Chapter 6

Transcriptome Sequencing for Precise and Accurate Measurement of Transcripts and Accessibility of TCGA for Cancer Datasets and Analysis

Bijesh George, Vivekanand Ashokachandran,
Aswathy Mary Paul and Reshmi Girijadevi

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70026

Abstract

Next-generation sequencing (NGS) technologies are now well established and have become a routine analysis tool for its depth, coverage, and cost. RNA sequencing (RNA-Seq) has readily replaced the conventional array-based approaches and has become method of choice for qualitative and quantitative analysis of transcriptome, quantification of alternative spliced isoforms, identification of sequence variants, novel transcripts, and gene fusions, among many others. The current chapter discusses the multi-step transcriptome data analysis processes in detail, in the context of re-sequencing (where a reference genome is available). We have discussed the processes including quality control, read alignment, quantification of gene from read level, visualization of data at different levels, and the identification of differentially expressed genes and alternatively spliced transcripts. Considering the data that are freely available to the public, we also discuss The Cancer Genome Atlas (TCGA), as a resource of RNA-Seq data on cancer for selection and analysis in specific contexts of experimentation. This chapter provides insights into the applicability, data availability, tools, and statistics for a beginner to get familiar with RNA-Seq data analysis and TCGA.

Keywords: RNA-Seq, transcriptome data analysis, NGS data analysis, TCGA

1. Introduction

Genetic and epigenetic features encompassed in the genome are the basic determinants of fate and functions of cells. At the human interface, qualitative and/or quantitative differences in transcripts are the first level readout of these features in any specific context of their identification

[1]. These contexts may refer to a diseased state or the influence of stimulation such as intrinsic ligands or response to immunogens. With the total transcripts often referred to as transcriptome, the stage-specific or cell type-specific transcriptome of cells are valuable to evaluate the genetic and epigenetic features characteristic to them. From high- to low-input RNA, the RNA sequencing methods have considerably improved to appreciate the inter- and intra-level population heterogeneity of cells. Not restricted to messenger RNA (mRNA), these technologies are also being increasingly exploited to analyze other transcription-based products such as microRNAs and lncRNAs, reaching out to the identification of over 10–30 pg of a human cell or tissue [2]. RNA or transcripts are of two categories, protein coding mRNAs which synthesize protein and non-coding RNAs involved in regulating gene expression and in cell structure maintenance. mRNA makes up only 6% of the total RNA content of a cell or tissue; a number of methods and kits are available for RNA extraction from the cell [2, 3].

The human genome has more than 99.5% sequence identity to each other at the genomic level when analyzed in toto. However, they are also paradoxically personalized and are amenable to somatic variations. Hence, the cells could also be heterogeneous at genome level within an individual, and the genomic sequence variations are necessary to be accounted whenever they are analyzed at the transcriptome level. Toward this, the sequence obtained by RNA sequencing also reflects their coding sequence in the genome, kept aside, the RNA editing. Further, there are a plethora of other sequence determinants that could also be analyzed by sequence-based identification of transcripts. These determinants include the isoforms, gene fusions and identification of transcripts from putative pseudogenes. Unarguably, human cancer cells or tissues of diverse origins and stages in different populations are the most explored differential genome and transcriptome to date accounting the amount of data derived by RNA sequencing [4]. The Cancer Genome Atlas (TCGA) is probably the most extensive resource of providing access to cancer data especially from next-generation sequencing (NGS) platform. TCGA provides a number of options to perform analysis on cancer-related experimental data and stands as a major data repository for cancer data.

2. Transcriptomics

2.1. Gene expression

Gene expression at transcript level is a temporal dynamics event that involves turn “on” or “off” mechanism constituted by the coordinated action of epigenetic factors and transcriptional regulators. Since gene products are part of metabolic pathways in the organism, the inefficiency of protein synthesis control mechanism can lead to an abnormal behavior of metabolic pathways and then lead to diseases [5]. Determining or quantifying the amount of transcripts in a biological condition provides a clear picture about the involvement of that gene in a particular condition. It is necessary to use the quantitative methods to understand normal cell development, disease mechanisms and to determine when, where, and how much a gene is showing divergence with different biological condition [1]. Identification of key genetic factors/marker/a set of genes responsible for a certain biological process can make a sizable change to existing treatment mechanism approach [6].

2.2. Applicability of transcriptome data

Functions of each gene are not completely defined, information about the involvement of genes in functional pathways is identified and available from biological databases which provide clues on how each gene behaves in different metabolic pathways. Estimating the genes expressed in a particular biological condition allows comparing with the existing annotations. Only a small percentage of the genome is expressed in each cell, and a portion of the RNA synthesized in the cell is specific for that cell type [4], identifying the genes which are differentially expressed in similar tissue, but different context has therapeutic significance. Moreover, transcriptome sequencing allows identifying transcript level variations such as cassette exon, mutually exclusive exons, intron retentions, indels, alternative splice junctions, alternative promoters (**Figure 1**), and isoform-specific expression profiles [7].

2.3. Requirements

The number of biological/technical replicates, adequate sequencing depth, and essentially, the sequencing qualities are the major factors that should be accounted in a sequencing-based

