VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

Ing. Layal Abo Khayal

TRANSCRIPTOMIC CHARACTERIZATION USING RNA-SEQ DATA ANALYSIS

TRANSKRIPTOMICKÁ CHARAKTERIZACE POMOCÍ ANALÝZY RNA-SEQ DAT

ZKRÁCENÁ VERZE DISERTAČNÍ PRÁCE

ABRIDGED DOCTORAL THESIS

Obor: Biomedicínská elektronika a biokybernetika

Školitel: Prof. Ing. Ivo Provaznik, Ph.D.

Rok obhajoby: 2017

ABSTRACT

The high-throughputs sequence technologies produce a massive amount of data, that can reveal new genes, identify splice variants, and quantify gene expression genome-wide. However, the volume and the complexity of data from RNA-seq experiments necessitate a scalable, and mathematical analysis based on a robust statistical model. Therefore, it is challenging to design integrated workflow, that incorporates the various analysis procedures. Particularly, the comparative transcriptome analysis is complicated due to several sources of measurement variability and poses numerous statistical challenges. In this research, we performed an integrated transcriptional profiling pipeline, which generates novel reproducible codes to obtain biologically interpretable results. Starting with the annotation of RNA-seq data and quality assessment, we provided a set of codes to serve the quality assessment visualization needed for establishing the RNA-Seq data analysis experiment. Additionally, we performed comprehensive differential gene expression analysis, presenting descriptive methods to interpret the RNA-Seq data. For implementing alternative splicing and differential exons usage analysis, we improved the performance of the Bioconductor package DEXSeq by defining the open reading frame of the exonic regions, which are differentially used between biological conditions due to the alternative splicing of the transcripts. Furthermore, we present a new methodology to analyze the differentially expressed long non-coding RNA, by finding the functional correlation of the long non-coding RNA with neighboring differential expressed protein coding genes. Thus, we obtain a clearer view of the regulation mechanism, and give a hypothesis about the role of long non-coding RNA in gene expression regulation.

KEYWORDS:

RNA-Seq, Differential Gene Expression (DGE), Alternative splicing, Differential Exon Usage (DEU), long non-coding RNA (lncRNA).

ABSTRAKT

Vysoce výkonné sekvenční technologie produkují obrovské množství dat, která mohou odhalit nové geny, identifikovat splice varianty a kvantifikovat genovou expresi v celém genomu. Objem a složitost dat z RNA-seq experimentů vyžadují škálovatelné metody matematické analýzy založené na robustníchstatistických modelech. Je náročné navrhnout integrované pracovní postupy, které zahrnují různé postupy analýzy. Konkrétně jsou to srovnávací testy transkriptů, které jsou komplikovány několika zdroji variability měření a představují řadu statistických problémů. V tomto výzkumu byla sestavena integrovaná transkripční profilová pipeline k produkci nových reprodukovatelných kódů pro získání biologicky interpretovovatelných výsledků. Počínaje anotací údajů RNA-seq a hodnocení kvality je navržen soubor kódů, který slouží pro vizualizaci hodnocení kvality, potřebné pro zajištění RNA-Seq experimentu s analýzou dat. Dále je provedena komplexní diferenciální analýza genových expresí, která poskytuje popisné metody pro testované RNA-Seq data. Pro implementaci analýzy alternativního sestřihu a diferenciálních exonů jsme zlepšili výkon DEXSeq definováním otevřeného čtecího rámce exonového regionu, který se používá alternativně. Dále je popsána nová metodologie pro analýzu diferenciálně exprimované dlouhé nekódující RNA nalezením funkční korelace této RNA se sousedícími diferenciálně exprimovanými geny kódujícími proteiny. Takto je získán jasnější pohled na regulační mechanismus a poskytnuta hypotéza o úloze dlouhé nekódující RNA v regulaci genové exprese.

KLÍČOVÁ SLOVA

RNA-Seq, diferenciální genová exprese (DGE), alternativní splicing, diferenciální použití exonů (DEU), dlouhá nekódující RNA (lncRNA).

TABLE OF CONTENTS

1. INTRODUCTION TO RNA SEQUENCING

RNA-seq can be identify as an assembly of experimental and computational methods to determine the identity and abundance of RNA sequences in biological samples. The experimental methods involve isolation of RNA from cell, tissue, or whole-animal samples, preparation of libraries that represent RNA species in the samples, actual chemical sequencing of the library, and subsequent bioinformatic data analysis. A critical distinction of RNA-seq from earlier methods, such as microarrays, is the incredibly high throughput of current RNA-seq platforms, the sensitivity afforded by newer technologies, and the ability to discover novel transcripts, gene models, and small noncoding RNA species.

RNA-seq methods are derived from generational changes in sequencing technology. First-generation high-throughput sequencing typically refers to Sanger sequencing. With capillary electrophoresis being utilized to deal with nucleic acid fragment lengths. Second-generation sequencing, also known as next-generation sequencing (NGS), refers to methods using similar sequencing by synthesis chemistry of individual nucleotides, but performed in a massively parallel format, so that the number of sequencing reactions in a single run can be in millions. A typical NGS run could consist of 6000M sequencing reactions of 100 nucleotides yielding 600 billion bases of sequence information. Third-generation sequencing refers to methods that are also massively parallel and use sequence by synthesis chemistry but have as templates individual molecules of DNA or RNA. Third-generation sequencing platforms have fewer sequencing reactions per run, in the order of a few millions, but the length of sequence per reaction can be larger and can easily run into the 1500 nucleotide range [1].

Data obtained from an RNA-seq experiment can be substantially informative, ranging from the identification de novo protein coding transcripts in embryonic stem cells to characterization of gene regulation and alternative splicing. Questions that can be answered using RNA-seq data include: What are the differences in the levels of gene expression in normal and cancer cells? What happens to the gene expression levels in cell lines missing a tumor suppressor gene? Which genes are up-regulated during the development of brain? How is gene splicing changed during oxidative stress? What novel miRNAs can we discover in a human embryonic stem-cell sample?

New data derived from RNA-seq platforms showed a vast diversity for gene structure, identified novel unknown genes, and shed light on noncoding transcripts of both small and long lengths [2].

2. THE THEORETICAL REVIEW

2.1 ISOLATION OF RNAS

RNAs are typically isolated from freshly dissected or frozen cells or tissue samples using commercially available kits. High-throughput RNA isolation systems relies mainly on RNA attached to magnetic particles which facilitate their washing and isolation. To prevent degradation of RNA, samples can be immersed in RNA storage reagents, or processed partially and stored as a phenolic emulsion. At this stage, RNA samples can also be enriched for size-specific classes such as small RNAs, using column systems (miRVana; Ambion). Alternatively, samples can be isolated initially as total RNA and then size selected by polyacrylamide gel electrophoresis. [3]

In almost all cases of total RNA isolation, genomic DNA will contaminate the sample. This is unavoidable, and even if the contamination is minor, the sensitivity and throughput of RNA-seq will eventually capture these contaminants. Therefore, it is common practice that total RNA-isolated samples are treated with DNase, to digest contaminating DNA prior to library preparation. Most DNase kits provide reagents for inactivating DNase once the contaminating DNA has been removed. The amount of total RNA required for RNA-seq library preparation varies. Standard library protocols require 0.1–10 μg of total RNA, and high-sensitivity protocols can produce libraries from as little as 10pg of RNA. It is becoming common that RNA from single cells is isolated and specific kits for these applications are becoming available. [4]

2.2 QUALITY CONTROL OF RNA

It is required that RNAs are quality checked for degradation, purity, and quantity prior to library preparation. Nanodrop and similar devices measure the fluorescent absorbance of nucleic acid samples typically at 260 and 280nm. As the device measures absorbance of the sample, it is not able to distinguish between RNA and DNA, and therefore cannot indicate whether the RNA sample is contaminated with DNA. Moreover, degraded RNA will give similar readings as intact RNA, and therefore we cannot know about the quality of the sample. The 260/280 absorbance ratio will, however, provide some information about contamination by proteins.[5]

Agilent Bioanalyzer is a microfluidics capillary electrophoresis-based system to measure nucleic acids. It offers advantages of sensitivity and accuracy for performing RNA separation, detection, and quantitation, coupled with a rapid, automated system. The electrophoresis being used for sizing nucleic acid samples. When size standards are run, the sizing and quantitation of RNAs in the sample provides critical information not only on the concentration, but also on the quality of nucleic acid. Degraded RNAs will appear as a smear at low-molecular weights, whereas intact total RNA will show sharp 28S and 18S peaks. The Bioanalyzer system contains a microchip that is loaded with size controls and space for up to 12 samples at a time. Samples are mixed with a polymer and a fluorescent dye, which are then loaded and measured through capillary electrophoretic movement. The integrated data analysis pipeline on the instrument will also render the electrophoretic data into a gel-like picture for users more accustomed to traditional gel electrophoresis. The RNA profile of each sample is automatically displayed as individual electropherograms. [6]

2.3 LIBRARY PREPARATION

Before to sequencing, the RNAs in a sample are converted into a cDNA library, representing all the RNA molecules in the sample. This step is performed because in practice, RNA molecules are not directly sequenced, instead DNAs are sequenced due to their better chemical stability, and are also more amenable to the sequencing chemistry and protocols of each sequencing platform. Therefore, the library preparation has two purposes, the first is to adequately represent the RNAs in the sample and secondly to convert RNA into DNA. The major steps in library preparation can be found in J. Pease and R. Sooknanan research [7].

2.4 RNA-SEQ PLATFORMS

2.4.1 ILLUMINA

After libraries are made, ds cDNA is passed through a flow cell which will hybridize the individual molecules based on complementarity with adaptor sequences. Hybridized sequences held at both ends of the adaptor by the flow cell will be amplified as a bridge. These newly generated sequences will hybridize to the flow cell close by and after many cycles a region of the flow cell will contain many copies of the original ds cDNA. This entire process is known as cluster generation. After the clusters are generated, and one strand removed from the ds cDNA, reagents are passed through the flow cell to execute sequencing by synthesis. Sequencing by synthesis describes a reaction where in each synthesis round, the addition of a single nucleotide, which can be A, C, G, or T, as determined by a fluorescent signal, is imaged, so that the location and added nucleotide can be determined, stored, and analyzed. Reconstruction of the sequence of additions in a specific location on the flow cell, which corresponds to a generated ds cDNA cluster, gives the precise nucleotide sequence for a piece of ds cDNA [9].

