in doing so, degrade it. Bioremediation can be applied both ex situ and in situ. With ex situ bioremediation, the contaminated soil is excavated or the groundwater is extracted prior to treatment, while in situ remediation does not require excavation or extraction. As a result, the contaminated soil or groundwater serves as the bioreactor. The microorganisms may occur at the site naturally or be introduced from other locations. Microbial-mediated bioremediation has a great potential to effectively restore contaminated environment. A number of microorganisms are considered to be the best candidates among all living organisms to remediate most of the environmental contaminants into the natural biogeochemical cycle due to their diversity, versatility, and adaptability in the environment (Table 1.1).

Various types of organic pollutants are released in the environment; hence, a variety of microbes are needed for effective bioremediation of these pollutants (Table 1.1). *Pseudomonas putida* is the first bacteria to be patented for bioremediation (Prescott et al. 2002). Biodegradation of organic pollutants in the environment is either carried out in the presence of oxygen (oxidation) or under anaerobic condition by denitrification, methanogenesis, and sulfidogenesis.

Several microorganisms are reported to degrade PAHs in the environment (Table 1.1). Three mechanisms of PAH degradation by microbes are reported: (1) bacterial degradation, (2) lignolytic fungal degradation, and (3) non-lignolytic bacterial degradation. However, the "oxidation of the aromatic ring, followed by the systematic breakdown of the compound to PAH metabolites and/or carbon dioxide," is common in all three mechanisms (Bamforth and Singleton 2005).

Atrazine and organophosphate are the most commonly used pesticides in agriculture. Atrazine is commonly used as herbicide in crop systems such as maize, sorghum, and sugarcane. Despite having only one chlorine residue, atrazine is resistant to biodegradation. *Pseudomonas* sp. ADP degrades atrazine cyanuric acid using the AtzA, AtzB, and AtzC enzymes. Atrazine is converted to hydroxyatrazine by AtzA; then AtzB converts hydroxyatrazine to N-isopropylammelide by

Table 1.1 List of bacteria studied for bioremediation of pollutants

Pollutant	Organism	References
2, 4, 6-Trinitrotoluene (TNT)	*Methanococcus* sp.	Boopathy and Kulpa (1994)
Atrazine	*Pseudomonas* sp. (ADP)	Newcombe and Crowley (1999)
PCB formaldehyde	*Candida tropicalis*	Ijah (1998)
Polycyclic hydrocarbon	*Xanthomas*	Jogdand (1995)
Halogenated hydrocarbon	*Corynebacterium* spp.	Jogdand (1995)
Alkyl benzene sulfonate polycyclic aromatics	*Alcaligenes* spp.	Lal and Khanna (1996)
Chlorpyrifos	*Enterobacter* strain B-14	Singh et al. (2004)
Dibenzothiophene (DBT)	*Rhizobium meliloti*	Frassinetti et al. (1998)
Hexahydro-1,3,5- trinitro-1,3,5-triazine (RDX)	*Acetobacterium paludosum*	Sherburne et al. (2005)
	Clostridium acetobutylicum	Zhang and Hughes (2003)
PAHs	*Pseudomonas* sp.	Arun et al. (2008)
Phenanthrene, PAH	*Agrobacterium*, *Bacillus*	Aitken et al. (1998)
Polychlorinated biphenyl (PCB)	*Rhodococcus*	Chung et al. (1994)
Endosulfan	*Bacillus*	Mohamed et al. (2011)
Chlorpyrifos	*Enterobacter*	Niti et al. (2013)
Malathion	*Arthrobacter*	Hussaini et al. (2013)
Chlorpyrifos and methyl parathion	*Acinetobacter* sp., *Pseudomonas*	Ravi et al. (2015)
Heavy metals, lead, mercury, and nickel	*Saccharomyces cerevisiae*	Chen and Wang (2007), Infante et al. 2014
Cobalt, copper, Chromium	*Lysinibacillus sphaericus CBAM5*	Peña-Montenegro et al. (2015)
Cadmium	*Aspergillus* sp., *Microsporum* sp., *Cladosporium* sp.	Soleimani et al. (2015)
Fe(III), U(VI)	*Geobacter* spp.	Mirlahiji and Eisazadeh (2014)
Monocyclic aromatic hydrocarbons	*Pseudomonas putida*	Safiyanu et al. 2015

hydrolytic deamidation. Further, N-isopropylammelide is converted to cyanuric acid by AtzC which is finally converted to ammonia and carbon dioxide by other bacteria (Ang et al. 2005; Krutz et al. 2009).

Organophosphates are the most widely used pesticide in the agricultural sector. Several organophosphate-degrading microorganisms including bacteria and fungi are reported till date (Table 1.1). It is mainly degraded by the phosphotriesterase (PTE) group of enzymes which is found in microbes, plants, and animals.

PCB biodegradation in contaminated environments is very difficult due to its structural complexity and level of chlorination and is generally carried out by

co-metabolite mean. Dechlorination of PCBs removes *m*- and *p*-chlorines from highly chlorinated congeners and resulted in *ortho*-substituted mono- and tetrachlorobiphenyls (Wiegel and Wu 2000). PCB degradation by microbes occurs via a meta-cleavage pathway which produces tricarboxylic acid cycle intermediate and (chloro)benzoate (CBA) (Bruhlmann and Chen 1999).

Being resistant to degradation, metal ions are difficult to remove from the contaminated sites. However, reducing the bioavailability of these metal ions is an effective strategy for reducing their toxic effects. Microbes adopt different tactics to remove the metal ions from the environment such as (i) metal ion exclusion by permeability barrier, (ii) intra- and extra-cellular sequestration, (iii) active efflux pumping, (iv) enzymatic reduction, and (v) enzymatic reduction in the sensitivity of cellular receptors to metal ions (Ji and Silver 1995; Nies and Silver 1995; Rensing et al. 1999). The microbes mainly adopt two types of transformation strategies: (1) changing the oxidation state of inorganic forms and (2) conversion of inorganic form in to organic form (methylation and demethylation).

These microorganisms display a remarkable range of contaminant degradability that can efficiently restore natural environmental conditions (Pandey et al. 2003; Labana et al. 2005). By pure culture method, many microbes have been studied earlier about their remediation capacity, but there are enormous bacteria and fungi that are present in the environment that cannot be isolated but very potent for chemical degradation. To explore these types of microbes and their bioremediation strategies in minimum time, genomics and metagenomics are used. Further, system biology and functional genomic approach is applied to successful bioremediation of pollutants from the environment.

1.4 Role of System Biology in Bioremediation

The system biology approach for bioremediation requires the identification and characterization of microbial communities and the molecular processes involved, but it becomes difficult due to the fact that the toxic contaminants present are influencing the normal activity of microbial communities. The main objective of bioremediation is to degrade or detox the toxic chemicals, which is not possible without knowing all plausible environmental factors that influence the cell-to-cell interactions. Various bioremediation processes need different system biology approaches. If the purpose is to identify the microbial community structure, the DNA-based genomics tool such as 16S rRNA clone library, PhyloChip, or sequencing can be useful. If the aim is to identify the functional genes involved and to understand the associated cellular pathways with microbial bioremediation, various tools such as RNAseq, GeoChip (for RNA), and several mass spectroscopy methods (for proteins) are used. Similarly, if we want to identify and characterize the metabolites produced by the microbes during bioremediation, the metabolomics tools such as high-performance liquid chromatography (HPLC) and gas chromatography (GC) coupled with mass spectroscopy (MS), matrix-assisted laser desorption/

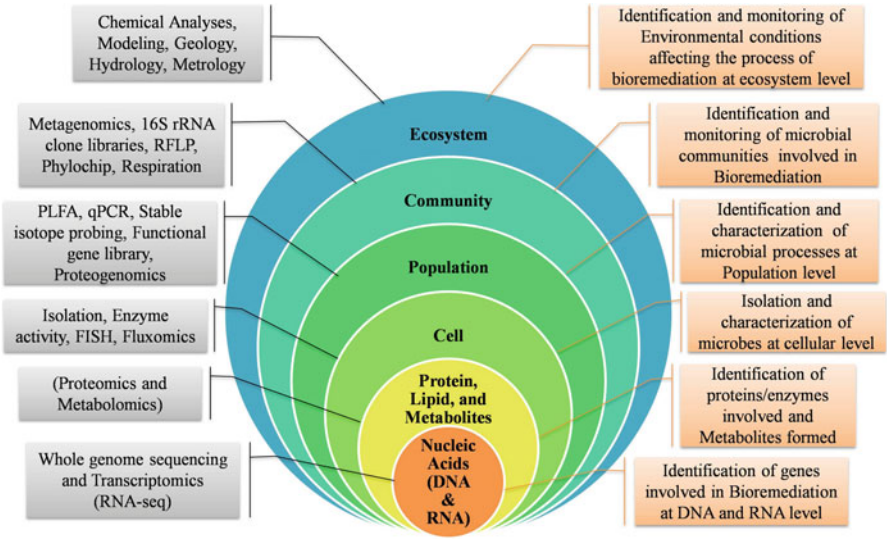

Fig. 1.1 System biology approach to bioremediation: Identification of governing factors from subcellular level to ecosystem level using the geochemical, ecological, genomic, proteomic, metabolomic, and computational techniques. The impact of environment on bioremediation at cellular level is understood by analyzing DNA, RNA, and protein. Communities and populations involved in bioremediation are analyzed to understand the effect on structure and function at ecosystem level

ionization (MALDI), nuclear magnetic resonance (NMR), and desorption electrospray ionization (DESI) are helpful. For the development of a useful conceptual model based on system biology, continuous monitoring of limiting nutrients, electron acceptors, electron donors, and hydrology is also needed.

It is important to understand the complex on-site bioremediation activities in order to apply the system biology approach. This can be achieved by various monitoring techniques that record and monitor terminal electron donors and electron acceptors, enzyme probes to access the functional activity in an ecosystem, functional genomic microarray, phylogenetic analysis, and metabolomics, proteomics, and quantitative PCR. These tools provide a greater understanding of microbial processes which are involved in bioremediation. An ecosystem generally consists of communities, populations, cells, proteins, lipids, and nucleic acids. The nucleic acids (DNA and RNA), proteins, and lipids can be analyzed at the cellular level to know their effect on the cells and are also used to analyze populations and community to figure out the consequences of bioremediation on structural and functional relationships in an ecosystem (Fig. 1.1).

The detoxification of some hazardous metals and organic compounds can be achieved by simply changing their redox state. This method is particularly useful for heavy metals and radioactive nuclei. Degradation of these chemicals is not possible, so biotransformation is an effective strategy against these contaminants as it reduces

their bioavailability, their mobility, and thus their harmful effect (Desjardin et al. 2002; Suzuki et al. 2005). In the presence of suitable electron donors, microbes can immobilize such contaminants by altering their redox state and using them as electron acceptor (Gadd 2000). The measurement of microbial respiration and metabolism provides a better insight during the bioremediation process. The enzyme activity measurement during bioremediation is a cost-effective, reliable way to monitor the microbial respiration and metabolism. TTC (triphenyl tetrazolium chloride) assays (Guochen et al. 2011) and INT (iodo-nitro-tetrazolium) (Mathew and Obbard 2001) are the commonly used dehydrogenase enzyme-based assays which are successfully implied in measuring microbial respiration during bioremediation of metals, polycyclic aromatic hydrocarbons (PAHs), and oils (Mosher et al. 2003; Bento et al. 2005). Several other molecular probes (based on enzyme or DNA/RNA) from key bioremediation pathways can also be used to track the metabolic pathways. Such probes are widely used to sense the degradation processes, e.g., in trichloroethylene (TCE), and of petroleum hydrocarbon degradation (Hazen et al. 2009; Bell et al. 2011; Beller et al. 2002). Rhee et al. (2004) use oligonucleotide microarrays to detect genes involved in biodegradation and biotransformation of naphthalene, PAHs, and nitrotoluene in microbial communities. The qPCR is a molecular technique used for identification of phylogenetic and metabolic genes linked to bioremediation process by microbes and can be used as a monitoring tool for metal detoxification and hydrocarbon degradation (Navarro et al. 2013; Kostka et al. 2011; Yergeau et al. 2012). To identify and characterize the microbial communities associated with bioremediation process, metagenomic approach such as 16S rRNA clone libraries is proved to be very useful. This approach is used to identify the microbial communities involved in bioremediation of metals, hydrocarbons, and chlorinated solvents (Yergeau et al. 2012; Militon et al. 2010). In recent years, it has been found that high-throughput microarrays (PhyloChip and GeoChip) are implied to study the bioremediation process of metal and organic (Conrad et al. 2010; van Nostrand et al. 2009; He et al. 2007). The PhyloChip (16S rRNA phylogenetic microarray) is used to identify, characterize, and monitor the fluctuation in microbial community, while the GeoChip (functional gene microarray) is used to monitor catabolic gene activity during bioremediation (He et al. 2007; DeSantis et al. 2007). Presently, proteomic and metabolomic approaches are very helpful in detailed understanding of microbial cellular processes and gene products involved in catabolism of contaminants in the environment (VerBerkmoes et al. 2009).

Recent advances in genomics, transcriptomics, proteomics, and metabolomics enable the precise prediction of microbial responses toward the pollutants and their controlling factors in totality (from cell to community level). The vast amount of useful data gathered by system biology approach is helpful in designing a cost-effective, sustainable, and low labor-intensive model for an effective bioremediation of organic and inorganic pollutants from soil and water.

1.5 Role of Functional Genomics in Bioremediation

Earlier most of the microbiological investigations of bioremediation were done by using treatability study in which the samples from contaminated sites were brought in the laboratory and the rates of degradation or immobilization are documented (Rogers and McClure 2003). Such studies provide an estimate of the degradation potential of the microbial community. The attempts are then made to isolate and purify the main degrader at the laboratory scale. The recovery of the isolates responsible for bioremediation process provides the opportunity to know not only their biodegradation reactions but also their physiology that may control their growth and activity rate in the contaminated site. With the advancement in technology, it was found that there is a conserved region in the bacterial genome, i.e., 16S rRNA, which could provide phylogenetic characterization of the microorganism that is involved in the bioremediation process (Watanabe and Baker 2000). Later it was seen that there are some limitations of the 16S rRNA technique, that is, the knowledge of the phylogeny of the organism associated with bioremediation does not necessarily predict the important aspect of their physiology (Pace, 1997; Achenbach and Coates 2000). With the commencement of advancement, the new technology era has arrived, i.e., the "genomic era," with the application of genome-enabled techniques to study not only pure culture but also environmental samples. Some software and database regarding the genomic analysis are summarized in Table 1.2. Nowadays, genomic tools are very useful for demystifying the biodegradation pathways using PCR, DNA hybridization, microarray analysis, isotope distribution analysis, molecular connectivity analysis, the assessment of mineralization process using metabolic footprinting analysis, and the improvement of the biodegradation process via metabolic engineering (Villas-Boas and Bruheim 2007). The main genomic approaches applied in the field of bioremediation include protein engineering, transcriptomics, whole genome sequencing, and metagenomics. Identification of whole protein and their expression level in a cell through various methods such as two-dimensional gel electrophoresis, HPLC, and mass spectrometry (MS) is known as proteomics. The membrane proteins play a crucial role in PAH biodegradation, where any mutation can affect membrane protein receptor; a variant of two-dimensional polyacrylamide gel electrophoresis (2-DE), multidimensional protein identification technology (MudPIT), is used (Paoletti et al. 2004). A recent advance in MS techniques enables the rapid identification of small peptides and proteins. Matrix-associated laser desorption-ionization time-of-flight mass spectrometry (MALDI-TOF-MS) is the commonly used method for peptide/protein identification after separation by 2-DE through peptide mass fingerprinting (Landry et al. 2000; Aitken and Learmonth 2002; Aebersold and Mann 2003). For identification of contaminants in water, liquid chromatography coupled with MS (LC-MS) has been used by Joo and Kim (2005). The recent advances in proteomics open a whole new possibility where identification as well as changes in composition of proteins during the process of bioremediation can be monitored (Vasseur et al. 1999; Wilkins et al. 2001). The comparative proteomic study is used to determine the degradation pathways, e.g., Tomas-Gallardo et al. (2006) used comparative

Table 1.2 Tools for comparative genomics

Tools/ database	Features/description	Web address/ URL	Reference (s)
Gene prediction tools			
GeneMark	Gene prediction in archaea, bacteria, and metagenomes	http://opal.biology.gatech.edu/GeneMark	Vallenet et al. (2009)
Glimmer	Microbial gene finding system	www.cbcb.umd/software/glimmer	Moriya et al. (2007)
FgenesB	Bacterial operon and gene prediction	http://linux1.softberry.com	Koski et al. (2005)
REGANOR	Gene prediction server and database	www.cebitec.uni bielefeld.de	Pellegrini et al. (1999)
Annotation pipeline			
KAAS	KEGG automatic server annotation	www.genome.jp/tools/kaas	Zhang et al. (2011)
BASys	Bacterial annotation system	http://basys.cal/basys cgi/submit.pl	Aziz et al. (2008)
RAST	Rapid annotation using subsystem technology	www.nmpdr.org/FIG/wikiiew.cgi	Romualdi et al. (2005)
Blast2GO	Annotation and sequence analysis tool	http://www.blast2go.com	
Database and resources			
NCBI	GenBank, RefSeq, TPA and PDB, databanks for storage and downloadable genomic information	http://www.ncbi.nlm.nih.gov	
KEGG	An integrated database resource, provides genomic, chemical, and systemic information	http://www.kegg.jp	
EMBL	Nucleotide sequence database	http://www.ebi.ac.uk/embl	
UniProt	Protein resource and functional information	http://www.uniprot.org	
InterProScan	Protein sequence analysis and classification	https://www.ebi.ac.uk/interpro	

proteomic study to determine the phthalate degradation pathway in the presence of phthalate and tetralin or naphthalene. The regulation of genes related to bioremediation can also be traced by using proteomic approach.

In protein engineering, the amino acid sequence of the target protein with desired activity is selected and altered. The active sites that are close to the substrate are selected for mutational analysis. Hybrid protein is then derived by site-directed

mutagenesis, site saturation mutagenesis, gene shuffling, and generation of insertions and deletion (Lutz and Patrick 2004). The three-dimensional structure of protein P450s-P50cam, P450terp, and P450bm-3 has been studied. The replacement of P450aldo residues with amino acid in P45011β hydroxylase activity concluded that the bioremediation can be enhanced by successful alteration of specific amino acid (Li and Wackett 1993). *Polaromonas* sp. strain JS666 is well studied by using transcriptomic approach for bioremediation of chlorinated solvents such as tetrachloroethylene and trichloroethylene under aerobic conditions (Jennings et al. 2009). Moreover, transcriptome profiling of this strain has also been done and led to the identification of new genes and enzymes involved in cDCE (cis-dichloroethene) which include HADs (Bpro0530 and Bpro5186), GST (Bpro0645), CMO (Bpro5565), Hlase (Bpro5566), and CO DHase (Bpro0577) (Jennings et al. 2009). This type of studies paves the way to identify the main candidate for degradation of specific compounds such as cDCE.

The study of microarray technology for the assessment of microbial communities has very much increased in recent years (Wu et al. 2006; Zhou 2003). An array systems named as PhyloChip used for community analysis in any environment allows detection of bacteria and archaeal taxa simultaneously (Hamady et al. 2010; Sagaram et al. 2009). PhyloChip analysis offers a range of advantages over conventional techniques like DGGE, SSCP, RFLP, RADP, etc. Another set of arrays named as GeoChip used for functional gene analysis that target the genes indulge in the geochemical cycling of N, C, and P; sulfate reduction, metal resistance and reduction, and contaminant degradation have also been found (Gentry et al. 2006). Nowadays, a new gene array known as functional gene arrays (FGA) having probes for genes containing enzymes for specific degradative pathways are in the limelight (Gentry et al. 2006; He et al. 2007; Wu et al. 2001). The GeoChip 2.0 is a broad range FGA now used for the study of various environmental metabolic processes (He et al. 2007). GeoChip 2.0 is successfully applied in the bioremediation of UV(I); diesel contamination; gene identification degradation of cellulose, phthalate, biphenyl, cyclohexanol, benzoate, and naphthalene; microbial N and C cycling in Antarctic sediment; metal resistance genes; stable isotope-probing experiments for microbial profiling; etc. (Gao et al. 2007; Leigh et al. 2007; Yergeau et al. 2007; van Nostrand et al. 2007; Liang et al. 2009).

With the genome-enabled techniques, we can easily identify and get into the deeper level study of the pathway and the genes involve in bioremediation. Earlier lag phase of genomics has been overcome by rapid advancement in sequencing technologies, assembling tools and efficient annotation pipeline. A few years ago, a very less number of whole genome sequence of microorganism were found in databases, but now we have witnessed an exponential increase in the whole genome sequence number in public databases. Some of the whole genome-sequenced bacteria that take part in bioremediation are listed in Table 1.3. With the culture-dependent work, we can understand only the genomics of culturable microorganisms, but now with the constant demand of scientists and researchers, updated technology and tools have been developed with which we can explore the genomics and diversity of the microorganisms directly by sequencing the DNA

Table 1.3 Whole genome sequence of some microbes and their relevance in bioremediation

Microorganism	Genome website	Bioremediation relevance	References
Desulfitobacterium hafniense	http://www.tigr.org	It is involved in reductive dechlorination of phenol and chlorinated solvent	Lanthier et al. (2001)
Dechloromonas aromatica	http://www.jgi.doe.gov	Capable of the anaerobic oxidation of benzene coupled to nitrate reduction	Coates et al. (1997)
Pseudomonas putida	http://www.tigr.org	Capable of aerobically degrading an organic contaminants. Used for genetic engineering with bioremediation capability	Nelson et al. (2002)
Rhodopseudomonas palustris	http://www.jgi.doe.gov	Anaerobic metabolism of aromatic compounds and regulation of this metabolism	Gibson and Harwood (2002)
Geobacter sulfurreducens Geobacter metallireducens	http://www.tigr.org http://www.jgi.doe.gov	Anaerobic oxidation of aromatic hydrocarbon and reductive precipitation of uranium	Lovely et al. (1991)
		Bioremediation of aromatic hydrocarbon and uranium	Lovely et al. 1989
Arthrobacter sp. strain LS16	http://www.ncbi.nlm.nih.gov	Degrade phenol-derived compound isolated from agricultural soil	Hassan et al. (2016)
Bacillus sp. strain UFRGS-B20	http://www.ncbi.nlm.nih.gov	Hydrocarbon degrader	Dörr de Quadros et al. (2018)
Pseudomonas alcaliphila JABI	http://www.ncbi.nlm.nih.gov	Organic pollutant degrader	Ridl et al. (2018)
Pseudomonas aeruginosa DNI	EMBL	Polycyclic aromatic hydrocarbon degradation	He et al. (2018)
Shewanella oneidensis		Metal ion-reducing bacteria	Heidelberg et al. (2002)

extracted from the contaminated site, a technique called "metagenomics" (Voget et al. 2003; Handelsman 2004). Global gene expression using DNA microarray technology very much depends on the amount of coverage of the cellular mRNA and proteins, whereas the coverage of the whole genome represents all the genes of an organism. The sample from the contaminated site is directly used for the gene function analysis in metagenomics. The DNA from the contaminated site sample is extracted and get sequenced. To begin with, when the data is retrieved from sequencing machines, the strategy is to assemble longer "contigs" from individual sequencing "reads"; a number of interactive tools work to close gaps between contigs; and the genomic sequences (draft or finished) are then subjected to gene (ORF) prediction tools (Table 1.2), to know the genes encrypted in the DNA

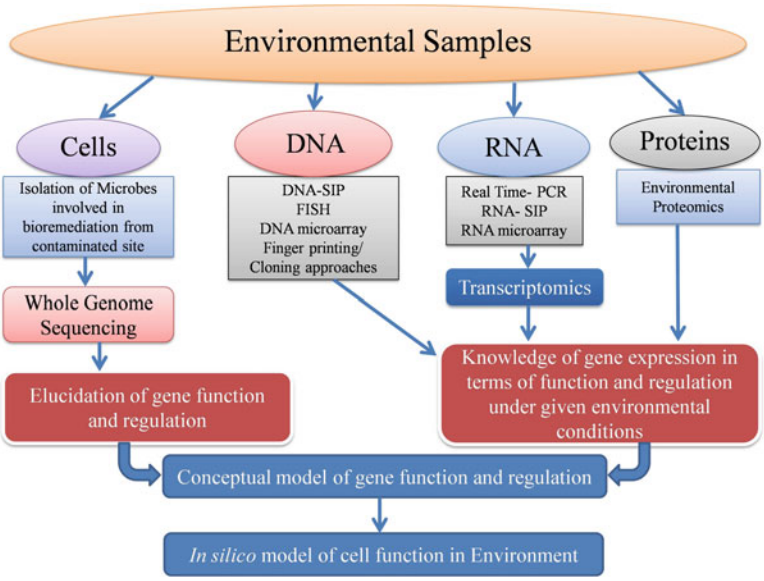

Fig. 1.2 Functional genomic approach to bioremediation**:** The whole genome sequence of the genomic DNA from microbial cells isolated from the environment provides the information of gene function and regulation. This information further used to analyze the mRNA and protein obtained directly from the environment. The environmental DNA used to determine the genetic capabilities of uncultured microorganisms, mRNA, and proteins derived from the environment provides crucial knowledge of gene expression under various environmental circumstances. The information thus obtained is used to construct the models of microbial function in the environment (adopted from Lovley 2003)

sequence. Automatic annotation pipelines are used to predict the structural properties of the putative coding sequences (CDSs) and to deduce functions of the encoded protein and RNAs (tRNA and rRNA) (Fig. 1.2).

1.6 Conclusions and Future Perspectives

Bioremediation is transforming the bioavailable contaminants or pollutants in an optimized environment to make it less toxic or other forms that is not hazardous for the environment and human beings. Various bacteria and fungi have been searched out that are playing a vital role in bioremediation of harmful chemical present in the soil and the water, but it still needs to be explored to achieve effective and reliable cleaning up of contaminants. For the microbiologists it is very challenging as most of the microorganisms in the environment are not easy to culture, so their functional biology is not fully studied. Undoubtedly, the system biology and functional genomic approach has revolutionized this field. It integrates various bioremediation approaches under one umbrella, which is very useful in understanding the complex process of bioremediation. By using this information, a better in situ and ex situ

models for bioremediation can be proposed. Genomics and metagenomics have potential applications in environmental bioremediation. For effective eradication of contaminants, diversification of microorganism at the contaminated site, identification of new enzymes and pathways, this technology seems to be very appropriate. Future concern on genomics update required in the field of bioremediation includes the development of methodologies for getting co-metabolic capability of the enzymes involved in degradation, identification of conditions that are conducive and facilitate the upregulation of degradative enzymes, and improvement of high bioavailability of mixed microbial populations.

References

Achenbach LA, Coates JD (2000) Disparity between bacterial phylogeny and physiology-comparing 16S rRNA sequences to assess relationship can be a powerful tool but its limitations need to be considered. ASM News 66:714–715

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207

Ahmad M, Khan MS (2012) Effects of pesticides on plant growth promoting traits of *Mesorhizobium* strain MRC4. Journal of Saudi Society of Agricultural Sciences 11:63–71

Aitken A, Learmonth M (2002) Protein identification by in-gel digestion and mass spectrometric analysis. Mol Biotechnol 20:95–97

Aitken MD, Stringfellow WT, Nagel RD, Kazunga C, Chen SH (1998) Characteristics of phenanthrene-degrading bacteria isolated from soils contaminated with polycyclic aromatic hydrocarbons. Can J Microbiol 44(8):743–752

Alrumman SA, Standing DB, Paton GI (2015) Effects of hydrocarbon contamination on soil microbial community and enzyme activity. J King Saud Univ Sci 27:31–41

Anderson SE, Meade BJ (2014) Potential health effects associated with dermal exposure to occupational chemicals. Environ Health Insights 8:51–62

Ang EL, Zhao H, Obbard JP (2005) Recent advances in the bioremediation of persistent organic pollutants via biomolecular engineering. Enzym Microb Technol 37:487–496

Arun A, Raja PP, Arthi R, Ananthi M, Kumar KS, Eyini M (2008) Polycyclic aromatic hydrocarbons (PAHs) biodegradation by basidiomycetes fungi, Pseudomonas isolate, and their co-cultures: comparative *in vivo* and *in silico* approach. Appl Biochem Biotechnol 151:132–142

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

Bamforth SM, Singleton I (2005) Bioremediation of polycyclic aromatic hydrocarbons: current knowledge and future directions. J Chem Technol Biotechnol 80:723–736

Bell TH, Yergeau E, Martineau C, Juck D, Whyte LG, Greer CW (2011) Identification of nitrogen-incorporating bacteria in petroleum contaminated arctic soils by using [15N] DNA-based stable isotope probing and pyrosequencing. Appl Environ Microbiol 77:4163–4171

Beller HR, Kane SR, Legler TC, Alvarez PJJ (2002) A real-time polymerase chain reaction method for monitoring anaerobic, hydrocarbon-degrading bacteria based on a catabolic gene. Environ Sci Technol 36:3977–3984

Bento FM, Camargo FAO, Okeke BC, Frankenberger WT (2005) Comparative bioremediation of soils contaminated with diesel oil by natural attenuation, biostimulation and bioaugmentation. Bioresour Technol 96:1049–1055

Boopathy R, Kulpa CF (1994) Biotransformation of 2,4,6-trinitrotoluene (TNT) by a Methanococcus sp. (strain B) isolated from a lake sediment. Can J Microbiol 40(4):273–278

Bruhlmann F, Chen W (1999) Tuning biphenyl dioxygenase for extended substrate specificity. Biotechnol Bioeng 63:544–551

Chakraborty R, Wu CH, Hazen TC (2012) Systems biology approach to bioremediation. Curr Opin Biotechnol 23:483–490

Chen C, Wang JL (2007) Characteristics of Zn^{2+} biosorption by *Saccharomyces cerevisiae*. Biomed Enviro Sci 20:478–482

Chung SY, Maeda M, Song E, Horikoshij K, Kudo T (1994) A Gram-positive polychlorinated biphenyl-degrading bacterium, *Rhodococcus erythropolis* strain TA421, isolated from a termite ecosystem. Biosci Biotechnol Biochem 58(11):2111–2113

Coates JD, Woodward J, Allen J, Phip P, Lovely DR (1997) Anaerobic degradation of polycyclic aromatic hydrocarbons and alkanes in petroleum contaminated marine harbor sediment. App Environ Microbiol 63:3589–3593

Conrad ME, Brodie EL, Radtke CW, Bill M, Delwiche ME, Lee MH, Swift DL, Colwell FS (2010) Field evidence for co-metabolism of trichloroethene stimulated by addition of electron donor to groundwater. Environ Sci Technol 44:4697–4704

de Lorenzo V (2008) Systems biology approaches to bioremediation. Curr Opin Biotechnol 19:579–589

DeSantis T, Brodie E, Moberg J, Zubieta I, Piceno Y, Andersen G (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb Ecol 53:371–383

Desjardin V, Bayard R, Huck N, Manceau A, Gourdon R (2002) Effect of microbial activity on the mobility of chromium in soils. Waste Manag 22:195–200

Desneux N, Decourtye A, Delpuech JM (2007) The sub-lethal effects of pesticides on beneficial arthropods. Annu Rev Entomol 52:81–106

Dörr de Quadros P, Fulthorpe R, Saati R, Cerqueira V, Bento FM (2018) Draft Genome Sequence of *Bacillus* sp. Strain UFRGS-B20, a Hydrocarbon Degrader. Genome Announc 6(8). pii: e00052-18

Dubus IG, Hollis JM, Brown CD (2000) Pesticides in rainfall in Europe. Environ Pollut 10:331–344

Frassinetti S, Setti L, Corti A, Farrinelli P, Montevecchi P, Vallini G (1998) Biodegradation of dibenzothiophene by a nodulating isolate of *Rhizobium meliloti*. Can J Microbiol 44(3):289–297

Gadd GM (2000) Bioremedial potential of microbial mechanisms of metal mobilization and immobilization. Curr Opin Biotechnol 11:271–279

Ganje TJ (1966) Selenium In: Chapman HD (ed.) Diagnostic criteria for plants and soils: Riverside, California University, Division of Agriculture Science, pp. 394–404

Gao H, Yang ZK, Gentry TJ, Wu L, Schadt CW, Zhou J (2007) Microarray-based analysis of microbial community RNAs by whole-community RNA amplification. Appl Environ Microbiol 73(2):563–571

Gentry TJ, Wickham GS, Schadt CW, He Z, Zhou J (2006) Microarray applications in microbial ecology research. Microb Ecol 52(2):159–175

Gibson J, Harwood CS (2002) Metabolic diversity in aromatic compound utilization by anaerobic microbes. Ann Rev Microbiol 56:345–369

Golyshin PN, Martins Dos Santos VA, Kaiser O, Ferrer M, Sabirova YS, Lünsdorf H (2003) Genome sequence completed of Alcanivorax borkumensis, a hydrocarbon- degrading bacterium that plays a global role in oil removal from marine systems. J Biotechnol 106:215–220

Guo H, Yao J, Cai M, Qian Y, Guo Y, Richnow HH, Blake RE, Doni S, Ceccanti B (2012) Effects of petroleum contamination on soil microbial numbers, metabolic activity and urease activity. Chemosphere 87:1273–1280

Guochen Z, Feng Z, Hang Y, Xue L, Jianzheng L (2011) Comparison between INT and TTC assay to determine the dehydrogenase activity of flocs. Water ResourEnviron Protect (ISWREP):1690–1693

Hamady M, Lozupone C, Knight RV (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J 4(1):17

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68(4):669–685

Hassan I, Eastman AW, Weselowski B, Mohamedelhassan E, Yanful EK, Yuan ZC (2016) Complete genome sequence of arthrobacter sp. Strain LS16, isolated from agricultural soils with potential for applications in bioremediation and bioproducts. Genome Announc 4 (1):1586–1615

Hazen T, Chakraborty R, Fleming J, Gregory I, Bowman J, Jimenez L, Zhang D, Pfiffner S, Brockman F, Sayler G (2009) Use of gene probes to assess the impact and effectiveness of aerobic in situ bioremediation of TCE. Arch Microbiol 191:221–232

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P et al (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J 1:67–77

He C, Li Y, Huang C, Chen F, Ma Y (2018) Genome sequence and metabolic analysis of a fluoranthene-degrading strain *Pseudomonas aeruginosa* DN1. Front Microbiol 9:2595

Heidelberg JF, Paulsen IT, Nelson KE, Gaidos EJ, Nelson WC, Read TD, Eisen JA, Seshadri R, Ward N, Methe B, Clayton RA (2002) Genome sequence of the dissimilatory metal ion–reducing bacterium *Shewanella oneidensis*. Nat Biotechnol 20(11):1118

Hussaini S, Shaker M, Asef M (2013) Isolation of Bacterial for Degradation of selected pesticides. Bull Environ Pharmacol Life Sci 2:50–53

Ijah UJJ (1998) Studies on relative capabilities of bacterial and yeast isolates from tropical soil in degrading crude oil. Waste Manag 18:293–299

Infante JC, De Arco RD, Angulo ME (2014) Removal of lead, mercury and nickel using the yeast *Saccharomyces cerevisiae*. Revista MVZ Córdoba 19:4141–4149

Jennings LK, Chartrand MM, Lacrampe-Couloume G, Lollar BS, Spain JC, Gossett JM (2009) Proteomic and transcriptomic analyses reveal genes upregulated by cis-dichloroethene in Polaromonas sp. strain JS666. Appl Environ Microbiol 75(11):3733–3744

Ji G, Silver S (1995) Bacterial resistance mechanisms for heavy metals of environmental concern. J Ind Microbiol 14:61–75

Jogdand SN (1995) Environ Biotechnol, 1st edn. Himalaya Publishing House, Bombay, pp 104–120

Joo WA, Kim CW (2005) Proteomics of Halophilic archaea. J Chromatogr B Analyt Technol Biomed Life Sci 815:237–250

Koski LB, Gray MW, Lang BF, Burger G (2005) AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics 6:151

Kostka JE, Prakash O, Overholt WA, Green SJ, Freyer G, Canion A, Delgardio J, Norton N, Hazen TC, Huettel M (2011) Hydrocarbon degrading bacteria and the bacterial community response in gulf of Mexico beach sands impacted by the deepwater horizon oil spill. Appl Environ Microbiol 77(22):7962–7974

Krutz LZ, Burke LC, Reddy KN, Zablotowicz RM, Price AJ (2009) Enhanced Atrazine degradation: evidence for reduced residual weed control and a method for identifying adapted soils and predicting herbicide persistence. Weed Sci 57:427–434

Kumar B, Gaur R, Goel G, Mishra M, Singh SK, Prakash D, Sharma CS (2012) Residues of pesticides and herbicides in soils from agriculture areas of Delhi region, India. EJEAFChe 328–338

Labana S, Pandey G, Paul D (2005) Plot and field studies on bioremediation of p-nitrophenol contaminated soil using *Arthrobacter protophormiae* RKJ100. Environ Sci Technol 39:3330–3337

Lal B, Khanna S (1996) Degradation of crude oil by Acinetobacter calcoaceticus and Alcaligenes odorans. J Appl Bacteriol 81:355–362

Landry F, Lombardo CR, Smith JW (2000) A method for application of samples to matrix-assisted laser desorption ionization time-of-flight targets that enhances peptide detection. Anal Biochem 279:1–8

Lanthier M, Villemur R, Lepin F, Bisailon JG, Beaudet R (2001) Geographic distribution of *Desulfitobacterium frappieri* PCP-1 and *Desulfitobacterium* spp. in soil from the province of Quebec. Canada FEMS Microbiol Ecol 36:185–191

Leigh MB, Pellizari VH, Uhlík O, Sutka R, Rodrigues J, Ostrom NE, Zhou J, Tiedje JM (2007) Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). ISME J 1(2):134

Letchinger M (2000) Pollution and water quality, neighbourhood water quality assessment. Project oceanography

Li S, Wackett LP (1993) Reductive dehalogenation by cytochrome P450CAM: substrate binding and catalysis. Biochemistry 32(36):9355–9361

Liang Y, Nostrand JDV, Wang J, Zhang X, Zhou J, Li G (2009) Microarray-based functional gene analysis of soil microbial communities during ozonation and biodegradation of crude oil. Chemosphere 75(2):193–199

Lopez-Errasquin E, Vazquez CV (2003) Tolerance and uptake of heavy metals by *Trichoderma atroviride* isolated from sludge. Chemosphere 50:137–143

Lopez-Mosquera ME, Barros R, Sainz MJ, Carral E, Seoane S (2005) Metal concentrations in agricultural and forestry soils in Northwest Spain: implications for disposal of organic wastes on acid soils. Soil Use Manag 21:298–305

Lovely DR, Phillips EJP, Gorby YA, landa ER (1991) Microbial reduction of uranium. Nature 350:413–416

Lovely DR, Baedecker MJ, Lonergan DJ, Cozzarelli IM, Phillips EJ, Siegel DI (1989) Oxidation of aromatic contaminants coupled to microbial iron reduction. Nature 339(6222):297

Lutz S, Patrick WM (2004) Novel methods for directed evolution of enzymes: quality, not quantity. Curr Opin Biotechnol 15(4):291–297

Mathew M, Obbard JP (2001) Optimisation of the dehydrogenase assay for measurement of indigenous microbial activity in beach sediments contaminated with petroleum. Biotechnol Lett 23:227–230

Megharaj M, Ramakrishnan B, Venkateswarlu K, Sethunathan N, Naidu R (2011) Bioremediation approaches for organic pollutants: a critical perspective, Environ Int. 37(8):1362–1375

Militon C, Boucher D, Vachelard C, Perchet G, Barra V, Troquet J, Peyretaillade E, Peyret P (2010) Bacterial community changes during bioremediation of aliphatic hydrocarbon-contaminated soil. FEMS Microbiol Ecol 74:669–681

Mirlahiji SG, Eisazadeh K (2014) Bioremediation of Uranium by *Geobacter* spp. J Res Dev 1:52–58

Mohamed AT, El Hussein AA, El Siddig MA, Osman AG (2011) Degradation of oxyfl uorfen herbicide by Soil microorganisms: Biodegradation of herbicides. Biotechnology 10:274–279

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:182–185

Mosher JJ, Levison BS, Johnston CG (2003) A simplified dehydrogenase enzyme assay in contaminated sediment using 2-(p-iodophenyl)-3(p-nitrophenyl)-5-phenyl tetrazoliumchloride. J Microbiol Methods 53:411–415

Navarro CA, von Bernath D, Jerez CA (2013) Heavy Metal Resistance Strategies of Acidophilic Bacteria and Their Acquisition: Importance for Biomining and Bioremediation. Biol Res 46:363–371

Nelson KE, Weinel C, Paulsen IT, Dodson RJ, Hilbert H, dos Santos VM, Fouts DE, Gill SR, Pop M, Holmes M, Brinkac L (2002) Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440. Environ Microbiol 4(12):799–808

Newcombe DA, Crowley DE (1999) Bioremediation of atrazine-contaminated soil by repeated applications of atrazine-degrading bacteria. Appl Microbiol Biotechnol 51(6):877–882

Nie M, Zhang X, Wang J, Jiang L, Yang J, Quan Z, Cui X, Fang C, Li B (2009) Rhizosphere effects on soil bacterial abundance and diversity in the Yellow River Deltaic ecosystem as influenced by petroleum contamination and soil salinization. Soil Biol Biochem 41:2535–2542

Nies DH, Silver S (1995) Ion efflux systems involved in bacterial metal resistances. J Ind Microbiol 14:186–199

Niti C, Sunita S, Kamlesh K (2013) Bioremediation: An emerging Technology for remediation of pesticides. Res J Chem Environ 17:88–105

NOFA Interstate Council: 9 (2004) The natural farmer. Ecologically sound nitrogen management. Mark Schonbeck

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740

Pandey G, Paul D, Jain RK (2003) Branching of o-nitrobenzoate degradation pathway in *Arthrobacter protophormiae* RKJ100: identification of new intermediates. FEMS Microbiol Lett 229:231–236

Paoletti AC, Zybailov B, Washburn MP (2004) Principles and applications of multidimensional protein identification technology. Expert Rev Proteomics 1:275–282

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96:4285–4428

Peña-Montenegro TD, Lozano L, Dussán J (2015) Genome sequence and description of the mosquitocidal and heavy metal tolerant strain *Lysinibacillus sphaericus* CBAM5. Stand Genomic Sci 10:1–10

Pierzynski GM, Sims JT, Vance GF (1994) Soils and environmental quality. Lewis Publishers, Ann Arbor

Prescott LM, Harley JP, Klein DA (2002) Microbiology, 5th edn. McGraw-Hill, New York, p 1014

Ravi RK, Pathak B, Fulekar MH (2015) Bioremediation of Persistent Pesticides in Rice field Soil Environment Using Surface Soil Treatment Reactor. Int J Curr Microbiol App Sci 4:359–369

Rensing C, Ghosh M, Rosen B (1999) Families of soft-metal-ion-transporting ATPases. J Bacteriol 181:5891–5897

Rhee SK, Liu X, Wu L (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. Appl Environ Microbiol 70:4303–4317

Ridl J, Suman J, Fraraccio S, Hradilova M, Strejcek M, Cajthaml T, Zubrova A, Macek T, Strnad H, Uhlik O (2018) Complete genome sequence of *Pseudomonas alcaliphila* JAB1 (= DSM 26533), a versatile degrader of organic pollutants. Stand Genomic Sci 13(1):3

Robinson DA, Hockley N, Dominati E, Lebron I, Scow KM, Reynolds B, Emmett BA, Keith AM, de Jonge LW, Schjønning P, Moldrup P (2012) Natural capital, ecosystem services, and soil change: Why soil science must embrace an ecosystems approach. Vadose Zone J 11(1)

Rogers SL, McClure N (2003) In: Head M, Singleton I, Milner MG (eds) A critical review in bioremediation. Horizon Scientific Press, Wymondham, pp 27–59

Romualdi A, Siddiqui R, Glöckner G, Lehmann R, Sühnel J (2005) GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. Bioinformatics 21:3669–3671

Rosen CJ, Horgan BP (2009) Preventing pollution problems from Lawn and Garden Fertilizers

Safiyanu I, Isah AA, Abubakar US, Rita Singh M (2015) Review on comparative study on bioremediation for Oil spills using microbes. RJPBCS 6:783–790

Sagaram US, DeAngelis KM, Trivedi P, Andersen GL, Lu SE, Wang N (2009) Bacterial diversity analysis of Huanglongbing pathogen-infected citrus, using PhyloChip arrays and 16S rRNA gene clone library sequencing. Appl Environ Microb 75(6):1566–1574

Schmidt JR, Shaskus M, Estenik JF, Oesch C, Khidekel R et al (2013) Variations in the microcystin content of different fish species collected from a eutrophic lake. Toxins (Basel) 5:992–1009

Sherburne LA, Shrout JD, Alvarez PJ (2005) Hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) degradation by Acetobacterium paludosum. Biodegradation 16(6):539–547

Shi JC, Wang HZ, Xu JM, Wu JJ, Liu XM, Zhu HP, Yu CL (2007) Spatial distribution of heavy metals in soils: a case study of Changxing, China. Environ Geol 52:1–10

Siddiqui ZS, Ahmed S (2006) Combined effects of pesticide on growth and nutritive composition of soybean plants. Pak J Bot 38:721–733

Singh BK, Walker A, Morgan JA, Wright DJ (2004) Biodegradation of chlorpyrifos by enterobacter strain B-14 and its use in bioremediation of contaminated soils. Appl Environ Microbiol 70 (8):4855–4863

Singh B, Singh Y, Sekhon GS (2006) Fertilizer-N use efficiency and nitrate pollution of groundwater in developing countries. J Contam Hydrol 20:167–184
Smith MJ, Flowers TH, Duncan HJ, Alder J (2006) Effects of polycyclic aromatic hydrocarbons on germination and subsequent growth of grasses and legumes in freshly contaminated soil and soil with aged PAHs residues. Environt Pollut 141:519–525
Soleimani N, Fazli MM, Mehrasbi M, Darabian S, Mohammadi J, Ramazani A (2015) Highly cadmium tolerant fungi: their tolerance and removal potential. JEHSE 13:1–9
Stevenson FJ (1994) Humus chemistry; genesis, composition, reactions, 2nd edn. Wiley, New York
Suzuki Y, Kelly SD, Kemner KM, Banfield JF (2005) Direct microbial reduction and subsequent preservation of uranium in natural near-surface sediment. Appl Environ Microbiol 71:1790–1797
Testiati E, Parinet J, Massiani C, Laffont-Schwob I, Rabier J, Pfeifer HR, Prudent P (2013) Trace metal and metalloid contamination levels in soils and in two native plant species of a former industrial site: evaluation of the phytostabilization potential. J Hazard Mater 248:131–141
Tomas-Gallardo L, Canosa I, Santero E, Camafeita E (2006) Proteomic and transcriptional characterization of aromatic degradation pathways in *Rhodococcus* sp. strain TFB. Proteomics 6: S119–S132
United States Environmental Protection Agency (US-EPA) (1996) A citizen's guide to bioremediation- technology fact sheet. Office of Solid Waste and Emergency Response. EPA 542-F-96-007
Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, Rouy Z, Roche D, Salvignol G, Scarpelli C, Médigue C. (2009) MicroScope a platform for microbial genome annotation and comparative genomics. Database 2009
Van Nostrand JD, Khijniak TV, Gentry TJ, Novak MT, Sowder AG, Zhou JZ, Bertsch PM, Morris PJ (2007) Isolation and characterization of four Gram-positive nickel-tolerant microorganisms from contaminated sediments. Microb Ecol 53(4):670–682
Van Nostrand JD, Wu W-M, Wu L, Deng Y, Carley J, Carroll S, He Z, Gu B, Luo J, Criddle CS et al (2009) GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer. Environ Microbiol 11:2611–2626
Vane CH, Kim AW, Beriro DJ, Cave MR, Knights K, Moss-Hayes V, Nathanail PC (2014) Polycyclic aromatic hydrocarbons (PAH) and polychlorinated biphenyls (PCB) in urban soils of Greater London, UK. Appl Geochem 51:303–314
Vasseur C, Labadie J, Hebraud M (1999) Differential protein expression by *Pseudomonas fragi* submitted to various stresses. Electrophoresis 20:2204–2213
VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems Biology: functional analysis of natural microbial consortia using community proteomics. Nat Rev Microbiol 7:196–205
Villas-Boas SG, Bruheim P (2007) The potential of metabolomics tools in bioremediation studies. OMICS 11:305–313
Voget S, Leggewie C, Uesbeck A, Raasch C, Jaeger KE, Streit WR (2003) Prospecting for novel biocatalyst in soil metagenome. Appl Environ Microbiol 69:6235–6242
Watanabe K, Baker PW (2000) Environmentally relevant microorganism. J Biosci Bioeng 89:1–11
Weber J, Karczewska A (2004) Biogeochemical processes and the role of heavy metals in the soil environment. Geoderma 122:105–107
Werner W (2002) Fertilizers, 6. Environmental aspects. Ullmann's Encyclopedia of Industrial Chemistry, Wiley-VCH, Weinheim
Wiegel J, Wu Q (2000) Microbial reductive dehalogenation of polychlorinated biphenyls. FEMS Microb Ecol 32:1–15
Wilkins JC, Homer KA, Beighton D (2001) Altered protein expression of *Streptococcus oralis* cultured at low pH revealed by two-dimensional gel electrophoresis. Appl Environ Microbiol 67:3396–3405
Wu L, Thompson DK, Li G, Hurt RA, Tiedje JM, Zhou J (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl Enviro Microbiol 67(12):5780–5790

Wu L, Liu X, Schadt CW, Zhou J (2006) Microarray-based analysis of sub nanogram quantities of microbial community DNAs by using whole-community genome amplification. Appl Environ Microbiol 72(7):4931–4941

Yergeau E, Kang S, He Z, Zhou J, Kowalchuk GA (2007) Functional microarray analysis of nitrogen and carbon cycling genes across an Antarctic latitudinal transect. ISME J 1(2):163

Yergeau E, Sanschagrin S, Beaumier D, Greer CW (2012) Metagenomic Analysis of the Bioremediation of Diesel-Contaminated Canadian High Arctic Soils. PLoS One 7(1):e30058. https://doi.org/10.1371/journal.pone.0030058

Zhang C, Hughes JB (2003) Biodegradation pathways of hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine (RDX) by Clostridium acetobutylicum cell-free extract. Chemosphere 50(5):665–671

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38:95–109

Zhao B, Poh CL (2008) Insights into environmental bioremediation by microorganisms through functional genomics and proteomics. Proteomics 8:874–881

Zhou J (2003) Microarrays for bacterial detection and microbial community analysis. Curr Opin Microbiol 6(3):288–294

Zhu YG, Shaw G (2000) Soil contamination with radionuclides and potential remediation. Chemosphere 41:121–128

Bacterial Pan-Genomics

2

Arash Iranzadeh and Nicola Jane Mulder

Abstract

Due to their tendency to have a high recombination rate, bacterial genomes are highly diverse across different strains. This diversity may even be in the form of the presence or absence of entire genes; therefore, each strain might have its own combination of genes. The pan-genome represents the complete gene pool of a species. It is made up of the core genome (genes shared by all strains) and the accessory genome (genes shared by some strains and not all). The pan-genome can be considered to be a comprehensive reference genome for computational biology, and several tools have been developed for pan-genomics applications. The tools enable scientists to explore bacterial genomes with more flexibility considering all types of genetic variations. Pan-genomics has many applications in medicine such as the development of vaccines and drugs against pathogenic bacteria. In this chapter, we discuss the fundamental principles and algorithms for pan-genome analysis and introduce and compare the most recent computational tools.

2.1 Introduction

Despite the fact that not all microbes are harmful, more than 17 million people are killed from infectious diseases caused by microbes each year ('WHO | Press release' 2013). There have been several deadly bacterial pandemics through history such as the *plague*, *cholera*, and *typhus* in which millions of people perished, reshaping the ancient and medieval world populations. Fortunately, the existence of microbes was

A. Iranzadeh (✉) · N. J. Mulder
Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
e-mail: arash.iranzadeh1980@gmail.com

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_2

discovered for the first time by *Robert Hooke* and *Antoni van Leeuwenhoek* in the seventeenth century (Gest 2004). Later, in the nineteenth century, the science of bacteriology was established. *Louis Pasteur* demonstrated the germ theory of disease and the relationship between microbes and diseases (Lanska 2014), and *Robert Koch* discovered the bacterium *Bacillus anthracis* as the cause of *anthrax* (Blevins and Bronze 2010).

Bacteria are one of the most important types of organisms that cause diseases. These single-celled prokaryotes can be found almost everywhere, and their existence on earth dates back to about 3.5 billion years ago (Kara and Robert 2018). Since they can live in extreme environmental conditions on earth, they might exist on other planets or even other galaxies in the universe (Grebennikova et al. 2018).

To understand the outbreak and pathogenesis of bacterial infections, their genomes must be studied and analyzed precisely. The first step is to sequence the genome, which contains information about the origin of the species and its evolution. Two major sequencing techniques have been developed: early sequencing techniques or *first-generation sequencing* developed in the 1970s, which include the *Maxam-Gilbert* and *Sanger* methods (Sanger et al. 1977), and modern sequencing or *next-generation sequencing* (*NGS*) technologies that have been developed in the twenty-first century. *NGS* is also called *deep sequencing*, *high-throughput sequencing*, or *massive sequencing*. Examples include *Illumina (Solexa)*, *Roche 454*, *and SOLiD* (Goodwin et al. 2016). *Sanger* sequencing can cover a long stretch of DNA with higher quality and is sometimes used to sequence small pieces of DNA such as bacterial plasmids or for validation. However, it is more expensive and time-consuming than *NGS* and is inefficient for sequencing entire genomes. For instance, *Sanger sequencing* took over a decade to deliver a draft human genome, while *NGS* takes only a single day to sequence an entire human genome (Behjati and Tarpey 2013). The whole genome of a bacterium can be sequenced by *NGS* technology for a few hundreds of dollars (Mengoni et al. 2015). As a result, today, an enormous amount of genetic sequences are publicly available in databases such as the *National Center for Biotechnology Information's (NCBI) GenBank* (Ostell and McEntyre 2007), the *DNA Databank of Japan (DDBJ)* (Miyazaki et al. 2004), and the *European Nucleotide Archive (ENA)* (Leinonen et al. 2011), among others.

Owing to the availability of large amounts of sequencing data, a bacterial species can be described through an inclusive reference genome called the *pan-genome.* Computational algorithms and tools have been developed for *pan-genome* building and analysis. The pan-genome is a kind of reference genome that represents all genes in a collection of bacterial isolates. The pan-genome is usually defined for bacteria and viruses because they are highly recombinogenic and have a small genome and many isolates can be cloned and sequenced quickly. Nonetheless, the pan-genome approach can also be applied for eukaryotes like plants to investigate similarities and differences in individuals of the same species. In this chapter, the bacterial pan-genome definition, algorithms, and the computational tools available for pan-genomics will be discussed.

2.2 Pan-Genome

For most studies in comparative genomics starting with NGS data, a reference genome is required, and it must be defined for data analysis. This reference genome can be:

1. The genome of one strain.
2. The consensus sequence drawn from all strains.
3. A comprehensive genome that contains all genetic variants.

The remarkable capability of bacteria to adapt to their environment is enabled by their ability to exchange their genetic material by homologous recombination and horizontal gene transfer. It allows bacteria to have a dynamic, adaptable, and diverse genome (Maloy 2013). This genomic plasticity is even considerable across different strains of the same bacterial species. Therefore, a single genome sequence cannot necessarily represent the entire range of genetic variation in bacteria. A pan-genome is actually a type of reference genome that displays all variants, including all possible genes. In 2005 and for the first time, the term, *comparative pan-genomics*, became official when eight strains of *Streptococcus agalactiae* were compared (Tettelin et al. 2005). Since then, the pan-genome has been defined as the following: "For a collection of closely related strains, pan-genome is the entire gene set that exists in those strains."

The pan-genome contains three types of genes according to their availability among strains (Fig. 2.1):

1. *Core genes* that exist in all strains.
2. *Accessory genes* (*dispensable genes*, *variable genes*, or *adaptive genes*) that are present in some strains but not all.
3. *Unique genes* (*specific genes*) that are a particular form of accessory genes that are present only in one strain.

The collection of core genes is called the *core-genome*, and the collection of accessory genes is called the *accessory-genome*. Therefore the pan-genome = core-genome + accessory-genome. The total gene number in the pan-genome is:

Total genes = core genes + accessory genes
Pan-genome = core-genome + accessory-genome

The genes in the core-genome are often the signals of identity and make a species what it is. Core genes, also called *the minimal gene set*, are essential for normal cell functions such as DNA replication, transcription, and translation and are universally conserved. The genes in the accessory-genome are not necessary for basic life, at least for all conditions that bacteria encounter. The existence of these genes causes some strains to gain specific traits such as virulence and antibiotic resistance or the

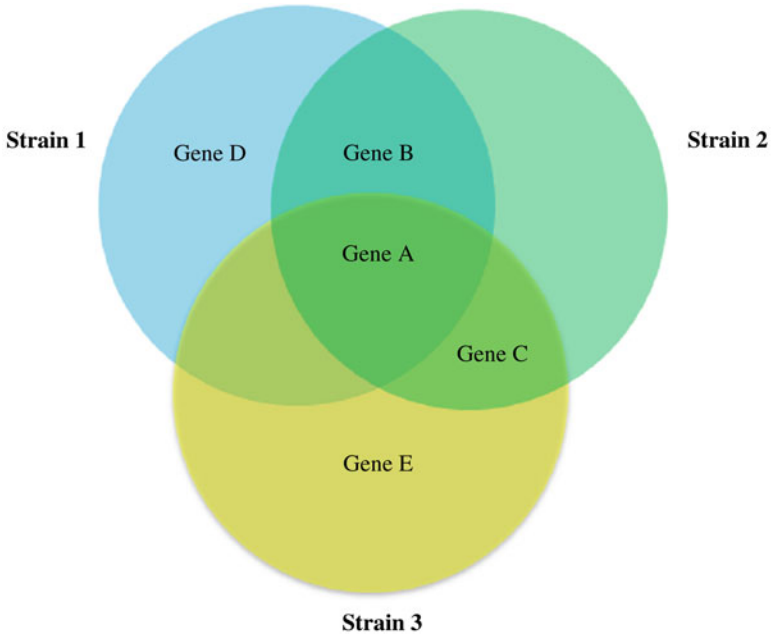

Fig. 2.1 Pan-genome of three strains. Gene A is a core gene as it exists in all strains. Gene B and C are accessory genes because they exist in two strains and not all. Gene D and E are unique genes specific to strain 1 and 3, respectively. This pan-genome has five genes A, B, C, D, and E in total

ability to occupy niche environments. The total number of genes in the pan-genome is usually larger than the number of genes in one single strain.

In some species, after adding a certain number of genomes from different strains, the total number of genes in the pan-genome does not increase further. This pan-genome that reaches a plateau is called a *closed* pan-genome. Meanwhile, for other species, every new strain adds new genes to the pan-genome. Such species have an *open* pan-genome that does not reach a plateau (Fig. 2.2). Species that are dormant and live in an isolated environment often have a closed pan-genome, whereas metabolically active species that have a diverse genome and horizontally transfer genes have an open pan-genome (Rouli et al. 2015).

It is worth noting that, although the pan-genome usually has a gene-based definition which means it refers to the entire gene set existing in different strains of one species, it can also have a sequence-based characterization which means it refers to all sequences found in different strains of one species. The gene-based definition considers variations at gene levels such as gene presence/absence and gene copy number variations (CNVs), while the sequence-based description is complete and considers all small-scale variants such as single nucleotide polymorphisms (SNPs), insertion/deletions (Indels), and structural variants (SVs) in coding and noncoding sequences. Nonetheless, the fundamental reason for the pan-genome definition is that a single individual genome is unable to show all genetic variants

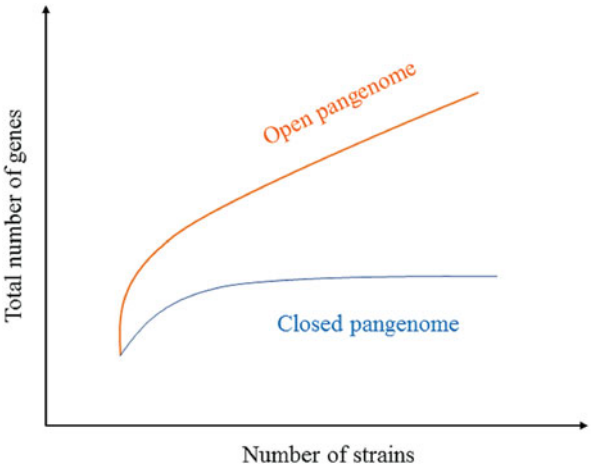

Fig. 2.2 In a closed pan-genome, the number of total genes will not increase after adding a certain number of strains. In open pan-genome, any new strain adds new genes to the pan-genome

found in the species. Therefore, a pan-genome is a hypothetical combination of variants that do not exist in reality. Thus, a more comprehensive pan-genome definition is:

Pan-genome is the entire set of all DNA sequences including genes and noncoding regions found in individuals of the species.

2.3 Computational Pan-Genomics

The discipline of computational pan-genomics refers to all computational principles that are applied for pan-genome visualization, statistical analysis, and software development.

2.3.1 Pan-Genome Graphical Representation

The main idea in pan-genomics is to replace traditional consensus and linear reference genomes by a pan-genome structure that captures all variants in one species. The two main methods to represent the pan-genome structure are:

1. A multiple sequence alignment (MSA)
2. A graph data structure

The structure of the pan-genome in Fig. 2.1 is visualized in Fig. 2.3 by an MSA and a graph.

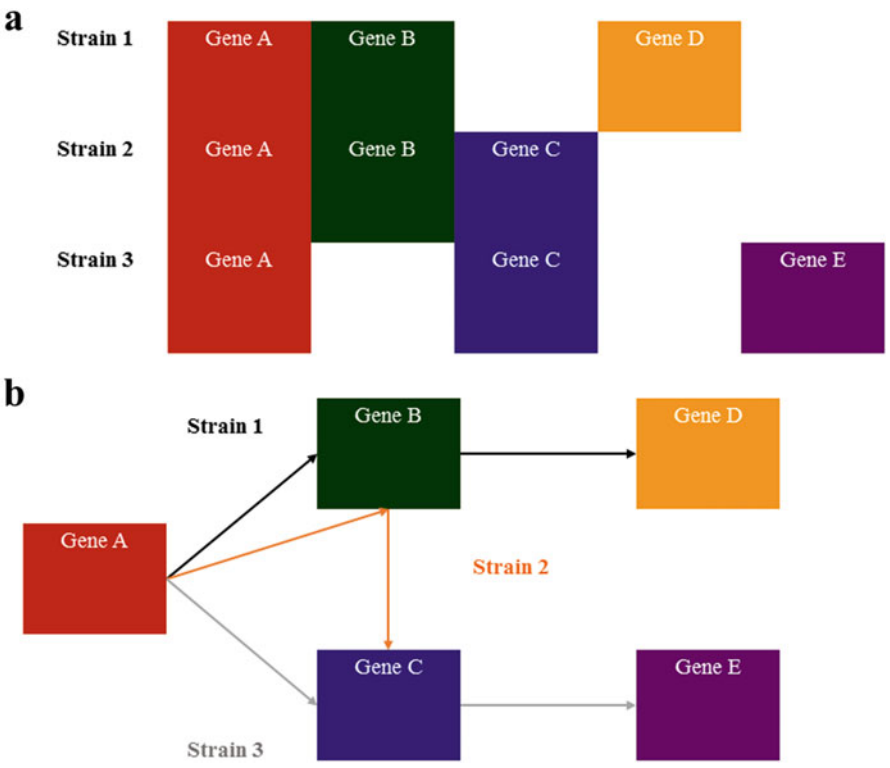

Fig. 2.3 (**a**) The pan-genome in Fig. 2.1 is shown by an MSA and (**b**) a directed graph. Each rectangle is a node that represents a gene, and each arrow is a directed edge that joins two nodes. Each path, which is a combination of edges on the graph, indicates the set of genes that are present in one strain; the black path shows genes found in strain 1, the orange path for strain 2, and the gray path for strain 3

2.3.1.1 Pan-Genome as an MSA

Representing a pan-genome as an MSA generates a large and sophisticated structure, the analysis of which is complicated. The reason is that each gene must be represented as many times as the number of isolates it exists in. For example, in Fig. 2.3a, Gene A is present in three strains, and Gene B is present in two strains. Therefore, Gene A appears three times and Gene B twice. Moreover, although the MSA can spot SNPs and Indels, it is not able to identify gene duplications and chromosomal structural variants. For almost all pan-genome analyses, researchers prefer graph data structures over an MSA.

2.3.1.2 Pan-Genome as a Graph

In molecular biology, the terms *graph* and *network* are used interchangeably (Huber et al. 2007). Before explaining how to represent a pan-genome as a graph data structure, it is helpful to introduce the basic definitions in the *graph theory* briefly.

The graph *G(V, E)* is a set of nodes or vertices (*V*) that are joined by a set of edges (*E*) (Fig. 2.3b). An edge that joins two nodes *u* and *v* is an *incident* on them and is denoted by *(u,v)*. Two nodes that are joined by an edge are called *adjacent nodes*, and two edges that are joined by a node are called *adjacent edges*. The edge *(u,u)* that joins the node *u* to itself is called a *loop*. The edge (*u*,*v*) that joins two different nodes u and v is named a *proper edge*. Multiple edges that join the same two nodes are called *multi-edges*. A complete graph is a graph where every pair of nodes is joined by an edge. A *directed edge* is an edge that has a specified direction and joins a start node called *tail node* to an end node called *head node*. The head node is the *successor* of the tail node, and the tail node is the *predecessor* of the head node. If all edges in a graph have direction, the graph will be called a *directed graph* or *digraph*. A *walk* is a way of getting from one node to another through a sequence of edges. A *path* is a walk in which every vertex appears only once. The length of the shortest path between two vertices is called the *distance* between them. A walk visiting every edge exactly once is called the *Eulerian walk* (Wilson 2006).

To design a graph that is able to represent all genetic variations in the pan-genome, the genome sequences of different strains are split into their substrings called *k-mers* which are subsequently arranged in a graph. A *k-mer* is a substring of certain length *k* (*k* is a natural number) that can be obtained from the sequence *S* of length *n* while $1 <= k < n$. The substring from position *i* to *j* is shown as *S*[*i*:*j*]. Therefore, any *k-mer* of sequence *S* is defined as below:

$$k\text{-}mers = S[i : i + k - 1] \quad \text{(Inclusive)} \quad (1 <= i <= n - k + 1) \& (1 <= k <= n).$$

The total number of *k-mers* will be $(n - k + 1)$.

Example

Sequence: S = "CGCTGAGCT"
Example of substrings: S[1:4] = "CGCT", S[2:4] = "GCT", S[5:5] = "G"
Sequence length: n = 9
K-mer length: k = 3
Kmers = S[i:i + k-1] = S[i:i + 2] (1 < = i < = 7)
A total number of k-mers of length 3 (3-mers) obtained from sequence S of length 9:
N – K + 1 = 9–3 + 1 = 7
All possible 3-mers obtained from S, note that there is a repeat of "GCT":
3-mers = ["CGC", "GCT", "CTG", "TGA", "GAG", "AGC", "GCT"]

The *k* value is critical here and must be selected carefully. It depends on the genome length, the available computing resources, and the type of analysis. To count all *k-mers* in a sequence, many space-efficient algorithms have been developed. Examples are *disk streaming of k-mers* (Rizk et al. 2013), *k-mer counter* (Kokot et al. 2017), and *Squeakr* (Pandey et al. 2018).

All *k-mers* derived from sequence *S* are arranged in a directed graph called a *de Bruijn graph* (DBG) denoted by *G(S,k)*. This graph contains a node for each distinct k-mer of *S*, and a directed edge *(u,v)* connects two nodes *u* and *v* if

$$u = S[i : i + k - 1] \text{ and } v = S[i + 1 : i + k]$$

The gene content of each strain is illustrated by an Eulerian walk on the graph.

To save memory and space, the DBG is compressed by merging its nodes to produce a *compressed DBG*. Two nodes *u* and *v* are allowed to be merged into a single node if

Node u is the only predecessor of node v, and node v is the only successor of node u. There may be multiple edges between them.

In a compressed DBG, every node (except the start node) has at least two different predecessors, or its single predecessor has at least two different successors, and every node (except the end node) has at least two different successors, or its only successor has at least two different predecessors (Beller and Ohlebusch 2016). A compressed DBG can be constructed by identifying maximal exact matches using a suffix tree (Marcus et al. 2014) or more efficiently by a combination of FM index, compressed suffix tree, and Burrows-Wheeler transform (Baier et al. 2015).

The k-mer-based representation of the pan-genome in a graph has many advantages such as simplicity, speed, and robustness. It is not always necessary to use fixed-length *k-mers* as the pan-genome can be arranged in acyclic and cyclic graphs (Marschall et al. 2016). The pan-genome graph is able to highlight all genetic diversities found in a species from SNPs and Indels to gene presence and absence. It renders a compact graphical portrait of the pan-genome that characterizes the variants among individuals. Moreover, graph-based pan-genomics provides access for retrieving data and defining a suitable coordinate system. It highlights the variable and conserved regions across the genomes. All genes are represented only once on the graph (no matter in how many strains they are present), and each strain is characterized by an exclusive walk on the graph (Fig. 2.4).

2.3.2 Pan-Genome Computational Analysis

High-performance and parallel computing are necessary for pan-genome computational analysis, particularly when a high number of strains are involved in the investigation. The computational pipeline often needs significant RAM and storage space.

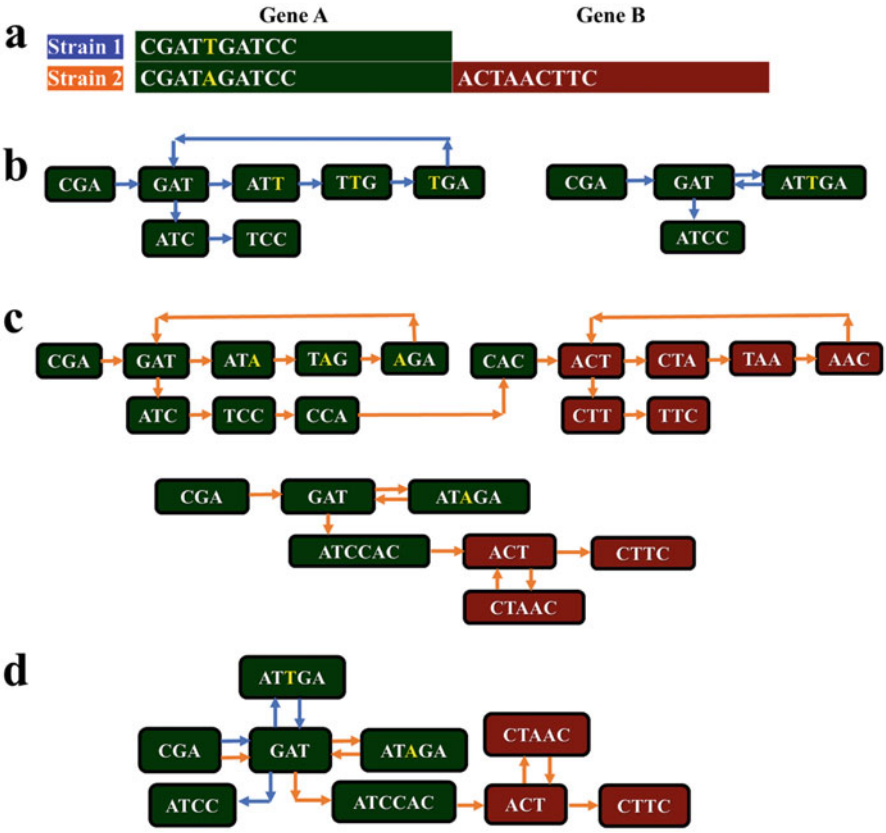

Fig. 2.4 (**a**) Genomes of strain 1 and strain 2; Gene A shown in green is a core gene, and gene B shown in red is an accessory gene; there is an SNP on position 5 of Gene A highlighted in yellow. (**b**) The genome of strain 1 is demonstrated by its 3-mers in the form of a DBG on the left and a compressed DBG on the right. (**c**) The genome of strain 2 is shown by its 3-mers in the form of a DBG on the top and a compressed DBG on the bottom. **(d)** The pan-genome of both strains is shown as a compressed DBG, the blue Eulerian walk indicates the genome of strain 1, and the orange Eulerian walk indicates the genome of strain 2; both variants in the form of SNP and gene presence/absence are identifiable on the graph; the yellow letter indicates the SNP, the green nodes show gene A, and the red nodes show gene B

A bacterial pan-genome analysis starts with a set of *whole genome sequencing (WGS)* short reads obtained from several closely related strains, preferably from the same species. The pipeline for a pan-genome analysis has four main steps (Fig. 2.5):

1. Reads quality control, preprocessing, and cleaning
2. Genome assembly and annotation
3. Pan-genome construction
4. Pan-genome downstream analysis

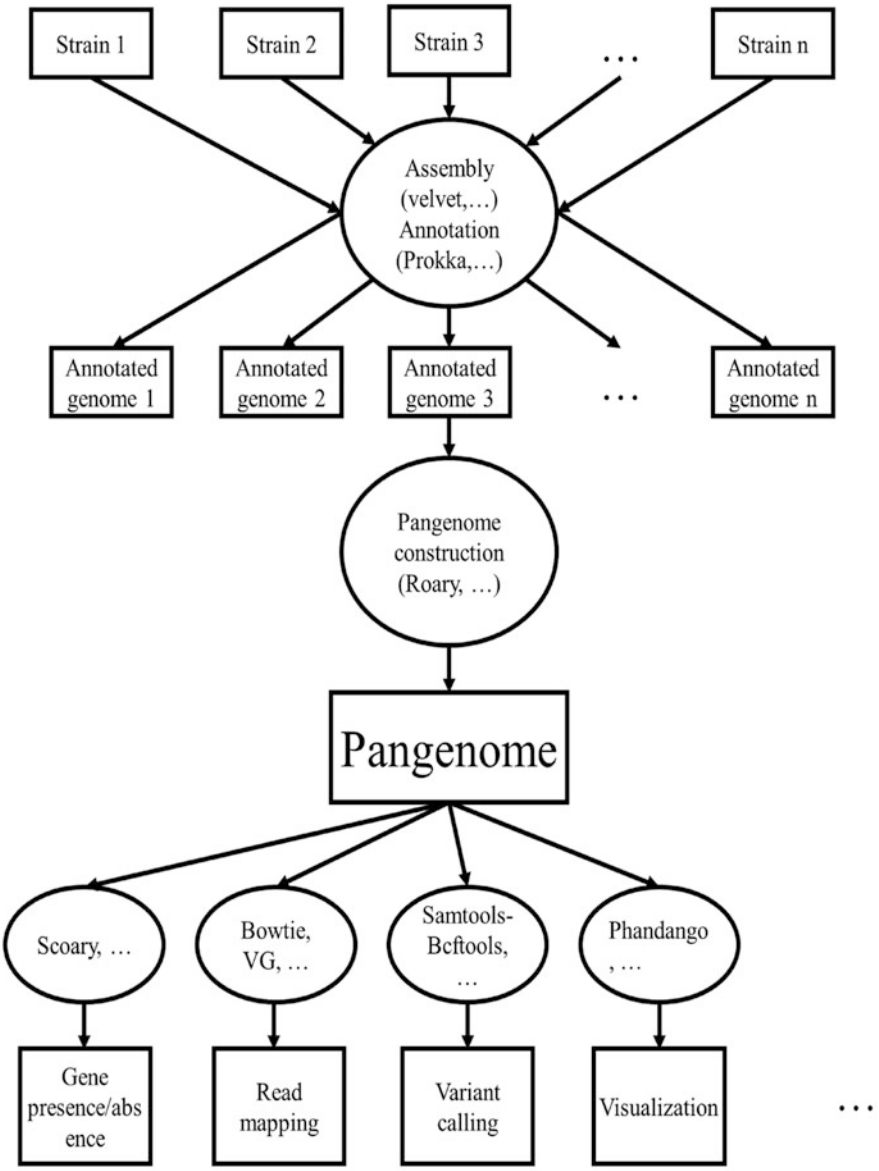

Fig. 2.5 Overview of pan-genomic analysis

2.3.2.1 Reads Quality Control, Pre-processing and Cleaning

The sequencing short reads are stored in standard file formats like *fastq* (Cock et al. 2009). The quality of the reads is evaluated by tools such as *fastqc* (Andrews 2010), the adapter sequences are trimmed, and sequences with low

quality are removed by tools like *FASTX toolkit* (Gordon and Hannon 2010). The clean sequences that are of high quality are supplied to the genome assemblers.

2.3.2.2 Genome Assembly and Annotation

The next step is to assemble the genomes of all strains. Typically, a pan-genome analysis is useful when working with a bacterial species whose genome is highly divergent across different strains. Defining a linear reference genome for such a diverse species is difficult. Thus, de novo *assembly* (Paszkiewicz and Studholme 2010), which is reference-free, is desired in computational pan-genomics. The genome assembly can be achieved by some publicly available tools such as *VelvetOptimiser* (Gladman and Seemann 2008) and *SOAPdenovo* (Luo et al. 2015). For a successful pan-genome analysis, the assembled genome should be of high quality, and contigs should have a minimum length of 500 base pairs. Tools like *Quast* (Gurevich et al. 2013) can be used to evaluate the quality of the assembled genomes.

To define the pan-genome and determine its core and accessory genes, all assembled genomes must be annotated coherently with a tool that is compatible with the pan-genome builder. Accurate assembly and annotation produce a pan-genome with high quality enabling a productive analysis. Annotated genomes are saved in standard file formats, such as *BED*, *GFT*, *GFF*, *and GFF3*. Many tools have been developed for bacterial gene prediction and annotation. Examples include *Glimmer* (Delcher et al. 2007) and *Prokka* (Seemann 2014). *Prokka* is specifically designed for prokaryotic genome annotation and works based on the integration of several tools and databases such as *SignalP* (Petersen et al. 2011), *Aragorn* (Laslett and Canback 2004), *HMMER3* (Finn et al. Finn et al. 2011), *Rfam* (Nawrocki et al. 2015), and *Infernal* (Nawrocki and Eddy 2013). *Prokka* is a well-run annotator that produces its outputs in various file formats, at least one of which will be compatible with one of the tools used for pan-genome construction. For annotation from scratch, *Pannotator* (Santos et al. 2013) is suitable, and to improve the available annotations, *eCAMBer* (Wozniak et al. 2014) and *Mugsy-Annotator* (Angiuoli et al. 2011) can be applied.

2.3.2.3 Pan-Genome Construction

As discussed earlier, the pan-genome can be defined either as a collection of genes or genome sequences from multiple strains of one species. For this reason, two types of tools have been developed for pan-genome construction and analysis (Zekic et al. 2018):

1. Gene-based tools
2. Sequence-based tools

To use the gene-based tools, all genomes must be annotated, and the gene content of each strain must be determined. These tools first use graph-based methods to assign orthologous genes found in strains and then construct the pan-genome. Some of the most popular gene-based tools developed so far are *EDGAR* (Blom et al. 2016),

PGAT (Brittnacher et al. 2011), *PGAP* (Zhao et al. 2012), *PanOCT* (Inman et al. 2018), *GET_HOMOLOGOUS* (Contreras-Moreira and Vinuesa 2013), *PanFunPro* (Lukjancenko et al. 2013), *ITEP* (Benedict et al. 2014), *PanGP* (Zhao et al. 2014), *LS-BSR* (Sahl et al. 2014), *Roary* (Page et al. 2015), *Micropan* (Snipen and Liland 2015), *Piggy* (Thorpe et al. 2018), *BPGA*, and *Pyseer* (Lees et al. 2018).

For a sequence-based pan-genome analysis, the sequences of different genomes are indexed. To increase efficiency regarding required time and memory, graph-based methods are applied. DBG is usually employed here as the analysis does not require a reference sequence or alignment. Examples of sequence-based tools are *Panseq* (Laing et al. 2010), *Harvest* (Treangen et al. 2014), *SplitMEM* (Marcus et al. 2014), *TwoPaCo* (Minkin et al. 2017), and *Bloom Filter Trie* (Holley et al. 2016).

2.3.2.4 Pan-Genome Downstream Analysis

Most of the tools mentioned above can perform some downstream analysis. The downstream analysis includes tasks such as multiple sequence alignment of the core-genomes, phylogenetic tree construction, alignment of the short reads to the pan-genome, variant calling, studying of genes in different metabolic pathways, pan-genome visualization, and various statistical analyses.

The multiple sequence alignment of core-genomes is sometimes produced by the pan-genome builder. This alignment is then used to extract the variant sites in the core genes which are used to draw an initial phylogenetic tree. This tree can be colored according to the sample phenotypes and provides an overview of the association between samples from different phenotypes. *Snp-sites* (Keane et al. 2016) is a tool which is useful to extract all variant sites from the multiple sequence alignment of all the core genes. The phylogenetic tree can be drawn by considering only those variant sites in the core-genome of the different strains. However, if sufficient computing resources are available, the phylogenetic tree can be drawn directly from the alignment of the core-genomes. Tools like *ClustalW* (Larkin et al. 2007) and *FastTree* (Price et al. 2010) are appropriate for this tree drawing. The tree coloring and visualization can be performed with the help of tools such as *Evolview* (He et al. 2016).

The software *Scoary* (Brynildsrud et al. 2016) was developed to score genes in the pan-genome for their association with an observed trait. This tool finds genes whose presence or absence are strongly associated with a phenotype and considers the influence of the population stratification. *Piggy* (Thorpe et al. 2018), on the other hand, is a tool that examines the variation in intergenic regions in bacteria. Apart from genes, the presence or absence of some intergenic regions affects the phenotypic behavior of the bacterium. Both *Scoary* and *Piggy* can use the output of *Roary* as their input.

To perform a pan-genome-based GWAS, the sequence of the genes in the pan-genome can be utilized as a reference to identify SNPs and Indels in each strain and determine whether they are in the core genes or accessory genes. In this case, the pan-genome is usually saved in a fasta file in which each record represents the consensus sequence of a gene drawn from the entire population. Then the short reads are aligned to this reference sequence and variants are called. The SNPs in the core

genes reflect the age of the species. To reduce the analysis workload, specific informative SNPs should be selected in a process called representative SNP selection (Hurgobin and Edwards 2017). As explained earlier, the pan-genome can be saved as a graph. Several tools have been developed to align short reads directly to the pan-genome graph. Examples are *BGREAT* (Limasset et al. 2016) and VG.[1] The pan-genome graph can also be used for reference-free variant calling (Iqbal et al. 2012). For further details about the tools, their algorithms, and performance, refer to (Xiao et al. 2015), (Vernikos et al. 2015), and (Zekic et al. 2018). Many scripts written in *R* and *Python* are available for pan-genome visualization; some of the more versatile tools for this are *Phandango* (Hadfield et al. 2018), *Panx* (Ding et al. 2017), and *Panviz* (Pedersen et al. 2017).

2.4 Pan-Genomics Research Examples

As mentioned earlier, the importance of pan-genomics in medicine and microbiology was first considered in 2005, when (Tettelin et al. 2005) utilized the term pan-genome for the first time and investigated six strains of *Streptococcus agalactiae*, which is the primary cause of neonatal infection in human. The research concluded that multiple strains must be sequenced to study a bacterial species because many of them have an open pan-genome. In some cases, hundreds or even thousands of strains must be sequenced. It is essential to consider their large accessory-genomes to identify potential candidates for the design of effective drugs or universal vaccines. To date, many studies have been conducted based on pan-genomics algorithms and principles.

As an example, (Rasko et al. 2008) carried out a pan-genome analysis on 17 samples of *E. coli*. Their dataset was composed of commensal, extraintestinal pathogenic, and diarrheagenic samples. They identified an open pan-genome which was made up of more than 13,000 genes including 2200 core genes. They found isolate-specific genes that led them to assume that each isolate can develop its virulence independently. They suggested that extraintestinal pathogenic samples share a significant level of similarity; however, in general, this study showed that the *E. coli* pathovars are not distinct on the molecular level (Rasko et al. 2008).

Pan-genomics can also be applied to investigate the genome of one bacterial species in association with other species. For example, (Donati et al. 2010) analyzed the genome of 44 strains of pathogenic *Streptococcus pneumoniae* and compared them with strains of nonpathogenic *Streptococcus mitis*. According to their results, *Streptococcus pneumoniae* has an open pan-genome that enables the bacterium to respond to the different environments. They determined that homologous recombination is the primary evolutionary process used by *Streptococcus pneumoniae*. The genetic materials can be exchanged within

[1] https://github.com/vgteam/vg

the species or with other species like *Streptococcus mitis*. They correlated the age of clones with the number of acquired genes (Donati et al. 2010).