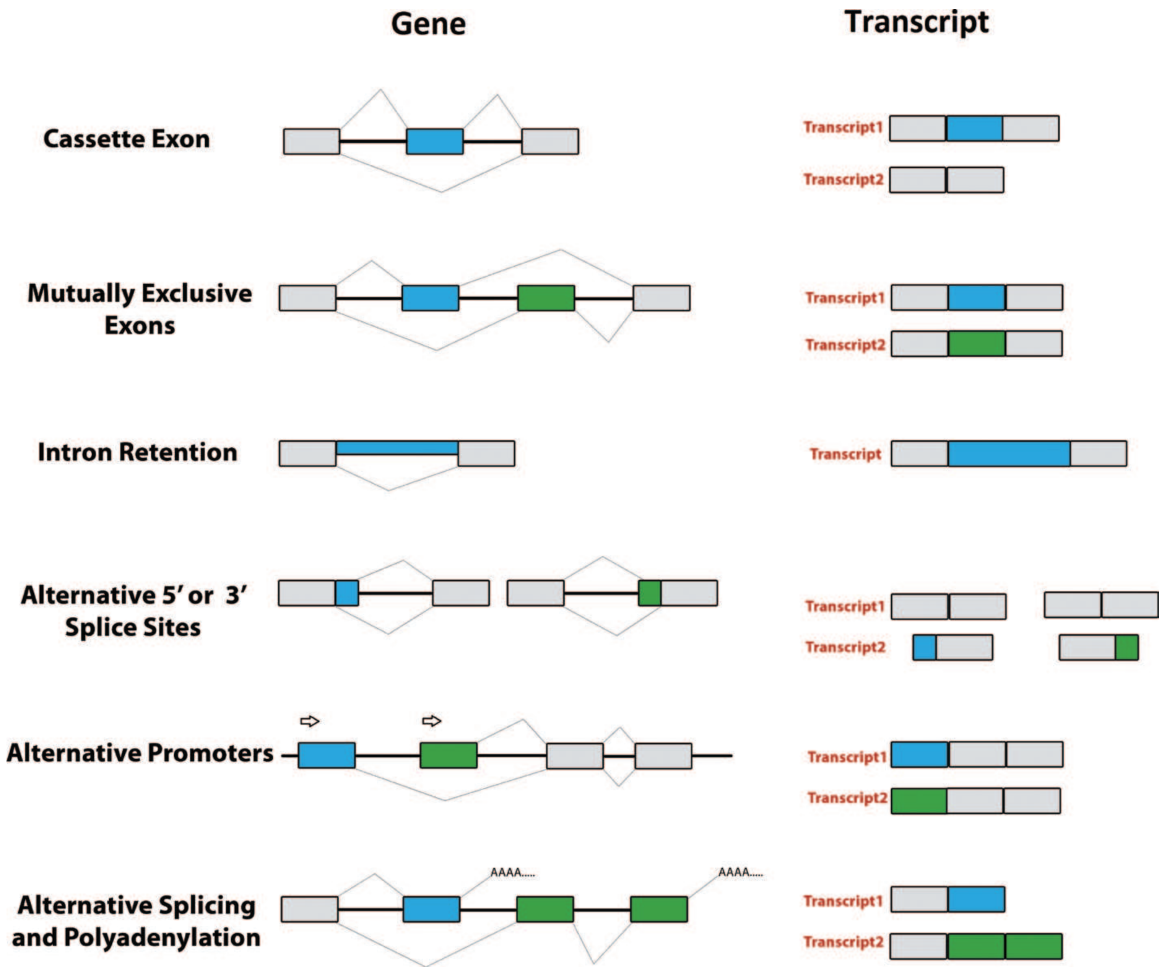

Figure 1. Alternative splicing. Here exons are boxes and lines are introns. Promoters represented by arrows and polyadenylation sites with AAA.

study. The parameters such as the availability of reference genome for the organism from which the sample is analyzed, information about the sequencer quality encoding, and whether multiplexing has been performed are also critical for the analysis. One should have a clear understanding of the biological sample, experimental conditions, and the biological questions that are in pursuit before starting a bioinformatics analysis of any transcriptome data [9].

Computational specifications have to be taken care to perform a genome assembly in a reasonable time without interruption. At least 8 core processor with 16 GB of RAM and enough fast storage system is required to perform a genome alignment within a reasonable time [7]. Genome assembly or alignment is the most computational resource consuming process, and the further downstream analysis such as variant calling or differential expression analysis can be performed using a desktop with an appreciable configuration.

Computational biologists prefer to use UNIX-based systems/servers for NextGen sequence analysis as large data can be handled more comfortably through command line by UNIX than a Windows OS [10].

2.4. Software requirements

A number of established and easily accessible one-shop sequence analysis tools [7, 11] are available online. However, it is important that one should understand the different steps involved in the analysis pipeline that are rather similar across them. There are various pieces of software in the pipeline, and each of them produces a number of output files. These include the main output file that can be used for further analysis and other supporting information such as the statistics of mapping, indicating the fraction of input data that had been successfully utilized by the algorithm (always get a higher fraction for good quality experiment) [7]. One should be aware of the files generated during each of the analysis steps that is fed into the next algorithm in the pipeline.

2.5. Precautions

A number of algorithms have been developed in recent years, and most of them are available as open-source algorithms. It is important to understand that the transcriptome analysis can be completed using open-source software and tools. Before starting the bioinformatics analysis on transcriptome data, one should decide the algorithms that can be used (**Figure 2**) including its release/version information in each successive step in the pipeline. Following the review articles that compare multiple algorithms and the research publications that have used specific algorithms, appropriate algorithms can be selected in each step [12]. Now, the next step is to select the annotation files to be used for the analysis.

Even though the information is same, data representation varies between annotation files from different biological data resources. An example given below represents human chromosome 22 from various biological data resources. Hence, one should confirm the annotation files such as genome file (.fasta) and gene transfer format (.gtf) files are compactible to each other.

Resource	Representation
NCBI reference genome GRCh38.p7	>gi|568801992|ref|NT_167212.2| chromosome 22 genomic scaffold, GRCh38.p7 primary assembly HSCHR22_CTG1_1
UCSC latest version GRCh38/hg38	>chr22
Ensembl	>22

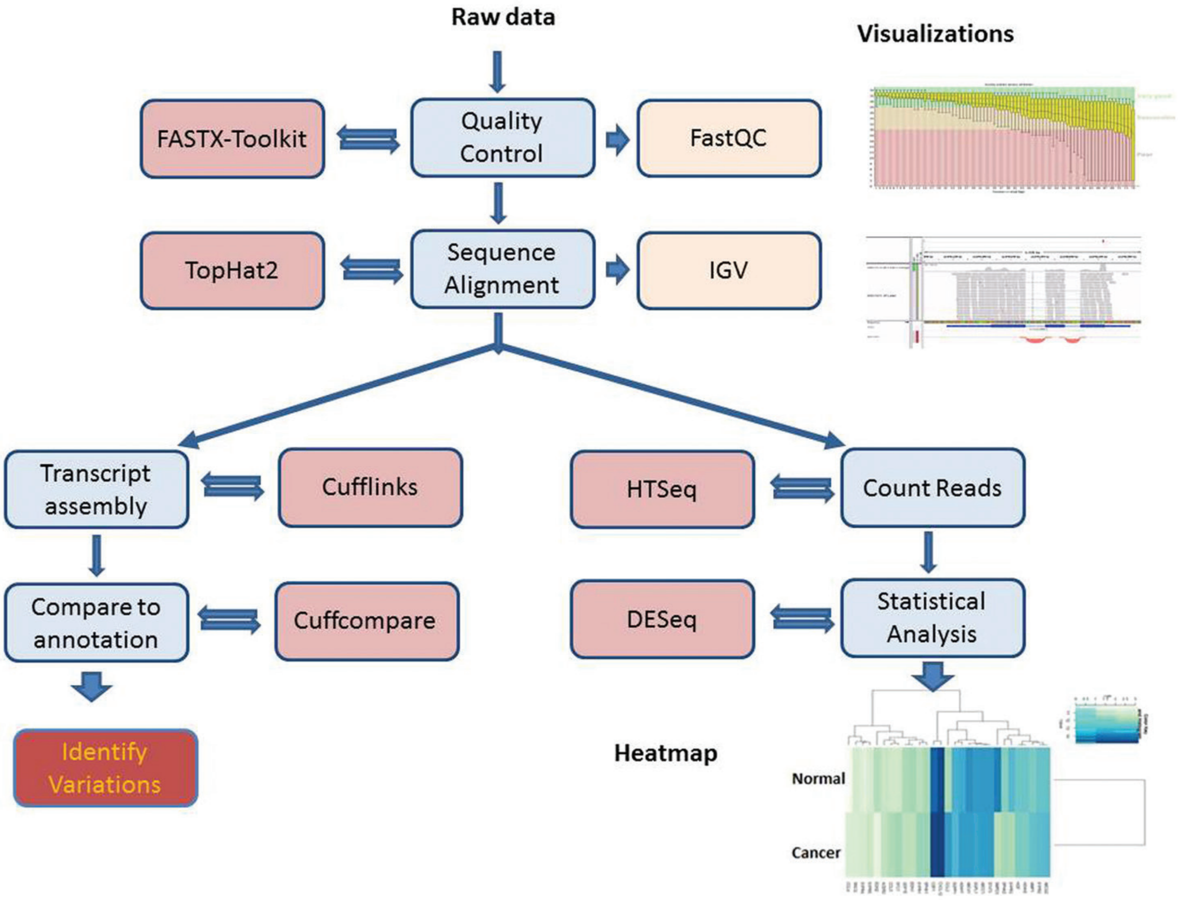

Figure 2. Transcriptomics workflow.