2.4.2 SOLID

SOLID stands for sequencing by oligonucleotide ligation and detection and is a platform. The sequencing chemistry is via ligation rather than synthesis. In the SOLID platform, a library of DNA fragments (originally derived from RNA molecules) is attached to magnetic beads at one molecule per bead. The DNA on each bead is then amplified in an emulsion so that amplified products remain with the bead. The resulting amplified products are then covalently bound to a glass slide. Using several primers that hybridize to a universal primer, di-base probes with fluorescent labels are competitively ligated to the primer. If the bases in the first and second positions of the di-base probe are complementary to the sequence, then the ligation reaction will occur and the label will provide a signal. Primers are reset five times by a single nucleotide, so at the end of the cycle, at least four nucleotides would have been interrogated twice due to the dinucleotide probes and the fifth nucleotide at least once. The ligation steps continue until the sequence is ready. The unique ligation chemistry allows for two checks of a nucleotide position and thus provides greater sequencing accuracy of up to 99.99%. While this may not be necessary for applications such as differential expression, it is critical for detecting single-nucleotide polymorphisms (SNPs). [8]

2.4.3 ROCHE 454

This platform is also based on adaptor-ligated ds DNA library sequencing by synthesis chemistry. ds DNA is fixed onto beads and amplified in a water–oil emulsion. The beads are then placed into picotiter plates where sequencing reactions take place. The massive numbers of wells in picotiter plates provide the massively parallel layout needed for NGS.

Free nucleotides and unreacted ATP are degraded by a PYRase after each addition. These steps are repeated until a predetermined number of reactions have been reached. Recording the light generation and well location after each nucleotide addition allows for reconstruction of the identity of the nucleotide and the sequence for each well. The advantage of this sequencing chemistry is that it permits for longer reads when compared to other platforms. Read lengths of up to 1000 bases can be achieved on this platform. Roche provides the current GS FLX+ system as well as a smaller GS junior system. With up to 1 million reads per run, and an average of 700nt per read, 700Mb of sequence data can be achieved in less than 1 day of run time. [9].

2.4.4 ION TORRENT

This newer platform utilizes the adaptor-ligated library followed by sequencing-by-synthesis chemistry of other platforms. However, it has a unique feature, instead of detecting fluorescent signals or photons, it detects changes in the pH of the solution in a well when a nucleotide is added and protons are produced. These changes are miniscule; however, the Ion Torrent device utilizes technologies developed in the semiconductor industry to achieve detectors of sufficient sensitivity and scales that are useful for nucleic acid sequencing. One limitation that has been pointed out is that homo-polymers may be difficult to read as there is no way to stop the addition of only one nucleotide if the same nucleotide is next in the sequence. Ion Torrent produces overall fewer reads than the others in a single run. For example, 60–80M reads at 200 bases per read are possible on the proton instrument in a run producing 10Gb of data. However, the run time is only 2–4h instead of 1–2 weeks on other platforms. The machine has a small footprint, can be powered down when not in use and easily brought back to use, and requires minimal maintenance. With the convenience, size, and speed, it has found sizable applications in microbe sequencing, environmental genomics, and clinical applications where time is critical. This platform is also very popular for amplicon sequencing and use of primer panels for amplicon sequencing developed by specific user communities. Its low-cost and small footprint have also made it attractive to laboratories wishing to have their own personal sequencer [10].

2.4.5 PACIFIC BIOSCIENCES

This is a platform representative of the third generation. The chemistry is still similar to second generation sequencing (SGS) as it is a sequencing-by-synthesis system; however, a major difference is that it requires only a single molecule, and reads the added nucleotides in real time. Single-molecule, real-time (SMRT) sequencing developed by Pacific BioSciences offers longer read lengths than the SGS technologies, making it well-suited for unsolved problems in genome, transcriptome, and epigenetics research, particularly assembly and determination of complex genomic regions, gene isoform detection, and methylation detection. [11]

PacBio sequencing captures sequence information during the replication process of the target DNA molecule. The template, called a SMRTbell, is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target double-stranded DNA (dsDNA) molecule[12]. When a sample of SMRTbell is loaded to a chip called a SMRT cell, a SMRTbell diffuses into a sequencing unit called a zero-mode waveguide (ZMW) [13].

SMRT uses zero-mode waveguides (ZMWs) as the basis of their technology. ZMWs are space-restricted chambers that allow guidance of light energy and reagents in the smallest available volume for light detection. In each ZMW, a single polymerase is immobilized at the bottom, which can bind to either hairpin adaptor of the SMRTbell, so a single DNA molecule is sequenced in real time, then start the replication. Four fluorescent labeled nucleotides, which generate distinct emission spectrums, are added to the SMRT cell and can be detected as a nucleotide chain is being synthesized [14].

The replication processes in all ZMWs of a SMRT cell are recorded by a ‘‘movie’’ of light pulses, and the pulses corresponding to each ZMW can be interpreted to be a sequence of bases (called a continuous long read, CLR). Because the SMRTbell forms a closed circle, after the polymerase replicates one strand of the target dsDNA, it can continue incorporating bases of the adapter and then the other strand. If the lifetime of the polymerase is long enough, both strands can be sequenced multiple times (called ‘‘passes’’) in a single CLR. [15]

2.4.6 NANOPORE TECHNOLOGIES

Despite the impressive gains in throughput and low per base cost of current sequencing, efforts continue to improve sequencing technologies. While current nanopore technologies are in development, they so far have had minimal impact on RNA-seq studies. However, their impact in the future may be greater. Nanopore sequencing is a third-generation single-molecule technique where a single enzyme is used to separate a DNA strand and guides it through a protein pore embedded in a membrane. Ions simultaneously pass through the pore to generate an electric current that is measured. The current is sensitive to specific nucleotides passing through the pore, thus A, C, G, or T disturb the current flow differently and produce a signal that is measured in the pore. The advantage of this system is its simplicity leading to small-platform device size (as USB stick-sized device), but the system is technically challenging due to the need to measure very small changes in current at single-molecule scale. The efforts to commercialize this technology are led by Oxford Nanopore, however Illumina also has nanopore sequencing under development. Oxford Nanopore technologies are slated to measure directly RNA, DNA, or protein as it passes through a manufactured pore. Although this technology is not widely available at a commercial level, it shows a lot of promise.[16]

2.5 RNA-SEQ APPLICATIONS

The purposes behind RNA-seq are to identify the sequence, structure, and abundance of RNA molecules in a particular sample. Identifying the structure means the gene structure (i.e., location of promoter, intron–exon junctions, 5′ and 3′ untranslated regions (UTRs), and polyA site). Secondary structure provides the locations of complementary nucleotide that forming stem-loop, or hairpin RNA [17]. Tertiary structure provides the three-dimensional shape of the molecule. However, identifying abundance means, the numerical amounts of each particular sequence both as absolute and normalized values. Sequence can be used to identify known protein-coding genes, novel genes, or long noncoding RNAs. Once sequence has been determined, folding into secondary structures can reveal the class of molecules such as tRNA or miRNA. Comparison of the abundance of reads for each RNA species can be made between samples derived from different developmental stages, body parts, or across closely related species. [2]

In the following is presented the common applications of using RNA-seq data.

2.5.1 PROTEIN CODING GENE STRUCTURE

Earlier transcriptomic methods such as microarray expression analysis, cloning and Sanger sequencing of cDNA libraries, and serial analysis of gene expression (SAGE), as well as computational prediction from genomic sequences, have already provides gene structures. These structure annotations have been archived in databases and provide an easily accessible source for comparing raw RNA-seq data with known protein coding genes. The first important step is to map the RNA-seq reads to known protein-coding genes.

Furthermore, RNA-seq data analysis can be used for confirming exon–intron boundaries, as well as the existence of completely novel exons. Therefore, using RNA-seq can define what is called a gene model, which is a collection of exons and introns that make up a gene. Since RNA-seq is quantitative, it can also specify within a sample the alternative exons usage: for example, when a specific exon is used five times more often than another one. The 5′ transcription start site (TSS) can be identified precisely using RNA-seq data. Similarly, at the 3′ end of the molecule, the 3′UTR can be identified as well, such that the site of polyadenylation can be observed in the RNA-seq reads. Alternative polyadenylation sites can also be observed in the same way as alternative TSS as well as their respective abundances. As RNA-seq is massively parallel, sufficient reads will permit these gene structures and their alternatives to be mapped for presumably every protein-coding gene in a genome. Thus, RNA-seq can provide the 5′TSS, 5′UTR, exon–intron boundaries, 3′UTR, polyadenylation site, and alternative usage of any of these if applicable[18]. A simplified scheme of gene structure illustrated in Figure 2.1.

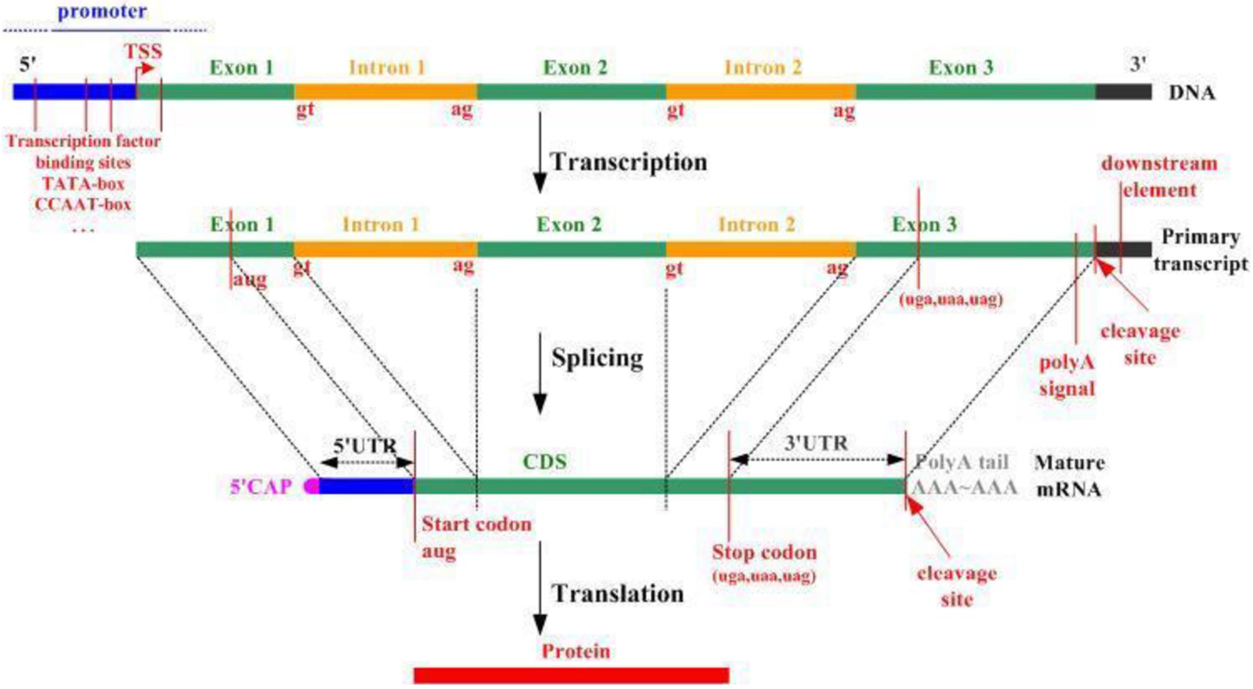

Figure 2.1 Schematic gene structure and simplified transcription, splicing and translation process[19]

2.5.2 NOVEL PROTEIN-CODING GENES

Previously, the annotations of protein-coding genes relied on computational predictions based on genomic sequences. This was fine as long as genome data were available, the gene model elements fit common expected size and distance parameters, and there were transcriptomic data in the form of expressed sequence tag (EST) data sets or orthology data available to verify the predictions. However, it was easy to see that these criteria fit well only a very limited number of organisms under scientific investigation. Therefore, RNA-seq, with its high throughput, could verify many of the previous predictions, but also in cases where no prediction existed, it could identify novel protein-coding genes. It was especially useful in cases where no genome sequence was available, so a transcriptome of an organism could be built entirely from RNA-seq data. A recent example of this application has been in the sequencing of the giant panda genome [20].