In another study carried out by (D'Auria et al. 2010), they built the pan-genome of five strains of *Legionella pneumophila*. They compared the gene content of a persistent strain from Spain to the genome of four other strains from other countries such as England, France, and the United States. Out of their constructed pan-genome, they identified 53 genes specific to the pathogenic strain from Spain. The research demonstrated that the accessory-genome contains new traits that can be exchanged through horizontal gene transfer and the virulence of the bacterium is promoted by part of its core-genome. In this study, pan-genomics was applied to compare samples from different geographical locations (D'Auria et al. 2010).

To improve the accuracy of a pan-genome study, especially for species that have a very diverse genome, hundreds or even thousands of samples must be included in the study. This type of data analysis requires excessive computing power and time. This issue has been resolved with the availability of *high-performance computing (HPC)* algorithms and clusters. A pan-genome study conducted by (Azarian et al. 2018) used 937 *Streptococcus pneumoniae* samples to investigate the evolutionary impact of vaccination. They concluded that the introduction of the *pneumococcal conjugate vaccine (PCV)* reduced the pan-genome size and the genetic diversity and changed frequencies of genes. However, the genetic diversity expanded again through in-migration of non-vaccine lineages, and frequencies of genes returned to the original value by selection (Azarian et al. 2018). In this study, pan-genomics was applied to compare samples from different time points.

In addition to bacteria, pan-genomics can be applied for eukaryotes and is particularly useful for plants. A study by (Gordon et al. 2017) explored 54 lineages of the grass *Brachypodium distachyon*. The number of genes in the pan-genome was twice the number of genes existing in an individual genome. As expected, they demonstrated that the core genes are essential for basic biological functions, while accessory genes are required for beneficial functions such as defense and development. They concluded that genes in the accessory-genome have critical roles in the phenotype of different individuals and transposable elements are significantly involved in the pan-genome evolution.

2.5 Conclusion

Pan-genomics has revolutionized our understanding of microbiology. Thanks to advances in sequencing technologies, researchers can access hundreds or even thousands of sequenced genomes from the same species. Different strains of a bacterial species often have different gene contents, and each of them has its combination of genes. The pan-genome has been developed to address this genomic diversification in bacteria. For some bacterial species, the diversity is high, and the ratio of *core-genome/pan-genome* is low. For example, only 3.7% of the pan-genome obtained from 2000 *Escherichia coli (E.coli)* strains (Land et al.

2015) and 3.5% of the pan-genome obtained from 53 strains of *Campylobacter concisus* (Gemmell et al. 2018) are core and shared by all strains.

On the other hand, in higher eukaryotic organisms, the gene contents of different individuals are almost the same, and their genomic diversity is mainly in the form of SNPs, Indels, SVs, and CNVs, which affect gene regulation or function instead of the gene content. For instance, the genetic diversity across human genomes is about 0.6% (Auton et al. 2015), and they share more than 99% of their gene repertoire. If we define a eukaryotic species as a group of individuals that share more than 99% of their genomes, we should redefine it for bacteria as many of them do not meet this threshold.

The considerable variation in gene content between bacterial strains is attributed to their capability to exchange their genetic material through transformation, transduction, and conjugation. They can use this ability to adapt themselves to their environment, survive in extreme conditions, become pathogenic, and acquire resistance to antibiotics. From the medical perspective, the pan-genome is an ideal reference genome to highlight all variants among bacterial strains. It helps researchers to differentiate pathogenic isolates in a large population and understand their pathology. The pan-genome can be applied as a reference to find genes or sequences that are significantly liked to the particular phenotypes such as invasiveness or antibiotic resistance and is thus crucial for facilitating bacterial disease control, vaccine development, and drug design.

References

Andrews S (2010) FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angiuoli SV et al (2011) Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinform. https://doi.org/10.1186/1471-2105-12-272

Auton A et al (2015) A global reference for human genetic variation. Nature. https://doi.org/10.1038/nature15393

Azarian T et al (2018) The impact of serotype-specific vaccination on phylodynamic parameters of Streptococcus pneumoniae and the pneumococcal pan-genome. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1006966

Baier U, Beller T, Ohlebusch E (2015) Graphical pan-genome analysis with compressed suffix trees and the burrows-wheeler transform. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv603

Behjati S, Tarpey PS (2013) What is next generation sequencing? Arch Dis Child Educ Pract Ed 98 (6):236–238. https://doi.org/10.1136/archdischild-2013-304340

Beller T, Ohlebusch E (2016) A representation of a compressed de Bruijn graph for pan-genome analysis that enables search. Algorithms Mol Biol. https://doi.org/10.1186/s13015-016-0083-7

Benedict MN et al (2014) ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC Genomics. https://doi.org/10.1186/1471-2164-15-8

Blevins SM, Bronze MS (2010) Robert Koch and the "golden age" of bacteriology. Int J Infect Dis. https://doi.org/10.1016/j.ijid.2009.12.003

Blom J et al (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw255

Brittnacher MJ et al (2011) PGAT: a multistrain analysis resource for microbial genomes. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr418

Brynildsrud O et al (2016) Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 17(1):238. https://doi.org/10.1186/s13059-016-1108-8
Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. https://doi.org/10.1128/AEM.02411-13
D'Auria G et al (2010) *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics. https://doi.org/10.1186/1471-2164-11-181
Delcher AL et al (2007) Identifying bacterial genes and endosymbiont DNA with glimmer. Bioinformatics. https://doi.org/10.1093/bioinformatics/btm009
Ding W, Baumdicker F, Neher RA (2017) panX: pan-genome analysis and exploration. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx977
Donati C et al (2010) Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biol. https://doi.org/10.1186/gb-2010-11-10-r107
Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr367
Gemmell MR et al (2018) Comparative genomics of campylobacter concisus: analysis of clinical strains reveals genome diversity and pathogenic potential. Emerg Microb Infect. https://doi.org/10.1038/s41426-018-0118-x
Gest H (2004) The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, fellows of the Royal Society. Notes Records R Soc. https://doi.org/10.1098/rsnr.2004.0055
Gladman S, Seemann T (2008) Velvet optimiser. Free Softw Found. https://doi.org/10.1016/S0925-8574(99)00040-3
Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. https://doi.org/10.1038/nrg.2016.49
Gordon A, Hannon GJ (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools, http://hannonlab.cshl.edu/fastx_toolkit/
Gordon SP et al (2017) Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat Commun. https://doi.org/10.1038/s41467-017-02292-8
Grebennikova TV et al (2018) The DNA of bacteria of the world ocean and the earth in cosmic dust at the international Space Station. Sci World J. https://doi.org/10.1155/2018/7360147
Gurevich A et al (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt086
Hadfield J et al (2018) Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx610
He Z et al (2016) Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw370
Holley G, Wittler R, Stoye J (2016) Bloom filter Trie: an alignment-free and reference-free data structure for pan-genome storage. Algorithms Mol Biol. https://doi.org/10.1186/s13015-016-0066-8
Huber W et al (2007) Graphs in molecular biology. BMC Bioinform. https://doi.org/10.1186/1471-2105-8-S6-S8
Hurgobin B, Edwards D (2017) SNP discovery using a Pangenome: has the single reference approach become obsolete? Biology 6(1):21. https://doi.org/10.3390/biology6010021
Inman JM et al (2018) Large-scale comparative analysis of microbial Pan-genomes using PanOCT. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty744
Iqbal Z et al (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. https://doi.org/10.1038/ng.1028
Kara R, Robert JK (2018) Bacteria | cell, evolution, & classification | Britannica.com. Encyclopaedia Britannica, Inc
Keane JA et al (2016) SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genom. https://doi.org/10.1099/mgen.0.000056

Kokot M, Dlugosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. Bioinformatics (Oxford, UK). https://doi.org/10.1093/bioinformatics/btx304
Laing C et al (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinform. https://doi.org/10.1186/1471-2105-11-461
Land M et al (2015) Insights from 20 years of bacterial genome sequencing. Funct Integrat Genom. https://doi.org/10.1007/s10142-015-0433-4
Lanska DJ (2014) Pasteur, Louis. In: Encyclopedia of the neurological sciences. https://doi.org/10.1016/B978-0-12-385157-4.00973-8
Larkin M et al (2007) ClustalW and ClustalX version 2. Bioinformatics. https://doi.org/10.1093/bioinformatics/btm404
Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. https://doi.org/10.1093/nar/gkh152
Lees JA et al (2018) pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty539
Leinonen R et al (2011) The European nucleotide archive. Nucleic Acids Res 39(Suppl 1). https://doi.org/10.1093/nar/gkq967
Limasset A et al (2016) Read mapping on de Bruijn graphs. BMC Bioinform. https://doi.org/10.1186/s12859-016-1103-9
Lukjancenko O et al (2013) PanFunPro: PAN-genome analysis based on FUNctional PROfiles. F1000 Res. https://doi.org/10.12688/f1000research.2-265.v1
Luo R et al (2015) Erratum to "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler" [GigaScience, (2012), 1, 18]. GigaScience. https://doi.org/10.1186/s13742-015-0069-2
Maloy S (2013) Bacterial genetics. In: Encyclopedia of biodiversity: second edition. https://doi.org/10.1016/B978-0-12-384719-5.00431-7
Marcus S, Lee H, Schatz MC (2014) SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu756
Marschall T et al (2016) Computational Pan-genomics: status, promises and challenges. bioRxiv. https://doi.org/10.1101/043430
Mengoni A, Galardini M, Fondi M (2015) Bacterial Pangenomics: methods and protocols. Methods Mol Biol. https://doi.org/10.1007/978-1-4939-1720-4
Minkin I, Pham S, Medvedev P (2017) TwoPaCo: an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. Bioinformatics (Oxford, UK). https://doi.org/10.1093/bioinformatics/btw609
Miyazaki S et al (2004) DDBJ in the stream of various biological data. Nucleic Acids Res 32 (Database issue):D31–D34. https://doi.org/10.1093/nar/gkh127
Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt509
Nawrocki EP et al (2015) Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. https://doi.org/10.1093/nar/gku1063
Ostell J, McEntyre J (2007) The NCBI handbook. NCBI Bookshelf:1–8. https://doi.org/10.4016/12837.01
Page AJ et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31 (22):3691–3693. https://doi.org/10.1093/bioinformatics/btv421
Pandey P et al (2018) Squeakr: an exact and approximate k-mer counting system. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx636
Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. Brief Bioinform. https://doi.org/10.1093/bib/bbq020
Pedersen TL et al (2017) PanViz: interactive visualization of the structure of functionally annotated pangenomes. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw761
Cock PJA et al (2009) The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. https://doi.org/10.1093/nar/gkp1137

Petersen TN et al (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. https://doi.org/10.1038/nmeth.1701

Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. https://doi.org/10.1371/journal.pone.0009490

Rasko DA et al (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol. https://doi.org/10.1128/JB.00619-08

Rizk G, Lavenier D, Chikhi R (2013) DSK: K-mer counting with very low memory usage. Bioinformatics. https://doi.org/10.1093/bioinformatics/btt020

Rouli L et al (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microb New Infect 7:72–85. https://doi.org/10.1016/j.nmni.2015.06.005

Sahl JW et al (2014) The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. Peer J. https://doi.org/10.7717/peerj.332

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Santos AR et al (2013) PANNOTATOR: an automated tool for annotation of pan-genomes. Genet Mol Res. https://doi.org/10.4238/2013.August.16.2

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14):2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Snipen L, Liland KH (2015) micropan: an R-package for microbial pan-genomics. BMC Bioinform. https://doi.org/10.1186/s12859-015-0517-0

Tettelin H et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci 102 (39):13950–13955. https://doi.org/10.1073/pnas.0506758102

Thorpe HA et al (2018) Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. GigaScience. https://doi.org/10.1093/gigascience/giy015

Treangen TJ et al (2014) The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. https://doi.org/10.1186/s13059-014-0524-x

Vernikos G et al (2015) Ten years of pan-genome analyses. Curr Opin Microbiol. https://doi.org/10.1016/j.mib.2014.11.016

'WHO | Press release' (2013) WHO. World Health Organization. Available at: http://www.who.int/whr/1996/media_centre/press_release/en/. Accessed 12 Sept 2018

Wilson RJ (2006) Graph theory. In: History of topology. https://doi.org/10.1016/B978-044482375-5/50018-3

Wozniak M, Wong L, Tiuryn J (2014) ECAMBer: efficient support for large-scale comparative analysis of multiple bacterial strains. BMC Bioinform. https://doi.org/10.1186/1471-2105-15-65

Xiao J et al (2015) A brief review of software tools for pangenomics. Genomics Proteom Bioinform. https://doi.org/10.1016/j.gpb.2015.01.007

Zekic T, Holley G, Stoye J (2018) Pan-genome storage and analysis techniques. Methods Mol Biol. https://doi.org/10.1007/978-1-4939-7463-4_2

Zhao Y et al (2012) PGAP: Pan-genomes analysis pipeline. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr655

Zhao Y et al (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu017

Phyllosphere and Its Potential Role in Sustainable Agriculture

3

Gulab Chand Arya and Arye Harel

Abstract

The Phyllosphere – the microbial composition of the aerial part of the plant – has coevolved with its plant host to populate one of the highly dominated places microbes are able to colonize. In contrast to root associated microbes which are engulfed by a buffering soil, the phyllosphere microbial community is highly affected by environmental factors such as climate variation. Considering the high diversity and abundance of foliar community consisting bacteria, fungi, protozoa and nematodes, the phyllosphere is subjected to complex ecological interactions (e.g., antibiosis, competition for resources, and symbiosis) among its members and the plant host. Similar to observation in human gut microbiome, these interactions are likely to affect plant interaction with pathogens, as partially demonstrated in studies of biocontrol agents. Thus, "Plants wear their guts on the outside" as previously suggested by Janzen DH (1985) The natural history of mutualisms. In: The biology of mutualism: ecology and evolution. Croom/Helm, London/Sydney, pp 40–99. In spite of the importance of this community, there are limited studies that deploy functional omics approaches to study the phyllosphere, and specifically the microbial biotic community associated with pathogenic organism - the Pathobiome. Thus, future studies should include functional analysis of the phyllosphere, role of its community members as biofertilizers and growth stimulators, the effect of nutrients (e. g., K, N, P, Fe) composition on its microbial population profile, and phyllosphere-host interactions. Empowered by "next generation sequencing", findings from these studies should enable to support agrotechnical practice and breeding programs that will improve crops production, quality, and resistance to biotic and abiotic stress.

G. C. Arya · A. Harel (✉)
Department of Vegetable and Field Crops, Institute of Plant Sciences, Agricultural Research Organization, Volcani Center, Rishon LeZion, Israel
e-mail: aryeharel@volcani.agri.gov.il

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_3

This chapter covers most important facets of knowledge accumulated from phyllosphere research: Environmental conditions affecting the establishment and composition of the phyllosphere. Advanced methodologies used for detection and study of the phyllosphere, following summary of its taxonomic composition. The role of the phyllosphere in plant fitness and health, including study of the pathobiome. And finally, the potential use of phyllosphere monitoring and manipulation in sustainable agriculture practices.

3.1 Introduction

Similar to humans and animals, plants are complex organisms that have coevolved with a plethora of soil, water, and airborne microbes. Growing evidences suggest that plants are associated with plenty of microorganisms including endophytes and epiphytes (Bulgarelli et al. 2013). The term phyllosphere was first introduced in 1956 to describe the microbial composition of the aboveground part of the plant including flower and foliage (leaves and stems) (Ruinen 1956). This definition was further modified to describe the microenvironment extending from the leaf surface outward to the outer edge of the boundary layer surrounding the leaf and inward into the leaf tissues (Morris 2002). Thus, phyllosphere can be defined as the microbial composition of the aerial part of the plant.

Among the aerial parts, leaves are one of the highly dominated places microbes are able to colonize. The total leaf surface estimated as $4x10^8$ km^2 may comprise approximately 10^{26} bacterial cells (Morris and Kinkel 2002; Lindow and Brandl 2003). The phyllosphere microbial community is populated mainly by bacteria including various phyla like *Proteobacteria*, *Bacteroidetes*, and *Actinobacteria* (Vorholt 2012); fungi including *Ascomycota*, and *Basidiomycota* (Izuno et al. 2016); yeast; algae; protozoa; and nematodes (Thapa and Prasanna 2018). In contrast to bacteria, the archaeal and fungal composition of the phyllosphere is much smaller (Andrews and Harris 2000; Lindow and Brandl 2003; Arnold and Lutzoni 2007; Jumpponen and Jones 2009); however, it is potentially involved in essential functions such as interactions with different pathogens and carbon/nitrogen dynamics (Vořížková and Baldrian 2013). In contrast to root-associated microbes, phyllosphere microbial community is highly affected by environmental factors such as climate variation affecting water availability and light intensity (Carvalho and Castillo 2018), anthropogenic factors (e.g., use of pesticides) (Karlsson et al. 2014; Glenn et al. 2015; Ottesen et al. 2015), and host genetic factors (e.g., age of the plant, anatomical location, and the host defense system) (Ottesen et al. 2013; Horton et al. 2014; Bringel and Couee 2015 Wagner et al. 2016).

Considering the abundance and variability of phyllosphere microbial community described above, the plant foliar pathogens (e.g., fungi or bacteria) are engulfed in multiple pairwise ecological interactions within the phyllosphere (i.e., mutualism, predation, parasitism, commensalism, and competition), which are mediated through

various mechanisms (e.g., antibiosis, signal interference, competition for resources, manipulating host growth and defense, and quorum sensing) (Lindow and Brandl 2003; Vorholt 2012; Bulgarelli et al. 2013). Study of these interactions and their underlying regulating molecular and physiological mechanisms is a new field holding great potential for developing new strategies to control plant pathogens.

This chapter describes factors allowing successful establishment of the phyllosphere microbial population, methods used for their detection and study, the phyllosphere taxonomic composition, their role in plant fitness and health, the importance of pathobiome study, and the potential use of phyllosphere monitoring and manipulation in sustainable agricultural practices (e.g., use of biocontrol agents, plant growth stimulators, and biofertilizers).

3.2 Environmental Conditions Affecting Microbial Community Composition

In contrast to rhizosphere microbes surrounded by buffering soil supplying a relatively stable environment, the phyllosphere bacteria are subjected to ephemeral and stressful environment. To establish and proliferate in this environment, the microbial community must cope with multiple stresses, such as climate or host-borne low water and nutrient availability, high radiation (mainly UV), warm and cold stresses including freezing, osmotic stress, plant defense activity (e.g., secretion of antimicrobial agents), and anthropogenic factors such as the use of pesticides (Karlsson et al. 2014; Glenn et al. 2015; Ottesen et al. 2015; Lindow and Brandl 2003; Gourion et al. 2006; Vorholt 2012; Bringel and Couee 2015; Sousa et al. 2018). As the anatomy of the leaf often has a significant effect on the microenvironment conditions, environmental stresses are often determined or modulated by the leaf architecture. In the following sections, we will discuss the leaf architecture in this context and the effect of major factors on microbial colonization of foliar organs.

3.3 Leaf Architecture

Physiological conditions resulting from leaf surface have a significant effect on their microbial community. The typical structure of the leaf consists of the cuticle, upper epidermis, palisade layer, veins, spongy layer, and lower epidermis comprising the stomata. The cuticle is a waxy hydrophobic layer that serves as a physical barrier against pathogens and prevents water loss (Yeats and Rosa 2013). Multiple studies illustrated that the majority of the microbial community "select" to avoid such unfavorable conditions by colonizing anatomical sites exhibiting more advantageous microenvironments. That is, the phyllosphere members often colonize the stomata, trichomes, epidermal cell wall junctions, veins, and the surrounding hydathodes (Table 3.1 and related citations). Proximity to these microenvironments could improve attachment, nutrient or water availability [e.g., hydathode exudates (Singh 2014)], and conductivity [e.g., veins (Thapa et al. 2018) and stomata].

Table 3.1 Common site of bacterial aggregation on different leaf surfaces

Site Plant	Epidermal cell wall junction	Cuticle	Stomata	Veins	Trichome
Apple			Mansvelt and Hattingh (1989)	Mansvelt and Hattingh (1989)	Mansvelt and Hattingh (1989)
Pepper					Bashan et al. (1982)
Soybean					Meyer and Wergin (1998)
Grapes	Davis and Brlansky (1991)				
Pear		Mansvelt and Hattingh (1987)	Mansvelt and Hattingh (1987)	Mansvelt and Hattingh (1987)	Mansvelt and Hattingh (1987)
Tomato		Roos and Hattingh (1983) and Timmer et al. (1987)	Roos and Hattingh (1983) and Timmer et al. (1987)	Roos and Hattingh (1983) and Timmer et al. (1987)	Roos and Hattingh (1983) and Timmer et al. (1987)
Rice			Mew et al. (1984)	Mew et al. (1984)	Mew et al. (1984)
Peach			Miles et al. (1977)	Miles et al. (1977)	
Sweet cherry			Roos and Hattingh (1983)		
Olive			Surico (1993)		

Furthermore, bacterial colonies often form aggregates and biofilm structures that increase resistance to stresses (e.g., low water availability) (Bogino et al. 2013), improve their attachment capacity, and supply interactions mediated by signaling and quorum sensing (Williams 2007), which could be beneficial for their survival.

3.4 Availability of Water

The availability of water on the leaf surface also determines the microbial community composition and is highly affected by transpiration, guttation (xylem exudation), (photo)respiration, and photosynthesis. The movement of water starts from its absorption through the roots via the xylem and travels all along through the stem to distribute among different branches of the plants and leaves. Environmental condition such as light intensity affects this process, as well as other water-related flows like transpiration (mediating transpiration), guttation, and (photo)respiration.

Because water accumulates near the stomata, bacterial communities were observed near this site (Table 3.1). The water diffuses through the stomata and condenses and forms a water film over the leaf (Burkhardt and Hunsche 2013), ultimately causing leaching of nutrients further supporting microbial colonization. Other minor aqueous pathways are present close the vascular tissues near the base of the trichrome and in anticlinal cell walls (Schönherr 2006) which are often colonized by microbes (Table 3.1). In addition, a microscopic layer of water observed to accumulate on the leaf cuticle may support microorganism division (Stevenson et al. 2015) and bacterial motility (swimming, twitching, and swarming) on the leaf surface (Beattie 2011). Study of flagellum-dependent *Pseudomonas putida* suggested that water layer thicker than 1.5 μm is suitable for bacterial motility over the leaf surface (Dechesne et al. 2010).

3.5 Photosynthesis

Leaf physiology has significant influence on phyllosphere composition. Physiological parameters are (in)directly regulated by photosynthesis among other processes, such as water and nutrient balance; and CO_2 and O_2 concentrations which have a significant effect on the microbial composition of the phyllosphere (Vacher et al. 2016). Plant nutrients, mainly carbohydrates, are made through the process of photosynthesis in palisade layer of cells and then transported to other parts of the plants. Some amount of these carbohydrates leaches out to the cuticle through diffusion. For example, chemical analysis on bean leaves suggested that about 0.2 to 10 μg of sugar is enough to support the growth of 10^7 to 10^8 cells/leaf (Mercier and Lindow 2000). The process of leaching not only excretes sugars but also other organic compounds, secondary metabolites, and volatile compounds potentially affecting the microbial composition of the phyllosphere (Tukey 1970; Blakeman 1972; Wildman and Parkinson 1981). Plants (and often microbes) emit a variety of volatile organic compounds (VOCs) including terpenoids, aromatic compounds, nitrogen-containing compounds, and volatile sulfur compounds. These compounds mediate bidirectional relationship between plants and microbes ultimately affecting the physiological characteristics of each other (Farré-Armengol et al. 2016). Plants, for example, secrete antimicrobial compounds and different carbon sources ultimately shaping microbial community. One interesting example is the secretion of methanol, which supports colonization of the methylotrophic bacteria *Methylobacterium extorquens*, residing on plant surface and utilizing methanol as the sole carbon source (Sy et al. 2005; Abanda-Nkpwatt et al. 2006). The capability of inhabitation of epiphytic microorganism on plants under ecologically ideal conditions is constrained by the wealth of carbon sources on the leaf surface (Wilson and Lindow 1994a, b). In this regard, it was demonstrated that the initial amount of sugar present on leaf surface before colonization of bacteria determines the subsequent total bacterial population size that it can support (Mercier and Lindow 2000). The total fungal population of the wheat and members of cotton microflora was found to increase with elevated CO_2 concentration; however, no significant

variations were observed in cotton foliar bacteria (Runion et al. 1994; Magan and Baxter 1996).

3.6 Genetic Background of the Host

The genetic background of the host plant species also contributes to variation in microbial community composition (Yang et al. 2001; Rasche et al. 2006a,b, c; Sessitsch et al. 2006; Correa et al. 2007; Redford et al. 2010; Kim et al. 2012; Balint et al. 2013; Bodenhausen et al. 2014; Dees et al. 2015; Müller et al. 2016; Wagner et al. 2016; Li et al. 2018). The effect of genetic background was demonstrated both in trees and field crops. *Trees*. Distinct bacterial communities were observed on six different species of tropical trees (Kim et al. 2012). Balint et al. (2013), who study fungal communities associated with the poplar leaves, suggested that host genotypes had a structuring effect on the composition of foliar fungal communities. In this regard, Redford et al. (2010), who studied bacterial distribution on *Pinus ponderosa* leaves, found that bacterial species variability was lower within a plant species than between different species, even over large geographical distances. Higher proportions of distinct microbial communities were found between different *Citrus* species compared to individuals of the same species (Yang et al. 2001). *Field Crops*. Rasche et al. (2006a, b, c) investigated the effect of plant genotype and identified that different cultivars of potato and pepper harbor genotype-specific bacteria on their leaf surface. In contrast, archaeal communities between wheat cultivars were found to be more similar when compared to bacterial communities (Stapleton and Simmons 2006). Klerks et al. (2007) identified variability in endophytic colonization of (the human pathogen) *Salmonella enterica* serovars among three lettuce cultivars. Sapkota et al. (2015) illustrated the importance of the host genotype by studying the leaf fungal community of the cereal crops, including winter wheat (*Triticum aestivum*), winter and spring barley (*Hordeum vulgare*), oat (*Avena sativa*), rye (*Secale cereale*), and triticale (a *Triticum* x *Secale*, hybrid). They identified a large diversity of nonpathogenic and pathogenic fungal species on different host genetic background. Wagner et al. (2016) studied the leaf and root microbiome of the *Boechera stricta* (Brassicaceae), a perennial wild mustard, and found that the leaf microbiome was highly affected by host genotype. They pointed out that site-specific environmental variables could amplify, overwhelm, or mask the effects of related host effect, giving rise to context-dependent expression of host genetic variation for functional traits influencing fungal communities. Finally, work of Dees et al. (2015) suggested a potential age effect, as they illustrated a shift in bacterial community composition and richness over time throughout the growing season in leafy green vegetables and suggested that host-microbe interactions play a role over time in shaping niches favoring the growth of particular taxa.

Few studies used analysis of mutants to provide functional support for the role of different genotypes in determining the phyllosphere structure and related mechanism. For example, total numbers of aerobic bacteria recovered from the leaves of the genetically modified potato expressing the antibacterial peptide magainin II

(isolated from the skin of the African clawed toad), its non-transgenic parent line, and an unrelated cultivar did not differ significantly (O'Callaghan et al. 2005). However, greater diversity in bacterial community was obtained in potato tubers expressing antibacterial attacin/cecropin (first isolated from the giant silk moth *Hyalophora cecropia*) or T4-lysozyme (form bacteriophage T4) agents compared to control plants (Rasche et al. 2006a, c). In another study, analysis of various *Arabidopsis* cuticular mutants revealed presence of "core" community (bacterial phylotypes that were ubiquitously present on all plant lines) and "plant line-specific" community (positively or negatively affected by the wax mutant phenotype) (Reisberg et al. 2013). The effect of toxins and cuticle on diversity of phyllosphere highlights the important role of plants' defense system in phyllosphere-host interactions. This complex system includes several layers of defense such as physical barriers (e.g., cutin), RNAi, systemic-acquired resistance, (SAR), and innate immunity system. The latter consists of recognition of microbes' conserved patterns (e.g., chitin) by R gene receptors, followed by activation of antimicrobial agents (e.g., the hypersensitivity response, HR, and phytotoxins) and its interaction with microbial effectors (Jones and Dangl 2006; Dangl et al. 2013; Saijo et al. 2018). These interactions are likely to have a significant effect on plant-microbe interactions, mainly for pathogens, and subsequently on the structure of the microbial population.

3.7 Anthropogenic Effect: Pesticides

Anthropogenic factors like fertilization and irrigation regimes, agrotechnical practice (e.g., growth in greenhouses or other structures affecting light temperature and humidity), and application of pesticides are likely to affect the phyllosphere microbial community composition and physiology. The effect of pesticides is potentially significant considering their extensive use for control of plant pathogens in monoculture practice. For example, Ottesen et al. (2015) studied the impact of copper pesticide applications on the phyllosphere microflora of tomatoes and found a significant decrease in the availability of *Gammaproteobacteria* between controls and copper-treated plants, suggesting that copper is effective at suppressing growth of certain taxa in this class. Zhang et al. (2009) studied the effect of cypermethrin insecticide application on the nontarget microbial community of the pepper plant phyllosphere and found an increase in *Bacteroidetes* and *Gammaproteobacteria* phyla. Karlsson et al. (2014) studied the effect of commonly used fungicides (e.g., azoxystrobin, bixafen, cyprodinil, and difenoconazole) on fungal phyllosphere composition in winter wheat. They identified moderate but significant effect on fungal saprotrophs' community composition, while the effect on common wheat pathogens was mixed. Increase (of 2–5 folds) in the microbial population of apple leaf surface treated with the insect repellent kaolin was observed by Glenn et al. (2015). Kaolin is a mineral used in organic agriculture, which is thought to provide microbial habitat due to the increased surface area on the leaf and reduced UV radiation at the leaf surface. Thus, this study highlights the complex interaction between the anthropogenic factors and plant-microbiome interactions.

3.8 Influence of Light on Phyllosphere Community

Plant contains specialized light-harvesting complex (photoreceptor) capturing light photons and subsequently driving photosynthesis and plant growth, while certain taxa of phyllosphere bacteria contain microbial photoreceptors (e.g., rhodopsin) (Atamna-Ismaeel et al. 2012). Since both the plant and certain microbial taxa use light as an energy source, light intensity and wavelength potentially have an important effect on plant-phyllosphere interactions. This assumption is supported by studies demonstrating that different light sources can affect substrate utilization, conidia formation, ROS production, DNA damage, circadian clock, and inhibition of sporulation on multiple phyllosphere harboring species (Carvalho and Castillo 2018). Moreover, studies of light-plant-phyllosphere interactions demonstrated that different light sources may reduce fungal sporulation, suppress pathogenic organism, and inhibit hyphal growth. Some examples of the effect of monochromatic and broader light spectrum on phyllosphere community members are provided in Table 3.2.

3.9 Detection and Study of the Phyllosphere Microorganisms

Identification of microbes from the environment has been routinely practised by scientists. A common classical lab approach is based on culturing methods, making use of a variety of selective media and growth conditions for isolation of target species. However, the estimation that less than 3% of the species are cultured (Schloss and Handelsman 2004; Yashiro et al. 2011; Jackson et al. 2013) emphasizes the importance of a culture-independent untargeted survey in uncovering the composition of phyllosphere community (Table 3.3).

Although limited in the ability to reveal the true nature of microbial diversity, classical methodologies enable deciphering of the physiological and mechanistic nature of selected microbial species and their interaction with the host. These methods include:

(a) Visualization (in situ) of microbes on the surface of the colonized plant parts through light, fluorescent, or scanning electron microscopy, SEM. Empowered by progressive imaging technology, i.e., fluorescence in situ hybridization (FISH) utilizing microbial fluorescence-specific probes, these methods enabled to study the distribution of selected leaf residence and their interaction with the leaf surface (Remus-Emsermann et al. 2014; Peredo and Simmons 2018). (b) Imprinting of selected host surface to adhesive tapes or culture, in order to assess the (in situ) spatial relationships between microbes and plant surface (Yadav et al. 2010). (**c**) Study of (ex situ) cultured isolate species utilize physiological, molecular, and functional genetic (e.g., generating mutants) approaches.

Culture-independent detection methods are practically useful in detecting non-culturable or slow-growing members of the microbiome and in describing the population structure of a sample. Based on the nucleotide, fatty acid, and metabolomics profile among the microbial communities, methods such as 16S

Table 3.2 Direct light effects on phyllospheric microorganisms and plant phyllosphere interaction*

Light Source	Species	Effect
Direct light effects on phyllosphere residing microorganisms		
White	*Pseudomonas* sp. DR 5–09	Changing capacity for substrate utilization
	Botrytis cinerea	Conidia formation
	Neurospora crassa	Conidia formation
UV	Multiple reports	DNA damage, production of ROS
Blue	*Pseudomonas* sp. DR 5–09	Changing capacity for substrate utilization
	Botrytis cinerea	Conidia formation
	Cercospora zeae	Biosynthesis of cercosporin
	Fusarium graminearum	Conidia formation
	Neurospora crassa	Conidia formation
	Peronospora effusa	Sporangial formation
	Trichoderma atroviride	Conidia formation
Red	*Pseudomonas* sp. DR 5–09	Changing capacity for substrate utilization
	Peronospora belbahrii	Inhibition of sporulation
Direct light effects in microorganism, examples of host-microorganism interaction		
White	Rose – *Podosphaera pannosa*	Reduction of number of spores
	Peanut – *Bacillus coagulans*	Predominance under UV exposure
	Peanut – *Clavibacter michiganensis*	Predominance under UV exposure
	Peanut – *Curtobacterium flaccumfaciens*	Predominance under UV exposure
	Rice – *Enterobacter cloacae*	Predominance under UV exposure
	Cucumber – *Podosphaera xanthii*	Suppression of powdery mildew
	Rose – *Podosphaera pannosa*	Suppression of powdery mildew
Blue	Maize – *Cercospora zeae*	Synchronization pathogenesis-maize photoperiodic responses
Red	Basil – *Peronospora belbahrii*	Inhibition of sporulation
	Broad bean – *Botrytis cinerea*	Inhibition of hypha formation and infection
	Onion – *Botrytis cinerea*	Inhibition of hypha formation and infection
	Rose – *Podosphaera pannosa*	Suppression of powdery mildew
R:FR	Rose – *Podosphaera pannosa*	Reduced suppression of powdery mildew by Far-Red

* Adapted from Carvalho and Castillo 2018

rRNA polymerase chain reaction-denaturing gradient gel electrophoresis (16S rRNA PCR-DGGE), phospholipid fatty acid analysis (PLFA), and community-level physiological profiling (CLPP) have been developed to study the diversity within phyllosphere microflora (Table 3.3). However, due to reduction in cost, improve in speed, accessible software, and developed procedures and computational pipelines, currently the most abundant methodology is based on ribosomal 16S

Table 3.3 Examples for culture-independent detection methods for identification of phyllosphere microbiome

Method	Plant and microbial taxa	References
16S rRNA polymerase chain reaction-denaturing gradient gel electrophoresis (PCR-DGGE)	**Rice**: *Proteobacteria*, *Firmicutes*, *Planctomycetes*, *Cyanobacteria*, *Actinobacteria*	Thapa et al. (2018) and Knief et al. (2012)
	Maize: *Sphingomonas*, *Acinetobacter*	Kadivar and Stapleton (2003)
	Seagrasses: *Cytophaga*, *Flavobacterium*, *Bacteroides*	Uku et al. (2007)
	Soybean: *Alphaproteobacteria*, Sphingomonas sp., *Methylobacterium*	Delmotte et al. (2009)
	Apple: *Sphingomonadales*, *Actinomycetales*,	Yashiro et al. (2011)
	Rhizobiales, *Pseudomonadales*, *Burkholderiales*	
	Lettuce: *Proteobacteria*, *Firmicutes*, *Actinobacteria*	Williams et al. (2013)
	Tomato: *Bacillus*, *Pseudomonadales*, *Curtobacterium*, *Sphingomonas*	Enya et al. (2007)
	Arabidopsis: *Proteobacteria*, *Bacteroidetes*, *Actinobacteria*	Horton et al. (2014)
	Common bean: *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Bacteroidetes*	De Oliveira et al. (2012)
Phospholipid fatty acid analysis	***Olea europaea*** L.: Arbuscular mycorrhizal fungi	Mechri et al. (2014)
Metaproteogenomics	**Rice, *Arabidopsis*, clover, soybean**	Knief et al. (2012) and Delmotte et al. (2009)

rRNA sequencing (for prokaryotes, or 18S for eukaryotes, and 18S internal transcribed sequencing, ITS, for fungi). Empowered by next-generation sequencing technology that increases output and precision, this technology enabled to reveal the interaction between microfloral profile and genetic background of the host among other environmental conditions (Knief et al. 2012; Dees et al. 2015) such as, different climates (Finkel et al. 2011; Rastogi et al. 2012), or seasons (Ercolani 1991; Thompson et al. 1993); and anthropogenic factors (Karlsson et al. 2014; Glenn et al. 2015; Ottesen et al. 2015). Extension of this approach by its incorporation with other high-throughput methods holds great potential in deciphering plant-microflora interactions. For example, incorporation of genome-wide association study (GWAS) with 16S rRNA analysis in *Arabidopsis thaliana* enabled to correlate diversity of microflora with host genetic variation (Horton et al. 2014), ultimately aiming to identify loci affecting plant-host interactions. Metagenomics of complete phyllosphere microbiome DNA sequences allows identifying their functional

capacity or even reconstructing novel genomes (Tyson et al. 2004) in addition to their population profile (Finkel et al. 2016). Single cell genomics, integrated with high-throughput cell sorting of microorganisms from the phyllosphere followed by genome sequencing, could enable to study functions of uncultured microbes in the context of microbial species (Yoon et al. 2011; Xu and Zhao 2018). Metaproteogenomics approach integrating both metagenomics (based on complete DNA sequencing, as opposed to 16S) and metaproteomics data was used to identify diversity and related functionality in phyllosphere of rice, *Arabidopsis*, and soybean (Knief et al. 2012; Delmotte et al. 2009). Similarly, integration of transcriptomic analysis with 16S/18S/ITS population survey offers great potential in correlating population profile with functional analysis. Future research integrating the described classical methods together with culture-independent strategies are likely to yield novel findings that include the role of unculturable microbial species in host-microbe-environment interactions.

3.10 Phyllosphere Composition

Multiple studies of the phyllosphere community using isolation and culture-independent approaches (e.g., sequencing of internal transcribed spacer [ITS], 16S, or 18S rRNA and metagenomics) have demonstrated a high bacterial species richness that is restricted to a small number of phyla, primarily *Proteobacteria* (70% abundance), *Bacteroidetes*, and *Actinobacteria* (Table 3.4) (Yang et al. 2001; Arnold and Lutzoni 2007; Delmotte et al. 2009; Finkel et al. 2011; Kim et al. 2012; Knief et al. 2012; Vorholt 2012; Bodenhausen et al. 2013). However, most of these surveys are not focused on higher variability existing in lower taxonomic levels, which could be correlated to potential functionality of the microbiome.

3.11 Phyllosphere Effect on Plant Health and Fitness: Toward Sustainable Agriculture

Similar to findings in the rhizosphere, phyllosphere microbial community interaction with its plant host is often beneficial. Symbiotic relationship provides the foliar microorganisms a shelter, habitat, and nutrient sources, whereas members of the phyllosphere community could serve as a *biofertilizer* (Finkel et al. 2017) **to increase plant productivity or** enhance the overall *protection* capacity against plant pathogens**.** A **biofertilizer** is a composite of living microorganisms (e.g., bacteria or fungi), capable of promoting plant growth by increasing the supply or availability of primary nutrients when applied to plant soil, seeds, or foliar surfaces (Sahoo et al. 2013). An effective biofertilizer can increase nutrient availability through natural processes such as nitrogen fixation (conducted by nitrogenase activity) (Favelli and Messini 1990); nitrification (Watanabe et al. 2016); increasing the uptake of micronutrients, e.g., solubilizing phosphorus (Rodríguez and Fraga

Table 3.4 Phyllosphere composition among plant species

	Abundance in plant species (%)[a]													
Taxa[b]	*Tge*	*Tca*	*Tcl*	*Cxa*	*Can*	*Stu*	*Cal*	*Gma*	*T pr*	*Ath1*	*Ath2*	*Osa*	*Po*	*Zma*
Archaea	–	–	–	–	–	–	–	0.35	–	–	–	0.5	–	–
Eukaryota	–	–	–	–	–	–	–	0.58	–	–	–	–	–	–
Bacteroidetes	17.0	12.9	23.7	20.6	–	2.2	–	12.1	11.0	21.0	16.0	2.0	18.0	6.0
Actinobacteria	4.0	–	–	1.2	5.3	8.0	5.3	9.8	6.0	25.0	4.0	38.1	4.0	7.0
Alphaproteobacteria	20.0	10.9	7.8	32.0	30.8	5.8	15.8	42.8	66.0	38.0	46.0	34.7	36.0	74.0
Betaproteobacteria	29.0	0.9	1.4	2.4	17.9	25.5	10.5	10.3	4.0	6.0	14.0	4.8	–	4.0
Gammaproteobacteria	12.0	75.2	63.7	11.4	25.6	38.6	60.5	5.3	7.0	2.0	16.0	6.0	17.0	6.0
Deltaproteobacteria	–	–	–	–	–	–	–	–	4.0	1.0	–	1.6	–	–
Chloroflexi	–	–	–	–	–	–	–	–	–	–	–	0.6	–	–
Oxalobacteraceae	–	–	–	–	–	–	–	–	–	–	–	–	8.0	–
Deinococcus-Thermus	–	–	–	–	–	–	–	–	–	–	–	–	0.9	
Cyanobacteria	–	–	–	14.5	–	–	2.6	–	–	–	–	–	–	
Firmicutes	12.0	–	–	13.9	20.5	19.8	5.3	–	–	–	–	–	4.0	
Acidobacteria	5.0	–	–	–	–	–	–	–	–		–	–	–	

[a] Initials in column headers indicate species names: *Tge – Thlaspi geosingense*, Idris et al. (2006); *Tca – Trichilia catigua*, Lambias et al. (2006); *Tcl – Trichilia claussenii*, Lambias et al. (2006); *Cxa – Campomanesia xanthocarpa*, Lambias et al. (2006); *Can – Capsicum annuum*, Rasche et al. 2006a, b; *Stu – Solanum tuberosum*, Rasche et al. (2006a); *Cal – Crocus albiflorus*, Reiter and Sessitsch (2006); *Gla – Glycine max*, Delmotte et al. (2009); *Tpr – Trifolium pratense*, Delmotte et al. (2009); *Ath1– Arabidopsis thaliana*, Delmotte et al. (2009); *Ath2 – Arabidopsis thaliana*, Reisberg et al. (2013); *Osa – Oryza sativa*, Vorholt (2012); *Po – Populous* sp. – Crombie et al. (2018); *Zma – Zea mays*, Wallace et al. (2018)

[b] As abundance of taxa derived from metagenomics/16S-based surveys often contains unknown species, columns do not sum up to 100%

1999) and iron [e.g., via siderophore production, (Stintzi et al. 2000)]; and stimulating plant growth through the synthesis of growth-promoting substances. An example for the latter was given by studies that illustrated that phyllosphere microorganism produce auxin-like growth substance (indole acetic acid, IAA) which is known to promote overall growth of the plant (Sun et al. 2014; Fu et al. 2016). Several reports illustrated that phyllosphere microbes can stimulate plant growth. For example, Estiken et al. (2010) observed increase in the plant growth and P, Fe, and Zn content of strawberry leaves upon application of plant growth-promoting microbes *Pseudomonas* and *Bacillus*. Yeast strain isolated from carnivorous plant *Drosera* has been shown to increase lateral root formation in *Arabidopsis* (Sun et al. 2014; Fu et al. 2016). The microbial community members of the phyllosphere can improve plant protection against microbes by indirect interactions such as increase of overall plant health as described above, stimulation of the plant defense system by activation of salicylic acid-mediated SAR (Jones and Dangl 2006), and competition on nutrient sources and niche. Alternatively, bacteria and fungi can participate in direct antibiotic interactions with plant pathogens such as release of secondary metabolites (e.g., toxins, antibiotics, or enzymes), which is capable of inhibiting pathogen growth and infection cycle, promoting pathogen cell death, ultimately leading to seize of the infection or inhibition of its negative effect on the plant host (Lindow and Brandl 2003; Vorholt 2012; Bulgarelli et al. 2013, Pusztahelyi et al. 2015; Vogel et al. 2016; Larousse and Galiana 2017; Prince et al. 2017; Bartoli et al. 2018; Chen et al. 2018; Hassani et al. 2018).

3.12 Biocontrol

There are several known phyllosphere members capable of controlling pathogen activity. While few of them serve as commercial components, studies report only on potential antipathogenic activity for many others (Table 3.5) (Andrews 1992). Antibiosis activity of *Pantoea agglomerans* (strain E325) that was studied on apple flower stigma demonstrated that growth of fire blight-causing agent (*Erwinia amylovora*) was significantly reduced (Pusey et al. 2011). Analysis of lemon's phyllosphere leads to identification of a pseudomonad bacteria *Pseudomonas protegens* CS1 that produced a siderophore pyochelin which acts against citrus canker-causing agent (*Xanthomonas citri* subsp. *citri*) (Michavila et al. 2017). A *Bacillus* (RAB4R) species from rice phyllosphere induced growth inhibited by multiple *Aspergillus flavus* strains and also inhibited aflatoxin production in rice fields as well as in maize seedlings (Chalivendra et al. 2018). Yeasts are also used as potential biocontrol agent because of their ability to persist for long periods in the environment. *Metschnikowia pulcherrima* and *Aureobasidium pullulans* evolved as promising candidates for biocontrol applications against fungal phyllosphere diseases isolated from apple phyllosphere (Gross et al. 2018). Biological control activities of rice-associated *Bacillus* species were also isolated from rice leaves showing protection against fungal sheath blight and bacterial panicle blight of rice (Shrestha et al. 2016). The severity of fire blight disease (in apple trees) caused by

Table 3.5 Examples of biocontrol agents

Control agent	Pathogen and disease	Host	References
Pantoea agglomerans	Fire blight-causing agent *Erwinia amylovora*	Apple, pear	Stockwell et al. (2002)
Metschnikowia pulcherrima	*Aspergillus*	Apple	Gross et al. (2018)
Aureobasidium pullulans	*Aspergillus*	Apple	Gross et al. (2018)
Bacillus species	Rice blight-causing agent *Rhizoctonia solani*	Rice	Shrestha et al. (2016)
Bacillus species	Sheath blight-causing agent *Burkholderia glumae*	Rice	Shrestha et al. (2016)
Pseudomonas graminis	Fire blight-causing agent *Erwinia amylovora*	Apple	Mikiciński et al. (2016)
Bacillus species	*Aspergillus niger*	Potato	Kumar et al. (2018)
Bacillus species	*Rhizoctonia solani*	Potato	Kumar et al. (2018)
Trichoderma viride **commercial**	Black spot caused by *Diplocarpon rosae*	Rose	Karthikeyan et al. (2007)
Pseudomonas fluorescens pf1 **commercial**	Black spot caused by *Diplocarpon rosae*	Rose	Karthikeyan et al. (2007)
Actinomycetes	Bacterial leaf blight caused by *Xanthomonas oryzae*	Rice	Ilsan et al. (2016)
Rhodotorula glutinis	Gray mold-causing agent *Botrytis cinerea*	Tomato	Kalogiannis et al. (2006)

Erwinia amylovora was significantly reduced upon application of *Pseudomonas graminis* identified from the apple phyllosphere (Mikiciński et al. 2016). *Bacillus* species isolated from phyllosphere of different plant species were shown to exhibit antifungal activity against *Aspergillus niger* and *Rhizoctonia solani*. These studies, among others, highlighted the potential use of several fungi and bacteria as biocontrol agents. However, considering multiple requirements (e.g., efficacy, sustainability to relevant agrotechnical practice) a biocontrol agent should fulfill (Pal and Gardner 2006), this field is still missing discovery of new agents, and better understanding of their mechanism of action and their application procedure, that will eventually enable their effective utilization in agriculture.

3.13 Nitrogen Fixation

Nitrogen fixation of free atmospheric N_2 is essential for its subsequent incorporation into organic molecules which serve as fundamental building blocks of life (e.g., nucleic and amino acids) (Kuypers et al. 2018). Nitrogen fixation, carried out by microbes (both free living and symbiotic) abundant in the rhizosphere, has a significant contribution to sustainable agriculture (Igiehon and Babalola 2018).

Thus, phyllosphere microbial community could potentially make a substantial contribution to the nitrogen requirement of field and crop vegetation. Several studies reported that the phyllosphere contains bacterial species capable of nitrogen fixation, and few illustrated nitrogen fixation in this habitat. For example, symbiotic nitrogen fixation has been reported in phyllosphere microbial community (Bulgarelli et al. 2013). Most studies reported on the presence of foliar-residing N_2-fixing species. Sengupta et al. (1981) studied the occurrence of phyllosphere nitrogen-fixing microorganisms in eastern India and identified the diazotroph *Klebsiella pneumoniae* on plants like orchids, *Scindapsus officinalis*, *Ficus*, and cucurbits. In a study in tropical rainforest of Costa Rica, nitrogen-fixing cyanobacterial member (*Nostoc* species) of the genus *Scytonema* was found on leaves of *Spathacanthus hoffmannii* (Freiberg 1998). Members of genera-containing species capable of nitrogen fixation (e.g., *Pseudotsuga menziesii*, *Pseudomonas*, *Bacillus*, *Achromobacter*, *Klebsiella*, and *Mycobacterium*) were found in the needle leaves of *Pinus nigra* and *Pseudotsuga menziesii* (Favilli and Messini 1990). Similarly, in tropical lowland rainforest in Costa Rica, cyanobacteria in addition to diazotrophic bacteria were found on plant species of *Carludovica drudei*, *Grias cauliflora*, and *Costus laevis* (Fürnkranz et al. 2008).

3.14 Phyllosphere Under Climate Changes

The global earth surface temperature is expected to increase to approximately 2–3°C by 2050 due to predicted climate changes (IPCC 2013). The major anthropogenic driving force is the increase in atmospheric gases such as carbon dioxide, nitrous oxide, and methane (Montzka et al. 2011). As a result, an increase in the area of arid and semiarid land is expected, i.e., leading to desertification and salinization of land used for agriculture (Le Houérou 1996; Pankova and Konyushkova 2013). Heat, drought, and salinity stresses resulting from these processes are likely to reduce quality and productivity of staple crops (Fahad et al. 2017), increase sensitivity to pathogens for some plant pathogen systems, and affect composition of foliar-residing microflora. Indeed, elevation in temperature was demonstrated to affect phyllosphere community profile. Campisano et al. (2017) showed a shift of the endophytic community of *Vitis vinifera* induced by temperature changes (imitating predicted climate change) and a stronger influence of the season on bacterial taxa in stems (phyllosphere compartment) compared to roots. The potential negative effect of long-term warming shift was demonstrated in *Galium album* phyllosphere (Aydogan et al. 2018). This study suggested that 2°C increase in temperature significantly reduced the population size of beneficial microbes like *Sphingomonas* and *Rhizobium* species and supports potential colonization of pathogenic bacteria in the phyllosphere of plants. Moreover, Compant et al. (2010) who reviewed the effect of elevated CO_2 suggested that plant-associated microorganisms impart a beneficial effect on plants under elevated CO_2 conditions. Considering the previously discussed beneficial roles of phyllosphere in the control of diseases and support of plant health, climate-induced changes in phyllosphere population can have a

negative effect on plant health. On the other hand, this microflora could prove beneficial in mediating plant resistance to climate change-borne stresses.

3.15 Next-Generation Agriculture

The concept of next-generation agriculture suggested by Schlaeppi and Bulgarelli is based on analogy to personalized medicine used in medical science (Schlaeppi and Bulgarelli 2015). The drugs or medicine tailored, keeping in mind the health risk, genetics of the individual patient, time of diagnosis, optimal treatment, and improvement of health are considered in personalized medicines (Mathur and Sutton 2017). Similarly, to improve the production and quality in agriculture, we should aim to design microbial inoculants customized to the requirements of the specific cultivar, climate, soil, and agricultural practice. Thus, next-generation agriculture concept is based on biogeochemical analysis of farm soil samples to characterize their nutrient (e.g., K, N, P, Fe) composition and microbial population profile. The latter can be empowered by next-generation sequencing to reduce cost speed, and improve production. Consequently, using analysis of these samples, and studies from scientific literature correlating sampled data with required parameters for crops, farmers could apply adequate microbial inoculants and fertilizer to the field. (2017) These studies should further include the potential impact of application of microbial inoculations to the natural population of the soil. Thus, risk assessments should be analyzed for the plant and related microbiome, when both naturally and genetically modified microorganisms are applied (Schlaeppi and Bulgarelli 2015).

This approach could be further expanded to facilitate maintenance and application of a beneficial holobiont plant-microbiome profile in both phyllosphere and rhizosphere. Translating of plant microbiome research into next-generation agriculture technology requires fundamental studies correlating related metadata (climate, cultivar, soil composition, and agricultural practice) with microbiome profile and desired crop parameters, e.g., yield, quality, and health.

3.16 The Pathobiome

The word pathobiome refers to the microbial biotic community associated with pathogenic organism. To date, various studies highlighted the beneficial role of symbiotic root flora (Arnold et al. 2003; Bulgarelli et al. 2013; Schlaeppi and Bulgarelli 2015; Parnell et al. 2016), and few members of the phyllosphere (Nuclo et al. 1998; Zhang and Yuen 1999; Braun-Kiewnick et al. 2000; Elad 2000; Stromberg et al. 2000; Volksch and May 2001; Yuen et al. 2001; Arnold et al. 2003; Innerebner et al. 2011; Parnell et al. 2016) in protecting against plant pathogens. One interesting observation is that axenic plants are more susceptible to infection, by analogy to animal model systems (Innerebner et al. 2011). These observations lead to the hypothesis that upon infection by a pathogen, the plant microbial community undergoes changes that contribute to adaptation to the plant

pathogen. Thus, the study of plant pathogen interactions in a broader perspective, which includes the entire biotic environment, i.e., the pathobiome-holobiont (Jefferson 1994; Vayssier-Taussat et al. 2014; Mitter et al. 2016), will lead to a better understanding of its pathogenicity, persistence, transmission, and underlying molecular mechanisms (Vayssier-Taussat et al. 2014; Jakuschkin et al. 2016).

Fungal and bacterial plant pathogens cause extensive annual yield losses of staple crops worldwide (Oerke 2006; Fisher et al. 2012), demonstrating accelerated evolution (mainly due to continuous fungicide utilization in monoculture practice and human or climate-dependent dispersal), which results in newly evolved resistant pathogens, calling for novel and revamped control strategies (Brown and Hovmoller 2002; Fisher et al. 2012). Similar to advanced practices in the field of root microflora (Arnold et al. 2003; Bulgarelli et al. 2013; Schlaeppi and Bulgarelli 2015; Parnell et al. 2016), and the human gut-microbiome (Pedron and Sansonetti 2008; Gaggia et al. 2010; Clemente et al. 2012), understanding the pathogenicity process in the context of the phyllosphere microbiome may provide new avenues for the development of improved disease control strategies.

3.17 Limited Omics-Based Studies of the Pathobiome

There are only few studies that deploy functional omics approaches to study the pathobiome. As a result, there is limited knowledge regarding pathobiome diversity, functional impact on pathogenicity, the barrier effect (including induction of SAR), and related ecological interactions among microbiome-pathogens-plant and the environment. Several recent studies demonstrated changes in the composition of the pathobiome during the infection process. A study of the resistance of the *A. thaliana bdg* cuticle mutant to *B. cinerea* identified changes in the composition of its microbial community compared to that of the wildtype (Ritpitakphong et al. 2016). The differences were also found at the genus level; the most abundant genera in the mutant were *Pseudomonas* and *Rhizobium*. This study did not demonstrate changes in microbiota composition between infected and noninfected plants. However, because protection of the mutant was not apparent in sterile plants, the data highlight the role of the pathobiome in protection against pathogens. To determine the effect of the pathogen on microbiota composition, a survey of the microbiota was performed on oak trees infected with the fungal obligate biotroph (powdery mildew) *Erysiphe alphitoides* (Jakuschkin et al. 2016). This analysis identified significant changes in the composition of the foliar fungal and bacterial communities as a result of *E. alphitoides* infection, and 26 operational taxonomic units (OTUs) were predicted to directly interact with the pathogen (based on microbial interaction networks). Analysis of the fungal part of the phyllosphere during powdery mildew infection on pumpkin illustrated that the abundance of *Ascomycota* and the *Podosphaera* genus (causing powdery mildew), and that the richness and diversity of the fungal community increased as disease severity increased, likely due to the symbiotic and competitive stresses (Zhang et al. 2018). A systematic survey of *A. thaliana* wild-type plants exposed to various levels of infection by *Albugo*

sp. showed that the biotrophic pathogen affected the diversity of the phyllosphere. In addition, the pathogen generated a hub in the network of organisms (i.e., connecting high number of community members) suggesting it affects colonization of the microbial inhabitants (Agler et al. 2016). Discovery of pathogen hubs may facilitate detection and research of pathogen interacting microbes. This study was followed by analyses that utilized proteomic-based functional analysis of the leaf secretome (apoplast fluid), accompanied by a 16S rDNA and fungal ITS survey to analyze the pathobiome of this system (i.e., *Albugo* sp. and *A. thaliana*) (Ruhe et al. 2016). These results show a decreased alpha diversity of bacteria in infected plants and that restructuring of the community is not mediated by host protein secretion. Because the plant immune system had (a mild) effect on *Albugo* sp., researchers have used an *A. thaliana* line that is mutant in immunity to different biotrophic pathogens (via hormonal regulation), to demonstrate that *Albugo* has adapted to a broad range of pre-existing defense fluctuations in the host. Thus, this functional approach allowed a deeper understanding of the interplay between plant defense, the pathogen, and the associated microbiota.

3.18 Perspectives and Open Questions

Recent advances, especially in high-throughput technologies, enabled to increase our understanding of environmental factors affecting phyllosphere composition, including the host and anthropogenic effects. However, there are gaps in our knowledge in this new field; thus, many questions remain open. An important example is the potential effect of expected climate changes on the phyllosphere composition and consequent effect on plant physiology and health. The potential use of next-generation agriculture is awaiting substantial studies of the effect of agricultural practice on the phyllosphere composition and the effect of the latter on plant health and physiology (e.g., yield and crop quality). Finally, to the best of our knowledge, studies associating plant microbiomes to the host/pathogen transcriptome or proteome [as previously shown in rhizospheres metaproteomes (Wang et al. 2011; Wu et al. 2011)] are currently scarce. The correlation of host functional data to associated microbiome and pathobiome could empower future breeding programs (Gopal and Gupta 2016) aimed to improve crop yield and resistance to diseases under abiotic stresses resulting from climate change.

References

Abanda-Nkpwatt D, Müsch M, Tschiersch J, Boettner M, Schwab W (2006) Molecular interaction between *Methylobacterium extorquens* and seedlings: growth promotion, methanol consumption, and localization of the methanol emission site. J Exp Bot 57(15):4025–4032

Agler MT, Ruhe J, Kroll S, Morhenn C, Kim ST, Weigel D, Kemen EM (2016) Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. Plos Biolo 14

Andrews JH (1992) Biological control in the phyllosphere. Annu Rev Phytopathol 30:603–635

Andrews JH, Harris RF (2000) The ecology and biogeography of microorganisms on plant surfaces. Annu Rev Phytopathol 38:145–180

Arnold AE, Lutzoni F (2007) Diversity and host range of foliar fungal endophytes: Are tropical leaves biodiversity hotspots? Ecology 88:541–549

Arnold AE, Mejía LC, Kyllo D, Rojas EI, Maynard Z, Robbins N, Herre EA (2003) Fungal endophytes limit pathogen damage in a tropical tree. Proc Natl Acad Sci U S A 100:15649–15654

Atamna-Ismaeel N, Finkel OM, Glaser F, Sharon I, Schneider R, Post AF, Spudich JL, Von Mering C, Vorholt JA, Iluz D, Beja O, Belkin S (2012) Microbial rhodopsins on leaf surfaces of terrestrial plants. Environ Microbiol 14:140–146

Aydogan EL, Moser G, Muller C, Kampfer P, Glaeser SP (2018) Long-term warming shifts the composition of bacterial communities in the phyllosphere of galium album in a permanent grassland field-experiment. Front Microbiol 9:144

Bálint M, Tiffin P, Hallström B, O'Hara RB, Olson MS, Fankhauser JD, Piepenbring M, Schmitt (2013) Host genotype shapes the foliar fungal microbiome of balsam poplar (*Populus balsamifera*). PLoS One 8:e53987

Bartoli C, Frachon L, Barret M, Rigal M, Huard-Chauveau C, Mayjonade B, Zanchetta C, Bouchez O, Roby D, Carrere S, Roux F (2018) In situ relationships between microbiota and potential pathobiota in *Arabidopsis thaliana*. ISME J 12:2024–2038

Bashan Y, Diab S, Okon Y (1982) Survival of *Xanthomonas campestris* pv. *vesicatoria* in pepper seeds and roots in symptomless and dry leaves in non-host plants and in the soil. Plant Soil 68:161–170

Beattie GA (2011) Water relations in the interaction of foliar bacterial pathogens with plants. Annu Rev Phytopathol 49:533–555

Blakeman JP (1972) Effect of plant age on inhibition of *Botrytis cinerea* spores by bacteria on beetroot leaves. Physiol Plant Pathol 2:143–152

Bodenhausen N, Horton MW, Bergelson J (2013) Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. PLoS One 8(2):e56329

Bodenhausen N, Bortfeld-Miller M, Ackermann M, Vorholt JA (2014) A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. PLoS Genet 10: e1004283

Bogino PC, Oliva MDLM, Sorroche FG, Giordano W (2013) The role of bacterial biofilms and surface components in plant-bacterial associations. Int J Mol Sci 14(8):15838–15859

Braun-Kiewnick A, Jacobsen BJ, Sands DC (2000) Biological control of *Pseudomonas syringae* pv. *syringae*, the causal agent of basal kernel blight of barley, by antagonistic *Pantoea agglomerans*. Phytopathology 90:368–375

Bringel F, Couee I (2015) Pivotal roles of phyllosphere microorganisms at the interface between plant functioning and atmospheric trace gas dynamics. Front Microbiol 6:486

Brown JK, Hovmoller MS (2002) Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. Science 297:537–541

Bulgarelli D, Schlaeppi K, Spaepen S, van Themaat EVL, Schulze-Lefert P (2013) Structure and functions of the bacterial microbiota of plants. Annu Rev Plant Biol 64:807–838

Burkhardt J, Hunsche M (2013) "Breath figures" on leaf surfaces-formation and effects of microscopic leaf wetness. Front Plant Sci 4:422

Campisano A, Albanese D, Yousaf S, Pancher M, Donati C, Pertot I (2017) Temperature drives the assembly of endophytic communities' seasonal succession. Environ Microbiol 19(8):3353–3364

Carvalho SD, Castillo JA (2018) Influence of light on plant-phyllosphere interaction. Front Plant Sci 9:1482

Chalivendra S, Derobertis C, Reyes Pineda J, Ham JH, Damann K (2018) Rice phyllosphere bacillus species and their secreted metabolites suppress *Aspergillus flavus* growth and aflatoxin production in vitro and in maize seeds. Toxins (Basel) 10:1–16

Chen Y, Wang J, Yang N, Wen Z, Sun X, Chai Y, Ma Z (2018) Wheat microbiome bacteria can reduce virulence of a plant pathogenic fungus by altering histone acetylation. Nat Commun 9:3429

Clemente JC, Ursell LK, Parfrey LW, Knight R (2012) The impact of the gut microbiota on human health: an integrative view. Cell 148:1258–1270

Compant S, Van Der Heijden MG, Sessitsch A (2010) Climate change effects on beneficial plant-microorganism interactions. FEMS Microbiol Ecol 73:197–214

Correa OS, Romero AM, Montecchia MS, Soria MA (2007) Tomato genotype and *Azospirillum* inoculation modulate the changes in bacterial communities associated with roots and leaves. J Appl Microbiol 102:781–786

Crombie AT, Larke-Mejia NL, Emery H, Dawson R, Pratscher J, Murphy GP, McGenity TJ, Murrell JC (2018) Poplar phyllosphere harbors disparate isoprene-degrading bacteria. Proc Natl Acad Sci U S A 115(51):13081–13086

Dangl JL, Horvath DM, Staskawicz BJ (2013) Pivoting the plant immune system from dissection to deployment. Science 341:746–751

Davis CL, Brlansky RH (1991) Use of immune gold labelling with scanning electron microscopy to identify phytopathogenic bacteria on leaf surfaces. Appl Environ Microbiol 7(10):3052–3055

de Oliveira CL, de Queiroz MV, Borges AC, de Moraes CA, de Araújo EF (2012) Isolation and characterization of endophytic bacteria isolated from the leaves of the common bean (*Phaseolus vulgaris*). Braz J Microbiol 43(4):1562

Dechesne A, Wang G, Gulez G, Or D, Smets BF (2010) Hydration-controlled bacterial motility and dispersal on surfaces. Proc Natl Acad Sci U S A 107:14369–14372

Dees MW, Lysoe E, Nordskog B, Brurberg MB (2015) Bacterial communities associated with surfaces of leafy greens: shift in composition and decrease in richness over time. Appl Environ Microb 81:1530–1539

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, Von Mering C, Vorholt JA (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci U S A 106:16428–16433

Elad Y (2000) Biological control of foliar pathogens by means of *Trichoderma harzianum* and potential modes of action. Crop Prot 19:709–714

Enya J, Shinohara H, Yoshida S, Tsukiboshi T, Negishi H, Suyama K, Tsushima S (2007) Culturable leaf-associated bacteria on tomato plants and their potential as biological control agents. Microb Ecol 53:524–536

Ercolani GL (1991) Distribution of epiphytic bacteria on olive leaves and the influence of leaf age and sampling time. Microbial Ecol 21:35–48

Esitken A, Yildiz HE, Ercisli S, Donmez MF, Turan M, Gunes A (2010) Effects of plant growth promoting bacteria (PGPB) on yield, growth and nutrient contents of organically grown strawberry. Sci Hortic 124(1):62–66

Fahad S, Bajwa AA, Nazir U, Anjum SA, Farooq A, Zohaib A, Sadia S, Nasim W, Adkins S, Saud S, Ihsan MZ, Alharby H, Wu C, Wang D, Huang J (2017) Crop production under drought and heat stress: Plant responses and management options. Front Plant Sci 8:1147

Farre-Armengol G, Filella I, Llusia J, Penuelas J (2016) Bidirectional interaction between phyllospheric microbiotas and plant volatile emissions. Trends Plant Sci 21:854–860

Favilli F, Messini A (1990) Nitrogen fixation at phyllospheric level in coniferous plants in Italy. Plant Soil 128:91–95

Finkel OM, Burch AY, Lindow SE, Post AF, Belkin S (2011) Geographical location determines the population structure in phyllosphere microbial communities of a salt-excreting desert tree. Appl Environ Microbiol 77:7647–7655

Finkel OM, Delmont TO, Post AF, Belkin S (2016) Metagenomic signatures of bacterial adaptation to life in the phyllosphere of a salt-secreting desert tree. Appl Environ Microb 82:2854

Finkel OM, Castrillo G, Herrera Paredes S, Salas Gonzalez I, Dangl JL (2017) Understanding and exploiting plant beneficial microbes. Curr Opin Plant Biol 38:155–163

Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ (2012) Emerging fungal threats to animal, plant and ecosystem health. Nature 484:186–194

Freiberg E (1998) Microclimatic parameters influencing nitrogen fixation in the phyllosphere in a Costa Rican premontane rain forest. Oecologia 117(1–2):9–18

Fu SF, Sun PF, Lu HY, Wei JY, Xiao HS, Fang WT, Cheng BY, Chou JY (2016) Plant growth-promoting traits of yeasts isolated from the phyllosphere and rhizosphere of *Drosera spatulata* Lab. Fungal Biol 120:433–448

Fürnkranz M, Wanek W, Richter A, Abell G, Rasche F, Sessitsch A (2008) Nitrogen fixation by phyllosphere bacteria associated with higher plants and their colonizing epiphytes of a tropical lowland rainforest of Costa Rica. ISME J 2:561–570

Gaggia F, Mattarelli P, Biavati B (2010) Probiotics and prebiotics in animal feeding for safe food production. Int J Food Microbiol 141(Suppl 1):S15–S28

Glenn DM, Bassett C, Dowd SE (2015) Effect of pest management system on 'Empire' apple leaf phyllosphere populations. Sci Hortic 183:58–65

Gopal M, Gupta A (2016) Microbiome selection could spur next-generation plant breeding strategies. Front Microbiol 7:1971

Gourion B, Rossignol M, Vorholt JA (2006) A proteomic study of *Methylobacterium extorquens* reveals a response regulator essential for epiphytic growth. Proc Natl Acad Sci U S A 103:13186–13191

Gross S, Kunz L, Muller DC, Santos Kron A, Freimoser FM (2018) Characterization of antagonistic yeasts for biocontrol applications on apples or in soil by quantitative analyses of synthetic yeast communities. Yeast 35:559–566

Hassani MA, Duran P, Hacquard S (2018) Microbial interactions within the plant holobiont. Microbiome 6:58

Horton MW, Bodenhausen N, Beilsmith K, Meng D, Muegge BD, Subramanian S, Vetter MM, Vilhjálmsson BJ, Nordborg M, Gordon JI, Bergelson J (2014) Genome-wide association study of *Arabidopsis thaliana*'s leaf microbial community. Nat Commun 5:5320

Idris R, Kuffner M, Bodrossy L, Puschenreiter M, Monchy S, Wenzel WW, Sessitsch A (2006) Characterization of Ni-tolerant methylobacteria associated with the hyper accumulating plant *Thlaspi goesingense* and description of *Methylobacterium goesingense* sp. nov. Syst Appl Microbiol 29:634–644

Igiehon NO, Babalola OO (2018) Rhizosphere microbiome modulators: Contributions of nitrogen fixing bacteria towards sustainable agriculture. Int J Environ Res Public Health 15:574

Ilsan NA, Nawangsih A, Wahyudi A (2016) Rice phyllosphere actinomycetes as biocontrol agent of bacterial leaf blight disease on rice. Asian J Plant Pathol 10:1–8

Innerebner G, Knief C, Vorholt JA (2011) Protection of *Arabidopsis thaliana* against leaf-pathogenic *Pseudomonas syringae* by *Sphingomonas* strains in a controlled model system. Appl Environ Microbiol 77:3202–3210

IPCC (2013) Climate change 2013 the physical science basis: working group I contribution to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge

Izuno A, Kanzaki M, Artchawakom T, Wachrinrat C, Isagi Y (2016) Vertical structure of phyllosphere fungal communities in a tropical forest in thailand uncovered by high-throughput sequencing. PLoS One 11:e0166669

Jackson CR, Randolph KC, Osborn SL, Tyler HL (2013) Culture dependent and independent analysis of bacterial communities associated with commercial salad leaf vegetables. BMC Microbiol 13:274

Janzen DH (1985) The natural history of mutualisms. In: The biology of mutualism: ecology and evolution. Croom/Helm, London/Sydney, pp 40–99

Jakuschkin B, Fievet V, Schwaller L, Fort T, Robin C, Vacher C (2016) Deciphering the Pathobiome: Intra- and Interkingdom Interactions Involving the Pathogen *Erysiphe alphitoides*. Microb Ecol 72:870–880

Jefferson R (1994) The Hologenome. In: Agriculture, environment and the developing world: A future of PCR. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York
Jones JD, Dangl JL (2006) The plant immune system. Nature 444:323–329
Jumpponen A, Jones KL (2009) Massively parallel 454 sequencing indicates hyper diverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. New Phytol 184:438–448
Kadivar H, Stapleton AE (2003) Ultraviolet radiation alters maize phyllosphere bacterial diversity. Microb Ecol 45:353–361
Kalogiannis S, Tjamos SE, Stergiou A, Antoniou PP, Ziogas BN, Tjamos EC (2006) Selection and evaluation of phyllosphere yeasts as biocontrol agents against grey mould of tomato. Eur J Plant Pathol 116:69–76
Karlsson I, Friberg H, Steinberg C, Persson P (2014) Fungicide effects on fungal community composition in the wheat phyllosphere. PLoS One 9:e111786
Karthikeyan M, Bhaskaran R, Mathiyazhagan S, Velazhahan R (2007) Influence of phylloplane colonizing biocontrol agents on the black spot of rose caused by *Diplocarpon rosae*. J Plant Interact 2:225–231
Kim M, Singh D, Lai-Hoe A, Go R, Abdul Rahim R, Ainuddin AN, Chun J, Adams JM (2012) Distinctive phyllosphere bacterial communities in tropical trees. Microb Ecol 63:674–681
Klerks MM, Franz E, Van Gent-Pelzer M, Zijlstra C, Van Bruggen AH (2007) Differential interaction of *Salmonella enterica* serovars with lettuce cultivars and plant-microbe factors influencing the colonization efficiency. ISME J 1:620–631
Knief C, Delmotte N, Chaffron S, Stark M, Innerebner G, Wassmann R, Von Mering C, Vorholt JA (2012) Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. ISME J 6:1378–1390
Kumar S, Chaudhary D, Jangra R (2018) Establishment of antifungal phyllospheric bacteria in potato (*Solanum tuberosum* L.). Int J Curr Microbiol Appl Sci 7:1048–1056
Kuypers MM, Marchant HK, Kartal B (2018) The microbial nitrogen-cycling network. Nat Rev Microbiol 16(5):263
Lambais MR, Crowley DE, Cury JC, Bull RC, Rodrigues RR (2006) Bacterial diversity in tree canopies of the Atlantic forest. Science 312:1917
Larousse M, Galiana E (2017) Microbial partnerships of pathogenic oomycetes. PLoS Pathog 13: e1006028
Le Houérou HN (1996) Climate change, drought and desertification. J Arid Environ 34(2):133–185
Li Y, Wu X, Chen T, Wang W, Liu G, Zhang W, Li S, Wang M, Zhao C, Zhou H, Zhang G (2018) Plant phenotypic traits eventually shape its microbiota: A common garden test. Front Microbiol 9:2479
Lindow SE, Brandl MT (2003) Microbiology of the phyllosphere. Appl Environ Microbiol 69:1875–1883
Magan N, Baxter ES (1996) Effect of increased CO_2 concentration and temperature on the phyllosphere mycoflora of winter wheat flag leaves during ripening. Ann Appl Biol 129 (2):189–195
Mansvelt EL, Hattingh MJ (1987) Scanning electron microscopy of colonization of pear leaves by *Pseudomonas syringae* pv. syringae. Can J Bot 65:2517–2522
Mansvelt EL, Hattingh MJ (1989) Scanning electron microscopy of invasion of apple leaves and blossoms by pseudomonas syringae pv. Syringae Appl Environ Microbiol 55:533–538
Mathur S, Sutton J (2017) Personalized medicine could transform healthcare. Biomed Rep 7:3–5
Mechri B, Attia F, Tekaya M, Cheheb H, Hammami M (2014) Colonization of olive trees (*Olea europaea* l.) with the arbuscular mycorrhizal fungus glomus sp. Modified the glycolipids biosynthesis and resulted in accumulation of unsaturated fatty acids. J Plant Physiol 171:1217–1220
Mercier J, Lindow SE (2000) Role of leaf surface sugars in colonization of plants by bacterial epiphytes. Appl Environ Microbiol 66:369–374
Mew TW, Mew IC, Huang JS (1984) Scanning electron microscopy of virulent and avirulent strains of *Xanthomonas campestris* pv. oryzae on rice leaves. Phytopathology 74:635–641

Meyer SLF, Wergin WP (1998) Colonization of soybean cyst nematode females, cysts, and gelatinous matrices by the fungus *Verticillium lecanii*. J Nematol 30(4):436

Michavila G, Adler C, De Gregorio PR, Lami MJ, Caram Di Santo MC, Zenoff AM, de Cristobal RE, Vincent PA (2017) *Pseudomonas protegens* CS1 from the lemon phyllosphere as a candidate for citrus canker biocontrol agent. Plant Biol 19(4):608–617

Mikiciński A, Sobiczewski P, Puławska J, Malusa E (2016) Antagonistic potential of *Pseudomonas graminis* 49M against *Erwinia amylovora*, the causal agent of fire blight. Arch Microbiol 198:531–539

Miles WG, Daines RH, Rue JW (1977) Presymptomatic egress of *Xanthomonas pruni* from infected peach leaves. Phytopathology 67(7):895–897

Mitter B, Pfaffenbichler N, Sessitsch A (2016) Plant-microbe partnerships in 2020. Microb Biotechnol 9:635–640

Montzka SA, Dlugokencky EJ, Butler JH (2011) Non-CO_2 greenhouse gases and climate change. Nature 476(7358):43

Morris CE (2002) Phyllosphere. In: Encyclopedia of life sciences. Wiley, Chichester, pp 1–8

Morris CE, Kinkel L (2002) Fifty years of phyllosphere microbiology: significant contributions to research in related fields. In: Lindow SE, Hecht-Poinar EI, Vern JE (eds) Phyllosphere microbiology. APS Press, St. Paul, Minn, pp 365–375

Muller DB, Vogel C, Bai Y, Vorholt JA (2016) The plant microbiota: Systems-level insights and perspectives. Annu Rev Genet 50:211–234

Nuclo RL, Johnson KB, Stockwell VO, Sugar D (1998) Secondary colonization of pear blossoms by two bacterial antagonists of the fire blight pathogen. Plant Dis 82:661–668

O'Callaghan M, Gerard EM, Waipara NW, Young SD, Glare TR, Barrell PJ, Conner AJ (2005) Microbial communities of *Solanum tuberosum* and magainin-producing transgenic lines. Plant Soil 266:47–56

Oerke EC (2006) Crop losses to pests. J Agric Sci 144(1):31–43

Ottesen AR, Peña AG, White JR, Pettengill JB, Li C, Allard S, Rideout S, Allard M, Hill T, Evans P, Strain E (2013) Baseline survey of the anatomical microbial ecology of an important food plant: *Solanum lycopersicum* (tomato). BMC Microbiol 13:114

Ottesen AR, Gorham S, Pettengill JB, Rideout S, Evans P, Brown E (2015) The impact of systemic and copper pesticide applications on the phyllosphere microflora of tomatoes. J Sci Food Agric 95:1116–1125

Pal KK, Gardener BM (2006) Biological control of plant pathogens. Plant Health Instructor 2:1117–1142

Pankova YI, Konyushkova MV (2013) Effect of global warming on soil salinity of the arid regions. Russ Agric Sci 39:464–467

Parnell JJ, Berka R, Young HA, Sturino JM, Kang Y, Barnhart DM, DiLeo MV (2016) From the lab to the farm: an industrial perspective of plant beneficial microorganisms. Front Plant Sci 7:1110

Pedron T, Sansonetti P (2008) Commensals, bacterial pathogens and intestinal inflammation: an intriguing menage a trois. Cell Host Microbe 3:344–347

Peredo EL, Simmons SL (2018) Leaf-FISH: Microscale imaging of bacterial taxa on phyllosphere. Front Microbiol 8:1–14

Prince DC, Rallapalli G, Xu D, Schoonbeek HJ, Cevik V, Asai S, Kemen E, Cruz-Mireles N, Kemen A, Belhaj K, Schornack S, Kamoun S, Holub EB, Halkier BA, Jones JD (2017) *Albugo*-imposed changes to tryptophan-derived antimicrobial metabolite biosynthesis may contribute to suppression of non-host resistance to *Phytophthora infestans* in *Arabidopsis thaliana*. BMC Biol 15:20

Pusey PL, Stockwell VO, Reardon CL, Smits TH, Duffy B (2011) Antibiosis activity of *Pantoea agglomerans* biocontrol strain E325 against *Erwinia amylovora* on apple flower stigmas. Phytopathology 101:1234–1241

Pusztahelyi T, Holb IJ, Pócsi I (2015) Secondary metabolites in fungus-plant interactions. Front Plant Sci 6:573

Rasche F, Marco-Noales E, Velvis H, van Overbeek LS, López MM, van Elsas JD, Sessitsch A (2006a) Structural characteristics and plant-beneficial effects of bacteria colonizing the shoots of field grown conventional and genetically modified T4-lysozyme producing potatoes. Plant Soil 289:123–140

Rasche F, Trondl R, Naglreiter C, Reichenauer TG, Sessitsch A (2006b) Chilling and cultivar type affect the diversity of bacterial endophytes colonizing sweet pepper (*Capsicum annuum* L.). Can J Microbiol 52:1036–1045

Rasche F, Velvis H, Zachow C, Berg G, Van Elsas JD, Sessitsch A (2006c) Impact of transgenic potatoes expressing anti-bacterial agents on bacterial endophytes is comparable with the effects of plant genotype, soil type and pathogen infection. J Appl Ecol 43:555–566

Rastogi G, Sbodio A, Tech JJ, Suslow TV, Coaker GL, Leveau JHJ (2012) Leaf microbiota in an agroecosystem: Spatiotemporal variation in bacterial community composition on field-grown lettuce. ISME J 6:1812–1822

Redford AJ, Bowers RM, Knight R, Linhart Y, Fierer N (2010) The ecology of the phyllosphere: Geographic and phylogenetic variability in the distribution of bacteria on tree leaves. Environ Microbiol 12:2885–2893

Reisberg EE, Hildebrandt U, Riederer M, Hentschel U (2013) Distinct phyllosphere bacterial communities on *Arabidopsis* wax mutant leaves. PLoS One 8:e78613

Reiter B, Sessitsch A (2006) Bacterial endophytes of the wildflower *Crocus albiflorus* analyzed by characterization of isolates and by a cultivation-independent approach. Can J Microbiol 52:140–149

Remus-Emsermann MN, Lucker S, Muller DB, Potthoff E, Daims H, Vorholt JA (2014) Spatial distribution analyses of natural phyllosphere-colonizing bacteria on *Arabidopsis thaliana* revealed by fluorescence in situ hybridization. Environ Microbiol 16:2329–2340

Ritpitakphong U, Falquet L, Vimoltust A, Berger A, Metraux JP, L'Haridon F (2016) The microbiome of the leaf surface of *Arabidopsis* protects against a fungal pathogen. New Phytol 210:1033–1043

Rodriguez H, Fraga R (1999) Phosphate solubilizing bacteria and their role in plant growth promotion. Biotechnol Adv 17:319–339

Roos IM, Hattingh MJ (1983) Scanning Electron Microscopy of *Pseudomonas syringae* pv, morsprunorum on sweet cherry leaves. J Phytopathol 108:18–25

Ruhe J, Agler MT, Placzek A, Kramer K, Finkemeier I, Kemen EM (2016) Obligate biotroph pathogens of the genus *Albugo* are better adapted to active host defense compared to niche competitors. Front Plant Sci 7:820

Ruinen J (1956) Occurrence of *Beijerinckia* species in the 'phyllosphere'. Nature 177:220–221

Runion GB, Curl EA, Rogers HH, Backman PA, Rodriguez-Kabana R, Helms BE (1994) Effects of free-air CO_2 enrichment on microbial populations in the rhizosphere and phyllosphere of cotton. Agric For Meteorol 70:117–130

Sahoo RK, Bhardwaj D, Tuteja N (2013) Biofertilizers: A sustainable eco-friendly agricultural approach to crop improvement. In: Tuteja N, Singh Gill S (eds) Plant acclimation to environmental stress. Springer, New York, pp 403–432

Saijo Y, Loo EP, Yasuda S (2018) Pattern recognition receptors and signaling in plant-microbe interactions. Plant J 93:592–613

Sapkota R, Knorr K, Jørgensen LN, O'Hanlon KA, Nicolaisen M (2015) Host genotype is an important determinant of the cereal phyllosphere mycobiome. New Phytol 207:1134–1144

Schlaeppi K, Bulgarelli D (2015) The Plant Microbiome at Work. Mol Plant-Microbe Interact 28:212–217

Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68:686–691

Schonherr J (2006) Characterization of aqueous pores in plant cuticles and permeation of ionic solutes. J Exp Bot 57:2471–2491

Sengupta B, Naudi AS, Samanta RK, Pal D, Sengupta DN, Sen SP (1981) Nitrogen fixation in the phyllosphere of tropical plants: occurrence of phyllosphere nitrogen-fixing microorganisms in eastern India and their utility for the growth and nutrition of host plants. Ann Bot 48:705–716

Sessitsch A, Hackl E, Wenzl P, Kilian A, Kostic T, Stralis-Pavese N, Sandjong BT, Bodrossy L (2006) Diagnostic microbial microarrays in soil ecology. New Phytol 171:719–736

Shrestha BK, Karki HS, Groth DE, Jungkhun N, Ham JH (2016) Biological control activities of rice-associated *Bacillus* sp. Strains against sheath blight and bacterial panicle blight of rice. PLoS One 11:e0146764

Singh S (2014) Guttation: quantification, microbiology and implications for phytopathology. In: Lüttge U, Beyschlag W, Cushman J (eds) Progress in botany. Progress in botany (genetics – physiology – systematics – ecology). Springer, Berlin/Heidelberg

Sousa LP, Da Silva MJ, Mondego JMC (2018) Leaf-associated bacterial microbiota of coffee and its correlation with manganese and calcium levels on leaves. Genet Mol Biol 41:455–465

Stapleton AE, Simmons SJ (2006) Plant control of phyllosphere diversity: genotype interactions with ultraviolet-B radiation. In: Microbial ecology of aerial plant surfaces

Stevenson A, Burkhardt J, Cockell CS, Cray JA, Dijksterhuis J, Fox-Powell M, Kee TP, Kminek G, Mcgenity TJ, Timmis KN, Timson DJ, Voytek MA, Westall F, Yakimov MM, Hallsworth JE (2015) Multiplication of microbes below 0.690 water activity: implications for terrestrial and extra-terrestrial life. Environ Microbiol 17:257–277

Stintzi A, Barnes C, Xu J, Raymond KN (2000) Microbial iron transport via a siderophore shuttle: a membrane ion transport paradigm. Proc Natl Acad Sci U S A 97:10691–10696

Stockwell VO, Johnson KB, Sugar D, Loper JE (2002) Antibiosis contributes to biological control of fire blight by *Pantoea agglomerans* strain eh252 in orchards. Phytopathology 92:1202–1209

Stromberg KD, Kinkel LL, Leonard KJ (2000) Interactions between *Xanthomonas translucens* pv. *translucens*, the causal agent of bacterial leaf streak of wheat, and bacterial epiphytes in the wheat phyllosphere. Biol Control 17:61–72

Sun PF, Fang WT, Shin LY, Wei JY, Fu SF, Chou JY (2014) Indole-3-acetic acid-producing yeasts in the phyllosphere of the carnivorous plant *Drosera indica* L. PLoS One 9:e114196

Surico G (1993) Scanning electron microscopy of olive and oleander leaves colonized by *Pseudomonas syringae* subsp. *savastanoi*. J Phytopathol 138(1):31–40

Sy A, Timmers AC, Knief C, Vorholt JA (2005) Methylotrophic metabolism is advantageous for *Methylobacterium extorquens* during colonization of *Medicago truncatula* under competitive conditions. Appl Environ Microb 71:7245–7252

Thapa S, Prasanna R (2018) Prospecting the characteristics and significance of the phyllosphere microbiome. Ann Microbiol 68:229–245

Thapa S, Ranjan K, Ramakrishnan B, Velmourougane K, Prasanna R (2018) Influence of fertilizers and rice cultivation methods on the abundance and diversity of phyllosphere microbiome. J Basic Microbiol 58:172–186

Thompson IP, Bailey MJ, Fenlon JS, Fermor TR, Lilley AK, Lynch JM, McCormack PJ, McQuilken MP, Purdy KJ, Rainey PB, Whipps JM (1993) Quantitative and qualitative seasonal-changes in the microbial community from the phyllosphere of sugar-beet (*Beta vulgaris*). Plant Soil 150:177–191

Timmer LW, Marios JJ, Achor D (1987) Growth and survival of *Xanthomonads* under conditions nonconductive to disease development. Phytopathology 77:1341–1345

Tukey HB (1970) The leaching of substances from plants. Annu Rev Plant Physiol 21:305–324

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Uku J, Bjork M, Bergman B, Diez B (2007) Characterization and comparison of prokaryotic epiphytes associated with three east African sea grasses. J Phycol 43:768–779

Vacher C, Hampe A, Porté AJ, Sauer U, Compant S, Morris CE (2016) The phyllosphere: microbial jungle at the plant–climate interface. Annu Rev Ecol Evol Syst 47:1–24

Vayssier-Taussat M, Albina E, Citti C, Cosson JF, Jacques MA, Lebrun MH, Le Loir Y, Ogliastro M, Petit MA, Roumagnac P, Candresse T (2014) Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. Front Cell Infect Microbiol 4:29

Vogel C, Bodenhausen N, Gruissem W, Vorholt JA (2016) The Arabidopsis leaf transcriptome reveals distinct but also overlapping responses to colonization by phyllosphere commensals and pathogen infection with impact on plant health. New Phytol 212(1):192–207

Volksch B, May R (2001) Biological control of *Pseudomonas syringae* pv. *glycinea* by epiphytic bacteria under field conditions. Microb Ecol 41:132–139

Vorholt JA (2012) Microbial life in the phyllosphere. Nat Rev Microbiol 10:828–840

Voříšková J, Baldrian P (2013) Fungal community on decomposing leaf litter undergoes rapid successional changes. ISME J 7:477–486

Wagner MR, Lundberg DS, Tijana G, Tringe SG, Dangl JL, Mitchell-Olds T (2016) Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. Nat Commun 7:1–15

Wallace J, Kremling KA, Kovar LL, Buckler ES (2018) Quantitative genetics of the maize leaf microbiome. Phytobiomes J.: In Press 2:208

Wang HB, Zhang ZX, Li H, He HB, Fang CX, Zhang AJ, Li QS, Chen RS, Guo XK, Lin HF, Wu LK, Lin S, Chen T, Lin RY, Peng XX, Lin WX (2011) Characterization of metaproteomics in crop rhizospheric soil. J Proteome Res 10:932–940

Watanabe K, Kohzu A, Suda W, Yamamura S, Takamatsu T, Takenaka A, Koshikawa MK, Hayashi S, Watanabe M (2016) Microbial nitrification in through fall of a Japanese cedar associated with archaea from the tree canopy. Springerplus 5:1596

Wildman HG, Parkinson D (1981) Seasonal changes in water-soluble carbohydrates on *Populus tremuloides* leaves. Can J Bot 59:862–869

Williams P (2007) Quorum sensing, communication and cross-kingdom signalling in the bacterial world. Microbiology 153:3923–3938

Williams TR, Moyne AL, Harris LJ, Marco ML (2013) Season, irrigation, leaf age, and *Escherichia coli* inoculation influence the bacterial diversity in the lettuce phyllosphere. PLoS One 8:e68642

Wilson M, Lindow SE (1994a) Ecological similarity and coexistence of epiphytic ice-nucleating (ice+) *Pseudomonas syringae* strains and a non-ice-nucleating (ice-) biological control agent. Appl Environ Microbiol 60:3128–3137

Wilson M, Lindow SE (1994b) Coexistence among epiphytic bacterial populations mediated through nutritional resource partitioning. Appl Environ Microbiol 60(12):4468–4477

Wu L, Wang H, Zhang Z, Lin R, Zhang Z, Lin W (2011) Comparative metaproteomic analysis on consecutively *Rehmannia glutinosa*-monocultured rhizosphere soil. PLoS One 6(5):e20611

Xu Y, Zhao F (2018) Single-cell metagenomics: Challenges and applications. Protein Cell 9:501–510

Yadav RKP, Kakamanoli K, Despoina V (2010) Estimating bacterial population on the phyllosphere by serial dilution plating and leaf imprint methods. Ecoprint: Int J Ecol 17:47–52

Yang CH, Crowley DE, Borneman J, Keen NT (2001) Microbial phyllosphere populations are more complex than previously realized. Proc Natl Acad Sci U S A 98:3889–3894

Yashiro E, Spear RN, McManus PS (2011) Culture- dependent and culture independent assessment of bacteria in the apple phyllosphere. J Appl Microbiol 110:1284–1296

Yeats TH, Rose JK (2013) The formation and function of plant cuticles. Plant Physiol 163:5–20

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332:714–717

Yuen GY, Steadman JR, Lindgren DT, Schaff D, Jochum C (2001) Bean rust biological control using bacterial agents. Crop Prot 20:395–402

Zhang Z, Yuen GY (1999) Biological control of *Bipolaris sorokiniana* on tall fescue by *Stenotrophomonas maltophilia* strain C3. Phytopathology 89:817–822

Zhang B, Bai Z, Hoefel D, Tang L, Wang X, Li B, Li Z, Zhuang G (2009) The impacts of cypermethrin pesticide application on the non-target microbial community of the pepper plant phyllosphere. Sci Total Environ 407:1915–1922

Zhang Z, Luo L, Tan X, Kong X, Yang J, Wang D, Zhang D, Jin D, Liu Y (2018) Pumpkin powdery mildew disease severity influences the fungal diversity of the phyllosphere. Peer J 6:e4559

Functional Genomics and Systems Biology Approach for Understanding Agroecosystems

4

Birendra Singh Yadav and Ashutosh Mani

Abstract

Plant metabolism is affected by several biotic and abiotic factors of our environment that leads to low yield in crops. The integrative approach of functional genomics and systems biology is one of the most promising tools for understanding the agroecosystems. In this chapter, we will discuss the role of functional genomics to study the effect of stress on plants. Various approaches and tools of systems biology will be also discussed to understand the alteration in biological networks, i.e., gene regulatory, protein-protein and metabolic networks, etc. Different tools available for studying the agroecosystems using omics and systems biology have been explored here in detail.

