New Jersey Institute of Technology
Digital Commons @ NJIT

Theses | Electronic Theses and Dissertations

Spring 5-31-2013

RNA-sequence analysis of human melanoma cells

Jharna Miya
New Jersey Institute of Technology

Follow this and additional works at: https://digitalcommons.njit.edu/theses

 Part of the Bioinformatics Commons, and the Computer Sciences Commons

Recommended Citation

Miya, Jharna, "RNA-sequence analysis of human melanoma cells" (2013). *Theses*. 165.
https://digitalcommons.njit.edu/theses/165

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at Digital Commons @ NJIT. It has been accepted for inclusion in Theses by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

NJIT New Jersey's Science & Technology University

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

RNA-SEQUENCE ANALYSIS OF HUMAN MELANOMA CELLS

by
Jharna Miya

RNA-sequencing refers to the use of high throughput sequencing technologies that are used to sequence cDNA in order to get the complete information of a sample's RNA content. The objective of this study is to analyze this data in different aspects and to characterize gene expression. Besides this characterization, the data was also used to investigate the effect of sequencing depth on gene expression measurements.

This research focuses on quantitative measurement of expression levels of genes and their transcripts. In this study, complementary DNA fragments of cultured human melanoma cells are sequenced and a total of 139,501,106 million 200-bp reads from two samples affected with the disease are obtained. The RNA-seq is performed by first mapping the sequence reads to the reference human genome sequence (NCBI 36.1 [hg19] assembly) using Tophat and Bowtie software's. Then, using Cufflinks software the alignments are assembled into gene transcripts and relative abundances are obtained. Finally, differentially expressed genes are found by comparing the affected samples with a control sample.

The findings are represented in the form of graphs, which best signify the gene expression. This graphical representation of the results will allow the readers to study the expression and structure of genes in human melanoma cells.

RNA-SEQUENCE ANALYSIS OF HUMAN MELANOMA CELLS

by
Jharna Miya

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

May 2013

APPROVAL PAGE

RNA-SEQUENCE ANALYSIS OF HUMAN MELANOMA CELLS

Jharna Miya

Dr. Zhi Wei, Thesis Advisor Assistant Professor of Computer Science, NJIT	Date
Dr. Usman W. Roshan, Committee Member Associate Professor of Computer Science, NJIT	Date
Dr. Dimitri Theodoratos, Committee Member Associate Professor of Computer Science, NJIT	Date

BIOGRAPHICAL SKETCH

Author: Jharna Miya

Degree: Masters of Science

Date: May 2013

Undergraduate and Graduate Education:

- Masters of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2013

- Bachelor of Technology in Biotechnology,
  College of Engineering & Technology, IILM Academy of Higher Learning, U.P, India, 2009

Dedicated to
my family, who's love and constant support
has been with me throughout my life
and continues to be.
Touchwood

ACKNOWLEDGMENT

I would like to thank my thesis advisor, Dr. Zhi Wei, for his valuable guidance and consistent encouragement throughout my research work. His mentoring surely kept me going even through the hard times. I am highly grateful to Dr. Usman Roshan and Dr. Dimitri Theodoratos for their moral support and valuable feedbacks time to time. I would specially thank them for being a part of my thesis committee as well.

I would also like to thank my friend Wei (soon to be, Dr. Wei Wang) for giving me inputs often. He has been a source of inspiration to me. Last but not the least; I thank NJIT for making my dream come true and for providing the cutting edge technology environment for my professional as well as personal growth.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

nt	Nanotesla
GB	Gigabytes
gHz	Gigahertz
PERL	Practical extraction and reporting language
UNIX	UNiplexed Information Computing System (UNICS)
bp	Basepair
RNA	Ribo Nucleic Acid
DNA	Deoxyribonucleic Acid
NGS	Next Generation Sequencing
cDNA	Complementary Deoxyribonucleic Acid
FPKM	Fragments Per Kilobase of transcript per. Million fragments mapped
RPKM	Reads Per Kilobase per Million
SAM	Sequence Alignment/Map
BAM	Binary version of the Sequence Alignment/Map (SAM) format
GFF	General Feature Format
GTF	General Transfer Format
R	R Statictical software

CHAPTER 1
INTRODUCTION

1.1 Background Biology

The basic unit of life in all living organisms is the cell. Every single cell possesses the complete information of its creator or parent cell. This information is passed on to the next set of cells or daughter cells during a process called cell division. Development and functioning instructions along with the genetic information are stored in a type of nucleic acid in the DNA. A gene is a specific segment of the DNA that contains all the necessary coding information to instruct the cell for the creation of a protein or functional ribonucleic acid (RNA). A protein is a biochemical compound facilitating biological functions.

All cells perform few common activities known as “housekeeping processes”. Additionally some specific activities are carried out by specialized cells. All of these cell types have identical genes, but they differ in their gene expression. Gene expression is the synthesis of proteins or functional RNA based on the information of the DNA.

Several steps in the gene expression process may be modulated, including the transcription, RNA splicing, translation and post-translational modification of a protein. This is known as ‘central dogma’ of molecular biology.

Figure 1.1 The central dogma.
Source: [2].

1.2 Next Generation Sequencing [NGS]

Much like the development of microarray technology for measuring gene expression in the late 1990s and early 2000s, the development of technologies for high-throughput sequencing, termed next-generation sequencing (NGS) technologies, is having an impact on the types of questions that biologists can ask these days.

The automated Sanger method is considered as a "first generation" sequencing technology. The newer methods are referred to the term next generation sequencing. One advantage of the technology is that it allows single-investigator labs to generate data that was previously the domain of large-scale sequencing center's [3].

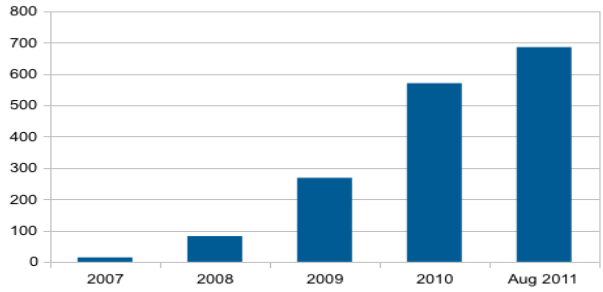

Figure 1.2 Number of publications by year deposited in PubMed referring to the term "Next Generation sequencing". (Figure generated by the author with PubMed queries.) Source: [3].

1.2.1 NGS Technologies

There are many NGS technologies, two of the most common are mentioned below with brief introduction. The main aim of the following techniques is to show how the incorporation of nucleotides into a DNA fragment is translated into a measurable signal which can be eventually quantified as gene expression.

1.2.1.a Ion Semiconductor Sequencing

It is a method of DNA sequencing based on the detection of hydrogen ions that are released during the polymerization of DNA. A microwell containing a template DNA strand to be sequenced is flooded with a single species of deoxyribonucleotide triphosphate (dNTP).

If the introduced dNTP is complementary to the next unpaired nucleotide in the leading template nucleotide, the nucleotide will be integrated through DNA polymerase. The incorporation of a deoxyribonucleotide (dNTP) into a growing DNA strand involves the formation of a covalent bond and the release of pyrophosphate and a positively charged hydrogen ion. With the release of the positively charged ion the reaction changes the pH of the solution, which is detected by a hypersensitive ion sensor (ISFET, ionsensitive field-effect transistor).

