Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2582-2845.8705

ISSN: 2582 – 2845

Ind. J. Pure App. Biosci. (2021) *9*(3), 39-47

Review Article

Indian Journal of Pure & Applied Biosciences

Peer-Reviewed, Refereed, Open Access Journal

Next Generation Sequencing Approaches Deciphering Hidden Microbial Treasure in Soil

Ruchi Srivastava[1,2], Alok Kumar Srivastava[2], Pramod W. Ramteke[1,3*], Vijai Kumar Gupta[4] and Csaba Vágvölgyi[5]

[1]Sam Higginbottom University of Agriculture Technology and Sciences, Prayagraj, 211007, India

[2]ICAR-National Bureau of Agriculturally Important Microorganisms, Mau, 275103, India

[3]Department of Molecular Biology & Genetic Engineering,
Rashtrasant Tukadoji Maharaj Napgur University, Nagpur, 440033, India

[4]Centre for Safe and Improved Food & Bio-refining and Advanced Biomaterials Research Centre
Scotland's Rural College (SRUC), Edinburgh EH9 3JG, Scotland UK

[5]Department of Microbiology, Faculty of Science and Informatics,
University of Szeged, Közép fasor 52, H-6726 Szeged, Hungary

*Corresponding Author E-mail: pwramteke@gmail.com

Received: 3.05.2021 | Revised: 7.06.2021 | Accepted: 12.06.2021

ABSTRACT

Soil is one of the most important and complex biological habitats on earth. As we know the microbes are important key players in every ecosystem, and biological and ecological processes. Thus, it is necessary to understand this microbial treasure to have information about their role in such processes. Initial culture dependent methods helped a lot but are insufficient to indentify all the microbial species present in the soil. It has been estimated that only ~1% of bacterial species are cultivable on culture medium and rest are still hidden in through such methods. On the other hands, soil metagenomics is a modern concept that allows us to recognize these hidden species without biasness of growing bacteria on to petri plates. In last two decades rapid improvements in modern techniques itself enhanced the human capabilities in not only identifying but also have an understanding of functional aspects of these microbes in soil. Present review describes the available culture dependent methods and emergence and improvement in modern sequencing approaches helping to explore soil microbial diversity of more detail.

Keywords: *Diversity, Metagenomics, Sequencing.*

INTRODUCTION

In 2020 humans know how to land and develop their colonies on mars but we are less friendly and less aware to our soil. What it consist how our activities usually called as anthropogenic constitute a set of microbial species or destroy the more suitable microbial network in a given set of environment.

Cite this article: Srivastava, R., Srivastava, A. K., Ramteke, P. W., Gupta, V. K., & Vágvölgyi, C. (2021). Next Generation Sequencing Approaches Deciphering Hidden Microbial Treasure in Soil, *Ind. J. Pure App. Biosci. 9*(3), 39-47. doi: http://dx.doi.org/10.18782/2582-2845.8705

Before jumping into methods and ways to study theses microbes first we have to understand what is microbial diversity and why do we study it. In terms of information technology diversity is defined as amount and distribution pattern of information in a community, while biologically a range of different kinds of organisms and the relative abundance in a community (Torsvik et al., 1998). Diversity can be described in three ways genetic, species and ecological diversity. Genetic diversity refers within different species, whereas, species diversity comparatively describe two different terms species richness refers total number of species and other is quantitative variation among species (Simental-Rodríguez et al., 2014).

Microbial diversity: Conventional approach

Identification of microbial species in rhizospheric zone is an essential and important step in answering the following questions, i.e. (i) group of microbial species associate with particular soil and crop type, (ii) recognizing the beneficial bacteria and factor influencing such species, (iii) identification of pathogenic microbial species and (iv) how a microbe communicate and affect other in the niche. Species, genetic and ecosystem diversity constitute the soil microbial diversity (Solbrig, 1991). Species richness corresponds to total number of species present in an ecosystem and environment factor greatly influence the presence of microbial species. An ecosystem may support to microbial species as compared to other and it varies. On the other hand species evenness denotes the distribution of a species in a given set of environment. Measuring the soil microbial diversity initially started with cultivation, identification and categorization of microbial group, but is limited due to taxonomic and methodological inadequacy (Fakruddin et al., 2013). As all the microbial species present in rhizospheric soil have to be culture in or on microbial media that reduced the chance of identification of microbial species in environment, because it is difficult to grow all microbes in lab and is time taking too. Hence, at this point appearance of advance sequencing methods, given a hope for rapid identification of almost all the microbial species, present in environment either cultivable or uncultivable.