2.5.3 QUANTIFYING AND COMPARING GENE EXPRESSION

Once the sequence and gene structure have been elucidated, it is logical that abundance values can be attributed to each gene as well as various features in their structures. As many studies would like to compare the abundance of RNA transcripts from healthy versus sick, nontreated versus treated, or time point 0 versus 1, it is logical that comparative studies are made. The range and types of comparative studies are virtually unlimited. In one of the earliest RNA-seq studies, transcripts from adult mouse brain, liver, and skeletal muscle were sequenced and compared [18].

More than 40M single-end reads at 25nt were sequenced on an Illumina platform and the authors found novel TSSs, alternative exons, and alternative 3′UTRs. The study demonstrated the shallowness of previous annotations of gene structure and thus highlighted how the breadth and depth of annotations provided by RNA-seq technology could change our view of gene structure. These results thus paved the way for subsequent RNA-seq studies. Another RNA-seq study two years later, followed the expression of RNA transcripts from mouse skeletal muscle cells during differentiation after 60 h, 5 days, or 7 days [21]. The technology improved so that more than 430M paired-end reads at 75nt were used to identify greater than 3700 previously unannotated transcripts. TSSs were also shown to change in more than 300 genes during differentiation. It is also possible to study RNA transcripts in whole animals. The total RNA from whole animals could be isolated and subjected to RNA-seq, in recent study Over 30M reads from water- or ethanol-treated animals were obtained [22]. Ethanol exposure could be seen to increase RNA transcripts of detoxification enzyme genes and decrease transcripts involved in endoplasmic reticulum stress.

2.5.4 EXPRESSION QUANTITATIVE TRAIT LOCI (EQTL)

RNA-seq studies have become so pervasive that they have been used to study quantitative traits, especially in the context of genome variation One of the most prominent directions One of the most prominent directions One of the most prominent directions has been the extensive set of studies on expression quantitative trait loci (eQTLs), namely, the discovery of genetic variants that explain variation in gene expression. Such studies have offered promise not just for the characterization of functional sequence variation but also for the understanding of basic processes of gene regulation and interpretation of genome-wide association studies. An eQTL is a locus that explains a fraction of the genetic variance of a gene expression phenotype [23].

2.5.5 SINGLE-CELL RNA-SEQ

RNA-seq is a variation of RNA-seq where the source of total RNA for sequencing comes from a single cell. Typically, total RNA is not isolated, but rather cells are individually harvested from their source and reverse-transcribed. Methodology for library preparation is similar to RNA-seq: RNA is reverse-transcribed to cDNA, adaptors are ligated, barcodes for each cell are added, and ds cDNA amplified. Due to the low complexity of RNA species, single isolated cells or individual libraries are sometimes pooled prior to sequencing. In one example of this approach, a single mouse blastomere was collected and RNA-sequenced from its contents. The authors found 5000 genes expressed and >1700 novel alternative splice junctions, indicating both the robustness of the approach as well as the complexity of splicing in a single cell [24]. In another example of the approach, single cells from the nematode C. elegans at an early multicell developmental stage were isolated and libraries prepared from total RNAs. New transcription of genes could be monitored at each individual stage of development via profiling the transcripts of individual cells [25].

2.5.6 FUSION GENES

As read numbers and lengths increased, and paired-end sequencing became available, the ability to identify rare, but potentially important transcripts increased. Such is the case with fusion genes, which are transcripts generated from the fusion of two previously separate gene structures. Fusion partners can contribute 5′UTRs, coding regions, and 3′polyadenylation signals. Conditions for this event to occur happen during genomic rearrangement found in cancer tissues and cells. Cytogenetic derangements such as genomic amplifications, translocations, and deletions can bring together two independent gene structures. For example, 24 novel and three known fusion genes were detected in three breast cancer cell lines using paired-end sequencing of libraries sized 100 or 200nt in length [26]. One of these fusion genes, *VAPB-IKZF3*, was found to be functional in cell growth assays[27]. Recent RNA-seq studies have found fusion genes to be present in normal tissue, suggesting that fusion gene events might have normal biological function as well[28].

2.5.7 GENE VARIATIONS

As the amount of RNA-seq data accumulates, it is possible to mine the data for gene variation. Mostly bioinformatic approaches by downloading publicly available data have been used to scan SNPs in transcriptomic data [29]. In this study, 89% of SNPs derived from RNA-seq data at a coverage of 10× were found to be true variants. SNP detection can also be obtained directly from original RNA-seq data. A group performed RNA-seq on muscle from *Longissimus thoraci* (Limousine cattle) muscle mRNAs [30]. They were able to identify >8000 high-quality SNPs from >30M paired-end reads. A subset of these SNPs was used to genotype nine major cattle breeds used in France, demonstrating the utility of this approach.

2.5.8 LONG NONCODING RNAS

Another application of RNA-seq has been to find transcripts that are present, but do not code genes. Long noncoding RNAs (lncRNAs) were known before RNA-seq technologies were available. However, the extent of their existence and pervasiveness was not fully appreciated until RNA-seq methods were able to uncover the many different species of lncRNAs in living cells. lncRNAs are generally described as transcripts that fall outside of known noncoding RNAs such as tRNAs, ribosomal RNAs, and small RNAs, do not overlap a protein-coding exon, and are >200nt in length [31]. lncRNAs can control transcription as enhancers (eRNA) epigenetically by binding and altering the function of histone proteins, as competitors to RNA-processing machinery [competitive endogenous RNA (ceRNA)], or as noise generated randomly. It can now be appreciated that lncRNAs may play a role in disease such as Alzheimer's disease [32].

2.5.9 SMALL NONCODING RNAS (MIRNA-SEQ)

RNA-seq can be used to identify the sequence, structure, function, and abundance of small noncoding RNAs. The most well-known example of these being miRNAs (miRNA-seq), but other small noncoding RNAs such as small nucleolar RNAs (snRNA), microRNA offset RNAs (moRNAs), and endogenous silencing RNAs (endo-siRNAs) can also be studied using miRNA-seq approaches. The methods used for miRNA-seq are similar to RNA-seq. The starting materials can be total RNA or size-selected/fractionated small RNAs. Most of the common sequencing platforms will sequence small RNAs once converted into ds cDNAs, such that much of the difference in the experimental protocols occur before sequencing. [33]. There are many applications for characterizing these molecules not only in the studies of basic biochemistry, physiology, genetics, and evolutionary biology, but also in medicine as a diagnostic tool for cancer or in aging processes. A recent study of the nematode Panagrellus redivivus has presented the identification of >200 novel miRNAs and their precursor hairpin sequences while also providing gene structure models, annotation of the protein-coding genes, and the genomic sequences in a single publication [34].

2.5.10 AMPLIFICATION PRODUCT SEQUENCING (AMPLI-SEQ)

It is sometimes the case that whole transcriptomes do not need to be sequenced, but only a small number of genes. While one can always obtain a subset of genes of interest from a whole transcriptome sequence analysis, the effort, time, and resources required may be more than necessary. By using a panel of PCR primers consisting of 10–200 pairs, one can perform reverse transcription-PCR (RT-PCR) and instead of cloning each individual product and isolating plasmid DNA for Sanger sequencing, one can sequence the pool of PCR products to obtain the sequence. This has practical applications where the number of samples to be interrogated is large, and the number of genes is small [35].

2.6 OSTEOBLAST CELLS DIFFERENTIATION

Skeletal component cells including osteoblasts, chondrocytes, adipocytes, myoblasts, tendon cells, and fibroblasts, are derived from mesenchymal stem cells [36]. while Osteoclast is a hematopoietic cell derived from CFU-GM (colony forming unit- granulocyte, monocyte) and branches from the monocyte-macrophage lineage early during the differentiation process [37].

Bone is constructed through 3 processes: osteogenesis, modeling, and remodeling. It is constantly being remodeled in a dynamic process where osteoblasts are responsible for bone formation (or ossification), and osteoclasts for its resorption. Osteoblast and osteoclast work in tight cooperation, and together constituting a "bone multicellular unit"[38]. Fine tuning of this system is crucial for the development of bones, for repairing fractures, and for the correct maintenance of the skeleton throughout life.

Osteoblast differentiation can be characterized in three stages[40]:

a) Cell proliferation
b) Matrix maturation
c) Matrix mineralization

In Stage 1 the cells continue to proliferate and express fibronectin, collagen, TGFb receptor 1, and osteopontin.In Stage 2 they exit the cell cycle and start differentiating, while maturating the extracellular matrix with Alp and collagen.In Stage 3 matrix mineralization occurs when the organic scaffold is enriched with osteocalcin, which promotes deposition of mineral substance. Osteocalcin is in fact the second most abundant protein in bone after collagen [41]. At this stage the osteoblast assumes its characteristic cuboidal shape [42].

2.7 ALTERNATIVE SPILCING

Genetic information of an organism is stored in the genes, this information is transcribed from DNA into a messenger RNA (mRNA) template by a process called transcription. However, in eukaryotes, before the mRNA can be translated into proteins, non-coding portions of the sequence, called introns, must be removed and protein-coding parts, called exons, joined by RNA splicing to produce a mature mRNA. Recent estimates indicate that the expression of nearly 95% of human multi-exon genes involves alternative splicing.[43]

Alternative splicing of precursor mRNA is an essential mechanism for gene regulation and for generating proteomic diversity, it produces different protein products that function in diverse cellular processes, including cell growth, differentiation, and organism development. Furthermore, it has a largely hidden function in quantitative gene control, by targeting RNAs for nonsense-mediated decay. In the Figure 2.2 illustrate the general concept of alternative splicing.