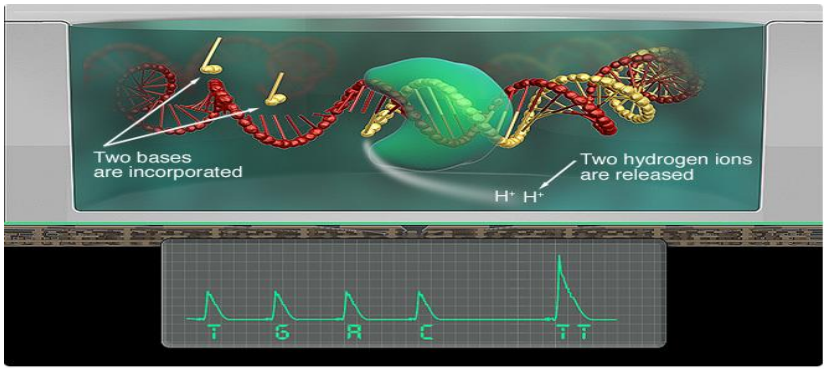

Figure 1.3 In ISS, as the new two nucleotide bases are incorporated into the DNA strand, 2 hydrogen ions are released. If there are two identical bases on the DNA strand, the voltage will be double, and the chip will record two identical bases called. Because this is direct detection - no scanning, no cameras, no light - each nucleotide incorporation is recorded in seconds.
Source: [4].

1.2.1.b Pyrosequencing

It is based on the detection of pyrophosphates that are released during the polymerisation of DNA. Like ion semiconductor sequencing, a small well containing a template DNA is sequentially flooded with one of four different dNTPs. If the introduced dNTP is complementary to the next unpaired nucleotide, the DNA polymerase incorporates the dNTPwith the release of pyrophosphate. Each incorporation event is accompanied by release of pyrophosphate (PPi) in a quantity to the amount of incorporated nucleotide.

To detect the released amount of pyrophosphates, they are translated to light emission in two enzymatic steps. In the first step the ATPsulfurylase converts pyrophosphate (PPi) to adenosine-triphosphate (ATP). The newly formed ATP is the fuel for the second step (luciferase reaction).

In the second step the luciferase enzyme converts luciferin to oxiluciferin and emitted light. The amount of the released pyrophosphate, the amount of the newly formed ATP and the the detected peak of the emitted light is proportional [5].

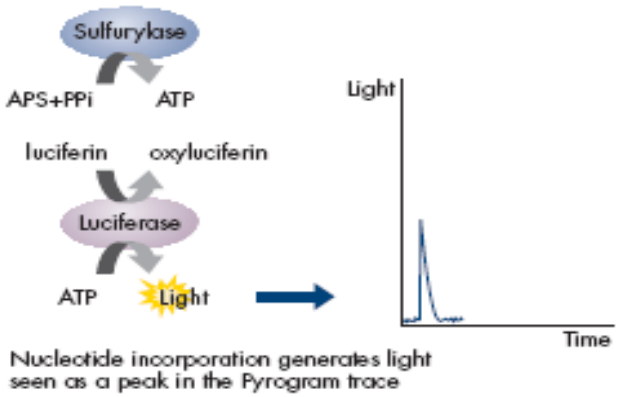

Figure 1.4 Nucleotide incorporation generates light seen as a peak in the Pyrogram trace. Source: [6]

1.2.2 NGS Applications

1.2.2a Transcriptome Sequencing

Transcriptome consists of all the three types of RNA molecules, mRNA, rRNA, tRNA and also any other non-coding RNA produced inside the cell. RNA-seq also known as "Whole Transcriptome Shotgun Sequencing" [7] or "a revolutionary tool for transcriptomics"[8], refers to the use of high-throughput sequencing technologies used to sequence and study a specific transcript or interactions between transcripts [9].

1.2.2b Genomic Sequencing

There are two different approaches of sequencing a genome, de novo sequencing or resequencing.

De novo sequencing is used if there is no reference genome available. A reference genome, also known as a reference assembly, is a digital nucleic acid sequence database as a representative example of a species genetic code. Once you have the reference sequence for an organism, you can perform resequencing, also called comparative sequencing, to characterise the genetic diversity within the organisms' species or between closely related species. This includes the analysis of copy-number variations (CNVs) and single-nucleotide polymorphism (SNP, point-mutations in the DNA). CNVs are variations in the count of copies of one or more section of the DNA. Understanding CNVs and SNPs can reveal information of drug response, human diseases and human genome evolution.

1.3 RNA-Sequencing

RNA is generated by transcription from DNA, the information is already present in the cell's DNA. RNA as such is less stable in the cell, and also more prone to nuclease attack experimentally. RNA-seq sequences cDNA in order to get information about a sample's RNA content. This technique reveals gene expression information, such as how different alleles of a gene are expressed, detects post transcriptional mutations, or identifies fusion genes. However, it is sometimes desirable to sequence RNA molecules as in the case of Helicos BioSciences.

Gene expression profiling, also called "transcriptomics", is the measurement to find out which genes are switched "on" or "off". If a gene is considered "on", it is used to produce mRNA. Measuring mRNA concentration is still a useful tool in determining the machinery of the cell. Altered levels of a specific mRNA sequence result into different protein structures. Expression profiling experiments, like RNA-Seq, involve measuring the relative amount of mRNA expressed in two or more experimental conditions.

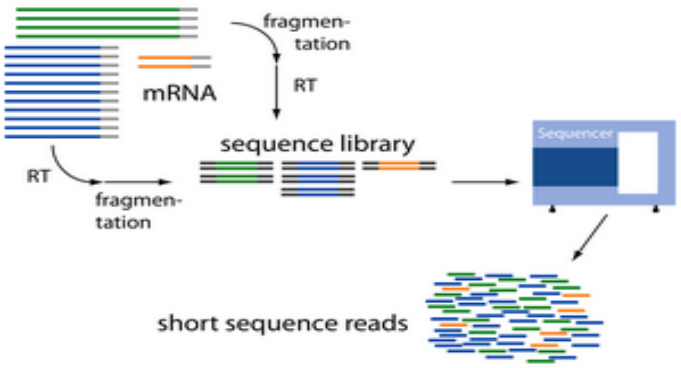

Figure 1.5 Here mRNA is converted to cDNA during Reverse Transcription[RT]. Sequencer generates short reads of these cDNA sequences that are further analysed.
Source: [10].

This section deals with the sequencing of RNA, in particular with the sequencing of mRNA, but this is not the only purpose of transcriptomics. Transcriptomics is also focusing on noncoding RNA (ncRNA). A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein (i.e. rRNA, tRNA, miRNA, siRNA, piRNA). These RNAs play an important role in how the genome is regulated and how traits are passed on or eliminated by environmental and genetic factors.

1.3.1 RNA-Seq Workflow

1.3.1a Sample Preparation

The amount of RNA required is dependent on the sequencing technology and the method of priming used. Larger RNA molecules, such as mRNA must be fragmented into smaller pieces (200-500 bp). Common fragmentation methods use either RNA fragmentation (before cDNA conversion) or cDNA fragmentation (after cDNA conversion). The population of mRNA is converted to a cDNA library with adapters attached to one or both ends (single- or paired-end-sequencing). The purpose of the adapter is to place a sequencing primers onto the adapter sequence [8].

1.3.1b Sequencing

A detailed explaination of different technologies and their protocols of sequencing machines is beyond the scope of this thesis. A general summarised example of sample preparation for RNA-seq experiments with Illumina Genome Analyzer II has been provided to illustrate the basic concept, which is alike to other technologies. Following protocol overview is summarised in brief from [Wilhelm and Landry 2009] [11].

- Fragmentation of cDNA
- Purification, end repair and tailing of cDNA fragments
- Adapter ligation
- Size-based purification of ligation products using agarose gel and extraction kits for separation
- PCR of ligation products
- Purification and sequencing of the fragments

1.3.1c Read Mapping

The main challenge in the data analysis step is to map the reads back to the reference genome. At first the data needs to be filtered by low quality reads. This filtering is not difficult and improves the quality and performance of further downstream analysis. The result of an RNA-seq workflow is the expression score for either genes or exons [3].

To derive expression scores a method has to map the reads back to the reference genome to assign each read to an exon or gene. Another question of read mapping is how to deal with repetitive sequences in the genome. For complex organisms, like human or mouse, repetitive sequences represent nearly 50% of the genome.

The simplest approach to reads which match to repetitive sequences, is to remove this reads from the results unless they can be matched unambiguously. One approach to improve the results of repetitive sequences is to use “paired-end” fragments. This approach involves sequencing from both ends of a single molecule of a known size. If one of the two reads of the paired-end sequencing approach maps to a highly repetitive region in the genome and the other one does not, it allows both reads to be mapped to the reference genome unambiguously.