2.6. File formats

In each step of the analysis pipeline, multiple file formats are generated or used. It is necessary to know the information contained in each type of files. Here, we discuss file types classified into three categories. The first category is the raw files that contain the information adopted from the sequencer to represent the raw sequences with a quality score for each base-pair identification [13]. The file formats can be .sff, .csfasta + .qual, .fastq, etc. The most common file format is the .fastq extension. Second file category is the alignment files that represent the information on how each read or the fragment had been aligned to the reference genome [14], these files can be in .sam, .bam, and .bed formats. The third category is the annotated data files that represent data readily available from standard biological databases such as reference genome sequences (in .fasta format) and the annotated gene information (.gtf, .gff formats). Apart from all the standard file formats listed above, there are algorithm specific files which contain additional information about the specific run of the each algorithm in the pipeline.

3. Transcriptome data analysis

The high-throughput methods previously described (RNA-Seq) are done by direct sequencing of complementary DNA (cDNA) and as a result gives insights into the gene expression profiling [12, 15–17], quantification of alternative splicing [8, 9, 18, 19], variant calling [20–23], novel transcripts [14, 24, 25], and several others. These quantitative measurements are done by the final data produced by each sequencing platforms. However, the process of sequencing involves different steps (reverse transcription, amplification, fragmentation, purification, adaptor ligation, and sequencing that the chance of error in any step is highly likely and could result in faulty outputs. It makes the data in the worst case not suitable for further analysis, so that the experiment may have to be repeated. Nonetheless, these errors can be monitored and necessary actions can be undertaken to rectify the errors prior to analysis. Such preliminary steps are often referred to as quality control analysis of sequencing data.

3.1. Quality control

This section of the chapter will discuss various reasons and statistical assessment of errors such as sequence read quality, read duplication, GC bias, nucleotide composition bias, adapter contamination, flow cell contamination, enrichment, and false positive errors [26, 27], and how those can be tackled using available tools. The data used for the analysis in this chapter are mainly in the ".fastq" format, the most common format output of runs on many platforms. However, there are many quality control analysis tools available that either come aligned with the machine itself or as standalone software (commercial and open source). The quality control analysis can be done using many software tools, and one of the popular open-source software is FastQC [28].

Data output from sequencing machine includes the information about the sequence fragment as well as a score corresponding to each base identification, we are considering ".fastq" format, widely used in many platforms, to explain the features. A single read is represented by four consecutive lines in .fastq format. The first and third line represent sequence identifiers and other optional information, such as machine version, flow cell information, etc., related to the specific run of the sample in the machine. The second line is the sequence bases, and fourth is the quality value for each base which is represented as ASCII characters.

This ASCII quality value or phred quality score gives the accurate measure of the base calling quality during sequencing. Phred quality score is mathematically defined as

$$Q = -10 \times \log_{10}(P) \text{ or } P = 10^{-Q/10} \tag{1}$$

Where Q is the phred quality score, and P is the probability of getting a faulty base.

In essence, a phred score of 30 is the probability of a base to be wrong is 1 in 1000. However, there are no standard methods to measure this exact quality; the phred score above 20–25 (**Figure 3a** and **b**) is considered as the average score to be acceptable for further analysis because phred quality assessments are probabilistically stable [13, 29].

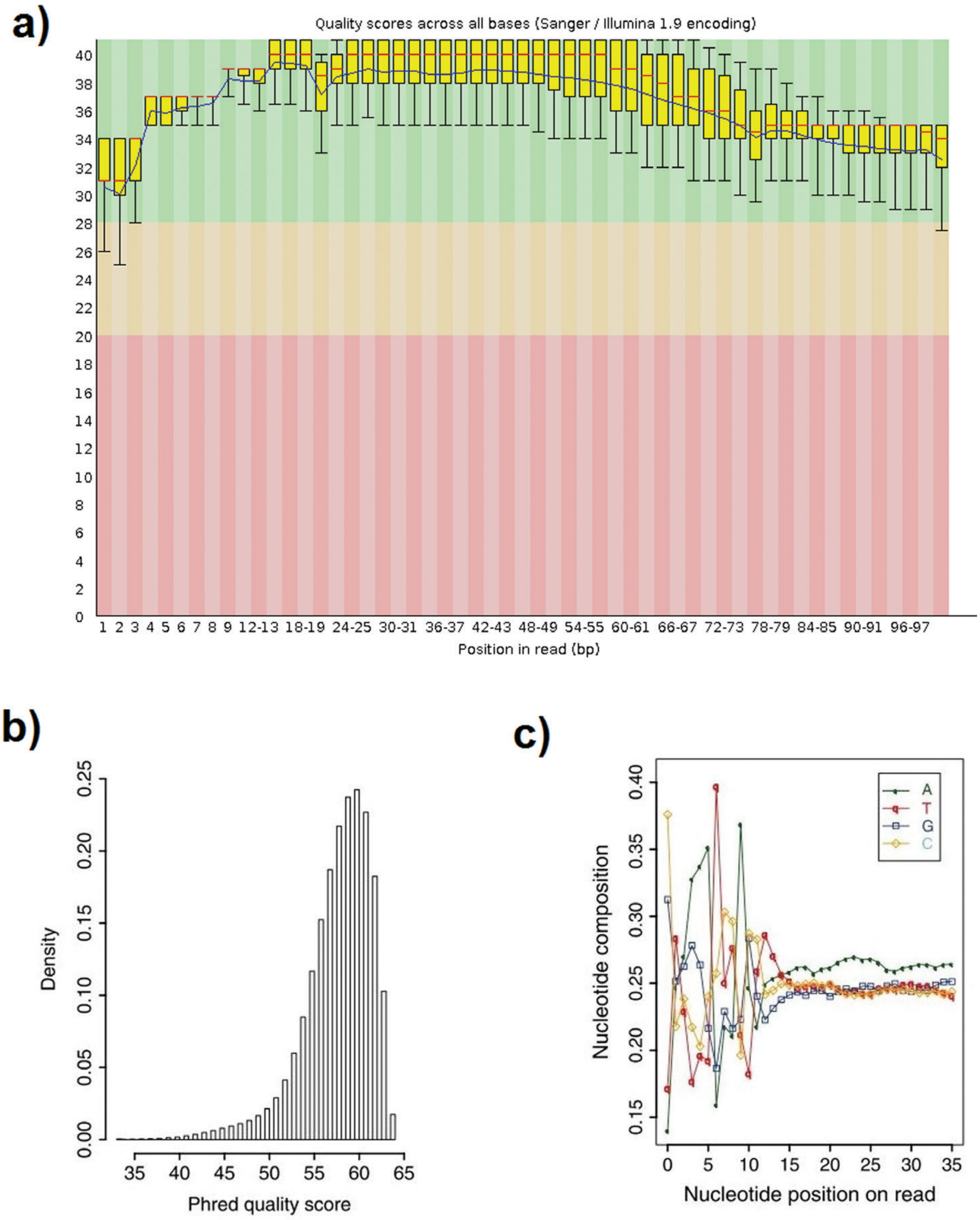

Figure 3. Quality control measures. (a) Per base sequence quality whisker plot: distribution of quality of bases all over the whole file, (b) distribution of percentage of sequences with different quality, and (c) distribution of bases in a .fastq file.