Regulation of alternative splicing is a complicated process in which numerous interacting components are involved. Additional molecular features, such as chromatin structure, RNA structure and alternative transcription initiation or alternative transcription termination, collaborate with these basic components to generate the protein diversity due to alternative splicing.

Splicing is carried out by the spliceosome, a massive structure in which five small nuclear ribonucleoprotein particles (snRNPs) (U1, U2, U4, U5 and U6), that are associated with a large number of auxiliary proteins cooperate to accurately recognize the splice sites and catalyze the steps of the splicing reaction. The auxiliary elements known as Exon Splicing Enhancers (*ESEs*), and Intron Splicing Enhancers (*ISEs*), in addition to Exon Splicing Silencers (*ESSs*), and Intron Splicing Silencers (*ISSs*) [44]. These auxiliary elements are involved in defining both constitutive and alternative exons.

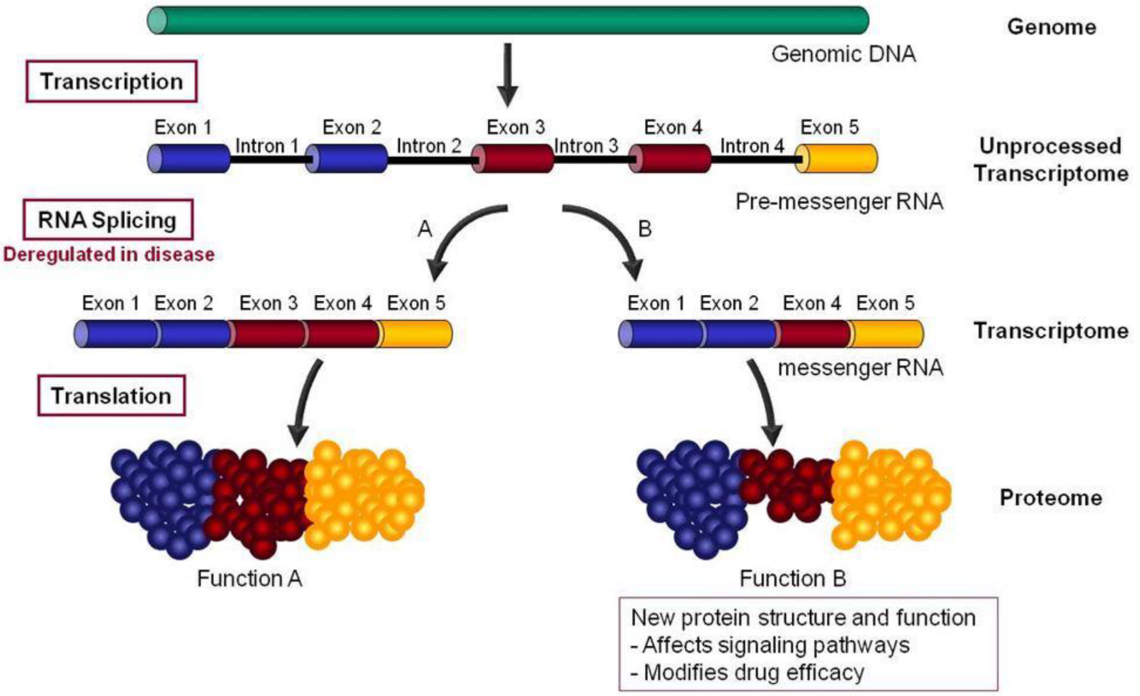

Figure 2.2 Genetic scheme of general concept of RNA transcription, splicing and translation[45].

2.8 QUALITY CONTROL AND PREPROCESSING

High throughput sequencers can generate tens of millions of sequences in each run. Before analyzing this RNA-seq data and using it for transcriptome study to draw biological conclusions, quality control must be performed to ensure that RNA-seq data are of high quality and suitable for subsequent analyses without biases. Quality problems typically originate either in the sequencing itself or in the preceding library preparation. They include low-confidence bases, sequence-specific bias, 3′/5′ positional bias, polymerase chain reaction (PCR) artifacts, untrimmed adapters, and sequence contamination. These problems can seriously affect mapping to reference, assembly, and expression estimates. Many of those defects can be corrected for by filtering, trimming, error correction, or bias correction. While some cannot be corrected for, but they must be taken into consideration when interpreting results.

During my research I dedicated considerable time for quality control of FASTQ files[46]. Either by using the available Quality Control (QC) tools as FastQC[47], RSeQC[48] and Trimmomatic[49]. Or by coding a number of tools using HTSeq library in Python [50], to check the eligibility of RNAseq data, as described in the following.

2.8.1 FASTQ FORMAT:

FASTQ has emerged as a common file format for sharing sequencing read data combining both the sequence and an associated per base quality score. It provides an additional extension to the FASTA format, it is the ability to store a numeric quality score associated with each nucleotide in a sequence. However, it is lacking the clear formal definition. Furthermore, there are three incompatible variants of FASTQ format; original Sanger standard, the Solexa, and Illumina variants. Over time, the FASTA format has developed by consensus; however, in the absence of an obvious standard. For example, some parsers will fail to handle the very long '>' title lines or very long sequences without line wrapping. There is also no standardization for record identifier [46].

Although Illumina initially continued to use the Solexa FASTQ variant, from Genome Analyzer Pipeline version 1.3 onwards, PHRED quality scores rather than Solexa scores were used [51]. The Illumina 1.3+ FASTQ variant encodes PHRED scores with an ASCII offset of 64, and so can hold PHRED scores from 0 to 62 (ASCII 64–126), although currently raw Illumina data quality scores are only expected in the range 0–40. (Table 2.1)

FASTQ variant	ASCII Characters		Quality Score	
	Range	Offset	Type	Range
Sanger standard *'fastq-sanger'*	33 - 126	33	PHRED	0 to 93
Solexa/early Illumina 'fastq-solexa'	59 - 126	64	Solexa	-5 to 62
Illumina 1.3+ 'fastq-illumina'	64 - 126	64	PHRED	0 to 62

Table 2.1 FASTQ variants between different sequencing platforms

2.8.2 QUALITY ASSESSMENT BY FASTQC

FastQC [47] generates QC report contains 12 analysis modules as follows:

1. Basic Statistics module: it summarizes statistical information about the sequencing reads.
2. Per Base Sequence Quality module: This module generates a plot, shows an overview of the range of quality values across all bases at each position in the FastQ file.
3. Per Sequence Quality Scores module: it allows to see if a subset of the sequences have universally low quality values.
4. Per Base Sequence Content module: it plots out the proportion of each base position in the sequencing reads for each of the four DNA bases has been called.
5. Per Sequence GC Content module: it measures the GC content distribution across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content. An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position.
6. Per Base N Content module: If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. This module plots out the percentage of base calls at each position for which an N was called.
7. Sequence Length Distribution module: Some high throughput sequencers generate sequence fragments of uniform length as in our raw RNA-Seq data the length of the reads is 101.

8. Duplicate Sequences module: In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias
9. Overrepresented Sequences module: A normal high-throughput library contains a diverse set of sequences. Finding that a single sequence is very overrepresented in the set, either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as it is expected.
10. Adapter Content module: The Kmer Content module will do a generic analysis of all the Kmers in the library to find those which do not have even coverage through the length of reads.
11. *Kmer Content module:* The analysis of overrepresented sequences will point an increase in any precisely duplicated sequences. However, this analysis suffers from few problems which might fail:
    - If we have very long sequences with poor quality, then random sequencing errors will dramatically reduce the counts for exactly duplicated sequences.
    - If we have a partial sequence which is appearing at a variety of places within our sequence, then this won't be seen either by the per base content plot or the duplicate sequence analysis.
12. *Per Tile Sequence Quality module:* This graph is available only if we use an Illumina library (which is the case of our data) that retains its original sequence identifiers. Encoded in these is the flow-cell tile from which each read came. The graph shows the quality scores from each tile across all the bases to see if there was a loss in quality associated with only one part of the flow-cell.

2.8.3 TRIMMING LOW QUALITY READS:

The raw data of the next generation sequencing usually suffered, beside the attached adapters which must be removed, from low quality sequencing bases along the reads, that can easily result in suboptimal downstream analyses. Nevertheless, it is considerable to precisely identify such sequences, including partial adapter sequences, while leaving valid sequence data pristine [52]. Trimmomatic is the optimal choice designed to work on NGS data for identification of adapter sequences and quality filtering. It is able to process paired-end samples and optimized for Illumina NGS data [53].

The trimming procedures that Trimmomatics performed classified in the following list:

2.8.3.1 Removing technical sequences:

Identifying adapter or other contaminant sequences within a dataset is inherently a tradeoff between sensitivity (ensuring all contaminant sequences are removed) and specificity (leaving all non-contaminant sequence data intact). This issue is even more critical when only a small part of the contaminant sequence is included within the read. [53]

2.8.3.2 Quality filtering:

Trimmomatic offers two main quality filtering alternatives; sliding window and maximum information. The Sliding Window uses a relatively standard approach. This works by scanning from the 5′ end of the read, and removes the 3′ end of the read when the average quality of a group of bases drops below a specified threshold. This prevents a single weak base causing the removal of subsequent high-quality data, while still ensuring that a consecutive series of poor-quality bases will trigger trimming.

The following equation gives a length threshold score:

$$Score_{LT}(l) = \frac{1}{(1 + e^{t-l})}$$

Whereas: *t* is a target length, and *l* is the putative length after trimming.

The second factor models ‘coverage’, it provides a linear score based on retained sequence length:

$$Score_{cov}(l) = l$$

The third factor models the ‘error rate’, it uses the error probabilities from the read quality scores to determine the accumulated likelihood of errors over the read. To calculate this score, we simply take the product of the probabilities that each base is correct, giving:

$$Score_{Err} = \prod_{i=l}^{l} P_{corr}[i]$$

The correctness probabilities P_{corr} of each base are calculated from the sequence quality scores. The error score typically begins as a high score at the start of the read, depending on the read quality, typically drops rapidly at some point during the read.

The Maximum Information algorithm determines the combined score of the three factors for each possible trimming position, and the best combined score determines how much of the read to trim. A strictness parameter s can be set between 0 and 1, controls the balance between the ‘coverage’ factor (for s = 0) and the ‘error rate’ factor (for s =1). This gives the following formula:

$$Score(l) = \frac{1}{(1+e^{t-l})} \cdot l^{(1-s)} \cdot \left(\prod_{i=l}^{l} P_{corr}[i] \right)^{s}$$

2.9 SEQUENCE ALIGNMENT

The purpose of sequence alignment is to figure out where the sequences are similar and how high the similarity is. Aligning or “mapping” reads to a reference genome or transcriptome allows us to estimate where the read originated from. Mapping reads to genome provides genomic location information, which can be used for discovering new genes and transcripts, and for quantifying expression. If a reference genome is not available, or if our target is to quantify only known transcripts, reads can be mapped to a transcriptome instead.