Paired-end sequencing maps 93% of the reads to 85% single-read sequencing [11] [12]. There are different algorithmic approaches to map RNA-Seq reads to a reference transcriptome. The first one is called “unspliced read alignment” and the second is called “spliced alignment”. In the unspliced read alignment approach are reads aligned to a reference genome (or transcriptome) without allowing large gaps. This approach is separated in two main categories, “seed methods” and “Burrows-Wheeler transform methods”.

Seed methods such as MAQ [12] and Stampy [13] find matches for short subsequences of the reads (seeds) which match perfectly the reference. Each seed is extended with a more sensitive method (i.e. Smith-Waterman). The Burrows-Wheeler transform methods such as BWA [14] and Bowtie [15] compact the genome into a data structure that is very efficient when searching perfect matches. The performance of this approach is exponentially decreased by allowing mismatches because the algorithm is based on a iteratively search [16], [3].

RNA-Seq read mapping

- unspliced read alignment
  - seed methods
  - Burrows-Wheeler transform methods
- spliced aligners
  - exon first
  - seed and extend

1.3.2 Power of RNA-Seq

RNA-Seq has evident advantages over already existing approaches and has changed the view of eukaryotic transcriptomes. Novel detection of transcriptomes is not challenge anymore for RNA-Seq in comparison to hybridisation- based approaches.

RNA-Seq is useful to study complex transcriptomes. Short reads (30-bp) can give information about how two exons are connected, whereas longer reads or pair-end short reads reveal connectivity between multiple exons. In comparison to Tiling arrays RNA-Seq has a very low (if any) background signal because of the explicit mapping of sequencing reads to unique regions of the genome.

Other advantages to Tiling arrays are the larger dynamic range and the high level of reproducibility in technical and biological replicates. RNA-Seq is also used to detect entirely novel genes, novel splice variants of existing genes and it is used for gene fusion detection [3].

1.4 Melanoma Cells

Melanoma is a malignant tumor of melanocytes. Melanocytes are the cells that produce the dark pigment, melanin, which is responsible for the color of skin. These cells mainly occur in skin, but are also found in other parts of the body, including the bowel and the eye. Melanoma can originate in any part of the body that contains melanocytes. [17]

Melanoma is least common than other skin cancers. However, it is much more dangerous if found in later stages. All cancers are caused by damage to the DNA inside cells. This damage can be inherited in the form of genetic mutations, but in most cases, it builds up over a person's lifetime and is caused by factors in their environment. DNA damage causes the cell to grow out of control in abundance, leading to a tumor. Melanoma is usually caused by damage from UV light from the sun, but UV light from sunbeds can also contribute to the disease [18].

Figure 1.6 The slide shows a melanoma on a patient's skin.
Source: [19].

CHAPTER 2
MATERIALS AND METHODS

2.1 Datasets – Human Melanoma.

Geo Expression Omnibus, a public functional genomics repository contains the datasets that I have used for my research under accession number - GSE33092. The raw data are available in SRA format.

The three datasets used in the study are:

GSM819489 - RNA-seq from primary human melanocytes infected with RFP lentivirus (control sample)

GSM819491 - RNA-seq from primary human melanoma sample1

GSM819492 - RNA-seq from primary human melanoma sample2

Each of sample1 and sample2 have around 70 +/- 10 million reads with average length of fragment 350bp +/- 50bp, Read length of 200bp, Pair-end reads.

The datasets were converted from SRA format to FASTQ format using the SRAToolkit software. For most of the analysis, the sequences from sample-1 and sample-2 were pooled to create a 140-million-read data set.

This pooled dataset was created using SAMtools. The BAM files [accepted_hits.bam] from the tophat results of both the samples1 & 2 were merged together. This merged BAM file was converted into SAM format [human readable format] and then PERL script was written to divide the 140 million (approx.) dataset into 14 fold with 10 million reads each. Thus, 14 sub-datasets were generated for this research.

Next Generation Sequencing [NGS] Pipeline

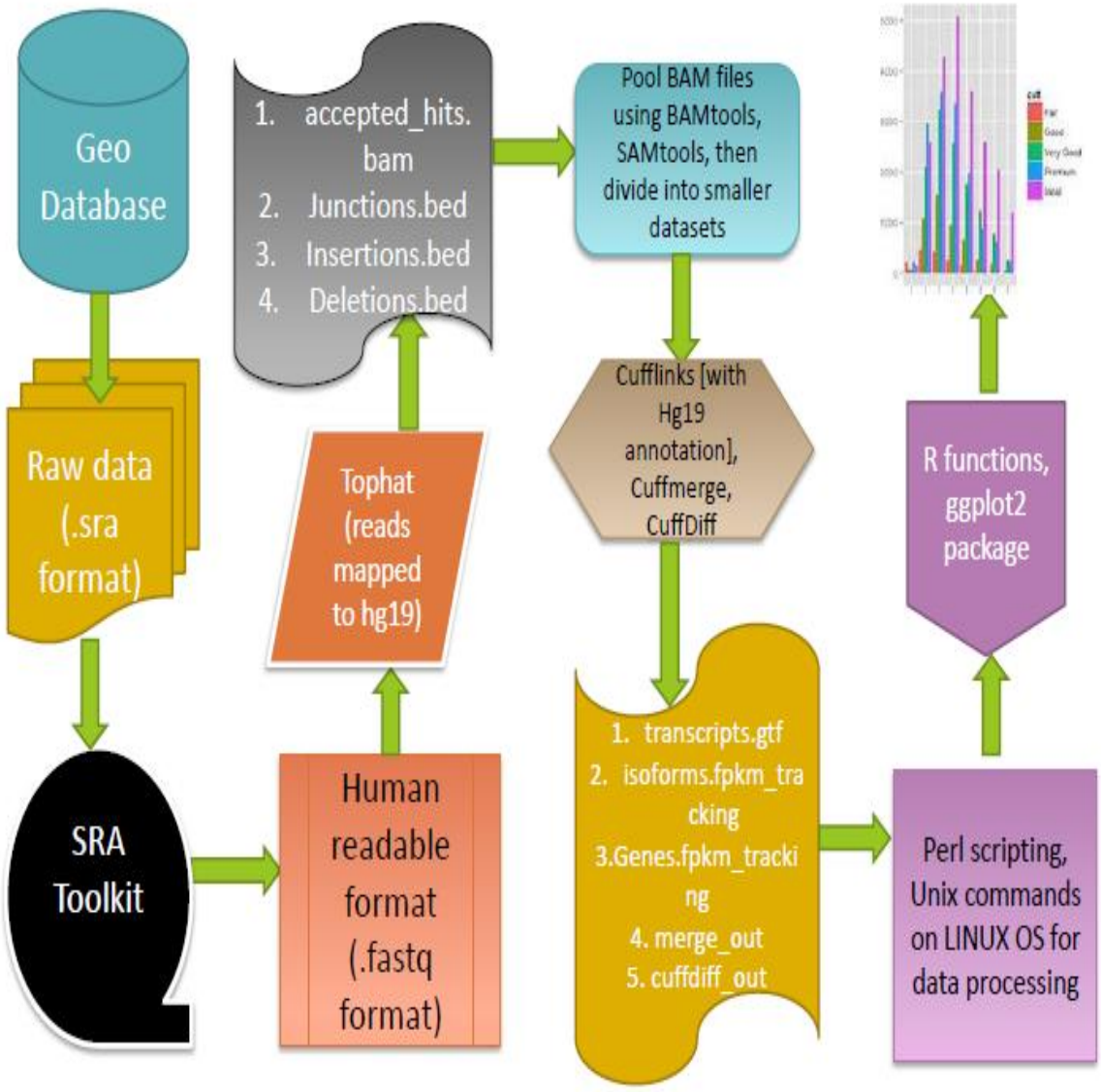

Figure 2.1 The next generation sequencing pipeline. Overview of the methodology developed for this research study.

