Crimson Publishers
Wings to the Research

Review Article

Bio-Informatics Analysis of Meta-Transcriptomics Sequencing

Siddharth Vats*

Faculty of Biotechnology, Institute of Biosciences and Technology, Shri Ramswaroop Memorial University, India

ISSN: 2689-2707

***Corresponding author:** Siddharth Vats, Faculty of Biotechnology, Institute of Biosciences and Technology, Shri Ramswaroop Memorial University, Barabanki- 225003, UP, India

Submission: April 7, 2021

Published: June 02, 2021

Volume 2 - Issue 5

How to cite this article: Siddharth Vats. Bio-Informatics Analysis of Meta-Transcriptomics Sequencing. Trends Telemed E-Health 2(5). TTEH. 000549. 2021. DOI: 10.31031/TTEH.2021.02.000549

Abstract

This mini review has focused on the bioinformatics based meta transcriptomics sequencing. Meta transcriptomics is the study of expression of RNAs, their regulation in and among communities of organisms. The level of complexity increases from one organism to community, from one species to community of species. The different types of software which are used and can help in sequencing are explained in detail here. How this different software is integral part of meta transcriptomics and helps in understanding their expression level, and how the expression level is influenced by the pathological, physiological and biological conditions in organisms.

Introduction

Genetic materials study has become much easier with the development of Bioinformatics [1-3]. Table 1 gives a complete description about bioinformatics and how it has helped the scientist community to have a clear understanding about the genetic biomolecules. Meta transcriptomics is the study of expression of RNAs, their regulation in and among communities of organisms. Meta-transcriptomics is a type of extension of RNA sequence from individual species to the level of communities of species. The level of complexity increases from one organism to community, from one species to community of species. This compounds challenge of data analysis. Meta transcriptomics sees a steady growth in next generation sequencing. To have better understanding about the microbes, it's always required to culture that microbe and study its characteristics and properties. There are microorganisms; some of them are easy to culture while some do not get cultured. Meta genomics study that deals with genetic material recovered/obtained from environmental samples especially soil. That is why this field also termed as eco-genomics, environmental genomics or community genomics. Meta-genomics explores the potential associated with these genomes while transcriptomics tells us about the genes which are most expressed or active in that specific environment. There are two parts that forms the DNA or RNA, exons and introns. Both exons and introns form the stretch of the DNA which gets transcribed (Transcription unit) into RNAs (Transcripts) and encodes at least one gene. This gene (RNA, the transcription produces) if encodes protein forms messenger RNA (mRNA). A single gene can produce multiple different RNAs (i.e., transcripts). But tissue type, time of development, hormonal factors, and environmental factors do have a major role to play. Generally, at a given time in a cell only single major transcripts will be expressed. The field of science that deals with functional activity and expression of all the transcripts (RNAs) from the environmental samples especially that belongs to microbial community is called Meta-transcriptomics. Each microorganism lives in its microbiome. Microbe and its microbiome have a close relationship, microbiome helps microbe to thrive within it while microbes help in maintenance of the characteristics of the micro biome. Meta-transcriptomics focus on gene activity diversity, highly active main genes and pathways and comparison of gene expression. Next Generation Sequences (NGS), started in the year 2006 Peimbert & Alcaraz, and they could only describe the diversity at taxonomic and community level, community membership of microbes, profiles of microbes, their taxonomic abundance.

Table 1: Bioinformatics of genomic material.