For each sequencer, they use different set of ASCII values to score each base calling and a maximum score of 41 which is almost 1 in 10,000 (99.99% accuracy) is the probability that a base is called incorrectly (**Table 1**). However, if the quality of any read falls to much lower scale, it is better to trim those regions off. There are many standard trimming tools available as open source. Few popular tools are FASTX-Toolkit [30], cutadapt [31], and trimgalore [32]. They cannot only be used for quality trimming but also has several other purposes, such as adapter trimming, demultiplexing, etc.

3.2. Evaluation of read quality

There are several statistical analysis pipelines available as open source to check the quality of the NGS data. This session explains the basic backgrounds of quality checks such as (1) base quality, (2) sequence content and distribution, and (3) duplicated sequences.

3.2.1. Base quality

As explained previously, base calling bias is strictly avoided because any error in base calling means the base is not correctly called. This analysis is done basically by the quality encoding values given to the reads in the file. This analysis is completely depending on the phred quality score throughout the base length. As an exception, the quality of reads will fall down toward the end of the reads, which is quite normal for long runs as the supplied base get reduced, and random calling of base leads to these false-positive errors.

Base quality analyzes are done for rectifying read errors could have happened during the run or library preparation. The data from the ".fastq" file can be plotted different ways based on the phred quality score of each bases, the proportion of reads being called wrong, N content distribution in the read, and finally, sequence length distribution. It is obvious that the sequence length would have uneven distribution in trimmed reads.

3.2.2. Sequence content and distribution

Evaluating GC content over the sequenced reads is as important as other modules because it leads to many biological reasoning. GC over AT is basically because of the stability of the

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Table 1. Phred quality score.

bonds between them, and the annealing process of PCR is based on the melting temperature of GC bonding. DNA methylation happens at cytosine, and comparatively, exons are high in GC content than introns.

In an NGS run, the bases are provided with an equal ratio, and the average of each base as output is expected to be 25% of each base (**Figure 3c**). Any fluctuation from this composition is considered as bias which is due to overrepresented sequences like adapter dimers or rRNA in the sample. However, it is expected that a little bias at the first few bases from 5′ which is essentially produced by the random hexamer priming from PCR amplification.

Before starting any analysis, adapters are trimmed off from the reads because the presence of adapters in the sample will lead to the expression of overrepresented sequences. This is more like a final check to be done to make sure the overrepresented sequences or enrichment identified is not spurious.

3.2.3. Duplicated sequences

As discussed in the GC content, there are few other ways to check the overrepresented sequences. These methods are used to confirm the sample is not contaminated, unless there is some kind of enrichment in the reads. The enrichment analysis is done basically on different scales. The length of the read is considered as the scale here. Creating K-mers of different length can make sure that how often an enrichment or overrepresented sequence can occur in the read, and this can be calculated to double check the presence of contamination or enrichment study.

3.3. Genome alignment

This is the second major step in transcriptomic data analysis. If the reference genome is available for the organism, it can be referred to as resequencing analysis else should be referred to as de novo sequencing analysis. In resequencing data, the analysis pipeline is comparatively easier compared to de novo sequencing. If reference genome is available, all we need is to map the fragments to the genome and find out the genes showing expression in the experiment. Although the amount of data generated from the sequencer is huge, it is short in length compared to the actual size of the genome. However, an advanced computationally efficient algorithm is required to perform this time consuming and banal process [5].

Genome alignment is the most important step in transcriptome analysis as all the downstream analysis, and the result accuracy is based on the efficiency of the alignment algorithm. As the data are obtained from transcriptome, the algorithm cannot directly map the reads to reference genome. An efficient splice aligner algorithm is required to complete the task [12], and most of these algorithms use a technique called hashing or indexing either in raw data or the genome data or both.

Read alignment algorithm has a number of parameters such as input and index as mandatory, and many other optional parameters also based on the computational resources available that can be set for the efficient mapping of reads. For example, we can set the number of multiple alignments for a single read and the maximum insertion or deletion length that can be allowed. A precise understanding of experimental conditions helps to set appropriate parameters according to a specific experiment. Moreover, default values provided to help and avoid confusions [7].

3.4. Gene quantification

Gene quantification is performed after alignment to a genome. The first step is to identify the amount of fragments or reads that could be mapped to each genomic location. Gene level or transcript level quantification can be performed according to user's choice. A number of software tools (coverageBED [33], htseq-count [34], and featureCounts [35]) are available for gene quantification. Quantification is performed against a reference annotation (GTF/GFF) file with coordinates for the gene, transcript, or exon. For example, htseq-count uses "--idattr=<id attribute>" that indicates GFF attribute to be used as feature ID from the ninth column where unique ids or accession numbers are available. Gene qualification has to be performed after normalization to avoid misleading measurements. Hence, gene level or sample level normalization of the data in terms of total number of reads mapped, read length, and coverage should be performed.

The reads per kilobase of exon model per million mapped reads (RPKM) measure normalizes with the sequencing depth that varies significantly between samples as well as the gene length. Fragments per kilobase of exon model per million mapped reads (FPKM) measure normalizes similar to RPKM but for the paired-end data and the transcripts per million (TPM) first normalizes by gene length, then by sequencing depth, preferably a better way of normalization [36].

3.5. Splice variation analysis

Transcriptome analysis can identify transcript sequence level features such as cassette exon, mutually exclusive exons, intron retentions, indels, alternative splice junctions, and hence, different possible isoforms all based on genome mapping (**Figure 1**). There are ~41,000 unique transcripts that are identified from a total of ~20,000 genes in human (NCBI RefSeq) [37].

Identification of transcripts from short and specific number of reads aligned across the gene, and the identification of splice junctions is a challenge in variation analysis. A number of algorithms such as Cufflinks [38], SLIDE [39], and StringTie [40] are available to analyze the alignment with user-provided existing annotations. Cufflinks [38] efficiently utilizes the advantage of paired-end sequencing data to annotate the splice variations (**Figure 4**).

