INTERNATIONAL JOURNAL OF ADVANCES IN PHARMACY, BIOLOGY AND CHEMISTRY

Review Article

Algorithms and databases: A key to solve genetic equation in Next Generation Sequencing

Shailesh Kumar[1*], Sumit Govil[2], Sheema Sadana[3], A. N. Pathak[1].

[1]Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur, Rajasthan, India - 303002.

[2]School of Life Sciences, Jaipur National University, Jaipur, Jagatpura, Rajasthan, India - 302025.

[3]Department of Bioinformatics, Hans Raj Mahila Maha Vidyalaya Jalandhar, Punjab, India - 144001.

Abstract

Next Generation Sequencing is a high-throughput DNA sequencing technology based on Non Sanger's methods of sequencing. These techniques involve emulsion PCR or bridge PCR for target amplification. These methods are used to reduce PCR bias during amplification. Once amplified sequences are generated by various methods it requires the alignment and assembly algorithms. The coverage technology is used which describes the number of short reads that overlap with each other within a specific genomic region. All these advancement in sequencing produces high accuracy (99.94% in SOLiD). High accuracy of the sequences can be used for exact assessment of sequences, gene expression, annotation and analysis. Various algorithms are developed for various purposes like aligning, mapping and identifying the specific sequence variants in a genome sequence. For visualization and other application like find differentially expressed genes by data mining on NGS data, Galaxy platform (Free public server for Bioinformatics Tools) is used. Storing information in Databases play a very important role for describing the complete knowledge genome and through SNP analysis, we may identify diseases and genetic variations in humans. RNA–seq a new technique based on NGS technology is used to find gene expression and disease diagnosis. This review gives information of databases and Tools that are used for data storage, retrieval and analysis of sequences produced by NGS techniques.

Keywords: Next Generation Sequencing, Transcriptome, Metagenomic, Genome assembly.

INTRODUCTION

Next-generation sequencing named for a new way of sequencing, different from Sanger method, comprises of Second generation and Third generation sequencing techniques. These techniques are High-Throughput DNA sequencing technologies producing billions of sequences in one day.

It is possible by development of fast amplification methods like emulsion PCR (ePCR) and Bridge PCR. These methods increase the accuracy of the amplification by reducing the PCR bias. The amplified DNA is sequenced by using automated sequencing methods like pyrosequencing[1]. The data analysis of these generated sequences requires fast, accurate, and efficient algorithms tools that can handle very short reads (25–500 bp), for sequence alignment, contig assembly, with a resolution of single base precision. This high (99.94%) accuracy of base calling is used for identification of sequence variations which can be used further to make inferences in a variety of applications, such as disease pathology and novel pathogen detection. Next-generation sequencing provides a quantitative approach to study aging, and enables us to survey the genome to find the exact regions and genes that are

affected and obtain information that was never available before. Next generation sequencing (NGS) platforms are currently being utilized for targeted sequencing of candidate genes or genomic intervals to perform sequence-based association studies. With the rapid advances of the NGS technologies, the cost of sequencing has dramatically decreased over the past few years and has made the sequencing of human genomes routine at the genome sequencing centers and core facilities in institutes[2].

Algorithms involved in next generation sequencing:

Algorithms and Tools involved in NGS are related to removal of error prone data, sequence alignment, assembly, base calling and visualization. At present number of integrated tools are also available which includes all process in single program. We are going to discuss about few of them in this review.

TagDust, a program identifying artifactual sequences in large sequencing runs. During library preparation, TagDust program identifies all reads on the basis of user-defined cutoff for the false discovery rate (FDR). This program is used to increase the accuracy of sequencing before alignment and assembly[3].