Molecule	Levels	Description	Software#, ##, ###
DNA (Genes)	Genome	Complete set of genetic material of an organism (Exons plus introns).	Homer,off-potter, GIGA,CrispRVariants, CRISPOR, Breaking CAS, LOVD 3.0, Phred, DIALIGN, off spotter, SpliceCenter, BioGPS
DNA (Genes)	Genomics	Study of all set of genetic material of an organism.	Off-Spotter, CrispRVariants, CRISPOR, BreakingCAS, GIGA, Phred, SpliceCenter, GIGA, DIALIGN, BioGPS, Genome Browser, Velvet, Celera, Metasim, Euler
DNA (Genes)	Metagenomics	Meta genomics study that deals with genetic material recovered/obtained from environmental samples especially soil.	CoMet, STAMP, CD-HIT-OUT, GAAS, Megan, MetaPhlAn, Mocat, Straine
RNA (Transcripts)	Transcriptome	Complete range of all types of RNAs present in an organism or a system or a tissue type.	Transcriptome Analysis Console (TAC) software, StringTie, Cufflinks, Scripture, Trinity, Oases, TransABySS
RNA (Transcripts)	Transcriptomics	Transcriptomics tells us about the genes which are most expressed or active in that specific environment.	SpliceSeq
RNA (Transcripts)	Meta-transcriptomics	Meta transcriptomics helps in understanding their expression level, and how the expression level is influenced by the pathological, physiological and biological conditions in organisms. This in simple terms means study of RNA expression among communities of organisms.	Bowtie, Bowtie2, tophat, BWA Velvet, Trinity cluster results, SOAPdenovo, Celera, Functional Annotation tool like BWA, BLAST, FAST and BLAT, BLASTX DIAMOND and USEARCH, QuickGO, Blast2GO, REVIGO, AmiGO, Gorilla, GO-FEAT, SAMSA2
Proteins	Proteome	All set of proteins produced by a living organism or any system.	LCMS/MS, MALDI-TOF
Proteins	Proteomics	The study of proteome or some proteins. It deals with the study on their abundance, location, turnover number, post translational modifications, their activity and how do they interact.	Predic Protein, PEAKS CMD, PEAKS Studio, PANTHER-PSEP, SIFT, Mutation Taster, MuPIT, LS-SNP, MaxQuant, Peaks, OpenMS
Proteins	Meta-proteomics	Study of all types of proteins recovered from environmental sources.	eggNOG mapper, MetaGOmics, ProPHAnE, Unipept, MEGAN, MPA, mPies

One of the major sequencing techniques was amplicon sequencing of 16s/18s rRNA. Amplicons are source/products of gene amplification/PCR/Gene duplication. The sequence analysis technique focuses upon "who is there" in that microbiome. Whole Genome Shotgun (WGS) Metagenomics (WGSM) became popular, after the introduction of 454 pyrosequencing. WGSM threw light on environmental taxonomic and functional diversity, to a limited extend. Metagenomics do play a role in the describing the potential outcome but to have deeper analysis of functional profile of microbial community. Metagenomics is useful in determining "what can they do". And meta-transcriptomics is associated with "who is doing what" in microbiome. It exploits the RNA sequences to understand genes as well as pathways which are active, what are their active functions and what are the taxa that are responsible for those active functions and pathways.

Meta Transcriptomics Sequencing

Meta transcriptomics allow studying the whole gene expression profiles of cultural and non-cultured microbial communities. Meta transcriptomics and meta-genomics both help each other. But there is a major difference as in metagenomics we study the genomic content, but then meta-transcriptomics take the lead in differentiating active and nonactive both genes, based on expression. Meta transcriptomics helps in understanding their expression level, and how the expression level is influenced by the pathological, physiological and biological conditions in organisms. Meta-transcriptomics also helps in differentiating functions of microbial communities, compositions of microbial communities, active vs non active microbial communities.

Software used in meta-transcriptomics

Rapid advance in sequencing technologies.

Kallisto: This software helps in estimation of abundance of RNA-seq transcript.

Kraken: There can be short and long sequences, but kraken software helps in taxonomic labeling to reads of short sequences.

HumanN2: Path analysis and gene transcription analysis across community. RNA and DNA sequencing of the same living sample are studied for getting a correct RNA transcription.

PanPhlAn: This software helps in analysis of strain specific transcriptomics. RNA and DNA sequencing of the same living sample are studied for getting a correct RNA transcription.

Advantages of Meta-Transcriptomics

Microarray was one of the New Generation Sequencing (NGS) technique that was used to study the expression of genome of an organism, but only if the genomic sequence of the organism is

known. Another NGS technique to study the transcripts (RNAs) was Expressed Sequence Tags (ESTs) and this forms the basis of most of available transcriptomics and transcriptome studies Meta-Transcriptomic sequencing technique provide an insight to both culturable and non-culturable microbial/living organisms transcriptome (set of all RNA transcripts, coding mRNA and non-coding sequencing from individual cells, tissues, or population of cells) information, from all microbial communities from specific/nonspecific samples from environment (Figure 1). Meta-transcriptomics mainly focus upon the mRNA out of all transcripts. As, focus on mRNAs give an overview of functional profile and gene expression of the microbiome.