2.2 Read Alignment to Reference Human Genome with Tophat and Bowtie

RAW reads were mapped to hg19 [reference human genome in fastq format] using Tophat software. RNA-seq reads were aligned to this human genome using the ultra-high throughput short read aligner Bowtie, which then analyzes the mapping results to identify splice junctions between exons. Alignment with Tophat produced the following files.

1. accepted_hits.bam. A list of read alignments in SAM format. SAM is a compact short read alignment format.

2. junctions.bed. A UCSC BED track of junctions reported by TopHat. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction.

3. insertions.bed and deletions.bed. UCSC BED tracks of insertions and deletions reported by TopHat.

2.3 Transcript Assembly Using Cufflinks

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols [20].

Cufflinks takes a text file of SAM alignments, or a binary SAM (BAM) file as input which in my case was the “accepted_hits.bam” file. The RNA-Seq read mapper TopHat produces output in this format.

Following are the output files from Cufflinks:

1. transcripts.gtf - This GTF file contains Cufflinks' assembled isoforms.There one GTF record per row, and each record represents either a transcript or an exon within a transcript.

2. isoforms.fpkm_tracking - This file contains the estimated isoform-level expression values in the generic FPKM Tracking Format.

3. genes.fpkm_tracking - This file contains the estimated gene-level expression values in the generic FPKM Tracking Format.

Using Cufflinks (Trapnell et al. 2010) the alignments were assembled into gene transcripts and their relative abundances were calculated. For the analysis; Cufflinks was provided with Gencode (version 3c NCBI36) (Harrow et al. 2006) gene annotations. The analysis was restricted to levels 2 and 3 Gencode genes that are annotated as ‘‘protein coding’’ or ‘‘processed transcript’’; in this study, this set of gene models is referred to as ‘‘Gencode.’’

To investigate the effect of sequencing depth on various expression profiling measurements, smaller subsets of the pooled data set were created, analyzing depths of 1 to 10 million reads (in intervals of 10 million reads), 10 to 20 million reads (in intervals of 10 million reads), and 20 to 30 million reads (in intervals of 10 million reads) and so on until 140 million reads. All these 14 sub-datasets were put through Cufflinks for further analysis.

Cuffmerge

Cufflinks includes a script called Cuffmerge that is used to merge together several Cufflinks assemblies and automatically filters a number of transfrags that are probably artfifacts. A reference GTF file is also provided to the script in order to gracefully merge novel isoforms and known isoforms and to maximize overall assembly quality. The main purpose of this script is to make it easier to make an assembly GTF file suitable for use with Cuffdiff.

2.4 Differential Gene Expression using CuffDiff.

Cufflinks includes a program, "Cuffdiff" was used to find significant changes in transcript expression, splicing, and promoter use. It takes the aligned reads from two or more conditions and reports genes and transcripts that are differentially expressed using a rigorous statistical analysis.

Cuffdiff takes the GTF2 file of transcripts as input (produced by Cuffmerge), along with two or more BAM files containing the fragment alignments for two or more samples. It produces a number of output files that contain test results for changes in expression at the level of transcripts, primary transcripts, and genes. It also tracks changes in the relative abundance of transcripts sharing a common transcription start site, and in the relative abundances of the primary transcripts of each gene.

2.5 Data Analysis and Plotting Graphs

All the data were processed in LINUX environment using PERL scripting language. Algorithms were designed in the script as per the requirement of the analysis. 'AWK' and 'SED' utilities were mostly used for the extraction and curing of data. This processing led to the simplified files that were to be used for generating data for tables and graphs.

R environment was extensively used as well for simplifying data further and also generation of graphs. Graphs were plotted for representation of my analysis results using the GGplot2 package in R.

GGPLOT2 library is a relatively new and one of the main graphing packages supported by R. It has great control features designed to create appropriate graphs. The data for tables has been generated after running the files through the PERL scripts. These tables were loaded to R using "read.table" function and were plotted using GGPLOT2 package.

2.6 Computer Specifications

PHI cluster at NJIT has been used for running all the jobs. PHI is virtual machine running on Vmware. LINUX OS was used for majority of data processing.

Table 2.1 PHI Cluster Specification.

Operating system	SL 5.5
Model	VMware[5]
Number of computer nodes	1
Processor * core per nodes	1 * 4
Processor type	AMD Opteron Model 8384
Processor speed, GHz	2.7
RAM per node, GB	32
RAM per Processor, GB	32
RAM per core, GB	8
Total RAM, GB	32

From LINUX OS (run on Phi) the processed data was transferred to the Windows OS, which was further loaded and modified on R x64 3.0.0 platform for generating plots. Following are the specifications of Windows machine [PC].

Table 2.2 Windows Machine Specifications

Operating system	**Processor**	**Memory**
WINDOWS 7 64-BIT	Intel(R) Core(TM) i5-2410M 2.30 GHz	Hard Disk : 640 GB RAM : 6 GB

CHAPTER 3

RESULTS AND DISCUSSION

For most of the analysis, the pooled dataset was used, and for the rest of it the 14 subsets obtained from this pooled dataset were used. Using all of the sequence reads, the expression levels in the genes were estimated. Expression levels are measured in "fragments per kilobase of exon model per million mapped reads" (FPKM) (Trapnell et al.2010), and the expression level for a gene is the sum of the FPKM values of its isoforms.

The analysis has been restricted to levels 2 and 3 Gencode genes that are annotated as ''protein coding'' or ''processed transcript''; in this study, this set of gene models is referred as ''Gencode.''

Analysis of all transcripts with expression levels greater than zero includes FPKM values that are very close to zero. Thus, an FPKM value of 0.05 has been set as the lower bound in subsequent analysis.