Sequence assembly in Next Generation Sequencing plays important role. Some assembler like Scaffolded and Corrected Assembly of Roche454 (SCARF) is a next-generation sequence assembly tool for evolutionary genomics that is designed especially for assembling 454 EST sequences against high-quality reference sequences from related species. SCARF is based on algorithm to match 454 contigs with reference sequences and generate a scaffolded contig. SCARF is capable of assembling raw 454 reads; reads are assembled with a de novo assembler prior to SCARFing[4]. There are basically three approaches are carried out for de novo assembly, a) The Overlap/Layout/Consensus (OLC) methods rely on an overlap graph b) de Bruijn Graph (DBG) methods use some form of K-mer graph c) greedy graph algorithms may use OLC or DBG. Number of software are developed using these approaches. Greedy assembler bases SSAKE (First short read assembler), SHARCGS (operates on uniform length and high coverage), VCAKE (an iterative extension algorithm combined in Newbler pipeline for Solexa 454 hybrid). OLC based software are optimized for large genomes like Archne, Celera Assembler, and CAP/PCAP. Celera assembler is evolved to CABOG. De Brujin Graph based software are Velvet, ABySS, Euler (uses a filter process called spectral alignment for removing sequencing errors)[5].

GSNAP program used by Sequencher for alignment of very short reads (14bp) to very long sequences. It supports Illumina-Solexa or Sanger standard FastQ data formats. GSNAP program can align single-end and paired-end next-generation data to a reference sequence. The reference sequence may be in the form of a FastA or GenBank file. GSNAP uses highly efficient methods for compressing the reference sequence and thereby speeding up the program[6]. Reptile tool is incorporated in pipeline of next generation sequence data for short read error correction to increase precision of assembly of sequences[7]. Due to Ultrafast speed i.e. alignment of short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory uses Burrows-Wheeler-Transformed (BWT) impact is high. This scheme is called the Burrows-Wheeler transform, Genomic Next-Generation Universal Mapper (GNUMAP) algorithm unbiased probabilistic mapping of oligonucleotides, from next-generation sequencing for searching genetic mutations. GNUMAP algorithm assigns a number to segments of the reference genome that are only several base pairs long. If read has more than a few base pairs with a lower probability, this algorithms will discard the entire read, which ignores probability and quality data that are supplied by next-generation sequencers. GNUMAP with competing next-generation mapping algorithms including Bowtie and MAQ mapped about 15% more reads than any other algorithm tested[8]. Meta-IDBA another de Novo assembler for metagenomic data which works on two steps. Firstly it tries to partition the de Bruijn graph into isolated components of different species based on an important observation. Then, for each component, it captures the slight variants of the genomes of subspecies from the same species by multiple alignments and represents the genome of one species, using a consensus sequence. When it is compared with other assembler it was observed that it has similar accuracy[9]. GRASS algorithm provides a mixed-integer programming formulation for contig scaffolding problems, which combines contig order, distance and orientation in a single optimization objective to make error free scaffolds of assembled sequences[10]. Tool such as FANSe used for long read algorithms. They mapped a dataset generated on the 454 GS FLX sequencing platform to the E. coli reference genome. For validation of mapping result of FANSe, they randomly chose 20 mapped reads (Indel-free and Indel-containing reads) and manually verified the unique and correct mapping of all these reads using the NCBI nucleotide BLAST tool. This makes is tool of high sensitivity and low ambiguity. It also uses hotspot score to prioritize the processing of highly possible matches and implements modified Smith–Watermann refinement with reduced scoring matrix to accelerate the calculation without

compromising its sensitivity[11]. Numbers of other tools are also used for assembly and alignment we had discussed only few of them refer table 1 for more tools.

RNA-seq is a new technique which is used to analyze gene expression in an organism. It is based on deep-sequencing technologies working on next generation sequencing methods. This is said to be the most revolutionary technique for Transcriptomic analysis till date. RNA-Seq experiments produce data on millions of short reads. The data report the base sequence of the reads and the positions on the genome to which the reads are mapped[12].