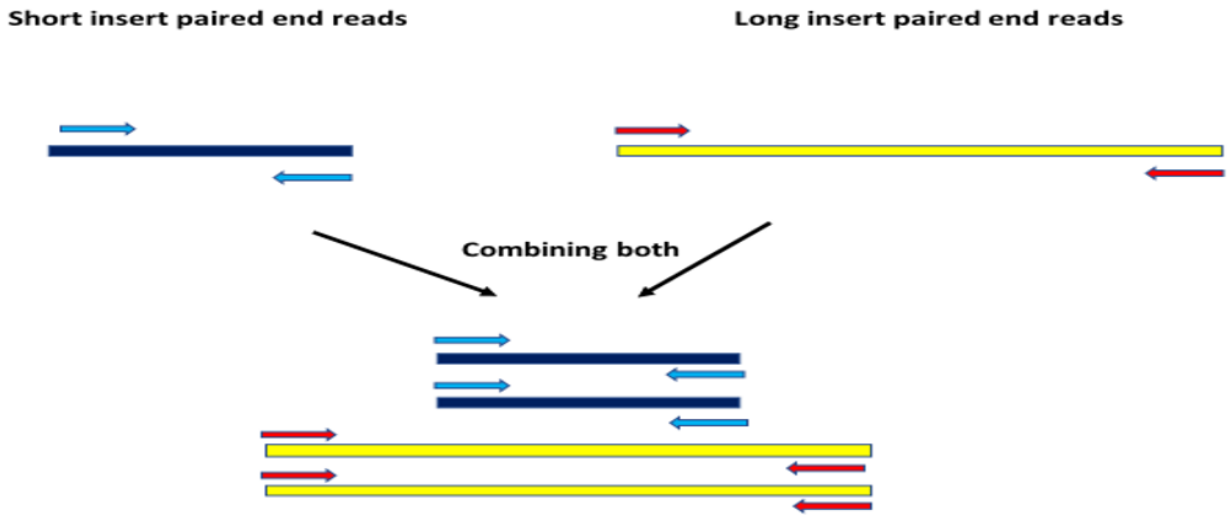

Figure 1: *De novo* assembly.

Preprocessing

There are steps which are crucial and important to perform met transcriptomics. These steps are termed as preprocessing steps (Figure 2). Gene expression of some model organism was done with microarray, but for meta-transcriptomics NGS technique is better.

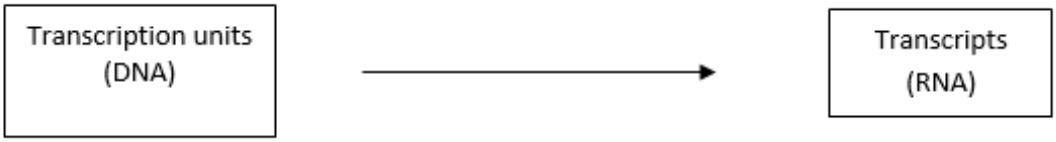

Figure 2: Formation of transcripts from the transcription Units.

De novo assembly

De novo assembly is carried out for novel meta-transcriptomes, with no reference sequence or transcriptomes available to make alignment with. In *de novo* sequencing the reads are arranged as contigs, the sequence which come out as consensus sequential region, decides the quality of the *de novo* sequence, which in turn depend upon the

a) Length of the contigs

b) Continuity of the contigs

c) Gaps of the contigs

To perform *de novo* sequencing in meta transcriptomics, is to choose diverse sequence in different sizes, contigs with Short Paired Reads (SPR, where the ends are known sequences, separated by short (length known) but unknown sequence) as well as Long Insert Paired End Reads (LIPER, whose Ends reads are known, length gap is known, but the gap sequence is unknown) (Figure 3). Combining of SPR and LIPER sequences is the best way to cover the genome. SPR are used to fill the gap in the LIPER help to sequence at both higher and lower read depths. There is various software available that are helpful in performing *de novo* assembly and alignment. Velvet is algorithm is developed by Daniel Zerbino and Ewan Birney. Velvet *de novo* assembler is used to build long continuous reads/sequ. This assembler builds contigs, SPR, LIPER and finally scaffolds from these reads. This software is very helpful in studying sequence data from unknown sources, unmapped reads or new organisms, which do not have any reference sequence or have not assembled yet. Velvet is focused upon de Bruijn graph building with the help of reads and removes the error from de Bruijn graph. Velvet tries to clarify the repeats based on the information, whether the reads are long reads or paired end. Velvet is based on two software *velvtg* and *velveth* and are used together. Zerbino [4], has described about velvet software for de-novo assembly in the article "using the velvet *de novo* assembler for short read sequencing technologies".