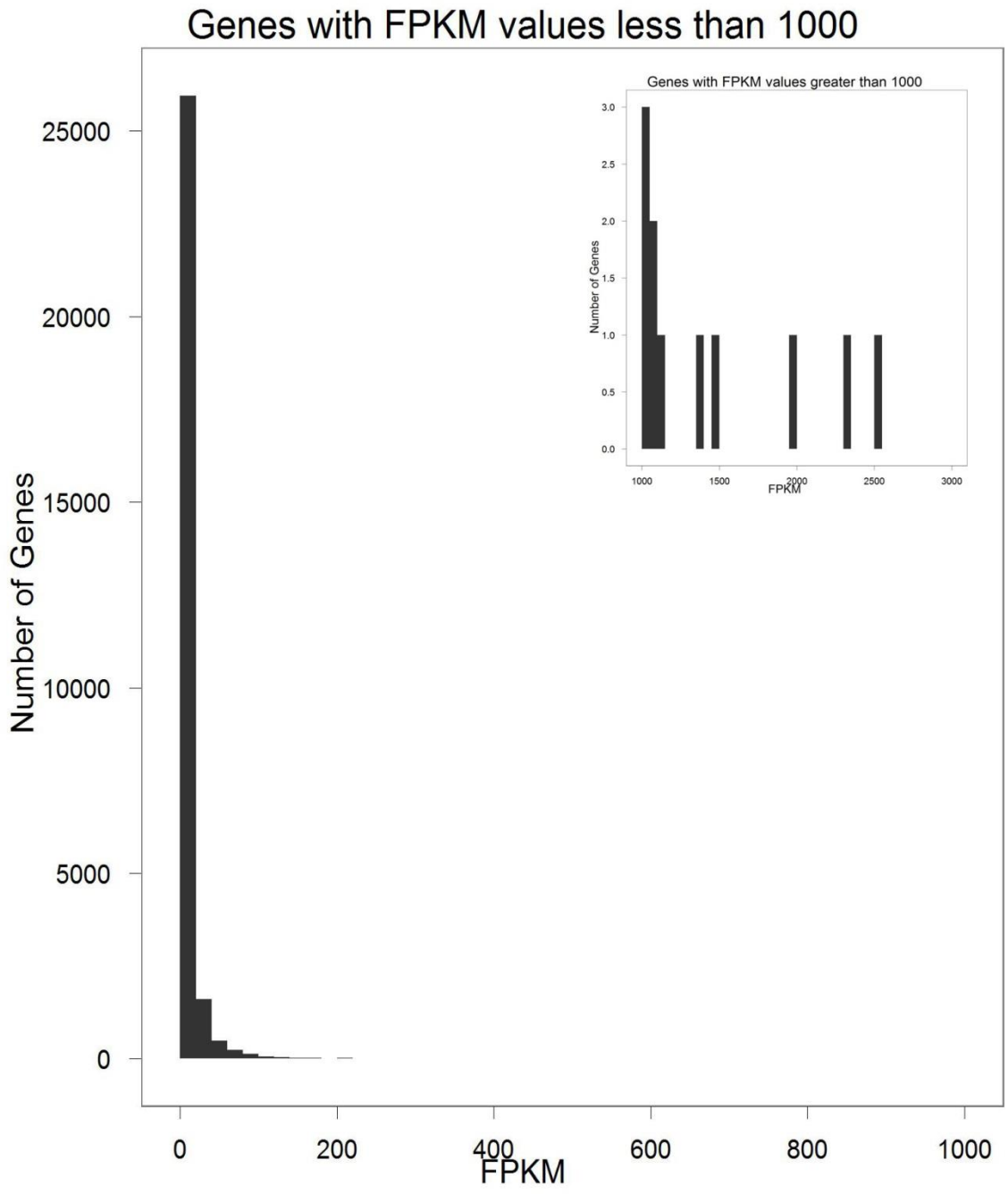

Figure 3.1 Distribution of FPKM values for Gencode genes. The distribution of gene expression values is skewed right; the median and mean FPKM values 0.72 and 8.15 respectively. The main figure shows genes FPKM values less than 1000. (Inset) Genes with FPKM values greater than 1000.

Expression landscape across chromosomes was surveyed by determining the fraction of genes that are expressed within 1-Mb intervals (Figure 3.2).

Figure 3.2 Distribution of expressed genes by chromosome. For each chromosome, number (y-axis) of Gencode genes residing in 1-Mb intervals along the chromosome (x-axis depicts physical distance in megabases) were plotted. (Red) The number of genes that are expressed (FPKM>=0.05); (blue) the number that are not expressed.

The RNA-seq data of Sample1 was compared with the RNA-seq data of sample2. Correlation coefficient calculation was done to measure the strength and direction of linear relationship between the two variable samples.

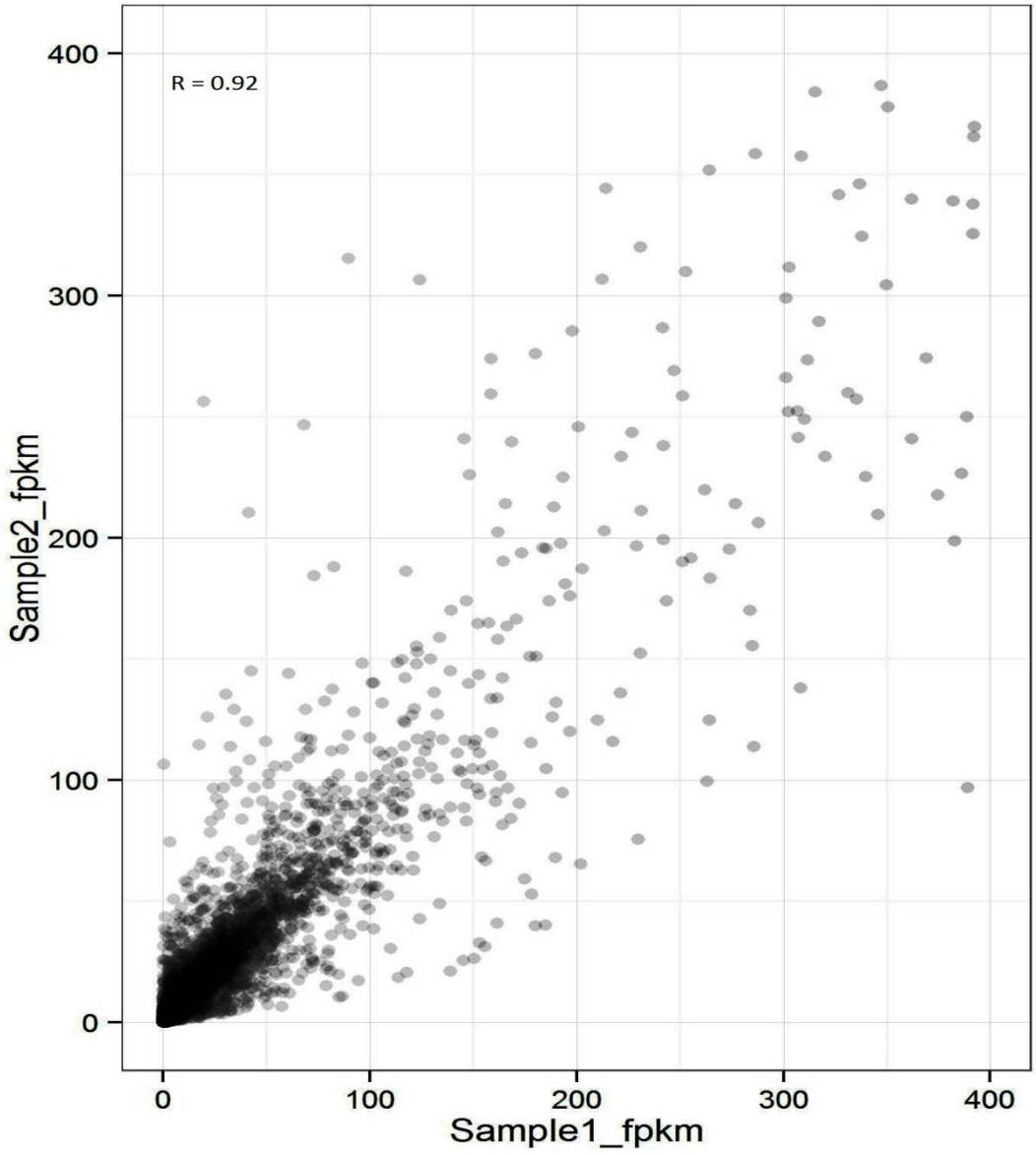

Figure 3.3 Scatter plot of expression values from RNA-seq of both samples. Comparison of FPKM for the genes detected for both samples. For each gene the average expression values across 2 samples were plotted. Higher value of R = 0.92 implies very high correlation within the samples.

In designing an RNA-seq study, a parameter of interest is the sequencing depth needed to address the relationship between sequencing depth and expression levels. For this analysis the 140 million 400 bp read dataset was divided into 14 sub-datasets and analyzed on how the detection of a gene and the measurement of its expression level varies with increasing sequencing depth.

First, it was assumed the the ~140 million read data set gives a comprehensive catalog of transcribed genes and then assessed how many genes are detected in fraction of those reads. It was found that with 50 million reads 80% of the genes (FPKM >= 0.05) were detected and 95% of their transcripts were detected.