The accuracy of sequencing and assembly can solve many problems related to unsolved mechanism of gene expression. RNA sequencing[23] approaches avoid the weaknesses which are associated with Microarrays for RNA profiling. The expression profiling Workflow includes mainly four steps (1) QC: Filter Short Reads, (2) Align and Assemble or Assemble and Align, (3) Computational Analysis: Quantify Expression, or other applications and (4) Data Visualization. At each step of Workflow they use various software for example FASTX Toolkit, Fast QC, R Short Read are used for filtering, TopHat[13] for aligning and for assembly they use Cufflinks, whereas Cuffcompare, Cuffdiff , SAMtools[14], BEDtools[15] , R: edger[16], DEGSeq[17] are used for quantifying the expression and IGV, or UCSC Genome Browser[18] is used for data visualization.

The software listed is based on very refined algorithms which are developed to increases the efficiency of experiments and accuracy in result interpretation. Gene fusion is a phenomenon which plays an important role in the onset and development of some cancers, such as lymphomas and sarcomas. Studies on Gene fusion are possible because of RNA Seq technique; FusionMap[19], TopHat-Fusion[20], FusionFinder[21], deFuse[22] are various software that are developed to detect gene fusion in RNA seq data with remarkable accuracy.

Databases:

There are number of databases are available to store Next Generation Sequence information and RNA-Seq Information. The SRA data model was designed in collaboration with the EBI and the DDBJ under the auspices of the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC's DDBJ/EMBL/GenBank database has been a critical resource in biomedicine. Sequence submission in several ways: first time and occasional submitters can use an interactive interface and upload smaller data sets through a web browser. High-throughput users can submit data via an automated submission pipeline that uses XML to describe metadata and the community-developed Sequence Read Format (SRF) as a common container file format, and all three SRAs use a high-speed file transfer protocol called fasp (Aspera, Inc., Emeryville, CA, USA) that allows users to transfer files at speeds up to 400 Mbps, many times faster than ftp[25]. Cancer Biomarker Database with Next Generation Sequencing Targets this database contains pre-clinical and clinical biomarkers that were identified or validated using patient samples. The biomarkers include in these studies are gene expression biomarkers, miRNA biomarkers, SNP/mutation/deletion biomarkers and protein biomarkers. The most important component of the database is the annotation. Ideally, the annotation should be provided in GFF3 format. Ensembl conveniently provides GTF files for their annotated genomes. NGSmethDB: Database for next-generation sequencing whole genome methylation allows the generation of whole genome methylation maps at single-cytosine resolution. This database uses a web interface based on GBrowse which allows visualizing the methylation data in a genomic context together with many other annotations. These databases allow CpG dinucleotides and second the CAG/CTG pattern, methylated in undifferentiated cells and displaying the methylation among different tissues in the promoter regions of RefSeq genes[26]. NGS Catalog is a database that deposits published information of various NGS studies and their mutation characteristics like simple nucleotide variations (SNVs), small insertions/deletions, copy number variations, and structural variants, as well as mutated genes and gene fusions detected by NGS. This information is an important key for Epigentic and genomic variation studies. Other facilities provided by NGS catalog is user data upload, NGS general analysis pipelines, and NGS software[27]. RNA Seq Atlas is a reference database for gene expression profiling in normal tissues by next generation sequencing[28]. SRA database is another source from where one can download data related to RNA seq studies and can use it to analyze for some new findings.

Platforms used by companies:

Modern high-throughput data collection methods are likely to be highly enabling and transformative to the biomedical sciences. Galaxy is an open-source, scalable, web-based framework for data and analysis tools integration. The Galaxy platform[29] empowers transparent and reproducible research by providing interactive access to popular next generation sequencing tools, genomic interval operations, and visualization at genome browsers. Galaxy on the cloud infrastructure is the feasible solution for us to