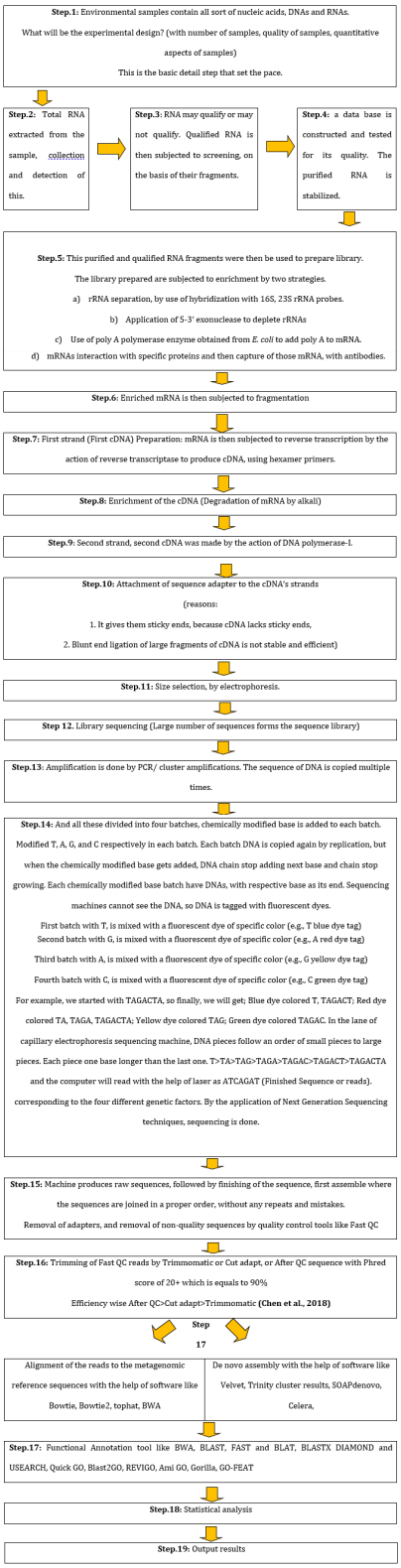

Figure 3:

SOAP *de novo*

This method is novel for short read assembly, which build *de novo* draft assembly, which are in the range of human genomes sizes. This program is very suitable for assembling for Illumina genome analyzer (GA) short reads. This software provide tool for creating sequences which can act as reference sequences for understanding unexplored sequences in a cost-effective way. SOAP *de novo* work well for large genomes of animal, plants, bacteria and fungi. It runs on Linux of 64 Bit. SOAP *de novo* method works one genomic sequence and represent the reads with de Brujin graph

and removal of error full connections on the graph by involving any of the following steps either single or in combination in any order.

a) Clipping the tips.

b) Removal of low average links.

c) Resolving the tiny repeats.

d) Merge bubbles

Then it breaks all the repeats and produce contigs (single/paired), which are then mapped to create a scaffold from them. The gaps present in theses scaffolds are then closed or filled. These scaffold graphs are then used to create transcripts.

Trinity

This software tool assembles transcripts reads from illumine RNA-Seq data. This tool is developed by the University of Jerusalem and Broad Institute. This tool is efficient, robust and fast in *de novo* reconstruction of transcriptomes. Trinity is the combination of three independent software, namely Butterfly Inchworm, Chrysalis which are applied in a sequence to process RNA-seq data. Trinity work by distributing or partitioning RNA seq Data into many/individuals de Bruijn graph. Where each graph represent complexity at transcriptional level for a given locus or gene. Each graph is then undergoing, into processing and produce independently processed extracts of spliced isoforms. RNA seq Data derived from paralogous genes is teased out.

Inchworm module: This work upon RNA seq Data, assembles them, into unique sequences. This generate transcripts of dominant isoform into full length transcripts, then provide information about spliced transcripts which have been alternatively spliced and their unique portions.

Chrysalis module: This is the step which use raw material (Contigs) from inchworm module. Clusters the contigs and construct de Brujin graph from them for each cluster. Each cluster have its transcriptional complexity for gene or genes with common sequence. All the reads form the full read set and are then partitioned among available disjoint graphs/Individual graphs.

Butterfly module: Individual graphs are then processed parallelly, locate and trace out the paths of the reads/pair of reads in those disjoint graphs. Report full length transcript isoforms (alternatively spliced) and teasing to separation those reads which are corresponding of paralogous genes.