4.1 Introduction

The natural ecosystems which are modified for the production of food and fiber are known agroecosytems. There are several biotic and abiotic factors that are also present in the natural ecosystems. Agroecosytems supports the production of many crops but the environmental factors affect the productivity of crops (Ptaszek 2013). The interaction of biotic and abiotic stress component of environment affects the life as well productivity of crops, and hence, it is quite necessary to study the role of interference and underlying mechanisms of plants to sustain against the challenges. The agroecosysystems also interact positively and negatively with insects, birds, and weeds and contribute in sustainability of crops (GANS 2005).

Functional genomics is the study of function and interactions of genes/proteins by using genome-wide approaches by integrating the data obtained from different

B. S. Yadav · A. Mani (✉)
Department of Biotechnology, Motilal Nehru National Institute of Technology, Allahabad, India
e-mail: amani@mnnit.ac.in

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_4

processes like DNA sequence, gene expression, transcriptome profiling, and DNA-protein, RNA-protein, and protein-protein interactions. On the basis of above-mentioned information, one can build the model of interaction that regulates the gene expression and other biological processes (Bunnik and Le Roch 2013). In 1998 Weinstein (Weinstein 1998) coined the term "omics" and classified it into genomics, transcriptomics, and proteomics. Functional genomics fills the research gap between the classical gene expression and genome-wide expression and it is the correlation with biological process (Gasperskaja and Kučinskas 2017).

Plants are ultimate source of energy, food, and other valuable compounds. Systems biology allows us to understand how plants used to synthesize the various valuable compounds as well as it also correlate the phenotype and genotype (Kell 2002; Dhondt et al. 2013). The productivity of various plants is decreasing day by day due to the biotic and abiotic factor of environment, and this includes cold, drought, heat, salinity, and heavy metals (Bebber et al. 2013). Plants are very much susceptible to stresses and sometimes all these stress acts simultaneously and plants act accordingly if they unable to process it may die (Ramegowda and Senthil-Kumar 2015; Mushegian 2017).

Various omics approaches as shown in Fig. 4.1 enable the researcher to identify the stress-responsive genes, pathways, and secondary metabolites using genomics and systems biology tools (Pandey et al. 2015). The integrative approach of transcriptomics, proteomics, metabolomics, phenomics, and interactomics with systems biology is used to understand the underlying mechanism in plants during various stress conditions (Kumar et al. 2015; Ben-Amar et al. 2016; Pandey et al. 2016). The systems of plants are well studied and understood by using the omics and systems biology approach and are useful for discovering the marker genes associated with stress and it is the very first step to develop the tolerance variety of crops (Saito and Matsuda 2010; Dhondt et al. 2013).

In this chapter, we have focused the application of different omics and systems biology approach during biotic and abiotic stress on plants to understand the tolerance strategy of plants.

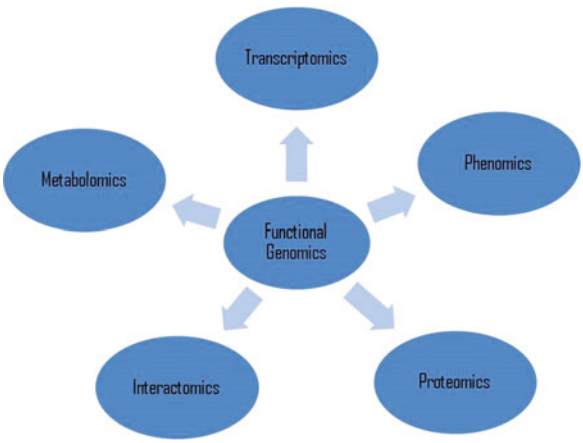

Fig. 4.1 Various aspects of functional genomics that are used to understand the effect of stress on plants to understand the mechanism of defence

4.2 Transcriptomics

Gene expression is also affected by modification in DNA by DNA methylation without changing the DNA sequence. This can be determined by using methylation-dependent restriction enzyme (MDRE) or bisulfite conversion and its PCR (Schuster 2007; Zilberman and Henikoff 2007). Other modifications like acetylation, phosphorylation, and ubiquitination also regulate the gene expression (Bannister and Kouzarides 2011), and it can be investigated using chromatin immunoprecipitation (CHIP) (Shendure and Aiden 2012). The information content of an organism is present in its genomic DNA and is expressed through transcription, where transcriptome requires the information content of genome at a particular time and allow us to study the differential expression pattern at a specific condition (Lowe et al. 2017). The term transcriptome was used in 1990 (Piétu et al. 1999), and before this scientists were using serial analysis of gene expression (SAGE) based on Sanger sequencing (Velculescu et al. 1995).

Next-generation sequencing (NGS) is the massively parallel sequencing of the genome under specific conditions to study the genomes using RNA sequencing (RNA Seq) which is high-throughput sequencing technology (Ozsolak and Milos 2011), and it intakes the RNA in nanogram (Hashimshony et al. 2012). It does not require the prior information of genomic sequencing whose sequencing/analysis has to be done (Stefano 2014) and has high accuracy up to 90% in sequencing. NGS methods are also able to detect the SNPs with high technical reproducibility of 99% (Marioni et al. 2008) as explained in Fig. 4.2 the various aspects of NGS.

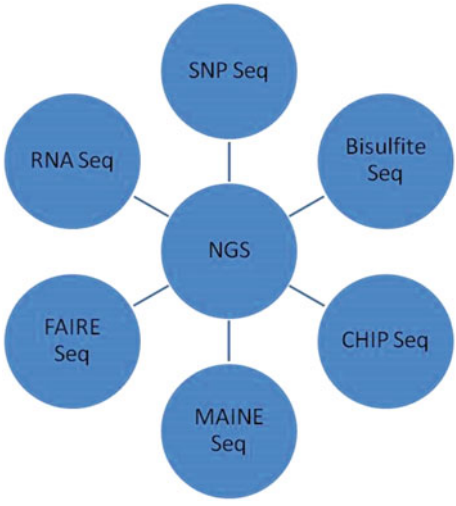

Fig. 4.2 The various role of Next generation sequencing used for analysis and understanding the biological process in plants and animals

DNA microarray has the ability to measure the simultaneous expression of thousands of genes in at the particular stage of cells (Schena et al. 1995). This technique is used to investigate the diagnostic or prognostic biomarkers, disease-associated genes, and the response of gene against a particular drug/stress in plants and animals to understand the mechanism involved in a particular condition (Aguan et al. 2000; Zhang et al. 2005). The first mRNA is isolated from a normal and experimental condition which are transcribed into cDNA and labeled with dye and finally allowed to hybridize with a probe attached on chip to measure the level of mRNA in control and experimental condition (Gibson and Muse 2009). Lots of studies have been performed on various plants during biotic/abiotic stress using transcriptomics approach, and some important studies are listed in Table 4.1.

Table 4.1 List of transcriptomics studies available on various plants to understand the abiotic stress using omics approach

S. No	Species	Stress treatment	Tissue	Comments	References
1.	*Arabidopsis thaliana*	Cadmium 5 μM,50 μM	Root and shoot	During cadmium treatment of 50 μM of cd in Arabidopsis, transcriptome study was performed and revealed that the sulfur assimilation pathway was increased leads to more production of GSH and phenylpropanoid biosynthesis was also enhanced	Herbette et al. (2006)
2.	*Glycine max*	Cadmium 40 μM	Root	Oxidative markers production was reduced and free radicals were generated in a large amount and it was also observed that GST was increased during cadmium exposure of 40 μM exposure	McLaughlin et al. (2000)
3.	Barely	Cadmium 80 μM	Roots, shoots	The minimum inhibitory concentration of cadmium was calculated and the treatment of cadmium was given. Exposed roots and shoots were selected and used for mRNA isolation and RNA-Seq was performed but the analysis is not done	Kintlová et al. (2017)

(continued)

Table 4.1 (continued)

S. No	Species	Stress treatment	Tissue	Comments	References
4.	*Zea mays*	Cadmium 100 μM	Root	Transcriptome analysis of *Zea mays* was performed under 100 μM of cadmium and it was observed that indole acetic acid (IAA), auxin biosynthesis and transporter genes were underexpressed	Yue et al. (2016)
5.	*Oryza sativa*	Cadmium 50 μM	Root, shoot	Seedlings were damaged and level of expressions of transporter genes was affected, revealed by transcriptome analysis of *Oryza sativa* under exposure of cadmium	Oono et al. (2016)
6.	*Oryza sativa*	Chromium 25 μM, 50 μM, 100 μM, 250 μM	Root	Different concentration of chromium was given to rice and phenotypic studies were performed. Root treated with 100 μM chromium was selected and transcriptome analysis was performed and reported an increased level of lipid peroxidation and proline synthesis. Glutathione plays important role in chromium detoxification as reported by them	Dubey et al. (2010)
7.	*Zea mays*	300 mg/ml Chromium	Leaves	The chromium stress associated genes were identified and that are responsible for ROS detoxification and defense response. Morphology of *Zea mays* was changed during chromium treatment	Wang et al. (2013)

(continued)

Table 4.1 (continued)

S. No	Species	Stress treatment	Tissue	Comments	References
8.	*Brassica napus*	Chromium 25 μM and 100 μM	Leaves	Photosynthesis efficiency, ATP synthesis, and transpiration are adversely affected during exposure of chromium and proteomics study also prove the above-mentioned results. It was also observed that phosphoglycolate production was enhanced during chromium exposure	D'Alessandro et al. (2013)
9.	*Crambe abyssinica*	Chromium 50 μM 100 μM 150 μM 250 μM	Seedlings	Various concentration of chromium is exposed to plant and 150 μM concentration of chromium was selected for further studies. Ion transporter, sulfur assimilation, photosynthesis, photosynthesis, and cell metabolism were affected due to chromium exposure	Zulfiqar et al. (2011)
10.	*Arabidopsis* plant	Arsenic 100 μM 200 μM 300 μM	Root	Different concentration of arsenic was given to tolerance and susceptible varieties of Arabidopsis and he revealed that ethylene-related pathway was changed. Heat shock genes and aqua transporter genes expression was varied	Fu et al. (2014)
11.	Barley	Arsenic 5 μM	Root	Exposure of arsenate and arsenite in barley was compared with other plants and concluded that barely having low uptakes. Phosphate transporter gene was affected and arsenate and arsenite were localized in xylem sap	Su et al. (2010)

(continued)

Table 4.1 (continued)

S. No	Species	Stress treatment	Tissue	Comments	References
12.	Maize	Arsenic 5 ppm 10 ppm	Leaves	Lipid peroxidation was increased due to exposure of arsenate at 10 ppm. SOD and peroxidase activity was also increased during As(V) exposure proved by biochemical assays	Kumar Yadav and Srivastava (2015)
13.	*Crambe abyssinica*	Arsenic 100 μM 150 μM 200 μM 250 μM 300 μM	Seedlings	For studying molecular mechanism involved for detoxification of as 250 μM concentration exposed to plant and subjected to microarray and reported that sulfur metabolism, heat shock protein, and metal transporter protein expression were altered	Paulose et al. (2010)
14.	*Oryza sativa*	Arsenic 5 μM 10 μM 25 μM 50 μM	Root	Root sample treated with 25 μM of as (V) was used to extract the mRNA and microarray was performed. The DEGs involved in cell wall biogenesis and cell cycle were downregulated. The ethylene response factor, heat shock factor, and myb-like genes were upregulated for detoxification of arsenic	Huang et al. (2012)
15.	*Medicago truncatula*	Arsenic 25 μM	Root	The microarray of Medicago treated with 25 μM of as(III) was performed and validated with q-RT PCR. Root growth and rate of photosynthesis were decreased	Lafuente et al. (2015)

It has been revealed by a transcriptomic study that many genes undergo differential exposure during exposure of environmental stress. The genes which are upregulated and downregulated are a key player in several biological processes and molecular function. Stress-responsive gene which is overexpressed is involved in different defense mechanism against the environmental stress while underexpressed genes are generally involved in the storage process. Certain metabolic pathways are overexpressed and underexpressed for providing tolerance against heavy metal stress in chickpea plant (Szklarczyk et al. 2017).

4.3 Proteomics

In understanding the biological process, it is necessary to understand the function of the protein in a biological process. Sometimes transcriptomics study does not correlate with proteomics study due to posttranslational modification which changes the function of the protein. Two-dimensional gel electrophoresis can be used to analyze the protein content on a cell where proteins are first separated by size followed by mass spectrometry. LC-MS is also one of the approaches where proteins are first separated by one-dimensional SDS PAGE, then protein is digested and separated by LC and analyzed by MS. Multidimensional protein identification technology is high-throughput technology used for separating protein from complex mixture by digesting proteins into peptides followed by separation on the basis of charge and hydrophobicity and finally analyzed by MS. (Washburn et al. 2001). The signal obtained from MS data is compared with a database for identification of the protein.

Stress associated proteins were differentially expressed during abiotic stress as reported in various literature contributing towards the stress tolerance (Witzel et al. 2009; Hossain et al. 2012; Pérez-Clemente et al. 2013). The different proteomic study has been performed for, e.g., drought stress (Caruso et al. 2008; Mirzaei et al. 2012; Mohammadi et al. 2012; Cramer et al. 2013; Zhang et al. 2016), salt stress (Nam et al. 2012; Zhu et al. 2012), water lodging (Komatsu et al. 2009, 2010, 2013a, b, 2014, Alam et al. 2010a, b), and heat stress (Rollins et al. 2013; Xuan et al. 2013). The complex biological process is analyzed using proteomics in plants during stress (Aghaei and Komatsu 2013; Ghosh and Xu 2014; Gong et al. 2015). The methodology of the proteomics approach for identifying candidate protein during stress condition is shown in Fig. 4.3.

4.4 Interactomics

The union of informatics, biochemistry, and engineering technology enables the researcher to understand the interactions of proteins used to study under interactomics. The omics technology enable the researcher to understand the biological system and interaction of expressed proteins in a cell (the proteome) and genome encoded product along with its interaction in complex biological

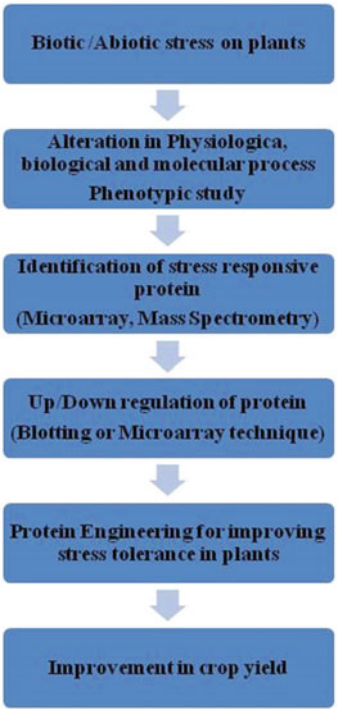

Fig. 4.3 Flow chart of methodology to perform the proteomics study on plants during stress conditions

network i.e. the interactome. It is a fast-growing area of systems biology for understanding the biological process and regulatory network during biotic and abiotic stress in plants. The study of interactions in signal transduction, transcriptional regulation, metabolic pathways, and other biological processes is said as interactomics. Lots of approaches have been developed for studying the interactome like in silico, in vivo, and in vitro (Rao et al. 2014). The first approach includes the computational analysis and text mining; in vivo includes Y2H hybrid system, while the in vitro method experiments are performed on the living organism to understand the biological functions interactome.

Uhrig, Williams, and Bowles in 2006 performed the protein-protein interaction network study on Arabidopsis that was based on literature mining and co-expression approach (Williams and Bowles 2004; Uhrig 2006). Jane Geisler-Lee in 2007, predicted Arabidopsis protein interactome based on the interolog method. They also concluded that the predicted proteins were co-localized and co-expressed by analyzing existing experimental data from Arabidopsis and decipher the significant role of signaling and cellular function by enabling hypothesis generation Arabidopsis interactome (Geisler-Lee et al. 2007).

4.5 Metabolomics

The study of metabolites at the particular instant in agroecosystem is known as metabolomics. It is important to understand the plant stress response in terms of metabolites. Lots of studies have been done for deciphering the role of metabolite during different abiotic stress condition. In 2004, Rizhky performed the metabolic profiling of plants during drought, heat, and combined stress and reported accumulation of sucrose and other sugar like starch (Rizhsky 2004). NMR based metabolic fingerprinting of plants during heavy metal stress were performed by Bailey in 2003 and responsible metabolic pathways were explored (Bailey et al. 2003). The microarray data of chickpea during heavy metal stress were analyzed and responsible pathways which were providing tolerance were also identified. It was seen that Nitrogen metabolism, Starch sucrose metabolism, and Riboflavin metabolism were altered. Several metabolomics approacheses was also used to understand the production of metabolite during different stress condition in plants like water salinity (Cramer et al. 2007), oxidative stress (Baxter et al. 2006), and heavy metal stress (Le Lay et al. 2006).

The metabolomics data is similar to transcriptomics and proteomics data. It requires lots of computational work including file handling, data mining, and finally comparative analysis. Lots of online server/databases/tools are available for analysis and visualization of metabolic pathway for understanding the agroecosystem shown in Table 4.2. Once the function of metabolite will be known, one can directly correlate with the function of the gene during particular stress condition. It is a rapidly growing technology to understand the metabolic pathways in plants by using different strategies like targeted analysis, metabolite profiling, and metabolic fingerprinting (Fiehn 2002; Halket et al. 2005; Shulaev 2006).

Table 4.2 List of various online tools and their web address that are used for analysis of various metabolic pathways in plants

Pathway database/ tools	Web address	References
KEGG	https://www.genome.jp/kegg/	Kanehisa et al. (2017)
BioCyc	https://biocyc.org/	Paley and Karp (2006)
MetaCyc	https://metacyc.org/	Caspi (2006)
AraCyc	https://www.plantcyc.org/databases/aracyc/15.0	Zhang (2005)
MapMan	https://mapman.gabipd.org/	Thimm et al. (2004)
KaPPA-view	http://kpv.kazusa.or.jp/	Tokimatsu (2005)
BioPathAT	http://www.murdockmetabolomics.wsu.edu/LangeLabHome.html	Lange and Ghassemian (2005)
MetNetDB	http://metnetweb.gdcb.iastate.edu/MetNet_db.htm	Wurtele et al. (2003)

4.6 Phenomics

Generally, plants interact with the biotic and abiotic component of the ecosystem. As a result, there is a change in genotype and phenotype of plants. The alteration in phenotype is due to combined interaction of genome with the environment. The study of phenotype during the interaction of biotic and abiotic factor of the environment at a particular instant is known as phenomics. Invasive and noninvasive techniques are available to understand the change in phenotype in model plants (Großkinsky et al. 2015). Phenomics technology is used in basic plant research during different stress conditions and crop breeding (Furbank and Tester 2011). A noninvasive method of plant phenomics methodology is shown in Fig. 4.4.

In image acquisition, digital image of plants is taken to study the effect of biotic and abiotic effect on plants. It is done by several approaches like tomography imaging (Bovik 2005), thermography imaging (Padhi et al. 2012), LIDAR (Lenco 1982), and time-of-flight camera (Klose et al. 2011). After image acquisition, processing of the image is done for removing the noise and improving the contrast (Hamuda et al. 2016). Image cropping, contrast improvement, and dimensional reduction are major operation done during image processing (Singh et al. 2016). After the image preprocessing, segmentation of image is done, in which objects are identified and isolated by removing irrelevant background present in the objects (Singh and Misra 2017). In the feature extraction process, the image of interest is used to extract the numerical value where different algorithms can be applied to understand the phenotype (Van Der Heijden et al. 2012). Finally, the generated data is analyzed by a machine-learning approach (Ghatak 2017).

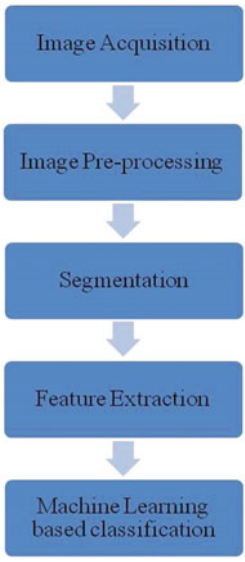

Fig. 4.4 Various steps involved in generation and analysis of Phenomics data in plants for understanding the effect of stress

4.7 Role of Systems Biology for Understanding Agroecosystem

For a comprehensive study of plants systems biology, the role of bioinformatics is crucial. "Omics" data analysis and interpretation required the good skill of bioinformatics and algorithm (Joyce and Palsson 2006). The omics data can be analyzed for understanding plants systems, and this requires lots of tools for visualization of networks and construction of pathways and analysis tools as shown in Fig. 4.5.

Systems biology represents the graphical view of biomolecules and their interactions in the form of a biological network (BN). In BN, nodes of the graph represent the gene/protein/DNA/RNA and edges represent the way of interaction. The interaction may be direct (flow of information from one to another biomolecule) and undirected (having interaction but the direction is not sure). The way of interaction may be physical interaction, metabolic reaction, and regulatory connection (Joyce and Palsson 2006). The most connected node in networks is termed as hub node, which is a key player in BN (Barabási and Oltvai 2004). During stress condition in plants, the interaction between biomolecule changes and it can be well studied and understood by using systems biology by constructing or reconstructing gene to metabolite network, protein-protein interaction networks, gene regulatory networks, and transcriptional regulatory network (Yuan et al. 2008).

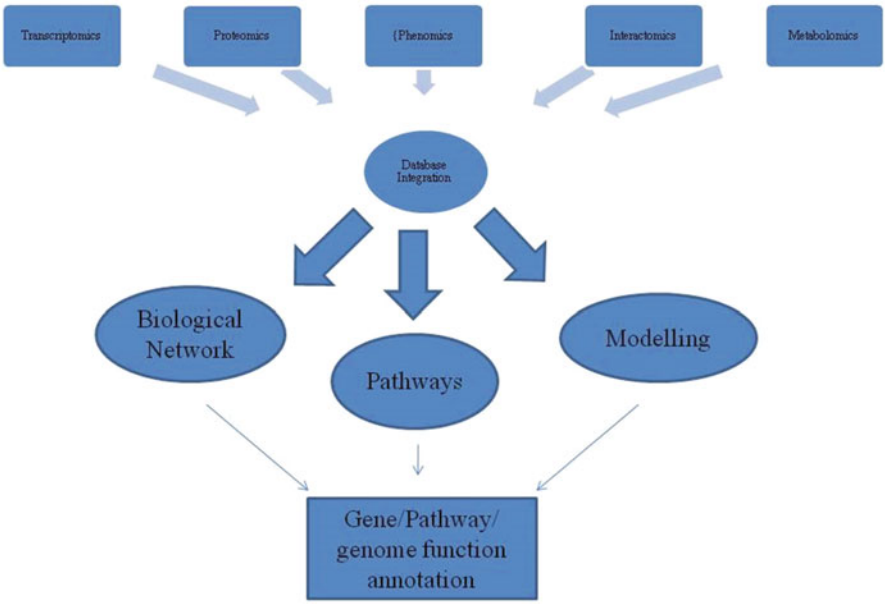

Fig. 4.5 Integrative approach using omics and systems biology to understand the plants system during stress conditions

4.8 Gene to Metabolite Networks

Gene to metabolite networks is made on the basis of the coefficient of correlation between genes. The way of interaction is represented on the basis of the correlation value. A different biological process, molecular functions, and gene functions can be explored to understand plants systems biology during normal and experimental conditions. Various studies on different plants have been performed to understand the gene to metabolite network and candidate gene responsible for over- or under-production of secondary metabolites were identified during stress and normal conditions (Goossens 2003; Scheible 2004; Zulak et al. 2007).

4.9 Protein-Protein Interaction Networks

In PPI networks, nodes are represented by the proteins and edges represents the physical or genetic interaction. The function of genes is explored on the basis of genetic interaction between the proteins (Boone et al. 2007) while physical interactions are used to understand protein-protein interaction and dimer formation (de Folter 2005). Analysis on the basis of biological networks construction in chickpea during cadmium and chromium exposure has been reported that most of the hub genes are involved in protein dimerization (Yadav and Mani 2018).

4.10 Transcriptional Regulatory Networks

Interaction of the transcription factor and downstream gene are studied in this type of networks. Here nodes are transcription factor or regulatory genes and represent the activation and deactivation (Babu et al. 2004). One transcription factor interacts with a large number of genes simultaneously. Different studies have been performed to understand the stress response in plants using transcriptional regulatory networks (Nakashima et al. 2009; Yun et al. 2010; Todaka et al. 2012).

4.11 Gene Regulatory Networks

In gene regulatory networks, node represents the genes/mRNA/proteins and edges are regulatory interaction like activation, repression, inhibition, or functional interactions (Long et al. 2008). Various studies reported that gene regulatory network having an important role to understand the underlying mechanisms during developmental and stress condition on the basis of gene regulatory network in different plants (Li et al. 2006; Meng et al. 2011; Pires et al. 2013). There are lots of tools available for construction and analysis of biological network in plants which are shown in Table 4.3.

Table 4.3 List of frequently used software and online tools used for construction of biological networks and their analysis using genomics data in plants

Software/tools	Description	Reference
Cytoscape	It is widely used software to construct, visualize, and analyze the biological network	Shannon, et al. (2003)
Genemania	It is an online tool for construction and analysis of biological network	Franz et al. (2018)
Cell designer	A software for analyzing the biological network	Funahashi et al. (2006)
Networks	Software used for biological network analysis	Team (2014)
Medusa	Software used for visualization of small network	Bosi et al. (2015)
BioLayout Express3D	A tool for biological network visualization and analysis	Wright et al. (2014)
String	An online tool for construction and visualization of a biological network	Szklarczyk et al. (2017)

4.12 Conclusion

In upcoming years, plants will be a solution of all problems like water and food scarcity. So, it is very important to understand the agroecosystem; for this one should know about functional genomics and systems biology. The functional genomics and systems biology can be used to develop a strategy to escape the plants from stress condition and new variety can also be developed. In addition to advantage, there are certain challenges like big data handling and its analysis. Due to the complexity of plants, wet lab experiment is not always possible; hence there is a need of lots of computational approaches that can be applied to functional genomics and systems biology to understand the agroecosystem in a well-defined manner.

References

Aghaei K, Komatsu S (2013) Crop and medicinal plants proteomics in response to salt stress. Front Plant Sci 4. https://doi.org/10.3389/fpls.2013.00008

Aguan K, Carvajal JA, Thompson LP, Weiner CP (2000) Application of a functional genomics approach to identify differentially expressed genes in human myometrium during pregnancy and labour. Mol Hum Reprod 6:1141–1145

Alam I, Lee DG, Kim KH et al (2010a) Proteome analysis of soybean roots under waterlogging stress at an early vegetative stage. J Biosci 35:49–62. https://doi.org/10.1007/s12038-010-0007-5

Alam I, Sharmin SA, Kim KH et al (2010b) Proteome analysis of soybean roots subjected to short-term drought stress. Plant Soil 333:491–505. https://doi.org/10.1007/s11104-010-0365-7

Babu MM, Luscombe NM, Aravind L et al (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14:283–291

Bailey NJC, Oven M, Holmes E et al (2003) Metabolomic analysis of the consequences of cadmium exposure in Silene cucubalus cell cultures via 1H NMR spectroscopy and chemometrics. Phytochemistry 62:851–858. https://doi.org/10.1016/S0031-9422(02)00719-7

Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. Cell Res 21:381–395. https://doi.org/10.1038/cr.2011.22

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113

Baxter CJ, Redestig H, Schauer N et al (2006) The metabolic response of heterotrophic Arabidopsis cells to oxidative stress. Plant Physiol 143:312–325. https://doi.org/10.1104/pp.106.090431

Bebber DP, Ramotowski MAT, Gurr SJ (2013) Crop pests and pathogens move polewards in a warming world. Nat Clim Chang 3:985–988. https://doi.org/10.1038/nclimate1990

Ben-Amar A, Daldoul S, Reustle GM et al (2016) Reverse genetics and high throughput sequencing methodologies for plant functional genomics. Curr Genomics 17:460–475. https://doi.org/10.2174/1389202917666160520102827

Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. Nat Rev Genet 8:437–449

Bosi E, Donati B, Galardini M et al (2015) MeDuSa: a multi-draft based scaffolder. Bioinformatics 31:2443–2451. https://doi.org/10.1093/bioinformatics/btv171

Bovik A (2005) Handbook of image and video processing

Bunnik EM, Le Roch KG (2013) An introduction to functional genomics and systems biology. Adv Wound Care 2:490–498. https://doi.org/10.1089/wound.2012.0379

Caruso G, Cavaliere C, Guarino C et al (2008) Identification of changes in Triticum durum L. leaf proteome in response to salt stress by two-dimensional electrophoresis and MALDI-TOF mass spectrometry. Anal Bioanal Chem 391:381–390. https://doi.org/10.1007/s00216-008-2008-x

Caspi R (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 34:D511–D516. https://doi.org/10.1093/nar/gkj128

Cramer GR, Ergül A, Grimplet J et al (2007) Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. Funct Integr Genomics 7:111–134. https://doi.org/10.1007/s10142-006-0039-y

Cramer GR, Van Sluyter SC, Hopper DW et al (2013) Proteomic analysis indicates massive changes in metabolism prior to the inhibition of growth and photosynthesis of grapevine (Vitis vinifera L.) in response to water deficit. BMC Plant Biol 13. https://doi.org/10.1186/1471-2229-13-49

D'Alessandro A, Taamalli M, Gevi F et al (2013) Cadmium stress responses in Brassica juncea: hints from proteomics and metabolomics. J Proteome Res 12:4979–4997. https://doi.org/10.1021/pr400793e

de Folter S (2005) Comprehensive interaction map of the Arabidopsis MADS box transcription factors. Plant Cell Online 17:1424–1433. https://doi.org/10.1105/tpc.105.031831

Dhondt S, Wuyts N, Inzé D (2013) Cell to whole-plant phenotyping: the best is yet to come. Trends Plant Sci 18:1360–1385

Dubey S, Misra P, Dwivedi S et al (2010) Transcriptomic and metabolomic shifts in rice roots in response to Cr (VI) stress. BMC Genomics 11. https://doi.org/10.1186/1471-2164-11-648

Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 48:155–171. https://doi.org/10.1023/A:1013713905833

Franz M, Rodriguez H, Lopes C et al (2018) GeneMANIA update 2018. Nucleic Acids Res 46:W60–W64. https://doi.org/10.1093/nar/gky311

Fu S-F, Chen P-Y, Nguyen QTT et al (2014) Transcriptome profiling of genes and pathways associated with arsenic toxicity and tolerance in Arabidopsis. BMC Plant Biol 14:94. https://doi.org/10.1186/1471-2229-14-94

Funahashi A, Matsuoka Y, Jouraku A, et al (2006) Celldesigner: a modeling tool for biochemical networks. In: Proceedings – Winter Simulation Conference. pp 1707–1712

Furbank RT, Tester M (2011) Phenomics - technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16:635–644

Gans C (2005) Checklist and bibliography of the amphisbaenia of the world. Bull Am Museum Nat Hist 289:1. https://doi.org/10.1206/0003-0090(2005)289<0001:CABOTA>2.0.CO;2

Gasperskaja E, Kučinskas V (2017) The most common technologies and tools for functional genome analysis. Acta medica Litu 24:1–11. https://doi.org/10.6001/actamedica.v24i1.3457
Geisler-Lee J, O'Toole N, Ammar R et al (2007) A predicted Interactome for Arabidopsis. Plant Physiol 145:317–329. https://doi.org/10.1104/pp.107.103465
Ghatak A (2017) Machine learning with R. Springer, Singapore
Ghosh D, Xu J (2014) Abiotic stress responses in plant roots: a proteomics perspective. Front Plant Sci 5. https://doi.org/10.3389/fpls.2014.00006
Gibson G, Muse SV (2009) Primer of genome science. Sinauer Associates, Sunderland
Gong F, Hu X, Wang W (2015) Proteomic analysis of crop plants under abiotic stress conditions: where to focus our research? Front Plant Sci 6. https://doi.org/10.3389/fpls.2015.00418
Goossens A (2003) Secretion of secondary metabolites by ATP-binding cassette transporters in plant cell suspension cultures. Plant Physiol 131:1161–1164. https://doi.org/10.1104/pp.102.016329
Großkinsky DK, Svensgaard J, Christensen S, Roitsch T (2015) Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. J Exp Bot 66:5429–5440
Halket JM, Waterman D, Przyborowska AM, et al (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. In: J Exp Botany 56(410). pp 219–243
Hamuda E, Glavin M, Jones E (2016) A survey of image processing techniques for plant extraction and segmentation in the field. Comput Electron Agric 125:184–199
Hashimshony T, Wagner F, Sher N, Yanai I (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep 2:666–673. https://doi.org/10.1016/j.celrep.2012.08.003
Herbette S, Taconnat L, Hugouvieux V et al (2006) Genome-wide transcriptome profiling of the early cadmium response of Arabidopsis roots and shoots. Biochimie 88:1751–1765. https://doi.org/10.1016/j.biochi.2006.04.018
Hossain Z, Nouri MZ, Komatsu S (2012) Plant cell organelle proteomics in response to abiotic stress. J Proteome Res 11:37–48
Huang T-L, Nguyen QTT, Fu S-F et al (2012) Transcriptomic changes and signalling pathways induced by arsenic stress in rice roots. Plant Mol Biol 80:587–608. https://doi.org/10.1007/s11103-012-9969-z
Joyce AR, Palsson BØ (2006) The model organism as a system: integrating "omics" data sets. Nat Rev Mol Cell Biol 7:198–210
Kanehisa M, Furumichi M, Tanabe M et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361. https://doi.org/10.1093/nar/gkw1092
Kell DB (2002) Genotype - phenotype mapping: genes as computer programs. Trends Genet 18:555–559
Kintlová M, Blavet N, Cegan R, Hobza R (2017) Transcriptome of barley under three different heavy metal stress reaction. Genomics Data 13:15–17. https://doi.org/10.1016/j.gdata.2017.05.016
Klose R, Penlington J, Ruckelshausen A (2011) Usability of 3D time-of-flight cameras for automatic plant phenotyping. Bornimer Agrartech Berichte 69:93–105. https://doi.org/10.1103/PhysRevC.92.024618
Komatsu S, Yamada E, Furukawa K (2009) Cold stress changes the concanavalin A-positive glycosylation pattern of proteins expressed in the basal parts of rice leaf sheaths. Amino Acids 36:115–123. https://doi.org/10.1007/s00726-008-0039-4
Komatsu S, Sugimoto T, Hoshino T et al (2010) Identification of flooding stress responsible cascades in root and hypocotyl of soybean using proteome analysis. Amino Acids 38:729–738. https://doi.org/10.1007/s00726-009-0277-0
Komatsu S, Makino T, Yasue H (2013a) Proteomic and biochemical analyses of the cotyledon and root of flooding-stressed soybean plants. PLoS One 8(6). https://doi.org/10.1371/journal.pone.0065301

Komatsu S, Nanjo Y, Nishimura M (2013b) Proteomic analysis of the flooding tolerance mechanism in mutant soybean. J Proteome 79:231–250. https://doi.org/10.1016/j.jprot.2012.12.023

Komatsu S, Nakamura T, Sugimoto Y, Sakamoto K (2014) Proteomic and Metabolomic analyses of soybean root tips under flooding stress. Protein Pept Lett 21:865–884. https://doi.org/10.2174/0929866521666140320110521

Kumar Yadav R, Srivastava SK (2015) Effect of Arsenite and arsenate on lipid peroxidation, Enzymatic and Non-Enzymatic Antioxidants in Zea mays Linn Biochem Physiol Open Access 4:. https://doi.org/10.4172/2168-9652.1000186

Kumar D, Chapagai D, Dean P, Davenport M (2015) Biotic and abiotic stress signaling mediated by salicylic acid. In: Elucidation of abiotic stress signaling in plants: functional genomics perspectives. pp 329–346

Lafuente A, Pérez-Palacios P, Doukkali B et al (2015) Unraveling the effect of arsenic on the model Medicago-Ensifer interaction: a transcriptomic meta-analysis. New Phytol 205:255–272. https://doi.org/10.1111/nph.13009

Lange BM, Ghassemian M (2005) Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. Phytochemistry 66:413–451

Le Lay P, Isaure MP, Sarry JE et al (2006) Metabolomic, proteomic and biophysical analyses of Arabidopsis thaliana cells exposed to a caesium stress. Influence Potassium Supply Biochimie 88:1533–1547. https://doi.org/10.1016/j.biochi.2006.03.013

Lenco M (1982) Remote sensing and natural resources. Nat Resour 18:2–9

Li S, Assmann SM, Albert R (2006) Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. PLoS Biol 4:1732–1748. https://doi.org/10.1371/journal.pbio.0040312

Long TA, Brady SM, Benfey PN (2008) Systems approaches to identifying gene regulatory networks in plants. Annu Rev Cell Dev Biol 24:81–103. https://doi.org/10.1146/annurev.cellbio.24.110707.175408

Lowe R, Shirley N, Bleackley M et al (2017) Transcriptomics technologies. PLoS Comput Biol 13. https://doi.org/10.1371/journal.pcbi.1005457

Marioni JC, Mason CE, Mane SM et al (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517. https://doi.org/10.1101/gr.079558.108

McLaughlin MJ, Zarcinas BA, Stevens DP, Cook N (2000) Soil testing for heavy metals. Commun Soil Sci Plant Anal 31:1661–1700. https://doi.org/10.1080/00103620009370531

Meng Y, Shao C, Wang H, Chen M (2011) The regulatory activities of plant MicroRNAs: a more dynamic perspective. Plant Physiol 157:1583–1595. https://doi.org/10.1104/pp.111.187088

Mirzaei M, Soltani N, Sarhadi E et al (2012) Shotgun proteomic analysis of long-distance drought signaling in rice roots. J Proteome Res 11:348–358. https://doi.org/10.1021/pr2008779

Mohammadi PP, Moieni A, Hiraga S, Komatsu S (2012) Organ-specific proteomic analysis of drought-stressed soybean seedlings. J Proteome 75:1906–1923. https://doi.org/10.1016/j.jprot.2011.12.041

Mushegian AA (2017) Stress signals in plants. Sci Signal 10

Nakashima K, Ito Y, Yamaguchi-Shinozaki K (2009) Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. Plant Physiol 149:88–95. https://doi.org/10.1104/pp.108.129791

Nam MH, Huh SM, Kim KM et al (2012) Comparative proteomic analysis of early salt stress-responsive proteins in roots of SnRK2 transgenic rice. Proteome Sci 10. https://doi.org/10.1186/1477-5956-10-25

Oono Y, Yazawa T, Kanamori H et al (2016) Genome-wide transcriptome analysis of cadmium stress in rice. Biomed Res Int 2016. https://doi.org/10.1155/2016/9739505

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12:87–98. https://doi.org/10.1038/nrg2934

Padhi J, Misra RK, Payero JO (2012) Estimation of soil water deficit in an irrigated cotton field with infrared thermography. F Crop Res 126:45–55. https://doi.org/10.1016/j.fcr.2011.09.015

Paley SM, Karp PD (2006) The pathway tools cellular overview diagram and Omics viewer. Nucleic Acids Res 34:3771–3778. https://doi.org/10.1093/nar/gkl334
Pandey A, Sharma M, Pandey GK (2015) Small and large G proteins in biotic and abiotic stress responses in plants. In: Elucidation of abiotic stress signaling in plants: functional genomics perspectives. pp 231–270
Pandey GK, Pandey A, Prasad M, Böhmer M (2016) Editorial: abiotic stress signaling in plants: functional genomic intervention. Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.00681
Paulose B, Kandasamy S, Dhankher OP (2010) Expression profiling of Crambe abyssinica under arsenate stress identifies genes and gene networks involved in arsenic metabolism and detoxification. BMC Plant Biol 10. https://doi.org/10.1186/1471-2229-10-108
Pérez-Clemente RM, Vives V, Zandalinas SI et al (2013) Biotechnological approaches to study plant responses to stress. Biomed Res Int 2013
Piétu G, Mariage-Samson R, Fayein NA et al (1999) The genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. Genome Res 9:195–209. https://doi.org/10.1101/gr.9.2.195
Pires ND, Yi K, Breuninger H et al (2013) Recruitment and remodeling of an ancient gene regulatory network during land plant evolution. Proc Natl Acad Sci 110:9571–9576. https://doi.org/10.1073/pnas.1305457110
Ptaszek M (2013) Progress in molecular biology and translational science. In: Progress in molecular biology and translational science. pp 59–108
Ramegowda V, Senthil-Kumar M (2015) The interactive effects of simultaneous biotic and abiotic stresses on plants: mechanistic understanding from drought and pathogen combination. J Plant Physiol 176:47–54
Rao VS, Srinivas K, Sujini GN, Kumar GNS (2014) Protein-protein interaction detection: methods and analysis. Int J Proteomics 2014:1–12. https://doi.org/10.1155/2014/147648
Rizhsky L (2004) When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. Plant Physiol 134:1683–1696. https://doi.org/10.1104/pp.103.033431
Rollins JA, Habte E, Templer SE et al (2013) Leaf proteome alterations in the context of physiological and morphological responses to drought and heat stress in barley (Hordeum vulgare L.). J Exp Bot 64:3201–3212. https://doi.org/10.1093/jxb/ert158
Saito K, Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 61:463–489. https://doi.org/10.1146/annurev.arplant.043008.092035
Scheible W-R (2004) Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of Arabidopsis in response to nitrogen. Plant Physiol 136:2483–2499. https://doi.org/10.1104/pp.104.047019
Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470. https://doi.org/10.1126/science.270.5235.467
Schuster SC (2007) Next-generation sequencing transforms today's biology. Nat Methods 5:16–18. https://doi.org/10.1038/nmeth1156
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski BIT (2003) Cytoscape: an open source platform for complex network analysis and visualization. Genome Res. http://www.cytoscape.org/
Shendure J, Aiden EL (2012) The expanding scope of DNA sequencing. Nat Biotechnol 30:1084–1094. https://doi.org/10.1038/nbt.2421
Shulaev V (2006) Metabolomics technology and bioinformatics. Brief Bioinform 7:128–139
Singh V, Misra AK (2017) Detection of plant leaf diseases using image segmentation and soft computing techniques. Inf Process Agric 4:41–49. https://doi.org/10.1016/j.inpa.2016.10.005
Singh A, Ganapathysubramanian B, Singh AK, Sarkar S (2016) Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci 21:110–124
Stefano GB (2014) Comparing Bioinformatic gene expression profiling methods: microarray and RNA-Seq. Med Sci Monit Basic Res 20:138–142. https://doi.org/10.12659/MSMBR.892101

Su YH, McGrath SP, Zhao FJ (2010) Rice is more efficient in arsenite uptake and translocation than wheat and barley. Plant Soil 328:27–34. https://doi.org/10.1007/s11104-009-0074-2

Szklarczyk D, Morris JH, Cook H et al (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45:D362–D368. https://doi.org/10.1093/nar/gkw937

Team N (2014) NetworkX. networkx.github.io

Thimm O, Bläsing O, Gibon Y et al (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939. https://doi.org/10.1111/j.1365-313X.2004.02016.x

Todaka D, Nakashima K, Shinozaki K, Yamaguchi-Shinozaki K (2012) Toward understanding transcriptional regulatory networks in abiotic stress responses and tolerance in rice. Rice 5:1–9

Tokimatsu T (2005) KaPPA-view. A web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. Plant Physiol 138:1289–1300. https://doi.org/10.1104/pp.105.060525

Uhrig JF (2006) Protein interaction networks in plants. Planta 224:771–781

Van Der Heijden G, Song Y, Horgan G et al (2012) SPICY: towards automated phenotyping of large pepper plants in the greenhouse. Funct Plant Biol 39:870–877. https://doi.org/10.1071/FP12019

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487. https://doi.org/10.1126/science.270.5235.484

Wang R, Gao F, Guo BQ et al (2013) Short-term chromium-stress-induced alterations in the maize leaf proteome. Int J Mol Sci 14:11125–11144. https://doi.org/10.3390/ijms140611125

Washburn MP, Wolters D, Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19:242–247. https://doi.org/10.1038/85686

Weinstein JN (1998) Fishing expeditions. Science 282:627

Williams EJB, Bowles DJ (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res 14:1060–1067. https://doi.org/10.1101/gr.2131104

Witzel K, Weidner A, Surabhi GK et al (2009) Salt stress-induced alterations in the root proteome of barley genotypes with contrasting response towards salinity. J Exp Bot 60:3545–3557. https://doi.org/10.1093/jxb/erp198

Wright DW, Angus T, Enright AJ, Freeman TC (2014) Visualisation of BioPAX networks using BioLayout Express3D. F1000Research. https://doi.org/10.12688/f1000research.5499.1

Wurtele ES, Li J, Diao L et al (2003) MetNet: software to build and model the biogenetic lattice of Arabidopsis. Comp Funct Genomics 4:239–245. https://doi.org/10.1002/cfg.285

Xuan J, Song Y, Zhang H et al (2013) Comparative proteomic analysis of the stolon cold stress response between the C4 perennial grass species Zoysia japonica and Zoysia metrella. PLoS One 8. https://doi.org/10.1371/journal.pone.0075705

Yadav BS, Mani A (2018) Analysis of bHLH coding genes of Cicer arietinum during heavy metal stress using biological network. Physiol Mol Biol Plants. https://doi.org/10.1007/s12298-018-0625-1

Yuan JS, Galbraith DW, Dai SY et al (2008) Plant systems biology comes of age. Trends Plant Sci 13:165–171

Yue R, Lu C, Qi J et al (2016) Transcriptome analysis of cadmium-treated roots in maize (Zea mays L.). Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.01298

Yun KY, Park MR, Mohanty B et al (2010) Transcriptional regulatory network triggered by oxidative signals configures the early response mechanisms of japonica rice to chilling stress. BMC Plant Biol 10. https://doi.org/10.1186/1471-2229-10-16

Zhang P (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant Physiol 138:27–37. https://doi.org/10.1104/pp.105.060376

Zhang X, Jafari N, Barnes RB et al (2005) Studies of gene expression in human cumulus cells indicate pentraxin 3 as a possible marker for oocyte quality. Fertil Steril 83:1169–1179. https://doi.org/10.1016/j.fertnstert.2004.11.030

Zhang H, Ni Z, Chen Q et al (2016) Proteomic responses of drought-tolerant and drought-sensitive cotton varieties to drought stress. Mol Gen Genomics 291:1293–1303. https://doi.org/10.1007/s00438-016-1188-x

Zhu Z, Chen J, Zheng HL (2012) Physiological and proteomic characterization of salt tolerance in a mangrove plant, Bruguiera gymnorrhiza (L.) Lam. Tree Physiol 32:1378–1388. https://doi.org/10.1093/treephys/tps097

Zilberman D, Henikoff S (2007) Genome-wide analysis of DNA methylation patterns. Development 134:3959–3965. https://doi.org/10.1242/dev.001131

Zulak KG, Cornish A, Daskalchuk TE et al (2007) Gene transcript and metabolite profiling of elicitor-induced opium poppy cell cultures reveals the coordinate regulation of primary and secondary metabolism. Planta 225:1085–1106. https://doi.org/10.1007/s00425-006-0419-5

Zulfiqar A, Paulose B, Chhikara S, Dhankher OP (2011) Identifying genes and gene networks involved in chromium metabolism and detoxification in Crambe abyssinica. Environ Pollut 159:3123–3128. https://doi.org/10.1016/j.envpol.2011.06.027

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).

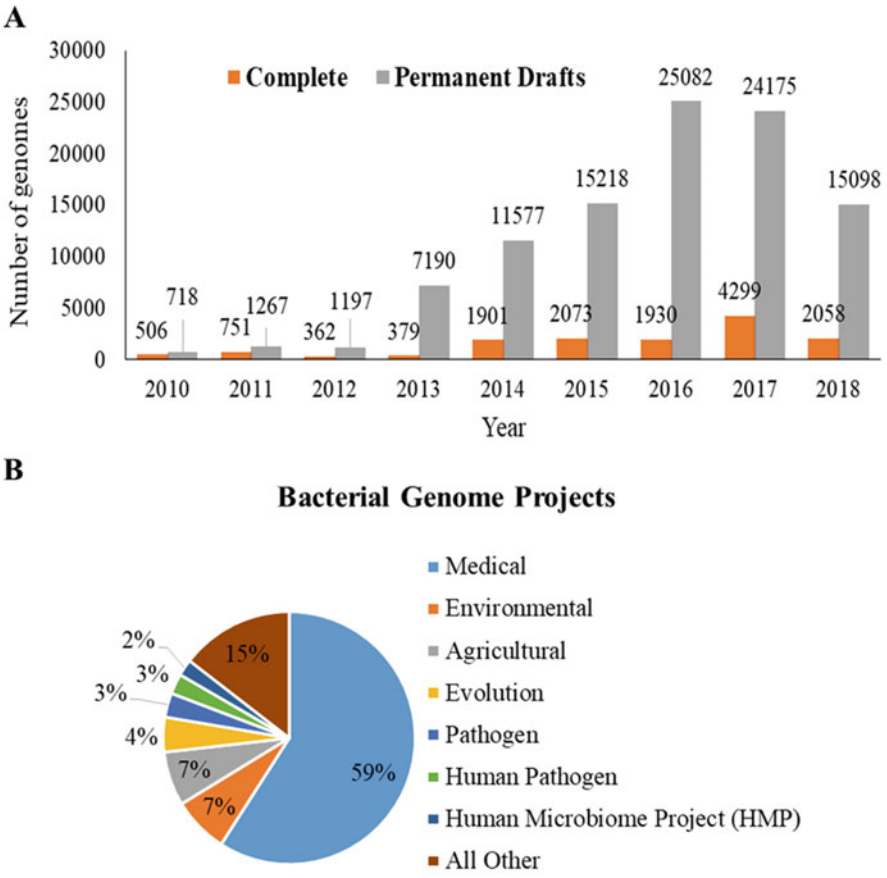

Fig. 5.1 The number of complete and permanent draft genomes (**a**) and projects' relevance to bacterial genome (**b**) in GOLD. Presented data accessed on December 26, 2018, from https://gold.jgi.doe.gov/

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	Availability	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 communities. 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

6 Bioinformatics and Microarray-Based Technologies to Viral Genome Sequence Analysis

Mayank Pokhriyal, Barkha Ratta, and Brijesh S. Yadav

Abstract

Identification of microbial pathogen is an important event which lead to diagnosis, treatment, and control of infections produce by them. The high-throughput technology like microarray and new-generation sequencing machine are able to generate huge amount of nucleotide sequences of viral and bacterial genome of both known and unknown pathogens. Few years ago it was the DNA microarrays which had great potential to screen all the known pathogens and yet to be identified pathogen simultaneously. But after the development of a new generation sequencing, technologies and advance computational approach researchers are looking forward for a complete understanding of microbes and host interactions. The powerful sequencing platform is rapidly transforming the landscape of microbial identification and characterization. As bioinformatics analysis tools and databases are easily available to researchers, the enormous amount of data generated can be meaningfully handled for better understanding of the microbial world. Here in this chapter, we present commentary on how the computational method incorporated with sequencing technique made easy for microbial detection and characterization.

M. Pokhriyal
Viral Research and Diagnostic Laboratory (VRDL), Government Medical College, Haldwani, Uttarakhand, India

B. Ratta
Division of Biochemistry, Indian Veterinary Research Institute, Bareilly, Uttar Pradesh, India

B. S. Yadav (✉)
Center of Bioengineering Research, University of Information Science and Technology (UIST), St. Paul, Republic of North Macedonia
e-mail: brijeshbioinfo@gmail.com

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_6

6.1 Introduction

6.1.1 Importance of Microbes

Microorganisms like viruses, bacteria, and fungus have evolved to survive in every type of conditions on our planet including human and animal bodies. Although many are not harmful, a few cause life-threatening diseases. Traditionally these are identified by culturing in appropriate media, biochemical analysis, and serological testing. However, more numbers of microbes are yet to be characterized than are known. Through the human body, microbes play vital roles in all of their ecosystems (Hentges 1993). In spite of being extremely small, the sheer numbers of microbes living on the planet have large effects on the cycling of nutrients and compounds, essential for the survival of all organisms.

Microbes are encountered in all walks of human life and there prevails a constant interaction. The vast majority of the bacteria in the body are rendered harmless by the protective effects of the immune system, and a few are beneficial. In fact, the relationship between microbes and humans is delicate and complex. Ten times as many microbes live on or inside your body. The microbes living in our digestive system break down food and produce useful vitamins. The millions of microbes that coat our skin and intestinal lumen form a protective barrier against more dangerous microbes. Without them, our bodies would be open to microbial attack. In spite of these benefits, a relatively small number of microbes are harmful to humans. Many diseases and epidemics are caused by microbes: the plague during the Middle Ages, smallpox, AIDS, influenza, food poisoning, and anthrax. These diseases result in severe illness, or even death. As scientists learn more about bacteria, fungi, and viruses, they are better able to treat and prevent these diseases (Yadav et al. 2013).

6.1.2 Viral Diseases in Humans and Livestock

Viral diseases of livestock can be devastating both to farmers and the wider community. Rinderpest, a disease of cattle caused by rinderpest virus, spread rapidly across Africa by 1892 and led to death of nearly 95% of the cattle in East Africa. In the early years of the twentieth century, rinderpest was common in Asia and parts of Europe, and its prevalence increased in Asia. In 1957, Thailand had to appeal for aid because many buffaloes had died, due to which paddy fields could not be prepared for rice. In India, spread of rinderpest was controlled by numerous eradication programs throughout the twentieth century (Barrett et al. 2006). As a result rinderpest was eradicated from India in the year 1995 (Barrett et al. 2006). In human and animals, viral diseases keep emerging very frequently.

Research on various aspects of animal diseases particularly diagnosis is carried out regularly to develop effective measures to guarantee animal health through containment of emerging diseases. Many viral diseases have been eradicated from globe by taking effective control measures. Smallpox was the first viral disease to be eradicated globally followed by Rinderpest or cattle plague. In the last 30 years or so,

about 40 new viral diseases have been identified (Zappa et al. 2009). A recent study has shown that by 2020, about 10 new human diseases would emerge compared to the past, and mobility of humans and animals have increased; with this, rate of spread of disease has also increased (Morens and Fauci 2013). It has become vital to have an arsenal of diagnostic techniques which can identify a pathogen whether new or old in shortest possible time to combat effective control measures. New technologies are being developed in molecular biology at a very rapid rate and many of them are being applied in diagnosis. PCR (polymerase chain reaction) is once such technique which was developed for amplification of a fragment of DNA in the early 1980s (Mullis 1990). The PCR and related techniques like real-time PCR are now the most widely used molecular method for the detection and identification of viruses (Espy et al. 2006). Multiplex PCR allows detection of more than one species of virus in a single assay (Elnifro et al. 2000). However, these techniques have limitations of multiplexibility and versatility and require extensive prior knowledge of the sequences to be amplified (Miller et al. 2009).

Another technique, microarray, developed in mid-1990s (Kostrzynska and Bachand 2006) for studying gene expression, is now being increasingly applied for diagnostics. Microarray consists of thousands to millions of oligonucleotide probes representing different genes of known or unknown pathogens deposited or directly synthesized on a surface in an ordered fashion (Anderson et al. 2008; Manoj 2009). These numbers give huge capability to microarray to detect and quantitate hundreds and thousands of genes simultaneously. The capability of microarray to detect a large number of genes was used for diagnosis at the beginning of this century when the SARS corona virus was identified using a diagnostic microarray chip (Wang et al. 2002). Since then several microarray chips have been developed and tested (Wang et al. 2002, 2003; Martín et al. 2006; Chiu et al. 2006; Quan et al. 2007; Gardner et al. 2010; Yadav et al. 2015).

6.2 Virus Identification Methods

A virus is identified by showing presence of its proteins or nucleic acids or via an immunological response to the host. Each method has its own advantages and drawbacks with respect to sensitivity, specificity, efficiency, and feasibility. Depending on the virus type, concentration, and circumstances under which the viral sample was collected, certain methods may be more effective than others. The most commonly used methods to detect and identify different viruses can be broadly divided into four categories, that is:

A. Electron microscopy: A direct method for detecting virus requires purified and high concentration of virus, but only experienced technician then discerns the virus by its physical structure features. Viruses such as poxviruses and herpes viruses can be easily identified using this technology (Nii 1971).
B. Virus isolation in cell culture: Another method is to grow the virus in cell cultures (Leland and Christine 2007) and observe for virus-induced changes such as cell

rounding, disorientation, swelling, shrinking, or death in the cells. This method is time-consuming, has low sensitivity, and also cannot be used for viruses like hepatitis B, parvovirus, and papillomavirus, which cannot grow in cell culture (Goldsmith and Miller 2009).

C. Immunofluorescence-based assays: ELISA (enzyme-linked immunosorbent assay) (Voller et al. 1978), one of the widely used methods for detecting viruses, relies on the presence of antigens or antibodies in bodily fluids.
D. Molecular techniques: Molecular biology techniques based on nucleic acid sequences are advanced and much faster than any other techniques and are now mostly used for virus diagnosis (Kreuze et al. 2009).

6.2.1 Nucleic Acid Sequence-Based Diagnosis

Large-scale availability of genomic and nucleotide sequences of disease-causing agents in public databases (GenBank of the USA, European molecular biology laboratory of Europe, and DNA databank of Japan) and progress in nucleic acid amplification techniques have enabled application of nucleic acid-based techniques for pathogen identification. These techniques include polymerase chain reaction (PCR) (Mullis 1990), loop-mediated isothermal amplification (LAMP) (Tomita et al. 2008), ligase chain reaction, nucleic acid sequence-based amplification/isothermal amplification (NASBA, (Liu et al. 2006), strand displacement amplification (Su et al. 2013), Qb replicase method (An et al. 2003), and branched DNA probes (Baumeister et al. 2012). The PCR is one of the most commonly used methods for virus detection (Cunningham 2004). Real-time PCR, an innovation of PCR, is one of the most sensitive methods of detecting pathogens (Mackay et al. 2002). Both PCR and real-time PCR have limited multiplexing capability.

Like PCR and real-time PCR, DNA microarray is a genome-based method which was originally developed for studying variation in gene expression but has since been adapted for pathogen detection. It has huge multiplex capability and can screen for all the known pathogen in one experiment (Wilson et al. 2002; Bryant et al. 2004). The DNA microarray also called DNA chip has high sensitivity and specificity (Kostrzynska and Bachand 2006). The microarray chip used for diagnostic purpose contains thousands of different oligonucleotide probes specific to respective pathogens. The probes designed for diagnostic assays are unique to a specific pathogen with respect to all the other pathogen genomes and also to host other nonspecific genome sequences present in the clinical samples. Specially designed software are used for designing probes and also for data analysis.

6.2.2 DNA Microarray-Based Virus Identification

DNA microarray has been used efficiently in clinical diagnostics for identifying disease-related genes with the help of its biomarkers (Loy and Bodrossy 2006) and also for disease diagnosis. The microarray-based virus diagnosis started at the

beginning of this century (Schena et al. 1995). Microarrays based on oligonucleotide probes representing nucleic acid sequences conserved between members of a taxonomic group were first used for detection of the then unknown SARS coronavirus (Wang et al. 2002, 2003) and since then has been used for detection of many viruses (Chiu et al. 2007; Chou et al. 2006; Quan et al. 2007; Martín et al. 2006, Chen et al. 2010). The detection of multiple rhinovirus serotypes in cell culture and clinical specimen (Wang et al. 2002), papillomavirus in cervical lesions (An et al. 2003), parainfluenza virus 4 in nasopharyngeal aspirates (Chiu et al. 2006), influenza virus from nasal wash and throat swabs (Lin et al. 2006), gammaretrovirus in prostate tumors (Urisman et al. 2006), foot and mouth disease virus from animal tissue (Martín et al. 2006), coronaviruses and rhinoviruses from nasal lavage (Kistler et al. 2007), metapneumovirus from bronchoalveolar lavage (Chiu et al. 2007), different respiratory pathogens including influenza virus and non-influenza agents in nasal swabs and lung tissue (Quan et al. 2007), and common food born viruses such as coxsackievirus, hepatitis A virus, norovirus, and rotavirus identified using tiling microarray (Chen et al. 2010). Grubaugh et al. (2013) identified 13 of 14 flaviviruses (*Culex* flavivirus, dengue-3, and Japanese encephalitis viruses) using microarray platform.

6.2.3 Animal Virus Diagnosis Using Microarray

Microarray technique is one of the most recent diagnostics in veterinary field (Feilotter 2004). Jack et al. (2009) developed a microarray assay for identifying viruses that cause vesicular or vesicular-like lesions in livestock animals. They were able to differentiate foot and mouth disease virus (FMDV), vesicular stomatitis virus (VSV), swine vesicular disease virus, vesicular exanthema of swine virus (VESV), BHV-1, orf virus, pseudocowpox virus, bluetongue virus serotype 1, and bovine viral diarrhea virus 1 (BVDV1). Leblanc and co-researcher (2009) used magnetic bead microarray for the rapid detection and identification of the four recognized species in the *Pestivirus* genus of the *Flaviviridae* family, i.e., classical swine fever virus, border disease virus, and BVDV1 and 2, which allowed specific and sensitive virus detection. They concluded that based on the simplicity of the assay, the protocols for hybridization and magnetic bead detection offer an emerging application for molecular diagnosis in virology that is amenable for use in a modestly equipped laboratory. Porcine reproductive and respiratory syndrome virus (PRRSV) and foot and mouth disease virus (FMDV) were detected in a cDNA microarray (Liu et al. 2006). GreenChip array facilitated the discovery of Ebola virus, in the porcine respiratory illness outbreak in the Philippines (Barrette et al. 2009). These chips have also been used for screening veterinary clinical samples (Mihindukulasuriya et al. 2008). Canine coronavirus (CCoV), feline infectious peritonitis virus (FIPV), feline coronavirus (FCoV), bovine coronavirus (BCoV), porcine respiratory coronavirus (PRCoV), turkey enteritis coronavirus (TCoV), transmissible gastroenteritis virus (TGEV), and human respiratory coronavirus (HRCoV) are identified based on microarray hybridization (Chen et al. 2010). Sharma et al. (2012) have designed

microarray chip for identification of animal viruses; the chip successfully identified the new castle disease virus in sheep and mixed infection of bovine viral diarrhea 2 and bovine herpes virus 1 (Ratta et al. 2013) in cattle.

6.3 Basic Concept of DNA Microarray

Since its inception DNA microarray has advanced biological sciences more profoundly than any other technique (Benedetti et al. 2000; Rockett and Dix 2000; Staudt and Brown 2000; Brown and Botstein 1999). The basic principle involved in DNA microarrays is the reverse of Southern blot (Southern 1975). Unlike Southern blot where targets are immobilized and a probe is labeled, in DNA microarray, probes are immobilized on a membrane and then hybridized against the labeled target population (Kurian et al. 1999). Since unlike Southern blot, in microarray, target is labeled not probes, target can be hybridized to a large number of probes and consequently microarray chips have the capacity to simultaneously detect tens or hundreds or thousands of specific nucleic acid targets present in biological samples in a single experiment (Schena et al. 1995). Microarrays have been used in a large number of applications such as genome-wide genotyping, expression profiling, RNA detection, protein arrays, and pathogen nucleic acid detections (Petricoin et al. 2002).

6.3.1 Probe Designing for Pathogen Identification

Probes, which are short DNA sequences that are similar to parts of the sequence of target, are the most important constituent of a microarray chip. Probe has to be specific to a pathogen only, and then only it can identify a pathogen unambiguously. Besides specificity probe designing requires consideration of many other factors, which influence hybridization processes such as guanine/cytosine (G/C) content, melting point, secondary structure, sequence specificity, polynucleotide tract, and probe length. Probe length affects the sensitivity and specificity of hybridization, while other factors contribute to nonspecific hybridization. Maximizing the specificity and the sensitivity are often conflicting goals in terms of achieving probe design. Some well-known examples for microarray chips are as follows.

The probes for rotavirus were selected by using the following criteria: lengths of about 20 nucleotides, melting temperatures between 65°C and 75°C, and two or more mismatches with homologous sequences in other genotypes (Chizhikov et al. 2002). For orthopoxviruses, C23L/B29R gene sequences of different orthopoxvirus species were aligned using clustalx software to find variable regions suitable to design species-specific oligoprobes. Criteria for oligoprobe design were as follows: one or more mismatches with other orthopoxvirus species, length of 13–21 nucleotides, and predicted melting temperatures between 36°C and 58°C (Majid et al. 2003).

One of the very first microarray chips to utilize a generalized computational algorithm to find conserved regions among viral pathogens was by Wang et al. (2002). The oligonucleotide probe length used in this chip was 70-mer. The chip based on this conserved sequence was used to identify a coronavirus from the then unknown SARS-coronavirus samples. Chou et al. (2006) developed a comprehensive algorithm for designing conserved probes. They assumed that a virus genus (G) is a collection of n viruses, in which each virus v_i ($i = 1, \ldots, n$) is associated with a subset of G. Comparison of this virus with another virus in the genus identifies similar sequences. The similarity was defined as having either (i) more than 75% local sequence similarity in a 50-bp window with any virus or (ii) >15 consecutive bases pairing. From these set of conserved sequences, those sets are picked up which would identify the entire virus in the genus alone or in combination. It was not necessary that a single conserved sequence would pick up the entire virus in the genus. From the conserved sequences, 70-mer conserved probe were selected based on the following criteria: (i) GC content between 40% and 60%, (ii) <5 continuous mononucleotide repeats, (iii) <25-bp BLASTN sequence identity matches, and (iv) $\leq$15 consecutive bases pairing with other viral sequences in the noncognate viral genus (Chou et al. 2006).

Jabado et al. (2008) used protein families database (Pfam) as the basis for designing conserved oligonucleotide probes. Their strategy for finding conserved region entails identifying short conserved region and corresponding nucleic acid region. First they identified most conserved nonoverlapping 20 amino acids sequences then extracted corresponding nucleotide sequences. If the gaps were found, the flanking regions were taken into account. The sequences which were not part of Pfam were extracted and homologous gene clustered. All clustered sequence was searched for common motifs with software such as MEME. Three motifs were selected for each sequence cluster. The nucleic acid sequence extracted for each protein motif was used for probe design. The conditions for probe design were $\leq$5 mismatches to the template, a T_m >60°C, no repeats exceeding a length of 10 nt, no hairpins with stem lengths exceeding 11 nt, and <33% overall sequence identity to non-viral genomes (Jabado et al. 2008).

Pan-Microbial Detection Array (MDA) is the most comprehensive array designed for virus diagnosis. This array includes probes for all the virus sequences present in the database at the time of designing. The designing strategy used for this array was family based. First, the sequence of all the viruses reported for virus families was grouped, and from the group, all the sequences similar to human and nonfamily virus sequences were removed. From the resultant viral sequences, probes were designed irrespective of the location or the gene using primer 3. The primers were selected on the basis of conservation within family and the number of probes per target sequence which was 50 probes per target sequence (Gardner et al. 2010).

6.3.2 Clinical Sample Processing and Hybridization

The clinical samples for identification of a virus are collected from different places, from different animals, and from different sources. The clinical samples could be blood, a swab from the nose, vagina, or mouth, tissue and stool, or any other. Most of these samples have been tested in microarrays. Foot and mouth disease virus (FMDV) was identified in ticks collected from a livestock market in Nairobi, Kenya (Sang et al. 2006). Vincent (2009) identified clinical porcine respiratory and reproductive syndrome virus in a nasal swab from pig samples. The clinical samples for microarray analysis are collected in solutions (like RNAlater) which makes RNA stable or directly in trizol for RNA isolation (Wang et al. 2002; Chiu et al. 2007). Samples for microarrays are generally processed by the methods adopted by Wang et al. (2002). This method employs anchored random nonamer primers for cDNA synthesis, nonspecific amplification, and introduction of amino-allyl nucleotide into the amplification product. Labeling is done by covalent binding of fluorescent dye (cyanine with aminoallyl group of nucleotide (Jabado et al. 2008; Wang et al. 2002; Gardner et al. 2010). Hybridization is done overnight, and the temperature of hybridization is defined by the probe sequences which vary between 65°C and 70°C (Tan et al. 2003; Wang et al. 2003; Jayaraman et al. 2006; Gardner et al. 2010). Hybridization signals are generated by using light of a predefined wavelength to stimulate the emission of the fluorescent signal. The amount of emission is determined by the amount of fluorescent dyes bound, which is correlated with the amount of targets-probe hybrids at the spot.

6.3.3 Data Processing on the Basis of Signal Intensity

Virus prediction from the signal intensity data has been the subject of intense study. Unlike gene expression microarrays, diagnostic microarrays have no up or down-regulated gene. In diagnostic microarray, a signal cannot be low, high, or unchanged; it has to be defined in binary numbers—present or absent, on or off. A signal is defined as present if it is above a predefined cutoff. This cutoff signal has arrived differently as the microarray technology for diagnosis developed. In the first broad-spectrum microarray chip, the control hybridization was carried out with the RNA from uninfected cell culture. In this particular chip, each spot was spiked with a known probe in a fixed ratio to normalize the expression of all the probes. The complementary sequence of the spike probe was labeled with Cy3 dye. Two color hybridization experiments were done for both infected and uninfected cell culture. The Cy5 signal for each probe was normalized against Cy3 signal. Cy5 labeled-infected and uninfected signal for each probe were calculated; from these values, a value of 1500 was arbitrarily defined as cutoff value, and the prediction was based on aggregate hybridization (Wang et al. 2003). In diagnostic arrays, it is not always possible to have controls for identifying cutoff value like Wang et al. (2003) where they used mock-infected cell culture as control. To overcome this control problem, random probes which have no sequence similarity with NR database have been

introduced (Gardner et al. 2010). If control samples are available, they are used for calculating the cutoff value; otherwise signal intensity of random probes is used for calculating cutoff value. The cutoff value is kept at median + 2 SD of the random probe signal intensity, 95–99% percentile of random signal intensity (Gardner et al. 2010). The virus identification from the signal intensity data makes use of different approaches. The simplest approach for predicting presence of virus in the sample is based on averaging signal intensity of all the probes of a virus or virus genus/family: if it is above a cutoff value, virus is predicted to be present; the other approach is based on the number/percentage of probes giving positive signal above a preset threshold (Sharma et al. 2012; Yadav et al. 2015). The first broad-spectrum virus-detecting microarray chip, ViroChip, made prediction based on the aggregate hybridization pattern (Wang et al. 2003). Chou et al. (2006) adopted similar criteria but with some modifications. Their method makes use of both signal intensity and a number of probes for making prediction. First, sum of all the signal intensity of a probe set is calculated, and then this is divided by the maximum intensity obtained for any group in the probe set. Based on the percentage, prediction is made. For the GreenChip (Palacios et al. 2007), virus prediction was done by a specially developed software called Green-LAMP. This software subtracts background values from the probe intensities, calculate Z-score from log intensities by dividing with standard deviation and compute tail probabilities (p-value) assuming log normality. Presence of a positive or negative signal is computed from a fixed P-value threshold, 0.1 for arrays with matched controls and 0.023 otherwise. The background levels are derived from matched control samples or from random 60-mer control probes if matched control samples are not present. The following assumptions were made for making predictions: (1) spot intensities are normally distributed, (2) spots represent independent observations, and (3) there are relatively few (<100) positive probes for any given virus. DetectiV (Watson et al. 2007) tool was developed for handling array data by selecting groups of probes comprising a species, genus, or family and computing a one-sample t-test with null hypothesis that log intensity ratio for each group is zero. PhyloDetect (Rehrauer et al. 2008), another tool for making virus prediction, converts the probe intensities to binary indicators (e.g., by thresholding against the median + 2 SD of the background intensities). In Pan-Microbial Detection Array (MDA) (Gardner et al. 2010), the threshold for positive signal was kept at 99th percentile of negative controls which were randomly generated probes. Prediction for presence or absence of a pathogen was based upon two conditional probabilities: the probability of observing a signal in the presence of a specific microbial target in sample and the probability of observing signal in the absence of microbial target in the sample. Liu et al. (2006) compared five different methods of virus prediction and concluded that hypergeometric distribution and log transform ranking method give good prediction, but other methods like the number of probes with a threshold or ratio method also give suitable prediction.

6.4 Recent Advances for Identification and Characterization of Microbes

Over the last two decades, sequence analysis of conserved genes has become a reliable, accurate, inexpensive, and scalable method of microbial identification in health and environmental sciences. These advantages have resulted in the routine use of sequencing methods to complement, and sometimes replace, traditional phenotypic methods of identification.

Various molecular techniques have emerged in the recent decades offering speed combined with specific and sensitive detection. They are simple, rapid and reliable, and dependent on the presence of nucleic acids (DNA and RNA) that code for the proteins. These methods include polymerase chain reaction (PCR), microarrays, metagenomics, next-generation sequencing, and many others. Detection of DNA is now possible on a single molecule, and high-throughput analysis allows thousands of detection reactions to be performed at once, thus allowing a range of characteristics to be rapidly and simultaneously determined. Some of the recent molecular detection methods can be performed in the laboratory or clinical settings and also at the farm site. Although some of these techniques provide immediate result, many require extensive computational approaches for analysis and interpretation of the data.

Metagenomics is recently introduced where we study the genomic content of an environmental sample of microbes. It is a derivation of conventional microbial genomics, with the key difference being that it bypasses the requirement for obtaining pure cultures for sequencing. Metagenomics holds the promise of revealing the genomes of the majority of microorganisms that cannot be readily obtained in pure form. Since the samples are obtained from communities rather than isolated populations, metagenomics may serve to establish hypotheses concerning interactions between community members. This process begins with sample and metadata collection and proceeds with DNA extraction, library construction, sequencing, read preprocessing, and assembly. Community composition analysis is employed at several stages of this workflow, and databases and computational tools are used to facilitate the analysis. Advances in the throughput and cost-efficiency of sequencing technology are fueling a rapid increase in the number and size of metagenomics datasets being generated. However, bioinformaticists are faced with the problem of how to handle and analyze these datasets in an efficient and useful way (Tringe and Rubin 2005). The goal of metagenomics studies is to get a basic understanding of the microbial world both surrounding us and within us.

Information from metagenomics studies will be fully exploited only if appropriate data-management and data-analysis methods are in place. One was that the data were immediately accessible in a form suitable for computer analysis; another was that the data were freely available, without impediment to all researchers, be they in academia or industry. The three nucleic acid sequence archives GenBank, EMBL-Bank, and DDBJ have spearheaded the cause of free availability of sequence information. In the process sequences of a large number of fragments have been registered in the international DNA databanks. However, the details of function of the sequence are

not available and are of limited use. Analysis and comparison of complex metagenomic data is driving the development of a new class of bioinformatics and visualization software. The field is moving forward rapidly, driven by enormous improvements in sequencing technology and the availability of many complementary technologies. Analysis and clustering of metagenomic sequences with the help of bioinformatics tools according to phenotypes and genomes might in future help in environmental preservation (Kunin et al. 2008).

6.4.1 New-Generation Sequencing

Sequencing is one technique that transformed biology from qualitative to a quantitative science and leads to the emergence of bioinformatics as an important discipline. Initially, sequencing started with radioisotope-labeled sequencing products analyzed on slab gels. This slow process was overtaken by fluorescent labeling and capillary electrophoresis that improved speed and data quality of sequencing. Recently the next-generation sequencing platforms have made possible massive parallel sequencing without the need for lengthy electrophoresis. There are various approaches for next-generation sequencing like sequencing by hybridization, microelectrophoresis, cyclic array sequencing and real-time observation of single molecules. These diverse approaches and sophistication of next-generation sequencing have brought great challenges for bioinformaticists to tackle alignment, sequence scoring, data assembly, storage, and release of huge amounts of data (Kunin et al. 2008).