perform NGS data analyses, such as whole exome, RNA and Chip- sequencing analyses. An RNA-Seq in Galaxy can be easily created to analyze and visualize transcriptome data. It can be used to find differentially expressed gene, to find alternatively spliced genes, to discover novel genes and exons, to identify novel splice junctions and most importantly for cancer studies to discover gene fusions. Machine specific steps needed to call base pairs and compute quality scores for those calls. This often results in a FASTQ file, which is a combination of the sequence data as a string of A, C, G and T characters and an associated Phred quality score for each of those bases. High throughput sequencing machines, such as the Illumina G1, allows providing their own alternatives to the standard primary analysis solution, called "The Illumina pipeline". Variant calling is an important process of accurate determination of variations (or differences) between a test sample and the reference genome. These variations may be in the various forms like single nucleotide variants, smaller insertions or deletions (also known as indels), or larger structural variants of various categorize such as transversions, translocations, and copy number variants.

Avidas is the data mining and visualization platform[30] at the core of all bioinformatics adding statistical analysis, machine learning, and ontological interpretation within the visualization-driven data analytics framework. The platform has been integrated with state-of-the-art algorithms for analysis and management of next-generation sequencing data on a wide range of computing infrastructures. It supports workflows for Alignment, RNA-Seq, DNA-Seq, ChIP-Seq, and Small RNA-Seq analysis.

Application of NGS in Epigenetics:

Epigenetics is the study of heritable changes in genes function that occur without a change in DNA sequence. NGS-based studies have provided detailed and comprehensive views of epigenetic modifications for the genomes of many species and cell types[31]. High-throughput DNA sequencing approaches promise to help the diagnosis and guide treatment decisions in many diseases and in many patients, and the combined analysis of the cancer genome and the epigenome promises to become a powerful diagnostic and therapeutic tool. DNA methylation has attracted much attention due to the discovery of 5-hydroxymethyl-cytosine and its role in epigenetic reprogramming and pluripotency. Eukaryotic DNA methylation, its role in metazoan genome evolution, epigenetic reprogramming, and its close ties with histone modifications in the context of transcription[26]. The epigenetic influences various processes like Replication, recombination, repair, cell-cycle progression, epigenetic silencing, transcription and chromosomal stability. In Epigenetics, DNA and Protein interactions can be studied using a technique called ChIP. In ChIP, DNA and associated proteins are chemically cross-linked (typically with formaldehyde) and the DNA is fragmented by sonication or digestion with micrococcal nuclease. Proteins cross-linked to DNA are then immunoprecipitated using an antibody specific to the protein of interest. Epigenetics have combined chromatin-immunoprecipitation (ChIP) with next-generation high-throughput sequencing technologies to describe the locations of histone post-translational modifications (PTM) and DNA methylation genome-wide. The ChIP-on-chip approach has proved to be productive for the genome-wide mapping of DNA-binding proteins, nucleosomes and histone modifications. A very advance approach, ChIP-Seq, which combines ChIP with large number of directed parallel sequencing. In ChIP seq technique enriched DNA is directly sequenced, by using Solexa or Illumina platforms and then the reads are mapped to the reference genome. ChIP-Seq has been employed to identify transcription factor binding sites in the human genome for neuron-restrictive silencing factor (NRSF) and signal transducer and activator of transcription[27]. There are number of commercial and open source pipelines like Avadis NGS, DNASTAR, GeneSifter, NextGENe, easyRNASeq[32], ExpressionPlot[33], GENE-Counter[34], RobiNA which works for RNA seq data analysis.