3.6. Differential expression analysis

Once the genome assembly is completed, the downstream analysis can follow two routes—the variation analysis and the differential expression analysis. Differential expression analysis refers

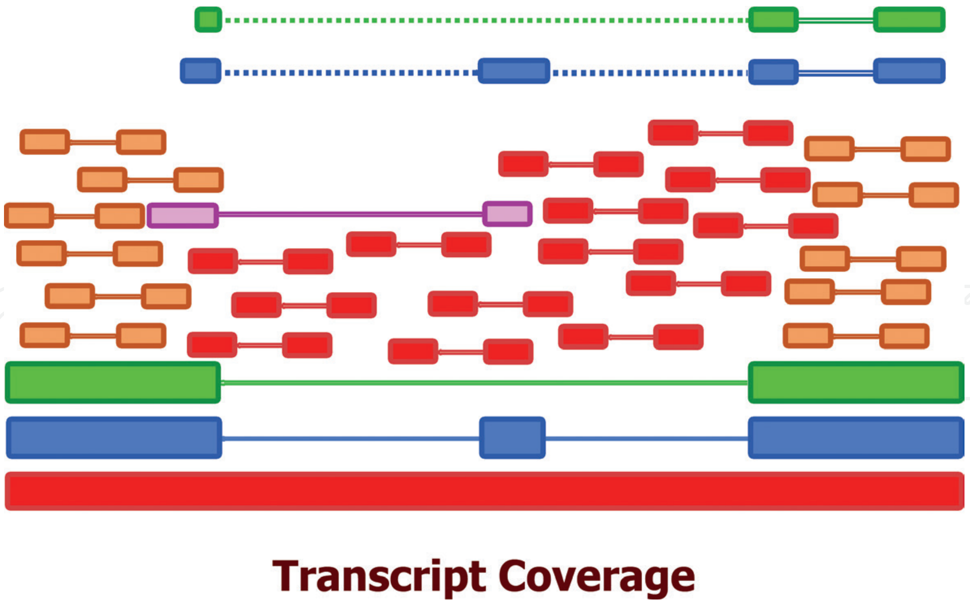

Figure 4. Transcript enrichment. Cufflinks identify three transcripts from reads mapped to the same genomic region.

the gene level expression difference between two or more samples. This can be performed using R packages like edgeR [9], DESeq [10] that can load gene quantification information from multiple samples and report the expression level difference for each transcript/gene. The above-mentioned R packages also can generate multiple figures such as heatmaps, histograms, dispersion plots, etc., which can be used for representing results as well as publications purposes. The comparison is performed after normalization of the data across samples that account the length of the fragments, sequencing depth, and the total number of reads mapped. RPKM, FPKM, and TPM are commonly used normalization values. Genes with at least 2-fold change are usually considered as differentially expressed, although a fold change of 1.5 is also considered in certain instances.

Types of graphical methods are available to visually represent the identified variations among experiments or samples used. Overview of gene expression studies can be represented by volcano plot, MA plot, heatmap, etc. Heatmap with hierarchical clustering clearly represents the trend of gene expression between samples.

Visualization is integral to NGS data from the evaluation of sequencing quality to the representation of the biologically significant results. Initially, the raw data have to undergo quality checking to assess the overall sequencing quality and decide quality measures (FastQC (**Figure 3a**) [28], NGSQC [41]). The next level of visualization is applicable to the alignment to the genome as perceived for the number of reads aligned to particular gene, exons, introns, and splice junctions with genome browsers such as UCSC browser [42], Integrative Genomics Viewer (IGV) [43], and Genome Maps [44]. Genome browsers load genome (.fasta), annotations (.gff, .gtf), variations (as bed files) to their interface to obtain clear visualization of collective data for a specified region along with the available annotation, identified evidence or mapped reads, and variations observed. They also host inbuilt tools to represent the data as plots and figures that can be used for publication [43].

4. TCGA: a genomic hub of cancer

The Cancer Genome Atlas well known as TCGA in short is a combined effort of National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) investing $50 million each to increase the better understanding of molecular basis of cancer using advanced genome analysis technology. The overall aim of launching such a big project was to improve the ability to diagnose, treat, and prevent cancer. The first phase of the study started in the year 2005 focused on the brain, lung, and ovarian cancers was aimed to test and develop the infrastructure for further research. The second phase of the study comprises of around 30 different type of cancers started in the year 2009 and analyzed by the year 2014.

The first phase of the study proved that an atlas specific for cancer can be created with a worldwide network of research and teams working on different cancer and develop a single platform for making the data publically accessible pooling all the data. The publicly available data from TCGA would also enable researchers around the world to make validate important discoveries. TCGA is supported by Genomic Data Commons (GDC) as one among the several programs at the NCI's Center for Cancer Genomics along with another program Therapeutically Applicable Research to Generate Effective Treatments (TARGET). Now, GDCs host genomic alterations of exactly 39 projects combining the TCGA and TARGET.

Data availability has categorized based on primary site of study, and they are kidney, adrenal gland, brain, colorectal, lung, uterus, bile duct, bladder, bone marrow, breast, cervix, esophagus, eye, head and neck, liver, lymph nodes, ovary, pancreas, pleura, prostate, skin, soft tissue, stomach, testis, thymus, and thyroid. Some of the primary sites are again divided into different subdivions. For example, kidney again divided into three different projects: kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma, and kidney chromophobe. So as the case with adrenal gland, brain, colorectal, lung, and uterus which all are divided again into two different sub categories as follows: pheochromocytoma & paraganglioma, adrenocortical carcinoma, glioblastoma multiforme, brain lower grade glioma, colon adenocarcinoma, rectum adenocarcinoma, lung adenocarcinoma, lung squamous cell carcinoma, uterine corpus endometrial carcinoma, uterine carcinosarcoma.

4.1. TCGA data and file formats

The main category of data available in TCGA are:

- Clinical
- Raw sequencing data
- Transcriptome profiling
- Simple nucleotide variation
- Biospecimen

- Copy number variation
- DNA methylation

Main categories of data type are:

- Aligned reads
- Gene expression quantification
- Annotated somatic mutation
- Raw simple somatic mutation
- Copy number segment
- Masked copy number segment
- Methylation beta value
- Isoform expression quantification
- miRNA expression quantification
- Biospecimen supplement
- Clinical supplement
- Aggregated somatic mutation
- Masked somatic mutation

These data that are generated from different experimental strategies such as WXS, RNA-Seq, and miRNA-Seq were studied under illumina platform, whereas Illumina Human Methylation 450 and Illumina Human Methylation 27 platforms were used for methylation array and genotyping array was carried out using Affymetrix SNP 6.0.

4.2. miRNA analysis

TCGA provides tissue-specific miRNA expression profiles, their isoforms, connection with diseases, and the discovery of unreported miRNAs. Alignment of the reads with BWA-aln is the very first step in the miRNA pipeline. Either the input can be FASTQ or BAM file format for alignment. The output after the alignment will be of BAM format. The alignment follows the expression workflow. The output from the expression workflow is raw read counts and normalized to reads per million mapped reads. There are two types of files, controlled and open. The aligned file which is having a controlled access, and the quantification files are open accessible (**Table 2**). The RPM comes in two separate files as "mirnas.quantification.txt" and "isoforms.quantification.txt." The mirna.quantification.txt data file describes the summed expression for each miRNA. The file contains the information:

- miRNA name
- raw read count

- reads per million miRNA reads
- cross-mapped to other miRNA forms (Y or N)

whereas the isoform.quantification.txt file contains every individual sequence isoform observed as follows:

- miRNA name
- alignment coordinates as <version>:<Chromosome>:<Start position>-<End position>:<Strand>
- raw read count
- reads per million miRNA reads
- cross-mapped to other miRNA forms (Y or N)
- region within miRNA

4.3. RNA-Seq analysis

TCGA uses an Illumina system as the basic platform. Information for nucleotide sequence and gene expression is found at TCGA. RNA sequence coverage, sequence variants (e.g., fusion genes), expression of genes, exon, or junction are different category of information available after the sequence alignment. The NCBI dbGaP database is the official repository for the actual sequence data [45]. After aligning the reads to reference genome, gene expression level is quantified in various forms such as HT-Seq raw mapping count, fragments per kilobase of transcript per million mapped reads (FPKM) and FPKM-UQ (upper quartile normalization) in TCGA mRNA quantification pipeline (**Table 3**). In case of mRNA analysis also the rules for data access are the same. Access for aligned reads file is controlled, whereas access for rest of the files is open.