The ability to simultaneously acquire huge amount of sequence data when applied to clinical and environmental samples helps in the identification of pathogenic microbes. Moreover, genome variability and evolution within the host can be tracked over short periods of time. These approaches were already being used in diagnostic virology for detection of novel pathogenic viruses and for mapping of resistance to antiviral drugs (Barzon et al. 2011).

6.5 Conclusion and Future Perspective

DNA microarray is a rapid method for virus identification. Proof of concept has already been shown in at least two cases where DNA microarray identified pathogen when all the other method for diagnosis failed. Currently the biggest problems in designing probes for diagnostic arrays are in designing conserved probes at genus level. In viral genera, getting conserved probes which would identify all the virus in the genus is next to impossible for almost all genera; as a consequence genus has to be arbitrarily subdivided based upon sequence conservation within the genus. This strategy has not been adopted though in MDA array the probes have been selected based upon conservation but not on uniqueness within a subgroup. Another problem with using microarray is the sequence heterogeneity within the virus species. All the unique probes of a virus species would not bind to all the isolates so a threshold has to be set for making prediction. The capacity of chips is going up, the

new-generation chips can incorporate a million probes, while the number of viruses reported in ICTV is just above 2000; however the number of virus sequence reported and stored in NCBI database is in hundreds of thousands. Thus it is possible to make probes for all the sequences and incorporate them in a chip, but doing that would create problems in interpretation of results because of cross-hybridization signals. The one way to avoid cross-hybridization signal is to reduce the size of probes which is currently set at about 70-mer. Microarray is costly as camper to PCR, and so it is generally restricted to the few commercial laboratories that can possess the capital, or those laboratories developing expertise in this field. This is a long procedure for the numbers of genes involved in this technique. ViroChip pan-viral microarray recently used deep sequencing technology to 17 respiratory samples collected from individuals infected with the 2009 H1N1 influenza virus early during the pandemic and deep sequencing which can test for thousands of potential pathogens simultaneously. Consequently diagnostic strategy of rapid ViroChip-based testing followed by deep sequencing could show to be a useful public health response to infectious disease outbreaks in the future (Yadav et al. 2014). Thus, identifying viral species using the previously reported viral microarray probe design strategy with new approaches is very impressive. The use of microarray in pathogen identification is still an intensive area of research. The new design strategies are constantly coming up. It is hoped that in near future a very precise and cost-effective chip would be developed, but to increase its practical usage in clinical microbiology laboratories, it has to become more affordable, be convenient to handle, and be accurate.

References

An HJ, Cho NH, Lee SY, Kim IH, Lee C, Kim SJ, Mun MS, Kim SH, Jeong JK (2003) Correlation of cervical carcinoma and precancerous lesions with human papillomavirus (HPV) genotypes detected with the HPV DNA chip microarray method. Cancer 97:1672–1680

Anderson BT, Ruane AC, Roads JO, Kanamitsu M, Salvucci G (2008) A new metric for estimating local moisture cycling and its influence upon seasonal precipitation rates. J Hydrometeorol 9:576–588

Barrett T, Pastoret P, Taylor W (2006) Rinderpest and peste des petits ruminants: virus plagues of large and small ruminants. Elsevier Academic Press, Amsterdam. ISBN 0-12-088385-6

Barrette RW, Metwally SA, Rowland JM, Xu L, Zaki SR, Nichol ST (2009) Discovery of swine as a host for the reston ebolavirus. Science 325:204–206

Barzon L, Lavezzo E, Militello V, Toppo S, Palù G (2011) Applications of next-generation sequencing technologies to diagnostic virology. Int J Mol Sci 12(11):7861–7884

Baumeister MA, Zhang N, Beas H, Brooks JR, Canchola JA, Cosenza C, Kleshik F, Rampersad V, Surtihadi J, Battersby TR (2012) A sensitive branched DNA HIV-1 signal amplification viral load assay with single day turnaround. PLoS One 7(3):e33295

Benedetti VMD, Biglia N, Sismondi P, De Bortoli M (2000) DNA chips: the future of biomarkers. Int J Biol Markers 15(1):1–9

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21(1):33–37

Bryant PA, Venter D, Robins-Browne R, Curtis N (2004) Chips with everything: DNA microarrays in infectious diseases. Lancet Infect Dis 4(2):100–111

Chen Q, Li J, Deng Z, Xiong W, Wang Q, Hu YQ (2010) Comprehensive detection and identification of seven animal coronaviruses and human respiratory coronavirus 229 E with a microarray hybridization assay. Intervirology 53(2):95–104

Chiu CY, Rouskin S, Koshy A, Urisman A, Fischer K, Yagi S, Schnurr D, Eckburg PB, Tompkins LS, Blackburn BG, Merker JD, Patterson BK, Ganem D, DeRisi JL (2006) Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult. Clin Infect Dis 43(8):e71–e76

Chiu CY, Alizadeh AA, Rouskin S, Merker JD, Yeh E, Yagi S, Schnurr D, Patterson BK, Ganem D (2007) Diagnosis of a critical respiratory illness caused by human metapneumovirus by use of a pan-virus microarray. J Clin Microbiol 45:2340–2343

Chizhikov V, Wagner M, Ivshina A, Hoshino Y, Kapikian AZ, Chumakov K (2002) Detection and genotyping of human group A rotaviruses by oligonucleotide microarray hybridization. J Clin Microbiol 40:2398–2407

Chou CC, Lee TT, Chen CH, Hsiao HY, Lin YL, Ho MS, Yang PC, Peck K (2006) Design of microarray probes for virus identification and detection of emerging viruses at the genus level. BMC Bioinformat 28(7):232–240

Cunningham CO (2004) Use of molecular diagnostic tests in disease control: making the leap from laboratory to field application. In: Current trends in the study of bacterial and viral fish and shrimp diseases. World Scientific Publishing Co., Singapore, pp 292–312

Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE (2000) Multiplex PCR: optimization and application in diagnostic virology. Clin Microbiol Rev 13:559–570

Espy MJ, Uhl JR, Sloan LM, Buckwalter SP, Jones MF, Velter EA, Yao JDC, Wengenach NL, Rosenblatt JE, Cockerill FR, Smith TF (2006) Real-time PCR in clinical microbiology: applications for routine laboratory testing. Clin Microbiol Rev 19(1):165–256

Feilotter HE (2004) Microarrays in veterinary diagnostics. Anim Health Res Rev 5(2):249–255

Gardner SN, Crystal JJ, Kevin SM, Tom RS (2010) A microbial detection array (MDA) for viral and bacterial detection. BMC Genomics 11:668

Goldsmith CS, Miller SE (2009) Modern uses of electron microscopy for detection of viruses. Clin Microbiol Rev 22(4):552–563

Grubaugh ND, Petz LN, Melanson VR, McMenamy SS, Turell MJ, Long LS, Pisarcik SE, Kengluecha A, Jaichapor B, O'Guinn ML, Lee JS (2013) Evaluation of a field-portable DNA microarray platform and nucleic acid amplification strategies for the detection of arboviruses, arthropods, and bloodmeals. Am J Trop Med Hyg 88(2):245–253

Hentges DJ (1993) The anaerobic microflora of the human body. Clin Infect Dis 16(Suppl 4):S175

Jabado OJ, Liu Y, Conlan S, Quan PL, Hegyi H, Lussier Y, Briese T, Palacios G, Lipkin WI (2008) Comprehensive viral oligonucleotide probe design using conserved protein regions. Nucleic Acids Res 36(1):e3

Jack PJ, Amos-Ritchie RN, Reverter A, Palacios G, Quan PL (2009) Microarray based detection of viruses causing vesicular orvesicular-like lesions in livestock animals. Vet Microbiol 133:145–153

Jayaraman A, Hall CK, Genzer J (2006) Computer simulation study of molecular recognition in model DNA microarrays. Biophys J 91:2227–2236

Kistler A, Avila PC, Rouskin S, Wang D, Ward T, Yagi S, Schnurr D, Ganem D, DeRisi JL, Boushey HA (2007) Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. J Infect Dis 196 (6):817–825

Kostrzynska M, Bachand A (2006) Application of DNA microarray technology for detection, identification, and characterization of food-borne pathogens. Can J Microbiol 52(1):1–8

Kreuze J, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1–7

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72(4):557–578

Kurian KM, Watson CJ, Wyllie AH (1999) DNA chip technology. J Pathol 187(3):267–271
Leblanc N, Gantelius J, Schwenk JM, Stahl K, Blomberg J, Andersson- Svahn H, Belak S (2009) Development of a magnetic bead microarray for simultaneous and simple detection of four pestiviruses. J Virol Methods 155:1–9
Leland DS, Christine CG (2007) Role of cell culture for virus detection in the age of technology. Microbiol Rev 20(1):49–78
Lin B, Wang Z, Vora GJ, Thornton JA, Schnur JM, Thach DC, Blaney KM, Ligler AG, Malanoski AP (2006) Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. Genome Res 16:527–535
Liu YC, Huang GS, Wu MC, Hong MY, Hsiung KP (2006) Detection of foot and mouth disease and porcine reproductive and respiratory syndrome viral genes using microarray chip. Vet Res Commun 30(2):191–204
Loy A, Bodrossy L (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. Clin Chim Acta 363:106–119
Mackay IM, Arden KE, Nitsche A (2002) Real-time PCR in virology. Nucleic Acids Res 30 (6):1292–1305
Majid L, Vladimir C, Maxim M, Sergei S, Konstantin C (2003) Detection and discrimination of orthopoxviruses using microarrays of immobilized oligonucleotides. J Virol Methods 112 (1–2):67–78
Manoj KR (2009) The widely used diagnostics "DNA microarray"- a review. Ame Jourof Inf Dis 5(3):207–218
Martín V, Perales C, Abia D, Ortíz AR, Domingo E, Briones C (2006) Microarray-based identification of antigenic variants of foot-and-mouth disease virus: a bioinformatics quality assessment. BMC Genomics 7:117
Mihindukulasuriya KA, Wu G, St. Leger J, Nordhausen RW, Wang D (2008) Identification of a novel coronavirus from a beluga whale by using a panviral microarray. J Virol 82:5084–5088
Miller SW, Avidor-Reiss T, Polyanovsky A, Posakony JW (2009) Complex interplay of three transcription factors in controlling the tormogen differentiation program of Drosophila mechanoreceptors. Dev Biol 329(2):386–399
Morens DM, Fauci AS (2013) Emerging infectious diseases: threats to human health and global stability. PLoS Pathog 9(7):e1003467
Mullis K (1990) The unusual origin of the polymerase chain reaction. Sci Am 4:56–65
Nii S (1971) Electron microscopic observations on FL cells infected with herpes simplex virus. I. Viral forms. Biken J 14(2):177–190
Palacios G, Quan PL, Jabado OJ, Conlan S (2007) Panmicrobial oligonucleotide array for diagnosis of infectious diseases. Emerg Infect Dis 13(1):73–81
Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA (2002) Clinical proteomics: translating benchside promise into bedside reality. Nat Rev Drug Discov 1:683–695
Quan PL, Palacios G, Jabado OJ, Conlan S (2007) Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. J Clin Microbiol 45(8):2359–2364
Ratta B, Yadav BS, Pokhriyal M, Saxena M, Sharma B (2013) Microarray chip based identification of a mixed infection of bovine herpesvirus 1 and bovine viral diarrhea 2 from Indian cattle. Curr Microbiol. https://doi.org/10.1007/s00284-013-0448-9
Rehrauer H, Schönmann S, Eberl L (2008) PhyloDetect: a likelihood-based strategy for detecting microorganisms with diagnostic microarrays. Bioinformatics 24:83–89
Rockett JC, Dix DJ (2000) DNA arrays: technology, options and toxicological applications. Xenobiotica 30(2):155–177
Sang R, Onyango C, Gachoya J, Mabinda E, Konongoi S, Ofula V, Dunster L, Okoth F, Coldren R, Tesh R, da Rossa AT, Finkbeiner S, Wang D, Crabtree M, Miller B (2006) Tickborne arbovirus surveillance in market livestock, Nairobi, Kenya. Emerg Infect Dis 12(7):1074–1080
Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467–470

Sharma B, Pokhriyal M, Rai GK, Saxena M, Ratta B, Chaurasia M, Liu BS, Sen A, Mondal B (2012) Isolation of Newcastle disease virus from a non-avian host (sheep) and its implications. Arch Virol 157(8):1565–1567

Staudt LM, Brown PO (2000) Genomic views of the immune system. Annu Rev Immunol 18:829–859

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98(3):503–517

Su H, Long J, Guo Q, Meng X, Tan Y, Cai Q, Chen Z, Meng X (2013) Long-tail probe-mediated cycled strand displacement amplification: Label-free, isothermal and sensitive detection of nucleic acids. Talanta 15(116):330–334

Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 31:5676–5684

Tomita N, Mori Y, Kanda H, Notomi T (2008 May) Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. Nat Protoc 3(5):877

Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. Nat Rev Gen 6:805–814

Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G (2006) Identification of a novel Gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. PLoS Pathol 2(3):e25

Vincent AL (2009) Characterization of influenza Avirus isolated from pigs during an outbreak of respiratory disease in swine and people during a county fair in the United States. Vet Microbiol 137:51–59

Voller A, Bartlett A, Bidwell DE (1978) Enzyme immunoassays with special reference to ELISA techniques. J Clin Pathol 31:507–520

Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL (2002) Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci 99 (24):15687–15692

Wang D, Urisman A, Liu Y-T, Springer M, Ksiazek TG (2003) Viral discovery and sequence recovery using DNA microarrays. PLoS Biol 1(2):e2

Watson M, Dukes J, Abu-Median AB, King DP, Britton P (2007) DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. Genome Biol 8(9):R190

Wilson WJ, Strout CL, DeSantis TZ, Stilwell JL, Carrano AV, Andersen GL (2002) Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. Mol Cell Probes 216(2):119–127

Yadav BS, Ronda V, Vashista DP, Sharma B (2013) Sequencing and computational approach to identification and characterization of microbial organisms. Biomed Eng Comput Biol 5:43–49

Yadav BS, Pokhriyal M, Vasishtha DP, Sharma B (2014) Animal viruses probe dataset (AVPDS) for microarray-based diagnosis and identification of viruses. Curr Microbiol. https://doi.org/10.1007/s00284-013-0477-4

Yadav BS, Pokhriyal M, Ratta B, Kumar A, Saxena M, Sharma B (2015) Viral diagnosis in Indian livestock using customized microarray chips. Bioinformation 11(11):489–492

Zappa A, Antonella A, Zanetti A (2009) Emerging and re-emerging viruses in the era of globalization. Blood Transfus 7(3):167–171

Application of Whole Genome Sequencing (WGS) Approach Against Identification of Foodborne Bacteria

7

Shiv Bharadwaj, Vivek Dhar Dwivedi, and Nikhil Kirtipal

Abstract

Food quality and safety along with their associated hazards risks present a major concern worldwide associated with relative economical losses as well as potential danger to consumer's health. In this context, antimicrobial resistance (AMR) surveillance is a critical step within risk assessment schemes, as it is the basis for informing global strategies, monitoring the effectiveness of public health interventions, and detecting new trends and emerging threats linked to food. A lack of measures and reliable assays to evaluate and maintain a good control of antimicrobial resistance foodborne pathogens may affect the food industry economy and shatter consumer confidence. Hence, surveillance of AMR is currently based on the isolation of indicator microorganisms and the phenotypic characterization of clinical, environmental, and food borne strains. However, this approach provides very limited information on the mechanisms driving AMR or on the presence or spread of AMR genes throughout the food chain. It is imperative to establish fast and reliable analytical methods that allow a good and rapid analysis of food products during the whole food chain. This chapter summarizes the information on the method developed and application of the whole-genome sequencing (WGS) in the past few years focusing on surveillance of AMR in foodborne pathogenic bacteria. Emphasis is also posed with respect to the routine implementation of these next-generation sequencing methodologies on characterization of well-known food pathogens. Besides, potential advantages and

S. Bharadwaj
Department of Biotechnology, College of Life and Applied Sciences, Yeungnam University, Gyeongbuk-do, Republic of Korea

V. D. Dwivedi
Center for Bioinformatics, Pathfinder Research and Training Foundation, Greater Noida, Uttar Pradesh, India

N. Kirtipal (✉)
Department of Biotechnology, Modern Institute of Technology, Rishikesh, Uttarakhand, India
e-mail: kirtipal.n@gmail.com

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_7

disadvantages of the WGS have been discussed on the surveillance of AMR in foodborne pathogens.

7.1 Introduction

Recent studies have documented that the earth is occupied with 7.5 billion human population while approx. 10% are malnourished (Zuker 2015). Food is the only energy source for doing day-to-day daily activities, and hence, factors in relation to food derived from both plant and animal sources such as food security, safety and quality become essential for the rising population. For instance, many microbes such as bacteria or yeast are used to process the raw food into consumable desired product (Smid and Kleerebezem 2014). The industries that use such microbes for the food processing, always monitored and optimized the strain performance to get the diversified desired products in terms of flavor and texture through known culture techniques (Smid and Kleerebezem 2014). In contrast, use of contaminated microbial culture reduces the quality of the desired product and arises major health concerns. Because of rapid feasting of processed and stored food, it becomes inevitable to monitored the quality of food using easy and rapid analytical approach (Piras et al. 2016).

Food safety is one of the major concerns in the developed countries as in recent years many deaths have been surfaced there due to consumption of spoiled food. A report published by the World Health Organization (WHO) stated that annually approximately 2.2 million people worldwide suffer from diarrheal diseases due to usage of contaminated food and water (Kuchenmüller et al. 2009). Additionally, environmental pollution and random changes in climate were also suggested to reduce the quality of consumable food by providing an ideal environment for many belligerent contaminants (Brambilla and Testa 2014; Seltenrich 2015). Foodborne pathogens mainly comprise the bacteria and fungi that have been frequently reported for the foodborne illnesses through consumption of spoiled food as well as epidemic and massive commercial damage worldwide. For instance, each year about 325,000 hospitalizations and 5000 deaths are registered in the United States caused by food poisoning (Gašo-Sokač et al. 2010). Among the reported bacterial strains, *Campylobacter jejuni*, *Staphylococcus aureus*, some virulent strains of *Bacillus* sp., *Staphylococcus* sp., *Escherichia coli*, and *Salmonella* sp., and toxin-producing fungi like *Penicillium*, *Claviceps* and *Aspergillus* spp. are marked as foodborne pathogens (Gašo-Sokač et al. 2010; Giacometti et al. 2013). Subsequently, preventive steps against food spoilage to control the foodborne pathogen-induced epidemics or diseases were concluded as an utmost priority in view of enormous economic and social importance (Havelaar et al. 2010).

7.2 Foodborne Pathogens Identification

7.2.1 Traditional Approaches for Classification of Pathogenic Strains

In the present scenario, various microbial techniques are used to isolate and characterize the pathogenic microbes as well as even differentiate them into several serotypes. The microbial characterization, however, varies for each species and mainly depends on the aim of the study. The most popular techniques employed in the study of foodborne pathogens are briefly discussed as follows:

(a) **Phage typing:** In this method, various types of bacteriophage (a group of viruses) are used and strains are distinguished based on their susceptibility to viral infection.
(b) **Multilocus sequence typing (MLST):** Herein, small variations in 400–500 base pair (bp) fragments of genome at seven different conserved genes in the species are detected using sequencing technique. However, this approach is time-consuming and expensive, while the correct selection of the genes can result in the discrimination between various strains of species.
(c) **Serotyping:** This method exploits the set of surface-specific receptors or molecules coating on the strains to distinguish them from others by means of molecule-specific antibodies.
(d) **Multilocus variable-number tandem repeat analysis (MLVA):** Some sequences in the genome of the species are monotonous whose number and loci vary among the different strains from same species. Analysis of such sequences can be performed with an easy and required less time in comparison to MLST. However, this approach is accompanied with validation and reproducibility issues.
(e) **Pulsed-field gel electrophoresis (PFGE):** Different strains of the same species occupied with some restriction sites at specific loci on the genome. Hence, genome is digested with specific restriction enzymes and then produced DNA fragments are electrophoretically separated on a gel based on DNA fragment size to identify the strain.

7.2.2 Disadvantage of Traditional Approaches for Foodborne Pathogen Strains Classification

The choice of typing method for distinguishing virulent strains primary relies on the hazard, aim for investigation, and most suitable method(s) available for the target. DNA fingerprinting method, such as PFGE and MLVA, and DNA sequence-based approach like MLST are concluded as an important approach for the investigation and recognition of foodborne pathogens (Swaminathan et al. 2001; Joseph and Forsythe 2012). However, PFGE and MLVA often failed to reflect suitable discriminatory results among the sporadic and outbreak-related cases for specific subtypes

of a given pathogenic species (Franz et al. 2016). Moreover, MLST suffered with short comes when applied against the pathogens that possess a high level of clonal and conserved population structure. In such occasions with less diversity in the strains of species, MLST might not always provide distinguishable results (Ranieri et al. 2013).

7.3 Whole Genome Sequencing (WGS)

Whole genome sequencing (WGS), a sequencing-based approach, has been recently purposed as a potential tool to assess, investigate and manage the diseases caused by foodborne pathogens. This approach can be used to identify and classify the pathogen with fine precision and hence, WGS is defined as an effective and efficient approach to manage the microbe-related food safety issues. Additionally, cost-effectiveness of this approach has greatly attracted its use in food safety management. Although many industrialized countries put forward this technology in food safety, but application of WGS is still particularly limited in transitional and developing countries.

7.3.1 Food Safety and Whole Genome Sequencing

With the advent of whole genome sequencing (WGS) technology, it provides worldwide substantial development and surveillance for food safety. In contrast to hitherto available plethora of molecular characterization and identification methods, WGS is simple to use regardless of the platform. Herein, genomic deoxyribonucleic acid (DNA) is isolated and sequenced with marked tags followed by visualization and analysis of generated data by the aid of bioinformatics tools. WGS is the universal approach that can be applied to any organisms and hence, beneficial in monitoring the production environment for food processing and quality. This technology has virtually mapped the whole genome of an organism that facilitates targeted comparison and exchange of the generated data. Moreover, these results can be used for foodborne disease surveillance, outbreak investigation and to check the food quality. Besides, this disclosed the critical steps where precautions and preventive steps needs to be taken for improved food safety through identification of transmission pathways for the foodborne pathogens. Moreover, whole genome data generated by WGS associated with some advantages such as sequence analysis can be applied simultaneously for the pathogen identification, subtyping, virulence biomarkers, predictions of antimicrobial resistance and genome-wide association studies. Furthermore, such data can be mined and stored in the form of databases which later can be retrieved or reanalyzed for the identification as well as management of emerging pathogens over time.

7.3.2 Whole Genome Sequencing Technologies

Based on the approach used in sequencing of the target, whole genome sequencing technique is broadly categorized as (i) short-read technologies that generate 500bp sequence fragments (e.g., Ion Torrent and Illumina) and (ii) long-read technologies, which produce longer DNA fragments of 1000–70,000 bp (e.g., Oxford Nanopore and Pacific Biosciences).

7.3.2.1 Short-Read Platforms

The majority of WGS platforms are classified under the group of short-read sequencers. These sequencers have a maximum read length of 1000 bases, but a more typical base range is between 50 and 400 bp (Deurenberg et al. 2017). The error rate for these platforms is quite low with accurate nucleotide calling score more than 99%. Traditionally, short-read platforms are further broken down into sequencing-by-synthesis (SBS) and sequencing-by-ligation (SBL) methods. There are currently two available SBL platforms; however, these instruments account for only a small percentage of sequencers in use because SBS has a higher output and is comparatively cost-effective. SBL will, therefore, not be discussed further in this chapter. There are two relevant SBS platforms in use: the Illumina suite of instruments and, to a lesser extent, the Ion Torrent. The Illumina suite of instruments relies on incorporating fluorescently labelled nucleotides in an elongating DNA strand. The nucleotides are modified such that only one base is incorporated at a time. As each base is incorporated into the elongating strand, the instrument identifies the nucleotide base in either two (MiniSeq, NextSeq) or four detection channels (MiSeq, HiSeq) (World Health Organization 2018). There are three broad types of Illumina instruments.

(a) ***HiSeq instrument.*** These machines offer the lowest cost per gigabase (Gb) of any currently available platform. Although the throughput of these instruments is quite useful for larger genomes (such as three billion nucleotides human genome) but their application to pathogen sequencing of a few million bases is not practical. For example, one lane of the HiSeq 2500 generates approximately 60 Gb of paired-end sequencing data of 125 bases in length or approximately one human genome at 18X coverage. This is equivalent to more than 66 *E. coli* genomes at similar coverage, far beyond what can reasonably be expected (or needed) at a typical pathogen sequencing center. This instrument's capabilities are more appropriate for activities carried out at human research institutions and eukaryotic genome centers.

(b) ***NextSeq instrument.*** Unlike the HiSeq instrument, this device does not require a minimum number of samples, making it a more practical option for sequencing centers that do not expect a regular influx of samples. NextSeq provides a per lane throughout like that of HiSeq 2500, and at similar costs per Gb sequenced. This makes the instrument a good fit for an academic sequencing core or a regional center focused on human sequencing. However, like HiSeq,

its throughput is likely to be beyond what is needed for a beginning pathogen sequencing center.

(c) ***MiSeq and MiniSeq instruments.*** These two techniques i.e. MiSeq and MiniSeq have 0.5–15 Gb and 1.5–7.5 Gb output, respectively. Like NextSeq, neither of these instruments requires a minimum number of samples. From a technical perspective, these instruments are therefore well suited for pathogen sequencing. The lowest-throughput setting (single reads of 36 bp in length on the MiSeq) can generate one *E. coli* genome at 100X coverage in as little as 4 h. With one of these instruments, a regional sequencing center could comfortably sequence 96 bacterial genomes per week. While the cost per Gb of these instruments is higher than that of the higher-throughput platforms, the initial instrument cost is substantially lower. Hence, this instrument is comparatively less used but, in a similar niche as the MiSeq and MiniSeq, is the suite of instruments from Ion Torrent (Thermo Fisher). Unlike the Illumina products, the Ion Torrent instruments detect bases through the release of a hydrogen ion during strand elongation rather than an optical signal. This method provides one notable advantage over the Illumina suite: Ion Torrent suite offers the fastest sequencing time (as little as 2.5 h) of any currently available instrument. However, this method of detection also leads to a higher error rate than the Illumina suite, especially in homopolymeric regions, principally more difficult in de novo assemblies. While the initial cost of the suite of instruments from Ion Torrent is comparable to that of the Illumina suite, the overall cost per billion bases sequenced is somewhat higher, owing to a lower number of reads produced per run. Furthermore, there is a smaller community of active users and publicly available software developers for the Ion Torrent and its downstream data analysis, which means fewer options for data pipelines and fewer resources available for troubleshooting. For these reasons, the Ion Torrent suite of products is best reserved for targeted sequencing projects (e.g., 16s RNA, virulence marker identification and transcriptome profiling).

7.3.2.2 Long-Read Platforms

Long-read sequencing platforms are typically considered to generate average read length more than 10 kilobases (kb), generally accepted minimum read length for high-quality and long-read assemblies. There are currently two available long-read platforms: the Pacific Biosciences (PacBio) suite and the Oxford Nanopore Technologies (ONT) suite. Both platforms can generate both short and long reads, final read length being defined by input DNA fragments rather than instrument itself. The PacBio suite consists of RS II and Sequel instruments. Like short-read platforms, PacBio relies on the sequencing-by-synthesis approach where fluorescently labelled nucleotides are detected as they are incorporated into the elongating DNA strand. The PacBio approach is called single-molecule real-time (SMRT) sequencing because single DNA molecules are monitored with no pausing steps for interrogation. SMRT technology allows to reads more than 60 kb, but with error rates as high as 15%. As a result, these long-read de novo assembly strategies must achieve higher cumulative coverage (approximately 120X) to overcome this error

rate. In terms of yield, RS II can generate approximately 1 Gb of data and Sequel approximately 5 Gb per SMRT cell but only about half of the yield contain reads over 10 kb. This means that two long-reads of *E. coli* genomes can be generated at approximately 120X coverage per cell on RSII and ten on the Sequel, make them well suited for low throughput sequencing center. While the output data from short-read platforms can result into nearly complete genome assemblies, though some gaps are still expected due to use of shorter read length for the sequencing. However, long-read length of PacBio offers an important and distinct advantage of SMRT sequencing where reads can produce high-quality genomic sequence that typically captures all the genetic material (e.g., closed genome). This becomes an essential part of the sequencing when regulatory authorities required a probable complete genomic sequence. The main drawbacks for the PacBio suite are high cost of the instrument (more than US$350,000 for the Sequel), reagents and infrastructure for the instrument installation. Both RS II and Sequel have footprints of many times larger compare to other sequencers and require a continuous supply of nitrogen that potentially limiting their application in developing countries (World Health Organization 2018).

The ONT suites of instruments, including MinION and PromethION, use a unique sequencing method: single DNA molecules translocate through a biological pore and small perturbations of current passing through the pore can be interpreted in terms of bases. The MinION instrument has a 3×10 cm^2 footprint and is more portable than other sequencing platforms. Also, this instrument works without the need for significant fluids or optics and hence occupies the negligible cost. The throughput per MinION instrument is between 5 and 10 Gb per 48 h run (with the option of shorter runtimes), and the PromethION is expected to generate as much 24-whole genome sequencing for foodborne disease surveillance as 60 Gb per flow cell. While this throughput may be somewhat high for a pathogen sequencing center, several benefits such as iterative loading, shorter runtimes, and real-time base calling abrogate required for the significant multiplexing. The limitations of ONT suite include a biased error profile, with indels being especially problematic in repetitive regions. Reagent costs are also higher than other platforms, though the low instrument cost and upcoming "flongle" flow cells may overcome this drawback (World Health Organization 2018).

7.3.3 Use of WGS in Identification of Foodborne Pathogens

The standardized method to monitor the susceptibility of bacteria against antimicrobial agents depends on various phenotypic tests which are later analyzed under different standardized assay guidelines, such as the European Committee on Antimicrobial Susceptibility Testing (EUCAST) (Kahlmeter et al. 2003). However, phenotypic tests don't reveal the genetic characteristics of the pathogen such as mobile genetic elements and genetic determinants conveying resistance which may enable their proliferation across the global. In this regard, WGS technique is not only helpful in early detection or epidemiological investigation of bacteria (Köser et al.

2014) but also shows usefulness to overcome the limitation of phenotypic tests. (Ellington et al. 2017). Also, it is important to mention that albeit numerous genomic studies have been conducted in relation to AMR since 2010, but only 12.6% of all WGS-related peer-reviewed publications documented the application of WGS against AMR food-related samples. Indeed, a few publications also covered the application of WGS to discover the resistance phenotypic profiles of some strains. For instance, analysis of 200 strains of *Salmonella enterica* serovar Typhimurium, *Escherichia coli*, *Enterococcus faecalis* and *Enterococcus faecium* strains isolated from Danish pigs revealed high concordance (99.74%) between predicted antimicrobial susceptibilities and phenotypes (Zankari et al. 2012). Hence, such reports provide the evidence and support the use of WGS to predict vulnerabilities within surveillance schemes. Moreover, use of WGS technique in monitoring the quality of food against AMR strains has been recently emerged with major focus on identification and characterization of foodborne pathogens such as *Campylobacter* spp. (Chen et al. 2013), *Listeria monocytogenes* or *Staphylococcus aureus* (Gordon et al. 2014), *Salmonella* (Allard et al. 2012) and Shiga toxin-producing *Escherichia coli* (Dallman et al. 2015). Also, some other bacterial species has been focused and accomplished that are not typically identified as foodborne illness such as *Klebsiella* sp. and *Enterococcus* sp. (Davis et al. 2015). Hence, essential initiatives have been made to distinguish the resistant bacterial isolates from the possible respective sources (Allard 2016), identification of genetic elements responsible for resistance or AMR mechanisms (Carroll et al. 2017) and source of infection in regard to food-related outbreaks that occurred due to AMR bacterial strains (Moran-Gilad 2017) as well as pursue the distribution of AMR microbes by resistant gene transfer (Karkman et al. 2018).

Herin, we have briefly summarized the recent advancements of WGS based prediction for the four major well-characterized foodborne AMR species i.e. *Salmonella* spp., *Campylobacter* spp., *Listeria monocytogenes* and *Escherichia coli* using WGS-based approaches.

7.3.3.1 Whole Genome Sequencing of *Salmonella* spp.

Salmonella spp. is reported as a major food pathogen in humans that causes gastroenteritis, and in 2016, the European Union (EU) reported a total of 94,530 cases by salmonellosis (European Food Safety Authority and European Centre for Disease Prevention and Control 2017). The different serotypes for *Salmonella* spp. have been reported that can be put on scale according to their frequency in humans on decreasing order as (i) *S. enterica* serovar Enteritidis, (ii) *S. enterica* serovar Typhimurium, (iii) monophasic *S. enterica* serovar Typhimurium, (iv) *S. enterica* serovar Infantis, and (v) *S. enterica* serovar Derby (European Food Safety Authority and European Centre for Disease Prevention and Control 2017). Moreover, multidrug resistance strains of *Salmonella* spp. have been frequently reported that were associated with feasting of raw undercooked or raw meat products (Hoffmann et al. 2014). Also, emergence of such serotypes was corelated with the practice of biocides, such as triclosan, and reported to reduce the antibiotics susceptibility by certain genes overexpression or mutations in *S. enterica* serovar Typhimurium

(Gantzhorn et al. 2015). Also, analysis of different *S. enterica* serovar Montevideo strains collected from repetitive contamination events as well as other sources such as laboratory, environmental, food or clinical samples studied by WGS show its potential in the molecular typing of *S. enterica* serovar Montevideo (Allard et al. 2012). Hence, use of WGS technique in salmonellosis outbreak epidemiological or surveillance investigations in food or food products has been preferred over other available techniques (Allard et al. 2018). Following use of WGS technique and phenotypic characterization of 113 *S. enterica* serovar Heidelberg isolates from poultry carcasses at the abattoir (n = 18), poultry meat (n = 44) and humans (n = 51) revealed the intertransmission of *S. enterica* serovar Heidelberg (Edirmanasinghe et al. 2017). This study also discovered the transmission of microbial resistance linked common AMR plasmid (CMY-2) against beta-lactam antibiotics among different genetic backgrounds *of S. enterica* serovar Heidelberg strains (Edirmanasinghe et al. 2017). Moreover, WGS data analysis discovered the blaCMY-2 gene located on largest contig (747 kb) equivalent to chromosome in two S. *enterica* serovar Heidelberg isolates, indicates the integration of CMY-2 plasmid in the bacterial chromosomal DNA of two strains. Additionally, 10 plasmid subtypes were also discovered from the sequence analysis that exhibited high homology to the *S. enterica* serovar blaCMY-2 beta-lactamase gene and Kentucky pCVM29188_101 plasmid (Edirmanasinghe et al. 2017). Another WGS analysis was also conducted on 288 collected *S. enterica* serovar *Typhimurium* isolates of different sources including food samples from four continents in 31 countries between 1911 and 1969 year (Tran-Dien et al. 2018). This reconsidering study emphasized that 4 % of the strains carried several beta-lactamase genes, including blaTEM-1, on different plasmids comprising *S. enterica* serovar Typhimurium virulence plasmid that induced resistance against the ampicillin. Moreover, a total of three groups with eleven ampicillin-resistant *S. enterica* serovar Typhimurium genomes were clustered. The major group contained seven ampicillin-resistant isolates from different sources including food samples in France (1959–1969) and embraced four types of beta-lactamase genes (blaOXA-1, blaOXA-2, blaTEM-1A, and blaTEM-1B). This study concluded that the multiple independent attainments of blaTEM gene-carrying plasmids in different bacterial strains caused the emergence and diffusion of ampicillin resistance in *S. enterica* serovar Typhimurium over the years (Tran-Dien et al. 2018).

7.3.3.2 Whole Genome Sequencing of *Campylobacter* spp.

Campylobacter spp. is another major pathogen reported to cause foodborne diarrhea in humans, while *C. jejuni* and *C. coli* are majorly responsible for campylobacteriosis and intensively studied using WGS (Yao et al. 2017). Most of the isolated strains of the *Campylobacter* spp. were reported to attain resistance against commonly used antimicrobials like aminoglycoside and gentamicin (Qin et al. 2012) and beta-lactams and fluoroquinolones (Liu et al. 2016). The strains of *Campylobacter spp.* have been frequently isolated from various food products such as raw meat from poultry as well as raw milk (Liu et al. 2016). Recently, WGS application was applied to assess the phylogenetic relationship between the

non-resistant and resistant strains of *Campylobacter* spp. Such studies revealed the role of WGS to expose the function of pTet-like plasmids such as pN29710-1 which possess properties of self-transmissible as well as insertion of multi-antibiotic-resistant gene and provide the drug resistant in wild bacterial strains (Chen et al. 2013). In another study conducted under the National Antimicrobial Resistance Monitoring System (NARMS) surveillance program, 114 strains of *Campylobacter* spp. (82 *C. coli* and 32 *C. jejuni*) were isolated and analyzed from different sources including food products between 2000 and 2013 in the United States (Zhao et al. 2016). It was reported that a total of 18 AMR genes (*blaOXA-61*, *tet(O)*, *lnu(C)*, *catA*, *aph(2")-Ic*, *aph(2")-Ib*, *aph(20)-If*, *aph(2")-Ih*, *aac(60)-Ie/aph(2")-Ia*, *aph(2")-Ig*, *aac(60)-Im*, *aadE*, *sat4*, *aac(60)-Ie/aph(2")-If*, *aad9*, *ant(60)*, *aph(30)-IIIa* and *aph(30)-Ic)*) as well as genetic alternations in *gyrA* and 23S rRNA house-keeping genes was found responsible for the resistance against different antibiotic families including aminoglycosides, tetracycline, lincomycin, chloramphenicol and beta-lactams (Zhao et al. 2016). Moreover, high degree of correlation between given antibiotic and phenotypic classified resistant strain, and incidence of respective antibiotic-resistant genes indicates the potential of WGS technique to accurately classified the antibiotic-resistant phenotypes of *Campylobacter* spp. (Zhao et al. 2016).

7.3.3.3 Whole Genome Sequencing of *Listeria monocytogenes*

Listeria monocytogenes, a zoonotic agent, is another foodborne bacterial pathogen that is generally spread by feasting of contaminated food and food products including dairy products, meat and vegetables such as radishes, cabbage and fresh lettuce (Gandhi and Chikindas 2007). Additionally, these pathogens easily inhabit and persist in the food processing as well as production facilities for years, and such conditions are sometimes attributed to the development of resistance in them against antimicrobials and other disinfectants. Hence, many authors have reported the utilization of WGS technique in the genetic characterization of their respective adaptability and factors responsible for their virulence, with major concern over the listeriosis outbreaks (Lim et al. 2016). Recently, 520 *L. monocytogenes* strains isolated from various sources including isolates from food samples were analyzed using WGS at the Microbiological Diagnostic Unit Public Health Laboratory of Australia and identified the divergent nested clusters within groups of isolated strains which were indistinguishable by other current accessible typing methods (Kwong et al. 2016). Moreover, this study also predicted the point in genetic variance that leads to *L. monocytogenes* outbreaks. In another similar attempt, WGS was used in the determination of *L. monocytogenes* for serogroup (Hyden et al. 2016). WGS has demonstrated the potential in the surveillance of *L. monocytogenes*, and application of WGS has been recently used to identify the AMR in *L. monocytogenes* isolates from both clinical and food products. Also, WGS was employed to identify the factors associated with persistence of *L. monocytogenes* at food processing units and contribution of antimicrobial agents or other disinfectant in the survival status (Ortiz et al. 2016). In this context, investigation was conducted on the two meat processing units at two places (Plant A and B) in Spain that included a pork processing unit

(Plant A deals with the fresh and cured meat products as well as export the product to the United States) and processed food product storage unit received from plant A (plant B) (Ortiz et al. 2016). Following WGS analysis of the isolated strains, it was concluded that several isolates had resistant against benzalkonium chloride (BAC), related to PFGE types S1 and S10-1 strains from plant A and S2-2, S2-3 and S10-3 strains from plant B. This study concluded that isolated BAC-resistant strains also shared large genomic differences i.e. S1 PFGE type (ST31 by MLST) isolates contained *inlA* and *prfA* genes with mutations that lead to low-virulence type and presence of stress survival islet1 (SSI-1) for the facilitated environmental persistence. Whilst the other four BAC-resistant isolates were linked to the ST121 type that possessed the various stress resistance and AMR genes. Also, genome analysis of ST121 strains revealed the presence of Tn6188 transposon which provides resistance against quaternary ammonium compounds (QAC), Tn5422 transposon was detected for cadmium resistance and *clpL* genes were also predicted to engage in stress response.

Moreover, analysis of two indistinguishable *L. monocytogenes* isolates via PFGE revealed the presence of genes responsible for non-specific multidrug efflux pumps; resistance gene against antibiotics such as lincomycin, tetracycline, beta-lactams and quinolone as well as various genes responsible for non-specific heavy metal resistance (Fox et al. 2017). Also, this study demonstrated the potential discriminatory utility of WGS for genomes with 99% nucleotide sequence identity against other typing techniques such as MLST or PFGE. More recently, 100 *L. monocytogenes* isolates collected at different sources including dairy farm environment (1), seafood (Allard 2016), meat (Allard et al. 2018), food (Allard et al. 2018), vegetables (Allard et al. 2018) and dairy (Wilson et al. 2018) from Australian food production chains between 1988 and 2016 were analyzed for their AMR profile using drafted genomes (Wilson et al. 2018). The analysis of data revealed the presence of AMR genetic markers for lincomycin resistance (*lmrB*) and fosfomycin resistance (*fosX*) genes while one strain (Lm16-001) was identified to contain erythromycin resistance gene (*ermB*) that can be linked with the phenotypic observations on erythromycin resistance.

7.3.3.4 Whole Genome Sequencing of *Escherichia coli*

In recent decades, *E. coli*, is well known as endogenous microbiota of humans, has been classified as another foodborne pathogen predominantly correlated with the consumption of meat and meat products and hence, also identified as reservoir of AMR gene in the food chain (Hussain et al. 2017). It was reported that *E. coli* strains isolated from free-range chicken outlets at various cities in different parts of India showed antimicrobial susceptibility pattern (Hussain et al. 2017). Moreover, 168 *E. coli* isolates in this study were also analyzed through WGS against strains from human *E. coli* pathotypes to establish the genetic relationship, and establish the two lineages of emergent *E. coli* human pathogens (Hussain et al. 2017). Similar study put major interest stress particular on Shiga toxin-producing *E. coli* (STEC) strains due to their capability to harbor novel antibiotic resistance plasmids (Losada et al. 2016). Among the twenty-six isolated *E. coli* strains (twenty-two STEC and

four non-STEC strains) from various sources and analyzed through WGS lead to the identification of 39 new plasmids which includes two plasmids with six genes linked to the resistance against major antibiotics such as cephalosporins, carbapenems, penicillins, aminoglycosides, chloramphenicol, sulphonamides and tetracyclines. Additionally, a pair of novel IncHI2 plasmids was also identified from the acquisition of AMR genes (Losada et al. 2016). Another study reported the three *E. coli* strains isolated from dairy cattle in the Chinese Jiangsu province with *blaNDM-5* gene located on a pNDM-MGR194-like plasmid identified through WGS application. (He et al. 2017). Interestingly, one of the isolated *E. coli* strain was also reported to possess *mcr-1* colistin resistance gene coexisted with *blaNDM-5* gene in the plasmid (He et al. 2017).

7.3.4 Advantages and Drawbacks Associated with Use of WGS in Food Safety Management

7.3.4.1 Key Advantages

The major benefits accompanying WGS use in surveillance and safety management for food are discussed as under:

(a) **Sensitivity and specificity:** Due to the generation of virtual whole genome sequence for the target organism in WGS, it provides the clear and sensitive information on the pathogen under consideration against conventional technologies. This information can be used specifically to elucidate the linkage of isolated strain with environmental or food and human cases that results in strong prediction of source of pathogen or illness. This helps regulatory authorities to act against outbreak with more pointed approach which includes restricting the distribution of affected food products as well as safety precautions in the size of outbreaks.

(b) **Cost:** The sequence data generated from the application of WGS for the pathogen can be used to distinguish between different serotypes, virulence factors and AMR of an isolate. Hence, this approch is cost-effective against traditional methods where different approaches are required for the identification of different serotypes for the same pathogen (Joensen et al. 2014).

(c) **Speed**: WGS technique can generate the data for the target pathogen within few days under optimized conditions which is faster against the current traditional approaches. Moreover, the complete analysis of the generated data by WGS can be useful in complete characterization of the pathogen, which provides better evidences for tracking and cause of outbreak in short time. Additionally, WGS provides specific links among the isolated pathogen with few cases and hence, suggested as useful in the determination of putative epidemics. WGS is associated with real-time and speed characteristics against the traditional method that ensure appropriate steps can be taken rapidly in food safety management and to protect public health.

(d) **Universality**: Unlike traditional methods which required sophisticated laboratories for the identification and typing methodologies of specific pathogen species, WGS universality is associated with cost-effectiveness and time efficiency. Such universal technology was suggested to be important especially in developing countries of food safety management.
(e) **Simple learning and efficiency of use**: WGS techniques are not complicated to learn and use like conventional methods such as PFGE. Additionally, its limited experimental work as numerous pathogens can be put for analysis simultaneously in a single laboratory.
(f) **Ease of sharing**: WGS provides the virtual sequence data of the pathogen that can be easily shared across the globe through electronic communication and even can be deposited in the relevant database. The data from such repositories can be retrieved and reanalyzed across the globe at any time.
(g) **Elastic and open to reanalysis**: The genome data is the most basic information decoded by the WGS for an organism. Interestingly, several sequencing and analytical bioinformatic platforms can be used for the generated genome at the same time. With the advancement and emergence of new sequencing methods or bioinformatics analytical approaches, previously generated sequence data can be employed for comparison against newly generated data through new approaches or historical data can be reanalyzed with newly developed bioinformatic methods.
(h) **Easier access to trade and markets**: The application of WGS is likely to help competent agencies with certification to follow standard international trade practices and treaties. Such practices will result in increased confidence in trade associates for the nation's food safety and management system.

7.3.5 Potential Disadvantages

Like other introduced innovations and techniques, WGS also comes with certain disadvantages while applied to food safety control management which are as follows:

(a) **Cost**: Although with the introduction of new sequencing methods that lead to reduction in WGS cost and likely to decline down in the near future, but application of WGS may be restricted in some developing countries due to real cost of consumables and equipment required in sequencing. Moreover, lack of established outbreak surveillance system in some countries to supply isolates for sequencing further decreases the cost-effectiveness of WGS approach. Such situations predominantly can be observed in developing countries because application of WGS may distract the vital resources from more persistent priorities such as establishment of basic surveillance and monitoring systems, improvement in food safety circumstances or amended water quality. However, some developing countries may select the implication of current subtyping

approaches for a while along with establishment of WGS, which will probably increase the surplus burden to both financial resources and human.

(b) **Data storage**: The application of WGS produced data in range of tera- and petabytes which requires both virtual and physical space estimated to be costly for storage in local data repositories. Though such data can be deposit in the global data repositories and make them publicly as conceivable solution, but again such databases demands for well-precise global data distribution mechanism.

(c) **Requirement of infrastructure and high-speed Internet connection**: The distribution of WGS data through global repositories required high-speed Internet facilities for the benefits of global community. In this context, limited Internet packages or bandwidth as well as interruption in the power along with limited infrastructure facilities can be a major drawback in many developing countries.

(d) **WGS sequence data and interpretation**: Most of the laboratories in developing countries do not have well-skilled bioinformaticians and thus, cannot fully yield the benefits of WGS application from their generated data. Such hindrance can be solved by doing analysis on online servers and platforms or by collaboration with experienced researchers in WGS analysis. However, these possible steps can be assumed when the data is available on global databases along with sufficient Internet connectivity. Moreover, under the suitable internet connectivity to essential genomics software or online servers, sometimes the data interpretation may not be easy such as when doing interpretation with relevance epidemiological information. In such occasions, training of both microbiologists conducting the WGS and end user of the data plays a significant part in the application of this sequencing technology.

(e) **Sustainability**: WGS may not be sustained when socioeconomic and local advantages are not well transferred and presented.

(f) **Probable unwarranted trade break**: Let us suppose that some countries used the WGS application in food safety system and such countries domineering the same application on imported food. But it may be possible that some developing countries cannot afford same level of WGS analysis on their exported food due to limited capacity and resources with them. Thus, under such conditions, one partner may not be able to do shared business with trade partners with abundant resources. Hence, a common agreement at the global level is required for the countries with established WGS application in food safety and security to assist the countries who lack the technology due to limited capacity and resources.

7.4 Conclusion

WGS is a well-established technique with great potential in the food safety and surveillance against foodborne pathogens and AMR, but its implementation in the respective routine laboratory is restricted to countable countries with established WGS approaches in epidemiology and public health surveillance (Sekse et al. 2017). In this regard, the Food and Drug Administration (FDA) has developed a database

from the collection of foodborne pathogens as Genome Trakr, with an ambition to assist the scientists to recognize the food source of an outbreak (Allard et al. 2016). Moreover, culture-independent data for the foodborne pathogens generated by WGS can be used to monitor the occurrence and distribution of AMR determinants in a range of environments including food and food products. Further, this information can be compiled in metadata collected from food and clinical samples that allows the implication of quantitative risk assessment frameworks for modelling resistance determinants distribution and occurrence. For instance, prediction of AMR distribution in common environmental niches has been thoroughly studied (Amos et al. 2015). Due to significant established role of WGS technique for surveillance programs, their monotonous usage demands its transformation into cost-effective and user-friendly approaches for real-time application on site by personnel who is not skilled in big data management (Oniciuc et al. 2018). To fulfill such requirements for easy use of WGS, recently miniaturized prototypes were developed that showed benefits of onsite application and generating the results in real time (Oniciuc et al. 2018). Also, due to availability of publicly available AMR databases specially designed for the food microbiology ecosystems, updated in the real time and freely accessible to the global, can be used to improve the exploitation of these molecular tools (Taboada et al. 2017).

References

Allard MW (2016) The future of whole genome sequencing for public health and the clinic. J Clin Microbiol:01082–01016

Allard MW, Bell R, Ferreira CM, Gonzalez-Escalona N, Hoffmann M, Muruvanda T, Ottesen A, Ramachandran P, Reed E, Sharma S (2018) Genomics of foodborne pathogens for microbial food safety. Curr Opin Biotechnol 49:224–229

Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW (2012) High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 13(1):32

Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, Timme R (2016) Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol 54(8):1975–1983

Amos GC, Gozzard E, Carter CE, Mead A, Bowes MJ, Hawkey PM, Zhang L, Singer AC, Gaze WH, Wellington EM (2015) Validated predictive modelling of the environmental resistome. ISME J 9(6):1467

European Food Safety Authority, European Centre for Disease Prevention and Control (2017) The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016. EFSA J 15(12):e05077

Brambilla G, Testa C (2014) Food safety/food security aspects related to the environmental release of pharmaceuticals. Chemosphere 115:81–87

Carroll LM, Wiedmann M, den Bakker H, Siler J, Warchocki S, Kent D, Lyalina S, Davis M, Sischo W, Besser T (2017) Whole-genome sequencing of drug-resistant Salmonella enterica isolated from dairy cattle and humans in New York and Washington states reveals source and geographic associations. Appl Environ Microbiol AEM:00140–00117

Chen Y, Mukherjee S, Hoffmann M, Kotewicz ML, Young S, Abbott J, Luo Y, Davidson MK, Allard M, McDermott P, Zhao S (2013) Whole-genome sequencing of gentamicin-resistant

Campylobacter coli isolated from U.S. retail meats reveals novel plasmid-mediated aminoglycoside resistance genes. Antimicrob Agents Chemother 57(11):5398–5405

Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A (2015) Whole-genome sequencing for national surveillance of shiga toxin–producing Escherichia coli O157. Clin Infect Dis 61(3):305–312

Davis GS, Waits K, Nordstrom L, Weaver B, Aziz M, Gauld L, Grande H, Bigler R, Horwinski J, Porter S (2015) Intermingled Klebsiella pneumoniae populations between retail meats and human urinary tract infections. Clin Infect Dis 61(6):892–899

Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, Kooistra-Smid AM, Raangs EC, Rosema S, Veloo AC (2017) Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 243:16–24

Edirmanasinghe R, Finley R, Parmley EJ, Avery BP, Carson C, Bekal S, Golding G, Mulvey MR (2017) A whole-genome sequencing approach to study cefoxitin-resistant salmonella enterica serovar heidelberg isolates from various sources. Antimicrob Agents Chemother 61(4)

Ellington M, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, Grundman H, Hasman H, Holden M, Hopkins KL (2017) The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect 23(1):2–22

Fox EM, Casey A, Jordan K, Coffey A, Gahan CG, McAuliffe O (2017) Whole genome sequence analysis; an improved technology that identifies underlying genotypic differences between closely related Listeria monocytogenes strains. Innovative Food Sci Emerg Technol 44:89–96

Franz E, Gras LM, Dallman T (2016) Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. Curr Opin Food Sci 8:74–79

Gandhi M, Chikindas ML (2007) Listeria: a foodborne pathogen that knows how to survive. Int J Food Microbiol 113(1):1–15

Gantzhorn MR, Olsen JE, Thomsen LE (2015) Importance of sigma factor mutations in increased triclosan resistance in Salmonella Typhimurium. BMC Microbiol 15(1):105

Gašo-Sokač D, Kovač S, Josić D (2010) Application of proteomics in food technology and food biotechnology: process development, quality control and product safety. Food Technol Biotechnol 48(3)

Giacometti J, Tomljanović AB, Josić D (2013) Application of proteomics and metabolomics for investigation of food toxins. Food Res Int 54(1):1042–1051

Gordon N, Price J, Cole K, Everitt R, Morgan M, Finney J, Kearns A, Pichon B, Young B, Wilson D (2014) Prediction of Staphylococcus aureus antimicrobial resistance from whole genome sequencing. J Clin Microbiol JCM:03117–03113

Havelaar AH, Brul S, De Jong A, De Jonge R, Zwietering MH, Ter Kuile BH (2010) Future challenges to microbial food safety. Int J Food Microbiol 139:S79–S94

He T, Wei R, Zhang L, Sun L, Pang M, Wang R, Wang Y (2017) Characterization of NDM-5-positive extensively resistant Escherichia coli isolates from dairy cows. Vet Microbiol 207:153–158

Hoffmann M, Zhao S, Pettengill J, Luo Y, Monday SR, Abbott J, Ayers SL, Cinar HN, Muruvanda T, Li C (2014) Comparative genomic analysis and virulence differences in closely related Salmonella enterica serotype Heidelberg isolates from humans, retail meats, and animals. Genome Biol Evol 6(5):1046–1068

Hussain A, Shaik S, Ranjan A, Nandanwar N, Tiwari SK, Majid M, Baddam R, Qureshi IA, Semmler T, Wieler LH (2017) Risk of transmission of antimicrobial resistant Escherichia coli from commercial broiler and free-range retail chicken in India. Front Microbiol 8:2120

Hyden P, Pietzka A, Lennkh A, Murer A, Springer B, Blaschitz M, Indra A, Huhulescu S, Allerberger F, Ruppitsch W (2016) Whole genome sequence-based serogrouping of Listeria monocytogenes isolates. J Biotechnol 235:181–186

Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM (2014) Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol 52(5):1501–1510

Joseph S, Forsythe S (2012) Insights into the emergent bacterial pathogen Cronobacter spp., generated by multilocus sequence typing and analysis. Front Microbiol 3:397

Kahlmeter G, Brown DF, Goldstein FW, MacGowan AP, Mouton JW, Österlund A, Rodloff A, Steinbakk M, Urbaskova P, Vatopoulos A (2003) European harmonization of MIC breakpoints for antimicrobial susceptibility testing of bacteria. J Antimicrob Chemother 52(2):145–148

Karkman A, Do TT, Walsh F, Virta MP (2018) Antibiotic-resistance genes in waste water. Trends Microbiol 26(3):220–228

Köser CU, Ellington MJ, Peacock SJ (2014) Whole-genome sequencing to control antimicrobial resistance. Trends Genet 30(9):401–407

Kuchenmüller T, Hird S, Stein C, Kramarz P, Nanda A, Havelaar A (2009) Estimating the global burden of foodborne diseases-a collaborative effort. Eurosurveillance 14(18):19195

Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP (2016) Prospective whole-genome sequencing enhances national surveillance of Listeria monocytogenes. J Clin Microbiol 54(2):333–342

Lim SY, Yap K-P, Thong KL (2016) Comparative genomics analyses revealed two virulent Listeria monocytogenes strains isolated from ready-to-eat food. Gut Pathog 8(1):65

Liu KC, Jinneman KC, Neal-McKinney J, Wu WH, Rice DH (2016) Genome sequencing and annotation of a Campylobacter coli strain isolated from milk with multidrug resistance. Genom Data 8:123–125

Losada L, DebRoy C, Radune D, Kim M, Sanka R, Brinkac L, Kariyawasam S, Shelton D, Fratamico PM, Kapur V (2016) Whole genome sequencing of diverse Shiga toxin-producing and non-producing Escherichia coli strains reveals a variety of virulence and novel antibiotic resistance plasmids. Plasmid 83:8–11

Moran-Gilad J (2017) Whole genome sequencing (WGS) for food-borne pathogen surveillance and control–taking the pulse. Eurosurveillance 22(23)

Oniciuc E, Likotrafiti E, Alvarez-Molina A, Prieto M, Santos J, Alvarez-Ordóñez A (2018) The present and future of Whole Genome Sequencing (WGS) and Whole Metagenome Sequencing (WMS) for surveillance of antimicrobial resistant microorganisms and antimicrobial resistance genes across the food chain. Genes 9(5):268

World Health Organization (2018) Whole genome sequencing for foodborne disease surveillance: landscape paper. WHO, Geneva

Ortiz S, López-Alonso V, Rodríguez P, Martínez-Suárez JV (2016) The connection between persistent, disinfectant-resistant Listeria monocytogenes strains from two geographically separate Iberian pork processing plants: evidence from comparative genome analysis. Appl Environ Microbiol 82(1):308–317

Piras C, Roncada P, Rodrigues PM, Bonizzi L, Soggiu A (2016) Proteomics in food: quality, safety, microbes, and allergens. Proteomics 16(5):799–815

Qin SS, Wang Y, Zhang QJ, Chen X, Shen ZQ, Deng FR, Wu CM, Shen JZ (2012) Identification of a novel genomic island conferring resistance to multiple aminoglycoside antibiotics in campylobacter coli. Antimicrob Agents Chemother 56(10):5332–5339

Ranieri ML, Shi C, Switt AIM, Den Bakker HC, Wiedmann M (2013) Comparison of typing methods with a new procedure based on sequence characterization for Salmonella serovar prediction. J Clin Microbiol 51(6):1786–1797

Sekse C, Holst-Jensen A, Dobrindt U, Johannessen GS, Li W, Spilsberg B, Shi J (2017) High throughput sequencing for detection of foodborne pathogens. Front Microbiol 8:2029

Seltenrich N (2015) New link in the food chain? Marine plastic pollution and seafood safety. Environ Health Perspect 123(2):A34

Smid E, Kleerebezem M (2014) Production of aroma compounds in lactic fermentations. Annu Rev Food Sci Technol 5:313–326

Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, Force CPT (2001) PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis 7(3):382

Taboada EN, Graham MR, Carriço JA, Van Domselaar G (2017) Food safety in the age of next generation sequencing, bioinformatics, and open data access. Front Microbiol 8:909

Tran-Dien A, Le Hello S, Bouchier C, Weill FX (2018) Early transmissible ampicillin resistance in zoonotic Salmonella enterica serotype Typhimurium in the late 1950s: a retrospective, whole-genome sequencing study. Lancet Infect Dis 18(2):207–214

Wilson A, Gray J, Chandry PS, Fox EM (2018) Phenotypic and genotypic analysis of antimicrobial resistance among listeria monocytogenes isolated from Australian food production chains. Genes 9(2):80

Yao H, Liu DJ, Wang Y, Zhang QJ, Shen ZQ (2017) High prevalence and predominance of the aph (2″)-if gene conferring aminoglycoside resistance in campylobacter. Antimicrob Agents Chemother 61(5)

Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, Larsen MV, Aarestrup FM (2012) Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. J Antimicrob Chemother 68 (4):771–777

Zhao S, Tyson GH, Chen Y, Li C, Mukherjee S, Young S, Lam C, Folster JP, Whichard JM, McDermott PF (2016) Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in campylobacter spp. Appl Environ Microbiol 82(2):459–466

Zuker CS (2015) Food for the brain. Cell 161(1):9–11

Functional Metagenomics for Rhizospheric Soil in Agricultural Systems

8

Estefanía Garibay-Valdez, Kadiya Calderón, Francisco Vargas-Albores, Asunción Lago-Lestón, Luis Rafael Martínez-Córdova, and Marcel Martínez-Porchas

Abstract

The study of plant microbiota has been stimulated by recognizing the fundamental role that the genetic capacity of the associated microbial communities has to modulate the phenotypic expression of plants, which is crucial for its health, physiology, and productivity. All genes in a metagenome can be described by whole-metagenome shotgun sequencing, but it is time-consuming, and a high level of experience is required. Alternatively, the amplification and high-throughput sequencing of the16S-rRNA gene allow describing the microbial composition. Then functional activities can be inferred by listing the abundance of each gene. Also, the identification and the quantification of microbiome transcripts are now accessible to determine the profile and changes in gene expression occurring in a microbial community in response to environmental or experimental variations. Considering the beneficial role of microbial communities in soil environments, it is important increasing the understanding of plant-microbe relationships to provide biotechnological information for control and management with sustainable practices.

Authors Estefanía Garibay-Valdez, Kadiya Calderón have been equally contributed to this chapter.

E. Garibay-Valdez · F. Vargas-Albores · M. Martínez-Porchas (✉)
Centro de Investigación en Alimentación y Desarrollo, A.C. Coordinación de Tecnología de Alimentos de Origen Animal, Hermosillo, Sonora, Mexico
e-mail: marcel@ciad.mx

K. Calderón (✉) · L. R. Martínez-Córdova
Departamento de Investigaciones Científicas y Tecnológicas de la Universidad de Sonora, Universidad de Sonora, Hermosillo, Sonora, Mexico
e-mail: kadiya.calderon@unison.mx

A. Lago-Lestón
Centro de Investigación Científica y de Educción Superior de Ensenada, Ensenada, BC, Mexico

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_8

8.1 Introduction

Agriculture constitutes a cornerstone activity to produce food at large-scale and also biomaterials for biotechnological purposes (Béné et al. 2016; Jones and Ejeta 2016). However, the increasing food demand together with the environmental impacts caused by agro-industries represents a challenge that requires more intensive yet environmentally safe production, as well as the sustainability of the resources involved in this activity (Gebbers and Adamchuk 2010).

Microbial communities play essential roles for the excellent performance of agriculture through the decomposition, solubilization, and recycling of nutrients and toxic compounds, competing against pathogens and producing desirable physicochemical and biological conditions for the cultured species. Plants are colonized by a number of microbes that together can reach higher cell loads than the plant itself. Most of these microbes thriving in the rhizosphere and nearby are directly or indirectly associated to the plant (Mendes et al. 2013), which can be considered as a harboring host and neighbor mutualist body for those microbes that are not in contact with the plant but interacting with it. Beneficial rhizospheric microbes can alter the morphology of plants while enhancing their growth and increasing mineral availability (Lakshmanan et al. 2014). Microbiome soil in plants is also fundamental because it can secrete growth hormones while inducing the immune system. Diseases in crops or other food-producing environments are sometimes strongly linked with the changes in the environmental microbiome.

The study and characterization of the microbiota associated with plants have recently considered the genetic stock of microorganisms as endophytes and epiphytes (in and on plants, respectively), as an extension of the host genome and with a fundamental role in its phenotype. The sum of all genetic information or hologenome (Theis et al. 2016) enables the adaptation process to new or changing environmental conditions and the ability to resist pathogens. However, some circumstances could affect these microbes.

Anthropogenic activities causing changes in soil affect both the soil microbiome and the organisms living in and on them. These changes can consequently alter the functional profile of the microbiome, affecting the occurrence and abundance of essential metabolic pathways that maintain a balance within the systems. Herein, there are three groups of microbes present in the rhizosphere, including commensal, beneficial, and pathogenic (Berendsen et al. 2012). Microbiomes can activate complex signaling pathways in response to biotic or abiotic stress, leading to localized and systemic defenses (Lakshmanan et al. 2014), but these functions may be impaired by the occurrence of pathogenic blooms. Therefore, understanding the taxonomic and functional microbiome profiles caused by the occurrence of microbial pathogens in food production systems may provide valuable information to understand the negative effects of these phenomena and devise strategies to alleviate the effects or eradicate the pathogens; in addition, crops can be engineered through their microbiota favoring desirable responses. Genomic disciplines, but particularly functional metagenomics, can provide adequate insights of microbiomes.

8.2 From Functional Metagenomics to Food Production

Microbes associated to plants are essential for their adequate development; in this regard, plant microbiome is a key determinant of their health, physiology, and productivity (Mendes et al. 2013; Berendsen et al. 2012), and at the same time, much of these microbes participate in biogeochemical cycles. Particularly, these microbes can influence seed germination, seedling vigor, plant growth and development, nutrition, diseases, and development (Qin et al. 2011; Mendes et al. 2013). At some extent, the functions of this microbiota can be extrapolated to that of animals, where these microbes are considered as an annexed organ or a host's genome extension, and therefore the entire comprehension of the biology of a plant also depends on the knowledge gathered about its microbiota. Microbes including bacteria, archaea, fungi, protozoa, and algae are usually part of these microbial consortia.

This is a symbiotic relationship where plants depend on specific functions of the microbes while producing photosynthetically fixed carbon and other exudate components for the microbes thriving in the spermosphere, phyllosphere, rhizosphere, and mycorrhizosphere (Mendes et al. 2013). Despite a considerable number of studies having demonstrated these associations, the diversity and complexity of these diversity and interaction networks suggest that our current knowledge about this subject is still limited. Understanding the plant microbiome is an unavoidable cornerstone not only to comprehend the biology of plants but also to identify microorganisms that can be exploited for biotechnological purposes (e.g., improving plant growth and health) or regulated to avoid undesirable scenarios (disease).

Advances in genomics science shed light to some biological processes as virulence or resistance to antibiotics of pathogenic microorganisms, which play important roles in agricultural system industries (Lazarevic and François 2013). Current technologies not only provide better pictures about the taxonomic structure of microbial communities but also show accurate predictions about the functional capabilities and even real-time activities of microbes.

The study of the microbiome from this perspective is biotechnologically relevant because the information associates some of the translational applications with basic biology. Despite metabolic and regulatory networks being difficult-to-approach subjects for microbiologists, this kind of information may serve to identify targets for molecular therapy, particularly microbiomes (Vargas-Albores et al. 2018). For example, crops affected by pathogenic bacteria may require alleviating strategies including the use of antibiotics. However, functional metagenomics can provide information about potential resistance to particular kinds of antibiotics and suggest a more targeted and efficient strategy (dos Santos et al. 2017), allowing the use of less antibiotic while alleviating the collateral damage to the environment. Nevertheless, achieving this level of understanding about particular features of a microbiome requires highly specialized sets of tools. Another studies have revealed that manipulating the soil microbiome may result in improved soil health and increased plant fertility (Chaparro et al. 2012); however, knowing the taxonomic profile and functional capabilities of soil and rhizospheric microbiota will provide greater certainty about the parameters or conditions to be manipulated.

In addition, rhizospheric soils may also be reservoirs of opportunistic pathogens; these can disrupt in host microbiota by competition for resources, release of antimicrobial compounds, or antagonism against beneficial microorganisms. Understanding pathogen functionality, health, and productivity of important agriculture species can be improved by elucidating the taxonomic diversity and functional potential of their respective microbiota. Pathogen outbreaks usually occur when the conditions allow their proliferation; this means that pathogens can either be introduced into the systems or can be part of the environmental microbiota in a dormant or fully active state but not being virulent (Mendes et al. 2013). Conditions in the outside-host environment are prone to fluctuate over time affecting the microbiota and the activity of some pathogens.

One of the main fluctuating conditions affecting the microbiota is the soil nutrient profile, which is not for the plant but rather for the microbiota; for example, N-fertilization is commonly used in several crops particularly for modern varieties; however, this fertilization reduces microbial biomass and diversity, and consequently some functional capabilities may be lost (Ramirez et al. 2010). These kinds of strategies may have short-time beneficial effect on production but may have consequences in the long term because the soil microbiota may lack adequate functional profile to stimulate the responses in the plants of future crops or inclusively may fail to suppress pathogens. In addition, these modern plant varieties may have lost their ability to support microbiomes that degrade organic nitrogen and solubilize mineral nutrients such as phosphorus (Wallenstein 2017). In this regard, taxonomic and functional metagenomics insights could offer clues for performing strategies minimizing this collateral damage to the rhizospheric microbiota. For example, Wallenstein (2017) suggests engineering rhizospheres through inoculants that form connections with the native microbiome or soil amendments that stimulate microbial activity (Fig. 8.1).

Rhizospheres are highly structured, and the microbiome-plant interconnections through different signaling pathways as well as nutrient interchange (including root exudations) constitute a common process during the life cycle of a plant. As crops have been historically selected considering traits for intensive management, these connections were decreased. However, Wallenstein (2017) argued: "in the future, a systems approach to rhizosphere engineering could restore some features of natural rhizospheres through soil amendments, inoculants and plant traits that support beneficial microbiomes" (Fig. 8.1).

The improvement and increased use of high-throughput sequencing methods and the availability of generating genome information have significantly contributed to achieving deep approaches in functional metagenomics (Loman and Pallen 2015). Functional metagenomics begins with the isolation of DNA from microbial samples of environmental systems (Lam et al. 2015). In this context, there are two pathways to understand how these genes are involved with the environment; the first one is performing a molecular shotgun sequencing or also denominated whole-metagenome shotgun sequencing (WMS) from genomic DNA fragments of a metagenome community. The second one requires the amplification and high-throughput sequencing of a taxonomic biomarker gene, for example, the16S-

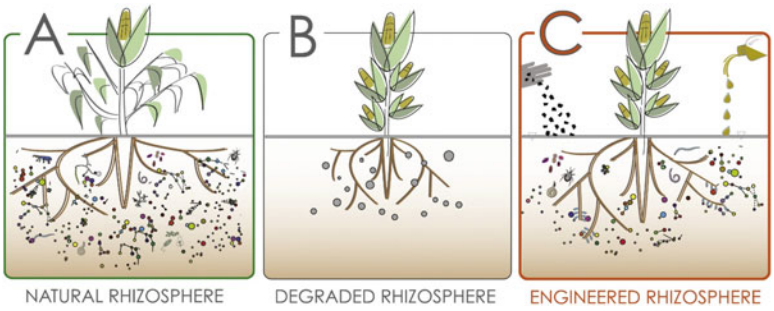

Fig. 8.1 Hypothesized conceptual models of rhizosphere systems in (**a**) natural ecosystems, (**b**) degraded systems such as those under intensive management and high fertilization, and (**c**) rhizospheres that have been engineered through inoculants that form connections with the native microbiome or soil amendments that stimulate microbial activity. (Figure obtained from Wallenstein 2017 (13). Rhizosphere 3(2):230–232 [w/ editorial & author permissions])

rRNA gene, avoiding the need of shotgun metagenomics while allowing, if required, the continuation of the metatranscriptomic and metaproteomic studies. Bioinformatics tools are absolutely required for any of these two approaches to process all of the recovered information (Morgan and Huttenhower 2012; Ortiz-Estrada et al. 2019).

8.3 Describing Microbial Communities Through the 16S rRNA Gene

Because the 16S rRNA gene is the most ubiquitous gene in the prokaryotic world and contains a combination of conserved and variable regions, it has been used for the classification of ribosomal RNA (rRNA) sequences and become a common approach for taxonomic identification of bacteria-forming complex communities (Glöckner et al. 2017). The 16S approach is mainly focused to elucidate the bacterial diversity, structure, and organization of microbiota as a community (Morgan and Huttenhower 2012). Nowadays, shotgun metagenomics is a very reliable technique involving the reconstruction of gene sets or genomes requiring a high level of experience and is time-consuming. For these and other reasons, studying microbial communities through 16S rRNA-targeted sequencing remains a valid approach.

Despite the many packages available on the market containing bioinformatics pipelines to analyze 16S rRNA gene amplicons, the most used open-code software tools are *Mothur* and Quantitative Insights Into Microbial Ecology (*QIIME*). These are based on a clustering-first approach, and the most recent pipelines including Kraken, CLARK, and One Codex use an assignment-first approach (Siegwald et al. 2017). These tools called pipelines are used to process the sequences by demultiplexing and quality filtering, classifying, aligning, and assigning sequences to operational taxonomic units (OTU) followed by the capability to process the information by microbial or ecological analyses like α- and β-diversity based on 16S rRNA gene (Glöckner et al. 2017; Almeida et al. 2018). The tools mentioned above

use SILVA as first-option reference datasets for taxonomic classification, but there are also other databases including the Greengenes rDNA, the NCBI Taxonomy, and the Ribosomal RNA Database project (RDP) (Balvočiūtė and Huson 2017).

Nonetheless, this is an approach to study bacteria and archaea but does not target eukaryotes. For targeting these last microbes, the 18S rRNA gene is required. This taxonomic biomarker gene is part of the structural RNA for the small component of eukaryotic cytoplasmic ribosomes. Unfortunately, the tools and protocols to study and valuate this biomarker are not as developed and extended as those for the 16S rRNA; however, there are some recommended software and useful pipelines to address this topic, particularly for amplification, sequencing, and taxonomic classification (Yang et al. 2013; Wang et al. 2014; Popovic and Parkinson 2018), whereas the rest of the statistical analyses have pretty much the same basis.

8.4 Drawing the Metabolic Potential

8.4.1 Targeted Metagenomics

A decade ago, providing a picture of the taxonomic profile and structure through 16S rRNA sequencing was considered as the limit to where this technique could lead; however, the feasibility of connecting databases containing information from simple biomarkers to those containing information on complete genomes made possible to infer the functional capabilities of the microbial community based in 16S rRNA sequencing. The correlation between taxonomy and metabolic function allows carrying out work focused on the identification of community structure and composition through analysis of taxonomic biomarker genes.

There are bioinformatics tools that construct an approach of functional metagenomics of a microbiome (Langille et al. 2013) (Fig. 8.2). These use the information from databases about the microbes constituting any niche and infer the metagenome of such microbial community and constructing a gene catalog with the abundance of different gene families. For example, tools like Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) are used for this purpose. In this kind of approach, the efficiency of functional predictions depends on the taxonomic classification accuracy, the genomes used as reference, and the information collected in databases from the closest ancestors. The databases of genes used for the metagenome construction are the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Clusters of Orthologous Groups of proteins (COG) (Kanehisa et al. 2008).

This, of course, is an imperfect association but can provide valuable information to evaluate factors influencing the functionality of microbial communities. One of the disadvantages of this approach is that this software uses the Greengenes database, which has not received actualizations since several years ago. However, there are other options such as Tax4fun, an R package using the SILVA database as

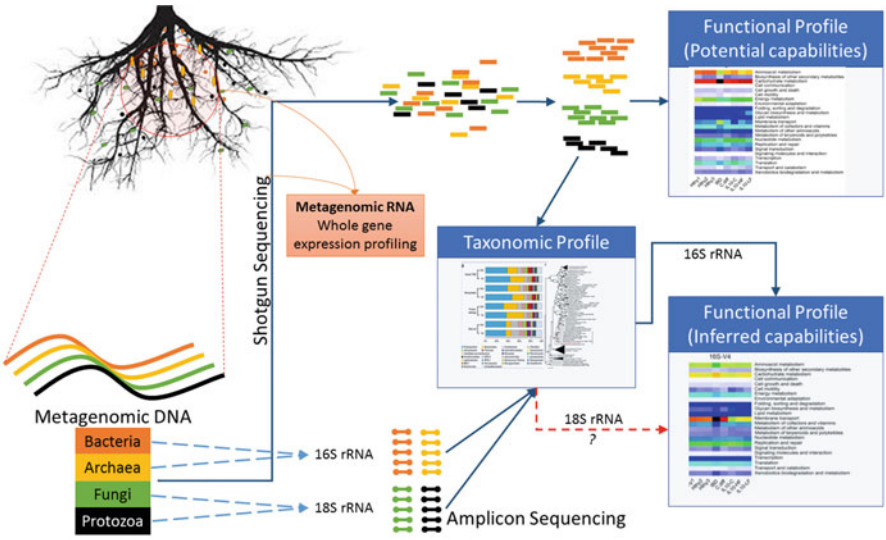

Fig. 8.2 General scheme for obtaining information about the taxonomic and functional profiles of microbiomes through metagenomics (shotgun sequencing) and targeted metagenomics (amplicon sequencing) approaches. To date, there are no bioinformatics tools for inferring functional profiles of eukaryotes using 18S rRNA; however, it is theoretically possible

reference and reported to be more accurate for some microbial communities (Aßhauer et al. 2015); however, neither PICRUSt nor Tax4fun has extensions to consider 18S rRNA data, and therefore the study of the microbiome may be incomplete.

Another issue to consider is that despite the easy generation of biomarker libraries, the manual exploration of the metabolisms associated with all observed OTUs is impractical; therefore, the researcher has no other option but to trust the gene family ordering process performed by the software.

On the other hand, there is evidence suggesting that in spite of the bias introduced by this imperfect correlation, shotgun metagenomics has validated the results obtained through targeted metagenomics using taxonomic biomarkers (Jovel et al. 2016).

Studies evaluating the rhizosphere have used these tools. Herein, Zhu et al. (2016) not only evaluated the effect of urea-based fertilizer rate on maize root exudation, the associated rhizosphere microbial community, and nitrogen-use-efficiency but also explored via PICRUSt the metagenomics contribution of bacterial OTUs obtained from 16S rRNA sequencing. Briefly, results revealed that some nitrifying and denitrifying genes were significantly influenced by the N rate, whereas some nitrogen-fixing and urease genes were not. However, on a total abundance basis, all N-cycle genes increased significantly as the N rate increased.

8.4.2 Metagenomics

The genomics discipline studying the genetic material recovered directly from environmental (or experimental) samples is known as metagenomics. To detect all of the genes contained in a microbial community, a shotgun metagenomics sequencing is performed, followed by annotations providing detailed output sets of metabolic and functional profiles (Vargas-Albores et al. 2018) (Fig. 8.2). Despite bioinformatics backgrounds being required to perform mandatory pipelines including quality control, assemblage, and annotation, there are user-friendly computing (public and private) software allowing the analysis of these data by most biologists. For instance, MG-RAST and CLC Genomics Workbench (Microbial Metagenomics Module) and other programs involve user-friendly end-user systems to perform complete mandatory pipeline processes. HUMAnN is another computational pipeline developed to determine the relative abundance of gene families and metabolic pathways from short-read sequence datasets (Aßhauer and Meinicke 2013), while STAMP is a friendly software to perform statistics of data obtained from targeted or shotgun metagenomics (Parks et al. 2014). There are plenty of free tools available; however, the sets of tools to use to establish a valid pipeline depend on the objectives of the study and the computing skills of the biologist.

Even though targeted metagenomics (16S rRNA) is an approach for the metagenome, shotgun sequencing offers not only more complete and precise results but also allows digging more into the functional potential of any microbial community, obtaining the genomic information from different taxa, and inferring metaproteomes from raw metagenomics sequences (Rooijers et al. 2011).

Studies related to agricultural sciences are still few but rapidly growing since microbiologists working in these disciplines are adopting these tools in an accelerated manner. Mendes et al. (2014) performed shotgun metagenomics to investigate the taxonomic and functional profiles of microbial communities in the bulk soil and in the rhizosphere of soybean plants and tested the validity of neutral and niche theories to explain the rhizosphere community assembly processes. This approach demonstrated that the assembly of the microbial community in the rhizosphere is based on niche-based processes as a result of a selective pressure exerted by the plant itself and other environmental factors.

Unlike targeted metagenomics, this approach can collect information about all of the microbial taxa thriving in the rhizosphere but requires additional processes for cleaning, ordering, and annotating sequences (Quince et al. 2017).

8.4.3 Metatranscriptomics

Transcription is the first phase of gene expression in which a particular segment of DNA is duplicated/transcribed into RNA language (especially mRNA). All of the living microbes thriving in any niche that is metabolically active are contributing to the activities of the microbiome. These functions can be estimated and tracked through metatranscriptomics.

The identification and quantification of microbiome transcripts are now possible. Metatranscriptomics is an "omics" discipline enabling researchers to determine the profile and changes in gene expression occurring in a microbial community in response to environmental or experimental variations. Metatranscriptomics estimates the function and activity of complete sets of transcripts (RNA-seq) obtained from any sample. Therefore, the "metatranscriptome" (messenger and noncoding RNAs) offers a vision framework about the regulatory networks and gene expression occurring at the time of sampling. This is a non-genomics discipline but that can complement and strengthen the results obtained through metagenomics.

Relevant biological processes depending on small RNAs such as virulence, quorum sensing, and stress/immune responses can be accessed through this approach (Bejerano-Sagie and Xavier 2007). This approach allows understanding some of the biomolecular networks dictating emergent phenotypes in the microbiome and their roles in the agricultural systems. In addition, because of microorganisms and mechanisms involved in relevant functions including disease suppression in soil and microbiome-plant interaction are still poorly known, these kinds of disciplines with their respective techniques may unveil the role of genes responsible for these mechanisms (Kothari et al. 2017).

For example, Hayden et al. (2018) performed a comparative metatranscriptomic approach assessing the taxonomic and functional characteristics of the rhizosphere microbiome of wheat plants grown in adjacent fields previously known as suppressive and nonsuppressive to the pathogen *Rhizoctonia solani* AG8 (major pathogen of grain crops). The authors reported that suppressive samples showed greater expression of a polyketide cyclase, a terpenoid biosynthesis backbone gene, and several cold shock proteins, whereas nonsuppressive samples showed higher expression of antibiotic genes.

Furthermore, the combination of data obtained from metagenomics and metatranscriptomics can provide information about of the changes in gene expression that are accompanied by the changes in the microbiome structure, enabling to know what species are present in a given niche, their functional capabilities, and their real-time activities under particular scenarios.

8.5 Future Perspectives

Microbial communities play an important role for agriculture progress. Alterations of these types of ecosystems allow the entry of pathogens resulting in diseases that disrupt the industry development. In recent years, interest in understanding the activities of these microbial communities has increased. However, one of the main challenges in this area is discriminating and/or interpreting this massive amount of information. Bioinformatics tools have become indispensable in metagenomics research, allowing to know the structure and metabolic potential of environmental microbes. Improving the accuracy of the results is encouraged. One option is the single-molecule real-time sequencing (SMRT) developed by Pacific Biosciences (PacBio), which produces long readings and provide large scaffolds (Rhoads and

Au 2015). Despite having closely similar error rate to Illumina Miseq platform and Roche 454, it is expected to increase its efficiency and eventually displace small fragment sequencing by synthesis (Wagner et al. 2016).

A few metagenome studies by sequence-based metagenomics approaches have been performed for plants and associated microorganisms. Busby et al. (2017) mentioned the study of the interaction of plant-microbes in a global context could lead to find a natural mechanism facilitated by microorganisms controlling diseases by their ecological process. For example, pathogens may develop resistance mechanism that complicates their elimination in soil systems. Functional metagenomics has facilitated the detection of new antibiotic resistance mechanisms and find out the source of antibiotic resistance from total microbiome. Understanding these mechanisms will provide strategies in antibiotic improvement and alternatives in antibiotic resistance (Pehrsson et al. 2013). In addition, a total comprehension of microbial effects in natural alliance at culture systems will improve productions beyond genetic alteration (Lakshmanan et al. 2014).

Considering the beneficial role of microbial communities in soil environments, it is important to increase the understanding of microbial pathogens and provide biotechnological information for their control and management with sustainable practices.