SPAdes: It is the tool that has been developed by collaboration among St. Petersburg State University, Russia; St. Petersburg Academic University, Russia and University of California, San Diego, USA. This software is free to use and works with PacBio, Oxford Nanopore, Ion torrent and Illumina (Paired end, single reads and mate pairs). SPAdesis basically for Single Cell Sequencing (SCS). SPAdes use the K-mers to develop de Bruijn Graph and carry out graph theoretical operations associated with structure, reads length and coverage in the graph. The assembling of sequence by SPAdes are performed in four steps. Starting with construction of graph from assembly and using multi-sized graph for detecting of chimeric reads, bulges or bubbles and then their removal. Which is followed by adjustment of k-pairs, and then de Bruijn graph construction followed by contigs construction. Gurevich [5] compared the efficiency of various assembler with SPAdes and found that SPAdeshas highest percentage of assembling the genome and highest number of completed genes assembly i.e., 97% and 4017/4324 respectively.

de Novo assembly tools

There are a number of *de novo* assembly tools available. Bioinformatics and High Throughput Analysis team of Faculty of Mathematics and Computer Science, at University of Jena, Thuringia, Germany, available *de novo* assembly tools were compared. *De novo* assembly faces the problem of repetitive regions, different splicing events and non-uniform coverage distribution. Martin, one of the members of that team did the comparative study for nine different RNA sequence data sets from different kingdoms life systems, 1. Homo sapiens 97.5 million reads (pe)/100 bp, strand specific; 2. Human + EBOV 3h 17.2 million reads (pe) 100 bp, unstranded; 3. Human + EBOV 7h 24.6 million reads (pe) 100 bp unstranded; 4. Human + EBOV 23h 26.4 million reads (pe) 100 bp unstranded; 5. Human Chr 1 simulated 60 million reads (pe) 100 bp, unstranded; 6. M. musculus 52.6 million reads (pe) 76 bp, strand- specific; 7. A. thaliana 16.9 million reads (se) 101 bp, unstranded; 8. C. albicans 11.5 million reads (pe) 51 bp, unstranded; 9. E. coli 7.9 million reads (se) 94 bp, strand specific) on ten *de novo* assembly software (namely Trinity, Oases, Trans-ABySS, SOAP denovo-Trans, IDBA-Tran, Bridger, Bin Packer, Shannon, SPAdes-sc, SPAdes-rna). It has been found that Trinity, RNA SPAdes and Trans-ABySS were the software that were better and outperformed other tools in comparison. But it also come out from the comparison that there was no single software that can be used for all kingdom life system.

Challenges in Meta-Transcriptomics Sequencing

There is a basic difference between DNA and RNA sequencing (Table 2). From 16S rRNA genes based traditional techniques of community membership and taxonomic profiling of microbiomes and microbes. Then we moved to shotgun metagenomics which perform random sequencing but of all genomic content. The disadvantage associated with these two techniques is they only reveals the presence or absences of gene or an organism instead their active nature and activity. New techniques are needed to assess the effect of environmental conditions with change in time. Even in the availability of promising meta-transcriptomics techniques, there are still various obstacles that need to be overcome. Most of the meta-transcriptomics generation is based on the data obtained from RNA and it's the ribosomal RNA that is the major chunk of the total RNA harvested. The abundance of one type of RNA under mines the coverage of other types of RNA like mRNA, the prime target of the meta transcriptomics study. Whatever percentage of mRNA is harvested, because of its unstable nature integrity of meta-transcriptomics is somewhere compromised. Before sequencing we have to try to avoid both the problems. Another major problem that haunts the meta-transcriptomics is the challenge of differentiating host and pathogen mRNA.

Table 2: Comparison between DNA and RNA sequencing.

DNA Sequencing	RNA Sequencing
This involves DNA high throughput sequencing, DNA deep sequencing, DNA sequencing, massive parallel sequencing. This also involves whole exome sequencing, targeted sequencing and whole genome sequencing. Whole Genome Sequencing study of entire DNA (double strand), Massive Parallel sequencing study both strands, exome sequencing focus upon mRNA coding region i.e., 1-3% of the entire DNA. Targeted sequence studies subset of genes to analyze DNA regions of interests. If the desired DNA is less in quantity, then the PCR products are used to quantify the DNA.	This called deep RNA sequencing, RNA high throughput sequencing, Massive RNA sequencing, mRNA sequencing, transcriptome sequencing, whole transcriptome sequencing. Targeted RNA sequence can be of small RNA region, or mRNA sequencing. The major difference between DNA sequencing and RNA sequencing is that in RNA sequencing require the collected RNA to be first reversed transcribed to produce cDNA and then amplified. Complimentary DNA or cDNA is the outcome of action of reverse transcriptase acting upon mRNA. cDNA is also used in heterologous expression. This is the type of expression of a desired protein by the help of cDNA of the protein. cDNA is also termed as mRNA Transcripts.