Application of RNA-Seq Data:

The wide application of NGS based RNA seq is to find gene expression, Differential gene expression, to identify genetic annotation, to find protein-protein interactions and Pathway analysis in an organism. It helps in disease classification and their diagnosis, RNA-Seq provides a very powerful tool for high-resolution genomic studies of tissues and cell populations to detect novel mutations and transcripts in cancers, to classify various tumors based on their gene expression patterns. This method can also identify microbial pathogens based on sequence identification[27]. As RNA-Seq methods increased in speed tremendously and significantly reduces cost, sequence-based microbial diagnosis could become approachable and accurate. This technique could be used to identify global changes in microbial populations within humans, such as the gut microbiome, or could be used to identify novel pathogens. In a study on neurological disorders in 2014, with the help highly reliable Sequencing methods researchers had decoded mutations associated with Mendelian as well as more complex

neurological diseases[35]. These methods are used as molecular diagnostic kit for identification of various diseases[36]. These NGS based methods are very important and are helping in decoding the code of life present in the grammar of genome. With the intervention of Computer based tools and algorithms that are developed world wide are making it easier to find most difficult results which are beyond the interpretation level. These tools have also increased the level of accuracy to approximate 100 percent. These in-silico methodologies are playing remarkable role in genomic and clinical based research.