Morphological and Biochemical Identification of soil microbial population

The identification of bacteria was based on morphological characteristics and biochemical tests. Morphological characteristics observed for bacterial colony included colony appearance; shape, elevation, edge, optical characteristics, consistency, colony surface and pigmentation and microscopic examination of isolates with simple staining and differential staining (Gram- and spore-staining). Biochemical characterizations were done according to the method of Bergey's Manual viz. Catalase test (Hayward, 1960), Deep glucose agar test Voges-Proskauer (VP) test, Methyl red test, Hydrolysis of casein, Hydrolysis of starch, Utilization of Citrate (Simmon, 1926), and propionate, Nitrate reduction test, Production of indole, Oxidase test (Gaby & Hadley, 1957), Motility test by wet mount method, Urease production test and Potassium hydroxide solubility test. First line of identification of microbial species based on morphological observation visually and under microscope, based on colony shape, elevation, edge, colour, etc. and their stain identity (Gram +ve or –ve), spore structure and position under microscope. The next step to identify these microbes based on biochemical characteristic they utilized or produced. Common test for identification of microbes up to genus level includes oxidase, peroxidase, catalase, Voges-Proskauer (VP), indole production, casein and starch hydrolysis, urease, citrate etc. Preliminary molecular methods for identification were also culture dependent. Extraction of genomic DNA and amplification of rDNA followed by sequencing of amplified product and phylogenic analysis are the basic step to be followed for identifying any microbes. In addition to rDNA amplification other conserved genes e.g. RNA polymerase beta subunit (rpoB), gyrase beta subunit (gyrB), and recombinase A (RecA), have also been explored to study the soil microbial community profile (Ghebremedhin et al., 2008). Such amplified gene products can be easily analyzed using various genetic fingerprinting techniques i.e. denaturing gradient gel electrophoresis (DGGE), terminal restriction fragment length polymorphism (T-RFLP), amplified ribosomal DNA restriction analysis (ARDRA), BOX and ERIC-PCR, and RAPD-PCR fingerprinting to generate microbial community profile on the basis of either sequence or length polymorphism. These techniques demonstrate the differences between microbial communities but fail to provide direct taxonomic identities.

Metagenomics

The modern way of identification of microbes in a given set of environment is to use metagenomics approach. In last two decade thousands of researches has been conducted for identification of microbes based on environmental DNA sequencing. This approach not only useful in agricultural soil (Velázquez-Sepúlveda et al., 2012; & Peiffer et al., 2013), municipal septic tanks, or biogas plant biodiversity but also enabled us to explore hidden microbial treasure in extreme natural condition. Various researchers have reported diverse microbial communities in following environment; rhizospheric soil of Antarctic region (Teixeira et al., 2010; & Renu et al., 2020).

Modern sequencing platform

After 3 billion dollar first Human Genome Project which was completed in 2003 and used Sanger sequencing capillary electrophoresis method. Advent of high throughput short reads parallel sequencing is a way more different and advance method of sequencing often called Next generation sequencing. These techniques sincerely revolutionized the sequencing of modern genomics studies. Next Generation method has several advantages as they don’t require a prior knowledge of genome. In addition to this can resolve single gene alteration, spliced transcripts, single nucleotide polymorphism (SNP), and any allelic gene variants too. Moreover their template requirement for DNA/RNA is very less in nanogram, and has high reproducibility. Various platforms popular now a days are as follow:

Illumina Sequencing

From mid 1990s with the initial thought provoking in Cambridge Scientists and formation of Solexa in 1998, followed by first Genome analyzer by Solexa in 2005, and further Solexa takeover by Illumina in 2007 are the historical developments of next generation technology we are exploring now. Illumina offers flexibility in application of this technology in genomics, and transcriptomics as well. Basic steps in Illumina sequencing are sample preparation, cluster generation, sequencing and data analysis. The very first step in illumina process is to breakdown the environmental DNA in to short fragments of nearly 200-600bp, followed by attachment of adapter sequence to above fragments. The next step is to add primers on complementary strand and wash out unattached primers. Cluster generation is isothermal amplification of molecules in flow cell which is a glass slide coated with different oligos through a process bridge amplification. Polymerase amplified to strand attached with oligos and original DNA fragments washed away. The sequencing start with the identification of first base by generating a fluorescence signal respective for each nucleotide in the chain.

Roche 454 sequencing

Initial library preparation is slightly different in this method from Illumina, in roche we used spray method in order to breakdown DNA into shorter fragment of around 300-800bp. Addition of adapters and, amplification, followed by denaturation, cloning into specific vectors and generation of single stranded DNA are the steps in Roche 454 sequencing. Amplification of DNA in this method carried out by immobilization of DNA on to beads immersed in emulsion. It provide a larger space for amplification of DNA. In order to perform this all component mix with emulsion oil rotate to generate water droplets and at high speed, and ideally each droplet can carry one DNA template and PCR component. Beads wrapped with such droplets, and oligos present over bead can form a complex with complementary adapter. Pyrosequening required single strand DNA binding protein and polymerase for prior to sequencing. A PTP plate comprised of nanopore where each pore can accommodate only one bead that is suitable for sequencing. In pyrosequencing when a dNTP is attached to template DNA a pyrophosphate group is release, which further reacts with ATP sulfuric acid chemical enzymes to produce ATP trigger fluoresce detected by CCD camera. Each nucleotide produces different fluorescence color of which data is recorded in computer. Longer sequencing read is a major advantage of Roche 454 and average read length is 400bp. Although the inability of Roche 454 technology to accurately measure the homopolymer length is a leading disadvantage that ultimately generate insertion and deletion sequencing error.

Ion Torrent technology:

Life Technologies commercialized the Ion Torrent semiconductor sequencing technology in 2010 (https//www.thermofisher.com/us/en/home/brands/ion-torrent.html). It is similar to 454 pyrosequencing technology but it does not use fluorescent labeled nucleotides like other second-generation technologies. It is based on the detection of the hydrogen ion released during the sequencing process (Rotheberg et al., 2011). Specifically, Ion Torrent uses a chip that contains a set of micro wells and each has a bead with several identical fragments. The incorporation of each nucleotide with a fragment in the pearl, a hydrogen ion is released which change the pH of the solution. This change is detected by a sensor attached to the bottom of the micro well and converted into a voltage signal which is proportional to the number of nucleotides incorporated. The Ion Torrent sequencers are capable of producing reads lengths of 200 bp, 400 bp and 600 bp with throughput that can reach 10 Gb for ion proton sequencer. The major advantages of this sequencing technology are focused on read lengths which are longer to other SGS sequencers and fast sequencing time between 2 and 8 hours. The major disadvantage is the difficulty of interpreting the homopolymer sequences (more than 6 bp) (Loman et al., 2012) which causes insertion and deletion (indel) error with a rate about ~1%.

SMAT (Pacific Bio)

Single Molecule, Real-Time (SMRT) is an initiative of Pacific Biosciences was introduced in 2011 (www.pacificbiosciences.com). It is considered the first of a third generation of DNA sequencing instruments because it can sequence single DNA fragments without the need for PCR amplification. Single molecules of DNA polymerase are immobi-lized at the bottom of nanometer scale aperture chambers called zero-mode wave guides (ZMWs) (Korlach et al., 2010). The template-directed primer extension reaction with nucleotides labeled with fluorescence at the end of the phosphate moiety is monitored in real time. The zero-mode wave guides restrict the observation to the volume containing the DNA polymerase. This dramatically reduces the perturbation due to background fluorescence of the labeled nucleotides that are not involved in the nucleotide incorporation reaction. Because the individual base incorporation error rate is quite high, DNA molecules are circularized and read several times to generate a consensus sequence. Reads of several kbases in length have been achieved by the developers, although total throughput of the system is still in the order of Mbases