Transcript Taxonomy

New generation of sequencing platforms and bioinformatics tools are the basis for our advancement in meta-transcriptomics and its success. An insight into the microbial communities at organism and species level give us better understanding to create genomic components and taxonomic profile.

Functional annotation

Functional annotation depends upon sequence similarity search tools like BWA, BLAST, FAST and BLAT. These tools rely on matching comparison with reference genomes. The sequences and reads/contigs which does match to the reference genome are then sent for analysis, directly to differential expression study. But there can be possibility for reads/contigs to not to match to any of the reference match/genome. But the diversity at nucleotide level is far greater than what is at the protein level. If the reads/contigs do not match with reference, then BLASTX search is carried out which is for protein NR databases. BLASTX search for matching with protein databases and is a very time taking process. This also require cluster computing. Other tools like DIAMOND and USEARCH can also be used, but cost is one of the factors to keep into consideration. Functional annotation can also be carried out by application of InterProscan5 and KEGG by Sequencing by Hybridization (SBH) method. Reads and contigs which matches are then processed based on enzyme prediction and protein interaction data. Further processing involves two major two major studies, one can be taxonomic databases for taxonomic distribution while other for network representation.

GOFEAT

Functional annotation attaches biological information to genomic element, which is a time consuming and laborious method. Biological information and analysis for functional annotation is done by the help of Gene Ontology database (GO, Database). There are tools which are helpful in finding functional annotation like QuickGO, Blast2GO, REVIGO, AmiGO, Gorilla etc., [6]. GO databases is used as a dictionary for gene functions. Gene functional enrichment can be performed by bioinformatics tools integration like that of NCBI, SEED, KEGG, InterPro, Pfam or UniProt but also come with some limitations [6]. Limitations like cost, no visual interface, complex configuration methods and commands, limited capacity, not easy to export or share the results, etc. Araujo [6] developed a free web based, user friendly online platform for functional annotation and enrichment by homology search. The name of the tool is GO-FEAT, which means Gene Ontology Functional Enrichment Annotation Tool (GO-FEAT) [7,8].

Conclusion

Bioinformatics hold a great potential in meta-transcriptomics. Meta transcriptomics is the study of expression of RNAs, their regulation in and among communities of organisms. Meta-transcriptomics is a type of extension of RNA sequence from individual species to the level of communities of species. The level of complexity increases from one organism to community, from one species to community of species. With new microbial diseases coming up, meta transcriptomics can help in the analysis of their genetic material.

References

1. Bhargava P, Khan M, Verma A, Singh A, Vats S, et al. (2019a) Metagenomics as a tool to explore new insights from plant-microbe interface. In: Plant Microbe Interface (1st edn), Springer, Cham, Switzerland, pp: 271-289.
2. Bhargava P, Vats S, Gupta N (2019b) Metagenomics as a tool to explore mycorrhizal fungal communities. In: Mycorrhizosphere and Pedogenesis Springer, Singapore, pp: 207-219.
3. Gupta N, Vats S, Bhargava P (2018) Sustainable agriculture: role of metagenomics and metabolomics in exploring the soil microbiota. In: Silico Approach for Sustainable Agriculture Springer, Singapore, pp: 183-199.
4. Zerbino DR (2010) Using the velvet *de novo* assembler for short-read sequencing technologies. Curr Protoc Bioinformatics 31(1): 11-5.
5. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8): 1072-1075.
6. Araujo FA, Barh D, Silva A, Guimarães L, Ramos RTJ (2018) GO FEAT: a rapid web-based functional annotation tool for genomic and transcriptomic data. Sci Rep 8(1): 1-4.
7. Chen S, Liu M, Zhou Y (2018) Bioinformatics analysis for cell-free tumor DNA sequencing data. Methods Mol Biol 1754: 67-95.
8. Yang IS, Kim S (2015) Analysis of whole transcriptome sequencing data: workflow and software. Genomics Inform 13(4): 119-125.