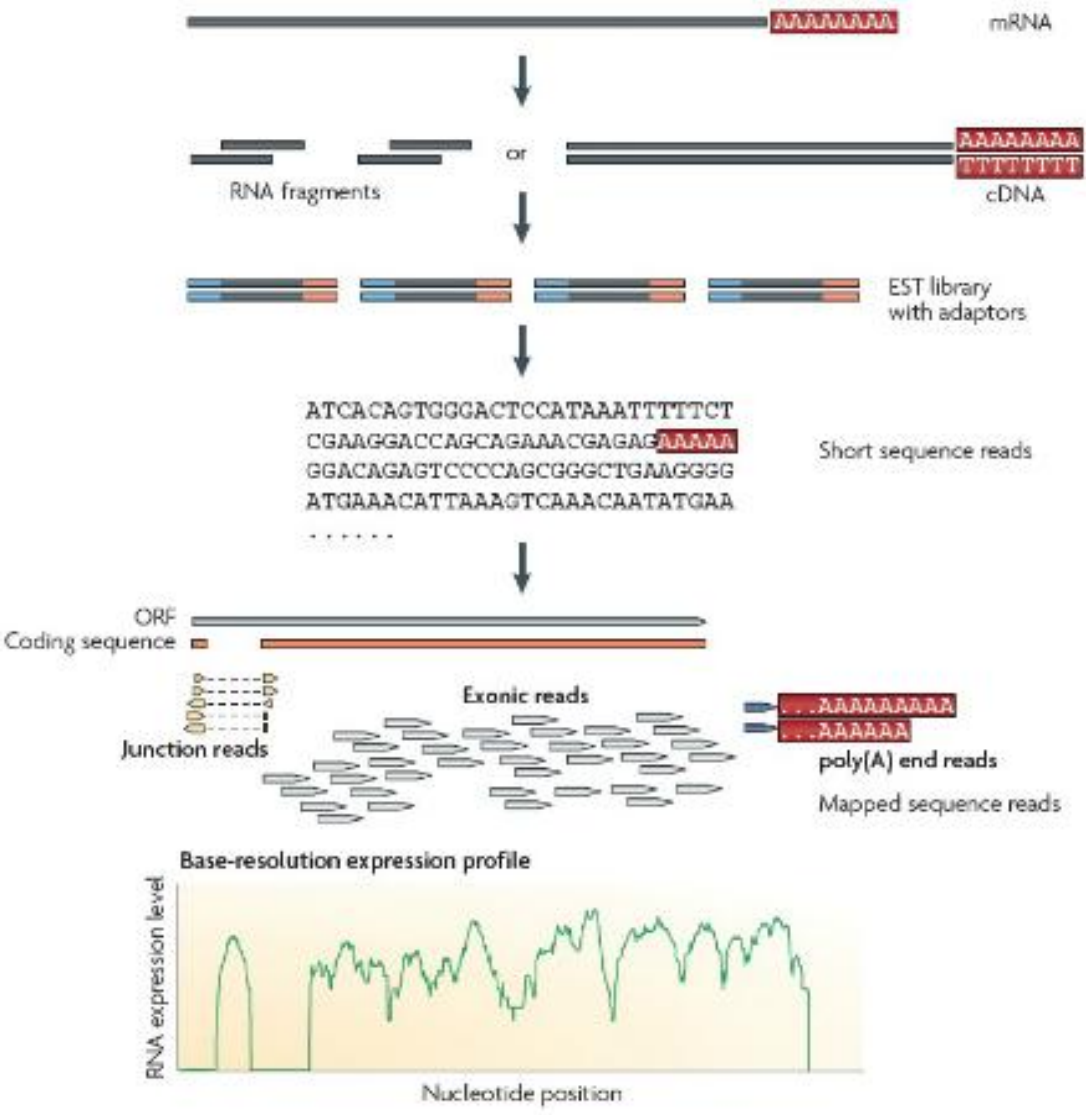

Figure 1: RNA seq Technique. [24]