Pipelines for big genomic data analysis

Advance platform simplified the sequencing process but the massive data generated from these methods also need to be analyzing fast. Simultaneously researchers' also are developing automated pipelines and small tools useful for step wise biological data analysis. Various available sequencing analysis pipelines are the computer scripts and programs comprised of multiple software organized in a definite sequence to execute their respective work. These scripts are generally written in perl, python, R, or UNIX shell. Complexity of the script increases with increasing the steps and features in pipeline. Scripts are arranged in pipeline framework that reduced user's effort to put huge data at each step. Class based framework are codes available for different functions while on the other hands, server based workbench comprised of predefined modules which is also user friendly. MG-RAST, Galaxy, CLC Genomics, Taverna, Meta WRAP, ANASTASIA etc are the common server base pipelines. Initially algorithms were developed to recognize the similarity between sequences (Altschul et al., 1990), and this capacity revolutionized the biological data science research. Large data set has been used for functional analysis because their sequences as probe were already available in NCBI databank. High throughput technologies usually generate large data commonly known as FASTQ file and this text data actually serve as raw data for further downstream processing, and require alignment, trimming, and comparison of raw data to either reference genome or so called de novo assembly. For amplicon or metagenomic analysis extracted from environmental samples various pipelines are available today. Processing of data starts with the conversion of raw reads into fastq file format, and now these raw aligned based on pair ends followed by combination and merging of pair ends in order to obtain a complete sequence for further processing of data. The next step is to check the quality of reads aligned with pair end, and to perform through individual program i.e. USEARC or in analysis pipelines such as QIIME.

Metagenomic analysis: It is used for microbial diversity analysis primarily target 16S rRNA ribosomal subunit, which is ~1500 bp long nucleotide sequences in prokaryotes, and comprised of nine conserved and hyper variable regions i.e. V1-V9. Illumina and MiSeq sequencing platform targets V3-V4 region that results in amplicon size of around 459bp, followed by pair end analysis of 250bp overlapping above region. Illumina using reporter proprietary software performed initial QC check. Now this raw data can be processed and analyzed further in any pipeline available. Various tools and pipelines for metagenomic data analysis are listed in Table 1. Metagenomic data generated from Oxford Nanopore Technology (ONT) can be analyzed using EPI2ME, or individual software available in nanotools. EPI2ME is a cloud based platform for Nanopore sequencing data gives freedom of data analysis in real-time. What's in my pot (WIMP) is workflow offer rapid identification of microbes i.e. bacteria, fungi and virus and generate interactive phylogenetic tree in real-time. Another useful workflow for RNA Sequencing by ONT is Master Of Pores (Cozzuto et al., 2020). ONT is a product of third generation sequencing technology that can perform the estimation of DNA and RNA without bias of amplification. Beside this important method to decipher novel catabolic genes in metagenome from different environment, is Substrate Induced Gene Expression (SIGEX) screening (Yun & Ryu, 2005). ONT and PacBio generate much longer read and thus there are different workflow for long read analysis. NanoGalaxy is a workflow with multiple tools specifically for long read analysis (Koning et al., 2020). Another important achievement is development of PICRUSt to predict function based on marker gene sequence. PICRUSt interpret function from input dataset by comparing it OTUs of large database it contains and predict the function.

RNASeq pipeline**:** RNASeq analysis is an advance effective method of comparing transcriptome in two different states. FASTQ file read of 30-400bp as per the sequencing platform used served as raw material for RNASeq pipeline. FASTQC file aligned and trimmed using *Cutadopt* and *trimmoatic* and further aligned through tools i.e. Star and *Tophat*. Downstream analysis can be performed via *Fragments Per Kilobase of transcript per Million* (FPKM) analysis that uses *cufflinks* for transcript assembly, *cuffmerge* to merge the files and cuffdiff for differential analysis.. On the other hand count based analysis can be carried out by read count using HtSeq, genomic alignment, and differential expression using *DESeq*, *Limma*, *EdgeR* etc.