References

Almeida A, Mitchell AL, Tarkowska A, Finn RD (2018) Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience 7:giy054

Aßhauer KP, Meinicke P (2013) On the estimation of metabolic profiles in metagenomics. In: On the estimation of metabolic profiles in metagenomics, OASIcs-OpenAccess Series in Informatics: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik

Aßhauer KP, Wemheuer B, Daniel R, Meinicke P (2015) Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31:2882–2884

Balvočiūtė M, Huson DH (2017) SILVA, RDP, Greengenes, NCBI and OTT, how do these taxonomies compare? BMC Genomics 18:114

Bejerano-Sagie M, Xavier KB (2007) The role of small RNAs in quorum sensing. Curr Opin Microbiol 10:189–198

Béné C, Arthur R, Norbury H, Allison E, Beveridge M, Bush S, Campling L, Leschen W, Little D, Squires D, Thilsted S, Troell M, Williams M (2016) Contribution of fisheries and aquaculture to food security and poverty reduction: assessing the current evidence. World Develop 79:177–196

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

Busby PE, Soman C, Wagner MR, Friesen ML, Kremer J, Bennett A, Morsy M, Eisen JA, Leach JE, Dangl JL (2017) Research priorities for harnessing plant microbiomes in sustainable agriculture. PLoS Biol 15:e2001793

Chaparro JM, Sheflin AM, Manter DK, Vivanco JM (2012) Manipulating the soil microbiome to increase soil health and plant fertility. Biol Fertil Soils 48:489–499

Dos Santos DFK, Istvan P, Quirino BF, Kruger RH (2017) Functional metagenomics as a tool for identification of new antibiotic resistance genes from natural environments. Microb Ecol 73:479–491

Gebbers R, Adamchuk VI (2010) Precision agriculture and food security. Science 327:828–831
Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, Bruns G, Yarza P, Peplies J, Westram R, Ludwig W (2017) 25 years of serving the community with ribosomal RNA gene reference databases and tools. J Biotechnol 261:169–176
Hayden HL, Savin K, Wadeson J, Gupta V, Mele PM (2018) Comparative metatranscriptomics of wheat rhizosphere microbiomes in disease suppressive and non-suppressive soils for Rhizoctonia solani AG8. Front Microbiol 9:859
Jones AD, Ejeta G (2016) A new global agenda for nutrition and health: the importance of agriculture and food systems. Bull World Health Organ 94:228–229
Jovel J, Patterson J, Wang W, Hotte N, O'keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL (2016) Characterization of the gut microbiome using 16S or shotgun metagenomics. Front Microbiol 7:459
Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:D480–D484
Kothari V, Kothari C, Rank J, Joshi A, Singh RP, Kothari R (2017) Metatranscriptomic 16 Rhizosphere for finding of the plant. Understanding host-microbiome interactions – an omics approach: Host-Microbiome Association 1:267p
Lakshmanan V, Selvaraj G, Bais HP (2014) Functional soil microbiome: belowground solutions to an aboveground problem. Plant Physiol 166:689–700
Lam KN, Cheng J, Engel K, Neufeld JD, Charles TC (2015) Current and future resources for functional metagenomics. Front Microbiol 6
Langille MGI, Zaneveld J, Caporaso JG, Mcdonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotech 31:814–821
Lazarevic V, François P (2013) Functional genomics of microbial pathogens. Brief Funct Genomics 12:289–290
Loman NJ, Pallen MJ (2015) Twenty years of bacterial genome sequencing. Nat Rev Microbiol 13:787
Mendes R, Garbeva P, Raaijmakers JM (2013) The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS Microbiol Rev 37(5):634–663
Mendes LW, Kuramae EE, Navarrete AA, Van Veen JA, Tsai SM (2014) Taxonomical and functional microbial community selection in soybean rhizosphere. ISME J 8:1577–1587
Morgan XC, Huttenhower C (2012) Human microbiome analysis. PLoS Comput Biol 8:e1002808
Ortiz-Estrada ÁM, Gollas-Galván T, Martínez-Córdova LR, Martínez-Porchas M (2019) Predictive functional profiles using metagenomic 16S rRNA data: a novel approach to understanding the microbial ecology of aquaculture systems. Rev Aquacult 11:234–245
Parks DH, Tyson GW, Hugenholtz P, Beiko RG (2014) STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 30:3123–3124
Pehrsson EC, Forsberg KJ, Gibson MK, Ahmadi S, Dantas G (2013) Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. Front Microbiol 4:145
Popovic A, Parkinson J (2018) Characterization of eukaryotic microbiome using 18S amplicon sequencing. In: Press SH (ed) Microbiome analysis. Springer, New York, pp 29–48
Qin S, Xing K, Jiang JH, Xu LH, Li WJ (2011) Biodiversity, bioactive natural products and biotechnological potential of plant-associated endophytic actinobacteria. Appl Microbiol Biotechnol 89(3):457–473
Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844
Ramirez KS, Lauber CL, Knight R, Bradford MA, Fierer N (2010) Consistent effects of nitrogen fertilization on soil bacterial communities in contrasting systems. Ecology 91:3463–3470

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289

Rooijers K, Kolmeder C, Juste C, Doré J, De Been M, Boeren S, Galan P, Beauvallet C, De Vos WM, Schaap PJ (2011) An iterative workflow for mining the human intestinal metaproteome. BMC Genomics 12:6

Siegwald L, Touzet H, Lemoine Y, Hot D, Audebert C, Caboche S (2017) Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. PLoS One 12:e0169563

Theis KR, Dheilly NM, Klassen JL, Brucker RM, Baines JF, Bosch TC, Cryan JF, Gilbert SF, Goodnight CJ, Lloyd EA (2016) Getting the hologenome concept right: an eco-evolutionary framework for hosts and their microbiomes. Msystems 1:e00028–e00016

Vargas-Albores F, Martinez-Cordova LR, Martinez-Porchas M, Calderon K, Lago-Leston A (2018) Functional metagenomics: a tool to gain knowledge for agronomic and veterinary sciences. Biotechnol Genet Eng Rev:1–23

Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J (2016) Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol 16:274

Wallenstein MD (2017) Managing and manipulating the rhizosphere microbiome for plant health: a systems approach. Rhizosphere 3:230–232

Wang Y, Tian RM, Gao ZM, Bougouffa S, Qian P-Y (2014) Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. PLoS One 9: e90053

Yang C, Ji Y, Wang X, Yang C, Douglas WY (2013) Testing three pipelines for 18S rDNA-based metabarcoding of soil faunal diversity. Sci China Life Sci 56:73–81

Zhu S, Vivanco JM, Manter DK (2016) Nitrogen fertilizer rate affects root exudation, the rhizosphere microbiome and nitrogen-use-efficiency of maize. Appl Soil Ecol 107:324–333

Microbial Genomics in Carbon Management and Energy Production

9

Shatabisha Bhattacharjee and Tulika Prakash

Abstract

Microbial genomics has helped us understand the diversity of the entire microbial world. Moreover, the recent development in the field of metagenomics has greatly facilitated the exploration of the unculturable microbes. These fields provide us with a detailed investigation on various functional features and metabolic pathways of microbes. Based on this information, the search of potential microbe(s) for energy production has developed significantly. This development may lead to a limited usage of fossil fuels in the near future. Therefore, researchers are in continuous efforts to develop the potent technologies to harness microbial communities for energy production. One example of such technology is the microbial fuel cells (MFCs). A few promising energy-producing case studies are discussed using this technology. The next application of microbes toward energy is the production of biohydrogen, which is considered as a promising biofuel in the near future. Furthermore, a brief section of the role microbial world, ecosystem, and their relationship with the climatic change is also discussed.

9.1 Introduction

The advancement in the field of microbial genomics has helped us understand the phylogenetically diverse microbial world, which comprises of different microbial communities such as archaea, bacteria, viruses, fungi, protozoa, microscopic metazoan, and microalgae (Martin 2002). The detailed investigation of these communities provides leads toward the sizes of their individual genomes which range from a few kilo to hundred mega bases and their genetic material (Massana

S. Bhattacharjee · T. Prakash (✉)
School of Basic Sciences, Indian Institute of Technology (IIT), Mandi, Himachal Pradesh, India
e-mail: tulika@iitmandi.ac.in

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_9

et al. 1997). Further, the information stored in their genomes in the form of DNA and RNA (in case of viruses) sequences has helped to explore the knowledge on structural, functional, and phylogenetic relation, metabolic pathways, growth kinetics of microbial communities, and their intertwined relationship with geochemical process of different environmental niches (Demirbas 2008).

Initially, the microbial genomics field was used to focus on the genomes of pathogens. In 1995, the first complete genome of *Haemophilus influenzae* was published (Fleischmann et al. 1995). Almost 1554 pathogenic bacterial genomes and 112 archaeal genomes have been sequenced, and many more are in progress at present (Fleischmann et al. 1995). Completed genomes of 41 eukaryotes, including 19 fungi, have been sequenced, and the sequencing of about 1100 genomes is in progress (Stover et al. 2000). The complete sequencing of 2675 kinds of viral species has been performed including partial and fully sequenced viral strains. Sequencing of 300,000 strains of HIV type 1 virus and 40,000 strains of influenza virus has also been performed (Squires et al. 2007).

Different literature studies have revealed that only 1% of the entire microbial community can be cultured in the laboratory environment, whereas 99 % cannot (Carr et al. 1998). Recently, researchers have developed high-throughput technologies to unveil the uncultivable microbial communities which are reported to participate synergistically with the cultivable ones in different environmental niches. Different DNA-based methods have been developed which includes 16S rRNA gene analysis. This method provides extensive information about the natural microbiome, including various taxa and species, present in the environments (Streit and Schmitz 2004). Thus, metagenomics is one of the new high-throughput technologies which help to analyze the complex microbial genomes found in different environmental niches (Schmidt et al. 1991). The study of metagenomic analysis is initiated by the isolation of DNA from different environmental niches (Fig. 9.1). Therefore, this approach aids in pooling the genetic material of the entire microbiome including those cultivable and uncultivable. However, there are a few difficulties associated with the metagenomic approach which include low quantity of DNA from some environment and contamination of purified DNA with compounds of polyphenol. Sometimes it becomes difficult to remove these polyphenolic compounds from the isolated DNA which interferes with the enzymatic modifications (Tsai and Olson 1992; Nelson 2003).

Presently, several researchers have reported in-depth metagenomic analysis of different environmental samples such as soil (Sangwan et al. 2012), water (Gomez-Alvarez et al. 2012), air (Cha et al. 2017), wastewater treatment plants (Balcom et al. 2016), and human associated (Mitra et al. 2015). However, major challenges are still present in case of annotating genes and sequencing error detection followed by correction, prediction of gene products, and interaction of different microbial communities from their sequenced raw genomic data (Furnham et al. 2012).

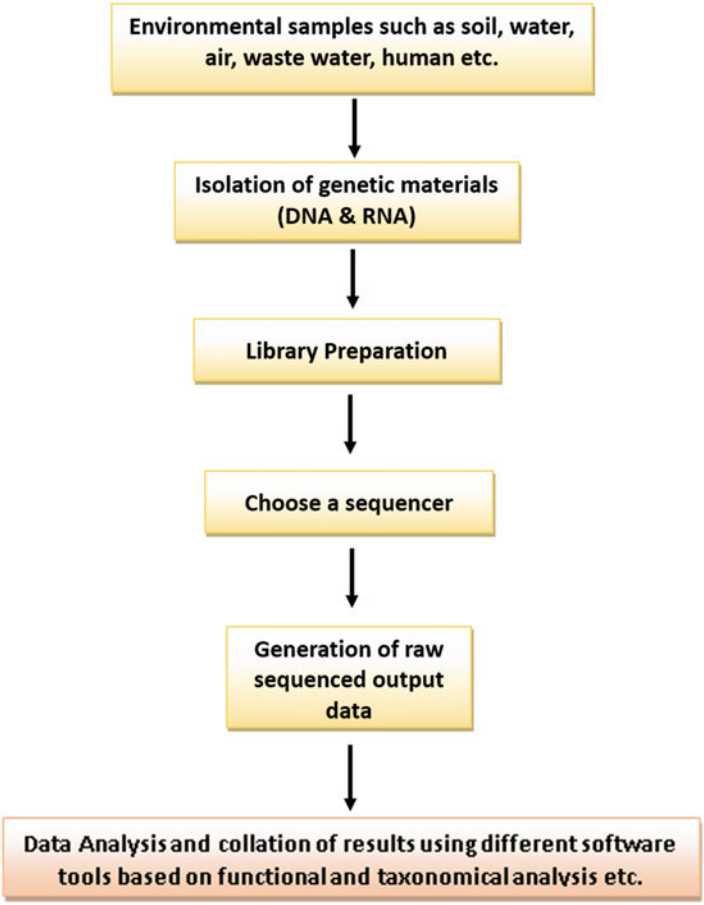

Fig. 9.1 Flowchart demonstrating the typical steps in metagenomic analysis

9.2 Role of Microbes in Carbon Management

9.2.1 Microbes, Ecosystem, and Carbon Cycle Intertwined with Climatic Change

There are countless numbers of ways in which different microbial communities, especially the soil microbes, influence the carbon cycle through their metabolic activities. These ways are often classified into two groups: one is known to affect

the concentration of carbon dioxide and the uptake of methane in the ecosystem and the other controls the loss of carbon from soil due to respiration and methane production. Through different literature studies, it has been known that climatic change, whether direct or indirect, is known to affect the soil microbiome activities and their response toward the emission of greenhouse gases in the atmosphere which further leads to global warming. There are two types of effects on the soil microbiome. Firstly, the direct effects such as the change in drastic climatic events, precipitation, temperature, and greenhouses gases that are produced have an impact on soil microbiome. Secondly, an indirect effect which leads to climatic change can alter the productivity of plants which contribute in alteration of soil's physiochemical nature. It might also be able to alter the carbon content of soil, thus affecting those microbiomes, which are responsible in the decomposition activities (Bardgett et al. 2008).

9.2.1.1 Carbon Cycling and Soil Microbiome

Microbes are known to possess a remarkable enzyme system which has the ability to operate the Earth's biogeochemical processes or cycles (Weiman 2015). With the help of this enzyme system, microbes often extract and break down the dead organic matter and transfer it into soil in usable form for the other living organisms. One of the major biogeochemical cycles is the global carbon cycle, which is known to depend on microbial communities (Falkowski et al. 2008). Microbes especially help in fixing the atmospheric carbon, plant growth, and degradation and transformation of dead organic matter and its decomposition into the environment. A huge amount of organic carbons are known to be stored in the grassland soil, permafrost, tropical forest, and many other ecosystems. Likewise, the terrestrial carbon cycle is produced from two biological metabolic processes such as photosynthesis and respiration (Prosser 2007). The autotrophic microorganisms help to transfer atmospheric carbon via carbon fixation process into the soil. Along with these organisms, chemoautotrophic ones are known to synthesize the atmospheric carbon dioxide into reusable organic forms (Gougoulias et al. 2014).

Soil microbiome plays an important part in carbon cycling. A huge number of microbial diversity is observed in just a handful of soil (Gans et al. 2005). Apart from the microbial communities, various fauna groups are also observed in soil. Among them, nematodes and mites are the highest observed species (Brussaard 1997). The presence of varied organisms leads toward the functional redundancy of soil. It is observed that specific fungal communities are responsible for the decomposition of specific carbon sources, which include cellulose, lignin, etc., in the soil (Cox et al. 2001). Through different literature studies, it has been observed that soil microbiome is not only responsible for decomposition but also for heterotrophic respiration (Nielsen et al. 2011). Thus, the shift in microbial communities in the soil will directly influence the carbon cycle (Fig. 9.2) (Orwin et al. 2006). According to the prediction models of different ecosystem, the climatic change will directly affect the microbial decomposition of carbon sources present in the soil (Friedlingstein et al. 2006). This will lead to the rise in global temperature or global warming process.

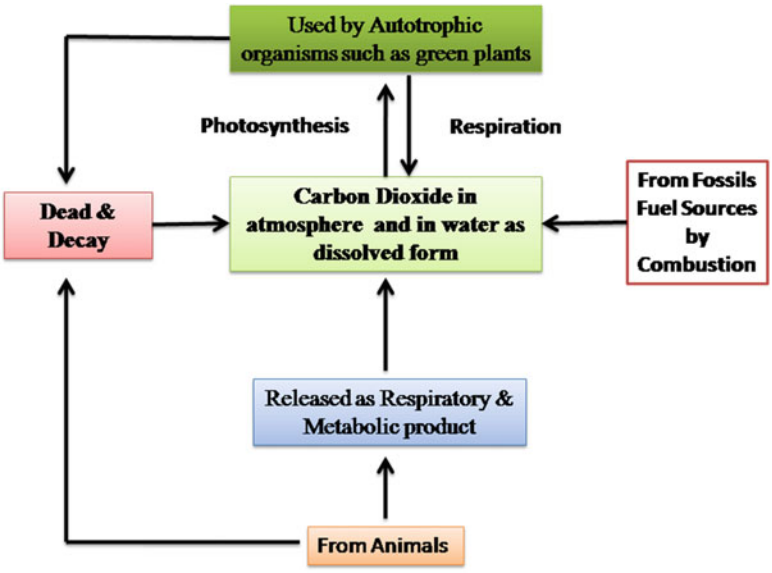

Fig. 9.2 Schematic representation of basic steps of carbon cycle

Thus, the efficiency of using carbon compounds by soil microbes leads toward the climatic change or global warming (Allison et al. 2010).

9.3 Microbial Genomics in Energy Production

The hunt for several options for the production of renewable energy from different natural sources is increasing day by day. For the sake of the society, sustainable energy systems are much needed. This need is fulfilled by biological methods for the production of fuels, such as biohydrogen, biomethane, bioethanol, biodiesel, etc. Among these renewable energy sources, biohydrogen is regarded as one of the promising fuels which can be produced from organic wastes and wastewater. With the advancement in the field of microbial genomics, the information of various microbes has been stored in different public databases (Quast et al. 2012). The approach of genomic data mining has immensely helped to explore the potent biohydrogen producers (Kalia and Purohit 2008).

9.3.1 Biogas as a Renewable Source of Fuel

Waste consists of organic complex compounds in the form of biomass such as sewage sludge, human and animal wastes, industrial effluents, etc. This waste is further broken down via anaerobic digestion into carbon dioxide and methane which is known as "biogas" (Kapdi et al. 2005). The anaerobic digestion is one of the

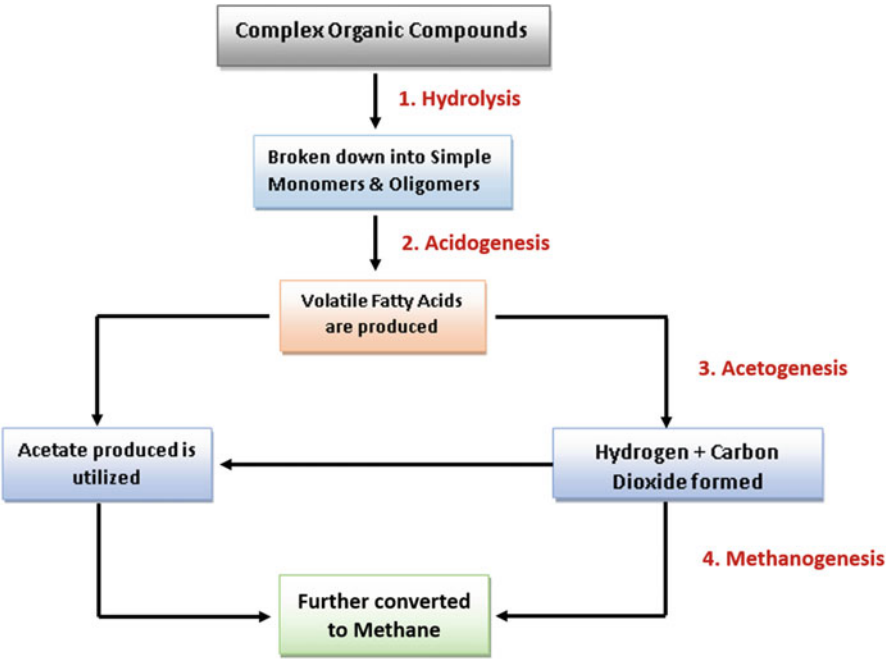

Fig. 9.3 Schematic representation of anaerobic dark fermentation

complex dark fermentative processes in which methane is produced in the absence of oxygen by the degradation of complex organic compounds. A number of hydrolytic microorganisms help in the degradation of organic compounds to monomers and oligomer; thus, the first set of reaction is hydrolysis (Fig. 9.3). Further, the fermentative or facultative bacteria convert the simpler monomers or oligomers into volatile fatty acid in the acidogenesis step. This is followed by acetogenesis, in which acetogens use volatile fatty acids to produce biohydrogen and carbon dioxide initially. These acetogens further consume the produced biohydrogen to form acetate. Acetate is finally converted to methane by CO_2-reducing methanogens as well as acetoclastic methanogens through methanogenesis. Different factors can affect the anaerobic digestion such as pH, partial pressure, requirements of micronutrients, temperature, etc., in which temperature is the most important factor. The microbial communities, which participate in this process, can withstand temperatures ranging from below freezing to more than 330.4 K, and optimum range is within 309.9 K (mesophilic) and 327.6 K (thermophilic) (Demirbas 2007).

Microbial community plays a vital part in the production of biogas from different substrates present in different forms of waste or wastewater. The sludge that comes from the anaerobic digester contains various kinds of microbial communities. Researchers have found 20 different bacterial phyla present in the wastewater sludge which are mostly methanogenic (Goswami et al. 2016). These bacterial phyla include *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Actinobacteria*, *Acidobacteria*,

Spirochetes, and *Chloroflexi* (Chojnacka et al. 2015). These microbial communities, especially those with hydrogenase enzymes, help to discard the surplus amount of electrons that are accumulated during the fermentation process (Elsharnouby et al. 2013; Traversi et al. 2012; Chang et al. 2011). Thus, a proper balance of H_2 producers as well as H_2 consumers exists in the medium. Different studies have shown that the H_2 consumers mainly consist of hydrogenotrophic, nitrate, or sulfate-reducing bacteria and homoacetogenic bacteria. These bacteria require H_2 to reduce the substrate that is formed after the degradation of complex compounds present in the wastewater. When the concentration of nitrates and sulfates is low, the population of homoacetogenic bacteria and the methanogenic archaea predominates (Wang et al. 2013; Oehler et al. 2012). The homoacetogenic bacteria are composed of genera such as *Butyribacterium*, *Acetobacterium*, *Clostridium*, *Eubacterium*, *Peptostreptococcus*, and *Sporomusa*. They can grow faster to form spores and are characterized as obligate or strict anaerobes. They also possess several adaptations related to hydrogen producers (Oehler et al. 2012).

For a higher yield of biohydrogen, several pretreatment strategies are employed to the sludge including heat treatment, aeration, acid treatment, alkaline treatment, etc. (Chaitanya et al. 2016). These pretreatment processes are mainly applied to the sludge for enriching biohydrogen-producing bacteria from mixed anaerobic microbiome. Several pretreatment methods, apart from the abovementioned, have also been used for improving biohydrogen production. These methods include heat shock and sodium 2-bromoethanesulfonate (Shrestha et al. 2017; Mohan and Pandey 2013). Different studies have been performed to evaluate the efficacy of the various pretreatment strategies for enhancing the production of biohydrogen from mixed microflora (Mohan et al. 2008). Hence, during the pretreatment procedures, these microbial communities get shifted according to the need and available substrate in the environment. Likewise, for specific biogas production such as biohydrogen, biomethane, or biohythane (both hydrogen and methane), the microbial diversity also changes accordingly (Mohan et al. 2007).

9.3.2 Biohydrogen as a Promising Source of Biofuel

Biohydrogen is a natural end product of the microbial-driven metabolic process. Biohydrogen, as the name signifies, is the generation of hydrogen gas from organic biomass or feedstock. There are several biological pathways which have been identified for the production of biohydrogen. Worldwide researches are going on from the past two decades for the exploration of various routes of biohydrogen production which may be used as a promising renewable source of energy. Biohydrogen production can be classified into two different fermentative processes: (Martin 2002) light-independent or dark fermentative process and (Massana et al. 1997) light-dependent or photosynthetic process (Mohan et al. 2008). The light-dependent photosynthetic process can be further divided into two groups: (Martin 2002) photosynthetic or fermentation based on carbon source and (Massana et al. 1997) type of inoculums used as a source of biocatalyst. The light-dependent process

takes place in two ways either through biophotolysis of water using microorganism (e.g., cyanobacteria) or via photosynthetic bacteria. The dark fermentation process, however, is confined to anaerobic metabolism in which acidogenic bacteria generate biohydrogen via an acetogenic process followed by the generation of various volatile fatty acids and carbon dioxide (Mohan and Pandey 2013).

Different research groups are trying to generate biohydrogen as one of the promising source of biofuels from various waste materials, especially different wastewater. In one of the study dairies, wastewater has been used as the substrate of carbon source for generating biohydrogen (Mohan et al. 2008). Similarly, chemical wastewater has also been used as a primary substrate for biohydrogen production. It has been investigated that effective biohydrogen production is observed at pH 6 (1.25 mmol H_2/g COD) as compared to pH 5 (0.71 mmol H_2/g COD) and pH 7 (0.27 H_2/g COD). The addition of co-substrate such as glucose and domestic sewage water along with the chemical wastewater has shown higher biohydrogen production (Mohan et al. 2007).

9.4 Microbial Fuel Cells

In microbial fuel cells (MFCs), the activity of biological catalysis redox reaction is combined with electrochemical systems (Du et al. 2007). An active microbe is incorporated in the compartment containing anaerobic anode as a catalyst. It has been reported in several studies that MFCs contain two chambers, namely, the anode and cathode, which are further divided by a proton exchange membrane (PEM) (Logan et al. 2006). The active microbes present as a catalyst inside the MFCs oxidize the organic substrate by generating electrons and protons (Rahimnejad et al. 2015). Further, the protons that are generated are moved through PEM, and the electrons are moved through the external circuits. The microbes that are used in MFCs include *Geobacillus*, *Clostridium*, *Pseudomonas*, etc. (Antonopoulou et al. 2010). However, MFCs are facing a lot of challenges based on the sustainability, process feasibility, high cost of materials, etc. (Rahimnejad et al. 2011).

The conception and idea of MFCs was developed by Michael C. Potter in 1910. Many researchers have been using this technology to study the microbial biochemical, electrochemical, and reaction based on material surface under certain optimized conditions. This technology will further help to explore the effect of material surface, chemical compounds, and substrates among others. Different research groups have studied MFCs in the form of reactor for the treatment of wastewater, starting from lab scale to pilot scale. One of the research groups has extracted energy from MFC reactor to detect compounds as well as to study the degradation of organics along with energy recovery. After 2 years, the first robot, named EcoBot I, was developed, which was directly powered by MFCs. These MFCs were fed with substrate like glucose without using any other kind of conventional power system. Thus, MFCs are regarded as one of the platform technologies with multiple developing applications that can generate electricity without using external power sources (Santoro et al. 2017).

9.5 Usage of Metagenomics Analysis in the Exploration of Microbial Communities

Recently, metagenomic analysis has proved to be one of the promising techniques to investigate and explore the presence of the entire microbiome in the environmental samples. In different literature studies, it has been reported that only 1% of microorganism are cultivable in normal laboratory conditions (Carr et al. 1998). Hence, the introduction of metagenomics analysis in these cases is advantageous. With the help of these techniques, pooling of the entire microbiome's genetic material (DNA and RNA) is carried out. Therefore, striking diversities based on taxonomic and functional aspects of several habitats have been investigated (Eckburg et al. 2005). Two types of analysis can be carried out based on the extraction of genetic material from the environmental samples including whole genome shotgun sequencing of all the available genes present in the microbial communities and by targeting a single gene, i.e., the hypervariable conserved region of the prokaryotic 16S ribosomal RNA (Staley and Sadowsky 2016).

9.6 Next-Generation Technologies Platforms Used in the Analysis of Environmental Samples

The traditional DNA-sequencing method was first developed by Sanger et al. in 1977. In that method, from a single specimen, only one kilobyte of sequenced data could be recovered at a time. Later, some advancement was implemented in the Sanger sequencers which had an ability to recover one kilobyte (kb) from 96 different samples at a time. As time passed by, a series of next-generation sequencing technologies (NGS) were introduced commercially with different sample detection methodologies as well as their unique chemistries. Millions of raw sequenced data or reads were generated from these NGS techniques. This process was also termed as "massively parallel throughput sequencing techniques." These techniques have the ability to generate sequenced reads from a particular genome through genome sequencing. The fragments of cDNA libraries are produced from the process of reverse transcription of RNA molecules, termed as RNAseq, or by pooling of PCR-based products, referred to as amplicons (Shokralla et al. 2012).

NGS technology has faced a lot of challenges after it is commercially introduced in the year 2005. The first challenge has been related to the read length and its accuracy. The second challenge has been in terms of labor and cost expenditure along with the output that is generated from the sequences. The third challenge has been faced in the amplification steps related to PCR bias leading to the formation of chimeric sequences (Gilbert and Dupont 2010; Pareek et al. 2011). But in spite of these challenges, NGS technology has proved to be one of the promising techniques of genomics, metagenomics, and metatranscriptomics-based studies.

Though diverse in chemistries, the basic steps of all the available NGS technologies are almost the same, which include preparation of libraries of amplicons or preparation of fragmented libraries, detection of various nucleotides

which are incorporated, etc. (Shendure and Ji 2008; Zhang et al. 2011). Furthermore, the NGS technologies can be classified into two main groups. The first technology is based on PCR including Roche 454 Genome Sequencer (Roche Diagnostics Corp., Branford, CT, USA), HiSeq 2000 (Illumina Inc., San Diego, CA, USA), AB SOLiD System (Life Technologies Corp., Carlsbad, CA, USA), and Ion Personal Genome Machine (Life Technologies, South San Francisco, CA, USA). And the second technology is based on single molecular sequencing technologies (these are basically non-PCR based in which the amplification step is absent). Such platforms include HeliScope (Helicos BioSciences Corp., Cambridge, MA, USA) and PacBio RS SMRT system (Pacific Biosciences, Menlo Park, CA, USA) (Glenn 2011).

9.7 Future Scope

Microbial communities' genomes can be harnessed using high-throughput technologies and metagenomics analysis to investigate bioenergy-producing potential microbe(s) which would further help to replace fossils fuels and reduce the emission of greenhouse gases into the environment. The genomes of certain potential microbe(s) can be mined to search for potent energy producers. This genome can be further harnessed to treat industrial wastewater containing complex compounds. Integration of new technology with proper microbial consortium would help us to develop further sustainable renewable source of bioenergy which might lead toward the development of new manufacturing model (Quan et al. 2004).

References

Allison SD, Wallenstein MD, Bradford MA (2010) Soil-carbon response to warming dependent on microbial physiology. Nat Geosci 3(5):336

Antonopoulou G, Stamatelatou K, Bebelis S, Lyberatos G (2010) Electricity generation from synthetic substrates and cheese whey using a two chamber microbial fuel cell. Biochem Eng J 50(1-2):10–15

Balcom IN et al (2016) Metagenomic Analysis of an ecological waste water treatment plant's microbial communities and their potential to metabolize pharmaceuticals. F1000Res 5:1881

Bardgett RD, Freeman C, Ostle NJ (2008) Microbial contributions to climate change through carbon cycle feedbacks. ISME J 2(8):805

Brussaard L (1997) Biodiversity and ecosystem functioning in soil. Ambio:563–570

Carr JK, Salminen MO, Albert J, Sanders-Buell E, Gotte D, Birx DL, McCutchan FE (1998) Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology 247(1):22–31

Cha S, Srinivasan S, Jang JH, Lee D, Lim S, Kim KS et al (2017) Metagenomic analysis of airborne bacterial community and diversity in Seoul, Korea, during December 2014, Asian Dust Event. PLoS ONE 12(1):e0170693

Chaitanya N, Sivaramakrishna D, Kumar BS, Himabindu V, Lakshminarasu M, Vishwanadham M (2016) Selection of pretreatment method for enriching hydrogen-producing bacteria using anaerobic sewage sludge with three different substrates. Biofuels 7(2):163–171

Chang S, Li JZ, Liu F (2011) Evaluation of different pretreatment methods for preparing hydrogen-producing seed inocula from waste activated sludge. Renew Energy 36(5):1517–1522

Chojnacka A, Szczęsny P, Błaszczyk MK, Zielenkiewicz U, Detman A, Salamon A, Sikora A (2015) Noteworthy facts about a methane-producing microbial community processing acidic effluent from sugar beet molasses fermentation. PLoS One 10(5):e0128008

Cox P, Wilkinson SP, Anderson JM (2001) Effects of fungal inocula on the decomposition of lignin and structural polysaccharides in Pinus sylvestris litter. Biol Fertil Soils 33(3):246–251

Demirbas A (2007) Progress and recent trends in biofuels. Prog Energy Combust Sci 33(1):1–18

Demirbas A (2008) Biofuels sources, biofuel policy, biofuel economy and global biofuel projections. Energy Convers Manag 49(8):2106–2116

Du Z, Li H, Gu T (2007) A state of the art review on microbial fuel cells: a promising technology for wastewater treatment and bioenergy. Biotechnol Adv 25(5):464–482

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. Science 308 (5728):1635–1638

Elsharnouby O, Hafez H, Nakhla G, El Naggar MH (2013) A critical literature review on biohydrogen production by pure cultures. Int J Hydrog Energy 38(12):4945–4966

Falkowski PG, Fenchel T, Delong EF (2008) The microbial engines that drive Earth's biogeochemical cycles. Science 320(5879):1034–1039

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269(5223):496–512

Friedlingstein P, Cox P, Betts R, Bopp L, von Bloh W, Brovkin V, Cadule P, Doney S, Eby M, Fung I, Bala G (2006) Climate–carbon cycle feedback analysis: results from the C4MIP model intercomparison. J Clim 19(14):3337–3353

Furnham N, de Beer TA, Thornton JM (2012) Current challenges in genome annotation through structural biology and bioinformatics. Curr Opin Struct Biol 22(5):594–601

Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309(5739):1387–1390

Gilbert, J.A. and Dupont, C.L., 2010. Microbial metagenomics: beyond the genome.

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11(5):759–769

Gomez-Alvarez V et al (2012) Metagenomic analysis of drinking water receiving different disinfection treatments. Appl Environ Microbiol 78(17):6095–6102

Goswami R, Chattopadhyay P, Shome A, Banerjee SN, Chakraborty AK, Mathew AK, Chaudhury S (2016) An overview of physico-chemical mechanisms of biogas production by microbial communities: a step towards sustainable waste management. 3 Biotech 6(1):72

Gougoulias C, Clark JM, Shaw LJ (2014) The role of soil microbes in the global carbon cycle: tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems. J Sci Food Agric 94(12):2362–2371

Kalia VC, Purohit HJ (2008) Microbial diversity and genomics in aid of bioenergy. J Ind Microbiol Biotechnol 35(5):403–419

Kapdi SS, Vijay VK, Rajesh SK, Prasad R (2005) Biogas scrubbing, compression and storage: perspective and prospectus in Indian context. Renew Energy 30(8):1195–1202

Logan BE, Hamelers B, Rozendal R, Schröder U, Keller J, Freguia S, Aelterman P, Verstraete W, Rabaey K (2006) Microbial fuel cells: methodology and technology. Environ Sci Technol 40 (17):5181–5192

Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl Environ Microbiol 68(8):3673–3682

Massana R, Gasol JM, Bjørnsen PK, Blackburn N, Hagstrøm Å, Hietanen S, Hygum BH, Kuparinen J, Pedrós-Alió C (1997) Measurement of bacterial size via image analysis of epifluorescence preparations: description of an inexpensive system and solutions to some of the most common problems

Mitra S et al (2015) In silico analyses of metagenomes from human atherosclerotic plaque samples. Microbiome 3:38

Mohan SV, Pandey A (2013) Biohydrogen production: an introduction. In: Biohydrogen, pp 1–24

Mohan SV, Bhaskar YV, Krishna PM, Rao NC, Babu VL, Sarma PN (2007) Biohydrogen production from chemical wastewater as substrate by selectively enriched anaerobic mixed consortia: influence of fermentation pH and substrate composition. Int J Hydrog Energy 32 (13):2286–2295

Mohan SV, Babu VL, Sarma PN (2008) Effect of various pretreatment methods on anaerobic mixed microflora to enhance biohydrogen production utilizing dairy wastewater as substrate. Bioresour Technol 99(1):59–67

Nelson KE (2003) The future of microbial genomics. Environ Microbiol 5(12):1223–1225

Nielsen UN, Ayres E, Wall DH, Bardgett RD (2011) Soil biodiversity and carbon cycling: a review and synthesis of studies examining diversity–function relationships. Eur J Soil Sci 62 (1):105–116

Oehler D, Poehlein A, Leimbach A, Müller N, Daniel R, Gottschalk G, Schink B (2012) Genome-guided analysis of physiological and morphological traits of the fermentative acetate oxidizer Thermacetogeniumphaeum. BMC Genomics 13(1):723

Orwin KH, Wardle DA, Greenfield LG (2006) Ecological consequences of carbon substrate identity and diversity in a laboratory study. Ecology 87(3):580–593

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52(4):413–435

Prosser JI (2007) Microorganisms cycling soil nutrients and their diversity. In: Modern soil microbiology. CRC Press, Boca Raton

Quan X, Shi H, Liu H, Lv P, Qian Y (2004) Enhancement of 2, 4-dichlorophenol degradation in conventional activated sludge systems bioaugmented with mixed special culture. Water Res 38 (1):245–253

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41(D1):D590–D596

Rahimnejad M, Najafpour G, Ghoreyshi AA (2011) Effect of mass transfer on performance of microbial fuel cell. In: Mass transfer in chemical engineering processes. InTech

Rahimnejad M, Adhami A, Darvari S, Zirepour A, Oh SE (2015) Microbial fuel cell as new technology for bioelectricity generation: a review. Alex Eng J 54(3):745–756

Sangwan N, Lata P, Dwivedi V, Singh A, Niharika N, Kaur J et al (2012) Comparative metagenomic analysis of soil microbial communities across three hexachlorocyclohexane contamination levels. PLoS ONE 7(9):e46219

Santoro C, Arbizzani C, Erable B, Ieropoulos I (2017) Microbial fuel cells: from fundamentals to applications. A review. J Power Sources 356:225–244

Schmidt TM, DeLong EF, Pace NR (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol 173(14):4371–4378

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135

Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Mol Ecol 21(8):1794–1805

Shrestha S, Fonoll X, Khanal SK, Raskin L (2017) Biological strategies for enhanced hydrolysis of lignocellulosic biomass during anaerobic digestion: current status and future perspectives. Bioresour Technol 245:1245–1257

Squires B, Macken C, Garcia-Sastre A, Godbole S, Noronha J, Hunt V, Chang R, Larsen CN, Klem E, Biersack K, Scheuermann RH (2007) BioHealthBase: informatics support in the elucidation of influenza virus host–pathogen interactions and virulence. Nucleic Acids Res 36 (suppl_1):D497–D503

Staley C, Sadowsky MJ (2016) Application of metagenomics to assess microbial communities in water and other environmental matrices. J Mar Biol Assoc U K 96(1):121–129

Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FSL, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL (2000) Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature 406(6799):959

Streit WR, Schmitz RA (2004) Metagenomics–the key to the uncultured microbes. Curr Opin Microbiol 7(5):492–498

Traversi D, Villa S, Lorenzi E, Degan R, Gilli G (2012) Application of a real-time qPCR method to measure the methanogen concentration during anaerobic digestion as an indicator of biogas production capacity. J Environ Manag 111:173–177

Tsai YL, Olson BH (1992) Rapid method for separation of bacterial DNA from humic substances in sediments for polymerase chain reaction. Appl Environ Microbiol 58(7):2292–2295

Wang H, Paul DR, Chung TS (2013) Surface modification of polyimide membranes by diethylenetriamine (DETA) vapor for H2 purification and moisture effect on gas permeation. J Membr Sci 430:223–233

Weiman S (2015) Microbes help to drive global carbon cycling and climate change. Microbe Mag 10(6):233–238

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38(3):95–109

Microbial Genome Diversity and Microbial Genome Sequencing

10

Aditi Jangid and Tulika Prakash

Abstract

The role of microbiome in agroecosystems has evolved due to improvements in microbial diversity analysis methods. The journey of microbial diversity estimation progressed from culture-dependent to culture-independent methods. The culture-dependent methods are important in finding the microbial diversity of different environments; however, they are immensely biased toward the dominant microorganisms present in a community. With the advancement in sequencing techniques and genomics, the community exploration using culture-independent methods has commenced a new understanding of microbial interactions with their surroundings. Molecular studies of different environmental communities have uncovered <1% of the total number of prokaryotic species representing the cultivable fraction. This chapter summarizes the different methods to acquire a microbial diversity that may eventually enhance plant growth in sustainable agriculture and may often play a role in the management of environmental problems. The merits and demerits of the commonly used molecular methods to investigate microbial communities are discussed. The potential applications of next-generation sequencing techniques for a comprehensive assessment of microbial diversity have been illustrated.

10.1 Introduction

In the era of climate change and ecosystem degradation, the aggregate agricultural output has become significantly altered. The biotic and abiotic stresses along with land degradation leading to lesser productivity and sustainability are the main

A. Jangid · T. Prakash (✉)
School of Basic Sciences, Indian Institute of Technology (IIT), Mandi, Himachal Pradesh, India
e-mail: tulika@iitmandi.ac.in

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_10

challenges in agriculture. In the current scenario, excessive use of chemical pesticides and fertilizers greatly impact the sustainability of agriculture. Moreover, human-generated wastes through industrialization and urbanization present a serious warning to the ecosystems. The restoration of the ecosystems has been implemented through various approaches, but with minimal success. Recently, the identification of the beneficial soil microbes came into picture which are considered as suitable candidates that may improve the sustainability of the environment. Several mechanisms exhibited by these microorganisms can be utilized commercially for solving the critical environmental issues. In agroecosystems, currently, the beneficial microbe-based products have shown remarkable success. For the future enhancement of worldwide crop production, it becomes necessary to maximize the microbial functions in agroecosystems. In this chapter, we have summarized different methods to obtain a microbial diversity that may eventually enhance plant growth in sustainable agriculture and may often play a role in the management of environmental problems.

10.2 Relevance of Microbiota in Agroecosystems

Agroecosystems are complex systems consisting of entwined networks of interactions formed amid numerous microorganisms and macroorganisms. Microorganisms such as bacteria, fungi, protists, and archaea interact with macroorganisms like plants and insects, with various functions. The nitrogen-fixing bacteria (Masciarelli et al. 2014) and mycorrhizal fungi (Van der Heijden et al. 1998), for example, supply solubilized nitrogen (N) and/or phosphorus (P) to host plants, thereby decreasing chemical fertilizer utility into agroecosystems. Some of these symbionts are often known to increase host plants' resistance to salinity and drought stress (Calvo et al. 2014). Nematophagous/entomophagous fungi (Behie et al. 2017; Zavala-Gonzalez et al. 2017), disease-suppressive bacteria (Mendes et al. 2011), and phyllosphere endophytes (Arnold et al. 2003; Morris and Monier 2003) may play essential, but often overlooked, roles in suppressing populations of soilborne or airborne pathogens and pests. Several sources of microbial functions remained to be explored not only in the agroecosystems but also in the natural ecosystem (Narisawa et al. 2007; Usuki and Narisawa 2007). Therefore, it is very important to understand the mutational changes due to the environmental factors which play a significant role in microbial diversity.

10.3 Exponential Growth of Microbial Genomes

The first bacterial genome, viz., *Haemophilus influenzae*, was completely sequenced way back in 1995 by Fleischmann et al. (1995) and Fraser et al. (1995). The technological improvements over the last 23 years allowed a marked increase in the understanding of microbial genomes. The revolutionized speed, productivity, and significant cost reduction of genome sequencing have become advantages in the

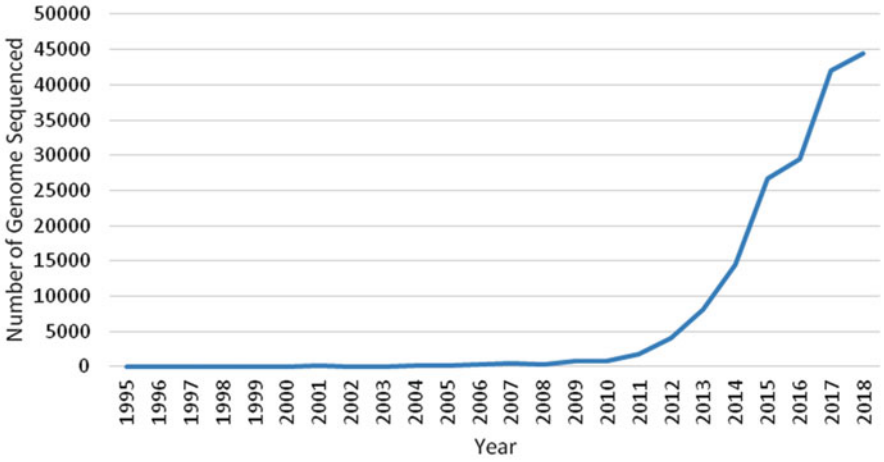

Fig. 10.1 Exponential increase in archaeal and bacterial genomes sequenced per year submitted to NCBI. (Source: Genbank, prokaryotes.txt file downloaded on 21 November 2018)

scientific world. These advantages have enabled not only the sequencing of many new genomes but also the widespread resequencing efforts to analyze genomic diversity. The first decade was dedicated to the first-generation sequencing method, and later, the second-/next-generation sequencing (Roche 454) was commercially introduced in 2005. Thereafter, in 2010, the third-generation (single molecule) sequencing, PacBio, was introduced commercially which has overcome several limitations of the second-generation sequencing such as an improved average read length. The *next-generation sequencing* (NGS) utilizes massive *parallel sequencing* to generate thousands of megabases of *sequence* information per day. The rapid growth of data shifts the cost from sequencing to management, analysis, and assembly of data. With the exposure of modern sequencing techniques, the number of sequenced archaeal and bacterial genomes is enhanced from 2 (1995) to more than 40,000 (2018) which represents the growth of more than 20,000-fold (Fig. 10.1). At present, more than 40,000 sequenced archaeal and bacterial genomes are available publically (NCBI 2018).

As the number of genomes has been increasing significantly, the phylogenetic classification has become essential to know their ancestral relatedness. For phylogenetic tree construction and taxonomic assignment, the ribosomal RNA genes, particularly the 16S rRNA gene, have been extensively utilized for a long time (Mizrahi-Man et al. 2013). The 16S rRNA gene remains extensively used due to its conserved sequence (at least one copy of this gene is present in all the taxa). The conserved nature of this gene permits simple specimen recognition by polymerase chain reaction, and its sequence gives authenticated data on different taxonomic ranks such as family, genus, or species in most cases. Nowadays, this single gene differentiation is being supplanted by advanced technologies where whole genome sequencing is done rather than single gene. To comprehensively examine and

classify hundreds and thousands of genomes, whole genome sequencing, in addition to several analysis tools, can be used. These new sequencing methods have established new visions in the understanding of hereditary connections which the 16S rRNA gene just approximates.

Over the last decade, a striking improvement in the bacterial genome sequencing has been the emergence of metagenomics, which beautifully covers all the genetic material present in a given environment or sample (Mende et al. 2012). The SRA maintained by NCBI currently serves approximately 34,000 metagenomic projects publically and deliver countless sequencing data. The colossal growth of metagenomics data improves the understanding of microorganisms related with its biological system, such as human microbiomes, microbiome associated with soils of all types, microbiome of fresh and saline water samples, and plant microorganism association frameworks.

Over the past few years, it has been observed that the diversity of microorganisms keeps on growing at an astonishing rate (Lagesen et al. 2010). The revolutionized sequencing efforts currently resulted in, for example, thousands of *Escherichia coli* genomes rather than 20 *Escherichia coli* genomes known previously (Cook and Ussery 2013). The growth of genomic data still resulted in new insights into the plasticity and diversity of bacterial genomes.

Different sequencing applications, for instance, bacterial genomes, proteomes, transcriptomes, metatranscriptomes, etc., have resulted in a vast majority of data. To make sense out of this huge data, efficient bioinformatics tools are required. Most biologists end up in an excess of sequence data and are in urgent requirement for bioinformatics tool to enable them to understand their enormous quantity. It is clear that these patterns of data generation will continuously increase in the near future as genome data become low-cost and plentiful. Numerous new methods are available for data interpretation, although the ever-increasing amount of sequencing data will demand new handy and convenient bioinformatics tools.

The advancement of new sequencing platforms gave rise to a tremendous amount of data, to manage the yield, various new assembly algorithms, and software being developed. As the third-generation sequencing technology serves to the scientific community, the algorithms have continued to evolve (El-Metwally et al. 2013). The rate of sequence production has increased exponentially, and therefore, the conventional genome annotation pipelines are no longer ready to scale the data, and new methodologies are now being considered regularly (Nielsen et al. 2014; Pop 2009). High-performance clusters and development of highly efficient algorithms can only permit the comparison, visualization, and communication of the sequencing results.

10.4 Microbial Diversity

For the recovery of different microbial species, conventionally, the culture-based methodology utilizing diversified culture media has been used to maximize the yield (Hill et al. 2000). However, it is impossible to grow all the microorganisms in vitro (Rappe and Giovannoni 2003), possibly due to the limitations of growth conditions

provided by culture media (Trevors 1998). Although these techniques indicate that phylotypes of some bacterial divisions are refined and can adjust in a number of natural environments, others seem limited to specific habitats (Hugenholtz et al. 1998). The estimation of prokaryotic diversity in contemporary studies regularly utilize techniques dependent on the investigation of nucleic acid sequences of 16S rRNA genes, which identifies and measures the phylotypes that are hard to culture (Reysenbach et al. 1992).

In an assemblage or community, biodiversity has been described as the gamut of significantly distinct types of organisms and their relative abundance. According to information theory, the diversity of a community has also been defined in terms of the quantity and distribution of information (Torsvik et al. 1998). In order to understand the microbial association with the ecological system, it is essentially important to comprehend the genomic alterations in the microorganism with time, in addition to their response to environmental conditions. Based on this, one can explain the microbial diversity at three levels, viz., within species (genetic), species number (species), and community (ecological) (Harpole 2010). The species diversity mainly comprises of two components, namely, (i) species richness, which is referred as the total number of species, and (ii) species evenness, which is the distribution of individuals among these species. Oftentimes, species evenness remains unknown in the bacterial systems because individual cells, which are very rare, remain undistinguished at species level. An appealing plausibility for the estimation of biodiversity is to use divergence in the molecular characters, especially the percentage of either the nucleic acid homology or the difference in base sequence. Previously, the diversity has been elucidated based on taxonomic species which may confine the extent of data and relationship prevailed. The diversity of operational taxonomic unit (OTU) or even communities may give us a superior estimation of the functioning of an ecosystem. Diversity studies can be utilized to recover ecological data about community structures. Species diversity is a parameter associated with the degree of stability of a community. Basically, any diversity index measure provides information about the heterogeneity of a community. The stability of a community is a function of its ability to withstand (resistance) and come back (resilience) from change (Yannarell and Triplett 2005). Diversity can hence be utilized to screen successions and impact of perturbations.

10.5 The Rationale Behind Studying Microbial Diversity

Within the natural microbial populations, a large amount of genetic data remains unknown, and it has been observed that the culturable bacteria represent an insignificant proportion of the aggregate bacterial population present (Giovannoni et al. 1990). However, it is vital to explore both the non-culturable and the culturable bacteria from different environments, where the samples can be compared using diversity studies. Another imperative purpose behind studying the microbial community diversity is insufficient information about the wiped out and surviving microorganisms. No harmony between the number of species that exist on the planet,

their vanishing or developing rate, or their metabolic utility is identified. The capacity of an ecological system to resist outrageous stress or perturbation conditions can somewhat be reliant on the microbial diversity within the system. Diversity analyses are in this manner vital in order to:

1. Increase the understanding of the microbial diversity of hereditary assets and comprehend the organism's distribution
2. Increase the understanding of the functional potential of microbial diversity
3. Understand the synchronization of biodiversity
4. Understand the outcomes of biodiversity, that is, to what extent does the sustainability and functioning of an ecosystem rely on preserving a specific level of microbial diversity?

10.6 Factors Controlling Microbial Diversity

In a bacterial assemblage, various microbes will play out similar processes and most likely be constituted in a similar niche (Zhao et al. 2012). The biotic and abiotic factors are the two characteristic groups that influence microbial diversity. The chemical and physical factors, for example, water accessibility, saltiness, temperature, oxic/anoxic conditions, pH, chemical contamination, pressure, pesticides, heavy metals, antibiotics, and so on, are considered as abiotic factors (Bååth et al. 1998). The accessory DNA, viz., phages, plasmids, and transposons, which impact the hereditary characteristics, such as the phenotypic changes of their host, are considered as biotic factors, which greatly affect microbial diversity (Zhao et al. 2012). Moreover, protozoans are likewise announced as affecting the microbial diversity (Clarholm 1994). The diversity profile of an ecological system is generally influenced by all the abovementioned factors in various manners and to various degrees and may shift the bacterial community.

10.7 Methods for Describing the Diversity of Microorganisms in Agroecosystems

The conventional techniques for microbial community analysis have been replaced by the molecular techniques, as the conventional techniques were highly slow and laborious. These new methodologies give a better understanding of the complex interactions occurring inside indigenous microbiota of different environments, a couple of which include host and host-pathogens, plants, and soil microbiota. With the advancement of next-generation sequencing (NGS) methods, it has turned out to be significantly faster and more economical to assess the complexity of microbiota in an environment extensively. This section focuses on these molecular methodologies for understanding complex microbial communities and their implications along with a brief introduction of the conventional methods (Fig. 10.2).

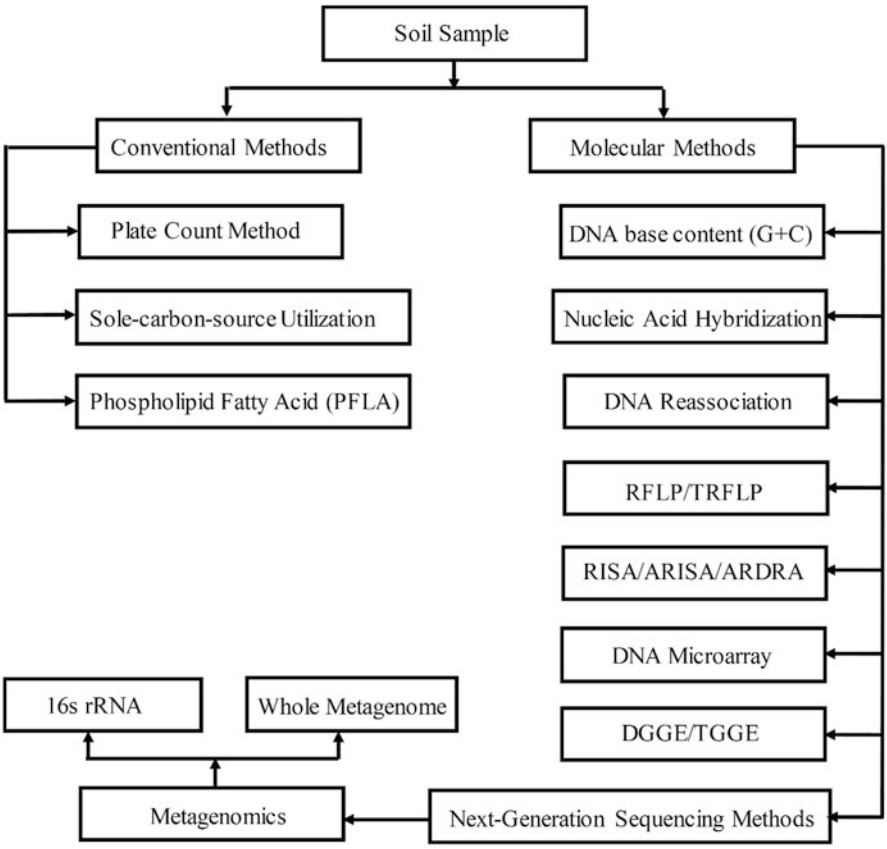

Fig. 10.2 Conventional and molecular methods to characterize the microbial diversity in the agroecosystems

The techniques to measure the microbial diversity in soil can be classified into two groups, viz., biochemical (Table 10.1) and molecular (Table 10.2).

10.7.1 Traditional and Biochemical Methods for Elucidation of Microbial Community Diversity

For the investigation of the microbial diversity, both traditional and biochemical techniques are highly important. Diversity measures based on physiological parameters can also be employed since such measures bypass the challenges that may emerge due to the grouping of analogous bacteria into species or clusters. These measures incorporate different parametric indices (resistance, mode of nourishment,

Table 10.1 Merits and demerits of the biochemical and traditional methods to study microbial community diversity (Kirk et al. 2004)

Method	Merits	Demerits
Plate counts	Economical	Partial toward fast-growing microbes
	Quick method	Only culturable microorganisms detected
		Bias toward large spore-producing fungal species
Sole-carbon-source utilization (SCSU) pattern/community-level physiological profiling (CLPP)	Quick	Only culturable microorganisms detected
	Comparatively less expensive	Support fast-growing microorganisms
	Highly reproducible	Only favors carbon-utilizing microorganisms
	Able to recognize a variety of microbes present in the communities	Considerate toward inoculum density
	Produce huge data	No in situ diversity, show potential metabolic diversity
	Option of using fungal and bacterial plates or site-specific carbon resources (Biolog)	
Fatty acid methyl ester analysis (FAME)/phospholipid fatty acid (PLFA) analysis	Culture not required	Amount of material needed is more, sample is fungal spores
	Allow direct isolation from soil	Other microorganisms can confound results is possible
	Follow specific communities or organisms	External factors can influence

and so on). In order to extract relevant information from large datasets, multivariate data analysis have also been regularly used (Sørheim et al. 1989). The early microbiologists often used various energy sources such as nitrogen and carbon along with the required growth factors to distinguish the different types of microbes in order to study their metabolic properties. However, the taxonomic classification based on various carbon and energy metabolisms may or may not follow the evolutionary conserved patterns of rRNA.

10.7.1.1 Microbial Diversity Estimation Using Plate Count Method

The selective and differential plating method to achieve the viable counts is the most traditional technique used for the estimation of microbial diversity; however, several challenges persisted with this method. These methods include (1) removal of

Table 10.2 Merits and demerits of some of the molecular methodologies to explore soil microbial community diversity (Kirk et al. 2004)

Method	Merits	Demerits
Mol % guanine plus cytocine (G +C)	Includes total isolated DNA	Extraction and lysing efficiency need extreme supervision
	Polymerase chain reaction biases do not affect the outcome	Requirement of plenty of DNA sample
	Quantitative	Resolution is not up to the mark
	Ability to detect rare members	
Nucleic acid reassociation	Total DNA isolated	High copy number of sequences required for detection
	Can study RNA or DNA	Less sensitive
	PCR biases do not influence	Extraction and lysing efficiency need extreme supervision
	In situ analysis can be possible	
DNA microarrays and DNA hybridization	Similar to nucleic acid hybridization	Culture of microorganisms is an essential step
	If using DNA fragments or genes, achieved specificity would be high	Most abundant species can only be detected
	Investigation of thousands of genes can be possible	Only in low diversity systems accurate results inferred
Denaturing and temperature gradient gel electrophoresis (DGGE and TGGE)	Reproducible, rapid, and reliable	Extraction and lysing efficiency need extreme supervision
	Parallel analysis of high number of samples can be done	Biases generated from PCR may affect the results
		Greater than one species can be represented by single band
		Sample handling can influence community structure, i.e., if sample stored too long before isolation, then community may change
		Fast-growing species can only be detected
Restriction fragment length polymorphism (RFLP)	Discern structural dynamics in microbial diversity	Complex banding patterns
		PCR biases
Terminal restriction fragment length polymorphism (T-RFLP)	Banding patterns simpler than RFLP	Influenced by PCR biases
	Can handle high number of samples simultaneously	Extraction and lysing efficiency need extreme supervision
	Can be automated	

(continued)

Table 10.2 (continued)

Method	Merits	Demerits
		Community fingerprint will change with change in restriction enzymes
	Between communities differences can be observed	Different forms of *Taq* lead to increased variability
	High reproducibility	
Ribosomal intergenic spacer analysis (RISA)/automated ribosomal intergenic spacer analysis (ARISA)	Community profiles are highly reproducible	PCR biases
		High DNA sample required (in case of RISA)

microorganisms from soil particles; (2) use of appropriate growth medium (Tabacchioni et al. 2000); (3) specific environmental conditions such as temperature, pH, and light; and (4) inability to detect all bacterial species (Trevors 1998).

10.7.1.2 Microbial Diversity Estimation Using Sole-Carbon-Source Utilization (SCSU)

In 1991, Garland and Mills introduced the community-level sole-carbon-source utilization (SCSU) method. The API and Biolog technologies are examples of biochemical identification systems. This system has been employed to identify the bacterial species based on their metabolic properties. Biolog employs a simple universal single color redox chemistry to detect reactions of bacterial cells. It consists of 96 wells and each well is a different test. Cell suspension is added and incubated over a period of hours. Cells will generate energy in some of the wells, reducing the redox dye and forming a purple color in all positive wells. SCSU have been quoted in some previous research to identify the microbial metabolic potential in arctic soils (Derry et al. 1999).

10.7.1.3 Microbial Diversity Estimation Using Phospholipid Fatty Acid (PFLA)

The microbial community characterization has been extensively carried out using microbial fatty acid composition. The fatty acids lying in the range of C2–C24 have provided the largest information about taxonomy and show their presence across a wide range of microorganisms (Banowetz et al. 2006). The composition of fatty acid remains stable and not affected by the presence of plasmids and mutations. The method is cheap, robust, and quantitative and has high reproducibility. The method allows the identification of bacterial populations based on the analysis of the fatty acid patterns, also referred to as the FAME (fatty acid methyl ester) analysis (Fig. 10.2).

10.7.2 Molecular Methods for Elucidation of Microbial Community Diversity

Conventional techniques for describing the microbial structure have been considering the culturable part of microbes. In natural environments, enormous undiscovered biodiversity exists. However, culture-based methods address very small part of this natural biodiversity which ends up in not uncovering the significant portion of microorganisms (Đokić et al. 2010). Nowadays, community analysis measurement, centered around the genetic information which do not require prior cultivation, has been used which is termed as molecular approach. In these methods, gene sequences (16S rRNA for prokaryotes) are used to distinguish the variable microbial community composition (Amann et al. 1995). Various methodologies have been evolved to infer microbial diversity at molecular level. Some of these are discussed in following sections.

10.7.2.1 Microbial Diversity Estimation Using Base-Content of DNA

Deduction of mole percentage guanosine and cytosine (mol% G+C) represents a traditional tool of genomic content estimation (Sarethy et al. 2014). To establish the phylogenetic rank of an organism, this method is now being used along with other genotyping methods (Sarethy et al. 2014). The G+C content varies within 3% and 10% in case of species and genera, respectively (Sarethy et al. 2014). Overall, the G +C content varies from 25–75%. However, similar G+C content does not confirm a taxonomic relationship. On the other side, if there is a distinction in G+C content, then it can be a key indicator of missing relationship. Thermal denaturation of DNA can determine the mol% G+C. Advantages of this method are its quantitative nature, incorporation of all extracted DNA, unbiasness from PCR artifacts, and the ability to uncover the hard-to-detect members in the microbial community. The initial requirement for several samples of DNA, which can be up to 50 μg, is a main drawback of this method (Tiedje et al. 1999).

10.7.2.2 Microbial Diversity Estimation Using Nucleic Acid Hybridization

For the interrogation of the molecular community within a sample or natural surroundings, the nucleic acid hybridization method proved to be a useful tool as it provides both the quantitative and qualitative measures (Clegg et al. 2000). These hybridization methods can be done in situ or on isolated RNA or DNA. A specific oligonucleotide or polynucleotide probe derived from known sequences can be labeled with markers at the 5′ end to determine the community diversity (Goris et al. 2007)

10.7.2.3 Microbial Diversity Estimation Using DNA Reassociation

The DNA reassociation kinetics reflect sequence variations of the microbial community present in the environment. DNA reassociation determines the microbial diversity by estimating the genetic heterogeneity of the population (Torsvik et al. 1996). From environmental samples, total DNA is extracted, purified, denatured, and

permitted to reanneal. The similarity of sequences will decide the rate of hybridization. This method identifies the microbial diversity by analyzing the decreased rate of reassociation of DNA sequences (Theron and Cloete 2000).

10.7.2.4 RFLP (Restriction Fragment Length Polymorphism)/T-RFLP (Terminal Restriction Fragment Length Polymorphism)

This method employed PCR-based amplification of bacterial-conserved genes, viz., 16S rRNA, from the total community DNA with the help of complementary oligonucleotides. PCR amplification is followed by tandem tetrameric restriction enzyme digestion (Moyer et al. 1994). The fragments generated are then separated by size using electrophoresis and display different migration profiles which ultimately give rise to the restriction fragment length polymorphism patterns of 16S rRNA genes showing distinct operational taxonomic units (OTUs) (Moyer et al. 1994).

For the study of microbial diversity, usage of RFLP method has been shown in the last couple of years (Moyer et al. 1996). RFLP analysis in which RNA gene has been used as a probe to investigate intraspecies variation is also known (Kauppinen et al. 1994). Additionally, closely related strains are often distinguished using enzyme electrophoresis and DNA–DNA hybridization (Goodfellow and O'Donnell 1993). A particular species may have more than five restriction fragments. In diverse communities, the screening of these banding markings becomes very difficult to analyze by RFLP (Tiedje et al. 1999). Moreover, a similar banding pattern between the organisms being compared is not always equivalent to their close relatedness (Tiedje et al. 1999). Further, this method is not useful in the identification of specific taxonomic groups (Liu et al. 1997).

T-RFLP is an add-on of RFLP in which the beginning steps such as DNA extraction, PCR-based amplification, and fragmentation remain similar; however, one of the primers is associated with a fluorescent tag (6-FAM) so that the sizes of only the terminal restriction fragment can be identified (Liu et al. 1997). The length of the unique restriction fragment can be counted in OTU terms, and the occurrence of these OTUs is calculated (Fakruddin and Mannan 2013). The species richness, evenness, and similarities between samples can be measured using banding patterns (Fakruddin and Mannan 2013; Liu et al. 1997). Restrictions of T-RFLP include limited options for universal primers, which can only detect the numerically dominant species present and underrepresent the true diversity (Rudi et al. 2007). The overestimation of diversity often occurs due to incomplete digestion by restriction enzymes (Osborn et al. 2000).

10.7.2.5 RISA (Ribosomal Intergenic Spacer Analysis)/ARISA (Automated Ribosomal Intergenic Spacer Analysis)

The working mechanism of RISA and ARISA is similar to RFLP and T-RFLP, viz., microbial diversity estimation occurs based on ribosomal fingerprinting. PCR-based amplification of ribosomal intergenic region (region between the 16S and 23S ribosomal subunits) is carried out in this method. The DNA melting and segregation of ribosomal intergenic region is done with polyacrylamide gel electrophoresis under

specific denaturing conditions. The heterogeneity in the length and sequence of these ribosomal intergenic regions provides the opportunity to distinguish interrelated strains and intra-related species (Fisher and Triplett 1999). The sequence polymorphisms are detected using silver staining and fluorescently tagged primer in RISA and ARISA, respectively (Fisher and Triplett 1999). RISA has been applied to differentiate microbial diversity of soil system (Borneman and Triplett 1997) and in plant rhizosphere (Borneman and Triplett 1997).

10.7.2.6 DNA Microarrays

Reverse sample genome probing (RSGP) is a DNA microarray method that characterizes community composition based on whole-genome DNA–DNA hybridization (Bae et al. 2005). This technique follows a four-step process: (1) isolation of chromosomal DNA from pure cultures, (2) cross-hybridization testing to define species with limited genomic cross- hybridization, (3) preparation of genome arrays by spotting known amounts of denatured genomic DNAs from all identified standards on a solid support, and (4) random labeling of a defined mixture of total community and internal standard DNA (Greene and Voordouw 2003; Kirk et al. 2004). This method has been developed to determine the culturable part of microbial composition in soil and in oil fields (Greene and Voordouw 2003).

10.7.2.7 Denaturing Gradient Gel Electrophoresis (DGGE)/Temperature Gradient Gel Electrophoresis (TGGE)

Denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) molecular techniques are used to investigate microbial community composition and dynamics. These molecular techniques explore microbial diversity by sequence-specific (16s rRNA) separation of double-stranded (ds) PCR products of the same size with distinct sequences based on their melting characteristics. DNA fragments have their own melting temperature based on their nucleotide sequence. The GC bindings are stronger than AT bindings because GC bindings contain three hydrogen bonds where AT bindings contain two hydrogen bonds; therefore, GC-rich sequences have a higher melting point. DGGE uses an extra sequence called "GC-clamp," which prevents a complete separation of the DNA strand. The GC-clamp is GC-rich sequences and is placed at the end of the PCR products. DGGE uses a linear gradient of denaturant, a substance composed of urea and formamide, which triggers the DNA to denature, while in TGGE the gradient is temperature (Naum and Lampel 2011). Amplification of the microbial community members is done using sequence-specific primers (16S). Species are then distinguished by comparing the migration distance of the PCR products in gels with those of reference strains (Naum and Lampel 2011).

Benefits include reproducibility, reliability, low cost, and rapidness along with back-tracing in reaction to any adverse conditions (Muyzer 1999). Limitations include errors in PCR (von Wintzingerode et al. 1997), variable DNA isolation productivity (Theron and Cloete 2000), species having their high number in community can only detected (MacNaughton et al. 1999), sample handling is quite back-breaking (Muyzer 1999), and identical melting characteristics of unlike sequences

leads to false results (Niemi et al. 2001). PCR-DGGE method has been applied for the deduction of bacterial community composition in soil with the antibiotic cefuroxime treatment and inoculated with multidrug-resistant bacteria *Pseudomonas putida* (Orlewska et al. 2018).

10.7.2.8 Next-Generation Sequencing Technologies

The next-generation sequencing (NGS) techniques like Illumina sequencing and pyrosequencing have provided an opportunity to discover new microorganisms and to explore their complexities. The real-time sequencing techniques enable the study of complex behavior of microorganisms in different environments and within their populations (BARTRAM et al. 2011). A higher resolution of the complex microbiota composition along with the elucidation of niche function with higher accuracy can be achieved using NGS. Next-generation sequencing techniques provide deep insights into the complex microbial communities under different environmental conditions with the advent of its high throughput nature (Fakruddin and Mannan 2012).

Several sequencing methods are available until now, and the technology is being continuously evolved to come up with an improved version. The sequencing methods are mainly grouped into three types: sequencing by synthesis (SBS), sequencing by ligation (SBL), and single-molecule sequencing (SMS).

10.7.2.8.1 Sequencing by Synthesis

NGS techniques which use chemiluminescence produced by nucleotide incorporation during the complementary strand synthesis by polymerase enzyme are grouped into sequencing by synthesis methods. One example of such techniques is the Sanger sequencing method, which uses chemiluminescence to determine the base composition. Sanger sequencing method uses dideoxynucleotide to determine the nucleotide sequence of a DNA strand with the requirement of varied template DNA strand sizes. In this technique, DNA is cleaved to reasonable sizes, adaptor-ligated, and cloned to amplify the fluorescence or chemiluminescence, and templates are immobilized in preparation for flow-cell cycles. The basic principle of sequencing chemistry differs in various methods, although the sequential washing steps of nucleotides remain the same for all methods. The SBS methods also differ by template amplification, read length, and immobilization process to the flow cell.

10.7.2.8.1.1 454 Sequencing

In this sequencing method, the DNA template sticks on microbeads, and subsequent amplification is done by emulsion PCR. Microbeads are placed in PicoTiterPlate with each well containing a single bead, which is incubated with apyrase, luciferase, DNA polymerase, and ATP sulfurylase along with luciferin and adenosine 5-′-phosphosulfate (ASP) onto the flow cell. The incorporation of dNTPs with the help of DNA polymerase into the newly formed strand liberates pyrophosphate (Ppi). The existing ASP reacts with this Ppi and results into adenosine triphosphate (ATP). ATP acts as the catalyzer for luciferase enzyme to convert luciferin to oxyluciferin, and in turn light is emitted which is equivalent to ATP production (Ronaghi et al. 1998). Sequential washing is performed to remove all unused ATP

and nucleotides removed by apyrase; a new chemical mixture is again added over the DNA templates, and the process is repeated until the DNA template has elongated. As the nucleotide incorporation progresses, fluorescent light is emitted and detected by a camera which is eventually analyzed to produce sequence of bases to form the DNA sequence. The Roche 454 technique has started with low read lengths (~100 bp), but the technology has now improved and produce ~500 bp (Ambardar et al. 2016). This platform is being used in genome or transcriptome sequencing due to its longer read length, when de novo assembly is involved (Buggs et al. 2012; Egan et al. 2012; He et al. 2012; Lai et al. 2012; Strickler et al. 2012; Zalapa et al. 2012).

10.7.2.8.1.2 Illumina

Illumina Genome Analyzer was initially developed by Solexa. It uses solid-phase bridge amplification in which adapters are ligated to 5′ and 3′ end of a DNA template (Egan et al. 2012). The substrate further binds to one end of the fragment. The immobilized forward or reverse primers are hybridized by adaptors to create a bridge for amplification. This bridge facilitates amplicon generation and forms identical template clusters, which provide high-quality chemiluminescent detection. Each channel of flow cell generates millions of such clusters. DNA amplicons are denatured and primed after amplification, and cycles of washes are performed to conduct elongation. The wash cycle includes four nucleotides mixture, modified as 3′-*O*-azidomethyl, a different fluorophore and reversible terminator (Bentley et al. 2008; Guo et al. 2008). Further, the image is captured, but the elongation continues after the cleavage of fluorescent dye moiety. The 3′-OH group is restored by reacting with tris(2-carboxyethyl)phosphine (Egan et al. 2012). This cyclic process is repeated until the desired DNA fragment length is achieved. The HiSEq 2000 sequencer can use dual flow cell to produce ~6 billion paired-end reads. Illumina platform is the most commonly and widely used method (Azam et al. 2012; Buggs et al. 2012; Ilut et al. 2012; McKain et al. 2012; Steele et al. 2012).

10.7.2.8.1.3 Ion Torrent

Another unique sequencing technique is the Ion Torrent sequencer, which interrogates base calling through change in pH rather than fluorescent dyes. DNA sequencing is performed using a semiconductor chip, which has millions of wells. As the DNA sequencing progress, the chemical information is captured in the wells and translated into base calls. The sequencing process starts when a sample of DNA is fragmented into millions of segments, and each segment then attaches to its own bead and is copied until it covers the bead. This automated process covers millions of beads with millions of different fragments. These beads then flow across the chip, each depositing into a well. The chip is loaded with one of the four types of bases. As a nucleotide is incorporated into growing DNA strand, H+ ion is liberated. The release of H+ ion changes the pH of solution in the well and converted into voltage (Rothberg et al. 2011). The important factor with Ion Torrent is its read length (~250) and its low sample preparation cost in comparison with other sequencing platforms. The utility of Ion Torrent has been reported for mutational discovery in the flax genome (Galindo-González et al. 2015).

10.7.2.8.2 Sequencing by Ligation

In case of sequencing by synthesis, DNA polymerase is used as the elongation machine for DNA sequence determination, while in case of sequencing by ligation methods, DNA ligase mismatch sensitivity acts as the main engine for nucleotide sequencing (Landegren et al. 1988). The characteristic of these methods include a variety of oligonucleotide probes with different fluorescent tags. The template DNA fragments are attached to the anchor sequence. This sequence helps in hybridization with the probes, and DNA ligase joins the probes to the primer and template. Further, the fluorescence image is captured to determine the probe incorporation, and the process is repeated until the target length is achieved. The methods elaborated here differ in the read length and probe usage.

10.7.2.8.2.1 SOLiD

Support oligonucleotide ligation detection (SOLiD) sequencer has been created by Applied Biosystems/Life Technologies, which uses a ligation sequencing method to determine base composition of DNA. Microbead-associated emulsion PCR is used for the enrichment of fragmented or mate-paired and primed libraries which are then supported onto a glass slide. An eight nucleotide long probe is added to the flow cell. At the left of the probe, two nucleotides exist that are the actual bases (dinucleotides), and each one is one of the four canonical bases. Each dinucleotide permutation corresponds to a dye color, and 16 possible dinucleotide permutations are available, so each dye color corresponds to four of these. A total of four different colors (red, green, blue, and orange) are available. The middle three universal bases are separated from right six to eight universal bases through phosphorothiolate linkage. The fragment that is being sequenced is annealed with primer which corresponds to the universal adapter P1. After primer annealing, the probe is attached with primer by DNA ligase. The fluorescence snapshot is produced, and the phosphorothiolate bond is cleaved, and the 5′ phosphate group is regenerated for successive ligation (McKernan et al. 2016). This cyclic process is repeated until the desired length of complementary strand is achieved. SOLID sequencing has been used in genomic sequencing, transcriptomics (e.g., Shulaev et al. (2011)) and resequencing studies (e.g., Ashelford et al. (2011)).

10.7.2.8.2.2 Polonator

The Polonator G.007 system is available through Azco Biotech and was developed at the Harvard University in Dr. George Church's laboratory in collaboration with Dover Systems (Egan et al. 2012). The library preparation for base sequencing in the Polonator system is achieved using emulsion PCR to amplify the template DNA, followed by loading of beads on flow cells and use of automated polymerase colony sequencing by ligation method (also called as polony) (Shendure et al. 2005). A full eight-flow-cell Polonator sequencing "run" can generate up to 240 million mappable reads of sequence, and a read length of 40 bases can be achieved in ~80 (Egan et al. 2012). The Polonator system is a benchtop, open-source platform. This platform is considered as a complete package since it provides all system aspects including machine and software. Additionally, the platform is open source which allows laboratories or even individual researchers to develop or explore protocols and

applications according to their requirements. This system has been reviewed by many researchers as a NGS sequencing platform (Deschamps and Campbell 2010; Moorthie et al. 2011; Myllykangas et al. 2012; Pareek et al. 2011), mainly for Personal Genome Project, but its use in plant genomics remains limited.

10.7.2.8.3 Single-Molecule Sequencing

The third-generation sequencing methods or commonly referred to as single-molecule sequencing (SMS) methods are another promising NGS technologies available. These technologies are based on the detection of chemiluminescence which is produced by the incorporation of a single nucleotide. It has eliminated the use of DNA template amplification during sequencing because this method can detect single nucleotide incorporation during DNA synthesis. This method has an advantage over other NGS methods as this method has simplified sample preparation process and can be used for sequencing of degraded or low-concentrated samples (e.g., Orlando et al. 2011). As PCR amplification is not required, it also avoids the PCR errors and biasness introduced during template amplification or during cDNA amplification in RNA-seq (Ozsolak et al. 2009). SMS methods use fluorescence imaging to detect nucleotide incorporation during base sequencing. The SMS technologies differ in several processes like detection of emitted light, minimization of background fluorescence, chemistry used, and the immobilization process of template and other molecules onto the flow cell. These technologies are comparatively new in the NGS field and further developing to become more readily available.

10.7.2.8.3.1 Helicos

In the SMS technologies described above, the Helicos Genetic Analysis System was the first system which was commercially available in the market (Harris et al. 2008). This technique uses two methods: in one method, denatured and fragmented DNA templates are directly immobilized to the template on a solid surface by linking covalently and universal primers are used for priming, while in the other method, they are hybridized to oligonucleotide primers which are immobilized to a solid surface (Thompson and Steinmann 2010). The DNA template molecules are being washed with DNA polymerase and "virtual terminator" nucleotides which are fluorescently labeled. The fluorescence imaging is detected by the incorporation of nucleotides, one at a time. After imaging, a cleavage step removes the fluorescent labels, and the process continues through each of the remaining bases. This process is repeated until the desired read length is achieved (Bowers et al. 2009; Egan et al. 2012). The average read length of 35 bp is produced by this method across 600 million to 1 billion reads, generating 21–35 Gb per run at a rate of >1 Gb/h (Egan et al. 2012). Another remarkable feature of this platform is multiplexing where up to 4800 samples per run or 96 samples per channel can be sequenced.

10.7.2.8.3.2 Pacific BioSciences

The PacBio *RS* SMS platform has been introduced in 2010. This system has a special feature, viz., zero mode waveguide (ZMW). As ZMV is illuminated, the wavelength of light is too large to allow it to pass through the waveguide. Attenuated light from the excitation beam penetrates the lower 20–30 nm of ZMV, creating the world's

most powerful microscope with a detection volume of 20 zeptoliters. A DNA template-polymerase complex is immobilized at the bottom of the ZMV (Eid et al. 2009; Levene et al. 2003). The ZMW allows the nearest nucleotide to be actively incorporated and has the strongest fluorescent signal. The DNA polymerase is bound to the DNA template for base sequencing. The four nucleotides are mixed and each nucleotide is labeled with a different fluorescent dye. The nucleotide sequence is determined by fluorescence imaging. The real-time methods can be accommodated with immobilized DNA polymerase model to analyze DNA templates which are much longer, and one can achieve potential length of base pairs in thousands (Metzker 2010). The laser excitation eventually degrades immobilized DNA polymerase and limits the read lengths. This method produces the highest error rate among all NGS methods (~15%) and requires several runs of sequencing for the same molecule to reduce or remove errors (Eid et al. 2009; Metzker 2010). The new SMRTbell protocol with the circularized templates and repeated sequencing has increased accuracy to 99.999% for 30X coverage.

10.7.2.8.3.3 Nanopore Sequencing

Oxford Nanopore Technologies Limited is the main developer of nanopore sequencing and has developed the GridION system. The main alluring feature of this technology is that it can directly analyze a strand molecule (DNA or RNA) without the need of amplification steps in library preparation. Nanopore technology uses the inherent properties like electronic or chemical of nucleotide for base sequencing. Currently, nanopore sequencing technology uses two methods for a DNA molecule: one is strand sequencing where a single strand of DNA is passed through a nanopore with the help of a specific enzyme for base encoding and detection (Clarke et al. 2009; Egan et al. 2012). The other method is exonuclease sequencing (Howorka et al. 2001; Lieberman et al. 2010) where at a time one base is cleaved and subsequently passed through the nanopore. A remarkable speed of nanopore sequencing platform (GridION) to pass a DNA molecule enables rapid sequencing throughput production. Dr. Mark Akeson's laboratory at the University of California is using this technology in an NHGRI-funded project in Santa Cruz as part of the "$1,000 genome program" (Egan et al. 2012).

10.7.2.9 Metagenomic Analysis of Microbial Communities

In order to investigate the collective genome of an environmental sample, the branch of science called metagenomics is used to provide information on the collective microbial diversity (Zeyaullah et al. 2009). The microbial communities (Ghazanfar et al. 2010) are randomly sampled from the environment and subsequently sequenced (Schloss and Handelsman 2003). The straight analysis of the metagenome has the potential to give a comprehensive view of the species composition, genetic diversity, interactions with the environment of natural microbial communities, and evolution (Simon and Daniel 2011). The metagenomic analysis followed by gene expression and proteomic studies can be used to determine how functions are distributed, and resources are utilized among the microbial community. Eventually, the field of genomics can reveal the net contribution of individual species and strains in the community (Allen and Banfield 2005).

10.7.2.10 Community Genomics Analyzing Tools and Databases

The natural microbial phenomena such as population ecology, biogeochemical activities, microbial interactions, and evolutionary processes like lateral gene transfer events are accessed using community genomics (Allen and Banfield 2005). The analysis of regulatory networks and global gene expression patterns of community genomic data is produced using DNA microarray in a parallel and rapid mode. In the microbial community, complex linkages between distinct gene and gene families and the distribution of metabolic functions exist which can be uncovered by community microarray analyses. Various genome assemblers such as CAP, EULER, CELERA, PHRAP, JAZZ, TIGR, and ARACHNE are currently accessible for community genomics data analysis (Tyson et al. 2004).

Traditionally, metagenomics sample has been characterized using richness and diversity measure (Colwell 2009; Ondov et al. 2011). Examples of different diversity measures are Shannon, Chao, and Simpson that quantify the evenness of the distribution of the abundances of the taxa, often incorporating distance measures, for example, Jaccard, Sorenson, and Bray–Curtis (Colwell 2009). These indices offer measures of complexity of the community but disclose little about interactions within the community, which requires more complex downstream analyses.

Several web-based tools have been developed for the visualization of taxonomic profiles. The Visualization and Analysis of Microbial Population Structure (VAMPS) tool (Huse et al. 2014) can visualize and measure the differences and similarities among various taxonomic profiles of complex microbial communities along with their statistical inference. On the contrary, the web-based tool Krona provides visualization as a pie chart with an embedded hierarchy for taxonomy analysis (Ondov et al. 2011)

Microbial profiles of the natural environment are determined using metagenomics, but additional relevant information is extremely valuable for their deeper insights. Based on this, many approaches have been using the phylogenetic information to enhance the classification of sequences, like Amphora2. The Amphora2 provides phylogenetic inference based on phylum-specific marker databases (Wu and Scott 2012). Moreover, the same type of inference can be achieved algorithmically through squash clustering edge principal component analysis (PCA) (Matsen and Evans 2013).