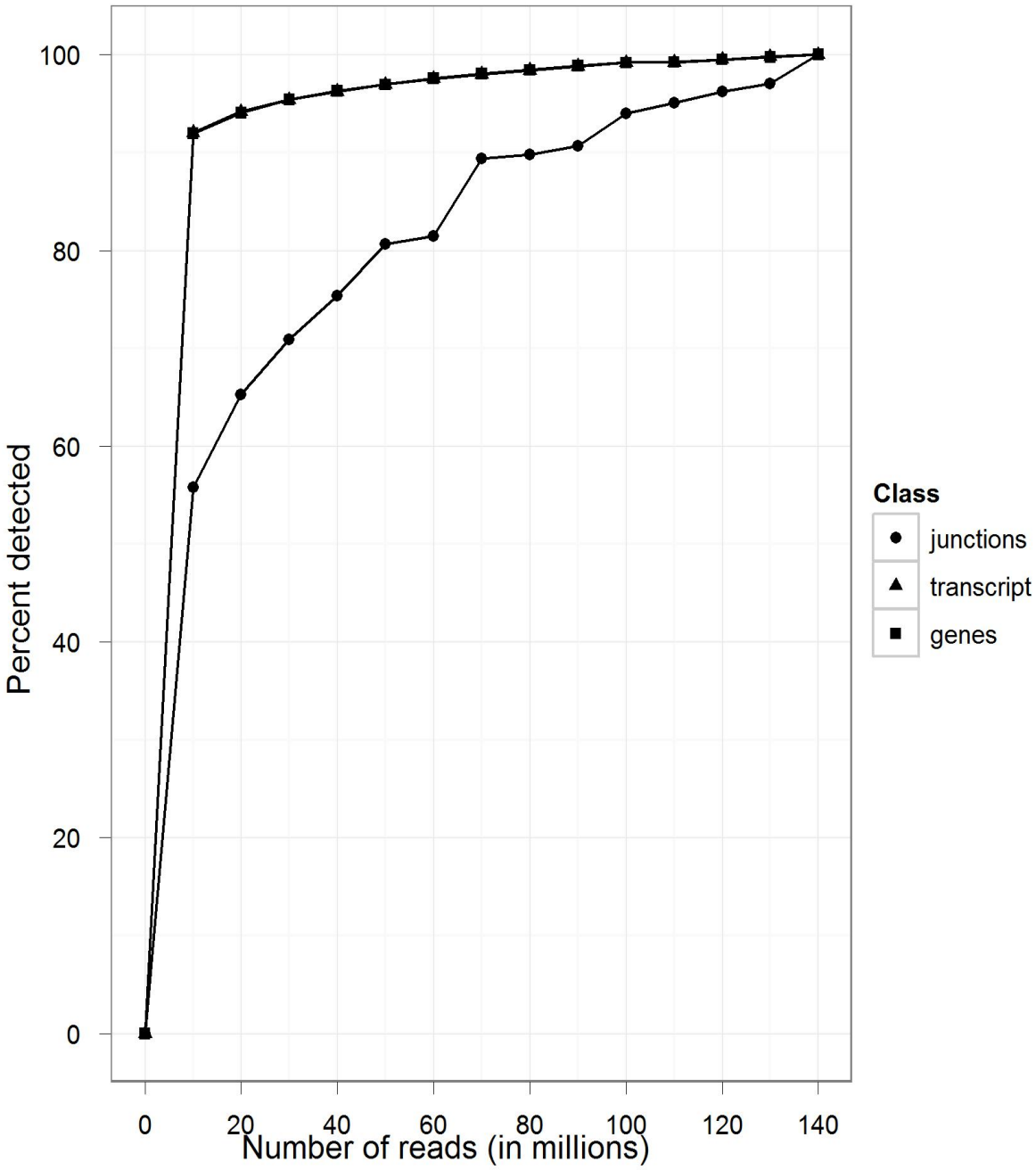

Figure 3.4 Number of junctions, transcripts and genes detected at different sequencing depths. The number of genes, transcripts and genes detected in the 140 million-data-set are assumed to be "final" values. Then, the percentages of these "final" values detected at various sequencing depths were determined. For example at 100 million reads 93% of genes and 98% of their transcripts were detected.

For most studies, information beyond whether a gene is expressed or not is important; accurate expression levels are needed. To study the robustness of expression levels it was assumed that the expression levels in the ~140-million-read data set were the "best estimates" and were then analysed the sequencing depth necessary to achieve these "final" values.

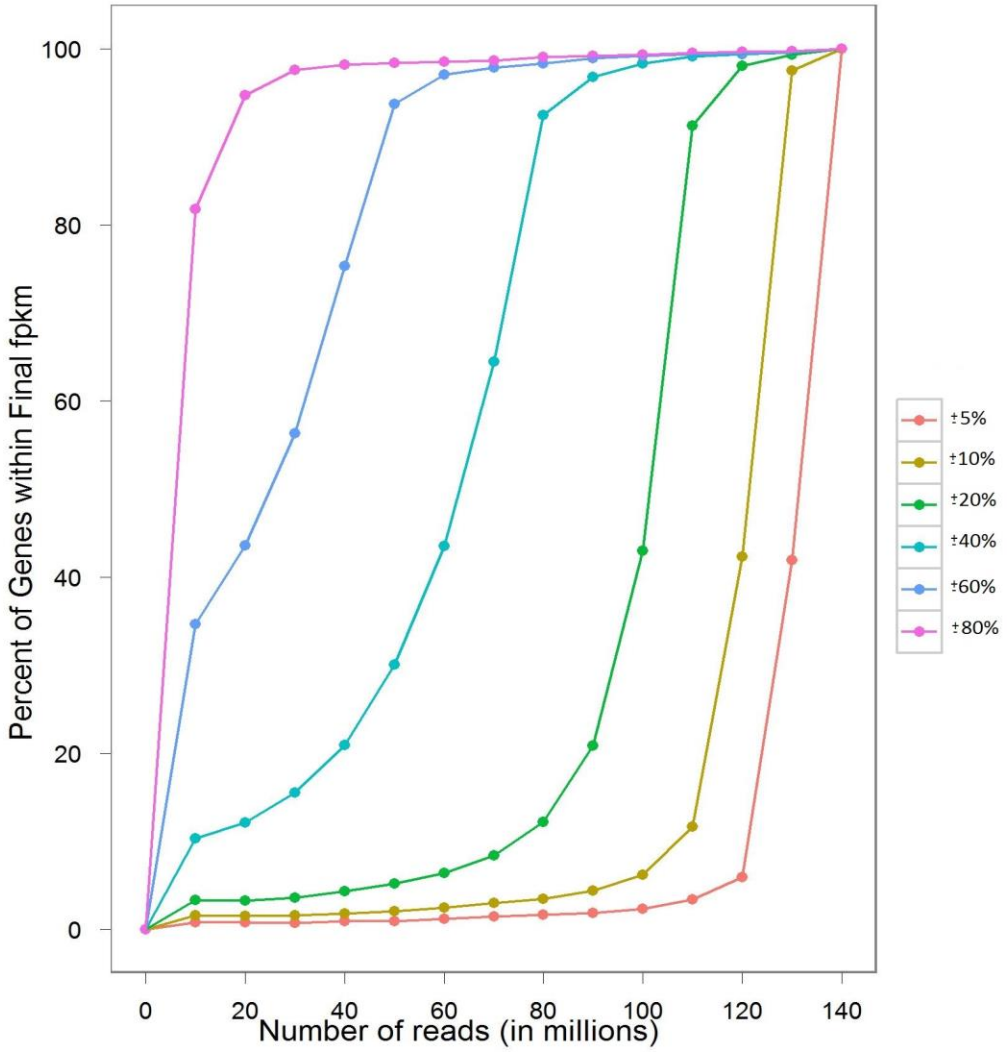

Figure 3.5 Gene expression levels at different sequencing depths. The percentages of genes that reach values within different percentages of the "final" level obtained at a depth of ~140 million reads were determined. With 100 million reads, only 6% of genes have FPKM measurements that are within 10% of their "final" value as compared to 97% at a depth of 130 million reads.

Furthermore, the coverage needed to study the relative abundance of alternatively spliced forms of genes was studied. Here, again it was found that deep sequencing depths are crucial. For example CD74 (Figure 3.6a) is a gene with three isoforms: NM_001025158, NM_001025158 and NM_004355. Observing the graph trend we can see that for isoform NM_004355 which has the highest FPKM values at 10 million reads gets more expressed as the sequencing depth increases.