ChipSeq analysis: Chromatin immune-precipitation sequencing is an important method in epigenomic studies. Histone modification analysis describes the role of epigenomic in cell structure and function.

Bioinformatic and statistical tools for diversity and functional analysis

The next step after getting raw reads from various sequencing platforms is to perform downstream analysis of microbial data. Various computational tools are available for microbes and plant associate microbiome analysis (Sarim & Patel, 2017). The fundamental steps in every bioinformatics tools are aligning, annotation and comparison of sequencer generated microbial data to available database or denovo assembly of sample data. Despite the platform, pipeline or tool used for sequencing and analysis of genomic data the fundamental questions are what kind of microbes is present and which genus or species is more abundant in a particular set of environment. In addition to this which gene or pathway is over expressed in such microbes that actually making theses microbe enable to sustain and grow in an environmental condition. Here the various statistical methods give users a better understanding of above mentioned questions. Alpha and beta diversity are the initial estimation of microbial species. Estimation and representation of alpha diversity can be visualize using R package, box plot, QIIME, USEARCH etc., and statistical validation can be performed using Analysis of Variance (ANOVA), and Kruskal Wallis test. The rarefaction curve describes both within sample (Edward et al., 2015; & Zhang et al., 2019). Venn diagram is another way to represent common or unique microbes in samples (Ren et al., 2019). Statistical tools for beta diversity are dendrogram and principal component analysis (Zhang et al., 2018; & Chen et al., 2019). Relative abundance of microbes at species level can be seen through stacked bar plot (Srivastava et al., 2020). In order to estimate similarity in samples correlation analysis can be performed using tools scatter plot etc. Heatmap are the widely used method to represent the correlation either in microbes, metabolic response, and gene and protein expression in any type of stress/environment. Difference in two or more groups can be represented using *volcano* and *Manhattan* plot. The former presents scatter plot exhibiting no of fold changes, no. of difference and p-value (Shi et al., 2019) and later also representing similar feature including different biomarkers and taxonomic information. Another very interactive way to showcase genetic relationship in operational taxonomic unit (OTUs) is phylogenetic tree, treemap and many other software package can performed the task and present information in multiple ways. Beside all these tools often used for statistical validation of genomics data is statistical analysis of metagenomic profile *STAMP*, *MetaViz* (Wagner et al., 2017), *MetaComp* (Zhai et al., 2017), *TAMER* (Jiang et al., 2012). Moreover, a variety of methods are now available for exploring the metabolic network when plant and microbes interact in an environment (Sarim et al., 2020).

Table 1: Tools and Pipelines for genomic data analysis

Tool	Type	Description	Availability
MG-RAST	GUI & CL	MG data analysis automated pipeline	https://www.mg-rast.org/
QIIME2	GUI & CL	Microbiome data analysis pipeline	https://qiime2.org/
MOTHUR	GUI & CL		https://mothur.org/
MEGAN	GUI, CL, macOS	MG data analysis software	http://ab.inf.uni-tuebingen.de/software/megan6/
USEARCH	CL, GUI	Sequence alignment tool comprised of many algorithm	https://www.drive5.com/usearch/
CD-HIT		Tool for rapid clustering of nucleotide or protein sequences	http://weizhongli-lab.org/cd-hit
MetaQUAST	CL, macOS	For unknown species, Huge diversity and detecting chimeric sequences	http://bioinf.spbau.ru/metaquast
Bowtie2	CL	Tool for rapid alignment of large reads of genomic data. It also support the local, gapped, and pair end alignment	https://github.com/BenLangmead/bowtie2
EzMAP	GUI	Java and R based user friendly platform for microbiome analysis including structural,	https://github.com/gnanibioinfo/EzMAP

		community comparison and functional prediction as well	
tidyMicro	CL	R packaged based analysis pipeline for microbiome data with interactive visualization of results	https://github.com/CharlieCarpenter/tidyMicro
ViromeScan	CL	Viral metagenomic data analysis	http://sourceforge.net/projects/viromescan/.
Metavir2		Viral metagenomic data analysis perform taxonomic identification, automatic construction of phylogentic tree, gene richness estimation and comparative virome analysis	http://metavir-meb.univ-bpclermont.fr
Salmon			http://combine-lab.github.io/salmon
Trimmomatic		Software for pair end, NGS read quality analysis,	http://www.usadellab.org/cms/index.php?page=trimmomatic
PICRUSt		Developed to predict the functional potential of bacterial population from 16S rRNA	https://github.com/picrust/picrust2