The taxonomic classification of microbial communities gives a brief introduction of "who" is present in the community, but "what" these microbes did is essentially important to understand as it builds insights into the underlying biological processes. This phenomenon is possible to manifest with the help of the functional annotations of the genes to which reads are mapped (Meyer et al. 2008). The useful resources for annotation include Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000), Clusters of Orthologous Groups (COG) (Tatusov et al. 1997), and Gene Ontology (GO) (Ashburner et al. 2000). For HMP, the functional and metabolic potential of WMS data was performed using HUMAnN (Abubucker et al. 2012).

Many other existing tools are available, such as IMG/M (Markowitz et al. 2011), CAMERA (Seshadri et al. 2007), METAREP (Goll et al. 2010), MEGAN (Huson et al. 2011), and CoMet (Lingner et al. 2011), which are often used to generate

microbial functional profiles. MEGAN is a standalone computer program, while METAREP, CoMet, and IMG/M are web-based tools. CAMERA provides computational structure for high-performance network access and grid computing. Alignment of query sequences is performed using BLAST for GO and KEGG database in METAREP and CoMet, whereas the NCBI taxonomy is used in MEGAN. METAREP often provides the option for taxonomic annotations, and IMG/M uses BLAST to generate phylogenetic information. However, IMG/M is more focused on protein-related information by annotating the results with resources, such as Pfam, COG, ENZYME, KEGG, and TIGRFAMs. METAREP produces heat maps, graphical summaries, and hierarchical clustering plots. In addition, METAREP also performs statistical tests and multidimensional scaling (MDS) to infer significant results. MEGAN uses the naïve and weighted lowest common ancestor algorithm to assign the reads and has been applied in some of the pioneer metagenomic studies, such as the data from mammoth bone and Sargasso Sea. Finally, CoMet combines Pfam domain family-derived assignments and ORF finding with comparative statistical analysis, providing the user with visualizations in the form of hierarchical clustering and MDS comprehensive tabular data files. This was applied to 454 data.

In targeted approaches, typically the sequence data do not directly provide the functional inference of a microbial community. However, now it is possible after the development of the tool, PICRUSt, which directly provides functional inference from taxonomic profiles of marker genes, such as the 16S rDNA, and a database of reference genomes (Langille et al. 2013).

10.7.3 Progress in Plant Breeding

Lai et al. (2012) took a variety of tissues for 11 weeds in the sunflower family and generated 22 EST libraries using Sanger, 454, and Illumina sequencing. They further compared the quality of sequence assemblies and coverage. Eventually, they performed expression profiling to five of them through NimbleGen microarrays. For the detection and quantification of hybridization and interrogation, they have also compared the distributions of Ks values between orthologs of congeneric taxa. Using this method, substantial introgression was detected among congeneric taxa of Centaurea and Helianthus, but not in Ambrosia and Lactuca; such data will be useful for long-standing questions of whether hybridization is a cause or consequence of range expansions. They found that gene discovery was enhanced by sequencing from multiple tissues, normalization of cDNA libraries, and especially greater sequencing depth. Some of the applications of NGS in plant breeding is depicted in Fig. 10.3.

10.7.4 Conclusion and Future Scope

The molecular genomic tools have been progressively developing, and their application in the field of agriculture microbial diversity is undergoing unrivalled

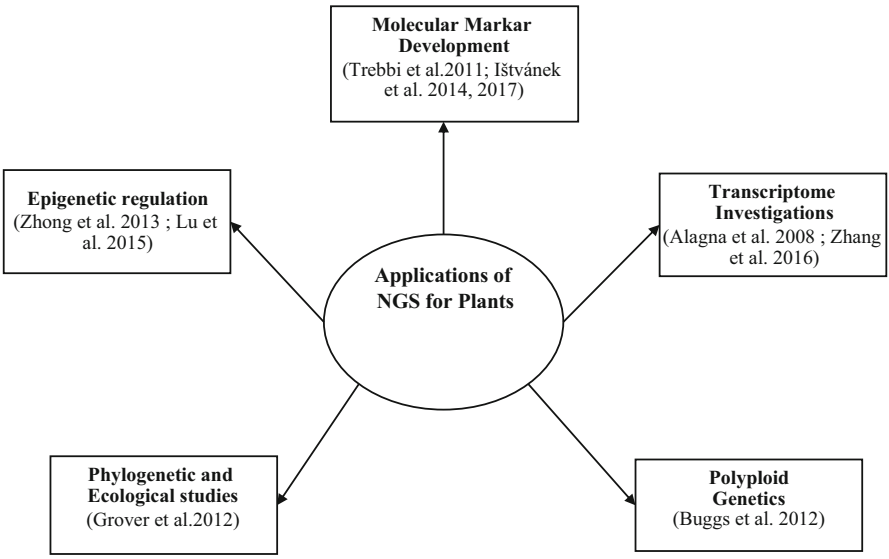

Fig. 10.3 Applications of next-generation sequencing technologies for plant breeding

changes. The examination of functional and taxonomical diversity of microbial communities through the postgenomic molecular approaches made us realize that we have only interpreted a minimal volume of the metabolic and genetic diversity present on Earth. In the present era, several research-oriented queries such as "How many microbial species exist on the earth particularly in context of agroecosystems?", "What is the range of metabolic diversity?", and "How microbial diversity is regulated by chemical, physical and biological factors?" remain to be inferred. The number of identified uncultured organisms significantly has been increasing, and the understanding of their functional roles has become a daunting task because of the unavailability of homologous genes in the databases or repositories. Breakthrough research has been done in the elucidation of microbial diversity by the application of next-generation sequencing methods. However, many technical issues remain including mRNA instability and extraction of genetic material. The development in NGS methods remains in progress to address some of these issues. As metagenomics, including many other NGS applications, tremendously produce high amount of data, the development of highly efficient bioinformatics tools are much needed for their evaluation. All of the molecular approaches accessible for microbial diversity and function analysis entail some pros and cons. None of these explore the genetic and functional diversity in complex microbial communities completely. The integration of various approaches should be applied to investigate the dynamics of microorganisms. Culture-independent and culture-based molecular techniques are neither excluding nor contradictory and should be performed interdependently. Toward this, a new emerging field for the identification of bacterial species, called culturomics, is being developed, which utilizes MALDI-TOF mass

spectrometry, multiple culture conditions, and 16S rRNA sequencing. This method results in the increment of cultured bacteria in the human gut (Lagier et al. 2018). A multifaceted systems approach holding various "omics" techniques will be needed to develop accurate and in-depth understanding into agroecosystems in order to uncover the interrelationship among environmental factors, proteins, and genes.

References

Abubucker S et al (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 8:e1002358

Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. Nat Rev Microbiol 3:489

Amann RI, Ludwig W, Schleifer K-H (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J (2016) High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol 56:394–404

Arnold AE, Mejía LC, Kyllo D, Rojas EI, Maynard Z, Robbins N, Herre EA (2003) Fungal endophytes limit pathogen damage in a tropical tree. Proc Natl Acad Sci 100:15649–15654

Ashburner M et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25

Ashelford K et al (2011) Full genome re-sequencing reveals a novel circadian clock mutation in Arabidopsis. Genome Biol 12:R28

Azam S et al (2012) Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (Cicer arietinum; Fabaceae), a crop species without a reference genome. Am J Bot 99:186–192

Bååth E, Díaz-Raviña M, Frostegård Å, Campbell CDJA (1998) Effect of metal-rich sludge amendments on the soil microbial community. Appl Environ Microbiol 64:238–245

Bae J-W et al (2005) Development and evaluation of genome-probing microarrays for monitoring lactic acid bacteria. Appl Environ Microbiol 71:8825–8835

Banowetz GM, Whittaker GW, Dierksen KP, Azevedo MD, Kennedy AC, Griffith SM, Steiner JJ (2006) Fatty acid methyl ester analysis to identify sources of soil in surface water. J Environ Qual 35:133–140

Bartram A, Lynch M, Stearns J, Moreno-Hagelsieb G, Neufeld J (2011) Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. Appl Environ Microbiol 77(11):3846–3852

Behie SW, Moreira CC, Sementchoukova I, Barelli L, Zelisko PM, Bidochka MJ (2017) Carbon translocation from a plant to an insect-pathogenic endophytic fungus. Nat Commun 8:14245

Bentley DR et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53

Borneman J, Triplett EW (1997) Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. Appl Environ Microbiol 63:2647–2653

Bowers J et al (2009) Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods 6:593

Buggs RJ et al (2012) Next-generation sequencing and genome evolution in allopolyploids. Am J Bot 99:372–382

Calvo P, Nelson L, Kloepper JW (2014) Agricultural uses of plant biostimulants. Plant Soil 383:3–41

Clarholm M (1994) The microbial loop in soil. In: Ritz K, Dighton J, Giller KE (eds) Beyond the biomass. Wiley-Sayce, Chichester, pp 221–230

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265

Clegg CD, Ritz K, Griffiths BS (2000) % G+ C profiling and cross hybridisation of microbial DNA reveals great variation in below-ground community structure in UK upland grasslands. Appl Soil Ecol 14:125–134

Colwell RK (2009) Biodiversity: concepts, patterns, and measurement. In: Levin SA (ed) The Princeton guide to ecology. Princeton University Press, Princeton, NJ, pp 257–263

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

Derry A, Staddon W, Kevan P, Trevors J (1999) Functional diversity and community structure of micro-organisms in three arctic soils as determined by sole-carbon-source-utilization. Biodivers Conserv 8:205–221

Deschamps S, Campbell MA (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol Breed 25:553–570

Đokić L, Savić M, Narančić T, Vasiljević B (2010) Metagenomic analysis of soil microbial communities. Arch Biol Sci 62:559–564

Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. Am J Bot 99:175–185

Eid J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

El-Metwally S, Hamza T, Zakaria M, Helmy M (2013) Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput Biol 9:e1003345

Fakruddin M, Mannan KSB (2012) Next generation sequencing technologies-principles and prospects. Res Rev Biosci 6:240–247

Fakruddin M, Mannan K (2013) Methods for analyzing diversity of microbial communities in natural environments. Ceylon J Sci 42

Fisher MM, Triplett EWA (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol 65:4630–4636

Fleischmann RD et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM et al (1995) The minimal gene complement of mycoplasma genitalium. Science 270:397–404

Galindo-González L, Pinzón-Latorre D, Bergen EA, Jensen DC, Deyholos MK (2015) Ion Torrent sequencing as a tool for mutation discovery in the flax (*Linum usitatissimum L.*) genome. Plant Methods 11:19

Ghazanfar S, Azim A, Ghazanfar MA, Anjum M, Begum I (2010) Metagenomics and its application in soil microbial community studies: biotechnological prospects. J Anim Plant Sci 6:611–622

Giovannoni SJ, Britschgi TB, Moyer CL, Field KGJN (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60

Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methé BA, Yooseph S (2010) METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. Bioinformatics 26:2631–2632

Goodfellow M, O'Donnell AG (1993) Handbook of new bacterial systematics. Academic Press, London/San Diego

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91

Greene EA, Voordouw G (2003) Analysis of environmental microbial communities by reverse sample genome probing. J Microbiol Methods 53:211–219

Guo J et al (2008) Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci 105:9145–9150

Harpole W (2010) Neutral theory of species diversity. Nat Educ Knowl 3(10):60

Harris TD et al (2008) Single-molecule DNA sequencing of a viral genome. Science 320:106–109
He R et al (2012) Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, Phragmites australis (Poaceae), reveals genes involved in invasiveness and rhizome specificity. Am J Bot 99:232–247
Hill G et al (2000) Methods for assessing the composition and diversity of soil microbial communities. Appl Soil Ecol 15:25–36
Howorka S, Cheley S, Bayley H (2001) Sequence-specific detection of individual DNA strands using engineered nanopores. Nat Biotechnol 19:636
Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765–4774
Huse SM, Welch DBM, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML (2014) VAMPS: a website for visualization and analysis of microbial population structures. BMC bioinformatics 15:41
Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. Genome research: gr. 120618.120111
Ilut DC, Coate JE, Luciano AK, Owens TG, May GD, Farmer A, Doyle JJ (2012) A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. Am J Bot 99:383–396
Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
Kauppinen J, Pelkonen J, Katila M-L (1994) RFLP analysis of Mycobacterium malmoense strains using ribosomal RNA gene probes: an addition tool to examine intraspecies variation. J Microbiol Methods 19:261–267
Kirk JL, Beaudette LA, Hart M, Moutoglis P, Klironomos JN, Lee H, Trevors JT (2004) Methods of studying soil microbial diversity. J Microbiol Methods 58:169–188
Lagesen K, Ussery DW, Wassenaar TM (2010) Genome update: the 1000th genome–a cautionary tale. Microbiology 156:603–608
Lagier J-C et al (2018) Culturing the human microbiota and culturomics. Nat Rev Microbiol:1
Lai Z et al (2012) Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression. Am J Bot 99:209–218
Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. Science 241:1077–1080
Langille MG et al (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814
Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299:682–686
Lieberman KR, Cherf GM, Doody MJ, Olasagasti F, Kolodji Y, Akeson M (2010) Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. J Am Chem Soc 132:17961–17972
Lingner T, Aßhauer KP, Schreiber F, Meinicke P (2011) CoMet—a web server for comparative functional profiling of metagenomes. Nucleic Acids Res 39:W518–W523
Liu W-T, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl Environ Microbiol 63:4516–4522
MacNaughton SJ, Stephen JR, Venosa AD, Davis GA, Chang Y-J, White DC (1999) Microbial population changes during bioremediation of an experimental oil spill. Appl Environ Microbiol 65:3566–3574
Markowitz VM et al (2011) IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122
Masciarelli O, Llanes A, Luna V (2014) A new PGPR co-inoculated with Bradyrhizobium japonicum enhances soybean nodulation. Microbiol Res 169:609–615
Matsen FA, Evans SN (2013) Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. Plos ONE 8:e56859

McKain MR et al (2012) Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). Am J Bot 99:397–406
McKernan K, Blanchard A, Kotler L, Costa G (2016) Reagents, methods, and libraries for bead-based sequencing. Google Patents
Mende DR et al (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. Plos one 7:e31386
Mendes R et al (2011) Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science:1203980
Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31
Meyer F et al (2008) The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinf 9:386
Mizrahi-Man O, Davenport ER, Gilad Y (2013) Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. Plos one 8:e53608
Moorthie S, Mattocks CJ, Wright CF (2011) Review of massively parallel DNA sequencing technologies. HUGO J 5:1–12
Morris CE, Monier J-M (2003) The ecological significance of biofilm formation by plant-associated bacteria. Annu Rev Phytopathol 41:429–453
Moyer CL, Dobbs FC, Karl DM (1994) Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. Appl Environ Microbiol 60:871–879
Moyer CL, Tiedje JM, Dobbs FC, Karl DM (1996) A computer-simulated restriction fragment length polymorphism analysis of bacterial small-subunit rRNA genes: efficacy of selected tetrameric restriction enzymes for studies of microbial diversity in nature. Appl Environ Microbiol 62:2501–2507
Muyzer G (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. Curr Opin Microbiol 2:317–322
Myllykangas S, Buenrostro J, Ji HP (2012) Overview of sequencing technology platforms. In: Bioinformatics for high throughput sequencing. Springer, Berlin, pp 11–25
Narisawa K, Hambleton S, Currah RS (2007) Heteroconium chaetospira, a dark septate root endophyte allied to the Herpotrichiellaceae (Chaetothyriales) obtained from some forest soil samples in Canada using bait plants. Mycoscience 48:274–281
Naum M, Lampel KA (2011) Analytical methods | DNA-based assays. In: Fuquay JW, ed. Encyclopedia of dairy sciences, 2nd edn. Academic Press, San Diego pp 221–225
NCBI (2018) National center for biotechnology information genome browser. https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt. Accessed 2018
Nielsen HB et al (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822
Niemi RM, Heiskanen I, Wallenius K, Lindström K (2001) Extraction and purification of DNA in rhizosphere soil samples for PCR-DGGE analysis of bacterial consortia. J Microbiol Methods 45:155–165
Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC bioinformatics 12:385
Orlando L et al (2011) True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res 21(10):1705–1719
Orlewska K, Piotrowska-Seget Z, Cycon M (2018) Use of the PCR-DGGE Method for the Analysis of the Bacterial Community Structure in Soil Treated With the Cephalosporin Antibiotic Cefuroxime and/or Inoculated With a Multidrug-Resistant Pseudomonas putida Strain MC1. Front Microbiol 9:1387
Osborn AM, Moore ER, Timmis KN (2000) An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environ Microbiol 2:39–50

Ozsolak F et al (2009) Direct RNA sequencing. Nature 461:814
Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52:413–435
Pop M (2009) Genome assembly reborn: recent computational challenges. Brief Bioinform 10:354–366
Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394
Reysenbach A-L, Giver LJ, Wickham GS, Pace NR (1992) Differential amplification of rRNA genes by polymerase chain reaction. Appl Environ Microbiol 58:3417–3418
Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. Science 281:363–365
Rothberg JM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348
Rudi K, Zimonja M, Trosvik P, Naes T (2007) Use of multivariate statistics for 16S rRNA gene analysis of microbial communities. Int J Food Microbiol 120:95–99
Sarethy IP, Pan S, Danquah MK (2014) Modern taxonomy for microbial diversity. In: Biodiversity-The dynamic balance of the planet. InTech Open, London
Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. Curr Opin Biotechnol 14:303–310
Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biol 5:e75
Shendure J et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732
Shulaev V et al (2011) The genome of woodland strawberry (Fragaria vesca). Nat Genet 43:109
Simon C, Daniel R (2011) Metagenomic analyses: past and future trends. Appl Environ Microbiol 77:1153–1161
Sørheim R, Torsvik VL, Goksøyr J (1989) Phenotypical divergences between populations of soil bacteria isolated on different media. Microb Ecol 17:181–192
Steele PR, Hertweck KL, Mayfield D, McKain MR, Leebens-Mack J, Pires JC (2012) Quality and quantity of data recovered from massively parallel sequencing: examples in Asparagales and Poaceae. Am J Bot 99:330–348
Strickler SR, Bombarely A, Mueller LA (2012) Designing a transcriptome next-generation sequencing project for a nonmodel plant species1. Am J Bot 99:257–266
Tabacchioni S, Chiarini L, Bevivino A, Cantale C, Dalmastri C (2000) Bias caused by using different isolation media for assessing the genetic diversity of a natural microbial population. Microb Ecol 40:169–176
Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637
Theron J, Cloete T (2000) Molecular techniques for determining microbial diversity and community structure in natural environments. Crit Rev Microbiol 26:37–57
Thompson JF, Steinmann KE (2010) Single molecule sequencing with a HeliScope genetic analysis system. Curr Protoc Mol Biol 92:7.10. 11–17.10. 14
Tiedje JM, Asuming-Brempong S, Nüsslein K, Marsh TL, Flynn SJ (1999) Opening the black box of soil microbial diversity. Appl Soil Ecol 13:109–122
Torsvik V, Sørheim R, Goksøyr J (1996) Total bacterial diversity in soil and sediment communities—a review. J Ind Microbiol 17:170–178
Torsvik V, Daae FL, Sandaa R-A, Øvreås L (1998) Novel techniques for analysing microbial diversity in natural and perturbed environments. J Biotechnol 64:53–62
Trevors J (1998) Bacterial biodiversity in soil with an emphasis on chemically-contaminated soils. Water Air Soil Pollut 101:45–67
Tyson GW et al (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37

Usuki F, Narisawa K (2007) A mutualistic symbiosis between a dark septate endophytic fungus, Heteroconium chaetospira, and a nonmycorrhizal plant, Chinese cabbage. Mycologia 99:175–184

Van der Heijden MG et al (1998) Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. Nature 396:69

von Wintzingerode F, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21:213–229

Wu M, Scott AJ (2012) Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033–1034

Yannarell AC, Triplett EW (2005) Geographic and environmental sources of variation in lake bacterial community composition. Appl Environ Microbiol 71:227–239

Zalapa JE et al (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99:193–208

Zavala-Gonzalez EA et al (2017) Arabidopsis thaliana root colonization by the nematophagous fungus Pochonia chlamydosporia is modulated by jasmonate signaling and leads to accelerated flowering and improved yield. New Phytol 213:351–364

Zeyaullah M et al (2009) Metagenomics-An advanced approach for noncultivable micro-organisms. Biotechnol Mol Biol Rev 4:49–54

Zhao L et al (2012) Characterization of microbial diversity and community in water flooding oil reservoirs in China. World J Microbiol 28:3039–3052

Molecular Biology Techniques for the Identification and Genotyping of Microorganisms

11

Nisarg Gohil, Happy Panchasara, Shreya Patel, and Vijai Singh

Abstract

The advent in molecular biology techniques has enabled, to a great extent, numerous identification and detection techniques of microorganisms by amplifying specific conserved DNA sequences. From time-to-time, the tremendous modifications have been employed in search of a rapid and inexpensive microbial identification method and now a whole genome to be sequenced is possible. In this chapter, we have described the different molecular identification techniques including 16S/18S rRNA, ITS and whole-genome sequencing as well as genotyping techniques such as pulse field gel electrophoresis, AFLP, RAPD, RFLP, ribotyping, BOX, ERIC, rep-PCR and multi-locus sequence typing with their basic principle, procedure, strengths and weaknesses.

11.1 Introduction

Microorganisms are the smallest living entity on the earth. These simplest and ubiquitous tiny microbes came before us and are potentially engaged in decomposition, fermentation, oxygen production, and generating vaccines, antibiotics, metabolites, chemicals, and other valuable products. They also play a crucial role

N. Gohil · H. Panchasara · S. Patel
School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

V. Singh (✉)
School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

Present address: Department of Biosciences, School of Sciences, Indrashil University, Kadi, Gujarat, India
e-mail: vijai.singh@indrashiluniversity.edu.in

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_11

in our body. Some of them are pathogenic and cause different diseases, while others are beneficial and aids in digestion, providing micronutrients, and helping to digest complex compounds into a simpler form which can be easily available to the host. In order to study and identify their crucial roles in health, agriculture, and food industries, their identification and characterization has become utmost importance.

Characterization and identification of microorganisms have become crucial for microbiologists ever since the recognition of the first bacterial isolate was obtained. The scientific study of the identification and classification of microorganisms is known as microbial systematics. This identification of microorganisms is delimited to biochemical and molecular techniques that involve profiling of its phenotype and genotype, respectively. The profiling of phenotypic characteristic is attributed to the organism's metabolic activities, whereas the genotypic profile is attributed by the genetic material (DNA). On the one hand, the biochemical assays are tedious, labor-intensive, and not always conclusive, which make them less practical. On the other hand, the molecular techniques are quite fast, reliable, sensitive, and reproducible. A pressing need has arisen to identify the microorganisms based on the molecular assays.

During the evolution, many of the genes are highly conserved that can be used for identification. Such highly conserved genes including 16S rRNA, 23S rRNA, internal transcribed spacer region (ITS), housekeeping genes (HKGs) (e.g., RNA polymerase, DNA gyrase), and virulence protein-encoding genes (e.g., hemolysin, lipase, protease) are widely used (Fox et al. 1992; Ashelford et al. 2005). Even sequencing of a whole-genome project has also gained more scientific and public attention because of the reduced price of the sequencing. Additionally, a wide range of bioinformatics tools have been devised that can align the multiple sequences and create a phylogenetic tree which shows the relationship among the various organisms at species and on a strain level (Ludwig et al. 2004). These tools perform a single or multi-locus sequencing wherein either one highly conserved gene (16S, 23S, 18S, or ITS) or a set of genes (gyrases or subunits of RNA polymerases) are sequenced, respectively.

For further discrimination of the microorganisms at the strain level, different genotyping or fingerprinting techniques have been developed wherein a distinct genetic profile of the microorganism can be generated via formation of chromosomal DNA fragments and their size-dependent separation. The fingerprint can be unique to the microorganisms and may show a similarity with its closely related strains (Singh et al. 2010a, 2012a, b). If the fingerprint of a test organism matches with any organism, then the two organisms are related at the species level. The fundamental application of sequencing and typing techniques is in systematics such as meticulous classification and nomenclature purposes. Apart from this, the methods are routinely used for several applications by applied microbiologists. The advances in the molecular techniques permitted more feasible industrial application of microorganisms (Donelli et al. 2013). Another significant role is in clinical microbiology. The identification of infectious agents and its pathogenicity also plays a pivotal role during an epidemic (Ecker et al. 2005; Schürch and Siezen 2010). In the present chapter, we highlight the techniques developed for sequencing are a breakthrough for microbial systematics, and some of these techniques for identification and genotyping of the microorganisms are discussed.

11.2 Molecular Biology Techniques for Identification of Microorganisms

The molecular techniques for identification of microbes are single gene sequencing (e.g., by sequencing ribosomal RNA encoding 16S, 23S, 18S, and ITS gene), multiple gene sequencing (housekeeping and pathogenic genes), and whole-genome sequencing.

11.2.1 The 16S rRNA and 23S rRNA Gene Sequencing

Ribosomes or ribonucleoproteins are responsible for synthesizing the proteins via translation process. It is highly conserved and found in each living organism. The prokaryotic ribosome consists of two subunits: a large subunit 50S (comprising of 23S rRNA and 5S rRNA) and a small subunit 30S (16S rRNA). The ribosomal operon is present in the genome as scattered form with varying the number. These sequences have been used for the molecular level identification and established the phylogenetic relationship among the microorganisms. Their ubiquity and functional and structural conservation make them applicable for bacterial systematics for the classification, nomenclature, and identification.

The sequence analysis of these genes is performed by amplifying the genes from genomic DNA by polymerase chain reaction (PCR) and then sequencing the amplicon. For the amplification of the ribosomal DNA, universal primers are available and can be used. It binds to the end regions of the gene to amplify it. The sequencing data is further aligned with the data stored in databases such as NCBI, EzTaxon-e (Kim et al. 2012), Ribosomal Database Project (Maidak et al. 1994), SILVA (Pruesse et al. 2007), and Greengenes (DeSantis et al. 2006). Similarly, 23S rRNA sequencing has been used for the identification of taxa of an isolate. However, 16S rRNA sequencing is more prevalent as the data for full-length sequences of 23S rRNA is less in number owing to its large size. Moreover, the inadequacy of the full-length sequence of 23S rRNA with respect to 16S rRNA causes difficulties in primer designing (Ludwig et al. 1992).

The 16S rRNA sequencing technique is highly useful and more reliable for the identification of bacterial isolates. However, it has some limitations including lower resolution at the species level and misinformation in databases (Fox et al. 1992; Ashelford et al. 2005). Consequently, the 16S rRNA, 23S rRNA, and ITS sequences can be routinely used for the molecular identification of microorganisms from diverse sources. Sometimes, these sequences are insufficient to identify microorganisms (Janda and Abbott 2007). Therefore, to identify the isolate at the species level, other biomarkers such as housekeeping and virulence protein-encoding genes can prove to be efficient.

11.2.2 Housekeeping Genes for Molecular Identification

Genes are small DNA sequences and have the ability to produce a protein under a control of a promoter. Some genes are essential for normal cell functioning and are highly conversed throughout the evolution process (i.e., HKGs). In order to refine the taxonomic identification at the species level and to acquire better resolution in phylogenetic relationships, an alternative approach of sequencing HKGs can be used that is quite reliable and simple. The HKGs such as DNA gyrase and RNA polymerase have been successfully used for the molecular identification and taxonomy of microorganisms.

One example of the HKG is *gyrB*. Organisms from the same genus consist of a set of HKGs, where the genomes show higher genetic variation with another genus as compared to 16S rRNA sequences. Therefore, by sequencing of HKGs, a presumptive isolate can be determined. Different strains can be characterized on the basis of the set of HKGs. Each strain contains a specific number of HKGs they possess. Overall, seven genes are sequenced for taxonomic purposes. In brief, the genes are PCR amplified and sequencing of the set of genes is performed and data are collected for analysis. A study shows that generating phylogenetic relationships by sequencing of housekeeping has an advantage of better resolution as compared to the 16S rRNA. Analysis of *gyrB* gene and 16S rRNA gene sequence of eight species of *Bacillus subtilis* was performed, and results showed 75.4–95.0% sequence similarities in the *gyrB* gene, whereas the 16S rRNA gene showed 98.1–99.8% (Wang et al. 2007).

A study manifests the importance of sequencing of HKGs for the identification of unknown isolates at the species level. Additionally, another HKG *rpoB* encoding β-subunit of RNA polymerase has been used for the identification of closely related species (Adékambi et al. 2009). These are simple, fast, accurate, and reliable molecular markers for identification and characterization. However, these molecular techniques fail to give clear explanation on the pathogenicity and disease conditions.

11.2.3 Detection of Microorganisms Using Virulence Factor-Encoding Genes

The genomes of pathogenic bacteria or infectious microorganisms consist of pathogenicity islands which are known to encode virulence factors and mobile genetic elements. It has high G + C content than the rest of the genome (acquired by horizontal gene transfer). A number of studies have reported the identification of pathogenic isolates by PCR amplification and sequencing of such genetic elements. Identification of methicillin-resistant *Staphylococcus aureus* isolates was performed by amplifying *mecA* gene. The identification based on such virulence factors also facilitates distinguishing of resistant and virulent *S. aureus* strains from susceptible strains, thereby demonstrating its application as a molecular diagnostic tool (Salisbury et al. 1997). A number of virulence proteins are available such as lipase (Singh et al. 2010a), hemolysin (Wang et al. 2003; Singh et al. 2007, 2008a, 2011),

and many more, which have been used for identification in a wide range of microorganisms.

11.2.4 Identification of Eukaryotic Organisms Using 18S rRNA and ITS Sequences

A fungus is responsible for many diseases in plants, animals, and human. The phylogenetic studies and molecular identification of a broad range of fungal species which form the second largest kingdom of eukaryotes is successfully done by sequence analysis of 18S rRNA gene and ITS region of the ribosome encoding operon of its genome. The 18S rRNA is a hypervariable sequence encoding small subunit rRNA in fungal genome. This highly repetitive region makes primer designing more feasible; hence, the identification and taxonomic procedure become ideal.

ITS region is an internally transcribed spacer region located between the sequences encoding small subunits, i.e., between 18S, 5S, and 28S rRNA. The ITS sequences are more variable than 18S rRNA in some manner; it is the more suitable biomarker for delineation at the species level. *Fusarium oxysporum* has been identified using the ITS region that was isolated from the different regions of India (Mishra et al. 2013a, b). Similar to the sequencing of other biomarker genes, this gene has been amplified and sequenced for identification. These techniques require universal primers for amplification of 18S rRNA gene (Borneman and Hartin 2000) and ITS region (Martin and Rygiewicz 2005). This technique is one of the better molecular markers for the identification of eukaryotic organisms in a simple, fast, sensitive, and specific manner. The DNA sequencing cost is decreasing gradually, which allows us to perform the whole-genome sequencing.

11.2.5 Whole-Genome Sequencing

In the 1970s and 1980s, the manual DNA sequencing methods such as Maxam Gilbert sequencing (Maxam and Gilbert 1977) and Sanger sequencing (Sanger et al. 1977) have been developed. Thereafter, the sequencing was shifted to a more rapid and automated method which made whole-genome sequencing (WGS) possible. Whole-genome sequencing is the technique that identifies the complete DNA sequence of an organism's genome. It is the most ideal identification technique among all the abovementioned techniques as they all have some advantage and disadvantage. WGS is quite expensive as compared to single gene sequencing, but it has more data that provides clearer understanding of the microorganisms. The two high-throughput DNA sequencing methods that have been used for whole-genome sequencing to date are as follows:

- Whole-genome shotgun sequencing (WGS) method.
- Next-generation sequencing (NGS) method.

11.2.5.1 Whole-Genome Shotgun Sequencing

Whole-genome shotgun sequencing (WGS) technique involves shredding of DNA into smaller overlapping fragments (of size ~2 kb) using restriction enzymes (an enzyme that identifies specific nucleotide sequences and cleaves the DNA molecule) and sequencing of the individual fragments of the DNA by cloning it into a plasmid vector to generate a library of clone. Randomly, some clones are selected from the library and sequenced from both ends to obtain the overlapping read. The sequences obtained are arranged into overlapping sections called contigs. While assembling the sequences into contigs, gaps (i.e., unidentified sequences) and single-stranded regions are identified, which are targeted again to produce the full-length sequence. The first organism to be sequenced using this method was *Haemophilus influenzae*. A tool named TIGR Assembler was used to assemble the sequence of the bacterium (Fleishchmann et al. 1995). The assembler performs pairwise comparisons of fragments, labels fragments as repeats or non-repeats, and finally constructs a genome by optimizing clone length and match criteria of repeats and non-repeat regions. It enables to overcome the hurdles in assembling of random sequence data obtained in the shotgun sequencing projects and allowing the assembling of data with more accuracy in reduced time (Sutton et al. 1995).

After assembling the sequences, bioinformatics analysis can be performed for annotation of the genome by different databases such as BLAST and other gene prediction programs (e.g., GLIMMER) using hidden Markov model (HMM) or interpolated Markov model (Delcher et al. 2007). This technique can be significantly used for demonstrating the phylogenetic diversity and analyzing genetic features (Ikeda et al. 2003; Venter et al. 2004). The technique is rapid and more cost-effective as compared to Sanger's sequencing method. It does not require any genome mapping resource. However, it produces more gaps and makes it not suitable for sequencing of organisms that have larger genomes with more repetitive sequences.

11.2.5.2 Next-Generation Sequencing

The next-generation sequencing (NGS) technology is also known as high-throughput sequencing. NGS is a breakthrough and potent alternative to all conventional methods because of its ability to provide high-throughput data. NGS can be utilized to perform the sequencing at the population scale in a single run at a faster rate and cheaper cost. Four varying platforms of next-generation sequencing technologies are available on the basis of template preparation and the principle of sequencing. The platforms for next-generation sequencing are the following:

1. 454/Roche technology
2. Illumina/Solexa technology.
3. Applied Biosystems (ABI)/SOLiD technology.
4. Helicos system.

11.2.5.2.1 The 454/Roche Technology

The 454 sequencing was the first platform of next-generation sequencing to be made commercially available. It was developed by 454 Life Sciences of Branford, CT, USA. It involves template preparation by emulsion PCR, and then the template DNA is fragmented by nebulization or sonication, and two different adapters are ligated to both ends of the fragments. One of the adapters is biotinylated (a process of covalently attaching biotin to a protein or other biomacromolecule) and permits the collection of single-stranded template DNA. The template is later diluted to a single-molecule concentration and immobilized to a bead. The template is then amplified by emulsion PCR (i.e., beads are encapsulated in droplets formed in a water-oil solution) so that many copies of the same template sequence are obtained in a single droplet. The emulsion is then disrupted, and beads having covalently attached clonally amplified template are enriched and again diluted to enable compartmentalization of individual bead per well in picotiter plate. After deposition of beads, sequencing is performed on the basis of pyrosequencing in which release of pyrophosphate group is detected by chemiluminescence. The sequencing reaction involves the synthesis of the complementary strand by incorporation of deoxyribonucleotide triphosphates (dNTPs) using polymerase.

When a dNTP is incorporated, a pyrophosphate group is released and used for the synthesis of ATP by ATP-sulfurylase. Further, detection is facilitated by luciferase enzyme that causes light to emit on the generation of ATP (Rothberg and Leamon 2008). The advantage of this technology over WGS is that it does not necessitate cloning in vivo, i.e., bacterial cloning is not required and can provide long read lengths. Unfortunately, it cannot accurately determine a stretch of homopolymer. Its feature of performing massively parallel sequencing in a faster rate makes it suitable for whole-genome sequencing.

11.2.5.2.2 Illumina/Solexa

Illumina, a short-read sequencing technology, reached into the market in 2006 and a year later the 454 sequencing. It works on the principle of sequencing-by-synthesis. The protocol involves the template preparation, cluster generation, sequencing, and data analysis. The template is prepared by fragmentation and addition of adapters to both ends. The adapters comprise of a sequence complementary to the oligonucleotides on the flow-cell anchors to consequently allow immobilization of template strand on a flow cell. In contrast to 454 sequencing, here the template is amplified by bridge amplification.

The template strand forms a bridge-like structure when multiple structures are formed (in a step called cluster generation), and in the flow cell, it looks like a bacterial colony; thus, the technique is also called polony sequencing. The clonal amplification is followed by sequencing using fluorescently labeled nucleotide. The sequence information is subsequently detected by its fluorescence as the nucleotide is fluorescent labeled. By the incorporation of these nucleotides, the label is cleaved to give fluorescence signal. Unlike 454 sequencing, the Illumina is used for de novo

assembly of the genomes as well as for re-sequencing of the genome of an organism whose closely related and previously sequenced reference genome is available (Farrer et al. 2009).

11.2.5.2.3 Applied Biosystems (ABI)/Supported Oligonucleotide Ligation and Detection (SOLiD) Technology

SOLiD (supported oligonucleotide ligation and detection) technology is also known as Applied Biosystems technology (as Applied Biosystems refined the technology and distributed the instrument commercially). It is a short-read sequencing technology that works on the principle of ligation. Similar to other technologies, the process incorporates template preparation, amplification, and sequencing. The template preparation and amplification (emulsion PCR) is identical to that of 454 sequencing technology. After the amplification of template, beads are enriched and immobilized to a glass slide for sequencing.

The sequencing occurs on the basis of ligation chemistry. For ligation reaction to commence, primer, octamer probes tagged with a dye and ligase to adjoin the probe to the primer are required. From the eight bases of the probe, the first two represent any 1 combination of 16 possible combinations that can be formed by the combination of any two bases out of four. While the other six bases are universal that possess complementarily to any bases, three out of the six bases have a fluorescent label attached to the 5′ end. When the probe ligates to the primer, the last three fluorescent-labeled bases are cleaved; hence, the attached probe has only five bases. The same process is repeated several times until the whole fragment is sequenced and every time the primer offsets by one base. The technique has an issue when palindromic sequences are present in the template, as they result in the formation of a hairpin structure. Currently, this technique is widely used for identification of microorganisms (Gulig et al. 2010).

11.2.5.2.4 Helicos

Helicos is the first single DNA molecule sequencing platform. The platform is useful for the identification of unique bacterial strains by comparing and analyzing the closely related reference genomes. It differs from other platforms in a way that the amplification step is eliminated. In this technique, DNA is fragmented and polyadenylated at 3′ end. The polyadenylated strands are denatured and hybridized to poly(dT) oligonucleotides present on the flow-cell surface. The positions of the immobilized strand in the flow cell are recorded, and then each of the four Cy5-labeled dNTPs is added individually to produce fluorescent. The fluorescence is imaged for further data analysis (Steinmann et al. 2011). It offers non-bias representation of templates and used for sequencing-based methods. However, it produces high error rates as compared to Illumina sequencing technology (Metzker 2010).

11.3 Molecular Biology Techniques for Genotyping of Microorganisms

The basic principle underlying all the genotypic techniques is collection, amplification, and separation of organism's DNA. However, the amplification step is optional (i.e., some techniques do not involve the amplification of DNA). Hence, on the basis of this step of amplification, genotyping techniques can be categorized as amplifying (e.g., RAPD, RFLP, AFLP, rep-PCR, ERIC, BOX, and MLST) and non-amplifying DNA techniques (e.g., PFGE and ribotyping). Moreover, it can be classified as restriction enzyme-based (e.g., PFGE, RFLP, AFLP, ribotyping), DNA sequencing-based (MLST, 16S rRNA, 18S rRNA, ITS, and WGS), and DNA hybridization-based (ribotyping, RFLP) methods. A summary of steps involved in each molecular method is briefly shown in Fig. 11.1. Major features of the individual molecular methods are given in Table 11.1.

11.3.1 Pulsed-Field Gel Electrophoresis (PFGE)

The development of a gel electrophoresis technique to separate the different charged particles with the use of an electric field has been one of the most remarkable molecular biological achievements to date. Since its origin, the technique has been used far apart to separate macromolecules such as DNA, RNA, and proteins in a

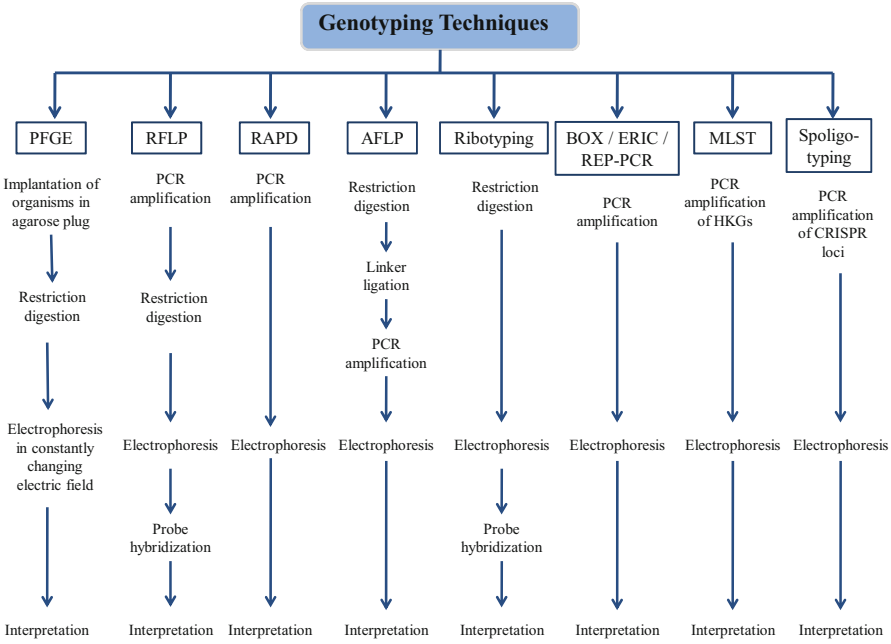

Fig. 11.1 Schematic representation of steps involved in different molecular genotyping methods

Table 11.1 Summary of molecular biology techniques used for genotyping of microorganisms

Technique	Advantages	Disadvantages	References
PFGE	High discrimination power and reproducibility	Tedious, laborious, and expensive	Sharma-Kuinkel et al. (2014)
RFLP	High reproducibility and reliability, cost-effective	Tedious and laborious	Yang et al. (2013) and Adzitey et al. (2013)
RAPD	Rapid, inexpensive, and effortless, prior knowledge of sequence not needed	Low reproducibility	Vila et al. (1996) and Ashayeri-panah et al. (2012)
AFLP	Considerable reproducibility, high resolution, and sensitivity	Laborious	Mueller and Wolfenbarger (1999)
Ribotyping	Considerable reproducibility and discrimination power, automation available	Tedious, expensive if involves automation	Bouchet et al. (2008)
Rep-PCR	Simple, rapid, and cost-effective	Low reproducibility, procedure standardization, and interpretation needed	Wu and Della-Latta (2002)
ERIC	Simple, rapid, and cost-effective	Low reproducibility	Adzitey et al. (2013) and Bilung et al. (2018)
BOX	Simple, rapid, and cost-effective	Less reproducibility	Wu and Della-Latta, (2002) and Bilung et al. (2018)
MLST	Highly standardized, easy to compare the results among different laboratories, high reproducibility	Tedious, difficult to detect organisms who share the same lineages	Larsen et al. (2012) and Sullivan et al. (2005)
Spoligotyping	Considerable discrimination power, rapid, and cost-effective	Ineffective in organisms that do not possess CRISPR array	Shariat and Dudley (2014)

Abbreviations: *PFGE* pulsed-field gel electrophoresis, *RFLP* restriction fragment length polymorphism, *RAPD* random amplified polymorphic DNA, *AFLP* amplified fragment length polymorphism, *Rep* repetitive extragenic palindromic, *ERIC* enterobacterial repetitive intergenic consensus, *MLST* multilocus sequence typing, *CRISPR*, clustered regularly interspaced short palindromic repeats

size-dependent manner. It is an extremely straightforward technique that employs the constant electric field across the agarose gel and enables the macromolecule to pass through the pores of the gel according to its size or mass. Electric current forces negatively charged DNA to move forward through the agarose gel matrix. Smaller molecules can pass easily through the average-sized pores and move faster than the larger ones. As a result, the molecules are separated in a size-dependent manner and

make distinct bands. However, the technique can only separate the restriction fragments up to ~20 kb (Birren and Lai 1993; Maule 1998).

To circumvent this obstacle, in 1984, Schwartz and Cantor (Schwartz and Cantor 1984) have designed and revealed the pulsed-field gel electrophoresis (PFGE) that have high discriminate potential of up to ~10 Mb (Herschleb et al. 2007; Nassonova 2008) by applying the periodic changes in the direction of electric field. PFGE in microbiology is a "gold standard" of molecular genotyping methods. The method incorporates the isolation of complete chromosomal DNA by bacterial cells lysis, and it is embedded in an agarose plug to protect from the mechanical shearing of the DNA. Subsequently, the DNA is subjected to cleave by the unique restriction enzyme(s). As a consequence, differently sized DNA fragments are formed. These resultant DNA fragments are then used for fractionation depending upon the sizes by constantly changing the direction of the electric field that facilitates DNA fragments to reorient and move through the sieves of gel (Schwartz and Cantor 1984). The DNA band patterns were obtained after electrophoresis run that can be analyzed and interpreted depending upon the number of band differences as follows:

1. Same pattern – "indistinguishable".
2. 1–3 band differences – "closely related"
3. 4–6 band differences – "possibly related"
4. $\geq$6 band differences – "unrelated" (Wu and Della-Latta 2006).

The establishment of the DNA pattern depends upon the concentration of the gel, buffer, voltage, pulse angle, switch time, temperature, runtime, and the directions of an electric field (Wu and Della-Latta 2006; Nassonova 2008; Chen and Ugaz 2008). Principally, the PFGE runs can take a couple of days (Ribot et al. 2001; Heng et al. 2009). In this light, it is obligatory to maintain the volume of the buffer and temperature. Altogether, PFGE has high discrimination potential as compared to conventional agarose gel electrophoresis technique. It has been widely used for typing of numerous bacterial pathogens for genotyping and epidemiological studies (Gosiewski and Brzychczy-Wloch 2015; Parizad et al. 2016) and for yeast (Schwartz and Cantor 1984; Maringele and Lydall 2006) and fungi (Wu et al. 1996; Lukácsi et al. 2006).

11.3.2 Restriction Fragment Length Polymorphism (RFLP)

Every living organism has differences in the number and location of the restriction sites into their genome. When the genomic DNA of organisms is digested with the restriction enzymes, differently sized DNA fragments will be formed. Those DNA extracts can be used for studying the polymorphism (a genetic variation occurs at a particular locus in the DNA). The restriction fragment length polymorphism (RFLP) uses the same principle. It is the first conventional, restriction enzyme analysis-based technique (Moser and Lee 1994; Ripamonti et al. 2009; Yang et al. 2013) used chiefly for exploiting variations in the types of individuals among the single species.

Apart from that, it has been also used for forensic science (Budowle et al. 1994; Góes et al. 2002; Panneerchelvam and Norazmi 2003; Roewer 2013), identification of gene location (Drayna et al. 1984; Singh et al. 2008b), diagnosis of hereditary disease (Bolhuis et al. 1987; Cooper and Schmidtke 1993), and parental tests (Smouse and Chakraborty 1986; Moser and Lee 1994; Dubey et al. 2010). After the advent of PCR, the RFLP was upgraded into RFLP-PCR for improving the resolution even with the very low concentration of DNA. In RFLP-PCR, ITS and intergenic spacer (IGS) of the ribosomal DNA are amplified by the PCR and subsequently digested by the RE for banding patterns (Mishra et al. 2013a, b). The generated unique banding pattern has been used predominantly for constructing the dendrogram in order to find a relationship among a wide range of the microorganisms.

11.3.3 Random Amplified Polymorphic DNA (RAPD)

Random amplified polymorphic DNA (RAPD), sometimes also called arbitrary primed PCR (AP-PCR), is a PCR-implemented DNA fingerprinting technique brought into light by Williams et al. (1990) and Welsh and McClelland (1990) independently for the amplification of discrete DNA fragments using random decamer primer(s). It has been widely used to study the polymorphism. RAPD markers are random, and it does not require the knowledge of pre-DNA sequences of the targeted organisms. In the PCR reaction, the DNA product formation depends upon the presence of priming sites within an amplifiable distance of each other. Due to variation in the annealing sites, differently sized amplified DNA fragments are formed that are subjected to agarose gel electrophoresis and the polymorphisms of different organisms that can be visualized and compared in a form of bands.

As the discovery of this technique, it has been employed in numerous fields such as genetic mapping (Krutovskii et al. 1998; Wang et al. 1999), developing genetic markers linked to a trait (Moon et al. 2016), population (Rocco et al. 2007; Ma et al. 2012) and evolutionary genetics (Stêpniak et al. 2002; Han et al. 2017), as well as plant (Debener et al. 1996; Morales et al. 2011; Guasmi et al. 2012; Nadeem et al. 2018) and animal breeding (Ali et al. 2004). As the technique depends on the PCR reaction, the conditions such as annealing temperature, the concentration of PCR components, as well as the template DNA influence the polymorphism. The technique is facing the disadvantage of low reproducibility.

11.3.4 Amplified Fragment Length Polymorphism (AFLP)

Amplified fragment length polymorphism (AFLP) is a robust, improved, multiple loci targeted DNA fingerprinting technique that was developed by Zabeau and Vos (1993). It applies principles of RFLP, RAPD, RE-based digestion, PCR, and agarose gel electrophoresis. It uses short synthetic DNA adapter molecules having a core sequence that can be used for primer design and enzyme-specific DNA sequences.

The technique starts with the isolation of genomic DNA and its digestion using two different restriction enzymes: a frequent cutter (e.g., MseI, NotI) and rare cutter (e.g., EcoRI). The frequent cutter restriction enzyme generates numerous small fragments, while the rare cutter limits the number of fragments. The adaptors are then incorporated and ligated with the generated DNA fragments. Selective PCR amplification of subset adapter molecules will be done using the primers. Subsequently, the PCR-amplified fragments are visualized by using gel electrophoresis (Mueller and Wolfenbarger 1999; Meudt and Clarke 2007; Lall et al. 2010). Apart from genotyping, the technique can be also applied for criminal investigation (Savelkoul et al. 1999) and paternity test (Savelkoul et al. 1999; Arenas et al. 2017). AFPL has been widely used for discrimination of microbial strains and also used for genotyping of agriculturally important microorganisms for the improvement of crop production and quality.

11.3.5 Ribotyping

Similar to RFLP and AFLP, ribotyping is a restriction enzyme digestion-based molecular technique that deals with the unique and conserved ribosomal genes. In other words, ribotyping is a fusion of RFLP and southern blotting techniques, and the developed bands represent number and sequence diversity of rRNA genes to the strain level (Olive and Bean 1999; Wu and Della-Latta 2002). In the typing method, the genomic DNA of the cells are isolated and then digested by using restriction enzymes. These fragments are used afterward for agarose gel electrophoresis and southern blotting on a nylon or nitrocellulose membrane. Subsequently, it is hybridized with labeled DNA probes that have complementarily to the rRNA sequences, and the generated pattern can be analyzed with the reference organisms. The technique is labor-intensive, time-consuming, and expensive. The updated automated ribotyping technique resolved these limitations with systems such as RipoPrinter® microbial characterization system (Dupont Qualicon Inc., Wilmington, DE). These systems provide a high-throughput, rapid, accurate, reliable, and standardized ribotype patterns that make it easy (Okatani et al. 2004; Pavlic and Griffiths 2009).

11.3.6 Use of Repetitive Sequences for Genotyping

All the organisms have repetitive sequences in their genome for genome stability, function, proper running, and processing of cellular machinery. These are generally noncoding DNA and conserved throughout evolution. Currently, these DNA sequences are used to make the distinct fingerprint profiles by amplifying the region using primers. These repetitive DNA sequences have been used for discrimination of microorganism at strain level and typing in order to study and evaluate the microbial population lineages and outbreaks. The molecular techniques such as REP-PCR, ERIC-PCR, and BOX-PCR that use the same principle are given as follows.

11.3.6.1 Repetitive Extragenic Palindromic (REP)

Repetitive extragenic palindromic (REP) elements are the first and most studied palindromic repeated sequences that exist throughout the bacterial genome in varying sizes (Versalovic et al. 1998). It was initially identified in *Escherichia coli* and *Salmonella typhimurium* (Stern et al. 1984; Dimri et al. 1992) and subsequently found throughout the bacterial system (Lupski and Weinstock 1992). In the early 1990s, Versalovic and colleagues have developed a PCR-based DNA profiling technique for yielding the differently sized REP elements (Versalovic et al. 1991; Lupski and Weinstock 1992). In this method, the unique primers were designed in a manner that has complementarity to oligonucleotide REP sequences and used to amplify REP sequences from the target genomic DNA. Amplicons are then separated using agarose gel electrophoresis and the generated DNA profiles can be used to discriminate the bacterial strains from different sources. This technique is not limited to bacteria genotyping; it is well functioning in fungal species. It has been efficiently used for typing of plant pathogenic fungal species *Botryosphaeriaceae* (Abdollahzadeh and Zolfaghari 2014). The technique is far apart used to generate species and strain-specific DNA fingerprints (Versalovic et al. 1991) to produce a rapid, reliable, and reproducible DNA profiles.

11.3.6.2 Enterobacterial Repetitive Intergenic Consensus (ERIC)

Enterobacterial repetitive intergenic consensus (ERIC) is 126-bp-long and highly conserved repetitive sequences restricted in intergenic regions (a subset of noncoding DNA) of the bacterial genomes. The first ERIC elements were observed in *E. coli*, *S. typhimurium*, other *Enterobacteriaceae* members, and *Vibrio cholerae* (Hulton et al. 1991). Subsequently, it has been also found in *Corynebacterium pseudotuberculosis* (de Sá Guimarães et al. 2011; Dorneles et al. 2014), *Pseudomonas aeruginosa* (Han et al. 2014; Khosravi et al. 2016; Auda et al. 2017), *Acinetobacter baumannii* (Presterl et al. 1997; Aljindan et al. 2018), *Sinorhizobium meliloti* (Niemann et al. 1999; Elboutahiri et al. 2010), *Haemophilus parasuis* (Olvera et al. 2006; Macedo et al. 2011), *Aeromonas hydrophila* (Szczuka and Kaznowski 2004), *Staphylococcus aureus* (Ye et al. 2012; Abdollahi et al. 2014), *Fusarium oxysporum* fungus (Mishra et al. 2015), and many more.

The number of copies and the locations of the ERIC sequences greatly vary among the species. By using primers complementary to ERIC elements, the distinctive banding pattern has been obtained (Wilson and Sharp 2006). Compared to other genotyping methods, it is fast, cheap, and has greater discrimination potential (Sampaio et al. 2006; Kosek et al. 2012). Interestingly, the ERIC typing has an upper hand over the PFGE in that it can show species-specific band (Stephenson et al. 2009).

11.3.6.3 BOX Element

Alike REP and ERIC, the BOX elements are a group of highly conserved, repetitive DNA sequences located within the intergenic regions. It is for the first time found in the chromosome of gram-positive *Streptococcus pneumoniae* (Martin et al. 1992). These repeated sequences are composed of three discriminate regions (boxA, boxB,

and boxC) with a molecular length of 59, 45, and 50 nucleotides, respectively. The BOX-PCR has been applied to a great extent and employed to many bacterial genera such as *Aeromonas* (Tacão et al. 2005; Singh et al. 2010b), *Pseudomonas* (Marques et al. 2008), *Bacillus* (Kumar et al. 2014), *Streptomyces* spp. (Lanoot et al. 2004), *Geobacillus* (Meintanis et al. 2008), and *Fusarium oxysporum* fungus (Mishra et al. 2015) in different combinations of three BOX regions (van Belkum and Hermans 2001).

Based on these sequences, the BOX-PCR DNA fingerprinting method has been developed, in which the BOX elements are amplified using a single primer and run agarose gel electrophoresis for generating a DNA pattern (Marques et al. 2008; Singh et al. 2010b). This band could be used for generating the relationship among microorganisms. This conventional method faces the limitation of poor band resolution that can be resolved by using fluorescence labeling (Versalovic et al. 1995). A number of agriculturally important pathogens have been identified and discriminated using BOX-PCR technique.

11.3.7 Multilocus Sequence Typing (MLST)

An ideal genotyping scheme should be rapid, feasible, cheaper, stable, reproducible, easy to implement, and widely applicable to a wide range of organisms. It should allow global comparisons of typing results from different research laboratories. However, the traditional typing methods like RAPD, REP, and ERIC face poor reproducibility (Ashayeri-panah et al. 2012; Wu and Della-Latta 2002). In 1998, Maiden et al. (1998) have invented a bacterial typing method named multilocus sequence typing (MLST) to overcome the hurdle of reproducibility among the laboratories. MLST is an unambiguous, universal, definitive, and highly discriminatory technique for characterizing the bacterial and fungal species (Taylor and Fisher 2003) based on nucleotide sequences of the internal conserved regions, typically 4 to 10 HKGs for each test organism (Pavón and Maiden 2009; MLST allelic profiles and sequences 2018).

Originally, the technique was developed for human pathogen *Neisseria meningitis* (Maiden et al. 1998); subsequently, it has been expanded for more than 125 organisms to date (MLST allelic profiles and sequences 2018). Approximately 450–500 bp internal nucleotide sequences per gene are used. The nucleotide variation rate in HKGs is relatively very slow. Even the variation is still considered to identify and genotype the organism, making it ideal to use in this emerging gold-standard typing method. The technique quantifies the unique sequences (alleles) in HKGs and provides unique integer number for each allele. In this manner, alleles at each HKG will get defined the allelic profile, and the combinations of that make distinct sequence type (Larsen et al. 2012).

As most of the organisms have enough differences, it is possible to create billions of unique allelic profiles. The number and type of HKGs are used by MLST that may vary from species to species. It can be obtained from a universal MLST databases. The steps of MLST involves the genomic DNA isolation, PCR amplification of the

HKGs using loci-specific primers, DNA sequencing of amplified HKGs, and MLST analysis (Maiden et al. 1998). The universal primer sequences for a particular organism can be also obtained from the universal MLST databases. The MLST database gets updated on newly found alleles and sequence types after some necessary verification (Aanensen and Spratt 2005). This technique can be further explored in agriculturally important microorganisms in order to properly manage and control for improving crop productivity and yields.

11.3.8 Spoligotyping

Clustered regularly interspaced short palindromic repeats (CRISPR) are DNA repeats present in the broad spectrum of bacteria and almost all archaea as an adaptive immune system against phage infection, viruses, and plasmids (Jinek et al. 2012, 2013; Mali et al. 2013; Singh et al. 2017). Although it is a bacterial defense system, it has been developed and extensively used as a form of technique named CRISPR-cas9 for genome editing in a wide range of organisms for biomedical (Singh et al. 2018), antimicrobial (Bikard et al. 2014; Bikard and Barrangou 2017; Pursey et al. 2018), metabolic engineering (Cho et al. 2018; Behler et al. 2018; Lian et al. 2018; Gohil et al. 2017), improving the crop (Upadhyay et al. 2013), and other biological purposes. Apart from that, the dynamic nature of the CRISPR has also unwrapped new avenues in detecting diseases (Gootenberg et al. 2017; Khambhati et al. 2018) and for genotyping (Shariat and Dudley 2014).

In spacer-oligonucleotide typing (spoligotyping), the CRISPR repeats are PCR amplified with labeled primers and then hybridized to probes using nylon membrane. Hence, different hybridization patterns can be obtained (Shariat and Dudley 2014). For enabling higher throughput, microbeads can be used instead of using membrane (Cowan et al. 2004). Owing to the fact that nearly half of the bacterial population contains CRISPR-cas9 system (Barrangou and Marraffini 2014), the spoligotyping is possible only in such bacteria as well as archaea. Still, the technique has been used significantly, especially for human pathogens such as *Mycobacterium tuberculosis* (Sola et al. 2015), *Salmonella enterica* (Sola et al. 2015), and the plant pathogen *Erwinia amylovora* (McGhee and Sundin 2012). Advances in CRISPR technique allow us to explore more to improve production and yield of the crop for sustainable agriculture.

11.4 Conclusion and Future Remarks

From the time when the microorganisms were differentiated according to their phenotypic characteristics to the techniques that have been used to date for the identification and characterization of the microorganisms, it has required tremendous improvements and changes to develop the techniques to be more rapid, reliable, convenient, cheaper, and widely acceptable to the world. Such everlasting progresses have made medical, plant disease diagnostics more accurate by faster

identification of the pathogens, easy determination of the contamination in the food and water bodies, population lineage study facile, evaluation of disease outbreak convenience, and so on. These molecular techniques will continue to provide the prime role in the better use of the microorganisms. Nowadays, the advanced technology such as WGS has opened up new avenues in microbial systematics and has become a routine tool in microbiology. Even though the cost of WGS continues to decline, it remains highly expensive as compared to other techniques. In the near future, we predict that the WGS will emerge as a gold standard for systematics of microbes and provide results in a real-time fashion. These molecular techniques have the potential to identify and discriminate plant pathogens in order to control the disease for improvement of crop productivity and yields.

Acknowledgments This work was supported by Puri Foundation for Education in India.

References

Aanensen DM, Spratt BG (2005) The multilocus sequence typing network: mlst. net. Nucleic Acids Res 33:W728–W733

Abdollahi S, Ramazanzadeh R, Kalantar E, Zamani S (2014) Molecular epidemiology of *Staphylococcus aureus* with ERIC-PCR method. Bull Env Pharmacol Life Sci 3:158–165

Abdollahzadeh J, Zolfaghari S (2014) Efficiency of rep-PCR fingerprinting as a useful technique for molecular typing of plant pathogenic fungal species: *Botryosphaeriaceae* species as a case study. FEMS Microbiol Lett 361(2):144–157

Adékambi T, Drancourt M, Raoult D (2009) The *rpoB* gene as a tool for clinical microbiologists. Trends Microbiol 17(1):37–45

Adzitey F, Huda N, Ali GRR (2013) Molecular techniques for detecting and typing of bacteria, advantages and application to foodborne pathogens isolated from ducks. 3. Biotech 3(2):97–107

Ali BA, Huang TH, Qin DN, Wang XM (2004) A review of random amplified polymorphic DNA (RAPD) markers in fish research. Rev Fish Biol Fish 14(4):443–453

Aljindan R, Alsamman K, Elhadi N (2018) ERIC-PCR genotyping of *Acinetobacter baumannii* isolated from different clinical specimens. Saudi J Med Med Sci 6(1):13–17

Arenas M, Pereira F, Oliveira M, Pinto N, Lopes AM, Gomes V, Carracedo A, Amorim A (2017) Forensic genetics and genomics: much more than just a human affair. PLoS Genet 13(9): e1006960

Ashayeri-panah M, Eftekhar F, Feizabadi MM (2012) Development of an optimized random amplified polymorphic DNA protocol for fingerprinting of *Klebsiella pneumonia*. Lett Appl Microbiol 54(4):272–279

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Env Microbiol 71(12):7724–7736

Auda IG, Al-Kadmy I, Kareem SM, Lafta AK, A'Affus MHO, Khit IAA, Kheraif A, Abdullah A, Divakar DD, Ramakrishnaiah R (2017) RAPD-and ERIC-based typing of clinical and environmental *Pseudomonas aeruginosa* isolates. J AOAC Int 100(2):532–536

Barrangou R, Marraffini LA (2014) CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. Mol Cell 54(2):234–244

Behler J, Vijay D, Hess WR, Akhtar MK (2018) CRISPR-based technologies for metabolic engineering in cyanobacteria. Trends Biotechnol 36(10):996–1010

Bikard D, Barrangou R (2017) Using CRISPR-Cas systems as antimicrobials. Curr Opin Microbiol 37:155–160

Bikard D, Euler CW, Jiang W, Nussenzweig PM, Goldberg GW, Duportet X, Fischetti VA, Marraffini LA (2014) Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol 32(11):1146–1150

Bilung LM, Pui CF, Su'ut L, Apun K (2018) Evaluation of BOX-PCR and ERIC-PCR as molecular typing tools for pathogenic *Leptospira*. Dis Markers 1351634

Birren BW, Lai E (1993) Pulsed field gel electrophoresis: a practical guide. Academic Press Inc, San Diego

Bolhuis PA, Defesche JC, Van der Helm HJ (1987) Differential diagnosis of genetic disease by DNA restriction fragment length polymorphisms. Clin Chim Acta 165(2–3):271–276

Borneman J, Hartin RJ (2000) PCR primers that amplify fungal rRNA genes from environmental samples. Appl Environ Microbiol 66(10):4356–4360

Bouchet V, Huot H, Goldstein R (2008) Molecular genetic basis of ribotyping. Clin Microbiol Rev 21(2):262–273

Budowle B, Sajantila A, Hochmeister MN, Comey CT (1994) The application of PCR to forensic science. In: Mullis KB, Ferré F, Gibbs RA (eds) The polymerase chain reaction. Birkhäuser, Boston, pp 244–256

Chen X, Ugaz VM (2008) Investigating DNA migration in pulsed fields using a miniaturized FIGE system. Electrophoresis 29(23):4761–4767

Cho S, Shin J, Cho BK (2018) Applications of CRISPR/Cas system to bacterial metabolic engineering. Int J Mol Sci 19(4):1089

Cooper DN, Schmidtke J (1993) Diagnosis of human genetic disease using recombinant DNA. Hum Genet 92(3):211–236

Cowan LS, Diem L, Brake MC, Crawford JT (2004) Transfer of a Mycobacterium tuberculosis genotyping method, spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. J Clin Microbiol 42(1):474–477

De Sá Guimarães A, Dorneles EMS, Andrade GI, Lage AP, Miyoshi A, Azevedo V, Gouveia AMG, Heinemann MB (2011) Molecular characterization of *Corynebacterium pseudotuberculosis* isolates using ERIC-PCR. Vet Microbiol 153(3–4):299–306

Debener T, Bartels C, Mattiesch L (1996) RAPD analysis of genetic variation between a group of rose cultivars and selected wild rose species. Mol Breed 2(4):321–327

Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23(6):673–679

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72(7):5069–5072

Dimri GP, Rudd KE, Morgan MK, Bayat H, Ames GF (1992) Physical mapping of repetitive extragenic palindromic sequences in *Escherichia coli* and phylogenetic distribution among *Escherichia coli* strains and other enteric bacteria. J Bacteriol 174(14):4583–4593

Donelli G, Vuotto C, Mastromarino P (2013) Phenotyping and genotyping are both essential to identify and classify a probiotic microorganism. Microb Ecol Health Dis 24(1):20105

Dorneles EM, Santana JA, Ribeiro D, Dorella FA, Guimarães AS, Moawad MS, Selim SA, Garaldi ALM, Miyoshi A, Ribeiro MG, Gouveia AM (2014) Evaluation of ERIC-PCR as genotyping method for *Corynebacterium pseudotuberculosis* isolates. PLoS One 9(6):e98758

Drayna D, Davies K, Hartley D, Mandel JL, Camerino G, Williamson R, White R (1984) Genetic mapping of the human X chromosome by using restriction fragment length polymorphisms. Proc Natl Acad Sci U S A 81(9):2836–2839

Dubey AK, Hussain N, Mittal N (2010) HindIII-based restriction fragment length polymorphism in hemophilic and nonhemophilic patients. J Nat Sci Biol Med 1(1):25–28

Ecker DJ, Sampath R, Blyn LB, Eshoo MW, Ivy C, Ecker JA, Libby B, Samant V, Sannes-Lowery KA, Melton RE, Russell K, Freed N, Barrozo C, Wu J, Rudnick K, Desai A, Moradi E, Knize DJ, Robbins DW, Hannis JC, Harrell PM, Massire C, Hall TA, Jiang Y, Ranken R, Drader JJ, White N, McNeil JA, Crooke ST, Hofstadler SA (2005) Rapid identification and strain-typing of respiratory pathogens for epidemic surveillance. Proc Natl Acad Sci U S A 102(22):8012–8017

Elboutahiri N, Thami-Alami I, Udupa SM (2010) Phenotypic and genetic diversity in *Sinorhizobium meliloti* and *S. medicae* from drought and salt affected regions of Morocco. BMC Microbiol 10(1):15

Farrer RA, Kemen E, Jones JD, Studholme DJ (2009) *De novo* assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads. FEMS Microbiol Lett 291(1):103–111

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223):496–512

Fox GE, Wisotzkey JD, Jurtshuk P Jr (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Evol Microbiol 42(1):166–170

Góes ACDS, Silva DAD, Domingues CS, Marreiro Sobrinho J, Carvalho EFD (2002) Identification of a criminal by DNA typing in a rape case in Rio de Janeiro, Brazil. Sao Paulo Med J 120 (3):77–79

Gohil N, Panchasara H, Patel S, Ramírez-García R, Singh V (2017) Book review: Recent advances in yeast metabolic engineering. Front Bioeng Biotechnol 5:71

Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia N, Daringer NM, Freije CA, Myhrvold C (2017) Nucleic acid detection with CRISPR-Cas13a/C2c2. Science 356(6336):438–442

Gosiewski T, Brzychczy-Wloch M (2015) The use of PFGE method in genotyping of selected bacteria species of the *Lactobacillus* genus. Methods Mol Bio 1301:225–240

Guasmi F, Elfalleh W, Hannachi H, Fères K, Touil L, Marzougui N, Triki T, Ferchichi A (2012) The use of ISSR and RAPD markers for genetic diversity among south tunisian barley. ISRN Agron 952196

Gulig PA, de Crécy-Lagard V, Wright AC, Walts B, Telonis-Scott M, McIntyre LM (2010) SOLiD sequencing of four *Vibrio vulnificus* genomes enables comparative genomic analysis and identification of candidate clade-specific virulence genes. BMC Genomics 11:512

Han MM, Mu LZ, Liu XP, Zhao J, Liu XF, Liu H (2014) ERIC-PCR genotyping of *Pseudomonas aeruginosa* isolates from *haemorrhagic pneumonia* cases in mink. Vet Rec Open 1(1):e000043

Han Y, Liu Y, Wang H, Liu X (2017) The evolution of *Vicia ramuliflora* (Fabaceae) at tetraploid and diploid levels revealed with FISH and RAPD. PLoS One 12(1):e0170695

Heng SK, Heng CK, Puthucheary SD (2009) Stacking gels: a method for maximising output for pulsed-field gel electrophoresis. Indian J Med Microbiol 27(2):142–145

Herschleb J, Ananiev G, Schwartz DC (2007) Pulsed-field gel electrophoresis. Nat Protoc 2 (3):677–684

Hulton CSJ, Higgins CF, Sharp PM (1991) ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. Mol Microbiol 5(4):825–834

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Ōmura S (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat Biotechnol 21(5):526–531

Janda JM, Abbott SL (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45(9):2761–2764

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 337 (6096):816–821

Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J (2013) RNA-programmed genome editing in human cells. elife 2:e00471

Khambhati K, Bhattacharjee G, Singh V (2018) Current progress in CRISPR-based diagnostic platforms. J Cell Biochem 120:2721. https://doi.org/10.1002/jcb.27690

Khosravi AD, Hoveizavi H, Mohammadian A, Farahani A, Jenabi A (2016) Genotyping of multidrug-resistant strains of *Pseudomonas aeruginosa* isolated from burn and wound infections by ERIC-PCR. Acta Cir Bras 31(3):206–211

Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, Na H, Park SC, Jeon YS, Lee JH, Yi H, Won S (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. Int J Syst Evol Microbiol 62(3):716–721

Kosek M, Yori PP, Gilman RH, Vela H, Olortegui MP, Chavez CB, Calderon M, Bao JP, Hall E, Maves R, Burga R (2012) Facilitated molecular typing of *Shigella* isolates using ERIC-PCR. Am J Trop Med Hyg 86(6):1018–1025

Krutovskii KV, Vollmer SS, Sorensen FC, Adams WT, Knapp SJ, Strauss SH (1998) RAPD genome maps of Douglas-fir. J Hered 89(3):197–205

Kumar A, Kumar A, Pratush A (2014) Molecular diversity and functional variability of environmental isolates of *Bacillus* species. Springerplus 3(1):312

Lall GK, Darby AC, Nystedt B, MacLeod ET, Bishop RP, Welburn SC (2010) Amplified fragment length polymorphism (AFLP) analysis of closely related wild and captive tsetse fly (*Glossina morsitans morsitans*) populations. Parasit Vectors 3(1):47

Lanoot B, Vancanneyt M, Dawyndt P, Cnockaert M (2004) BOX-PCR fingerprinting as a powerful tool to reveal synonymous names in the genus *Streptomyces*. Emended descriptions are proposed for the species *Streptomyces cinereorectus*, *S. fradiae*, *S. tricolor*, *S. colombiensis*, *S. filamentosus*, *S. vinaceus* and *S. phaeopurpureus*. Syst Appl Microbiol 27(1):84–92

Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Pontén TS, Ussery DW, Aarestrup FM, Lund O (2012) Multilocus sequence typing of total genome sequenced bacteria. J Clin Microbiol 50(4):1355–1361

Lian J, HamediRad M, Zhao H (2018) Advancing metabolic engineering of *Saccharomyces cerevisiae* using the CRISPR/Cas system. Biotechnol J 13(9):e1700601

Ludwig W, Kirchhof G, Klugbauer N, Weizenegger M, Betzl D, Ehrmann M, Hertel C, Jilg S, Tatzel R, Zitzelsberger H, Liebl S, Hochberger M, Shah J, Lane D, Wallnöfer PR, Scheifer KH (1992) Complete 23S ribosomal RNA sequences of gram-positive bacteria with a low DNA G+C content. Syst Appl Microbiol 15(4):487–501

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar BA, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32(4):1363–1371

Lukácsi G, Tako M, Nyilasi I (2006) Pulsed-field gel electrophoresis: a versatile tool for analysis of fungal genomes. A review. Acta Microbiol Immunol Hung 53(1):95–104

Lupski JR, Weinstock GM (1992) Short, interspersed repetitive DNA sequences in prokaryotic genomes. J Bacteriol 174(14):4525–4529

Ma X, Chen SY, Bai SQ, Zhang XQ, Li DX, Zhang CB, Yan JJ (2012) RAPD analysis of genetic diversity and population structure of *Elymus sibiricus* (Poaceae) native to the southeastern Qinghai-Tibet Plateau, China. Genet Mol Res 11(3):2708–2718

Macedo NR, Oliveira SR, Lage AP, Santos JL, Araújo MR, Guedes RMC (2011) ERIC-PCR genotyping of *Haemophilus parasuis* isolates from Brazilian pigs. Vet J 188(3):362–364

Maidak BL, Larsen N, McCaughey MJ, Overbeek R, Olsen GJ, Fogel K, Blandy J, Woese CR (1994) The ribosomal database project. Nucleic Acids Res 22(17):3485–3487

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95(6):3140–3145

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. Science 339(6121):823–826

Maringele L, Lydall D (2006) Pulsed-field gel electrophoresis of budding yeast chromosomes. Methods Mol Bio 313:65–73

Marques AS, Marchaison A, Gardan L, Samson R (2008) BOX-PCR-based identification of bacterial species belonging to *Pseudomonas syringae*: *P. viridiflava* group. Genet Mol Biol 31(1):106–115

Martin KJ, Rygiewicz PT (2005) Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts. BMC Microbiol 5(1):28

Martin B, Humbert O, Camara M, Guenzi E, Walker J, Mitchell T, Andrew P, Prudhomme M, Alloing G, Hakenbeck R, Morrison DA, Boulnois GJ, Claverys JP (1992) A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res 20(13):3479–3483

Maule J (1998) Pulsed-field gel electrophoresis. Mol Biotechnol 9(2):107–126

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74 (2):560–564

McGhee GC, Sundin GW (2012) *Erwinia amylovora* CRISPR elements provide new tools for evaluating strain diversity and for microbial source tracking. PLoS One 7(7):e41706

Meintanis C, Chalkou KI, Kormas KA, Lymperopoulou DS, Katsifas EA, Hatzinikolaou DG, Karagouni AD (2008) Application of rpoB sequence similarity analysis, REP-PCR and BOX-PCR for the differentiation of species within the genus *Geobacillus*. Lett Appl Microbiol 46(3):395–401

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46

Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends Plant Sci 12(3):106–117

Mishra RK, Pandey BK, Singh V, Mathew AJ, Pathak N, Zeeshan M (2013a) Molecular detection and genotyping of *Fusarium oxysporum* f. sp. *psidii* Isolated from different agro-ecological regions of India. J Microbiol 51(4):405–412

Mishra RK, Pandey BK, Singh V, Pathak N, Zeeshan M (2013b) Genetic characterization of *Fusarium oxysporum* isolated from Guava in northern India. Afr J Microbiol Res 7 (33):4228–4234

Mishra RK, Pandey BK, Pathak N, Zeeshan M (2015) BOX-PCR-and ERIC-PCR-based genotyping and phylogenetic correlation among *Fusarium oxysporum* isolates associated with wilt disease in *Psidium guajava* L. Biocatal Agric Biotechnol 4(1):25–32

MLST allelic profiles and sequences, University of Oxford 2018 <https://pubmlst.org/data/>. Accessed on 08 Sept 2018

Moon BC, Lee YM, Kim WJ, Ji Y, Kang YM, Choi G (2016) Development of molecular markers for authentication of the medicinal plant species *Patrinia* by random amplified polymorphic DNA (RAPD) analysis and multiplex-PCR. Hortic Environ Biotechnol 57(2):182–190

Morales RGF, Resende JTV, Faria MV, Andrade MC, Resende LV, Delatorre CA, Silva PRD (2011) Genetic similarity among strawberry cultivars assessed by RAPD and ISSR markers. Sci Agric 68(6):665–670

Moser H, Lee M (1994) RFLP variation and genealogical distance, multivariate distance, heterosis, and genetic variance in oats. Theor Appl Genet 87(8):947–956

Mueller UG, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. Trends Ecol Evol 14 (10):389–394

Nadeem MA, Nawaz MA, Shahid MQ, Doğan Y, Comertpay G, Yıldız M, Hatipoğlu R, Ahmad F, Alsaleh A, Labhane N, Özkan H, Chung G, Baloch FS (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32(2):261–285

Nassonova ES (2008) Pulsed field gel electrophoresis: theory, instruments and application. Cell Tissue Biol 2(6):557–565

Niemann S, Dammann-Kalinowski T, Nagel A, Pühler A, Selbitschka W (1999) Genetic basis of enterobacterial repetitive intergenic consensus (ERIC)-PCR fingerprint pattern in *Sinorhizobium meliloti* and identification of *S. meliloti* employing PCR primers derived from an ERIC-PCR fragment. Arch Microbiol 172(1):22–30

Okatani AT, Ishikawa M, Yoshida SI, Sekiguchi M, Tanno K, Ogawa M, Horikita T, Horisaka T, Taniguchi T, Kato Y, Hayashidani H (2004) Automated ribotyping, a rapid typing method for analysis of *Erysipelothrix* spp. strains. J Vet Med Sci 66(6):729–733

Olive DM, Bean P (1999) Principles and applications of methods for DNA-based typing of microbial organisms. J Clin Microbiol 37(6):1661–1669
Olvera A, Calsamiglia M, Aragon V (2006) Genotypic diversity of *Haemophilus parasuis* field strains. Appl Environ Microbiol 72(6):3984–3992
Panneerchelvam S, Norazmi MN (2003) Forensic DNA profiling and database. Malays J Med Sci 10(2):20–26
Parizad EG, Parizad EG, Valizadeh A (2016) The application of pulsed field gel electrophoresis in clinical studies. J Clin Diagn Res 10(1):DE01–DE04
Pavlic M, Griffiths MW (2009) Principles, applications, and limitations of automated ribotyping as a rapid method in food safety. Foodborne Pathog Dis 6(9):1047–1055
Pavón ABI, Maiden MC (2009) Multilocus sequence typing. Methods Mol Biol 551:129–140
Presterl E, Nadrchal R, Winkler S, Makristathis A, Koller W, Rotter ML, Hirschl AM (1997) Molecular typing of *Acinetobacter baumannii* from ten different intensive care units of a university hospital. Eur J Clin Microbiol Infect Dis 16(10):740–743
Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35(21):7188–7196
Pursey E, Sünderhauf D, Gaze WH, Westra ER, van Houte S (2018) CRISPR-Cas antimicrobials: challenges and future prospects. PLoS Pathog, 14(6):e1006990
Ribot EM, Fitzgerald C, Kubota K, Swaminathan B, Barrett TJ (2001) Rapid pulsed-field gel electrophoresis protocol for subtyping of *Campylobacter jejuni*. J Clin Microbiol 39 (5):1889–1894
Ripamonti C, Orenstein A, Kutty G, Huang L, Schuhegger R, Sing A, Fantoni G, Atzori C, Vinton C, Huber C, Conville PS (2009) Restriction fragment length polymorphism typing demonstrates substantial diversity among *Pneumocystis jirovecii* isolates. J Infect Dis 200 (10):1616–1622
Rocco L, Ferrito V, Costagliola D, Marsilio A, Pappalardo AM, Stingo V, Tigano C (2007) Genetic divergence among and within four Italian populations of *Aphanius fasciatus* (Teleostei, Cyprinodontiformes). Ital J Zool 74(4):371–379
Roewer L (2013) DNA fingerprinting in forensics: past, present, future. Invest Genet 4(1):22
Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26(10):1117–1124
Salisbury SM, Sabatini LM, Spiegel CA (1997) Identification of methicillin-resistant *staphylococci* by multiplex polymerase chain reaction assay. Am Jo Clin Pathol 107(3):368–373
Sampaio JLM, Viana-Niero C, De Freitas D, Höfling-Lima AL, Leão SC (2006) Enterobacterial repetitive intergenic consensus PCR is a useful tool for typing *Mycobacterium chelonae* and *Mycobacterium abscessus* isolates. Diagn Microbiol Infect Dis 55(2):107–118
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467
Savelkoul PHM, Aarts HJM, De Haas J, Dijkshoorn L, Duim B, Otsen M, Rademaker JLW, Schouls L, Lenstra JA (1999) Amplified-fragment length polymorphism analysis: the state of an art. J Clin Microbiol 37(10):3083–3091
Schürch AC, Siezen RJ (2010) Genomic tracing of epidemics and disease outbreaks. Microb Biotechnol 3(6):628–633
Schwartz DC, Cantor CR (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 37(1):67–75
Shariat N, Dudley EG (2014) CRISPRs: molecular signatures used for pathogen subtyping. Appl Environ Microbiol 80(2):430–439
Sharma-Kuinkel BK, Rude TH, Fowler VG (2014) Pulse field gel electrophoresis. Methods Mol Biol 1373:117–130
Singh V, Rathore G, Kumar G, Swaminathan TR, Sood N, Kapoor D, Mishra BN (2007) Detection of the hole forming toxin hemolysin gene of *Aeromonas hydrophila* isolates. Indian Vet J 84:900–902

Singh V, Rathore G, Kapoor D, Mishra BN, Lakra WS (2008a) Detection of aerolysin gene in *Aeromonas hydrophila* isolated from fish and pond water. Indian J Microbiol 48(4):453–458

Singh R, Tan SG, Panandam JM, Rahman RA, Cheah SC (2008b) Identification of cDNA-RFLP markers and their use for molecular mapping in oil palm (*Elaeis guineensis*). Asia Pac J Mol Biol Biotechnol 16(3):53–63

Singh V, Chaudhary DK, Mani I, Somvanshi P, Rathore G, Sood N (2010a) Molecular identification and codon optimization analysis of major virulence encoding genes of *Aeromonas hydrophila*. Afr J Microbiol Res 4(10):952–957

Singh V, Chaudhary DK, Mani I, Somvanshi P, Rathore G, Sood N (2010b) Genotyping of *Aeromonas hydrophila* by box elements. Microbiol 79(3):370–373

Singh V, Mani I, Chaudhary DK, Somvanshi P (2011) Molecular detection and cloning of thermostable hemolysin gene from *Aeromonas hydrophila*. Mol Biol 45(4):551–560

Singh V, Chaudhary DK, Mani I (2012a) Molecular characterization and modeling of secondary structure of 16S rRNA from *Aeromonas veronii*. Int J Appl Biol Pharm Technol 3(1):253–260

Singh V, Mani I, Chaudhary DK (2012b) Molecular assessment of 16S-23S rDNA internal transcribed spacer length polymorphism of *Aeromonas hydrophila*. Adv Microbiol 2(2):72–78

Singh V, Braddick D, Dhar PK (2017) Exploring the potential of genome editing CRISPR-Cas9 technology. Gene 599:1–18

Singh V, Gohil N, Ramírez García R, Braddick D, Fofié CK (2018) Recent advances in CRISPR-Cas9 genome editing technology for biological and biomedical investigations. J Cell Biochem 119(1):81–94

Smouse PE, Chakraborty R (1986) The use of restriction fragment length polymorphisms in paternity analysis. Am J Hum Genet 38(6):918–939

Sola C, Abadia E, Le Hello S, Weill FX (2015) High-throughput CRISPR typing of *Mycobacterium tuberculosis* complex and *Salmonella enterica* serotype *Typhimurium*. Methods Mol Biol 1311:91–109

Steinmann KE, Hart CE, Thompson JF, Milos PM (2011) Helicos single-molecule sequencing of bacterial genomes. Methods Mol Biol 733:3–24

Stephenson DP, Moore RJ, Allison GE (2009) Comparison and utilisation of repetitive element PCR for typing *Lactobacillus* isolated from the chicken gastrointestinal tract. Appl Environ Microbiol 75(21):6764–6776

Stêpniak E, Zagalska MM, Switonski M (2002) Use of RAPD technique in evolution studies of four species in the family Canidae. J Appl Genet 43(4):489–500

Stern MJ, Ames GFL, Smith NH, Robinson EC, Higgins CF (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. Cell 37(3):1015–1026

Sullivan CB, Diggle MA, Clarke SC (2005) Multilocus sequence typing. Mol Biotechnol 29 (3):245–254

Sutton GG, White O, Adams MD, Kerlavage AR (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1(1):9–19

Szczuka E, Kaznowski A (2004) Typing of clinical and environmental *Aeromonas* sp. strains by random amplified polymorphic DNA PCR, repetitive extragenic palindromic PCR, and enterobacterial repetitive intergenic consensus sequence PCR. J Clin Microbiol 42(1):220–228

Tacão M, Alves A, Saavedra MJ, Correia A (2005) BOX-PCR is an adequate tool for typing *Aeromonas* spp. Antonie Leeuwenhoek 88(2):173–179

Taylor JW, Fisher MC (2003) Fungal multilocus sequence typing—it's not just for bacteria. Curr Opin Microbiol 6(4):351–356

Upadhyay SK, Kumar J, Alok A, Tuli R (2013) RNA-guided genome editing for target gene mutations in wheat. G3 (Bethesda) 3(12):2233–2238

Van Belkum A, Hermans PW (2001) BOX PCR fingerprinting for molecular typing of *Streptococcus pneumoniae*. Methods Mol Med 48:159–168

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304(5667):66–74

Versalovic J, Koeuth T, Lupski R (1991) Distribution of repetitive DNA sequences in eubacteria and application to finerpriting of bacterial enomes. Nucleic Acids Res 19(24):6823–6831

Versalovic J, Kapur V, Koeuth T, Mazurek GH, Whittam TS, Musser JM, Lupski JR (1995) DNA fingerprinting of pathogenic bacteria by fluorophore-enhanced repetitive sequence-based polymerase chain reaction. Arch Pathol Lab Med 119(1):23–29

Versalovic J, de Bruijn FJ, Lupski JR (1998) Repetitive sequence-based PCR (rep-PCR) DNA fingerprinting of bacterial genomes. In: de Bruijn FJ, Lupski JR, Weinstock GM (eds) Bacterial genomes. Springer, Boston, pp 437–454

Vila J, Marcos MA, De Anta MJ (1996) A comparative study of different PCR-based DNA fingerprinting techniques for typing of the *Acinetobacter calcoaceticus*-A. baumannii complex. J Med Microbiol 44(6):482–489

Wang X, Miller AB, Lepine AJ, Scott DJ, Murphy KE (1999) Analysis of randomly amplified polymorphic DNA (RAPD) for identifying genetic markers associated with canine hip dysplasia. J Hered 90(1):99–103

Wang G, Clark CG, Liu C, Pucknell C, Munro CK, Kruk TM, Caldeira R, Woodward DL, Rodgers FG (2003) Detection and characterization of the hemolysin genes in *Aeromonas hydrophila* and *Aeromonas sobria* by multiplex PCR. J Clin Microbiol 41(3):1048–1054

Wang LT, Lee FL, Tai CJ, Kasai H (2007) Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA–DNA hybridization in the *Bacillus subtilis* group. Int J Syst Evol Microbiol 57(8):1846–1850

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18(24):7213–7218

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18 (22):6531–6535

Wilson LA, Sharp PM (2006) Enterobacterial repetitive intergenic consensus (ERIC) sequences in *Escherichia coli*: evolution and implications for ERIC-PCR. Mol Biol Evol 23(6):1156–1168

Wu F, Della-Latta P (2002) Molecular typing strategies. Semin Perinatol 26(5):357–366

Wu F, Della-Latta P (2006) Pulsed-field gel electrophoresis. In: Tang YW, Stratton CW (eds) Advanced techniques in diagnostic microbiology. Springer, Boston, pp 143–157

Wu S, Guo N, Yin Z, Chai J (1996) Characterization of pathogenic fungi genomes using pulsed field gel electrophoresis. Chinese Med Sci J 11(3):188–190

Yang W, Kang X, Yang Q, Lin Y, Fang M (2013) Review on the development of genotyping methods for assessing farm animal diversity. J Anim Sci Biotechnol 4(1):2

Ye Y, Jiang Q, Wu Q, Zhang J, Lu J, Lin L (2012) The characterization and comparison of *Staphylococcus aureus* by antibiotic susceptibility testing, enterobacterial repetitive intergenic consensus–polymerase chain reaction, and random amplified polymorphic DNA–polymerase chain reaction. Foodborne Pathog Dis 9(2):168–171

Zabeau M, Vos P (1993) Selective restriction fragment amplification:a general method for DNA fingerprinting. Publication 0 534 858 A1, bulletin 93/13. European Patent Office, Munich, Germany

RNA-Guided CRISPR-Cas9 System for Removal of Microbial Pathogens

12

Gargi Bhattacharjee, Khushal Khambhati, and Vijai Singh

Abstract

CRISPR-Cas9 technology has been cherished and well appreciated by the scientific community. The popularity of CRISPR-Cas9 technology is because it provides simple and efficient directions for genome engineering with feasible applications in a broad range of organisms. It stands to reason that the development of CRISPR-Cas9 is probably among the greatest revolution in the field of molecular biology, ever since the discovery of PCR. Genome engineering of microbes and other organisms may open up newer avenues to achieve a dynamic ecosystem. In this chapter, research on the use of CRISPR-Cas9 technology as an anti-phytopathogenic arsenal has been highlighted. Furthermore, the engineered organism developed using CRISPR-Cas9 technology has also been explained. Besides the applicative side, the background and molecular mechanisms of the CRISPR-Cas9 system have been mentioned and explained thoroughly.