Aligning reads to a reference genome is a challenging task for many reasons; reads are relatively short and there are millions of them, while genomes can be large and contain ambiguous sequence regions or indistinct such as repetitive regions and pseudogenes, this can impact the mapping to these areas. Furthermore, aligners have to cope with mismatches and indels (insertions-deletions) caused by genomic variation and sequencing errors. Eventually, many organisms have introns in their genes, so RNA-seq reads align to genome non-contiguously. Placing spliced reads across introns and determining exon–intron boundaries correctly is difficult, because sequence signals at splice sites are limited and introns can be thousands of bases long.

2.10 GENE EXPRESSION ANALYSIS

Once reads have been mapped to a reference genome, their mapping locations can be identified by genomic annotation. This enables us to quantitate gene expression by counting reads per genes, transcripts and exons. Quantitation of gene expression is an integral part of most RNA-Seq studies. In principle, calculating the count of mapped reads provides a direct way to estimate transcript abundance, it has been found that read count is approximately linearly related to the abundance of the target transcript[18], but in practice several complications need to be taken into account. Eukaryotic genes typically produce several transcript isoforms via alternative splicing and promoter usage. However, quantitation at transcript level is not easy with short reads, because transcript isoforms often have common or overlapping exons. Furthermore, the coverage along transcripts is not uniform because of mappability issues and biases introduced in library preparation. Because of these complications, expression is often estimated at the gene level or the exon level instead. However, gene level counts are not optimal for differential expression analysis for those genes which undergo isoform switching, because the number of counts depends on transcript length. This challenge can be overcome by applying the appropriate reads count normalization in addition to an effective statistical model for significant variability estimation, where A number of model-based methods have been developed that attempt to deconvolve the expression levels of individual transcripts for each gene from RNA-seq data, essentially by leveraging information from reads unambiguously assigned to regions where isoforms differ as RSEM [54], and cuffdiff from cufflinks package [55].

Differential expression analysis of RNA-seq data differs from microarray. In RNA-Seq the observed data are in the form of discrete counts generated from a sampling process, while microarray measurements are continuous measurements of a fluorescence signal.

3. AIMS OF THE DOCTORAL THESIS

In recent years several pipelines were founded for high throughputs sequence data analysis, many tools are available for quality control of RNA-Seq data, reads mapping, comparative analysis of gene expression and alternative exons usage, and for finding de novo long non-coding RNA. However, most of the analysis fail to deal the integrity, large datasets, and to produce descriptive results, that biologists can interpret without addition effort.

The main aim of this doctoral work is to introduce an integrated RNA-Seq data analysis pipeline, that produces illustrated outputs, especially in the transcriptomic characterization experiment. This type of experiment is based on an expanded investigation of a comparative gene expression between different biological conditions, where we have a numerous amount of outputs needed to be examined to get the significant and informative results. Based on those outputs, we can build hypothesis describes the gene regulation stands behind that biological mechanism. The main aims we achieved in this doctoral thesis can be listed as follows:

1- Introducing several codes for RNA-Seq data quality control, which provide tables of summarized reads statistics in samples, and give the mean of Phred quality scores across all the bases in a sample. In addition to plotting the mean quality of each base in all samples, we established a method to check the coverage uniformity. Especially when polyadenylated RNA library is used, there is usually concern that the coverage might vary across the gene's features.

2- Establishing a comprehensive framework for differential gene expression analysis, that produces descriptive outputs and facilitates the biological interpretation of the experiment.

3- Presenting an analyzing approach for multiple conditions experiment, we called it "ON/OFF genes", which can define the silent genes in a particular condition of the comparative analysis. This approach can highlight the functional roles, that genes can play in different conditions, and give a wider view of the genetical reasons behind the distinction between biological conditions.

4- Improving the performance Improving the performance of Bioconductor package DEXSeq [56]for differential exons usage, by specifying if the differentially used exonic part is within the ORF. This procedure will help to figure out if the differential used exons are involved in the alternative pathways, or distinctive functions of a gene's transcripts.

5- Suggesting a new approach to analyze differential expressed long non-coding RNA, by finding the functional correlation of lncRNA with neighboring differential expressed protein coding genes within the TAD (Topological associated Domain), to obtain an illustrated view concerning the regulation mechanism in a dataset.

4. RNA-SEQ DATA ANALYSIS WORKFLOW

4.1 OSTEOBLAST DIFFERENTIATION EXPERIMENT

The RNA-Seq data which our research based on, are from differentiated osteoblast cells. Although osteoblast differentiation was well characterized, a detailed transcriptional analysis of osteoblast differentiation based on RNA sequencing (RNA-seq) analyses is still missing. Therefore, we used RNA-seq to obtain a high-resolution transcriptome data set of murine osteoblast differentiation in vitro. The cells were harvested at four distinct time points: within proliferation, during maturation, terminal differentiation, and at the onset of mineralization.

As a consequence, we got 12 samples, 3 for each differentiation time point:

1- Day 0: The confluency of cells in the culture plate, before promoting the differentiation.

2- Day 3: Harvesting cells on the third day after the differentiation starts.

3- Day 6: Harvesting on the sixth day after the differentiation promoted.

4- Day 12: On the twelfth day of differentiation.

4.2 QUALITY CONTROL AND REPROCESSING METHODS

The first procedure needed to be implemented after we have the RNA-Seq data as FASTQ files, is checking the quality of the raw read sequencing. Once reads have been aligned to a reference genome, additional quality metrics can be investigated based on the location. These include coverage uniformity along transcripts, saturation of sequencing depth, ribosomal RNA content, and read distribution between exons, introns, and intergenic regions. Finally, once aligned reads have been counted per genes, sample relations and batch effects can be visualized with heatmaps and PCA plots.

4.2.1 FASTQC

FastQC is available as a standalone interactive Java application with a graphical user interface (GUI), and it can be run as well in a command line as non-interactive mode, where it would be suitable for integrating into a larger analysis pipeline for the systematic processing of large numbers of files.

Using the Bash Script, we coded a function to input the samples of original raw reads to FastQC to check the read qualities. This function can serve large experiment to pass samples to FastQC and get outputs by one command. User does not to care about how to input a set of samples, neither getting the outputs. As described in (Main dissertation: Code-box 3.1).

4.2.2 TRIMMOMATIC:

For the pair-end read, Trimmomatic requires as inputs both the reverse and forward reads and returns 4 outputs, 2 for the 'paired' output where both reads survived the processing, and 2 for corresponding 'unpaired' output where a read survived, but the partner read did not. In the end, we used the paired reads. Trimming process must be performed in an order of steps then the optional procedures can be added in the end of the command. It is recommended in most cases that adapter clipping is done as early as possible, since correctly identifying adapters using partial matches is more difficult. [49]

To perform the trimming on a set of samples and keep the paired samples, we coded two functions in Bash script for this purpose. One to insert the samples to Trimmomatic (Main dissertation: Code-box 3.2), and the second to keep merely the paired samples, which is used for mapping to the reference (Main dissertation: Code-box 3.3). Reading through the codes in the Cod-boxes explain the simple concept used to get a function with one line to input a set of samples. Despite the simplicity of our functions, they provide useful service for users with low or no knowledge with shell command-line, which is necessary to run this part of RNA-seq pipeline.

4.2.3 PYTHON _ HTSEQ

HTSeq is a Python library. It offers parsers for many common data formats in High-Throughput Sequencing (HTS) projects, as well as classes to represent data, such as genomic coordinates, sequences, sequencing reads, alignments, gene model information and variant calls. It also provides data structures that allow for querying via genomic coordinates. [50]

Python as a scriptural language is useful to abstract information from output reports as text files. we wrote two scripts for this purpose; the first one to get the basic statistic of Fastq files as the length of reads, sequenced reads number, and %GC. The second script to get the mean reads quality in each sample from FastQC report. Using HTSeq library we coded a script to plot the mean quality of the reads across the position. For checking coverage uniformity across the gene body in Poly-A libraries, we proposed a method based on HTSeq to get the coverage distribution in different gene features.

4.2.3.1 BasicStatistic Funnction:

Each FactQC output has in additional to html report, a plain text file called 'fastqc_data.txt'. In Osteoblast data set we have four differentiation time points (conditions) with three biological replicates for each condition, and each sample has two fastq files for forward and reverse reads, this means 24 FastQC reports in total. To extract the required information, we coded a useful function to read the data from such group of files. This function reads lines in a specific module (Basic Statistics module) in fastqc.data report, and then extract the required information about the reads, as sequence length, total sequenced numbers, encoding type, Sequences flagged as poor quality, and present of GC in the sequence. (Main dissertation: Code-box 3.4)

4.2.3.2 Mean Quality Function:

The function calculates the mean sequencing quality across all the bases in a sample. In FastQC report, there is module called "Per sequence quality scores", which gives how many reads have a specific quality value (from 2 to 40). I coded two functions for this task, the main mean quality function "QualityScore", to calculate the mean quality of all bases in a fastq (Main dissertation: Code-box 3.5). And "MQ_Av_Dataset" function, to get the mean quality of all bases in fastq files in a dataset, then calculate the average of mean quality of forward and reverse reads, eventually write the output as plaintext file (Main dissertation: Code-box 3.6).

4.2.3.3 Plotting Phred Quality Along Reads Positions:

Using the FastqReader function from Python_HTSeq package, generates objects of class FastqReader from the Fastq files. In this object each read in Fastq is a SequenceWithQualities object, and has three feature slots:

- name feature returns the name of the read (read.name)
- seq feature returns the sequence of the read (read.seq)
- qual feature returns the quality of each base in the read as an array of values (read.qual)

The function "read_qual" iterates over reads in FastqReader object and calculates the mean quality of all reads in each position coordinate, as described in the comments (after # symbol, or between *''' '''*) in the (Main dissertation: Code box 3.7). To accomplish this task, I coded another function "*pass_data_MQ"* to apply read_qual through a dataset (Main dissertation: Code box 3.8). All the codes are reproducible and helpful to deal with quality control of fastq files. We generated 70 million reads with good high-quality scores as Figure 4.1 illustrates the mean quality of raw reads. However, after we implied quality control enhancement techniques, we got a very good quality score along the positions in the reads, as it is shown in the Figure 4.2, after trimming the mean quality of the bases at the end of the reads improved to be over 28 scores, and across the middle of the reads, looks more linear with Phred scores between 33-38.