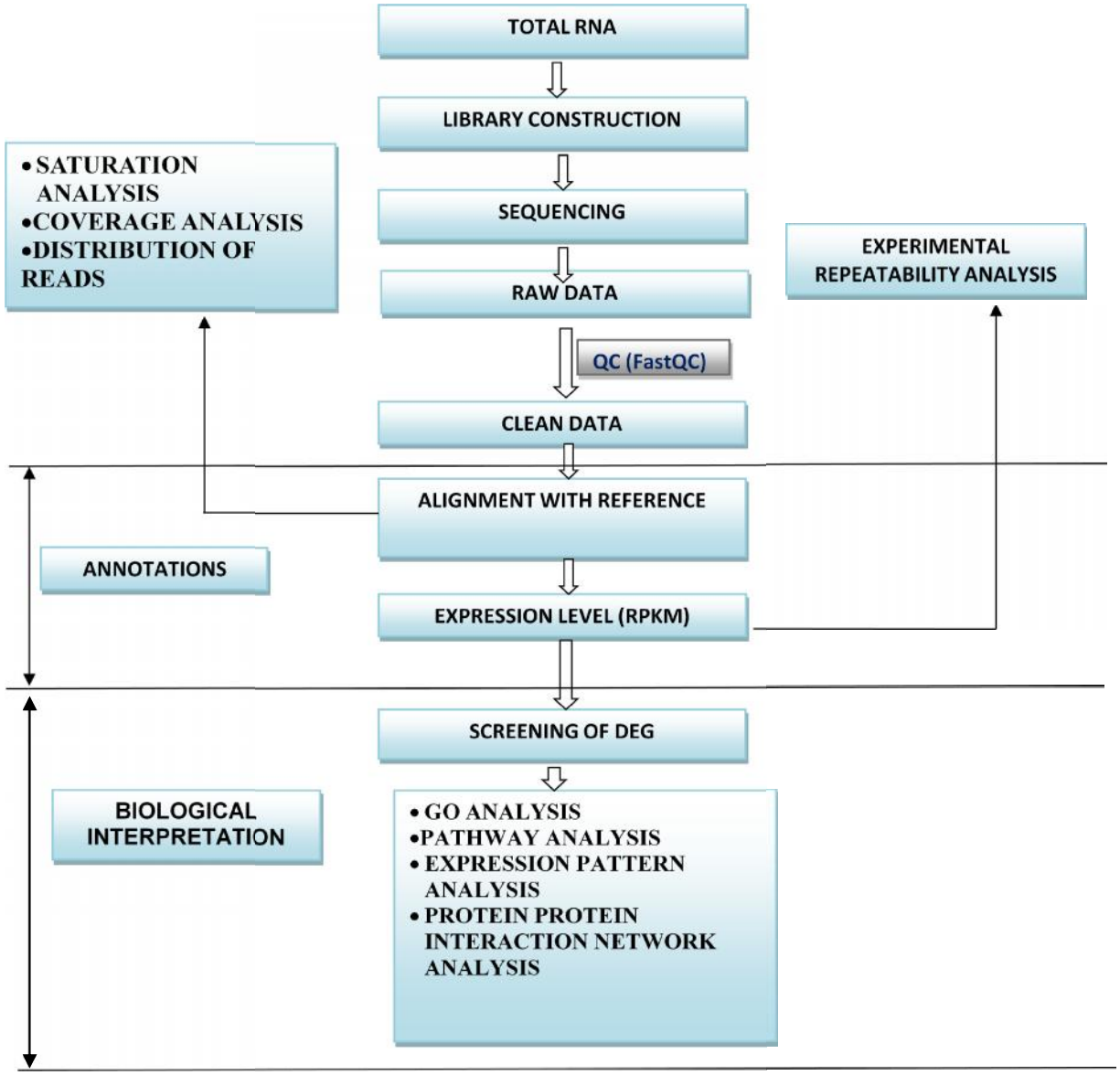

Figure 2: RNA seq Data Work flow

Table 1: List of various software used for Next Generation Sequence Data analysis.

S.No.	Process	Algorithm
1	Alignment	GASSST, MicroRazerS, B-SOLANA, RRBSMAP, ELAND
2	Assembly	Meta-IDBA, Gee Fu, QuRe, Bambus 2, FLASH, Gap5, Est2assembly, GeeFu , QSRA
3	Base Calling	Reptile, TagDust, Swift, SHREC, iCORN
4	Variant Detection	Slider, Slider, VARiD, SVDetect, ACCUSA, VarSifter, TREAT, SVseq, SomaticSniper
5	Visualization	Artemis, Savant, giraffe, CisGenome Browser, CummeRbund
6	Trimming Sequencing Quality Control	Clean reads, ConDeTri, Ea-utils
7	Complete Package	Geneious, Avadis NGS, Lasergene, Pipeline Pilot, SeqMan NGen

Table 2: List of various software used for RNA-seq Data Analysis.

S.No.	Process	Algorithm
1.	RNA seq Filtering	FASTX Toolkit, Fast QC, R Short Read, Flexbar, RSeQC, SAMStat, FLASH
2.	RNA seq alignment	TopHat, GMAP, RazerS, Mosaik , STAR, HMMSplicer
3.	RNA seq assembly	Cufflinks, iReckon, Flipflop, MITIE, RNAeXpress, Scripture
4.	RNA seq expression Quantification	Cuffcompare, Cuffdiff , SAMtools, BEDtools, R: edgeR, DEGSeq, DEXSeq, ERANGE, NPEBseq, rQuant
5.	RNA seq Data Visualization	IGV, UCSC Genome Browser, EagleView , Degust, SeqMonk
6.	RNA seq Annotation	HLAminer, seq2HLA, pasa