G. Bhattacharjee · K. Khambhati
School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

V. Singh (✉)
School of Biological Sciences and Biotechnology, Institute of Advanced Research, Gandhinagar, India

Present address: Department of Biosciences, School of Sciences, Indrashil University, Kadi, Gujarat, India
e-mail: vijai.singh@indrashiluniversity.edu.in

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_12

12.1 Introduction

The behavior of an organism is determined by the genetic materials stored within the cell. Just like changing the alphabets of the word(s), either disturbs, recasts, or changes, the sense or information that has to be conveyed through a given sentence, with the same notion, the permanent changes, or modifications of genetic instruction stored within the cell(s) alter the behavior of the subjected model. Genome engineering is one way that helps to achieve such "precise" manipulation. It can be explained as an approach which deliberately modifies or creates correction or deletion in the genome of a living organism, with an intention to fulfill a particular purpose (Carr and Church 2009). Few of these purposes may be to explore and learn the biology of the subjected model and to develop microbes which are capable of acting as biosensors or help in efficient bioremediation (Carr and Church 2009). In comparison to the classical chemical synthesis protocols, genetically engineered organisms have shown a profound capability to generate industrially and commercially important products in a cost-effective manner (Nielsen et al. 2014).

The use of chemicals, radioactive mutagens or transposon elements is not favored when experiments demand precise manipulation as such approaches create random mutations besides the desired alteration (Carroll 2017). Eventually, several tools have been devised and developed that help achieve the targeted genome manipulation, with each having its own pros and cons. For example, the lambda red-based recombination is a simple and straightforward tool to use, but it provides a low frequency of recombinants (Yu et al. 2000). However, the frequency can be slightly improved by using the strains already harboring certain endonuclease knockouts (Mosberg et al. 2012). Several other tools have also been developed which involves customizable nuclease that can help target the desired sequences. Zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) are examples of such customizable DNA scissors. TALEN exhibits superiority over ZFN as it is easy to design, can recognize longer sequences, and results in less off-target activity (Gupta and Musunuru 2014). Even though TALENs show precedent properties, it does lag in certain terms as compared to ZFN, such as cDNAs corresponding to pairs of TALENs are hard to deliver and express in cells as they tend to be larger in size (Gupta and Musunuru 2014). However, a new nuclease has to be designed if a new sequence is to be manipulated via ZFN or TALEN, making them costly, time consuming, and tedious to use.

Inspired by molecular mechanisms involved in CRISPR-Cas9 (clustered regularly interspaced short palindromic repeats-associated protein 9)-mediated immune system found in bacteria, the Cas9-gRNA complex has been exploited as a precise, scalable, cost-effective, and user-friendly genome engineering tool (Gasiunas et al. 2012; Jinek et al. 2012; Cong et al. 2013; Hsu et al. 2014). Considering the abovementioned features, CRISPR-Cas9 is a "tool of choice" among researcher for achieving the precise genome editing in an explicit way (Wiles et al. 2015). Reports suggest that more than 9000 research papers have been indexed in PubMed as regards "CRISPR" or "Cas9" since 2012 (Adli 2018). Initial in vitro studies highlight the caliber of Cas9 endonuclease for genome engineering purpose (Jinek et al. 2012). Over time, the use of Cas9 for genome editing has been explored in

human and mouse cell lines, along with a considerable number of a unicellular and multicellular organism, such as bacteria, yeast, mosquitos, zebrafish, and mice (Cong et al. 2013; DiCarlo et al. 2013; Gantz et al. 2015; Hisano et al. 2015; Jakočiūnas et al. 2015; Altenbuchner 2016; Singh et al. 2017; Burgio 2018). Diverse services are available that Cas9 or its recombinant form can provide for engineering purpose, some of which are listed as follows (Mali et al. 2013; Guo et al. 2016; Singh et al. 2018):

- Simplex or multiplex genome editing including double-strand breaks and nicks
- Efficient gene recombination
- Modulating genome architecture
- Gene(s) up- and downregulation

The significant increase of the world population calls for an escalating demand for food and shelter (Umesha et al. 2017). Therefore, to overcome the mentioned challenge, implementation of sustainable agricultural practices is required. A way to tackle the menace of food insecurity is to engineer microbes (Umesha et al. 2017). Analyzing the basic principle and mechanism of how microbes interact and behave in a given environment could possibly help to engineer microbes having the desired characters. The resultant engineered microbe would be meant for enhancing plant growth, for example, by increasing nutrient availability or to eradicate soil-borne diseases (Umesha et al. 2017). Eventually, several studies have been conducted which include the incorporation of CRISPR-Cas9-based systems for the evolution of engineered microbial tools that could possibly help to headway in achieving a sustainable agroecosystem.

In this chapter, we explore and explain how CRISPR-Cas9 technology can be used to eradicate several pathogenic microorganisms. By referring to such examples, the reader would be able to develop a better understanding about the use of CRISPR-Cas9 technology for dismissal of the pathogenic organism and subsequently be inspired to use such tactics for the development of a sustainable ecosystem. The chapter is divided into two sections. First, we explain the molecular mechanisms of the naturally occurring CRISPR-Cas9 system and the general working principle of CRISPR-Cas9 system. Second, we elaborate the applicative part of CRISPR-Cas9-based tools/systems in a range of microorganisms and parasites.

12.2 Background of CRISPR-Cas System

For an infection to strike and develop, phage has to go through several barriers that are imposed by the bacterial cell which help the bacteria to resist the infection (Shabbir et al. 2016). These barriers are broadly divided into innate and adaptive immune system. Tactics such as prevention of phage absorption or entry of its genetic material inside the cells are sensed or recognized by the bacterial innate immune system (Bikard and Marraffini 2012). In addition, digestion of the phage DNA by restriction endonuclease and aborting the infection by cell suicide are also examples of separate events performed as an inherent component of the innate immune system. The latter

mentioned strategy, named abortive infection, helps to protect the bacterial population rather than individual cell itself (Bikard and Marraffini 2012).

Another line of defense mechanism discovered in the bacterial cell is its adaptive immune system. CRISPR-Cas-mediated immunity is an example of such a defense mechanism against phage and other extrachromosomal elements. The CRISPR loci can be found in the archaeal and the bacterial genome that are marked as peculiar short repeats, spaced by short nucleotide (nt) sequence of encountered foreign DNA (Karginov and Hannon 2010). The "Cas" in "CRISPR-Cas" stands for CRISPR-associated sequence, and oftentimes, their product facilitates the CRISPR loci to demonstrate the adaptive nature of the immune system as well as helps to recognize and cleave the foreign sequence (either DNA or RNA depending on the type of CRISPR-Cas system).

The quest for CRISPR began in the year 1987 (Ishino et al. 1987), where some unique interspaced repeats were observed in the genome of *E. coli*. The term "clustered regularly interspaced short palindromic repeats" (CRISPR) was coined in 2002, and the possible biological role of CRISPR in the immune system was proposed in 2006 (Makarova et al. 2006; Karginov and Hannon 2010). However, in 2010 through in vivo studies, discovery as regards to the adaptive nature of CRISPR-Cas-mediated immune system was disclosed (Garneau et al. 2010). With the current knowledge, CRISPR-Cas systems are classified into class I and class II. Because of the frequent recombination and absence of universal Cas gene within the different CRISPR-Cas systems, classifying them becomes both tedious and difficult (Koonin et al. 2017). Currently, the classification is made by considering several factors including availability of different signature Cas genes and sequence similarities of shared Cas genes across different CRISPR-Cas systems. Moreover, the structure of CRISPR and organization of genes in the CRISPR loci are also taken into account while classifying CRISPR-Cas system (Koonin et al. 2017).

Class I consists of type I, III, and IV CRISPR-Cas system. Among all the types, types I and III are most commonly found, and contrary to this, type IV is a rare occurring CRISPR-Cas system. Generally, class I CRISPR-Cas system requires multiple subunits and proteins to demonstrate its defense mechanism, while class II CRISPR-Cas system does the same job by recruiting a single and simple multidomain protein rather than the involving multiple subunits and proteins. Furthermore, class II CRISPR-Cas system has very basic and uniform CRISPR-Cas loci as compared to class I system. CRISPR-Cas9-mediated immune system is an example of class II type II CRISPR-Cas system. In addition to type II, class II also includes type V and VI CRISPR-Cas system (Koonin et al. 2017). Research is on the progress about new discovery of CRISPR which is expanding every time. It can be further used for a variety of applications in environment, biofuels, agriculture, and more.

12.3 Molecular Mechanism of CRISPR-Cas9 System

CRISPR-Cas9-mediated immune system recruits a special kind of endonuclease that pursues for a target sequence (Rath et al. 2015). The hunt to cleave a particular sequence is guided by an RNA molecule which is called a guide RNA (gRNA). With

the involvement of a distinct kind of multi-domain endonuclease and gRNA, the CRISPR-Cas9 system is able to find a target sequence from the pool of genetic instructions present within the cell. On target recognition, the system activates itself and cleaves the recognized sequence (Rath et al. 2015). The target sequence is recognized based on the complementarity between the gRNA molecule and the DNA sequence. With that arises a question on how the cell gets to know as regards to the sequence of gRNA that would target a particular foreign sequence.

Being a part of the bacterial adaptive immune system, CRISPR-Cas9 has the ability to memorize the previously encountered sequence (Rath et al. 2015). Whenever a phage genome enters into the cell, the Cas proteins acquire a certain portion of foreign nucleotide sequences between two "repeat sequences" found within the CRISPR locus (Wei et al. 2015). Thereafter, the incorporated sequence would act as a memory for the host organism. The sequence of the phage genome selected to be linked within the CRISPR locus is known as protospacer. Once it gets integrated into the bacterial genome, it is termed as a spacer (Shah et al. 2013). However, the incorporation is complete in a way that the resultant product would lead to replication of the repeat sequence found adjacent to the leader sequence in the CRISPR locus. Furthermore, the incorporated spacer would be present between the newly duplicated repeat sequences (Wei et al. 2015). This scenario ensures the maintenance of the repeat-spacer architecture. The integration of 33–35-nt-long protospacer sequence is done by identification of the proximal end of the leader sequence and 5' end of its adjacent repeat sequence, through Cas1–Cas2 protein complex (Heler et al. 2015; Wei et al. 2015). The entire event of incorporating the protospacer sequence into the genome of the cell, specifically within CRISPR locus, is called spacer acquisition or more commonly adaption (Fig. 12.1).

However, any random sequence is not selected to act as a protospacer during the spacer acquisition. The sequence next to the protospacer adjacent motif (PAM) is selected and acts as a potential spacer sequence for CRISPR locus. PAM motif is specific short three- to five-nucleotide-long sequence that is recognized by a multidomain Cas9 nuclease (Mojica et al. 2009; Heler et al. 2015). Once recognized by Cas9, the Cas1–Cas2 protein complex would serve as an integrase and carry out steps necessary for spacer acquisition. Preference to a nonself protospacer over self seems to be hugely dependent on RecBCD-mediated DNA repair mechanism. Whenever there is stalled replication fork or DNA damage, RecBCD would cleave the dsDNA into short single-stranded DNA until it reaches the Chi site (Levy et al. 2015; Wang et al. 2015). Once it approaches Chi site, the 5' end has been degraded to some extent, producing 3' overhangs (Dillingham and Kowalczykowski 2008). The frequency of occurrence of Chi site would be less in the nonself DNA compared to the self DNA. Thus, during the RecBCD-mediated genome repair, the phage DNA could produce more degraded short single-stranded DNA in comparison to self DNA (Levy et al. 2015). It is assumed, through unknown mechanism, that the short ssDNAs are converted into dsDNA molecules. The potential PAM motif of protospacer is recognized by the Cas9 nuclease that recruits Cas1–Cas2 complex to integrate the 33–35 bp sequence (protospacer) into the CRISPR loci (Wright et al. 2016).

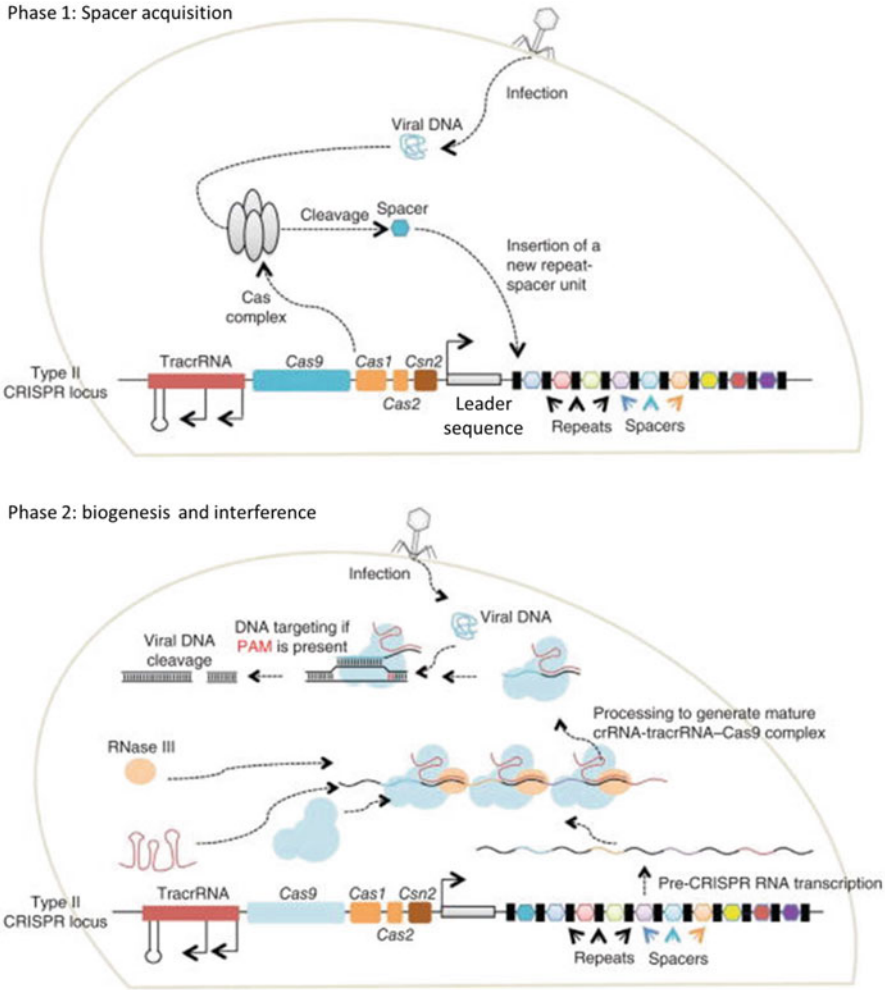

Fig. 12.1 CRISPR-Cas system defense mechanism by acquisition, RNA processing, and interference. Phase 1, on entry of viral DNA into the cell, the Cas1–Cas2 complex integrates the sequence adjacent to the PAM motif into the CRISPR loci. The first repeat unit and protospacer sequence is incorporated between the newly formed repeat sequence. Cas9 helps in the recognition of the PAM sequence. Cas1–Cas2 as a whole is shown as Cas complex. This step of protospacer incorporation is referred as spacer acquisition. In phase 2, the CRISPR loci are transcribed as a whole and are termed as pre-CRISPR RNA. With the formation of crRNA/tracrRNA-Cas9 complex and involvement of RNase III, the pre-crRNA is converted to mature CRISPR RNA, and this step is known as biogenesis. With the emergence of mature crRNA, Cas9 forms complex with it and hunts for the target sequence. The target sequence tends to lie next to the PAM motif, and it is complementary to the spacer sequence of mature crRNA. This step of DNA cleavage is called interference. (Figure reproduced with permission from Mali et al. 2013© (2013) Springer Nature)

The leader sequence of CRISPR locus not only helps in spacer acquisition but also helps to transcribe the CRISPR-RNA (crRNA). The leader sequence is present on the upstream of the repeat-spacer architecture. Furthermore, promoter sequence is embedded inside it (Carte et al. 2014). This sequence promotes the transcription of the so-called repeat-spacer architecture as a whole, and the resultant RNA obtained is called pre-crRNA. However, it is not the pre-crRNA but the mature crRNA (in applicative biology called as gRNA) that guides the multi-domain Cas9 nuclease to target the particular sequence (Brouns et al. 2008; Karvelis et al. 2013). RNase III and tracrRNA are responsible to convert the pre-crRNA into the mature crRNAs (Deltcheva et al. 2011; Garrett et al. 2015). A set of events that is performed following the transcription of pre-crRNA to its maturation is categorized as biogenesis. The sequence of tracrRNA complements the repeat sequence found within the pre-crRNA and is meant to form tracrRNA/pre-crRNA complex which is stabilized by the Cas9 nuclease (Garrett et al. 2015).

The transcribed repeat-spacer architecture has hydrogen-bonded tracrRNA on each and every repeat sequence of pre-crRNA. The gene corresponding to tracrRNA is located within the CRISPR-Cas loci (Chylinski et al. 2013; Chylinski et al. 2014). The role of RNase III in biogenesis is to cleave the tracrRNA:pre-crRNA complex at the repeat sequence, thus leading to the formation of several crRNAs (Deltcheva et al. 2011) (Fig. 12.1). The resultant crRNAs consist of a spacer and some flanking repeat sequences (on both ends of crRNA) which are attached to the nucleotide sequence of tracrRNA. The nucleotide sequences of repeats corresponding to the 5' of the produced crRNAs as well as some nucleotide sequence of the spacer are trimmed by some unknown nuclease to form a mature crRNA (Wright et al. 2016).

On the binding of mature crRNA to the Cas9 nuclease, there is a conformational shift from its auto-inhibited state, thus allowing the Cas9 nuclease to look for the target sequence (Jinek et al. 2014). Furthermore, two arginine residues found in the PAM-interacting domain of the Cas9 would be pre-positioned to search for the PAM motif in dsDNA (Wright et al. 2016). On recognizing the PAM motif, the nuclease loses the first 10–12 nt sequences starting from the proximal end of PAM motif found in dsDNA (Szczelkun et al. 2014). The first 10–12 nt sequences from the 3' end of the PAM motif is called the seed sequence (Szczelkun et al. 2014). Once the seed region and its distal end sequence have opened up, the mature crRNA forms complementarity with the seed over and above its distal end nucleotide sequence. Only a few, if not all, mismatches are tolerated in the seed region of the open DNA with the mature crRNA (Wright et al. 2016). Once a perfect or near-perfect match is achieved, the HNH domain of Cas9 nuclease is positioned in its catalytic active position. This, in turn, allosterically regulates and activates the RuvC domain of that particular nuclease (Sternberg et al. 2015). HNH and RuvC-like domain are meant to cleave target and the nontarget strand of dsDNA, respectively (Jinek et al. 2012). The entire process to hunt for the target sequence with the help of Cas9 nuclease is termed as interference (Fig. 12.1). Along with CRISPR-Cas9 system, the molecular mechanism of several other CRISPR-Cas systems is being studied, which promises not only effective genome engineering but also pathogen detection (Knott and Doudna 2018; Gootenberg et al. 2018; Khambhati et al. 2018).

12.4 The CRISPR-Cas9 Genome Editing Technology

Investigating the biological systems does satisfy one's curiosity, as it helps to give an answer to a particular question. The efforts put by researchers are highly appreciated as it helps unveil the behavior of the subjected system. However, if the knowledge gained by investigating the basic biology is somehow translated to some kind of applicative technology, then it would facilitate the upliftment of the human society. For example, increased understanding with regards to the behavior of restriction endonuclease has led to the development of several molecular cloning tools along with several crime investigating techniques. The developed molecular cloning tools have acted as a major support for biotechnology-based industries and also in exploring the basic biology (Roberts 2005).

Likewise, much knowledge has been gained about the working principle of the CRISPR-Cas9 system that has helped to develop an efficient and easy to use genome engineering tool(s). The in vitro proof-of-concept was demonstrated by Jinek and colleagues in 2012 (Jinek et al. 2012). The CRISPR-Cas9 can function as a tool for making precise double-strand breaks by expressing Cas9 protein inside a cell with the help of an expression vector. Simultaneously, gRNA can direct Cas9 towards the target sequence (Cong et al. 2013). The introduction of double-stranded break leads to the activation of repair mechanism by nonhomologous end joining (NHEJ) or by the homology-directed repair (HDR) pathway. The activation of such a mechanism facilitates the modification or replacement of the desired sequence or further can help to create gene knockouts (Singh et al. 2017; Singh et al. 2018).

Rather than expressing the tracrRNA and the spacer sequence individually, the single gRNA (sgRNA) can be chemically synthesized and expressed directly to cleave the desired sequence via Cas9. The sgRNA includes a tracrRNA and spacer sequence (12–13 nt target complementary sequence), which are physically linked as one with the help of a linker sequence. However, the only requirement is that the selected target sequence in dsDNA should be adjacent to the PAM motif (Jinek et al. 2012; Cong et al. 2013). A particular sequence of PAM motif corresponds to a particular kind of Cas9 protein, consequently, by identifying the different variants of Cas9 proteins and its corresponding PAM site (Singh et al. 2017) one can increase the proximity of using the tool to target the wide range of sequence (Fig. 12.2).

The popularity of Cas9 is not only confined to make precise double-strand breaks but also a mutant version named dead Cas9 (dCas9) is known for regulating the gene expression (La Russa and Qi 2015). The dCas9 is unable to produce double-strand breaks because of having mutation in the active domain of the protein (RuvC and HNH domains). As its applicative side, one can program the dCas9 with the help of gRNA molecule to recognize and anchor onto a particular promoter sequence. The promoter sequence of choice could be upstream to the gene sequence of which the downregulation is desired. The anchored dCas9 interferes with the RNA polymerase that is supposed to bind with the promoter sequence and prevents it from carrying out transcription. Furthermore, an ongoing transcription can be brought to a hold by targeting the gene sequence rather than the promoter sequence (Bikard et al. 2013). Thus, dCas9 can be used as a molecular tool to downregulate the gene expression. In a similar manner, it can also help in the upregulation of the desired gene by allowing

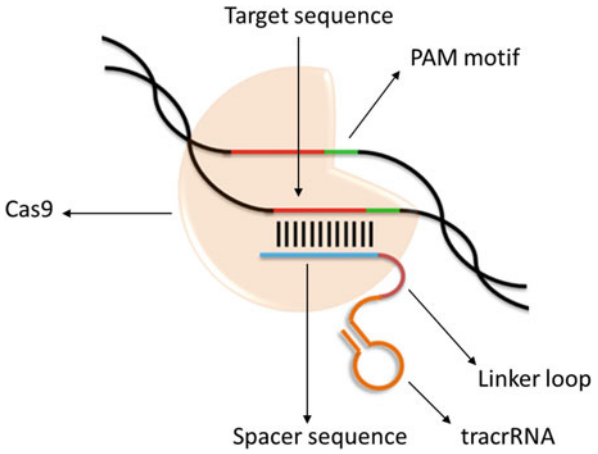

Fig. 12.2 Schematic representation of genome editing using CRISPR-Cas9 system. The guide RNA, which includes 20-nt spacer is linked to tracrRNA with the help of a linker sequence (forming a loop). The Cas9 nuclease in the presence of guide RNA forms a complex that can bind with the target region in the presence of PAM sequence that allows to create a double-strand breaks in order to generate a gene knockout. (Jinek et al. 2012)

the fused form of dCas9 that binds onto the promoter sequence. For example, the fusion of the trans-activator domain with dCas9 may allow us to activate the gene expression. The fused dCas9 could possibly have omega (ω) subunit of RNA polymerase (RNAP) linked to it, which allows the recruitment of RNAP to the promoter region. The recombinant form of the dCas9 protein would recruit the RNA polymerase onto the promoter sequence and activate the transcription of the desired gene (Bikard et al. 2013).

12.5 Potential Application of CRISPR-Cas9 System for Removal of Pathogens

12.5.1 Genome Editing of Fungi Using CRISPR-Cas9 System

Fungi are ubiquitous in nature that may either occur as a unicellular organism or as a highly complex multicellular organization. Depending upon their habitat, fungi delineate a range of infective properties. Being highly efficient decomposers, fungi feed on the dead and decomposed the matter, channelizing the important elements such as carbon, nitrogen, salts, and other organic matter back into the environment. Other than its involvement in human disease and infections, they are the general spoilers of food and crops. Fungi are also associated with synthesis of high value naturally available biologically active products in agro-based, food, and pharmaceutical industry. Parasitic fungi residing over plants and crops cause mildew and rust, resulting in huge monetary losses every year. As far as the higher organisms are concerned, the numbers of fungi involved in animal and plant diseases are relatively

less as compared to bacterial ones. Fungi are often characterized based on their structure or the fruiting bodies they form, their life cycle, and the type and arrangement of the spore (reproductive or distributional) they produce.

Predominantly, fungi are characterized into three major groups:

(i) Unicellular microscopic yeasts

Yeasts are single-celled, eukaryotic members of the fungus kingdom constituting about 1% of the total fungal population. They are small, round lemon-like cells sized ~5 μm in diameter (Duina et al. 2014). Yeasts multiply asexually by budding a daughter cell off from the parent cell (e.g., *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*).

(ii) Multicellular filamentous molds

Molds are composed of fine threadlike structures called hyphae. These hyphae divide repeatedly at the tip and form mesh-like arrangements by intertwining with neighboring hyphae. This mesh-like network is called mycelium. The digestive enzymes are present at the tips of the hyphae that digest the organic matter surrounding their habitat which is then utilized as the source of energy. Molds form spores on their aerial branches, which are nothing but their reproductive structures enclosed within a protective covering in order to protect them against harsh and unsuitable climatic condition or a state of starvation. The spores spread via wind, insects, or rain. When the conditions become favorable, the resistant spores germinate into a fresh new fungus and produce new hyphae (e.g., *Rhizopus nigricans* and *Spinellus fusiger*).

(iii) Macroscopic filamentous fungi bearing large fruiting bodies

Similar to filamentous molds, mushrooms also bear spores that are utilized to propagate and maintain their generations. However, a major difference that distinguishes mushrooms from molds is that mushrooms form visible fruiting bodies to hold the spores together. The fruiting body is commonly associated to a cap-like structure of mushroom, which is composed of densely packed hyphae, and the gills underneath the cap is where the spores reside (e.g., *Agaricus bisporus* (edible button mushrooms) and *Amanita phalloides* (deadly poisonous mushroom)).

The close associations of the filamentous fungi with humans have led to the thought of manipulation of the genome of higher fungi to extract high-value bulk and fine bioactive natural compounds (Thrane et al. 2007). Owing to their simple structure and genetic composition, yeast is used as model organism to study genetics. Yeasts *Saccharomyces* are extensively used in the baking and brewing industries. Considering their ability to synthesize large amounts of proteins along with the presence of a complex yet efficient post-translational processing system, a diverse range of fungus, namely, *Aspergillus niger*, *Aspergillus oryzae* and *Trichoderma reesei*, are used in the synthesis of enzymes and valuable proteins, the species being regarded safe by the US Food and Drug Administration (Ward 2012; Shi et al. 2017).

Filamentous fungi find applications in the synthesis of antibiotics, organic compounds, novel drugs, pigments, and so on. Filamentous fungi are applied in the synthesis of a variety of pharmaceutical products. Few of the cholesterol-

lowering drugs are statins, such as lovastatin and mevastatin, produced from *Aspergillus terreus* and *Penicillium citinium*, respectively (Barrios-González and Miranda 2010). Antifungal medication such as griseofulvin that is used to treat infections of the skin, scalp, nails, and groin is developed from the *Tinea* strains of fungi, and echinocandins, popularly known as "penicillin of antifungals" (Kumar and Jha 2017), along with penicillins, cephalosporins (Hamad 2010), and a few other antibiotics of the β-lactam class (Cho et al. 2014) that are derived from filamentous fungi. Conversely, some fungi produce mycotoxins that are pernicious to humans and plants. Some opportunistic pathogenic forms of yeasts such as *Candida* cause infections in immunocompromised individuals. Toxins such as fumonisins, trichothecenes, and aflatoxins from *Fusarium verticillioides*, *Fusarium graminearum*, and *Aspergillus flavus*, respectively, are to name a few (Woloshuk and Shim 2013), that contaminate the food products or infect crops, ultimately leading to huge economic losses (Harris et al. 2016; Lecellier et al. 2015). Therefore, to subside the adverse effects generated by fungi and potentiate their use for creating valuable yet economic products, it is necessary to understand their genetic build-up and reconstruct them accordingly. A pressing need has arisen to edit the fungal genome in order to produce large amounts of complex and useful chemicals for industrial and agricultural applications. Figure 12.3 shows a quick look of CRISPR-Cas9-assisted genome editing in microbes and parasite.

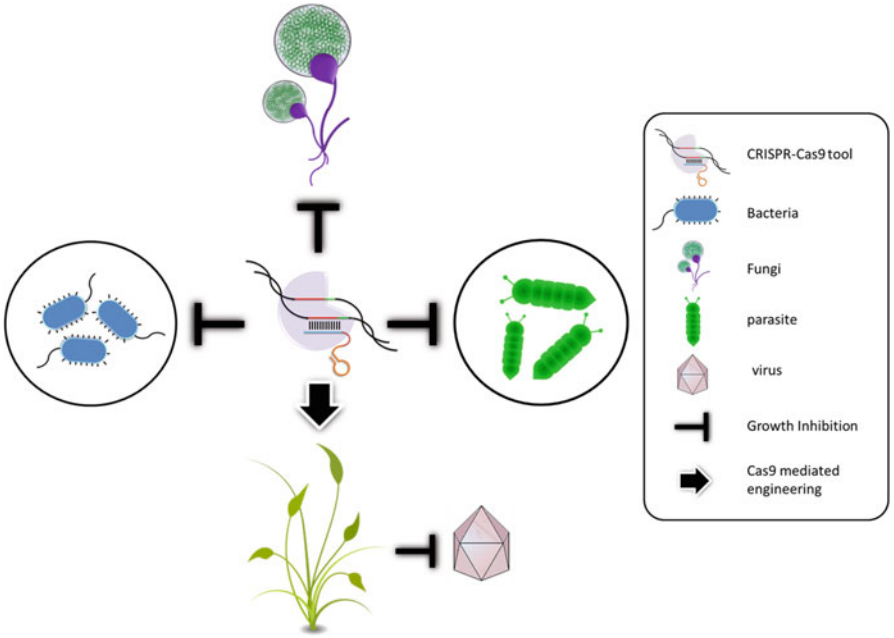

Fig. 12.3 Applications of CRISPR-Cas9 technology toward plant protection and better growth. It has demonstrated the inhibition of the plant pathogenic bacteria, fungi, and parasite. Plant has been modified using CRISPR-Cas9 system for viral resistance

The CRISPR-Cas9 platform offers to orchestrate the fungal genome and engineer its synthetic gene circuits accordingly. The platform thus finds application in decoding pathogenicity, synthesizing valuable metabolites and drugs, and boosting bioenergy processes (Deng et al. 2017). A key aspect of designing a highly efficient CRISPR-Cas9 tool is to have a coherent promoter. For the fungal system, often the poly III U6 promoter is used for expressing the sgRNA sequences, the promoter being excellent in transcription (Miyagishi and Taira 2002). In case of unavailability of the U6 promoter, *Saccharomyces cerevisiae* promoter SNR52 works fine in *Aspergillus fumigatus* (Fuller et al. 2015) and *Neurospora crassa* (Matsu-ura et al. 2015), or the T7 promoter of bacteriophage for activating sgRNA for *Trichoderma reesei* (Liu et al. 2015) and *Penicillium chrysogenum* (Pohl et al. 2016) has been described.

The desirable Cas9 expression can be attained by incorporating suitable inducible or constitutive promoters. Constitutive promoters that are used for the expression of Cas9 in filamentous fungi include gpdA (Zhang et al. 2016) and translation-elongation factor 1α (tef1) (Kuivanen et al. 2016; Liu et al. 2017). Codon optimization of Cas9 could readily enhance the genome editing ability in filamentous fungi, the proof of which has been observed in multiple strains including *Aspergillus* (Fuller et al. 2015; Katayama et al. 2016; Weber et al. 2016), β-lactam-producing *Penicillium chrysogenum* (formerly known as *Penicillium notatum*) (Pohl et al. 2016), *Trichoderma reesei* that is a rich source of industrially valuable cellulolytic enzyme (Liu et al. 2015), and corn smut causing *Ustilago maydis* strain (Schuster et al. 2016). Essentially, three main approaches are employed when it comes to transforming the Cas9-sgRNA complex into filamentous fungi which include polyethylene glycol (PEG) transformation, *Agrobacterium*-mediated transformation (AMT), and blastospore-mediated transformation (Deng et al. 2017). The PEG-$CaCl_2$ transformation is relatively simple, and the method makes use of protoplasts generated from hypha and germinated spores, whereas the *Agrobacterium* and the host are directly co-cultured for AMT-based transformations. Else ways, a blastospore transformation is facilitated by lithium acetate-mediated delivery of DNA.

The mechanics of the CRISPR-Cas9 system can be regulated in several ways. The inhibition of the system is delineated by a set of natural inhibitors, called anti-CRISPR (Acr) proteins. The proteins bind with Cas9 in the CRISPR-Cas9 complex, thereby functioning as its "off-switches" and thus get hold over the Cas9 activity. A bright side of the anti-CRISPR (Acr) proteins is their involvement in reducing the chances of CRISPR-Cas9 off-targeting and therefore diminishing the side effects (Shin et al. 2017). For example, a CRISPR inhibitory protein called AcrIIA4, which binds to the CRISPR-Cas9 complex involved in sgRNA-mediated targeting in sniping off the mutated hemoglobin gene responsible for sickle cell anemia, reduces the possibility of off-targeting by fourfold (Shin et al. 2017).

Light and chemical means have been used for corrections in the genome in *Aspergillus fumigatus* and many other eukaryotic strains (Zhang et al. 2016). Apart from this, it is also possible to maneuver genome editing by spatially and temporally regulating Cas9 activity through the anti-CRISPR system (Pawluk et al. 2016; Rauch et al. 2017). Majority of fungal infections occur as a function of the synergistic effect of multiple genes. CRISPR-Cas9 complex with utmost efficiency

can target the toxin-producing genes in most fungal pathogens. *Phytophthora sojae* is an oomycete (water mold) that attacks plants of agricultural and ornamental importance. The plant pathogen causes the stem and root in soybean plants to rot, leading to a whopping loss of soybean crops. Disrupting and replacing the *Avr4/6* gene belonging to the superfamily of RXLR virulence effector proteins using CRISPR-Cas9 enable a better control over the pathogen (Fang and Tyler 2016).

Attempts have also been taken to edit the genome of a commonly found saprophyte (an organism that feeds on dead and decaying organic matter) *A. fumigatus* (Zhang et al. 2016). Modifications such as disrupting the polyketide synthase gene (*pksP*), an important enzyme involved in toxin biosynthesis, hamper the ability of *A. fumigatus* to produce toxic compounds and consequently minimize the detrimental effects on the host (Fuller et al. 2015). Engineering entomopathogenic fungi *Beauveria bassiana*, a parasite that grows over arthropods and causes white muscardine disease, can serve the purpose of insect or pest management (Chen et al. 2017).

Cellulose is the most abundant biomass on earth and has proved to be a lucrative resource for the paper, wood, fiber, fodder, cosmetic, and pharmaceutical industries (Shokri and Adibkia 2013). The humongous quantity of agro-based waste can be utilized to generate a profitable product. The degradation of cellulose is catalyzed by an enzyme called cellulase. Some fungi belonging to the genus *Aspergillus*, *Rhizopus*, *Trichoderma*, *Fusarium*, *Neurospora*, and *Penicillium* can very easily degrade the cellulose into simple constituents (Sajith et al. 2016). Incorporating the CRISPR-Cas9 system to enhance the production of cellulase in several fungal strains has shown some assuring results in minimizing the size of the agro-waste generated. CRISPR-Cas9-mediated upregulation of transcriptional factor (clr-2) results in an increase of cellulases in *Neurospora crassa* (Matsu-ura et al. 2015).

Myceliophthora, an ascomycete (sac fungi), is a thermophilic fungus capable of hydrolyzing cellulose and hemicellulose and utilizes that as the energy source during the unavailability of proper carbon source, especially when the temperature rises and the soil becomes dry. It is possible to increase the production of cellulase by fivefold as compared to the parent strain just by disrupting 3–4 genes involved in the biosynthetic pathway (Liu et al. 2017). The CRISPR-Cas9 system is currently less explored in plant-beneficiary fungi, but most certainly that can be expanded to achieve better agricultural productivity and yields.

12.5.2 CRISPR-Cas9 as Antiviral Agent

Viruses are small particles that can infect all types of organisms. Oftentimes, it causes diseases and affects the health of plants and animals. According to the World Health Organization, HIV infects the life of about 35 million people with over 70 million cases of infection reported until 2017 (WHO 2018). Lifelong antiretroviral therapy (ART) may help manage the disease to some extent, but it cannot cure the disease completely. The virus delineates its infective properties by inserting its genes into the host genome which then continues to replicate latently. The main targets of the virus are the CD4+ cells, macrophages, and follicular dendritic cells. Restricting the spread of the infection is pretty challenging because the available antiviral

compounds fail to target the integrated proviral genome and the viruses are quick to rebound after ART cessation. Other than this, the viruses even tend to hide into tissue spaces of the central nervous system. Designing an RNA-guided CRISPR-Cas9 tool to target the regulatory genes of HIV-1 can be an effective solution. A lentiviral vector mode of transduction is a process where the gRNA cloned into lentiCRISPRv2 has been used to specifically target the regulatory genes *tat* and *rev* (Ophinni et al. 2018). Transduction of the tool into 293T and HeLa cell lines successfully eliminated the stably expressing Tat and Rev proteins. As a result, the functional assay of *tat* and *rev* genes revealed a significant reduction in the level of HIV-1 promoter-driven luciferase expression and inhibition of gp120 activity (Ophinni et al. 2018). Genome editing through CRISPR-Cas9 has also been used for herpes viruses (Chen et al. 2018) including the herpes simplex virus 1 and 2 (HSV-1 and HSV-2) (Johnson et al. 2014; Diner et al. 2016; Xu et al. 2016; Wang et al. 2018), cytomegalovirus (CMV) (Bierle et al. 2016), Epstein–Barr virus (EBV) (Kanda et al. 2016), and Kaposi's sarcoma herpesvirus (KSHV) (Avey et al. 2015; van Diemen and Lebbink 2017).

Annually, a huge sum of money is dissipated because of the loss of agriculturally important crops owing to viral infections. This matter is, therefore, a serious hurdle in assuring food security for the growing world population (Andolfo et al. 2016). A possible solution to this is to engineer the genome of host plants so as to improve their resistance against plant viruses (Khatodia et al. 2017). The CRISPR-Cas9 technology has the potential to serve as a novel antiviral agent for the protection of plants (Zhang et al. 2015). The CRISPR-Cas9-mediated virus resistance is broadly divided into two approaches: one is where the viral factors concerning the viral genome are targeted, while the other is where the host factors involved in supporting the viral cycle are meant to be targeted. However, using the CRISPR-Cas9 to target viral genes has been so far restricted just to the model species *Tobacco* and *Arabidopsis* (Khatodia et al. 2017).

Introducing the mutations at the attacking site of the virus through CRISPR-Cas9 protects the herbaceous plant *Nicotiana benthamiana* against the beet severe curly top virus (BSCTV) (Ji et al. 2015). Similarly, the resistance to bean yellow dwarf virus (BeYDV) has been also achieved by specifically knocking out the viral replication initiator protein (*Rep*) gene in transgenic *N. benthamiana* plants (Baltes et al. 2015). A broad spectrum resistance to a series of geminivirus including the tomato yellow leaf curl virus (TYLCV), beet curly top virus (BCTV), and Merremia mosaic virus (MeMV) is possible on Cas9-gRNA-mediated editing of the viral coat protein genes, Rep protein, and its conserved intergenic region (IR) (Ali et al. 2015). Thus, it can be said that CRISPR-Cas9 system has presented a number of ways to eradicate animal and plant viruses. However, with the limitless interactions of macromolecule found in the nature, more studies about them would definitely help to favor the efficient removal of plant pathogenic virus through Cas9-dependent arsenals.

12.5.3 Genome Editing of Parasites Using CRISPR-Cas9 System

The CRISPR-Cas9 tool has been implemented for the genome editing of a number of parasites including *Toxoplasma gondii* and *Plasmodium falciparum* (Ghorbal et al. 2014; Kuang et al. 2017; Payungwoung et al. 2018). On the other hand, CRISPR-Cas9 has made its way into the genome editing of *Trypanosoma cruzi* and *Leishmania*. Expressing the Cas9 endonucleases under the control of dihydrofolate reductase–thymidylate synthase (DHFR-TS) promoter and placing sgRNA under the direct control of U6snRNA promoter and terminator give rise to null mutants in *Leishmania* parasites (Sollelis et al. 2015). Another popular example is the Chagas disease-causing *T. cruzi* (Bern et al. 2011). These parasites spread through the biting of insects called Triatominae, commonly known as "kissing bugs." Knocking out genes (*Pfr1*, Pfr2, and *Gp72*) that are the key components of this particular parasite's flagellum revealed their association with flagellar attachment and cell motility (Lander et al. 2015). Repressing the expression of β-galactofuranosyl glycosyltransferase family of enzymes by multiplexing CRISPR-Cas9 in *T. cruzi* is another approach to reduce the outcome of the enzymatic product. Such kind of CRISPR-Cas9-based approaches may help to determine the drug and vaccine targets designed against kinetoplastid parasites (Chiurillo et al. 2017).

Another classic example of CRISPR-Cas9-mediated pest control is of *Plutella xylostella*. Popularly known as Diamondback moth, *P. xylostella* is responsible for damaging cruciferous crops (cauliflower, broccoli, cabbage, Brussels sprouts, etc.). Targeting the abdominal-A moth gene (*Pxabd-A*) involved in characterization and functioning of the abdominal segment results in inheritable defects and malformation of appendages in both sexes (Huang et al. 2016). Currently, CRISPR-Cas9 is less explored in plant pathogens. However, it can be further expanded in a wide range of plant pathogens such as fungus, bacteria, and viruses for controlling and managing diseases that allow us to improve the crop productivity.

12.5.4 CRISPR-Cas9 System for Removal of Bacteria

The Gram-negative bacterium *Escherichia coli* is among the most extensively studied organism from the genome editing perspective. *E. coli* is associated with 70–95% of urinary tract infections (UTIs), delineating its pathogenesis by forming a biofilm on the inner surfaces of the indwelling urinary catheter (Kucheria et al. 2005). Another member of the *Enterobacteriaceae* family, *Klebsiella pneumoniae*, is known to behave in a similar manner as *E. coli*. Both the uropathogenic strains of *E. coli* and *K. pneumoniae* trigger the catheter-associated urinary tract infections (CAUTIs), which is a very common nosocomial infection. The ability of these microbes to form biofilm over biotic and abiotic surfaces is principally regulated through a phenomenon called quorum sensing (QS). QS is the mechanism in which the bacterium establishes the cell-to-cell communication, senses the bacterial population, and regulates its gene expression accordingly (Rutherford and Bassler 2012; Gohil et al. 2018).

Once the QS mechanism is activated, the bacteria release the signaling molecules called autoinducers (AIs) into the intra- and extracellular environment. Once a threshold of AIs in the extracellular environment is reached, the microbes upregulate the biofilm formation or shape their protein expression accordingly (Sturbelle et al. 2015). To control this interaction, the *Lux* family of genes has been targeted through the CRISPR-Cas9 system. Of the many QS pathways involved, such as LuxR-SdiA, LuxS/AI-2, AI-3, and indole system, the LuxS/AI-2 system is reported to be directly linked to the central metabolism of *E. coli*, while the AI-2 is known to be involved in initiating the biofilm formation (De Keersmaecker et al. 2006). A precise deletion of the involved AI-2-dependent *LuxS* gene through CRISPR-Cas9 tool results in the downregulation of biofilm production (Kang et al. 2017).

The broad spectrum antibiotics tend to relentlessly kill the gut commensals. A possible solution to this problem is to design RNA-guided nucleases that distinctly target DNA sequences matching the organism of interest. The sgRNA-driven CRISPR-Cas9 plasmid, introduced into the bacterial population via bacterial or bacteriophage-based delivery, is designed in such a manner that selectively knocks down any of the undesirable gene, which may include the genes that confer virulence or those involved in antibiotic resistance. The tool works well for targeting the carbapenem-resistant *Enterobacteriaceae* as well as enterohemorrhagic *E. coli* (Citorik et al. 2014). A similar example of programmable removal of microbe, particularly *Staphylococcus aureus*, has been achieved by targeting sequence-specific guide-RNA-mediated antimicrobial action of CRISPR-Cas9 that snips off the targeted virulence genes in the virulent strains, leaving untouched the avirulent strains (Bikard et al. 2014). More often than not, the antibiotic-resistant genes reside within the inherent plasmids and are transferred between the strains through the exchange of such promiscuous plasmids. The abovementioned CRISPR-Cas9 system specifically targets and destroys the staphylococcal plasmids bearing antibiotic-resistant genes and prevents its spread among the avirulent staphylococcal strains (Bikard et al. 2014).

So far, the employment of the gRNA driven Cas9-mediated removal of bacterial pathogens that infect plants has been limited. However, by referring to the above examples, it can be stated that the Cas9 shows the potential of eliminating the undesired pathogens or their toxic and virulence property from a given environment. It would take time to reach at a reliable stage as few barriers need to be overcome. Delivery of the antimicrobial Cas9 vectors and the bacterial resistance (Pursey et al. 2018) against such vectors are the example of barriers that are to be conquered and require thoughtful consideration.

12.6 Conclusion and Future Remarks

Since its development from the early 2013, the CRISPR-Cas9 technique has been applied to a vast variety of biological studies. Compared to conventional transgenic techniques, CRISPR-Cas9 is undoubtedly an accurate and constructive way of genome editing. Therefore, employing the CRISPR-Cas9 system to manipulate the

genome of the model strains may help to overcome the pathogenicity and multiplication of the targeted organism or to derive important bioactive compounds from them. This may in turn help to speed up either the development of programmed novel strains with improved efficiency or to knockout undesirable genes. Even though gRNA-mediated genome correction is in its infancy for some microorganisms, this technology definitely promises a better future with functional benefits. The CRISPR-Cas9 system has a great future ahead in plant biotechnology for controlling microbial pathogens and allowing one to solubilize the complex nutrients into a simpler form which can be easily made available to plants in order to increase productivity and yields.

Acknowledgments This work was supported by Puri Foundation for Education in India.

References

Adli M (2018) The CRISPR toolkit for genome editing and beyond. Nat Commun 9(1):1911

Ali Z, Abulfaraj A, Idris A, Ali S, Tashkandi M, Mahfouz MM (2015) CRISPR/Cas9-mediated viral interference in plants. Genome Biol 16:238

Altenbuchner J (2016) Editing of the *Bacillus subtilis* genome by the CRISPR-Cas9 system. Appl Environ Microbiol 82(17):5421–5427

Andolfo G, Iovieno P, Frusciante L, Ercolano MR (2016) Genome-editing technologies for enhancing plant disease resistance. Front Plant Sci 7:1813

Avey D, Tepper S, Li W, Turpin Z, Zhu F (2015) Phosphoproteomic analysis of KSHV-infected cells reveals roles of ORF45-activated RSK during lytic replication. PLoS Pathog 11 (7):1004993

Baltes NJ, Hummel AW, Konecna E, Cegan R, Bruns AN, Bisaro DM, Voytas DF (2015) Conferring resistance to geminiviruses with the CRISPR–Cas prokaryotic immune system. Nat Plants 1:15145

Barrios-González J, Miranda RU (2010) Biotechnological production and applications of statins. Appl Microbiol Biotechnol 85(4):869–883

Bern C, Kjos S, Yabsley MJ, Montgomery SP (2011) *Trypanosoma cruzi* and Chagas' disease in the United States. Clin Microbiol Rev 24(4):655–681

Bierle CJ, Anderholm KM, Wang JB, McVoy MA, Schleiss MR (2016) Targeted mutagenesis of guinea pig cytomegalovirus using CRISPR/Cas9-mediated gene editing. J Virol 90 (15):6989–6998. https://doi.org/10.1128/JVI.00139-16

Bikard D, Marraffini LA (2012) Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. Curr Opin Immunol 24(1):15–20

Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA (2013) Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res 41(15):7429–7437

Bikard D, Euler CW, Jiang W, Nussenzweig PM, Goldberg GW, Duportet X, Fischetti V, Marraffini LA (2014) Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol 32(11):1146–1150

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, Van Der Oost J (2008) Small CRISPR RNAs guide antiviral defence in prokaryotes. Science 321(5891):960–964

Burgio G (2018) Redefining mouse transgenesis with CRISPR/Cas9 genome editing technology. Genome Biol 19(1):27

Carr PA, Church GM (2009) Genome engineering. Nat Biotechnol 27(12):1151–1162

Carroll D (2017) Genome Editing: Past, Present, and Future. Yale J Biol Med 90(4):653–659
Carte J, Christopher RT, Smith JT, Olson S, Barrangou R, Moineau S, Glover CV III, Graveley BR, Terns RM, Terns MP (2014) The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. Mol Microbiol 93 (1):98–112
Chen J, Lai Y, Wang L, Zhai S, Zou G, Zhou Z, Cui C, Wang S (2017) CRISPR/Cas9-mediated efficient genome editing via blastospore-based transformation in entomopathogenic fungus *Beauveria bassiana*. Sci Rep 7:45763
Chen YC, Sheng J, Trang P, Liu F (2018) Potential application of the CRISPR/Cas9 system against herpesvirus infections. Viruses 10(6):291
Chiurillo MA, Lander N, Bertolini MS, Storey M, Vercesi AE, Docampo R (2017) Different roles of mitochondrial calcium uniporter complex subunits in growth and infectivity of *Trypanosoma cruzi*. MBio 8(3):00574–00517
Cho H, Uehara T, Bernhardt TG (2014) Beta-lactam antibiotics induce a lethal malfunctioning of the bacterial cell wall synthesis machinery. Cell 159(6):1300–1311
Chylinski K, Le Rhun A, Charpentier E (2013) The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. RNA Biol 10(5):726–737
Chylinski K, Makarova KS, Charpentier E, Koonin EV (2014) Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res 42(10):6091–6105
Citorik RJ, Mimee M, Lu TK (2014) Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat Biotechnol 32(11):1141–1145
Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini L, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339 (6121):819–823
De Keersmaecker SC, Sonck K, Vanderleyden J (2006) Let LuxS speak up in AI-2 signaling. Trends Microbiol 14(3):114–119
Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471(7340):602–607
Deng H, Gao R, Liao X, Cai Y (2017) CRISPR system in filamentous fungi: Current achievements and future directions. Gene 627:212–221
DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM (2013) Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. Nucleic Acids Res 41(7):4336–4343
Dillingham MS, Kowalczykowski SC (2008) RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev 72(4):642–671
Diner BA, Lum KK, Toettcher JE, Cristea IM (2016) Viral DNA sensors IFI16 and Cyclic GMP-AMP synthase possess distinct functions in regulating viral gene expression, immune defenses, and apoptotic responses during herpesvirus infection. MBio 7(6):01553–01516
Duina AA, Miller ME, Keeney JB (2014) Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. Genetics 197(1):33–48
Fang Y, Tyler BM (2016) Efficient disruption and replacement of an effector gene in the oomycete *Phytophthora sojae* using CRISPR/Cas9. Mol Plant Pathol 17(1):127–139
Fuller KK, Chen S, Loros JJ, Dunlap JC (2015) Development of the CRISPR/Cas9 system for targeted gene disruption in *Aspergillus fumigatus*. Eukaryot Cell 14(11):1073–1080
Gantz VM, Jasinskiene N, Tatarenkova O, Fazekas A, Macias VM, Bier E, James AA (2015) Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. Proc Natl Acad Sci U S A 112(49):6736–6743
Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468(7320):67–71
Garrett RA, Shah SA, Erdmann S, Liu G, Mousaei M, León-Sobrino C, Peng W, Gudbergsdottir S, Deng L, Vestergaard G, Peng X (2015) CRISPR-Cas adaptive immune systems of the sulfolobales: Unravelling their complexity and diversity. Life 5(1):783–817

Gasiunas G, Barrangou R, Horvath P, Siksnys V (2012) Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A 109(39):2579–2586

Ghorbal M, Gorman M, Macpherson CR, Martins RM, Scherf A, Lopez-Rubio JJ (2014) Genome editing in the human malaria parasite *Plasmodium falciparum* using the CRISPR-Cas9 system. Nat Biotechnol 32(8):819–821

Global Health Observatory (GHO) data: HIV/AIDS (2018) World Health Organization http://www.who.int/gho/hiv/en/. Accessed 4 Oct 2018

Gohil N, Ramírez-García R, Panchasara H, Patel S, Bhattacharjee G, Singh V (2018) Book Review: Quorum Sensing vs. Quorum Quenching: A Battle With No End in Sight. Front Cell Infect Microbiol 8:106

Gootenberg JS, Abudayyeh OO, Kellner MJ, Joung J, Collins JJ, Zhang F (2018) Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science 360 (6387):439–444

Guo JC, Tang YD, Zhao K, Wang TY, Liu JT, Gao JC, Chang XB, Cui HY, Tian ZJ, Cai XH, An TQ (2016) Highly efficient CRISPR/Cas9-mediated homologous recombination promotes the rapid generation of bacterial artificial chromosomes of pseudorabies virus. Front Microbiol 7:2110

Gupta RM, Musunuru K (2014) Expanding the genetic editing toolkit: ZFNs, TALENs, and CRISPR-Cas9. J Clin Invest 124(10):4154–4161

Hamad B (2010) The antibiotics market. Nat Rev Drug Discov 9(9):675–676

Harris LJ, Balcerzak M, Johnston A, Schneiderman D, Ouellet T (2016) Host-preferential *Fusarium graminearum* gene expression during infection of wheat, barley, and maize. Fungal Biol Rev 120(1):111–123

Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA (2015) Cas9 specifies functional viral targets during CRISPR–Cas adaptation. Nature 519(7542):199–202

Hisano Y, Sakuma T, Nakade S, Ohga R, Ota S, Okamoto H, Yamamoto T, Kawahara A (2015) Precise in-frame integration of exogenous DNA mediated by CRISPR/Cas9 system in zebrafish. Sci Rep 5:8841

Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell 157(6):1262–1278

Huang Y, Chen Y, Zeng B, Wang Y, James AA, Gurr GM, Yang G, Lin X, Huang Y, You M (2016) CRISPR/Cas9 mediated knockout of the abdominal-A homeotic gene in the global pest, diamondback moth (*Plutella xylostella*). Insect Biochem Mol Biol 75:98–106

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A (1987) Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli,* and identification of the gene product. J Bacteriol 169(12):5429–5433

Jakočiūnas T, Bonde I, Herrgård M, Harrison SJ, Kristensen M, Pedersen LE, Jensen MK, Keasling JD (2015) Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. Metab Eng 28:213–222

Ji X, Zhang H, Zhang Y, Wang Y, Gao C (2015) Establishing a CRISPR–Cas-like immune system conferring DNA virus resistance in plants. Nat Plants 1:15144

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 337 (6096):816–821

Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science 343(6176):1247997

Johnson KE, Bottero V, Flaherty S, Dutta S, Singh VV, Chandran B (2014) IFI16 restricts HSV-1 replication by accumulating on the hsv-1 genome, repressing HSV-1 gene expression, and directly or indirectly modulating histone modifications. PLoS Pathog 10(11):1004503

Kanda T, Furuse Y, Oshitani H, Kiyono T (2016) Highly efficient CRISPR/Cas9-mediated cloning and functional characterization of gastric cancer-derived Epstein-Barr virus strains. J Virol 90 (9):4383–4393 90(9):4383-4393

Kang S, Kim J, Hur JK, Lee SS (2017) CRISPR-based genome editing of clinically important *Escherichia coli* SE15 isolated from indwelling urinary catheters of patients. J Med Microbiol 66(1):18–25

Karginov FV, Hannon GJ (2010) The CRISPR system: small RNA-guided defence in bacteria and archaea. Mol Cell 37(1):7–19

Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V (2013) crRNA and tracrRNA guide Cas9-mediated DNA interference in *Streptococcus thermophilus*. RNA Biol 10(5):841–851

Katayama T, Tanaka Y, Okabe T, Nakamura H, Fujii W, Kitamoto K, Maruyama JI (2016) Development of a genome editing technique using the CRISPR/Cas9 system in the industrial filamentous fungus *Aspergillus oryzae*. Biotechnol Lett 38(4):637–642

Khambhati K, Bhattacharjee G, Singh V (2018) Current progress in CRISPR-based diagnostic platforms. J Cell Biochem. https://doi.org/10.1002/jcb.27690

Khatodia S, Bhatotia K, Tuteja N (2017) Development of CRISPR/Cas9 mediated virus resistance in agriculturally important crops. Bioengineered 8(3):274–279

Knott GJ, Doudna JA (2018) CRISPR-Cas guides the future of genetic engineering. Science 361 (6405):866–869

Koonin EV, Makarova KS, Zhang F (2017) Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol 37:67–78

Kuang D, Qiao J, Li Z, Wang W, Xia H, Jiang L, Dai J, Fang Q, Dai X (2017) Tagging to endogenous genes of *Plasmodium falciparum* using CRISPR/Cas9. Parasit Vectors 10(1):595

Kucheria R, Dasgupta P, Sacks SH, Khan MS, Sheerin NS (2005) Urinary tract infections: new insights into a common problem. Postgrad Med J 81(952):83–86

Kuivanen J, Wang YMJ, Richard P (2016) Engineering *Aspergillus niger* for galactaric acid production: elimination of galactaric acid catabolism by using RNA sequencing and CRISPR/Cas9. Microb Cell Factories 15(1):210

Kumar A, Jha A (2017) Antifungals used against candidiasis. In: Kumar A, Jha A (eds) Anticandidal agents. Academic Press/Elsevier, New York, pp 11–39. ISBN 9780128113110

La Russa MF, Qi LS (2015) The new state of the art: CRISPR for gene activation and repression. Mol Cell Biol 35(22):3800–3809

Lander N, Li ZH, Niyogi S, Docampo R (2015) CRISPR/Cas9-induced disruption of paraflagellar rod protein 1 and 2 genes in *Trypanosoma cruzi*reveals their role in flagellar attachment. MBio 6 (4):01012–01015

Lecellier A, Gaydou V, Mounier J, Hermet A, Castrec L, Barbier G, Ablain W, Manfait M, Toubas D, Sockalingum GD (2015) Implementation of an FTIR spectral library of 486 filamentous fungi strains for rapid identification of molds. Food Microbiol 45:126–134

Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R (2015) CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature 520 (7548):505–510

Liu R, Chen L, Jiang Y, Zhou Z, Zou G (2015) Efficient genome editing in filamentous fungus *Trichoderma reesei* using the CRISPR/Cas9 system. Cell Discov 1:15007

Liu Q, Chen Y, Li Q, Wu L, Wen T (2017) Dcf1 regulates neuropeptide expression and maintains energy balance. Neurosci Lett 650:1–7

Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 1(1):7

Mali P, Esvelt KM, Church GM (2013) Cas9 as a versatile tool for engineering biology. Nat Methods 10(10):957–963

Matsu-ura T, Baek M, Kwon J, Hong C (2015) Efficient gene editing in *Neurospora crassa* with CRISPR technology. Fungal Biol Biotechnol 2(1):4

Miyagishi M, Taira K (2002) U6 promoter–driven siRNAs with four uridine 3′ overhangs efficiently suppress targeted gene expression in mammalian cells. Nat Biotechnol 20(5):497–500
Mojica FJ, Díez-Villaseñor C, García-Martínez J, Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155 (3):733–740
Mosberg JA, Gregg CJ, Lajoie MJ, Wang HH, Church GM (2012) Improving lambda red genome engineering in Escherichia coli via rational removal of endogenous nucleases. PloS ONE 7 (9):44638
Nielsen J, Fussenegger M, Keasling J, Lee SY, Liao JC, Prather K, Palsson B (2014) Engineering synergy in biotechnology. Nat Chem Biol 10(5):319–322
Ophinni Y, Inoue M, Kotaki T, Kameoka M (2018) CRISPR/Cas9 system targeting regulatory genes of HIV-1 inhibits viral replication in infected T-cell cultures. Sci Rep 8(1):7784
Pawluk A, Amrani N, Zhang Y, Garcia B, Hidalgo-Reyes Y, Lee J, Edraki A, Shah M, Sontheimer EJ, Maxwell KL, Davidson AR (2016) Naturally occurring off-switches for CRISPR-Cas9. Cell 167(7):1829–1838
Payungwoung T, Shinzawa N, Hino A, Nishi T, Murata Y, Yuda M, Iwanaga S (2018) CRISPR/Cas9 system in *Plasmodium falciparum* using the centromere plasmid. Parasitol Int 67 (5):605–608
Pohl C, Kiel JAKW, Driessen AJM, Bovenberg RAL, Nygard Y (2016) CRISPR/Cas9 based genome editing of *Penicillium chrysogenum*. ACS Synth Biol 5(7):754–764
Pursey E, Sünderhauf D, Gaze WH, Westra ER, van Houte S (2018) CRISPR-Cas antimicrobials: Challenges and future prospects. PLoS Pathog 14(6):1006990
Rath D, Amlinger L, Rath A, Lundgren M (2015) The CRISPR-Cas immune system: biology, mechanisms and applications. Biochimie 117:119–128
Rauch BJ, Silvis MR, Hultquist JF, Waters CS, McGregor MJ, Krogan NJ, Bondy-Denomy J (2017) Inhibition of CRISPR-Cas9 with bacteriophage proteins. Cell 168(1–2):150–158
Roberts RJ (2005) How restriction enzymes became the workhorses of molecular biology. Proc Natl Acad Sci U S A 102(17):5905–5908
Rutherford ST, Bassler BL (2012) Bacterial quorum sensing: its role in virulence and possibilities for its control. Cold Spring Harb Perspect Med 2(11):012427
Sajith S, Priji P, Sreedevi S, Benjamin S (2016) An overview on fungal cellulases with an industrial perspective. J Nutr Food Sci 6(1):461
Schuster M, Schweizer G, Reissmann S, Kahmann R (2016) Genome editing in Ustilagomaydis using the CRISPR–Cas system. Fungal Genet Biol 89:3–9
Shabbir MA, Hao H, Shabbir MZ, Wu Q, Sattar A, Yuan Z (2016) Bacteria vs. bacteriophages: parallel evolution of immune arsenals. Front Microbiol 7:1292
Shah SA, Erdmann S, Mojica FJ, Garrett RA (2013) Protospacer recognition motifs: mixed identities and functional diversity. RNA Biol 10(5):891–899
Shi TQ, Liu GN, Ji RY, Shi K, Song P, Ren LJ, Huang H, Ji XJ (2017) CRISPR/Cas9-based genome editing of the filamentous fungi: the state of the art. Appl Microbiol Biotechnol 101 (20):7435–7443
Shin J, Jiang F, Liu JJ, Bray NL, Rauch BJ, Baik SH, Nogales E, Bondy-Denomy J, Corn JE, Doudna JA (2017) Disabling Cas9 by an anti-CRISPR DNA mimic. Sci Adv 3(7):e1701620
Shokri J, Adibkia K (2013) Application of cellulose and cellulose derivatives in pharmaceutical industries. In: van de Ven T, Godbout L (eds) Cellulose-medical, pharmaceutical and electronic applications. InTechOpen. https://doi.org/10.5772/55178
Singh V, Braddick D, Dhar PK (2017) Exploring the potential of genome editing CRISPR-Cas9 technology. Gene 599:1–18
Singh V, Gohil N, Ramírez-García R, Braddick D, Fofié CK (2018) Recent advances in CRISPR-Cas9 genome editing technology for biological and biomedical investigations. J Cell Biochem 119(1):81–94
Sollelis L, Ghorbal M, MacPherson CR, Martins RM, Kuk N, Crobu L, Bastien P, Scherf A, Lopez-Rubio JJ, Sterkers Y (2015) First efficient CRISPR-Cas9-mediated genome editing in *Leishmania* parasites. Cell Microbiol 17(10):1405–1412

Sternberg SH, LaFrance B, Kaplan M, Doudna JA (2015) Conformational control of DNA target cleavage by CRISPR–Cas9. Nature 527(7576):110–113

Sturbelle RT, de Avila LFDC, Roos TB, Borchardt JL, Dellagostin OA, Leite FPL (2015) The role of quorum sensing in *Escherichia coli* (ETEC) virulence factors. Vet Microbiol 180 (3-4):245–252

Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T, Pschera P, Siksnys V, Seidel R (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc Natl Acad Sci U S A 111(27):9798–9803

Thrane U, Anderson B, Frisvad JC, Smedsgaard J (2007) The exo-metabolome in filamentous fungi. In: Nielsen J, Jewett MC (eds) Metabolomics. (Topics in current genetics), vol 18. Springer, Berlin, pp 235–252

Umesha S, Singh P, Singh R (2017) Microbiology biotechnology and sustainable agriculture. In: Singh RL, Mondal S (eds) Biotechnology for sustainable agriculture: emerging approaches and strategies. Woodhead Publishing, Cambridge, UK, pp 185–205

Van Diemen FR, Lebbink RJ (2017) CRISPR/Cas9, a powerful tool to target human herpesviruses. Cell Microbiol 19(2). https://doi.org/10.1111/cmi.12694

Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y (2015) Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. Cell 163(4):840–853

Wang D, Wang XW, Peng XC, Xiang Y, Song SB, Wang YY, Chen L, Xin VW, Lyu YN, Ji J, Ma ZW (2018) CRISPR/Cas9 genome editing technology significantly accelerated herpes simplex virus research. Cancer Gene Ther 25(5-6):93–105

Ward OP (2012) Production of recombinant proteins by filamentous fungi. Biotechnol Adv 30 (5):1119–1139

Weber J, Valiante V, Nødvig CS, Mattern DJ, Slotkowski RA, Mortensen UH, Brakhage AA (2016) Functional reconstitution of a fungal natural product gene cluster by advanced genome editing. ACS Synth Biol 6(1):62–68

Wei Y, Chesne MT, Terns RM, Terns MP (2015) Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. Nucleic Acids Res 43 (3):1749–1758

Wiles MV, Qin W, Cheng AW, Wang H (2015) CRISPR–Cas9-mediated genome editing and guide RNA design. Mamm Genome 26(9-10):501–510

Woloshuk CP, Shim WB (2013) Aflatoxins, fumonisins, and trichothecenes: a convergence of knowledge. FEMS Microbiol Rev 37(1):94–109

Wright AV, Nuñez JK, Doudna JA (2016) Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. Cell 164(1-2):29–44

Xu X, Fan S, Zhou J, Zhang Y, Che Y, Cai H, Wang L, Guo L, Liu L, Li Q (2016) The mutated tegument protein UL7 attenuates the virulence of herpes simplex virus 1 by reducing the modulation of α-4 gene transcription. Virol J 13:152

Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. Proc Natl Acad Sci U S A 97(11):5978–5983

Zhang D, Li Z, Li JF (2015) Genome editing: new antiviral weapon for plants. Nat Plants 1 (10):15146

Zhang C, Meng X, Wei X, Lu L (2016) Highly efficient CRISPR mutagenesis by microhomology-mediated end joining in *Aspergillus fumigatus*. Fungal Genet Biol 86:47–57

Biosorption-Cum-Bioaccumulation of Heavy Metals from Industrial Effluent by Brown Algae: Deep Insight

13

Priyanka Yadav, Jyoti Singh, and Vishal Mishra

Abstract

Biosorption by marine brown algae is considered to be very effective as the brown algae is found in diverse size and are having better efficiency in removing the heavy metals form wastewater which is one of the most critical problem now a days. Many micro and macroalgae are responsible for the recovery of different heavy metals. The brown seaweeds have the highest sorption capacity or higher rate of bioaccumulation for heavy metal ion than that of red and green seaweeds. Marine algae are fast growing algae and can perform relatively better as it requires a small amount of nutrients, CO_2 and sunlight for its survival. The present literature covers the biosorption by marine algae mainly the brown algae which can be used all around the year. The carboxyl acid group present in these biomass is found to be the most dominant as well as most abundant functional group that are followed by fucoidan.

13.1 Introduction

Algae are everywhere on the earth like in rivers, lakes, seas, on soil and walls, in plants and animals (as symbionts-partners collaborating together), or can say every place where lights are present to carry out photosynthesis (El Gamal 2010). The primary producer of the marine food chain is marine microalgae which show the toxic impact on a higher level when the toxicants are consumed by the same (Purbonegoro et al. 2018). Alga is considered to be an abundant and vastly accessible natural resource in a tropical ecosystem. It is observed that the brown algae have better uptake capacity as compared to red and green and considered to be one of the

P. Yadav · J. Singh · V. Mishra (✉)
School of Biochemical Engineering, IIT (BHU) Varanasi, Varanasi, Uttar Pradesh, India
e-mail: vishal.bce@itbhu.ac.in

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_13

best biosorbents for the retrieval of the heavy metals. Alginic acid and fucoidan are present in their cell wall, and at neutral pH, the alginic acid yields sulfate ion as well as carbohydrates (Sweetly 2014). Marine algae like *Sargassum* constitute of diverse multifunctional groups which are present on the surface and have even distribution of binding sites on the cell surface. There are many advantages of marine algae as biosorbent like the requirement of minimal preparatory steps, retention capacity is excellent and truly renewable, recyclable, and simply available all year around.

As biosorption is a passive mechanism, hence this process is faster than that of active or bioaccumulation (Bilal et al. 2018). Algal biosorption attributes the cell wall where complexation, as well as static attraction, plays a major role. Carboxyl group is considered to be the dominating binding groups in brown algae. Brown algae possess alginic acid and fucoidan in the cell wall matrix as well as in intercellular material. Seaweed is considered to be better than that of microbial biomasses because of the less variability in seawater than that of fermentation media. Marine algae carry a large number of biopolymers that are helpful in the metal binding.

Seaweeds are larger enough so that there is no need for any complicated or costly immobilization required (Schiewer and Volesky 2000). Seaweeds are accountable for recovery of heavy metals due to macroscopic structures which provides a basis for the formation of biosorbent particle. The brown alga *Sargassum* consists of gel materials in their call wall termed as alginates that are very porous in nature, are responsible for metal binding, and are also easily permeable to small ionic species. Volesky et al. reported *Sargassum* (brown algae) seaweeds as the best biomass for recovery of heavy metals (Vieira and Volesky 2000). In marine macroalgae, removal of heavy metals is done either by ion exchange on the surface of the cell or by means of intracellular transport of heavy metals (Sweetly 2014). Biosorption by marine algae is considered to be very effective as the marine algae are found in diverse size and are having better efficiency in removing heavy metals from wastewater which is one of the most critical problems nowadays. Numerous micro- and macroalgae are accountable for the recovery of different heavy metals. Marine algae or seaweeds like brown, red, and green have the highest sorption capacity or higher rate of bioaccumulation for the heavy metal ion. Out of these three seaweeds, brown algae are considered more proficient in biosorption. Marine algae are fast-growing algae and can perform relatively better as it requires a small amount of nutrients, CO_2, and sunlight for its survival.

Figure 13.1 shows three types of marine algae (also known as macroalgae) that are responsible for the removal of heavy metals, and Table 13.1 shows the classification of marine algae on the basis of their different characteristics.

13.2 Brown Algae

Brown algae are suitable for removing heavy metals as a consequence of its polysaccharide content (Volesky and Holan 1995). They have the capability to absorb heavy metals owing to chemical groups present on the surface, for instance, sulfonate, amino, carboxyl as well as sulfhydryl (Umar Mustapha and Halimoon 2015).

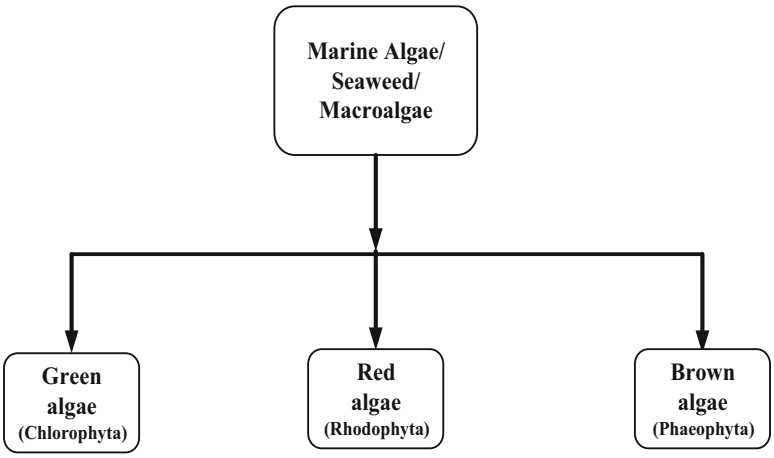

Fig. 13.1 Types of marine algae

Table 13.1 Characteristics of brown, red, and green algae

Characteristics	Brown algae	Red algae	Green algae
Examples	*Fucus*, *Laminaria*, *Macrocystis*, *Sargassum*	*Gracilaria*, *Porphyra*, *Palmaria*, *Chondrus*	*Caulerpa*, *Monostroma*
Habitat	Marine water	Marine water	Marine water
Photosynthetic pigments	Chlorophyll-a and -c, carotenoid, xanthophyll, fucoxanthin	Phycoerythrin, phycocyanin, chlorophyll-a and -d, xanthophyll, carotenoid	Chlorophyll-a and -b, carotenoids
Stored food	Laminarin, mannitol	Floridean starch	Starch
Cell wall	Cellulose and alginate	Cellulose, agar, and carrageen	Cellulose (mostly)
Key functional group for biosorption	Carboxylic group, sulfate	Carboxylic group, hydroxyl, amine, phosphate, C − O and C=O	Carboxylic group
Height	30 cm–30 m long	30 cm–30 m long	30 cm–30 m long

It is one of the most important plant groups that are successfully studied for the biosorption of heavy metals from industrialized wastes.

Owing to the existence of large amounts of carotenoid fucoxanthin that are located in the chloroplast of brown algae, the color becomes brown; these brown algae are grown in marine environments. There are about 13 divisions of Phaeophyta (a division of brown algae); out of which 2 orders called Laminariales and Fucales are most important that are abundantly available in nature. The order Laminariales are known as "kelp" and commercially applicable in the production of syrups, dessert gels, ceramics (for stabilizing property), cleanser, welding rods, and so on. The order Fucales is vast; therefore some of its species are mainly studied for the properties of their biosorption or metal binding ability. Carboxylic groups are

abundantly present in brown algae that are important in the biosorption process by reducing the cadmium and lead uptake (Fourest and Volesky 1996). After the carboxylic group, sulfonic acid plays a secondary role in metal binding at lower pH value.

Brown seaweed *Sargassum baculari* are useful for the biosorption of copper. Large amounts of seaweeds are harvested from oceans and can be further cultured for phycocolloid or food production. Lots of seaweeds are used for the testing of its biosorptive properties in the laboratory as well as on large-scale operations that can be easily conducted by the help of well-established activated carbon fixed-bed system. The process equipments and design procedure are already available, that's why it is beneficial to implement the operations on fixed-bed configurations. The investigation on biosorption of copper with brown seaweeds *Sargassum baccularia* immobilized onto polyvinyl alcohol (PVA) gel beads in fixed-bed experiment (Chu and Hashim 2007). The immobilization of seaweeds by PVA is done as it is easily accessible and inexpensive and possesses the best abrasion resistance properties. Chu et al. concluded that the immobilization process of seaweed biomass in PVA gel was suited for removal of toxic metals like that of copper in fixed-bed column operations as biosorbent exhibited favorable regeneration conditions and also the biosorption capacity remains unchanged all through three cycles of biosorption-desorption successfully (Chu and Hashim 2007).

In the recent years, detection of a huge number of heavy metal and low-cost sorbents but brown algae is recognized as the most promising as well as the most effective substrate for remediation of M^+ (metal ion) (Davis et al. 2003). The marine environment is considered to be the available source of the antimicrobial compounds as numerous sea organisms yield bioactive metabolites on the development of the chemical strategy and in response to the environmental stress (Maadane et al. 2017). The binding of M^+ on the surface of algae depends upon many factors like algal species and ionic charge of metal ions (Sulaymon 2014). Brown algae are most effective macroalgae because it contains a higher amount of alginate and on the other side, a carboxylic group that is responsible for capturing the cations present in the solution (Manuel et al. 2016). Figure 13.2 depicts the process of removal of heavy metals by metabolite-dependent as well as metabolite-independent phenomena. As we know that for the recovery of heavy metal, functional groups are responsible for binding as commonly the ion-exchange process is done by algae during biosorption. On the other side, bioaccumulation of heavy metals is either transformed or accumulates in the vacuoles or cytoplast of the algae.

Figure 13.3 depicts algin or alginate and fucoidan are mainly present in the outer layer, as well as the inner layer of brown algae due to which it is unique in comparison to the red and green and sorption efficiency and is also better than other ones. Alginic acid is a polymer of guluronic acid, mannuronic acid, salts of sodium, potassium, magnesium, calcium, and sulfated polysaccharides (Davis et al. 2003; Sweetly 2014) that offers sulfate ions and anionic carbohydrate at neutral pH.

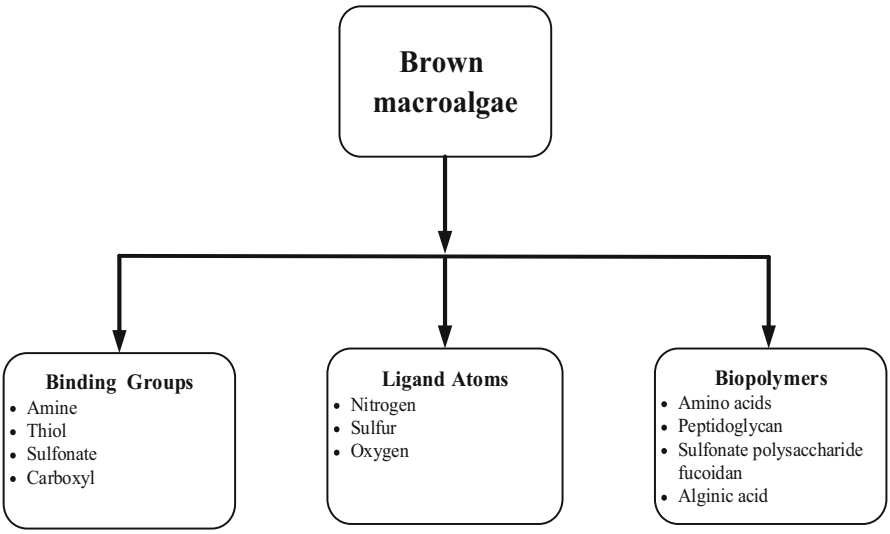

Fig. 13.2 Different binding groups of brown macroalgae

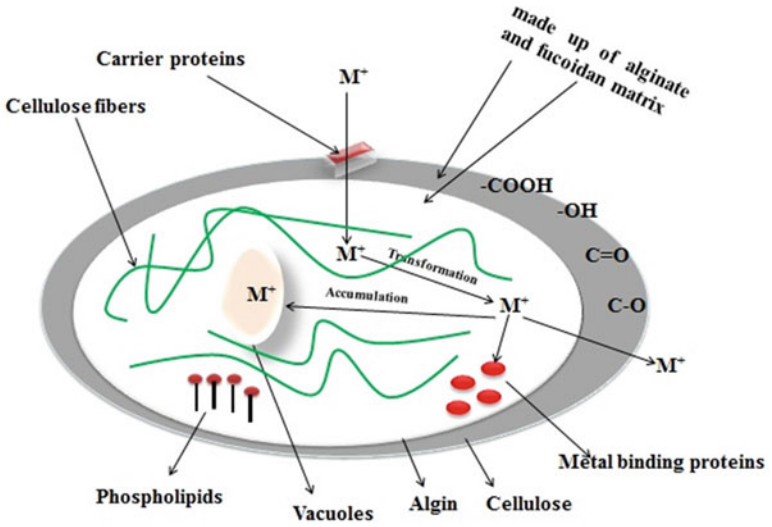

Fig. 13.3 Metal sorption by brown algae

13.3 Metabolic Pathway

The production of metabolites in brown algae was discussed by the help of "amino acid derivatives" and peptides metabolism and "energy and carbohydrate metabolism."