4.4. DNA-Seq analysis

Genomic diversity across different cancer types has been characterized by utilizing DNA sequencing systems based on Sanger Sequencing at different Genome Sequencing Centers.

Type	Description	Format
Aligned reads	miRNA-Seq reads that have been aligned to the GRCh38 build. Reads that were not aligned are included to facilitate the availability of raw read sets	BAM
miRNA expression quantification	A table that associates miRNA IDs with read count and a normalized count in reads per million miRNA mapped	TXT
Isoform expression quantification	A table with the same information as the miRNA Expression Quantification files with the addition of isoform information such as the coordinates of the isoform and the type of region it constitutes within the full miRNA transcript	TXT

Table 2. Data types and file formats.

Type	Description	Format
RNA-Seq alignment	RNA-Seq reads that have been aligned to the GRCh38 build. Reads that were not aligned are included to facilitate the availability of raw read sets	BAM
Raw read counts	The number of reads aligned to each protein-coding gene, calculated by HT-Seq	TXT
FPKM	A normalized expression value that takes into account each protein-coding gene length and the number of reads mappable to all protein-coding genes	TXT
FPKM-UQ	A normalized raw read count in which gene expression values, in FPKM, are divided by the 75th percentile value	TXT

Table 3. Gene quantification data formats.

Somatic variants from whole-genome sequencing are identified using this pipeline. Somatic variants are identified by comparing the tumor samples with the normal samples allele frequency. After annotating each mutation, one project is created combining files from multiple cases. Identification of somatic mutation has achieved through four pipelines. Identified somatic variants are then annotated. Information from multiple files is combined into one single MAF for each pipeline. Mutations are listed in a tab delimited format as Mutation Annotation Format (MAF). Two types of MAF files are produced for each variant calling in a project, i.e., the protected and the somatic or public MAF files. These MAF files are produced on the basis of annotated Variant Call Format (VCF) file. This VCF file contains variants reported in multiple transcripts. Only the critical ones are reported in the protected MAF file, whereas Public MAF are processed to remove the low quality and potential germline variants restricting the confidential information.VCF files are of two type, raw unannotated simple somatic mutations and annotated somatic mutation VCF files.

4.5. Single-nucleotide polymorphism

TCGA utilized SNP-based technology to analyze genome-wide variations. It also includes platforms to define CNV and loss of LOH across multiple samples.

4.6. DNA methylation sequencing

TCGA utilizes the Illumina platform for the DNA methylation study ensures single-base-pair resolution, high accuracy, easy workflows, and low input of DNA requirements. DNA methylation data files (**Table 4**) contain information of signal intensities (raw and normalized), detection confidence, and calculated beta values for methylated (M) and unmethylated (U) probes.

4.7. Reverse-phase protein array (RPPA)

Is a high throughput, functional, and quantitative proteomic method for large-scale protein expression profiling which helps in biomarker discovery and cancer diagnostics eventually.

Protein arrays consist of data representing protein expression and concentration. These data archives are deposited to the TCGA DCC and include original images of protein arrays, calculated raw signals, relative concentrations of proteins and normalized protein signals (**Table 5**).

4.8. Data processing workflow

TCGA have a well-organized structure from sample collection to bioinformatics analysis with involvement of several centers (**Table 6**).

Platform code	File type	Description
IlluminaDNAMethylation_OMA002_CPI	Tab-delimited, ASCII text (.txt)	Cy3 and Cy5 signals and detection confidence of methylated probes
IlluminaDNAMethylation_OMA002_CPI	Tab-delimited, ASCII text (.txt)	Calculated beta values
IlluminaDNAMethylation_OMA003_CPI	Tab-delimited, ASCII text (.txt)	Cy3 and Cy5 signals and detection confidence of methylated probes
IlluminaDNAMethylation_OMA003_CPI	Tab-delimited, ASCII text (.txt)	Calculated beta values
HumanMethylation27	Binary (.idat)	Intensity data file with statistics for each bead type in terms of bead count, mean and standard deviation per dye
HumanMethylation27	Tab-delimited, ASCII text (.txt)	Calculated beta values and mean signal intensities for replicate methylated and unmethylated probes
HumanMethylation27	Tab-delimited, ASCII text (.txt)	Calculated beta values, gene symbols, chromosomes and genomic coordinates (build 36). Some data have been masked (including known SNPs)
HumanMethylation450	Binary (.idat)	Intensity data file with statistics for each bead type in terms of bead count, mean and standard deviation per dye
HumanMethylation450	Tab-delimited, ASCII text (.txt)	Background-corrected methylated (M) and unmethylated (U) summary intensities as extracted by the methylumi package
HumanMethylation450	Tab-delimited, ASCII text (.txt)	Calculated beta values, gene symbols, chromosomes and genomic coordinates (hg18). Some data have been masked (including known SNPs)

Table 4. DNA methylation data files format.

File type	Description
Array Slide Image (tiff)	Black and white, high-resolution image of protein array
RPPA Slide Image Measurements (txt)	Raw signals from a black and white, high-resolution image of protein array
Super Curve Results (tab-delimited, txt)	Supercurve results, use dilution to calculate relative concentration
Normalized Protein Expression (MAGE-TAB data matrix, txt)	Signals for genes

Table 5. Protein data file format.

Project	Details	Source
Tissue Source Sites (TSSs)	Collection of the samples (blood and tissue from tumour and normal controls) and clinical metadata from patients (donors) Shipment of the annotated biospecimens to Biospecimen Core Resources (BCR) https://wiki.nci.nih.gov/display/TCGA/Tissue+Source+Site	https://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=tissue%20source%20site
Biospecimen Core Resource (BCR)	Coordination of sample delivery and data collection, cataloguing, processing, and verifying the quality and quantity Isolation and distribution of RNA and DNA from biospecimens to other institutions for genomic characterization and high-throughput sequencing http://cancergenome.nih.gov/abouttcga/overview/howitworks/bcr http://www.nationwidechildrens.org/biospecimen-core-resource-about-us	Research Institute at Nationwide Children's Hospital in Columbus, Ohio
Genome Sequencing Centers (GSCS)	High-throughput sequencing (data are available in TCGA Data Portal or at NIH's database of Genotype and Phenotype) Identification of the DNA alterations http://cancergenome.nih.gov/abouttcga/overview/howitworks/sequencingcenters	Broad Institute Sequencing Platform in Cambridge Human Genome Sequencing Center, Baylor College of Medicine in Houston The Genome Institute at Washington University
Cancer Genome Characterization Centers (GCCs)	Utilization of novel technologies and multiple platforms Comprehensive description of the genomic changes: alterations in miRNA and gene expression, SNP, CNV, and others http://cancergenome.nih.gov/abouttcga/overview/howitworks/characterizationcenters	Copy Number Alteration (Brigham and Women's Hospital and Harvard Medical School in Boston, The Broad Institute in Cambridge) Epigenomics (University of Southern California in Los Angeles, Johns Hopkins University in Baltimore) Gene (mRNA) Expression (University of North California at Chapel Hill) miRNA Analysis (British Columbia Cancer Agency in Vancouver) Targeted Sequencing Center (Baylor College of Medicine in Houston) Functional Proteomics (MD Anderson Cancer Center)
Proteome Characterization Centers (PCCs)	Identification of cancer-specific proteins http://cancergenome.nih.gov/abouttcga/overview/howitworks/proteomecharacterization	Cancer Proteomic Center Center for Application of Advanced Clinical Proteomic Technologies for Cancer Proteo-Genomic Discovery Prioritization and Verification of Cancer Biomarkers Proteome Characterization Centre and Vanderbilt Proteome Characterization Center
Data Coordinating Center (DCC)	Management of all generated data and transfer them to public databases (TCGA Data Portal and Cancer Genomics Hub) http://cancergenome.nih.gov/abouttcga/overview/howitworks/datasharingmanagement	University of California Santa Cruz

Table 6. TCGA centers and data processing.