GUI (Graphical User Interface) CL(Command Line)

CONCLUSION

Enumeration and identification of microbes in an environment is necessary to understand their role in ecosystem. Traditional approaches are effective but time consuming and unable to answer certain question e.g. hidden or unculturable microbial species actively present in the environment that may be a key player in driving various environmental processes. In order to understand almost all the microbial population and their functional role there is a dire need to upgrade the identification method. Advance OMICs technologies are capable to fulfill the above objective. Metagenomics and amplicon sequencing gives an insight of almost all microbes present in an environment and their interaction to others and how soil environment including biotic and abiotic stress affecting their microbial habitat. Advance statistical method i.e. STAMP, correlation analysis, and difference comparisons, ANOVA, various R statistical package increased the visualization and exact calculation of type of diversity in samples. In addition to this pipelines MG-RAST, Galaxy, EP2IME, QIIME etc. performing multiple data analysis in short time. Advent of ONT offer large reads of sequencing and visualization in real time. In addition to this prediction of functions of microbial community based on their marker gene sequence using PICRUSt is advancement in microbial functional analysis. Constant efforts of statistician, computational biologist, molecular biologist etc provided advance tools and pipeline for efficient, more precise and rapid identification of microbes and elucidation of their functional analysis. In near future possibly more cheaper technology for sequencing might be available and even more informative softwares for functional profiling of all microbes to each other and to environment may be developed that can also identify unidentified reads in OMICs data.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Chen, C., Chen, H. Y., Chen, X., & Huang, Z. (2019). Meta-analysis shows positive effects of plant diversity on microbial biomass and respiration. *Nature communications*, *10*(1), 1-10.

Cozzuto, L., Liu, H., Pryszcz, L. P., Pulido, T. H., Delgado-Tejedor, A., Ponomarenko, J., & Novoa, E. M. (2020). Masterofpores: A workflow for the analysis of Oxford nanopore direct RNA sequencing datasets. *Frontiers in genetics*, *11*, 211.

Cozzuto, L., Liu, H., Pryszcz, L. P., Pulido, T. H., Delgado-Tejedor, A., Ponomarenko, J., & Novoa, E. M. (2020). Masterofpores: A workflow for the analysis of Oxford nanopore direct RNA sequencing datasets. *Frontiers in genetics*, *11*, 211.

de Koning, W., Miladi, M., Hiltemann, S., Heikema, A., Hays, J. P., Flemming, S., & Stubbs, A. P. (2020). NanoGalaxy: Nanopore long-read sequencing data analysis in Galaxy. *GigaScience*, *9*(10), giaa105.

Edwards, D., Hawker, C., Carrier, J., & Rees, C. (2015). A systematic review of the effectiveness of strategies and interventions to improve the transition from student to newly qualified nurse. *International journal of nursing studies*, *52*(7), 1254-1268.

Fakruddin, M., & Mannan, K. (2013). Methods for analyzing diversity of microbial communities in natural environments. *Ceylon Journal of Science (Biological Sciences)*, *42*(1).

Gaby, W. L., & Hadley, C. (1957). Practical laboratory test for the identification of Pseudomonas aeruginosa. *Journal of Bacteriology*, *74*(3), 356.

Ghebremedhin, B., Layer, F., Konig, W., & Konig, B. (2008). Genetic classification and distinguishing of Staphylococcus species based on different partial gap, 16S rRNA, hsp60, rpoB, sodA, and tuf gene sequences. *Journal of clinical microbiology*, *46*(3), 1019-1025.