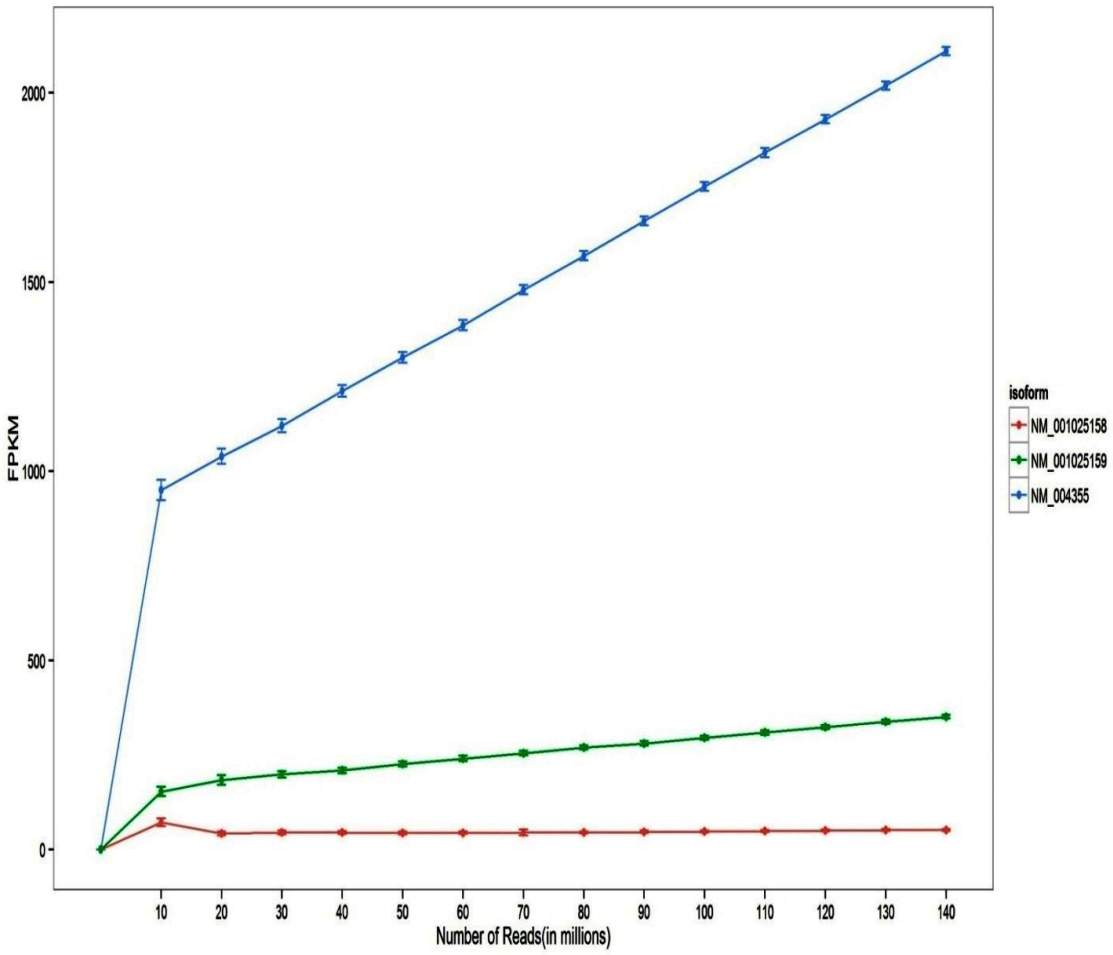

Figure 3.6a Expression values versus sequencing depth. FPKM values of three isoforms of CD74 are shown, the least abundant isoform (red line) reaches 20% of its "final" FPKM value with only 20 million reads; however, the expression values of the other two isoforms continued to increase with more reads. (Error bars represent 95% confidence interval)

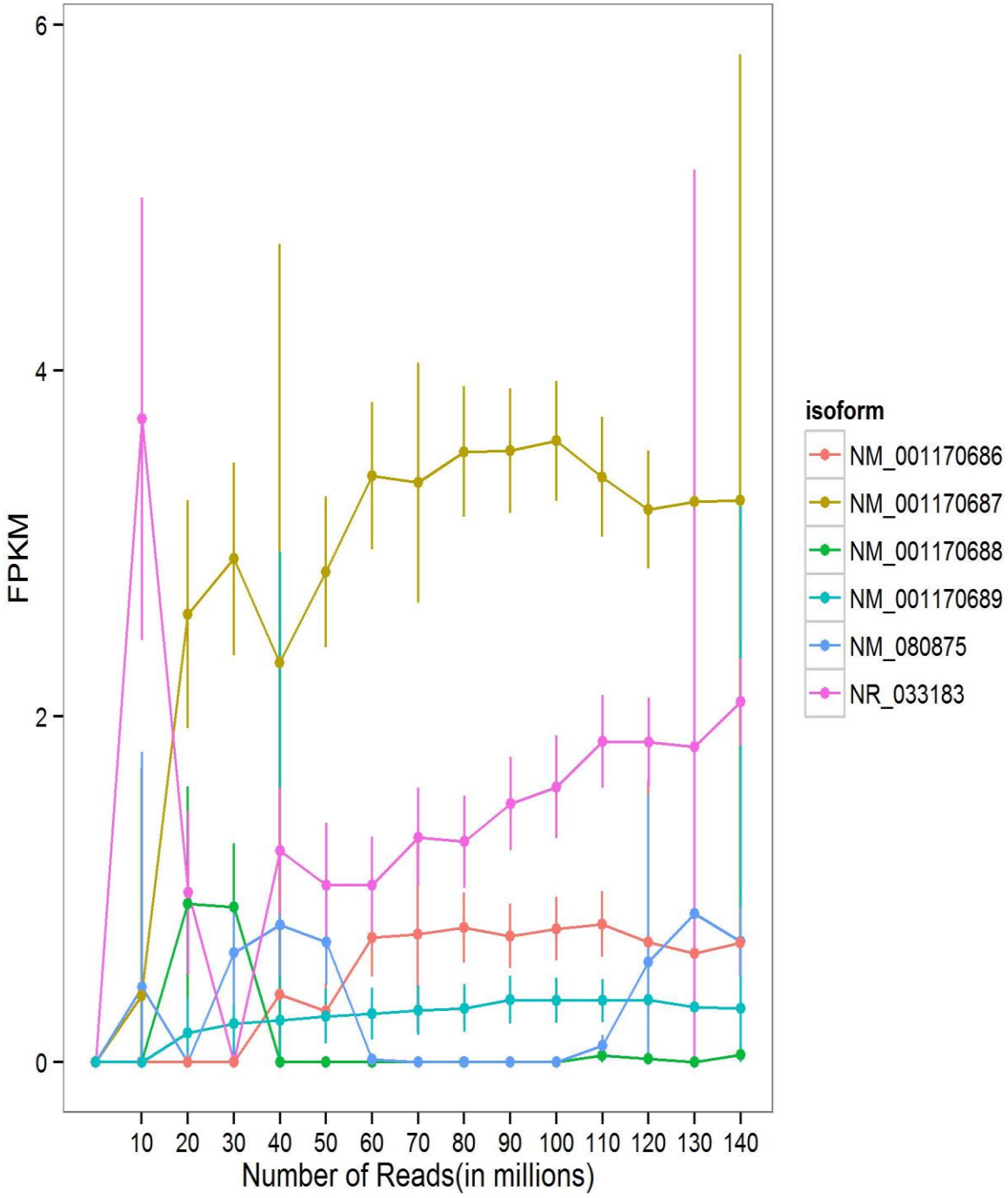

Figure 3.6b Expression values versus sequencing depth. FPKM values for gene MIB4 and its six transcripts at various sequencing depths. With 10 million reads, the expression level of NM_001170689 is underestimated and that of NR_033183 is overestimated. 20 million reads were needed to show that NM_001170687 (golden line) and not NR_033183 (pink line) is the most highly expressed isoform. (Error bars represent 95% confidence interval)

Differential analysis with cuffdiff:

Cufflinks includes a separate program, Cuffdiff, which calculates expression in two or more samples and tests the statistical significance of each observed change in expression between them. For this research, I ran Cuffmerge on the cufflinks result of the pooled dataset comprising sample-1 and sample-2. Thus, I provided two conditions as input to it.

Condition 1 – Control sample.