Eligible patient samples (blood and tissue) are collected by different Tissue Source Sites (TSSs) and delivered to the Biospecimen Core Resource (BCR). BCR catalogue, process, and verify the quality and quantity of these samples and then submit clinical data and metadata to the Data Coordinating Center (DCC). Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) then do the genomic characterization and high-throughput sequencing once the DCC provide molecular analytes. After sequencing, DCC again receives the sequence-related data from GSS. Trace files, sequences, and alignment mappings from Genome Characterization Centers are also submitted to the NCI's secure repository Cancer Genomic Hub (CGHub). Access to research community for these data is made available along with Genome Data Analysis Centers (GDACs). Information managed by DCC that has stored into public free-access databases (TCGA portal, NCBI's Trace Archive, CGHub), allows researchers to access the data and hence helps to advance in cancer studies.

4.9. TCGA data identifiers

Barcodes were initially used as the primary identifier for biospecimen data in TCGA during the beginning of the data. Tissue source site delivers the patient sample and the metadata to Biospecimen Core Resource (BCR). Once the sample is received by BCR, a human readable TCGA barcode was assigned. TCGA barcode was assigned to keep the navigation of the various results produced by the different data-generating centers for one particular sample connected. Sections of barcode also provide metadata information about the sample. Nowadays, BCR is also assigning universally unique identifiers (UUIDs) along with TCGA barcode to samples keeping UUIDs as the primary identifier instead of barcodes.

4.9.1. Barcodes

BCR generates the barcode for each sample received from TSS. Barcode initial numbers after the program code are assigned according to the TSS and the participant from which the tissue sample was received. The barcodes TCGA-02 and TCGA-02-0001 are assigned, respectively. Types of tissue are also differentiated with codes (**Table 7**). Next number in the barcode stands for the sample followed by the vial number; the sample was split into TCGA-02-0001-01 and TCGA-02-0001-01B. This vial number is again divided into different portions—TCGA-02-0001-01B-02. Analytes represented with barcode, e.g., TCGA-02-0001-01B-02D was extracted and distributed across one or more than one plates TCGA-02-0001-01B-02D-0182. Each well represented as, e.g., TCGA-02-0001-01B-02D-0182-06 is identified as an aliquot. These plates are later given to various characterize and sequencing centers.

4.9.2. Universally unique identifier (UUID)

UUIDs are randomly generated 32-digit hexadecimal value. TCGA became more complex, and the barcode was not enough to handle the generated data because there was not enough barcode combinations to represent the data. Also, flexibility in altering the barcode was also less when the associated metadata changes with a barcode. Considering all these factors, TCGA changed from using barcode for biospecimen and clinical data.

Tissue code	Letter code	Definition
1	TP	Primary Solid Tumor
2	TR	Recurrent Solid Tumor
3	TB	Primary Blood Derived Cancer—Peripheral Blood
4	TRBM	Recurrent Blood Derived Cancer—Bone Marrow
5	TAP	Additional—New Primary
6	TM	Metastatic
7	TAM	Additional Metastatic
8	THOC	Human Tumor Original Cells
9	TBM	Primary Blood Derived Cancer—Bone Marrow
10	NB	Blood Derived Normal
11	NT	Solid Tissue Normal
12	NBC	Buccal Cell Normal
13	NEBV	EBV Immortalized Normal
14	NBM	Bone Marrow Normal
20	CELLC	Control Analyte
40	TRB	Recurrent Blood Derived Cancer—Peripheral Blood
50	CELL	Cell Lines
60	XP	Primary Xenograft Tissue
61	XCL	Cell Line Derived Xenograft Tissue

Table 7. Tissue code.

The generated data are not only categorized based on the type but also the level at which these data can be accessed. In addition to the analyzed tumor data, TCGA also collects non-tumor samples aimed to analyze every patients germ line DNA to identify which alteration found in tumor sample responsible for the oncogenic process. For most of the tumors, TCGA collects and analyzes normal blood samples. In the absence of a matching normal blood sample, a normal tissue sample from the same patient is used as the germ line control for DNA assays. But in the case of RNA assays, using a normal blood sample as a control is not logically correct. Because RNA profile of blood sample is expected to be different from the RNA profile of tissues from other organs such as brain, breast, and lungs or ovary. Because of this reason, TCGA attempts to collect normal tissue matched to the anatomic site of the tumor not matched to the patient.

4.10. Accessibility of data

Access to the data is strictly controlled. There are two levels of data access:

- Open access data tier [raw, non-normalized data (Level I), processed data (Level II)].
- Controlled access data tier [segmented/interpreted data (Level III) apply to individual samples, while summarized data (Level IV)].

4.10.1. Open access data tier

The open access data level is composed of public data not unique to a patient. The open access data tier does not require any user certification [45].

Type of data accessible at open tier:

- Biospecimen
- Transcriptomic profiling
- Copy number variations
- DNA methylation
- Clinical
- Single-nucleotide variation

4.10.2. Controlled access data tier

Patient's unique information falls into the controlled access tier. Each data type has unique identifiers. In order to get the access to the data, user needs the certification.

Type of data accessible at controlled level:

- BAM and FASTQ files
- Level 1 and level 2 SNP6 array data
- Level 1 and level 2 exon array data
- Variant Call Format files
- Peculiar data of MAFs

In order to attain the access to these data, the researchers must:

- Complete the Data Access Request (DAR) form which is available electronically through the Database of Genotypes and Phenotypes (dbGaP).

Once the submitted request is evaluated and approved, researchers must

- Agree to restrict their use of the information to biomedical research purposes only
- Agree with the statements within TCGA Data Use Certification (DUC)
- Have their institutions certifiably agree to the statements within TCGA DUC

All patient samples are sworn to use for TCGA and there is no provision of sharing the material with a third party. Even this is not the case because 95% of material used up in different characterization. Even there is chance left to get the samples from the TSS centers. One can directly contact the TSS center for samples, and the decision lays on them.

4.11. TCGA data: visualization and data analysis

A huge amount of data accumulation demanding for advanced visualization technology and number of tools are available (**Table 8**). Visualization is essential to understand the data at ease.