Jiang, D., El-Din, T. M. G., Ing, C., Lu, P., Pomes, R., Zheng, N., & Catterall, W. A. (2018). Structural basis for gating pore current in periodic paralysis. *Nature*, *557*(7706), 590-594.

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., & Turner, S. W. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods in enzymology*, *472*, 431-455.

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., & Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology*, *10*(9), 599-606.

Peiffer, J. A., Spor, A., Koren, O., Jin, Z., Tringe, S. G., Dangl, J. L., & Ley, R. E. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proceedings of the National Academy of Sciences*, *110*(16), 6548-6553.

Renu., Gupta, S. K., Rai, A. K., Sarim, K. M., Sharma, A., Budhlakoti, N., Arora, D., & Singh, D. P. (2019). Metaproteomic data of maize rhizosphere for deciphering functional diversity. *Data in brief*, *27*, 104574.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., & Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348-352.

Sarim, K. M., & Patel, V. K. (2017). Deciphering the Effects of Microbiome on Plants Using Computational Methods. *In Plant Bioinformatics* (pp. 329-345). Springer, Cham.

Sarim, K. M., Srivastava, R., & Ramteke, P. W. (2020). Next-Generation Omics Technologies for Exploring Complex Metabolic Regulation During Plant-Microbe Interaction. In Microbial Services in Restoration Ecology (pp. 123-138). Elsevier.

Simental-Rodríguez, S. L., Quinones-Perez, C. Z., Moya, D., Hernandez-Tecles, E., Lopez-Sanchez, C. A., & Wehenkel, C. (2014). The relationship between species diversity and genetic structure in the rare Picea chihuahuana tree species community, Mexico. *Plos One*, *9*(11), e111623.

Simmons, J. S. (1926). A culture medium for differentiating organisms of typhoid-colon aerogenes groups and for isolation of certain fungi. *The Journal of Infectious Diseases*, 209-214.

Solbrig, O. T. (1991). From genes to ecosystems: a research agenda for biodiversity: report of a IUBS-SCOPE-UNESCO workshop, Harvard Forest, Petersham, Ma., USA, June 27-July 1, 1991.

Srivastava, R., Srivastava, A. K., Ramteke, P. W., Gupta, V. K., & Srivastava, A. K. (2020). Metagenome dataset of wheat rhizosphere from Ghazipur region of Eastern Uttar Pradesh. *Data in brief*, *28*, 105094.

Teixeira, L. C., Peixoto, R. S., Cury, J. C., Sul, W. J., Pellizari, V. H., Tiedje, J., &

Rosado, A. S. (2010). Bacterial diversity in rhizosphere soil from Antarctic vascular plants of Admiralty Bay, maritime Antarctica. *The ISME journal*, *4*(8), 989-1001.

Torsvik, V., Daae, F. L., Sandaa, R. A., & Øvreås, L. (1998). Novel techniques for analysing microbial diversity in natural and perturbed environments. *Journal of biotechnology*, *64*(1), 53-62.

Velázquez-Sepúlveda, I., Orozco-Mosqueda, M. C., Prieto-Barajas, C. M., & Santoyo, G. (2012). Bacterial diversity associated with the rhizosphere of wheat plants (Triticum aestivum): Toward a metagenomic analysis. *Phyton*, *81*, 81.

Wagner, J., Chelaru, F., Kancherla, J., Paulson, J. N., Zhang, A., Felix, V., & Corrada Bravo, H. (2018). Metaviz: interactive statistical and visual analysis of metagenomic data. *Nucleic acids research*, *46*(6), 2777-2787.

Yun, J., & Ryu, S. (2005). Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microbial cell factories*, *4*(1), 1-5.

Zhai, P., Yang, L., Guo, X., Wang, Z., Guo, J., Wang, X., & Zhu, H. (2017). MetaComp: comprehensive analysis software for comparative meta-omics including comparative metagenomics. *BMC bioinformatics*, *18*(1), 1-16.

Zheng, W., Zhao, Z., Gong, Q., Zhai, B., & Li, Z. (2018). Responses of fungal–bacterial community and network to organic inputs vary among different spatial habitats in soil. *Soil Biology and Biochemistry*, *125*, 54-63.