Condition 2 – pooled dataset of Melanoma samples1 and 2.

The output from this process was fed into Cuffdiff along with the BAM files obtained for both these conditions using Tophat previously.

The gene expression file was loaded into R and using the GGPLOT2 package, a Volcano plot was generated. It plots significance versus fold-change on the y- and x-axes.

A volcano plot is constructed by plotting the negative log of the p-value on the y-axis (usually base 10). This results in datapoints with low p-values (highly significant) appearing towards the top of the plot. The x-axis is the log of the fold change between the two conditions.

The genes with high significance appear in blue color above the dense population of genes the sit at the botton in red color. Graph 3.7 shows the volcano plot shown the differencially expressed genes in the two melanoma samples.

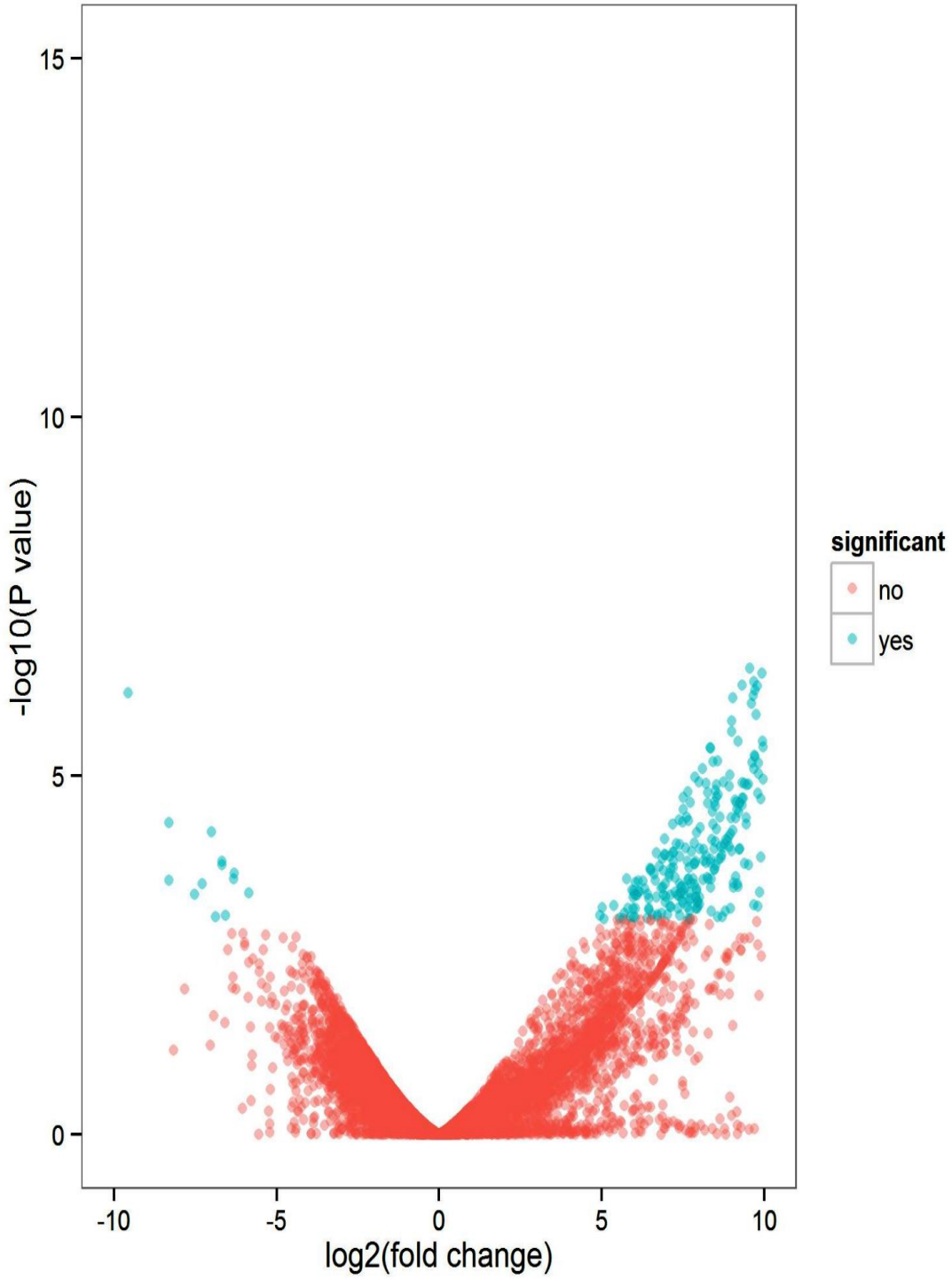

Figure 3.7 Volcano plot showcasing the differential expression of genes. A total of 308 significant genes were discovered in the two melanoma samples.

CHAPTER 4

CONCLUSION

More than 90% of the genes are alternatively spliced. For majority of the genes one isoform is predominantly expressed. While chromosome differs by gene density, the percentage of transcribed genes in each chromosome is variable.

The gene expression varies with increasing sequencing depth. With increasing sequencing depth also increases the number of genes, junctions and transcripts. Over increasing depth, we can find out the actual gene isoform expression pattern. There are 308 significant genes in the two melanoma samples.

It was found that with 50 million reads 80% of the genes (FPKM >= 0.05) were detected and 95% of their transcripts were detected. About 56% of total reads aligned uniquely to the human genome.

REFERENCES

1. [Morin et al. 2008] Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., und Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques, 45(1), 81–94.

2. The central dogma
http://guweb2.gonzaga.edu/faculty/cronk/CHEM198pub/L08-index.cfm
Accessed: 04/29/2013

3. Next generation sequencing
http://aboutme.biobyte.org/wp-content/uploads/2011/10/Next-Generation-Sequencing-and-its-_Applications-in-RNA-Seq.pdf
Accessed: 03/15/2013

4. Ion semiconductor sequencing
http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Semiconductor-Sequencing/Semiconductor-Sequencing-Technology/Ion-Torrent-Technology-How-Does-It-Work.html
Accessed: 03/18/2013

5. [Ronaghi et al. 1998] Ronaghi, M., Uhlen, M., und Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. Science (New York, N.Y.), 281(5375), 363, 365.

6. Pyrosequencing
http://www.pyrosequencing.com/DynPage.aspx?id=7454
Accessed: 03/18/2013

7. [Vera et al. 2008] Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., und Marden, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular ecology, 17(7), 1636–47.

8. [Wang et al. 2009] Wang, Z., Gerstein, M., und Snyder, M. (2009). RNASeq: a revolutionary tool for transcriptomics. Nature reviews. Genetics, 10(1), 57–63.

9. Transcriptome sequencing
http://en.wikipedia.org/wiki/RNA-Seq
Accessed: 03/15/2013

10. RNA sequencing
http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html
Accessed: 03/24/2013

11. [Wilhelm and Landry 2009] Wilhelm, B. T. und Landry, J.-R. (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods (San Diego, Calif.), 48(3), 249–57.

12. [Li et al. 2008] Li, H., Ruan, J., und Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research, 18(11), 1851–8.

13. [Lunter and Goodson 2010] Lunter, G. und Goodson, M. (2010). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome research.

14. [Li and Durbin 2009] Li, H. und Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–60.

15. [Langmead et al. 2009] Langmead, B., Trapnell, C., Pop, M., und Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology, 10(3), R25.

16. [Garber et al. 2011] Garber, M., Grabherr, M. G., Guttman, M., und Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nature methods, 8(6), 469–77.

17. Melanoma definition
http://en.wikipedia.org/wiki/Melanoma
Accessed: 04/14/2013

18. IARC Monographs on the evaluation of carcinogenic risks to humans. **55**. 1992. *Solar and ultraviolet radiation*.
Accessed: 04/22/2013

19. Human melanoma
http://en.wikipedia.org/wiki/File:Melanoma.jpg
Accessed: 3/24/2013

20. Cufflinks manual.
http://cufflinks.cbcb.umd.edu/
Accessed: 04/28/2013