4.2.3.4 Coverage Uniformity:

There is a concern that the coverage might vary across features of the RNA, since the used libraries are based on polyadenylated RNA. To provide an estimation of coverage across the genes feature, we separated the exons to 5' UTR exons, 3' UTR exon and the exons in the translated regions in between, as illustrated in the Python script "Separate Exons according to the translation regions" (Main dissertation: Code-box 3.9). Hence, we got 3 annotation GTF files of first (5' UTR), last (3' UTR) and middle exons in the genes. Then we coded a function to calculate the coverage of the reads using HTSeq library. Accomplishing this task was in the following steps:

1. Create GTF objects of the first exons, last exons and the middle exons from the gtf files generated in the previous step, by using GFF_Reader method from HTSeq library. (Main dissertation: Code-box 3.11)
2. Create alignment objects from BAM files, the outputs of TopHat Aligner, by using BAM_Reader method from HTSeq library. The BAM_Reader object yields for each alignment line in the BAM file an object of class BAM_Alignment. Every alignment object has a slot read, that contains a SequenceWithQualities object as described previously. (The read object has three features: read.name, reads.seq and read.qual). Furthermore, every alignment object aln has a slot iv (for "interval") that describes the positions on the genome where the read was aligned to (if it was aligned). This feature slot of the alignment object holds information of the start and end coordinates on the genome of that align object, the chromosome name, and the strand where that alignment located. (Main dissertation: Code-box 3.10)
3. Creating a GenomicArray data structure to store and retrieve information associated with a genomic position or genomic interval. The key of the GenomicArray is an genomic interval, which we retrieve from "align.iv" of each alignment object and simply iterate through all the reads and add the value 1 at the interval when each read was aligned.

4. To calculate the mean coverage of each aligned read, we passed the GenomicArray "ga" of the interval of each exon in the aligned read "iv" to list method, so we got a coverage values vector, then we calculated the mean of this vector.
5. Eventually we plotted the coverage density of 5' UTR, 3' UTR and exons within coding regions. As it is shown in the Figure 4.3.

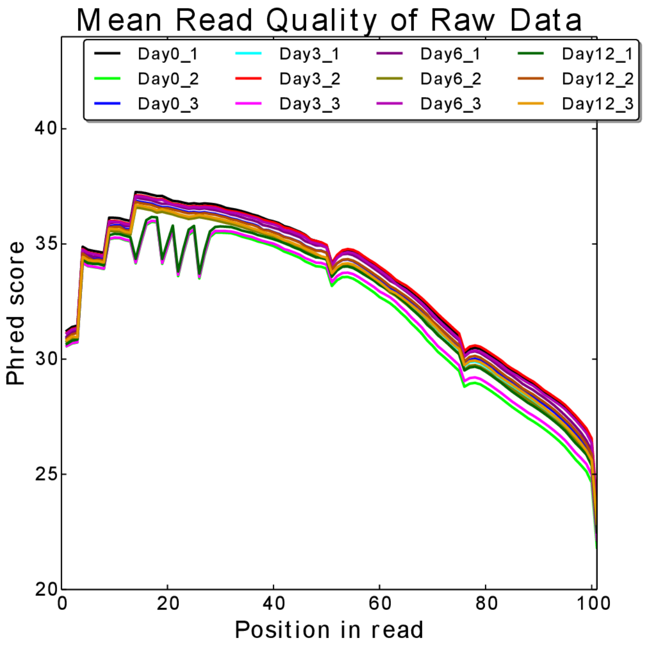

Figure 4.1 :Mean Phred Quality Scores of Raw RNA-Seq Reads.

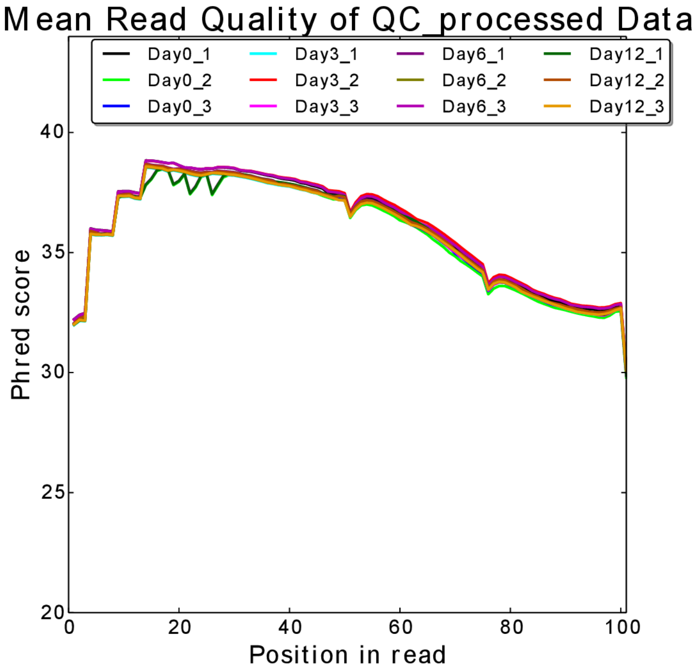

Figure 4.2: Mean Phred Quality Scores of RNA-Seq Reads after QC.

From the density distribution of the base mean which is the sequencing depth of each exon in Osteoblast dataset (Figure 4.3), we found that both 3′ UTR exons and the intervening coding sequences have a similar distribution of base mean values compared to 5′ UTR sequences. Although this slight change in the distribution across features, the level of coverage was similar for each, suggesting that the libraries capture transcriptional complexity across different elements of genes.

4.3 READS MAPPING USING TOPHAT ALIGNER

TopHat is an optimal fast junction aligner for RNA-Seq reads to mammalian-sized genomes, in order to identify exon-exon splice junctions. It is built on the ultra-high-throughput short read mapping program Bowtie, which is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches [57].

TopHat2 improved the performance of TopHat by mapping reads against the known transcriptome, this improves the overall sensitivity and accuracy of the mapping. The mapping procedure of TopHat2 consists of three major parts, optional transcriptome alignment, genome alignment, and spliced alignment. Paired-end reads are aligned individually first, and then combined to paired-end alignments by taking into account the fragment length and orientation.

To perform alignment by TopHat 2, first we need to prepare the reference genome index. Bowtie2 reference genome indexes are available for many organisms at the Bowtie2 website [58] and the Illumina iGenomes website [59]. However, it is better to build the reference index to be sure that genome index/FASTA files are from the same provider as GTF files and the same version. It is easy to build the index using bowtie2-build command.

When calling bowtie2-build command, we need to take into account the following notes: call the option -f to indicate that the reference is FASTA file with ".fa" extension (or .mfa , .fna), and the fasta file must be unzipped (Main dissertation : Code-box 3.12). TopHat2 accepts both FASTQ and FASTA files as input. Read files can be compressed (.gz). TopHat2 can also combine single-end reads in a paired-end alignment if needed [60]. Our Osteoblast samples are paired-end reads, so we used TopHat syntax of paired-end reads as stated in (Main dissertation : Code-box 3.13). The script is coding a function for calling TopHat2 on a dataset without concerning about inserting the correct name of each single sample.

4.4 DIFFERENTIAL GENE EXPRESSION ANALYSIS

4.4.1 METHOD

The fundamental process in RNA-Seq data analysis and transcriptome characterization, is to define a set of genes that have significant expression variance between conditions in an experiment. The comparative analysis of transcriptomic data in our research based on DESeq2 method [61], that takes a count matrix as an initial input. The count matrix composed of *n* rows; one row for each gene *I,* and *m* columns; one column for each sample j. The matrix elements K_{ij} indicate the number of sequencing reads that have been unambiguously mapped to a gene in a sample. The read count K_{ij} for gene *i* in sample *j* is modeled with a generalized linear model (GLM) [62] of the negative binomial family with a logarithmic link. Read counts K_{ij} follow a negative binomial distribution (a gamma-Poisson distribution) with mean μ_{ij}, the variance σ_{ij}^2, and dispersion α_i.

The mean parameter μ_{ij} is the expectation value of the observed counts for gene *i* in sample *j,* it is the product of a quantity q_{ij}, proportional to the concentration of cDNA fragments from the gene in the sample, and a normalization factor s_{ij}:

$$\mu_{ij} = s_{ij}.q_{ij}$$

The GLM fit returns coefficients indicating the overall expression strength of the gene and the log2 fold change between the conditions. DESeq2 uses GLMs with a logarithmic link:

$$\log_2 q_{ij} = \sum_r x_{jr}.\beta_{ir}$$

Where x_{jr} are design matrix elements, and β_{ir} coefficients.

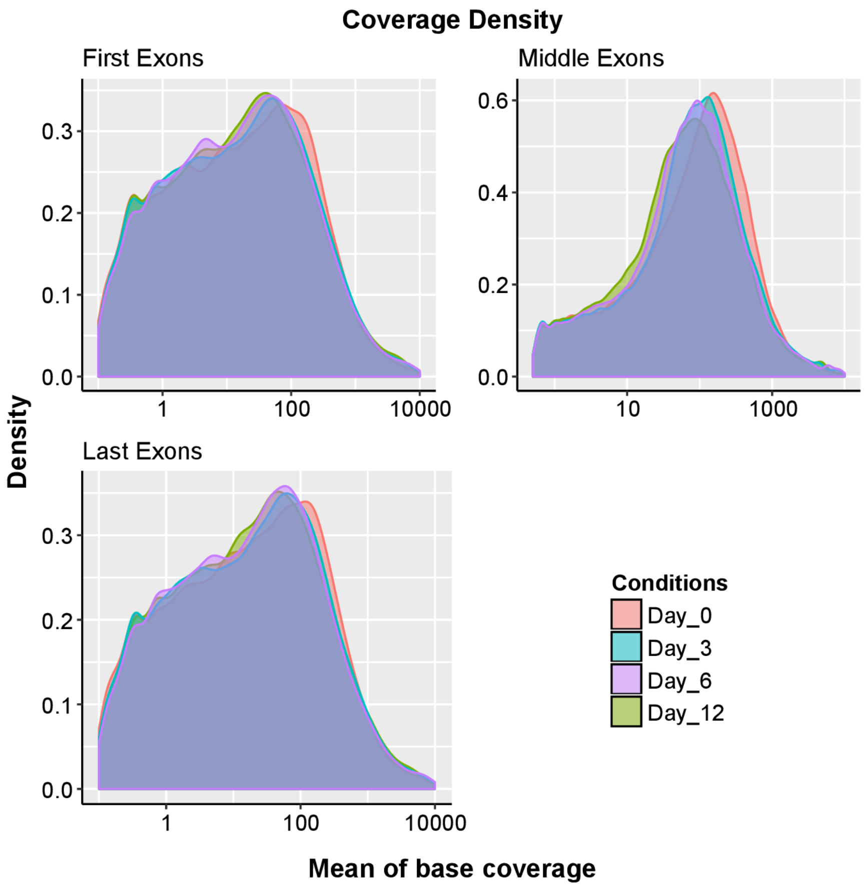

Figure 4.3: Coverage density for RNA-seq reads features. The plot shows three coverage density distribution of the base mean in the first exons (5' UTR exons), middle exons (within translation regions) and the last exons (3' UTR exons).