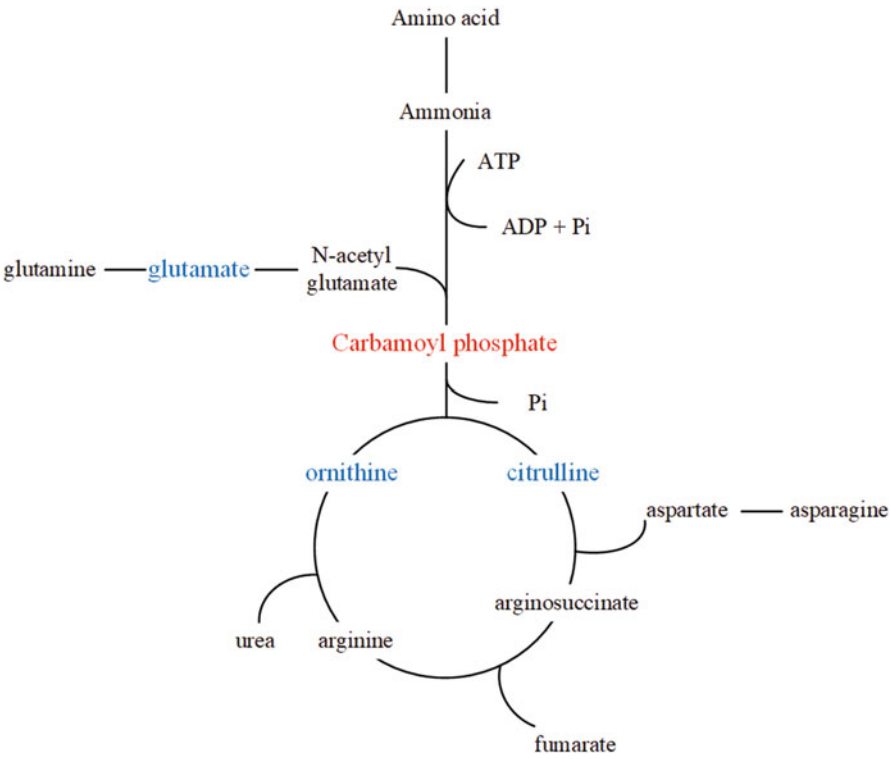

Fig. 13.4 Typical metabolic pathway for the formation of metabolites possessed by brown algae (glutamate, ornithine, and citrulline) that are involved in the urea cycle

13.3.1 Amino Acid Derivatives and Peptides

A study was conducted by Belghit et al. in which 70 compounds possess significant difference in relative abundance among 139 compounds related to amino acid derivatives (Fig. 13.4).

Many of these compounds ascribed to the stress response. Seaweeds are subjected to a variety of biotic and abiotic stress factors, and simultaneously they respond against this by regulating their physiological profile, especially carbon (C) and nitrogen (N) metabolism. N-enriched amino acids such as ornithine, glutamate, and citrulline are found in brown algae that help them to tolerate stress in different stress conditions (Belghit et al. 2017).

13.3.2 Energy and Carbohydrate Metabolism

Metabolites produced during glycolysis except glucose were more abundantly found in brown algae in comparison to red. Several Krebs cycle intermediates (citrate,

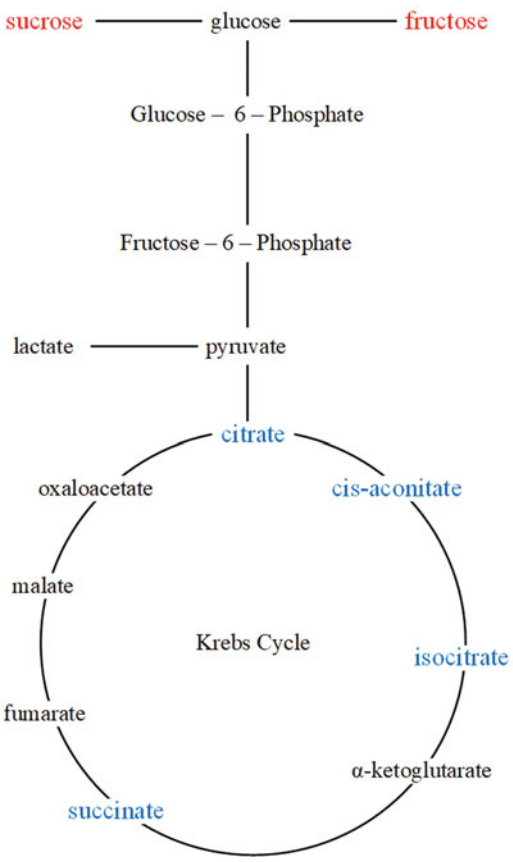

Fig. 13.5 Metabolic pathway of formation of Krebs cycle intermediates abundantly found in brown algae

succinate, cis-aconitate, and isocitrate) are highly concentrated in brown algae that protect them against oxidative stress. The metabolic pathway of their production is demonstrated in Fig. 13.5 (Belghit et al. 2017).

13.4 Heavy Metals and their Toxicity

Nowadays, toxicity of heavy metals is a serious concern that causes a lot of problems. The sources of such metals are mining industries, battery industries, fertilizer and chemical industries, and nuclear power plant. To overcome this serious problem, several methods were conducted like biosorption, bioaccumulation which are responsible for recovery of these toxic metals from industrial runoffs by using numerous biosorbents (Manuel et al. 2016).

Heavy metals like lead (Pb), zinc (Zn), arsenic (As), copper (Cu), mercury (Hg), boron (B), manganese (Mn), aluminum (Al), and nickel (Ni) have density that is five times higher than that of water. As a limited amount of metal is essential for human

Table 13.2 Sources and toxicity of heavy metals

Heavy metals	Toxic effects	Major sources	References
Zinc	Zinc fume causes damage to nervous membrane that possess corrosive effect on the skin	Metal plating, refineries, plumbing, brass manufacture	Alluri et al. (2007)
Arsenic	Dermatitis, bronchitis	Fungicides, pesticides, metal smelters	Alluri et al. (2007)
Lead	Affects nervous system and impairs mammalian spermatogenesis	Automobile batteries, lead batteries, lead paints, herbicides, insecticides	Bilal et al. (2018) and Umar Mustapha and Halimoon (2015)
Cadmium	Carcinogenesis, pulmonary irritation, kidney damage, renal disorder, human carcinogen	Metal industry, paint, food, cigarette smoke, phosphate fertilizers	Umar Mustapha and Halimoon (2015)
Barium	Intestinal problem	Aerospace, coal-burning factories	Shafiq et al. (2018)
Copper	Carcinogenesis, diabetes, liver damage, insomnia, Wilson disease	Electroplating industries, photovoltaic cells, leather industry (tanning), pesticides, fertilizers, electroplating industry	Umar Mustapha and Halimoon (2015)
Nickel	Cardiovascular diseases, chest pain, dizziness, dermatitis, shortness of breath, headache, nausea, lungs, and nasal cancer	Kitchen appliances, paints and powders, superphosphate fertilizers, batteries' processing units, metal refining, galvanization	Umar Mustapha and Halimoon (2015)

health but if consumes in more amount, it causes hazardous effects on living beings (Table 13.2). Copper is considered to be a component of several enzymes and proteins that participate in a different metabolic pathways in algae (Purbonegoro et al. 2018).

13.5 Biosorption

The term biosorption is a subclass of adsorption wherein the biological matrix is sorbent. This process provides an economical, eco-friendly, reversible, and rapid binding of M^+ from solution against functional groups available on the biomass surface. It is not dependent on cellular metabolism. It was found that among all the biosorbents, algae were found to be most appreciable (Michalal et al. 2013). Biosorption is responsible for removal of heavy metals and precious metals (Sweetly 2014). There are several conventional methods like precipitation, coagulation, ion exchange, and membrane separation for removal of heavy metals but consist of several disadvantages like requirements of higher energy, reagents, expensive, toxic

waste product generation, and not effective at low metal concentration. The adsorption mediated through activated carbon is considered to be the most effective method, but it also carries some disadvantages like high cost and non-recyclable in nature (Ramezani et al. 2013). The brown algae are considered to be one of the best biosorbents for recovery of heavy metals (Aziz 2004). It was reported that the green marine macroalgae possess the potential for removal of heavy metals such as Pb, Hg, and Cd from the aqueous solution. It was found that the highest Cd and Pb uptake is done by *Chaetomorpha* species and for Hg, *C. sertularioides* (Kumar and Goyal 2009). The excellent recovery of gold is shown by *Sargassum natans* (a brown alga) as reported in US Patent no. 4,769,223 (Volesky and Kuyucak 1988).

Table 13.3 shows different authors who have investigated different aspects like the usage of free/immobilized biomass in continuous/batch column experiment and includes parameters affecting the process (pH, temperature, and functional group). Thus, marine algae have a carboxylic group that is responsible for the heavy metal recovery.

13.6 Bioaccumulation

As soon as a portion of metal is taken by microorganism, then the process of bioaccumulation takes place. It is an active process wherein metal is metabolically controlled like energy production and transformation (Arunakumara and Zhang 2008). Brown algae show higher accumulation of heavy metals in comparison to green and red algae mainly because of the presence of polysaccharides and polyphenolic substances that constitute their cellular wall (Wallenstein et al. 2009). Brown algae are considered to be one of the better bioaccumulators of heavy metals and can also be used as a universal bioaccumulator (Sweetly 2014). Bioaccumulation is a complex process where metal level must be checked in the tissues from two adjacent tropical levels in animals (Jakimska et al. 2011). *Cladophora herpestica* (green algae) is considered as one of the dominating and abundantly growing on the Maruit lake surface can accumulate residual nutrients in addition to heavy metal ions from both atmospheric and aquatic environments (Al Maghraby and Hassan 2017). It was found that the accumulation of Hg, Cd, Zn, and Ag is done by various marine algae successfully (Fisher et al. 1984). Marine microalgae are considered to be a promising indicator species for inorganic as well as organic pollutants because of their abundance in the marine ecosystem that occupies the base of the food chain (Torres et al. 2008).

Table 13.4 depicts the bioaccumulation of various heavy metals by marine algae. In a report, it was found that the capacity of metal recovery like Hg has been checked by taking three algae, i.e., *Ulva lactuca* (green), *Gracilaria gracilis* (red), and *Fucus vesiculosus* (brown), and the result shows that the green algae have displayed the best performance in the recovery of Hg (Henriques et al. 2015). In a report, it was found that the green algae, i.e., *Ulva lactuca*, has the capability to remove Fe, Mn, Zn, Pb, Cr, and Cd metals except for Cu (Swaleh et al. 2016).

Table 13.3 Factors affecting biosorption of heavy metals using marine algae

Marine algae	Heavy metal	Waste water	Status	Functional group	pH	Temp (°C)	Efficiency	References
*Sargassum vulgare (*brown seaweeds*)*	Fe^{3+}	Coast of Tetouan (NW Morocco)	Dead	Carboxyl, alcohol-OH,-NH, -SO_3	2	60	94%	Benaisa et al. (2018)
Gelidium amansii	Pb^{2+}	General	Dead	Carbonyl, methylene, phosphate,carbonate, phenolic acid	4.5	45	100%	Salah et al. (2018)
Ulva lactuca sp.	Cd(II)	Aqueous solution	Dead	Amido, hydroxyl, C–O and C=O	5	20 ± 0.5	85%	Lupea et al. (2012)
Gelidium species	Cu(II)	–	Dead	–	5.3	–	13 mgg^{-1}	Vilar et al. (2008)
Chlorella vulgaris (green algae)	Cd(II)	General	Dead	_	_	_	98.12%	Cheng et al. (2016a, b)
Ulva fasciata	Pb	General	Dead	Amino, carboxyl, alcoholic groups	_	_	95%	Nessim et al. (2011)
Ulva fasciata (green algae)	Cu(II)	West coast of India and Singapore coast	Dead	(carboxyl groups from mannuronic and guluronic acids)	1–4	60	7.35%	Karthikeyan et al. (2007)
Sargassum sp. (brown algae)	Cu(II)	West coast of India and Singapore coast	Dead	(carboxyl groups from mannuronic and guluronic acids)	1–5	60	7.25%	Karthikeyan et al. (2007)
Sargassum siliquosum (brown algae)	Cd^{2+}	–	Dead	–	–	–	0.73 mmol/g	González et al. (2011)
Gracilaria salicornia (red algae)	Cr(VI)	Persian gulf coast	Dead	–	4	–	45.959 mg g^{-1}	Shams Khorramabadi and Darvishi Cheshmeh Soltani (2008)

Sargassum sp. (brown algae)	Cr(VI)	Persian gulf coast	Dead	–	4	–	33.258 mg g^{-1}	Shams Khorramabadi and Darvishi Cheshmeh Soltani (2008)
Sargassum polycystum ca-loaded biomass	La	–	Dead	–	3	–	(0.8–0.9) mmol g^{-1})	Diniz and Volesky (2005)
Sargassum polycystum ca-loaded biomass	Eu	–	Dead	–	4	–	(0.8–0.9) mmol g^{-1})	Diniz and Volesky (2005)
Sargassum polycystum ca-loaded biomass	Yb	–	Dead	–	5	–	(0.7–0.9) mmol g^{-1})	Diniz and Volesky (2005)
Fucus spiralis (brown algae)	Cd	–	Dead	–	6	–	114.9 mg/g	Romera et al. (2007)
Ascophyllum nodosum (brown algae)	Cd	–	Dead	–	6	–	75.2 mg/g	Romera et al. (2007)
Fucus spiralis (brown algae)	Ni	–	Dead	–	6	–	50 mg/g	Romera et al. (2007)
Sargassum vulgare (brown algae)	Ni^{2+}	–	Dead	–	–	–	0.085 mmol/g	González et al. (2011)
Sargassum vulgare (brown algae)	Pb^{2+}	–	Dead	–	–	–	1.10 mmol/g	González et al. (2011)
Ascophyllum nodosum (brown algae)	Ni	–	Dead	–	6	–	37.2 mg/g	Romera et al. (2007)
Fucus spiralis (brown algae)	Zn	–	Dead	–	6	–	53.2 mg/g	Romera et al. (2007)
Ascophyllum nodosum (brown algae)	Zn	–	Dead	–	6	–	45.7 mg/g	Romera et al. (2007)
Fucus spiralis (brown algae)	Cu	–	Dead	–	4	–	70.9 mg/g	Romera et al. (2007)
Ascophyllum nodosum (brown algae)	Cu	–	Dead	–	4	–	40.5 mg/g	Romera et al. (2007)

(continued)

Table 13.3 (continued)

Marine algae	Heavy metal	Waste water	Status	Functional group	pH	Temp (°C)	Efficiency	References
Fucus spiralis (brown algae)	Pb	–	Dead	–	3	–	204.1 mg/g	Romera et al. (2007)
Ascophyllum nodosum (brown algae)	Pb	–	Dead	–	3	–	204.1 mg/g	Romera et al. (2007)
Durvillaea potatorum	Cu(II)	–	Dead	–	4.5	–	1.3 mmol/g	Matheickal, Yu, and M Woodburn (1999)
Ecklonia radiata	Pb(II)	–	Dead	–	4.5	–	1.3 mmol/g	Matheickal et al. (1999)
Durvillaea potatorum	Pb(II)	–	Dead	–	4.5	–	1.6 mmol/g	Matheickal et al. (1999)
Ecklonia radiata	Cu(II)	–	Dead	–	4.5	–	1.1 mmol/g	Matheickal et al. (1999)
J. rubens (red macroalgae)	Pb^{+2}	–	Dead	Hydroxyl, amine, phosphate, C − O and C=O	5	60	91%	M Ibrahim et al. (2018)
J. rubens (red macroalgae)	Cd^{+2}	–	Dead	Hydroxyl, amine, phosphate, C − O and C=O	5	60	89%	M Ibrahim et al. (2018)
J. rubens (red macroalgae)	Ni^{+2}	–	Dead	Hydroxyl, amine, phosphate, C − O and C=O	5	60	85%	M Ibrahim et al. (2018)
Cystoseira barbata (brown algae)	Ni(II), Cd(II), Pb(II)	–	Dead	–OH groups (primarily phenolic, alcoholic, and carboxylic), $–CH_2$ groups	–	–	1 mmol/g-1.3 mmol//g	Yalcin et al. (2012)

Seaweeds *(Sargassum tenerrimum, Iyengariastellata, Lobophora variegata, Halimeda tuna, Cystoseiraindica, Sargassum cinereum, and Ulva lactuca)*	Cu^{2+}	–	Dead	Hydroxyl, carboxylic, amide, and phosphate group	5	–	60.97 mg/g	Patel et al. (2016)
U. pinnatifida	Cu^{2+}	Sea near Cheju Island, Korea	Dead	–	5.3–4.4	–	2.58 meq/g	Lee et al. (2002)
U. pinnatifida	Pb^{2+}	Sea near Cheju Island, Korea	Dead	–	5.3–4.4	–	2.6 meq/g	Lee et al. (2002)
U. pinnatifida	Zn^{2+}	Sea near Cheju Island, Korea	Dead	–	5.3–4.4	–	2.08 meq/g	Lee et al. (2002)
Dead consortium consisting of *Chlorella* and *Chlamydomonas*	Pb(II)	Biological oxidation pond, Wazirabad, New Delhi	Dead	–	–	–	15.95%	Kumar and Goyal (2009)
Ulva lactuca (green algae)	Cd(II)	–	Dead	Carboxyl, amino, C–O, C=O	5	–	85%	Lupea et al. (2012)

Table 13.4 Bioaccumulation of heavy metals from wastewater using marine algae

Marine algae	Heavy metals	Wastewater	Status	Efficiency	References
Chlorella vulgaris	Cd(II)	General	Live	95.7%	Cheng et al. (2016a, b)
Live consortium consisting of *Chlorella*, *Chlamydomonas*, *Lyngbya* species	Pb(II)	Biological oxidation pond, Wazirabad, New Delhi	Live	33.3 mg/g	Kumar and Goyal (2009)
Oscillatoria bornettia	Zn, Fe, Cu, Cd	_	Live	0.306, 0.302, 0.091, 0.276	Abirhire and Kadiri (2011)
Phacus curvicauda	Al	_	Live	0.439	Abirhire and Kadiri (2011)
U. lactuca	Hg	–	Live	99%	Henriques et al. (2015)
Chlorella vulgaris	Cu, Pb	–	Live	92.53% 98.70%	Luciana et al. (2013)

13.7 Advantages of Algal Biomass Over Conventional Methods

There are numerous advantages of algal biomass as a biosorbent such as used in the wastewater with high metal concentration, unlike the membrane process. Metal uptake capacity and efficiency of metal removal are also high. In addition to this, regeneration of biomass takes place and is cost-effective. These biomasses can be easily reused in different adsorption/desorption cycles. These biomasses can be used all year around. No generation of toxic chemicals takes place. Macroalgal biomass does not need to be immobilized. Few chemicals for the regeneration and desorption of biosorbent are needed. It is suitable for anaerobic as well as aerobic effluent treatment units used in continuous as well as in discontinuous regime for the selectivity of heavy metal ions. During acid treatment of algal biomass, the polysaccharides that are present on the cell wall can dissolve up to some extent, thus able to form additional binding sites generally the amino acids (González et al. 2011). As compared to the microbial biomasses that require immobilization for industrial-scale application, algal biomass can be used without any pretreatment (e.g., biosorption column) (Schiewer and Volesky 2000).

There are a lot of conventional methods used in the recovery of heavy metals like ion exchange, reverse osmosis, precipitation, membrane filtration, filtration, and coagulation, but each and every method has their own drawbacks like higher cost and time-consuming. To overcome all these drawbacks, here comes the process of

biosorption. Biosorption poses numerous advantages over conventional methods together with efficiency, cost-effectiveness, regeneration of biosorbent with the possibility of metal recovery, and requirements of additional nutrients minimization of biological/chemical sludge (Alluri et al. 2007).

13.8 Isotherm Models Used in the Biosorption Process

13.8.1 Langmuir Isotherm

This isotherm assumes a surface having equivalent sorption energies, homogeneous binding sites also, and no interactions between species that are sorbed. In this isotherm, once a site gets filled, then there will be no other sorption at that site (Langmuir 1916).

$$\frac{C_q}{Q_{eq}} = \frac{1}{bQ_{max}} + \frac{C_{eq}}{Q_{max}} \tag{15.1}$$

where

Q_{max}= Maximum amount of metal ions per unit weight of a bio sorbent(mg/g)
b = Langmuir constant that relates to the energy of adsorption

Langmuir isotherm can also be calculated in terms of separation parameters (dimensionless), i.e.,

$$R_L = 1/(1 + bC_0) \tag{15.2}$$

Equation 13.2 indicates the shape of isotherm that helps to predict whether adsorption is favorable or not.

Conditions:
1. Favorable when $0 < R_L < 1$
2. Unfavorable when $R_L > 1$
3. Linear when $R_L = 1$
4. Irreversible when $R_L = 0$

13.8.2 Freundlich Isotherm

These isotherms are applicable to adsorption on the surface that is heterogeneous with the interaction between the molecules that are adsorbed. On the basis of sorption on the heterogeneous surface, the equation is as under:

$$\log Q = logK_f + \frac{1}{n}\, logC_e \tag{15.3}$$

where.

$K_f$and n = Freundlich constant
n = indicator of the degree of nonlinearity between adsorption and concentration of the solution.

Freundlich equilibrium constants are determined by the plots of log Q_{eq}vs. log $C_{eq.}$

Conditions:
1. Linear adsorption occurs when n = 1
2. Physical process adsorption occurs when n > 1
3. Chemical process adsorption occurs when n < 1

13.8.3 Redlich-Peterson Isotherm

It is considered to be a special case of Langmuir when constant g becomes unity. It can be applied on the homogeneous surface or on a heterogeneous surface/system (Abdel-Ghani et al. 2015).

$$\text{Linear form } \ln\left[\left(\frac{AC_e}{q_e}\right) - 1\right] = g\,\ln(C_e) + \ln(B) \tag{15.4}$$

where A, B, and g($0 < g < 1$) that are represented are isotherm constant

At higher concentration, the isotherm equation reduced to form Freundlich isotherm, and when g = 1, then it reduced to Langmuir isotherm.

13.8.4 Sip Isotherm

An empirical formula was proposed by Sip and also termed as Langmuir-Freundlich isotherm, which is often represented as (Abdel-Ghani et al. 2015):

$$q_e = \frac{K_s C_e^{n_s}}{1 + a_s C_e^{n_s}} \tag{15.5}$$

where

K_s = Sip's constant /affinity constant (Lmg^{-1})
n_s = heterogeneity coefficient

At higher concentration of sorbate, Sip isotherm predicts as Langmuir isotherm.

At lower concentration of sorbate, Sip isotherm reduced to Freundlich isotherm and did not obeys Henry law.

13.8.5 Temkin Isotherm

This isotherm provides an equal distribution of binding energies on various exchange sites on the surface (Abdel-Ghani et al. 2015).

$$\text{Linear Temkin isotherm} \quad q_e = B \ln A + b \ln C_e \tag{15.6}$$

where

B = RT/b
R = universal gas constant (8.314 Lmol^{-1} K^{-1})
T = absolute temperature in Kelvin
B = heat of sorption
A = equilibrium binding constant

13.8.6 Dubinin-Radushkevich (D-R) Isotherm

This isotherm is a semi-empirical equation under which the adsorption follows a mechanism of pore filling. It is applicable to the process of physical adsorption and consists of van der Waals forces (Abdel-Ghani et al. 2015).

$$\text{Linear form} \quad \ln q_e = \ln q_d - \beta\epsilon^2 \tag{15.7}$$

Where,

q_d= D-R constant (mg g^{-1})
β = constant related to free energy
T = absolute temperature in Kelvin
ϵ = Polanyi potential

$$\epsilon = RT \ln [1 + 1/C_e] \tag{15.8}$$

13.9 Estimation of Equilibrium in Biosorption

The equilibrium in biosorption process is estimated by sorption isotherms which are beneficial in evaluating the relationship between equilibrium concentration (C_e) of the metal ions and the mass of metal ions bounded per (q_e) unit mass of biosorbent.

Most of the time, the equilibrium between solid and liquid is done by Langmuir isotherm (Michalak et al. 2013). A chemical speciation computer program termed as PHREEQCI 6.2 was used to calculate data corresponding to equilibrium condition and compared with experimental data. At equilibrium, the data of M^+ concentration remaining in the solution is entered, and software is useful for calculating both the number of unoccupied sites at equilibrium ($q_{max} - q$) and the amount of metal that is taken up by biomass (availability of number of active sites for the occupied metal in this condition). And finally, the program database was uploaded. This program excellently reproduces the experimental results, with a better correlation between calculated data and experimental data. Hence, PHREEQCI program has proven as one of the important means for predicting the behavior of biomass after equilibrium has attained (Romera et al. 2007).

13.10 Kinetics of Biosorption

Pseudo-first-order and pseudo-second-order kinetic model are utilized to study the kinetics of heavy metals (Tálos et al. 2012). The kinetic study of the biosorption process is done as it is useful in the determination of the time of contact that is helpful in assessing sorption equilibrium and for the analysis of process parameters like temperature, and pH that are helpful for the identification of the sorptive properties of the given biosorbents (Michalak et al. 2013).

The spirulina biomass can be proficiently utilized for removal of rhenium from the industrial effluent as well as batch solution. The biosorption of rhenium with the help of spirulina biomass fits better in pseudo-second-order kinetic model (Zinicovscaia et al. 2018). Different models related to kinetics are mentioned in Table 13.5.

Spirulina platensis has the maximum attainable biosorption that is 97.1%, and the equilibrium adsorption capacities of the adsorbent which are used for zinc ions were investigated using the two isotherms that are Langmuir and Freundlich isotherms, and Langmuir isotherm was found as a better correlation (Gaur and Dehankhar 2009)**.** The term pseudo-second order has reaction constant K_2 and was introduced in the mid-1980s, but it was not very popular until 1999 when McKay and Ho performed numerous experiments and concluded result. They analyzed that pseudo-second-order kinetics offers excellent correlation of experimental data (Liu and Shen 2008; Simonin and Bouté 2016). It was investigated successfully that the adsorption of metals like Cd^{2+} and Cu^{2+} has been defined more efficiently by pseudo-second order (Dang et al. 2008).

Table 13.5 Kinetic models for biosorption

Kinetic models	Uses	Equation	Nomenclature	References
Pseudo-first-order kinetic model	Also termed as Langergen's first-order reaction that can be applied only for the initial stage of adsorption. It depends upon the same concentration of only one of the two reactants	$log(q_e - q_t) = log\ q_e - \frac{tk_1}{2.303}$	k_1 = constant rate of adsorption; $q_e$and q_t = adsorption capacity at equilibrium and at time t; t = time	Moussout (2018)
Pseudo-second-order kinetic model	Used for depicting the kinetic process of M^{2+} onto adsorbent media. It offers the best correlation of experimental data	$\frac{t}{q_t} = \frac{1}{K_2 q_e^2} + \frac{t}{q_e}$	K_2(g mg^{-1} $min^{-1/2}$) = pseudo-second-order rate constant adsorption; q_e and q_t= equilibrium; t = time	Akbar et al. (2018)
Weber and Morris model	Analysis of the diffusion model takes place by this model	$q_t = k_{di}t^{1/2} + c$	k_{di}(mg g^{-1} $min^{-1/2}$) = IPD rate constant; c = intercept to the adsorption stage	Akbar et al. (2018)
Elovich model	Used for the determination of kinetics of chemisorption of gases on the surface of heterogeneous solids	$\frac{dq_t}{dt} = \alpha e^{-\beta qt}$	α = adsorption rate β = desorption constant	Dadwal and Mishra (2016)

13.11 Discussion

In this chapter, marine algae are considered as an appropriate biosorbent for removal of heavy metals attributable because of the presence of rich polysaccharides in their cell wall. To support this, various models have been discussed. The utilization of brown algae is mainly because of the reason that it possesses the best biosorptive as well as bioaccumulative properties in comparison to red and green algae. It is

concluded that the carboxylic group is the most abundant and dominant acidic functional group followed by fucoidan. Due to the abundance in the extracellular polymers and cell wall matrix polysaccharides, brown algae are most useful in the removal of heavy metals from the industrial effluents.

Acknowledgments Authors of this manuscript would like to thank the Council of Science and Technology Uttar Pradesh and School of Biochemical Engineering (IIT, BHU) Varanasi for providing their technical support.

The authors have declared no conflict of interest.

References

Abdel-Ghani NT, El-Chaghaby GA, Helal FS (2015) Individual and competitive adsorption of phenol and nickel onto multiwalled carbon nanotubes. J Adv Res 6(3):405–415

Abirhire O, Kadiri MO (2011) Bioaccumulation of heavy metals using microalgae. Asian J Micro Biotech Environ Sci 13:91–94

Akbar NA, Kamil NAFM, Zin NSM, Adlan MN, Aziz HA (2018) Assessment of kinetic models on Fe adsorption in groundwater using high-quality limestone. IOP Conf Ser: Earth Environ Sci 140(1):12–30

Alluri H, Ronda S, Settalluri V, Bondili J, Suryanarayana V, Venkateshwar P (2007) Biosorption: an eco-friendly alternative for heavy metal removal (Vol. 6)

Arunakumara KKIU, Zhang X. (2008). Heavy metal bioaccumulation and toxicity with special reference to microalgae (Vol. 7)

Aziz MA, Hashem MA, Ahmed KU, Haque MM (2004) Effect of salinity on growth and nitrogen fixation of cyanobacteria. Bangladesh J Prog Sci Tech 2(2):193–196

Belghit I, Rasinger JD, Heesch S, Biancarosa I, Liland N, Torstensen B, Bruckner CG (2017) In-depth metabolic profiling of marine macroalgae confirms strong biochemical differences between brown, red and green algae. Algal Res 26:240–249

Benaisa S, Arhoun B, El Mail R, Rodriguez-Maroto JM (2018) Potential of brown algae biomass as new biosorbent of Iron: kinetic, equilibrium and thermodynamic study. J Mater Environ Sci 9 (7):2131–2141

Bilal M, Rasheed T, Sosa-Hernandez JE, Raza A, Nabeel F, Iqbal HMN (2018) Biosorption: an interplay between marine algae and potentially toxic elements-A review. Mar Drugs 16(2). https://doi.org/10.3390/md16020065

Cheng J, Qiu H, Chang Z, Jiang Z, Yin W (2016a) The effect of cadmium on the growth and antioxidant response for freshwater algae Chlorella vulgaris (Vol. 5)

Cheng J, Yin W, Chang Z, Lundholm N, Jiang Z (2016b) Biosorption capacity and kinetics of cadmium(II) on live and dead Chlorella vulgaris (Vol. 29)

Chu KH, Hashim MA (2007) Copper biosorption on immobilized seaweed biomass: column breakthrough characteristics. J Environ Sci (China) 19(8):928–932

Dadwal A, Mishra V (2016) Review on biosorption of arsenic from contaminated water (Vol. 45)

Dang HV, Doan H, Dang Vu T, Lohi A (2008) Equilibrium and Kinetics of Biosorption of Cadmium(II) and Copper(II) Ions by Wheat Straw (Vol. 100)

Davis TA, Volesky B, Mucci A (2003) A review of the biochemistry of heavy metal biosorption by brown algae. Water Res 37(18):4311–4330

Diniz V, Volesky B (2005) Biosorption of La, Eu and Yb using Sargassum biomass. Water Res 39 (1):239–247

El Gamal AA (2010) Biological importance of marine algae. Saudi Pharm J 18(1):1–25

Fisher NS, Bohé M, Teyssié JL (1984). Accumulation and toxicity of Cd, Zn, Ag, and Hg in 4 marine phytoplankters (Vol. 18)

Fourest E, Volesky B (1996) Contribution of sulfonate groups and alginate to heavy metal biosorption by the dry biomass of Sargassum fluitans. Environ Sci Technol 30(1):277–282
Gaur R. Dehankhar N (2009) Equilibrium modelling and spectroscopic studies for the biosorption of zn+2 ions from aqueous solution using immobilized spirulina platensis (Vol. 6)
González F, Romera E, Ballester A, Blázquez ML, Muñoz J, García-Balboa C (2011) Algal biosorption and biosorbents. In (pp 159–178)
Henriques B, Rocha L, Lopes C, Figueira P, Monteiro JR, Duarte AR, Pereira E (2015) Study on bioaccumulation and biosorption of mercury by living marine macroalgae: Prospecting for a new remediation biotechnology applied to saline waters (Vol. 281)
Jakimska A, Konieczka P, Skóra K, Namieśnik J (2011) Bioaccumulation of metals in tissues of marine animals, part II: metal concentrations in animal tissues (Vol. 20)
Karthikeyan S, Balasubramanian R, Iyer CS (2007) Evaluation of the marine algae Ulva fasciata and Sargassum sp. for the biosorption of Cu(II) from aqueous solutions. Bioresour Technol 98 (2):452–455
Kumar R, Goyal D (2009) Comparative biosorption of Pb2+ by live algal consortium and immobilized and dead biomass from aqueous solution (Vol. 47)
Langmuir I (1916) The constitution and fundamental properties of solids and liquids. pp 2221–2267
Lee M-G, Lim J-H, Kam S-K (2002) Biosorption characteristics in the mixed heavy metal solution by biosorbents of marine brown algae. Korean J Chem Eng 19(2):277–284
Liu Y, Shen L (2008) From Langmuir Kinetics to first- and second-order rate equations for adsorption (Vol. 24)
Luciana R, Gervasio S, Troiani H, Gagneten AM (2013) Bioaccumulation and toxicity of copper and lead in Chlorella vulgaris (Vol. 4)
Lupea M, Bulgariu L, Macoveanu M (2012) Biosorption of Cd(II) from aqueous solution on marine green algae biomass (Vol. 11)
M Al Maghraby, D., & Hassan, I. (2017). Heavy metals bioaccumulation by the green alga Cladophera herpestica in Lake Mariut, Alexandria, *Egypt* (Vol. 1)
M Ibrahim W, Abdel Aziz YS, Hamdy S, Gad NS (2018) Comparative study for biosorption of heavy metals from synthetic wastewater by different types of marine algae (Vol. 09)
Maadane, A., Merghoub, N., El Mernissi, N., Tarik, A., Amzazi, S., Wahby, I., & Bakri, Y. (2017) Antimicrobial activity of marine microalgae isolated from Moroccan coastlines (Vol. 6)
Manuel J, Vigneshwasran C, Annadurai S, Prakash Kumar BG, Velmurugan S (2016) Algal biosorption of heavy metals
Matheickal J, Yu Q, Woodburn G (1999) Biosorption of Cadmium(II) from Aqueous solutions by pre-treated biomass of marine algae durvillaea potatorum (Vol. 33)
Michalak I, Chojnacka K, Witek-Krowiak A (2013) State of the art for the biosorption process--a review. Appl Biochem Biotechnol 170(6):1389–1416
Moussout H. (2018). Critical of linear and nonlinear equations of pseudo-first-order and pseudo-second order kinetic models (Vol. 4)
Nessim RB, Bassiouny A, Zaki HR, Moawad M, Kandeel KM (2011) Biosorption of lead and cadmium using marine algae (Vol. 27)
Nirmal Kumar J, Kumar RN, Oommen C (2009) Removal of cadmium, mercury, and lead from aqueous solution using marine macroalgae as low-cost adsorbents. PRAJÑ-J Pure Appl Sci:28
Patel G, Doshi H, Thakur M (2016) Biosorption and equilibrium study of copper by marine seaweeds from North West Cost of India. J Environ Sci Toxic Food Technol 10(7):54–64
Purbonegoro T, Suratno PR, Husna N (2018) Toxicity of copper on the growth of marine microalgae Pavlova sp. and its chlorophyll-a (Vol. 118)
Ramezani Moghaddam M, Fatemi S, Keshtkar A (2013) Adsorption of lead (Pb2+) and uranium (UO22+) cations by brown algae; experimental and thermodynamic modeling (Vol. 231)
Romera E, Gonzalez F, Ballester A, Blazquez ML, Munoz JA (2007) Comparative study of biosorption of heavy metals using different types of algae. Bioresour Technol 98 (17):3344–3353

Salah M, El-Naggar N, Hamouda R, Mousa I (2018) Biosorption optimization, characterization, immobilization, and application of Gelidium amansii biomass for complete Pb^{2+} removal from aqueous solutions (Vol. 8)

Schiewer S, Volesky B (2000) Biosorption by marine algae. In: Valdes JJ (ed) Bioremediation. Springer, Dordrecht, pp 139–169

Shafiq M, Alazba P, Amin M (2018) Removal of heavy metals from wastewater using date palm as a biosorbent: a comparative review (Vol. 47)

Shams Khorramabadi G, Darvishi Cheshmeh Soltani R (2008) Evaluation of the marine algae gracilaria salicornia and sargassum sp. for the biosorption of Cr (VI) from aqueous solutions (Vol. 8)

Simonin J-P, Bouté J (2016) Intraparticle diffusion-adsorption model to describe liquid/solid adsorption kinetics (Vol. 15)

Sulaymon A (2014) Biosorption of heavy metals: A review (Vol. 3)

Swaleh MM, Ruwa RK, Wainaina M, Ojwang LM, Shikuku SL, Maghanga JK (2016) Mariam M. Swaleha∗, Renison Ruwa b, Moses N. Wainainaa, Loice M. Ojwang'a, Samuel L. Shikuku a and Justin K. Maghangaa "Heavy Metals Bioaccumulation Assessment in acanthopleura gemmata from Fort Jesus Mombasa". J Environ Sci, Toxicol Food Technol 10:39–45

Sweetly J (2014) Macroalgae as a potentially low-cost biosorbent for heavy metal removal. A review. Int J Pharm Biol Arch 5(2)

Tálos K, Pernyeszi T, Majdik C, Hegedusova A, Páger C (2012) Cadmium biosorption by baker's yeast in aqueous suspension (Vol. 77)

Torres M, Barros M, Campos CG, Pinto E, Rajamani S, Sayre R, Colepicolo P (2008) Biochemical biomarkers in algae and marine pollution: a review (Vol. 71)

Umar Mustapha M, Halimoon N (2015) Microorganisms and biosorption of heavy metals in the environment: a review paper (Vol. 07)

Vieira RH, Volesky B (2000) Biosorption: a solution to pollution? Int Microbiol 3(1):17–24

Vilar VJ, Botelho CM, Loureiro JM, Boaventura RA (2008) Biosorption of copper by marine algae Gelidium and algal composite material in a packed bed column. Bioresour Technol 99 (13):5830–5838

Volesky B, Holan ZR (1995) Biosorption of heavy metals. Biotechnol Prog 11(3):235–250

Volesky B, Kuyucak N (1988). Biosorbent for gold. In: Google Patents

Wallenstein F, Couto R, Amaral A, Wilkinson M, Neto AI, Rodrigues A (2009) Baseline metal concentrations in marine algae from Sao Miguel (Azores) under different ecological conditions – Urban proximity and shallow water hydrothermal activity (Vol. 58)

Yalcin S, Sezer S, Apak R (2012) Characterization and lead(II), cadmium(II), nickel(II) biosorption of dried marine brown macroalgae Cystoseira barbata. Environ Sci Pollut Res Int 19 (8):3118–3125

Zinicovscaia I, Safonov A, Troshkina I, Demina L, German K (2018) Biosorption of Re(VII) from batch solutions and industrial effluents by cyanobacteria Spirulina platensis. CLEAN – Soil, Air, Water 46(7):1700576

Linking Microbial Genomics to Renewable Energy Production and Global Carbon Management

14

Neha, Abhishek Singh, Suman Yadav, and Yashpal Bhardwaj

Abstract

The diminishing concentration of available fossil fuels and increasing global demand of energy have obligated the need for the production of alternate fuels to current petroleum-based fuels. Microbes have the potential to render renewable and sustainable energy sources that are carbon-neutral to counter the elevated concentration of greenhouse gases in the substantial climate changes. Various advancements in sequencing technologies have enabled the study of the microbial diversity and interpreting the variations within the entire genome of organisms and concluding the most feasible pathway of substrate utilization in a comparatively cheaper and faster way. To completely exploit the biofuel-producing potential of these microbes, various genomes have been sequenced and are now available for study. Computational approaches like functional genomics, genome-scale metabolic engineering, and flux balance analysis can be used to improve the H_2-producing efficiencies of microbes. Many microorganisms like *Enterobacter* sp. IIT-BT 08 are reported to have a high rate of H_2 production, and its draft genome was generated at DOE Joint Genome Institute (JGI) using Illumina data. The *C. perfringens* strain JJC was sequenced using the Illumina MiSeq benchtop sequencer which uses a vast variety of carbohydrates producing acetate, butyrate, lactate, ethanol, H_2, and carbon dioxide and has various industrial applications. Access to multiple microalgal genome sequences now provides opportunities for application of "omic" approaches to decipher algal lipid metabolism and identify gene targets for the development of potentially engineered strains with optimized lipid content from which biofuel can be produced.

Neha and Yashpal Bhardwaj have been contributed equally with all other contributors.

Neha · A. Singh · S. Yadav · Y. Bhardwaj (✉)
Laboratory of Molecular Ecology, Centre of Advanced Study in Botany, Banaras Hindu University, Varanasi, Uttar Pradesh, India
e-mail: yashpalbot.bhu@gmail.com

V. Tripathi et al. (eds.), *Microbial Genomics in Sustainable Agroecosystems*,
https://doi.org/10.1007/978-981-13-8739-5_14

14.1 Introduction

Microorganisms are the most abundant and diverse forms of life and found in most habitats on the Earth including those conducive to extreme environments like hot springs, glaciers, miles beneath the soils, etc. The genetic, metabolic, and physiological diversity of microbial species is far greater than that found in plants and animals. The enormity of microbial species was estimated to be approximately 10^{11} to 10^{12} out of which most of the microbial species are still unknown. Of those species that have been described, their biological diversity is extraordinary, having adapted to grow under extreme temperature, pH, salt concentration, and oxygen levels. Currently, various advancements in sequencing technologies have enabled the study of microbial diversity and interpreting the variations within the entire genome of organisms and concluding the most feasible pathway of substrate utilization in a comparatively cheaper and faster way (Shendure and Ji 2008). The first bacterial genome of *Haemophilus influenzae* was sequenced in 1995 and took more than 13 months of effort to complete. Today, the entire genome of a microbial species can be sequenced in a very short span of time and take less than 30 h to sequence the entire genome. For example, presently sequencer like MiSeq produced by Illumina was delineated as a fast, personal benchtop sequencer, with very less run time as short as 10 h and outputs planned for targeted sequencing and small genome sequencing (Reuter et al. 2015). The whole genome sequencing of DNA extracted from culturable microorganisms or metagenome (genetic material isolated directly from environmental samples) unveils preliminary idea of the gene associated in numerous pathways related to energy production, metabolism, carbon sequestration, etc. (Fig. 14.1) (Yadav and Dubey 2018).

The key aim of carbon management is to develop options and mechanisms to reduce the causes and effects of climate change, including minimizing emissions and removal of various GHGs through natural and anthropogenic methods. Microbes have played a vital role in regulating the concentration of atmospheric GHGs (e.g., methane, carbon dioxide, nitrous oxide) that impact climate change. Study of microbial diversity involved in cycling of GHGs and integration of microbial genomics to global carbon management could enhance our capability to develop and evaluate microbial strategies for capturing and sequestering atmospheric CO_2. Autotrophic microorganisms and macroalgae are known to contribute significantly to CO_2 assimilation in aquatic systems such as the oceans and wetlands but have not generally been thought to have a key role in CO_2 fixation and sequestration in soils. This is despite the fact that microbial autotrophs have been reported in a number of soil studies.

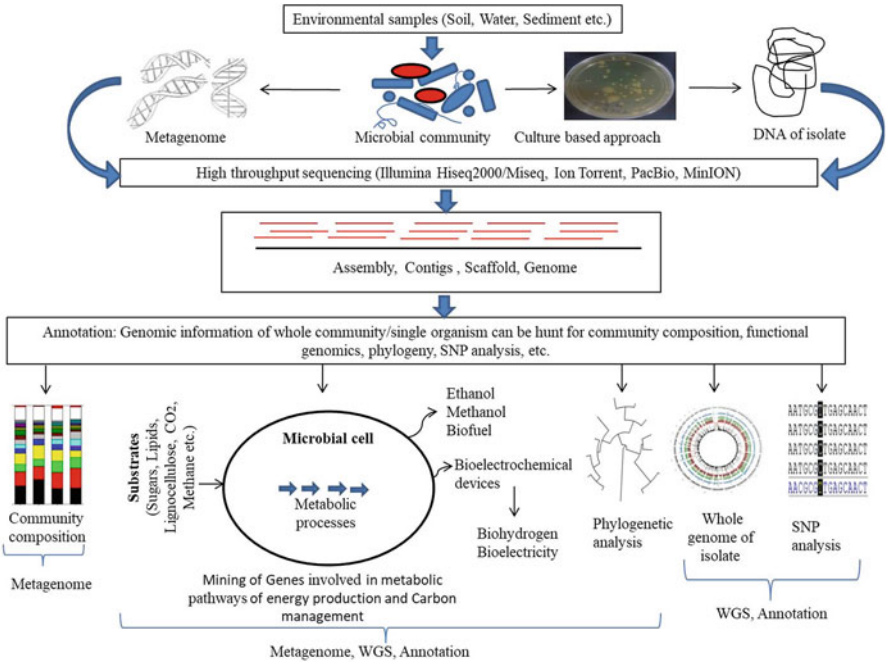

Fig. 14.1 Overview of application of genomics and metagenomic approaches for the energy production and carbon management

The world is presently focusing on the development of sustainable and nonpolluting energy sources, which will restore fossil fuels in the post-fossil fuel era (Rittmann et al. 2008). There are many alternative future fuels (e.g., hydrogen, methane, ethanol, methanol, gasoline, etc.) among which bio-hydrogen seems to be the most promising because it burns to water, which can be re-used in an environment-friendly manner (Nielsen et al. 2001). The preliminary idea for the production of renewable energy with microorganisms involves using communities of anaerobic microorganisms to transform the energy value in biomass to useful forms of energy. Waste and residues of agriculture, food processing, and other industries contain a large amount of biomass (Pfaltzgraff et al. 2013). Converting the biomass in these wastes in the form of energy provides two advantages at a time: first the generation of renewable energy and second the minimization of environmental pollution. The conversion of biomass to three valuable energy outputs (e.g., methane gas (CH_4), hydrogen gas (H_2), and electrons from bioelectricity) that are produced by a microbial fuel cell (MFC) can be achieved by various communities of anaerobic microorganisms (Logan 2004). Methanogenesis is already in widespread use today, and microbial sources of H_2 and electricity are being intensively investigated. The second approach exploits photoautotrophic microorganisms (e.g., cyanobacteria and eukaryotic algae) that capture the energy of sunlight to

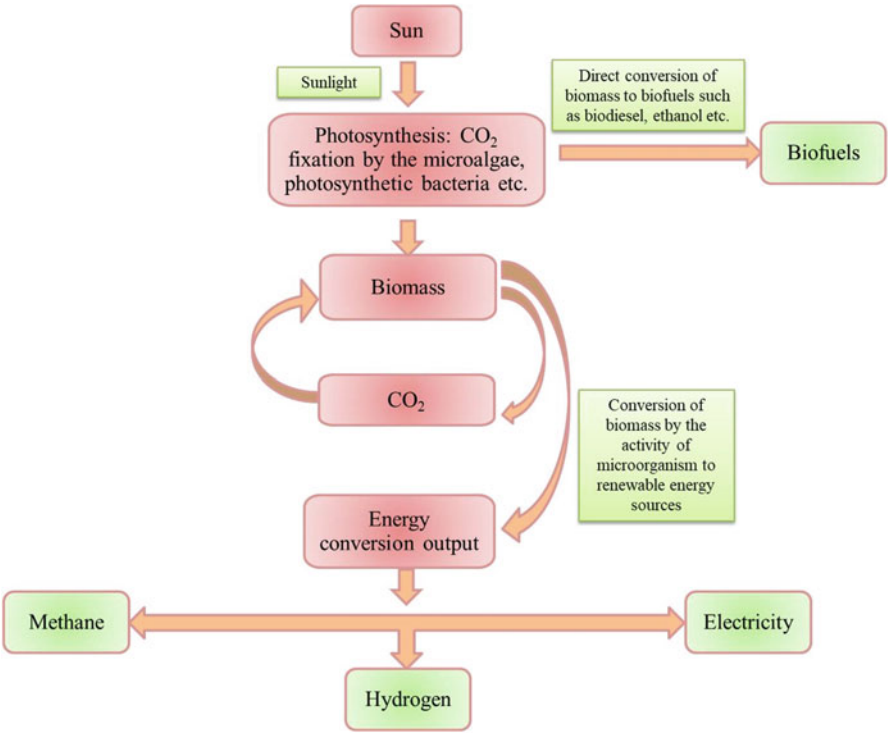

Fig. 14.2 Plant residues and microbial biomass can be used for renewable energy sources that aid in overall carbon flux in the atmosphere

grow and thereby produce biomass that can be harvested to augment the biomass produced from nature and agriculture (Fig. 14.2) (Liao et al. 2016). Owing to their high specific growth rates, year-round harvesting, and homogeneity, photosynthetic microorganisms can produce larger (by 100-fold or more) biomass-based energy stocks than plants (Misra et al. 2013). As a result, it might be feasible in production of sufficient microbial biomass to replace fossil fuels. Genomic-based study of microorganisms associated in energy production validated the base sequence of the whole DNA, and all the vital biological reactions of microorganisms can be decoded by the complete genome. On the basis of genomic data, the various metabolic pathways of microbes have been finished with speedy evolution in developing genome projects, a number of microorganisms have been sequenced completely and some are partially sequenced, and annotation of genes from the sequence information is done using bioinformatics (see http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/genome_table_cgi). The enzymes that can be coded by their genomes are discovered from the annotated sequence data, for example, the relation of hydrogenases to the redox proteins and enzymes has been demonstrated by biochemical work, and perhaps in the future through the compilation of sequence data. Currently, more than 100 sequences of hydrogenases are accessible, and

genomes of 80 microbes have been sequenced (Vignais et al. 2001). The imminent requirement to face the challenges associated in enabling these microorganisms to more realistic alternative for replacing the conventional fossil fuels using different omics-based techniques has been conferred in this chapter with possible future directions.

14.2 Carbon Management

To mitigate the effects of global environmental changes due to increasing emissions of greenhouse gases (GHGs), mainly CO_2 is one of the major problems of the twenty-first century (Schmidt et al. 2011). The carbon (C) emitted into the atmosphere was estimated to be 405 ± 30 Gigatons and primarily due to anthropogenic activities over the past 200 years, and as a result, the global atmospheric concentration of CO_2 has risen from 280 to 382 ppm in 2007, with a current annual increase of 0.88 ppm (Canadell et al. 2007). Thus to ensure global environmental security and to mitigate the effect of climate change, there is an immediate need to find cost-effective strategies for minimizing anthropogenic CO_2 emissions. There are three main methods that can be used to manage carbon in various stages of discovery and development: (i) near-term storage in the terrestrial biosphere where vegetation would fix the CO_2 and store it in biomass and soil; (ii) long-term storage in the earth's soil by pumping CO_2 into existing or drilled/excavated sub-surface reservoirs; and (iii) long-term storage in the earth's oceans where CO_2 would be injected thousands of feet deep and trapped by the water. Terrestrial carbon sequestration is an important step towards mitigating anthropogenic CO_2 emissions. Increased CO_2 concentrations in the atmosphere are thought to be partly contributed by the soil under agriculture. According to an approximation, soils have contributed 55–878 billion tons (GT) of carbon to the total atmospheric CO_2 (Kimble et al. 2002). Soil microorganisms are of great importance in soil carbon cycling. Carbon sources are added externally to soil as crop residues undergo microbial decomposition which results in loss of 2/3 of the carbon. A small portion of it is absorbed as the microbial biomass, and a major part of it is released into the environment as CO_2. As a result of this process accumulation of comparatively resistant SOC fraction, this is slightly transformed and may be attacked in the future by the microbial population. Carbon management requires the channelizing of CO_2 in the atmosphere into long-lived pools to mitigate or reduce their immediate remittance. Soil is a reservoir of huge stock of potentially volatile C which act both as a buffer against atmospheric CO_2 increase and as a possible sink for additional C depending on the balance between photosynthesis, the respiration of decomposer organisms, and stabilization of C in soils (Lal 2004; Woodward et al. 2009). Soil C sequestration can significantly contribute to the idea of mitigation and potentially offset a remarkable amount of diffuse CO_2 sources for which direct capture is not yet reasonable (King 2011). In terrestrial habitats, plants dominate the uptake of CO_2 from the atmosphere by net primary production (NPP), but microorganisms contribute largely to ecosystem C budgets with their roles as decomposers, plant symbionts, or pathogens, hence

modifying nutrient availability and affecting C turnover and absorption in soil (King 2011; Lal 2004). Methane (CH_4), the second most important greenhouse gas after CO_2, contributes to about 20% of the warming effects. Methane oxidation in aerobic soils is mediated by methane-oxidizing bacteria (MOB; methanotrophs), a subset of a physiological group of bacteria (methylotrophs). These bacteria utilize CH_4 as the sole carbon and energy source (Dubey 2005; Hanson and Hanson 1996). The role of methanotrophs can never be underestimated as these are the important contributors to attenuate CH_4 flux at the oxic-anoxic interfaces (Dutaur and Verchot 2007). The whole genome sequences of methanotrophic bacteria such as *Methylocystis* sp. Strain Rockwell (ATCC 49242) and *Methylomonas methanica* MC09 provide insights into the genomic and physiological information that can utilize further to optimize the use of these bacteria in industry or biotechnological purposes (Boden et al. 2011; Stein et al. 2011).

Land management and land use practices can manipulate the terrestrial ecosystem development of distinct microbial communities that support C sequestration (Bardgett et al. 2008; Singh et al. 2010). The fungal to bacterial ratio in soils has been related with C sequestration capability with greater fungal population being related to greater C storage. Recently, a cross-biome metagenomic research has revealed that the ratio of fungal/bacterial rRNA reads showed variation across different soils, among temperate and boreal forests having the highest fungal/bacterial ratios (Fierer et al. 2007). Higher C storage in fungal-dominated soils can be attributed to higher C use efficiency; longer retention of C in living biomass; and recalcitrant necromass resulting in longer resident time of C (Strickland and Rousk 2010). Challenges in manipulating microbial community for enhanced C sequestration arise from the enormous diversity and unculturability of soil microbial communities, which have precluded their comprehensive characterization and limited our understanding on their ecological functions. The new generation of omics methods (e.g., genomics, transcriptomics, proteomics, metabolomics, metagenomics) is proving instrumental in providing valuable information about the taxonomic, genetic, and functional properties of soil microbial communities. These techniques have begun to allow investigation of functional processes of terrestrial microbial communities involved in C cycling that can be incorporated into mechanistic and predictive ecological models (Larsen et al. 2012).

Based on the information gathered from full genome sequences, we infer that bacteria belonging to *Acidobacteria* and *Actinobacteria* possess an impressive array of genes allowing breakdown, utilization, and biosynthesis of diverse structural and storage polysaccharides and resilience to stressful soil conditions making them truly ubiquitous in terrestrial ecosystems. This finding supports the metagenomic evidence of higher SOC in *Acidobacteria*- and *Actinobacteria*-dominated communities and suggest that these groups promote soil C storage not only due to lifestyle (slow growth and lower metabolic activities) but also by producing polysaccharides for soil structural stability (Singh et al. 2010).

Metagenomic analysis has revealed that structure of microbial communities was markedly different between ambient CO_2 (aCO_2) and elevated CO_2 (eCO_2) as indicated by detrended correspondence analysis (DCA) of gene-based pyrosequencing data and functional gene array data. While the abundance of genes

involved in decomposing recalcitrant C remained unchanged, those involved in labile C degradation and C and N fixation were significantly increased under eCO_2. Here, using metagenomic technologies, we showed that 10 years of field exposure of a grassland ecosystem to eCO_2 dramatically altered the structural and functional potential of soil microbial communities (He et al. 2010).

Microbial communities living near the surface layers of oceans are the primary photosynthetic organisms driving the biological pump. Absorbing CO_2 and sunlight to produce most oceanic organic materials, the organisms make up the foundation of the marine food chain. Photosynthesis of phytoplankton such as diatoms, dinoflagellates, and cyanobacteria converts about as much atmospheric carbon to organic carbon in the ocean as plant photosynthesis does on land. Large variations in phytoplankton abundance, therefore, can greatly impact the oceans' ability to take up atmospheric carbon. Phytoplankton photosynthesis (Rivkin and Legendre 2001) fixes approximately 45 Pg C $year^{-1}$ (Falkowski et al. 2000). Dominant organisms in surface waters include such as cyanobacteria as *Synechococcus* sp. and *Prochlorococcus marinus*, which capture CO_2 and light to carry out photosynthesis. *Prochlorococci* now are thought to be the most abundant photosynthetic organisms on earth. Eukaryotic diatoms such as the recently sequenced *Thalassiosira pseudonana* also live in surface waters and convert CO_2 and other nutrients into hard silicates. This process carries organically complexed carbon to ocean depths, thus converting its relatively rapid cycling in surface waters (where it is returned to the atmosphere) to a considerably slower one in ocean sediments. The main goal of ocean carbon sequestration is to increase the export of deep ocean inventory of CO_2. Two approaches are taken into account: direct injection of a CO_2 stream into the ocean depths and iron fertilization to increase photosynthesis by phytoplankton in the biological pump and thus enhance the uptake of carbon.

14.3 Energy Production

To mitigate the increased discharge of greenhouse gases in the atmosphere and to fulfill the mounting global demand for energy to counter the decreasing concentration of fossil fuels have necessitated the production of alternate environment-friendly fuels. Different biological processes for the production of fuels such as ethanol, diesel, hydrogen (H_2), methane, etc. have capability to furnish sustainable energy system for the betterment of society (Angenent et al. 2004). In recent years, the interest in the production of different kinds of biofuels by exploiting microorganisms has been increasing steadily (Liao et al. 2016) especially because of the metabolic variety of different microorganisms that makes possible the production of biofuels from different substrates. For example, most of the bacteria can effortlessly transform sugars into ethanol, and cellulolytic microbes can use plant-driven substrates in the production of biofuels. Cyanobacteria and microalgae possess the capability to reduce the atmospheric CO_2 into biofuels photosynthetically, and methanotrophs can utilize methane to produce methanol (Liao et al. 2016). The genomic data of sequenced microbes can be connected to biofuel production yields. Genome sequence and metabolic pathway databases can be utilized for

Table 14.1 List of some of the important microorganisms involved in biofuel production with their whole genome sequence deposited to databases

Organism	Product	Gene Bank accession no.	Sequencing center	Taxonomy	References
Thermoanaerobacterium sp. strain PSU-2	H_2	MSQD00000000	Thaksin University, Thailand	*Firmicutes*	O-Thong et al. (2017)
Caldicellulosiruptor saccharolyticus DSM 8903	H_2	NC_009437	DOE Joint Genome Institute	*Firmicutes*	de Vrije et al. (2007)
Caldicellulosiruptor bescii DSM6725	H_2	NC_012036	DOE Joint genome Institute	*Firmicutes*	Kataeva et al. (2009)
Halanaerobium hydrogeniformans	H_2	CP002304	US DOE Bioenergy Research Center	*Firmicutes*	Brown et al. (2011)
Clostridium sp. strain Ade.Ty	H_2	AVSV00000000	Monash University, Malasiya	*Firmicutes*	Wong et al. (2014)
Zymomonas strain subsp. *mobilis* ATCC 29191	Ethanol	CP003704	US DOE-Joint Genome Institute	*Proteobacteria*	Desiniotis et al. (2012)
Eubacterium limosum KIST612	Ethanol	CP002273	College of life Sciences and Biotechnology, Korea University	*Firmicutes*	Roh et al. (2011)
Enterobacter sp. IIT-BT 08	H_2		US DOE-Joint Genome Institute	*Proteobacteria*	Khanna et al. (2013)
Brevundimonas naejangsanensis strain B1	H_2	JHOF00000000	Chinese National Human Genome Center	*Proteobacteria*	Su et al. (2014)
Clostridium thermocellum ATCC 27405	Ethanol	NC_009012	US DOE-Joint Genome Institute	*Firmicutes*	Rydzak et al. (2009)
Chlamydomonas reinhardtii	Biofuel	SRP053354	Phytozome DOE-Joint Genome Institutes	*Chlorophyceae*	Li et al. (2013)
Bacillus cereus ATCC 14579	Ethanol	NC_004721	Integrated Genomics Inc.	*Firmicutes*	Ouhib-Jacobs et al. (2009)
Clostridium acetobutylicum DSM 1731	Ethanol	CP002660-CP002662	Department of Human Genetics, University of California	*Firmicutes*	Bao et al. (2011)

screening microbes. The National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.ad.jp) databases greatly facilitate such analyses. Sequence analysis and pathway alignment of hydrogen metabolism in complete and incomplete genomes have led to the identification of potential hydrogen producers (Kalia et al. 2003) (Table 14.1). The work of Carere et al. (2012) demonstrated that the presence or absence of specific genes dictating carbon and electron flow towards end products may be used to infer end-product synthesis patterns and help to develop informed metabolic engineering strategies for optimization of H_2 and ethanol yields. Furthermore, certain genes may be used as suitable biomarkers for screening novel microorganisms' capability of producing optimal H_2 or ethanol and may be suitable targets for metabolic engineering strategies for optimization of either ethanol or H_2 yields.

14.3.1 Biological Hydrogen Production

The biological hydrogen production procedure makes utilization of microorganisms that tend to produce H_2 from lignocellulosic biomass and waste material (Bakonyi et al. 2014; Kumar et al. 2015). These materials are excellent source of fermentable sugar and are present in complex form and hardly digestible (Kumar et al. 2008). Direct or indirect biophotolysis, photo-fermentation, and dark fermentation methods are exploited for biological hydrogen production. The lower level yield of H_2 by biological hydrogen production methods is one of the major challenges that need to be addressed before it can be used for industrial purpose. Apart from wet lab experiments, in silico approaches which include functional genomics, genome-scale metabolic engineering, and flux balance analysis can be used to improve the H_2-producing capabilities. Many microorganisms are being explored for future biohydrogen generation at industrial scales. Among them, *Enterobacter* sp. IIT-BT 08 is reported to have a high rate of H_2 production, and its draft genome was generated at DOE Joint Genome Institute (JGI) using Illumina data. A complete genome sequence analysis was carried out for further enhancement of H_2 production by strain development (Khanna et al. 2013). *Halanaerobium hydrogeniformans* isolated from the haloalkaline environment is an obligately anaerobic, Gram-negative, nonmotile, nonsporulating, elongated rod. It can ferment a vast range of carbohydrates with optimal growth at pH 11, and 33°C, and produce acetate, formate, and H_2 as major metabolic end products. The *H. hydrogeniformans* genome was sequenced using a combination of Illumina and 454 technologies to improve assessment of its metabolic and bioenergy potential for robust H_2 production (Brown et al. 2011). Similarly, *Clostridium perfringens*, a Gram-positive and spore-forming strict anaerobe, can successfully utilize a vast variety of carbohydrates producing acetate, butyrate, lactate, ethanol, H_2, and carbon dioxide, which have industrial applications. The genome sequencing of *C. perfringens* strain JJC was performed using the Illumina MiSeq benchtop sequencer (2150-bp paired-end sequencing). The whole-genome shotgun project of *C. perfringens* strain JJC containing its assembly

and annotation has been deposited at DDBJ/EMBL/Gene Bank under the accession no.AWRZ00000000 (Wong et al. 2014) for further applications.

14.3.2 Liquid Biofuels

Liquid biofuels from plants and microalgae feedstock represent a renewable sustainable alternative to petroleum energy. The greatly minimized acreage estimates, high lipid or starch content, and biomass production rates that surpass those of terrestrial plants suggest that biodiesel or ethanol derived from lipids or starch produced by microalgae may circumvent many of the limitations ascribed to petroleum fuel and first-generation plant-based biofuels. An in-depth knowledge of microalgae genomics precludes these necessary increases in biological efficiency. Access to multiple microalgal genome sequences now provides a wealth of opportunities for application of "omic" approaches to unravel algal lipid metabolism and identify gene targets for the development of potentially engineered strains with optimized lipid content (Beer et al. 2009; Georgianna and Mayfield 2012; Mukhopadhyay et al. 2008; Rodríguez-Moyá and Gonzalez 2010; Yu et al. 2011). Bio-oil from microalgae can be used directly as fuel or chemically trans-esterified into biodiesel. Microalgae seem to be an attractive way to produce biofuel due to their ability to accumulate lipids and their very high actual photosynthetic yields; about 3–8% of solar energy can be converted to biomass, whereas observed yields for terrestrial plants are about 0.5%. The genetic information of the sequenced organisms has enabled the metabolic pathways for the lipid synthesis and which can be used in genetic engineering process efforts directed towards augmenting lipid accumulation in microalgae (Georgianna and Mayfield 2012). KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.jp/kegg/) is one of the most widely used comprehensive resources of metabolic pathways including for several organisms (Kanehisa et al. 2010). Currently, genome-wide studies have employed KEGG pathway database to identify genes and reconstruct major lipid biosynthetic pathways in various oleaginous microalgal species (Hashimoto et al. 2008; Misra et al. 2013; Rismani-Yazdi et al. 2011; Smith et al. 2012). Most of the omic-based studies undertaken so far have primarily addressed identification of gene targets for improving lipid production in microalgae. Now it is apparent that modification of the fatty acid profile to include more stearic acid (C18:0) and oleic acid (C18:1) is also indispensable for improving the algal-derived biofuel properties (Knothe 2009).

To date, ethanol accounts for up to 75% of the total biofuel use. Bioethanol dominates the market with a sale of 58 billion dollars per year. Nearly 50% of global sugar is utilized for ethanol production, and approximately 86,000 k ton/year ethanol so produced is majorly used for biofuel application (Burk 2010). The United States and Brazil are the leading producers dominantly using simple substrates such as corn and sugarcane, respectively (Aro 2016). Even though India is the second largest producer of sugarcane in the world, it contributes to only 2% of the global bioethanol production. Despite the abundant supply of lignocelluloses, their commercial con-

version to ethanol is limited to their recalcitrance (due to lignin sheath) to degradation and unique chemical composition (Zhao et al. 2012). In order to produce bioethanol from lignocellulosic biomass that is economically feasible, sustainable, and competitive with petroleum-based fuels, conventional process steps need to be integrated into the consolidated process to avoid maximum production of inhibitory sugar derivatives and achieve high ethanol titers. Environmental stresses and inhibitors encountered by *Saccharomyces cerevisiae* strains are the main limiting factors in bioethanol fermentation. Strains with different genetic backgrounds usually show diverse stress tolerance responses. An understanding of the mechanisms underlying these phenotypic diversities within *S. cerevisiae* populations could guide the construction of strains with desired traits. This kind of study, provided novel transcriptomic information on microbes and their RNA-seq data were useful in targeting genes involved in ethanol production for future genetic engineering (Wang et al. 2016).

14.3.3 Microbial Fuel Cell (MFC)

MFC is a biochemical-catalyzed system, which generates electricity by oxidizing soluble or dissolved organic wastes in the presence of either fermentative bacteria or enzyme. MFC technology relies on the electrogenic nature of certain bacteria while treating different wastewater and producing electrical energy. The microorganism generally presents in the anode chamber of fuel cell act as biocatalyst and generates electrons (e^-) and protons (H^+) by way of anaerobic respiration of organic substrate. The electron transfer through the anode is integrated with an external circuit to cathode and protons through the proton exchange membrane (which separates cathode and anode chamber) into the cathode chamber where they combine with the help of a mediator. The potential between the respiratory system and electron acceptor generates the current and voltage needed to make electricity (Logan 2004). Recently, a number of bacteria such as *Shewanella putrefaciens*, family of *Geobacteraceae*, *Rhodoferax ferrireducens*, *Bacillus subtilis*, *Geobacter sulfurreducens*, and *Escherichia coli* were reported in the literature and have the ability to transfer produced electrons from oxidized fuel (substrate) to the electrode without using artificial mediator, making it possible to establish mediator-less MFCs (Kaufmann and Lovley 2001; Kim et al. 1999). The 16S rDNA analysis of anode biofilm and suspended cells reveals predominance bacterial community involved in electron transfer. Patil et al. (2009) studied the activated sludge-based microbial fuel cell and analyzed the developed microbial community in the anode chamber and reported the predominance of *β-Proteobacteria* clones with 50.6% followed by unclassified bacteria (9.9%), *α-Proteobacteria* (9.1%), other *Proteobacteria* (9%), *Planctomycetes* (5.8%), *Firmicutes* (4.9%), *Nitrospora* (3.3%), *Spirochaetes* (3.3%), *Bacteroides* (2.4%), and *γ-Proteobacteria* (0.8%). Diverse bacterial groups represented as members of the anode chamber community. They suggest that chocolate wastewater has a potential for future MFC practical applications as it can provide a readily biodegradable waste source for electricity generation.

14.4 Conclusion and Future Prospective

Microbes have the potential to transform biomass into eco-friendly biofuels like ethanol, H_2, etc. through bioprocessing. To address the global energy crisis, genomics shall play an important role. Through knowledge of different omics technology, it will be easy to develop a better understanding to harness different renewable and carbon-neutral energy sources like lignocellulosic biomass, microalgae, and cyanobacteria. Moreover, the genetic engineering of different enzymes will be a vital factor in optimizing the development of sustainable energy in the form of biofuel. In the coming years, researchers will continue to look to nature for solutions to the global energy crisis. By applying genomic research and engineering to renewable fuel stocks and the bacteria and enzymes that convert those sources to energy, scientists can optimize billions of years of evolution to meet our growing energy needs in an environmentally friendly way.

References

Angenent LT, Karim K, Al-Dahhan MH, Wrenn BA, Domíguez-Espinosa R (2004) Production of bioenergy and biochemicals from industrial and agricultural wastewater. Trends Biotechnol 22 (9):477–485

Aro EM (2016) From first generation biofuels to advanced solar biofuels. Ambio 45:24–31

Bakonyi P, Nemestóthy N, Simon V, Bélafi-Bakó K (2014) Review on the start-up experiences of continuous fermentative hydrogen producing bioreactors. Renew Sust Energ Rev 40:806–813

Bao G, Wang R, Zhu Y, Dong H, Mao S, Zhang Y, Chen Z, Li Y, Ma Y (2011) Complete genome sequence of Clostridium acetobutylicum DSM 1731, a solvent-producing strain with multireplicon genome architecture. J Bacteriol 193(18):5007–5008

Bardgett RD, Freeman C, Ostle NJ (2008) Microbial contributions to climate change through carbon cycle feedbacks. ISME J 2:805

Beer LL, Boyd ES, Peters JW, Posewitz MC (2009) Engineering algae for biohydrogen and biofuel production. Curr Opin Biotechnol 20(3):264–271

Boden R, Cunliffe M, Scanlan J, Moussard H, Kits KD, Klotz MG, Jetten MSM, Vuilleumier S, Han J, Peters L, Mikhailova N, Teshima H, Tapia R, Kyrpides N, Ivanova N, Pagani I, Cheng J-F, Goodwin L, Han C, Hauser L, Land ML, Lapidus A, Lucas S, Pitluck S, Woyke T, Stein L, Murrell JC (2011) Complete genome sequence of the aerobic marine methanotroph *Methylomonas methanica* MC09. J Bacteriol 193:7001

Brown SD, Begemann MB, Mormile MR, Wall JD, Han CS, Goodwin LA, Pitluck S, Land ML, Hauser LJ, Elias DA (2011) Complete genome sequence of the haloalkaliphilic, hydrogen-producing bacterium *Halanaerobium hydrogeniformans*. J Bacteriol 193:3682–3683

Burk MJ (2010) Sustainable production of industrial chemicals from sugars. Int Sugar J 112:30–35

Canadell JG, Le Quéré C, Raupach MR, Field CB, Buitenhuis ET, Ciais P, Conway TJ, Gillett NP, Houghton RA, Marland G (2007) Contributions to accelerating atmospheric CO^2 growth from economic activity, carbon intensity, and efficiency of natural sinks. Proc Natl Acad Sci 104:18866–18870

Carere CR, Rydzak T, Verbeke TJ, Cicek N, Levin DB, Sparling R (2012) Linking genome content to biofuel production yields: a meta-analysis of major catabolic pathways among select H_2 and ethanol-producing bacteria. BMC Microbiol 12:295

de Vrije T, Mars AE, Budde MA, Lai MH, Dijkema C, de Waard P, Claassen PA (2007) Glycolytic pathway and hydrogen yield studies of the extreme thermophile *Caldicellulosiruptor saccharolyticus*. Appl Microbiol Biotechnol 74:1358–1367

Desiniotis A, Kouvelis VN, Davenport K, Bruce D, Detter C, Tapia R, Han C, Goodwin LA, Woyke T, Kyrpides NC, Typas MA, Pappas KM (2012) Complete genome sequence of the ethanol-producing *Zymomonas mobilis subsp. mobilis* centrotype ATCC 29191. J Bacteriol 194 (21):5966–5967

Dubey S (2005) Microbial ecology of methane emission in rice Agroecosystem: a review. Appl Ecol Environ Res 3:1–27

Dutaur L, Verchot LV (2007) A global inventory of the soil CH_4 sink. Glob Biogeochem Cy 21

Falkowski P, Scholes RJ, Boyle E, Canadell J, Canfield D, Elser J, Gruber N, Hibbard K, Högberg P, Linder S, Mackenzie FT, Moore B III, Pedersen T, Rosenthal Y, Seitzinger S, Smetacek V, Steffen W (2000) The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System. Science 290:291–296

Fierer N, Bradford MA, Jackson RB (2007) Toward an ecological classification of soil bacteria. Ecology 88:1354–1364

Georgianna DR, Mayfield SP (2012) Exploiting diversity and synthetic biology for the production of algal biofuels. Nature

Hanson RS, Hanson TE (1996) Methanotrophic bacteria. Microbiol Rev 60:439–471

Hashimoto K, Yoshizawa AC, Okuda S, Kuma K, Goto S, Kanehisa M (2008) The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes. J Lipid Res 49 (1):183–191

He Z, Xu M, Deng Y, Kang S, Kellogg L, Wu L, Van Nostrand JD, Hobbie SE, Reich PB, Zhou J (2010) Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO2. Ecol Lett 13(5):564–575

Kalia VC, Lal S, Ghai R, Mandal M, Chauhan A (2003) Mining genomic databases to identify novel hydrogen producers. Trends Biotechnol 21:152–156

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38(Database issue):D355–D360

Kataeva IA, Yang SJ, Dam P, Poole FL 2nd, Yin Y, Zhou F, Chou W-C, Xu Y, Goodwin L, Sims DR, Detter JC, Hauser LJ, Westpheling J, Adams MW (2009) Genome sequence of the anaerobic, thermophilic, and cellulolytic bacterium "*Anaerocellum thermophilum*" DSM 6725. J Bacteriol 191(11):3760–3761

Kaufmann F, Lovley DR (2001) Isolation and characterization of a soluble NADPH-dependent Fe (III) reductase from *Geobacter sulfurreducens*. J Bacteriol 183:4468

Khanna N, Ghosh AK, Huntemann M, Deshpande S, Han J, Chen A, Kyrpides N, Mavrommatis K, Szeto E, Markowitz V, Ivanova N, Pagani I, Pati A, Pitluck S, Nolan M, Woyke T, Teshima H, Chertkov O, Daligault H, Davenport K, Gu W, Munk C, Zhang X, Bruce D, Detter C, Xu Y, Quintana B, Reitenga K, Kunde Y, Green L, Erkkila T, Han C, Brambilla E-M, Lang E, Klenk H-P, Goodwin L, Chain P, Das D (2013) Complete genome sequence of *Enterobacter* sp. IIT-BT 08: A potential microbial strain for high rate hydrogen production. Stand Genomic Sci 9:359–369

Kim BH, Kim HJ, Hyun MS, Park DH (1999) Direct electrode reaction of Fe(III)-reducing bacterium, *Shewanella putrefaciens*. J Microbiol Biotechnol 9:127–131

Kimble JM, Lal R, Follett RF (2002) Agricultural practices and policies for carbon sequestration in soil. CRC Press, Boca Raton

King GM (2011) Enhancing soil carbon storage for carbon remediation: potential contributions and constraints by microbes. Trends Microbiol 19:75–84

Knothe G (2009) Improving biodiesel fuel properties by modifying fatty ester composition. Energy Environ Sci 2:759–766

Kumar R, Singh S, Singh OV (2008) Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. J Ind Microbiol Biotechnol 35(5):377–391

Kumar G, Bakonyi P, Periyasamy S, Kim SH, Nemestóthy N, Bélafi-Bakó K (2015) Lignocellulose biohydrogen: Practical challenges and recent progress. Renew Sust Energ Rev 44:728–737

Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. Science 304:1623–1627

Larsen PE, Field D, Gilbert JA (2012) Predicting bacterial community assemblages using an artificial neural network approach. Nat Methods 9:621

Li X, Huang S, Yu J, Wang Q, Wu S (2013) Improvement of hydrogen production of *Chlamydomonas reinhardtii* by co-cultivation with isolated bacteria. Int J Hydrog Energy 38:10779–10787

Liao JC, Mi L, Pontrelli S, Luo S (2016) Fuelling the future: microbial engineering for the production of sustainable biofuels. Nat Rev Microbiol 14(5):288–304

Logan BE (2004) Peer reviewed: extracting hydrogen and electricity from renewable resources. Environ Sci Technol 38:160A–167A

Misra N, Panda PK, Parida BK (2013) Agrigenomics for microalgal biofuel production: an overview of various bioinformatics resources and recent studies to link OMICS to bioenergy and bioeconomy. OMICS 17(11):537–549

Mukhopadhyay A, Redding AM, Rutherford BJ, Keasling JD (2008) Importance of systems biology in engineering microbes for biofuel production. Curr Opin Biotechnol 19(3):228–234

Nielsen AT, Amandusson H, Bjorklund R, Dannetun H, Ejlertsson J, Ekedahl L-G, Lundström I, Svensson BH (2001) Hydrogen production from organic waste. Int J Hydrog Energy 26:547–550

O-Thong S, Khongkliang P, Mamimin C, Singkhala A, Prasertsan P, Birkeland NK (2017) Draft genome sequence of *Thermoanaerobacterium* sp. strain PSU-2 isolated from thermophilic hydrogen producing reactor. Genom Data 12:49–51

Ouhib-Jacobs O, Lindley ND, Schmitt P, Clavel T (2009) Fructose and glucose mediates enterotoxin production and anaerobic metabolism of *Bacillus cereus* ATCC14579(T). J Appl Microbiol 107(3):821–829

Patil SA, Surakasi VP, Koul S, Ijmulwar S, Vivek A, Shouche YS, Kapadnis BP (2009) Electricity generation using chocolate industry wastewater and its treatment in activated sludge based microbial fuel cell and analysis of developed microbial community in the anode chamber. Bioresour Technol 100:5132–5139

Pfaltzgraff LA, De bruyn M, Cooper EC, Budarin V, Clark JH (2013) Food waste biomass: a resource for high-value chemicals. Green Chem 15:307–314

Reuter JA, Spacek DV, Snyder MP (2015) High-throughput sequencing technologies. Mol Cell 58 (4):586–597

Rismani-Yazdi H, Haznedaroglu BZ, Bibby K, Peccia J (2011) Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: Pathway description and gene discovery for production of next-generation biofuels. BMC Genomics 12:148

Rittmann BE, Krajmalnik-Brown R, Halden RU (2008) Pre-genomic, genomic and post-genomic study of microbial communities involved in bioenergy. Nat Rev Microbiol 6:604

Rivkin RB, Legendre L (2001) Biogenic carbon cycling in the upper ocean: effects of microbial respiration. Science 291(5512):2398–2400

Rodríguez-Moyá M, Gonzalez R (2010) Systems biology approaches for the microbial production of biofuels. Biofuels 1:291–310

Roh H, Ko H-J, Kim D, Choi DG, Park S, Kim S, Chang IS, Choi I-G (2011) Complete Genome Sequence of a Carbon Monoxide-Utilizing Acetogen, *Eubacterium limosum* KIST612. J Bacteriol 193:307–308

Rydzak T, Levin DB, Cicek N, Sparling R (2009) Growth phase-dependant enzyme profile of pyruvate catabolism and end-product formation in *Clostridium thermocellum* ATCC 27405. J Biotechnol 140(3–4):169–175

Schmidt MWI, Torn MS, Abiven S, Dittmar T, Guggenberger G, Janssens IA, Kleber M, Kögel-Knabner I, Lehmann J, Manning DAC, Nannipieri P, Rasse DP, Weiner S, Trumbore SE (2011) Persistence of soil organic matter as an ecosystem property. Nature 478:49

Shendure J, Ji HL (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Singh BK, Bardgett RD, Smith P, Reay DS (2010) Microorganisms and climate change: terrestrial feedbacks and mitigation options. Nat Rev Microbiol 8:779–790

Smith SR, Abbriano RM, Hildebrand M (2012) Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. Algal Res 1:2–16

Stein LY, Bringel F, DiSpirito AA, Han S, Jetten MSM, Kalyuzhnaya MG, Kits KD, Klotz MG, Op den Camp HJM, Semrau JD, Vuilleumier S, Bruce DC, Cheng J-F, Davenport KW, Goodwin L, Han S, Hauser L, Lajus A, Land ML, Lapidus A, Lucas S, Médigue C, Pitluck S, Woyke T (2011) Genome sequence of the methanotrophic *Alphaproteobacterium Methylocystis* sp. Strain Rockwell (ATCC 49242). J Bacteriol 193:2668

Strickland MS, Rousk J (2010) Considering fungal:bacterial dominance in soils – Methods, controls, and ecosystem implications. Soil Biol Biochem 42:1385–1395

Su H, Zhang T, Bao M, Jiang Y, Wang Y, Tan T (2014) Genome Sequence of a Promising Hydrogen-Producing Facultative Anaerobic Bacterium, *Brevundimonas naejangsanensis* Strain B1. LID - 10.1128/genomeA.00542-14 [doi] LID - e00542-14 [pii]. Genome, Announc

Vignais PM, Billoud B, Meyer J (2001) Classification and phylogeny of hydrogenases. FEMS Microbiol Rev 25(4):455–501

Wang J, Suzuki T, Dohra H, Takigami S, Kako H, Soga A, Kamei I, Mori T, Kawagishi H, Hirai H (2016) Analysis of ethanol fermentation mechanism of ethanol-producing white-rot fungus *Phlebia sp*. MG-60 by RNA-seq. BMC Genomics 17(1):616

Wong YM, Juan JC, Gan HM, Austin CM (2014) Draft Genome Sequence of *Clostridium perfringens* Strain JJC, a Highly Efficient Hydrogen Producer Isolated from Landfill Leachate Sludge. Genome Announc 2:e00064–e00014

Woodward FI, Bardgett RD, Raven JA, Hetherington AM (2009) Biological approaches to global environment change mitigation and remediation. Curr Biol 19:R615–R623

Yadav S, Dubey SK (2018) Cellulose degradation potential of *Paenibacillus lautus* strain BHU3 and its whole genome sequence. Bioresour Technol 262:124–131

Yu W-L, Ansari W, Schoepp NG, Hannon MJ, Mayfield SP, Burkart MD (2011) Modifications of the metabolic pathways of lipid and triacylglycerol production in microalgae. Microb Cell Factories 10:91

Zhao XQ, Zi LH, Bai FW, Lin HL, Hao XM, Yue GJ, Ho NWY (2012) Bioethanol from lignocellulosic biomass. Adv Biochem Eng Biotechnol 128:25–51