Tool	Application
The Cancer Imaging Archive, CIA (http://www.cancerimagingarchive.net)	TCIA hosts a large archive of medical images of cancer accessible for public download. Information regarding patients treatment details, outcomes, pathology and genomics are also provided as supporting information based on availability
Berkeley Morphometric Data (http://tcga.lbl.gov:9999/biosig/tcgadownload.do)	Characterize tumour histopathology, through the delineation of the nuclear regions, from hematoxylin and eosin (H&E) stained tissue sections. The advantages of such a database is that other samples can be cross-referenced for personalized therapy and precision medicine as it contains information regarding responses to therapies, molecular correlates and morphometric subtypes
The Cancer Digital Slide Archive, CDSA (http://cancer.digitalslidearchive.net/)	Is an integrated Web-based platform supporting whole-slide pathology image visualization and data integration of the TCGA data
The Broad GDAC Firehose (http://firebrowse.org/)	Is a powerful tool for exploring cancer data. FireBrowse helps researchers to easily find any of thousands of data archives generated by the same. A powerful RESTful API is provided, with bindings to the UNIX command line, Python and R for programmers. For easy access, graphical interface like viewGene to explore expression levels and iCoMut are provided to explore the mutation information of each TCGA disease study with an interactive figure
The MD Anderson GDAC's MBatch (http://bioinformatics.mdanderson.org/tcgabatcheffects)	Is designed to help researchers to assess, diagnose and correct for any batch effects in TCGA data. It first allows the user to assess and quantify the presence of any batch effects through Principal Component Analysis and Hierarchical Clustering algorithms. The results from these algorithms are presented graphically as diagrams
Cancer Genome Workbench, CGWB (https://cgwb.nci.nih.gov/)	NCI developed application which integrate and display genomic and transcription alterations across various cancers. Integrated tracks view, Heatmap view, Bambino are the major viewers
UCSC Cancer Genomics Browser (https://genome-cancer.soe.ucsc.edu/)	Is an open access suite integrate, visualize and cancer genomic data along with clinical data
Integrative Genomics Viewer, IGV (http://www.broadinstitute.org/igv)	Is a freely available visualization tool of the genome developed by Broad Institute
The cBioPortal for Cancer Genomics (http://cbioportal.org)	Is interactive open-access resource for the exploration of multidimensional cancer genomics data sets. The barriers between the genomic data and the researchers are reduced rapidly after the resources was established. This database stores DNA copy-number data (deep deletions or amplification), non-synonymous mutations, mRNA and microRNA expression data, protein level, phosphoprotein level (RPPA) data, limited de-identified clinical data and DNA methylation data
Regulome Explorer (http://explorer.cancerregulome.org/)	It explores the association between and molecular features of TCGA data. According to user-specified parameters the data can be filtered for the search and visualize

Table 8. Visualization and data analysis tools.

Author details

Bijesh George, Vivekanand Ashokachandran, Aswathy Mary Paul and Reshmi Girijadevi*

*Address all correspondence to: reshmisuresh@gmail.com

Cancer Research Program-9, Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, Kerala, India

References

[1] Institute NHGR. Transcriptome The National Human Genome Research Institute: The National Human Genome Research Institute. 2015 [updated August 27, 2015; cited 2015 August 27, 2015]. Available from: https://www.genome.gov/13014330/transcriptome-fact-sheet/

[2] Tuffaha MSA. Phenotypic and Genotypic Diagnosis of Malignancies: An Immunohistochemical and Molecular Approach. 1st ed. Weinheim: Wiley-Blackwell, 2008. DOI: 10.1002/9783527621521

[3] Ramalho AS, Beck S, Farinha CM, Clarke LA, Heda GD, Steiner B, et al. Methods for RNA extraction, cDNA preparation and analysis of CFTR transcripts. Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society. 2004;**3**(Suppl 2):11-15

[4] Gilbert SF. Differential Gene Expression. Developmental Biology. 6th ed.; Sunderland (MA): Sinauer Associates; 2000

[5] Hoopes L. Introduction to the gene expression and regulation topic room. Nature Education. 2008;**1**(1):160

[6] Alberts B, Johnson A, Lewis J, et al. Studying Gene Expression and Function. Molecular Biology of the Cell. 4th ed.; New York: Garland Science; 2002

[7] Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;**25**(9):1105-1111

[8] Zahler, A. M. Pre-mRNA splicing and its regulation in Caenorhabditis elegans The C. elegans Research Community WormBook. WormBook, 2012. 1551-8507. DOI:10.1895/wormbook.1.31.2

[9] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;**456**(7221):470-476

[10] Christie M. L. Linux: The glue that binds your next-generation sequencing analyses. The Molecular Ecologist. [Internet]. 2012. Available from: http://www.molecularecologist.com/2012/10/linux-the-glue-that-binds-your-next-generation-sequencing-analyses/ [Accessed:2017-07-18]

[11] Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research. 2016;**44**(W1):W3-W10

[12] Lindner R, Friedel CC. A comprehensive evaluation of alignment algorithms in the context of RNA-seq. PLoS One. 2012;**7**(12):e52403

[13] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research. 1998;**8**(3):186-194

[14] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotechnology. 2010;**28**(5):503-510

[15] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research. 2008;**18**(9):1509-1517

[16] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;**10**(1):57-63

[17] Wilhelm BT, Landry JR. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods. 2009;**48**(3):249-257

[18] Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods. 2010;**7**(12):1009-1015

[19] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010;**28**(5):511-515

[20] Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Research. 2012;**22**(1):142-150

[21] Park E, Williams B, Wold BJ, Mortazavi A. RNA editing in the human ENCODE RNA-seq data. Genome Research. 2012;**22**(9):1626-1633

[22] Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nature Biotechnology. 2012;**30**(3):253-260

[23] Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, et al. Identifying RNA editing sites using RNA sequencing data alone. Nature Methods. 2013;**10**(2):128-132

[24] Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & Development. 2011;**25**(18):1915-1927

[25] Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature Biotechnology. 2011;**29**(8):742-749

[26] Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research. 2012;**40**(10):e72

[27] Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research. 2010;**38**(12):e131

[28] Simon A. FastQC: A quality control tool for high throughput sequence data. [Internet]. 2016. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [18 Jul. 2017]

[29] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research. 1998;**8**(3):175-185

[30] FASTX-Toolkit

[31] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, [S.l.], may. 2011;**17**(1):10-12. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200. Date accessed: 18 Jul. 2017. doi:http://dx.doi.org/10.14806/ej.17.1.200

[32] Trimgalore

[33] Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;**26**(6):841-842

[34] Anders S, Pyl PT, Huber W. HTSeq—A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;**31**(2):166-169

[35] Liao Y, Smyth GK, Shi W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;**30**(7):923-930

[36] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;**17**:13

[37] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016;**44**(D1):D733-D745

[38] Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;**27**(17):2325-2329

[39] Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proceedings of the National Academy of Sciences of the United States of America. 2011;**108**(50):19867-19872

[40] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 2015;**33**(3):290-295

[41] Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, et al. NGSQC: Cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010;**11**(Suppl 4):S7

[42] Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC Genome Browser database: 2016 update. Nucleic Acids Research. 2016;**44**(D1):D717-D725

[43] Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013;**14**(2):178-192

[44] Medina I, Salavert F, Sanchez R, de Maria A, Alonso R, Escobar P, et al. Genome Maps, a new generation genome browser. Nucleic Acids Research. 2013;**41**(Web Server issue):W41-W46

[45] Health NIo. The Cancer Genome Atlas