Using linear models provides the flexibility to analyze more complex experimental designs, to serve the genomic studies.

4.4.2 RESULTS:

4.4.2.1 DESeqDataSet Object Design

To create DESeq object, it is necessary to provide with the count matrix, the experiment design. Which is the sample information table. For osteoblast, we designed the experiment with one conditions level (time points). The row names are the columns name of count matrix (the samples), the columns are the conditions, here we have one condition level with four variables. The sample information table looks like the following

Samples	Condition
Day_0_R1 Day_0_R2 Day_0_R3	Day_Zero Day_Zero Day_Zero
Day_3_R1 Day_3_R2 Day_3_R3	Day_Three Day_Three Day_Three
Day_6_R1 Day_6_R2 Day_6_R3	Day_Six Day_Six Day_Six
Day_12_R1 Day_12_R2 Day_12_R3	Day_Twelve Day_Twelve Day_Twelve

Table 4.1: Table of osteoblast experimental samples

4.4.2.2 Performing Differential expression analysis:

As described in methods we need first to estimate the size factors normalization, then estimate the count dispersion, and the final step is performing the negative binomial GLM fitting and Wald test statistics.

To figure out the number of comparisons between experiment conditions, we got to the conclusion that for n conditions, the number of comparisons is given by taking the Combination of number of conditions n and $n-2$:

$$C_{n-2}^{n} = \binom{n}{n-2} = \frac{n!}{2! \cdot (n-2)!}$$

The function *"results"* from DESeq Bioconductor package gives the results of these comparisons including log2 fold changes of expression values, p-values and adjusted p-values which calculated based on the used statistical test.

We use Bonferroni correction for the p-value of Wald test, where we use the estimated standard error of a log2 fold change to test if it is equal to zero. For the FDR cutoff, we set the argument alpha in *results* function to initial value of 0.01. This independent filtering based on the mean of normalized counts for each gene, optimizing the number of genes which will have an adjusted p value below a given FDR cutoff.

The common used cutoff of the adjusted p_value is (0.01), and fold change of 4, so the log2 fold change (LFC) is 2. Applying these cutoffs, we got the following number of significant differentially expressed genes in 6 comparisons:

Comparisons	Significance p_value < 0.01	Differentially expressed FC >4, *|LFC| > log2(4)*	Upregulated LFC >2	Downregulated LFC < -2
Day_12 vs Day_0	10058	2299	1231	1068
Day_6 vs Day_0	10026	2834	1702	1132
Day_3 vs Day_0	9695	2378	1466	912
Day_12 vs Day_3	5551	655	233	422
Day_12 vs Day_6	5427	626	214	412
Day_6 vs Day_3	1788	155	92	63

Table 4.2: Number ofdifferential expressed genes with adj. p_value <0.01 and |LFC| > 2.

Using default threshold of 0.01 for adjusted p-value by Bonferroni correction, we got 12932 genes significantly differentially expressed at least in one of the comparisons out of 29148 genes analyzed. We got (4012) genes as a union of genes that significantly expressed at least in one of the comparisons, with adjusted *p*-value less than 0.01 and absolute logarithm fold change greater than 2. From the BioMart data mining tool in Ensembl, we could define the DE protein coding genes, and the noncoding ones, writing a code in R to define the bio type of each gene in DGE data set. In the following table (Table 4.3) is the statistics of protein coding genes, non-coding and linc-RNA:

	All genes	Protein coding	Non-coding	Linc-RNA
All genes in DESeq	29148	17948	11200	2008
Significant Adj.P <0.01	12932	11773	1159	287

DE genes union Adj.P <0.01 & LFC >2	4012	3308	704	184

Table 4.3: Union of DE genes; upregulated and downregulated, protein coding and noncoding.

Although we used stricter p-value cutoff commonly used by researchers (alpha = 0.001), we still got huge number of significantly differentially expressed genes as it is illustrated in Table 4.4 and MA-plot Figure 4.4.

Comparisons	Significance p_value < 0.001	Upregulated LFC >2	Downregulated LFC < -2
Day_12 vs Day_0	9609	1170	1130
Day_6 vs Day_0	9555	1628	1077
Day_3 vs Day_0	9201	1399	869
Day_12 vs Day_3	5069	219	393
Day_12 vs Day_6	4922	197	387
Day_6 vs Day_3	1506	90	59

Table 4.4: Number of DGE with adj. p_value <0.001 and |LFC| > 2

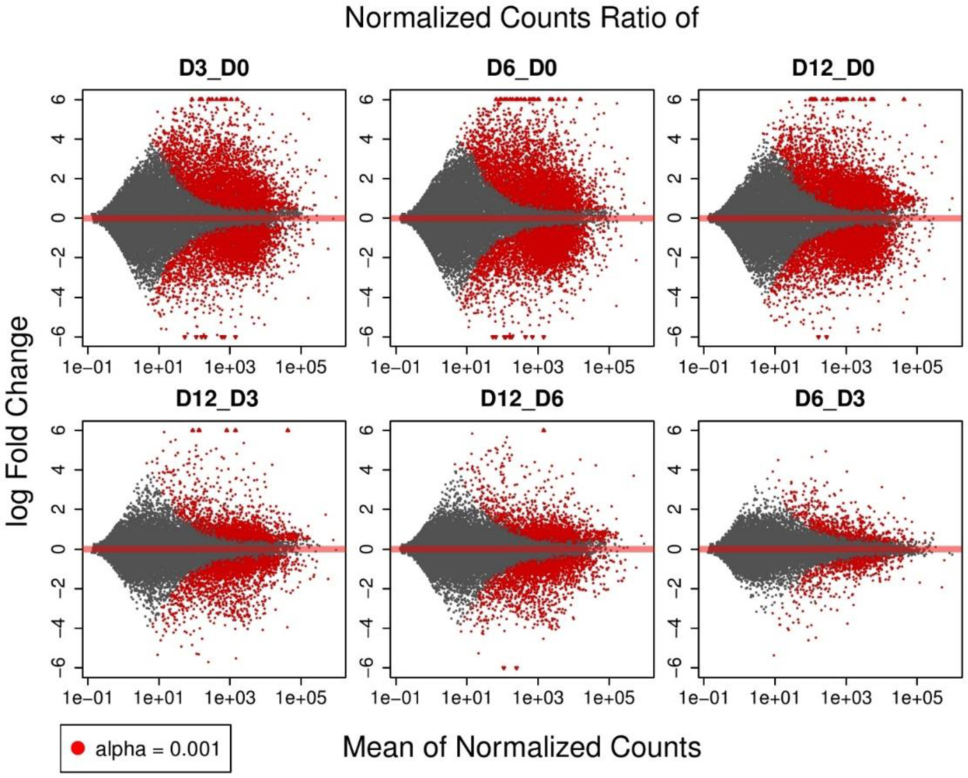

Figure 4.4 MA-plot of 6 comparisons between osteoblast differentiation time points. Red points indicate the genes if the adjusted p-value less than 0.001

The MA-plot is originally used to visualize DNA microarray gene expression data, however, it is also used to visualize high-throughput sequencing analysis. It plots the distribution of differences between normalized counts taken in two samples. M refers to log ratio (log2 Fold change) and A refers to mean average scales (mean of normalized counts).

4.4.2.3 Setting Thresholds:

Using the common used cutoffs of adjusted p-value and LFC, gives us huge number of genes not suitable for further Gene Ontology enrichment analysis or clusters visualization. Therefore, to choose the appropriate thresholds for our data, we use volcano plot. Usually we cut when the volcano arms start to open as the following illustrates.

In the panel (A) the cutoff of the adjusted *p*-value is 0.01, we can see most of the volcano dots are blue, which are the significant DE genes that have adjusted p-value less than 0.01, the small black line in the base of the volcano is the non-significant. In (B) we added the cutoff for the fold change of 4: $abs(log_2FC) > log_2(4)$. We chose the cutoff adjusted *p*-value 10^{-50} according to the volcano plot, as it is illustrated in (C). In (D) applying both cutoffs on the genes set of $|LFC| > \log_2 5$ and $-\log_{10} adjP < 50$.

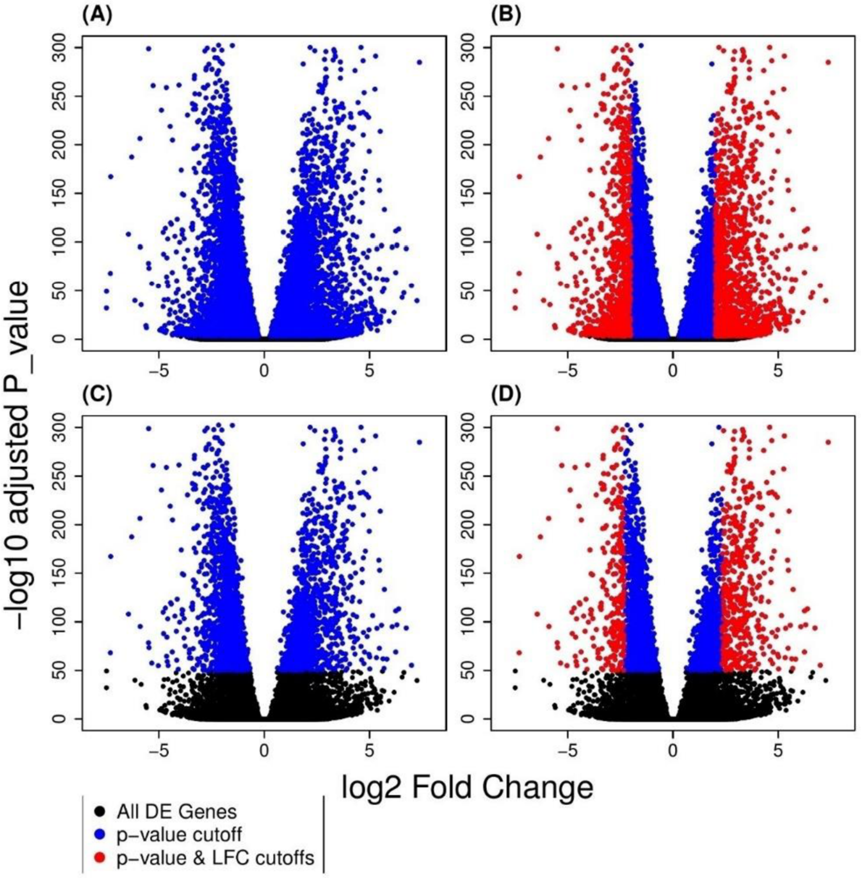

Figure 4.5 Example of volcano plots of DGE in day3 vs day 0, illustrates making the decision of choosing adjusted p-value cutoff = $10^{\wedge-50}$.

The following graph in Figure 4.6 of volcano plots of the six comparisons, proves that choosing the cutoffs was adequate decision for all the samples.