REFERENCES

1. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends in genetic, 2008;24(3):133-141.
2. Wang Q, Xia J, Jia P, Pao W, Zhao Z, Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. Briefings in bioinformatics, (2013); 14(4):506-519.
3. Lassmann T, Hayashizaki Y, Daub CO, TagDust—a program to eliminate artifacts from next generation sequencing data, Bioinformatics, 2009; 25(21):2839-2840.
4. Barker MS, Dlugosch KM, Reddy ACC, Amyotte SN, Rieseberg LH, SCARF: maximizing next-generation EST assemblies for evolutionary and population genomic analyses. Bioinformatics, 2009; 25(4): 535-536.
5. Miller JR, Koren S, Sutton G, Assembly algorithms for next-generation sequencing data, Genomics, 2010; 95(6):315-327.
6. Wu TD, Nacu S, Fast and SNP-tolerant detection of complex variants and splicing in short reads, Bioinformatics, 2010; 26(7): 873-881.
7. Yang X, Dorman KS, Aluru S. Reptile: representative tiling for short read error correction." Bioinformatics, 2010; 26(20): 2526-2533.
8. Clement NL et al., The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing, Bioinformatics, 2010; 26(1):38-45.
9. Peng Y, Leung HC, Yiu SM, Chin FY, Meta-IDBA: a de Novo assembler for metagenomic data, Bioinformatics, 2011;27(13):i94-i101.
10. Gritsenko AA, Nijkamp JF, Reinders MJ, de Ridder D, GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies, Bioinformatics, 2012;28(11):1429-1437.
11. Shumway M, Cochrane G, Sugawara H, Archiving next generation sequencing data, Nucleic acids research, 2010; 38:suppl 1; D870-D871.
12. Lee J, Ji Y, Liang S, Cai G, Müller P, On differential gene expression using RNA-Seq data. Cancer informatics, 2011; 10:205.
13. Trapnell C, Pachter L, Salzberg SL, TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009; 25(9):1105-1111.
14. Li H, et al., The sequence alignment/map format and SAMtools, Bioinformatics, 2009; 25(16): 2078-2079.
15. Quinlan AR, and Hall IM, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, 2010; 26(6):841-842.
16. Robinson MD, McCarthy DJ, Smyth GK, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010; 26(1):139-140.
17. Wang L, Feng Z, Wang X, Wang X, Zhang X, DEGseq: an R package for identifying differentially expressed genes from RNA-seq data, Bioinformatics, 2010; 26(1):136-138.
18. Karolchik D, et al, The UCSC genome browser database, Nucleic acids research, 2003;31(1):51-54.
19. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W, FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics, 2011; 27(14):1922-1928.
20. Kim D, Salzberg SL, TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol, 2011;12(8): R72
21. Francis RW, Thompson-Wicking K., Carter KW, Anderson D, Kees UR, Beesley AH, FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data, PloS one, 2012; 7(6), e39987.
22. McPherson A, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq

data, PLoS computational biology, 7(5), e1001138
23. Wang Z, Gerstein M, Snyder M, RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 2009; 10(1):57-63
24. http://bgiamericas.com/wp-content/uploads/2012/01/workflow2-1024x1000.png
25. Hackenberg M, Barturen G, Oliver JL, NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data, Nucleic acids research, 2011; 39(1): D75-D79.
26. Hackenberg, M, Rodríguez-Ezpeleta N, Aransay AM, miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments." Nucleic acids research, 2011; 39 (2): W132-W138.
27. Xia J, Wang Q, Jia P, Wang B, Pao W, Zhao Z, NGS Catalog: A database of next generation sequencing studies in humans." Human mutation, 2012; 33(6): E2341-E2355.
28. Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A, RNA- Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics, 2012;28(8):1184-1185.
29. Giardine B, et al., Galaxy: a platform for interactive large-scale genome analysis. Genome research 2005;15(10): 1451-1455.
30. Ofria C, Wilke CO, Avida: A software platform for research in computational evolutionary biology, Artificial life, 2004; 10(2):191-229..
31. Meaburn E, Schulz R, Next generation sequencing in epigenetics: insights and challenges. In Seminars in cell & developmental biology. 2012; 23(2) 192-199.
32. Delhomme N, Padioleau I, Furlong EE, Steinmetz LM, easyRNASeq: a bioconductor package for processing RNA-Seq data.Bioinformatics, 2012; 28(19): 2532-2533.
33. Friedman BA, Maniatis T, ExpressionPlot: a web-based framework for analysis of RNA-Seq and microarray gene expression data. Genome biology, 2011; 12(7): R69.
34. Cumbie JS, et al. GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences." PloS one , 2011; 6(10): e25279.
35. Guerreiro R, Brás J, Hardy J, Singleton A. Next generation sequencing techniques in neurological diseases: redefining clinical and molecular associations. Human molecular genetics, 2014. ddu203.
36. Zhang W, Hong C, and Lee-Jun CW. Chemical Diagnostics (Volume 336) "Application of next generation sequencing to molecular diagnosis of inherited diseases." . Springer Berlin Heidelberg, 2014. 19-45.