By using cutoff of adjusted p-value less than $10^{\wedge-50}$ and absolute value of logarithm fold change |LFC| greater than $\log_2 5$, we got 1441 genes including 1386 protein coding, 55 non-coding and 19 lincRNA (Table 4.5).

adjP< $10^{\wedge-50}$ |LFC|> $\log_2(5)$	All genes	Protein coding	Non-coding	lincRNA
DE genes union	1441	1386	55	19

Table 4.5 Significant differential expressed genes subsets with cutoff adjP< $10^{\wedge-50}$

4.4.2.4 Count data transformation:

For samples similarity analysis and visualization as clustering, it is important to use homoscedastic data (all random variables in the sequence have the same finite variance). Heteroscedasticity in RNA-Seq data causes a problem, when the original count scale is used in clustering or ordination algorithm, the result will be dominated by highly expressed, highly variable genes; if logarithm-transformed data are used, undue weight will be given to weakly expressed genes, which show exaggerated LFCs.

The purpose behind data transformation is to remove the dependence of the variance on the mean, particularly the high variance of the logarithm of count data when the mean is low. It produces transformed data on the $\log_2$ scale which has been normalized with respect to library size. In order to transform the data to remove the experiment-wide trend. The aim of this transformation is not that all the genes have exactly the same variance after transformation. However, after the transformations, the genes with the same mean do not have exactly the same standard deviations, but that the experiment-wide trend has flattened. It is those genes with row variance above the trend which will allow us to cluster samples into interesting groups.

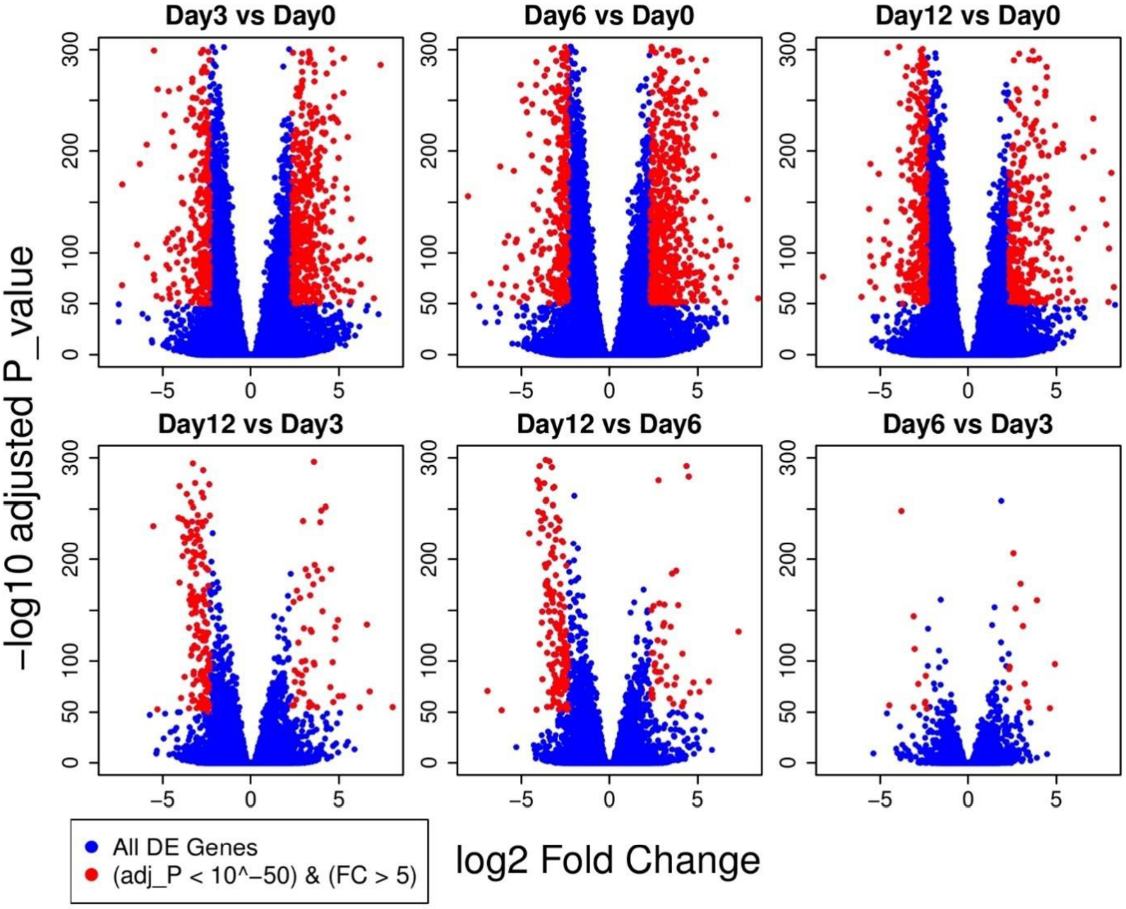

Figure 4.6: Volcano plots of the differential gene expression in osteoblast differentiation days.

4.4.2.5 Principal Component Analysis:

Principal component analysis (PCA) is a statistical procedure that can be used for exploratory data analysis. PCA uses linear combinations of the original data (gene expression values) to define a new set of unrelated variables (principal components). These new variables are orthogonal to each other, avoiding redundant information. [63]. Thus, PCA can be used to reduce the dimensions of a data set, allowing the description of data sets and their variance with a reduced number of variables. It is often sufficient to look at the first two components, as these describe the largest variability.

PCA plot is useful for visualizing the overall effect of experimental covariates and batch effects (technical sources of variation), it is used to get an impression on the similarity of RNA-sequencing samples. The variance in RNA-Seq data usually grows with the expression mean, using PCA on the transformed data matrix by regularized logarithm transformation will often lead to principal components that are dominated by the variance of a few highly expressed genes, and avoid the high random noise of low count data

The following graph in Figure 4.7 is the PCA plot of osteoblast data set that has 12 samples for 4 biological conditions. The replicates in each condition show similarity in the variances which proves that the experimental samples did not suffer from an abnormality in variance between biological replicated.

4.4.2.6 Sample to sample distance:

We computed the Euclidean distance between the samples, using the regularized logarithm transformed data count. From the distance matrix, we created dendrogram of the samples as Figure 4.8 illustrates, the count data in day_0 of osteoblast differentiation has greater variance to the other days, while day_3 and day_6 are closer, this means the genes in both time points have similar expression patterns.

A heatmap of the distance matrix gives an overview over similarities and dissimilarities between samples, so we used the Euclidean distance of rlog transformed data likewise to create a distance heatmap as the Figure 4.9.

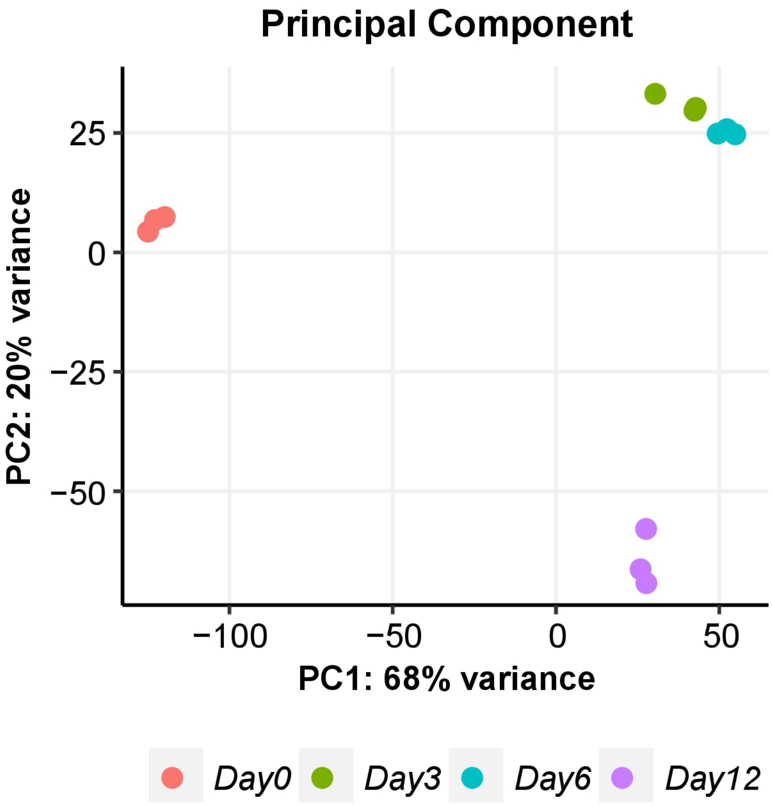

Figure 4.7 PCA of osteoblast dataset. The replicates in each condition show similarity in the variances

4.4.2.7 On / off genes subset analysis

The overall distribution of the fold change differences between the conditions was almost symmetric (Figure 4.10). However, we found groups of genes with on/off expression as in Table 4.6. We research each gene of them and defined a group of genes which have significant biological role in osteoblast differentiation and ossification.

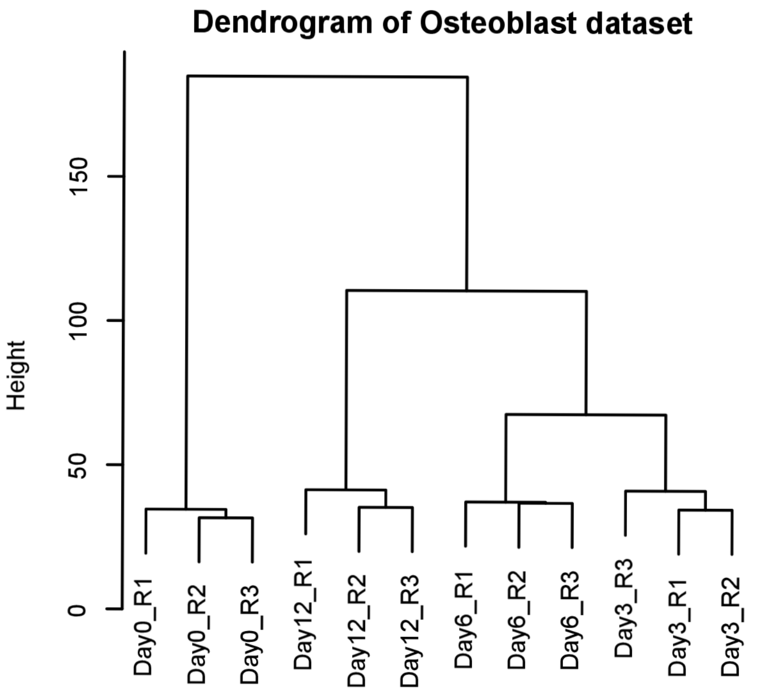

Figure 4.8 Hierarchical clustering dendrogram of osteoblast dataset

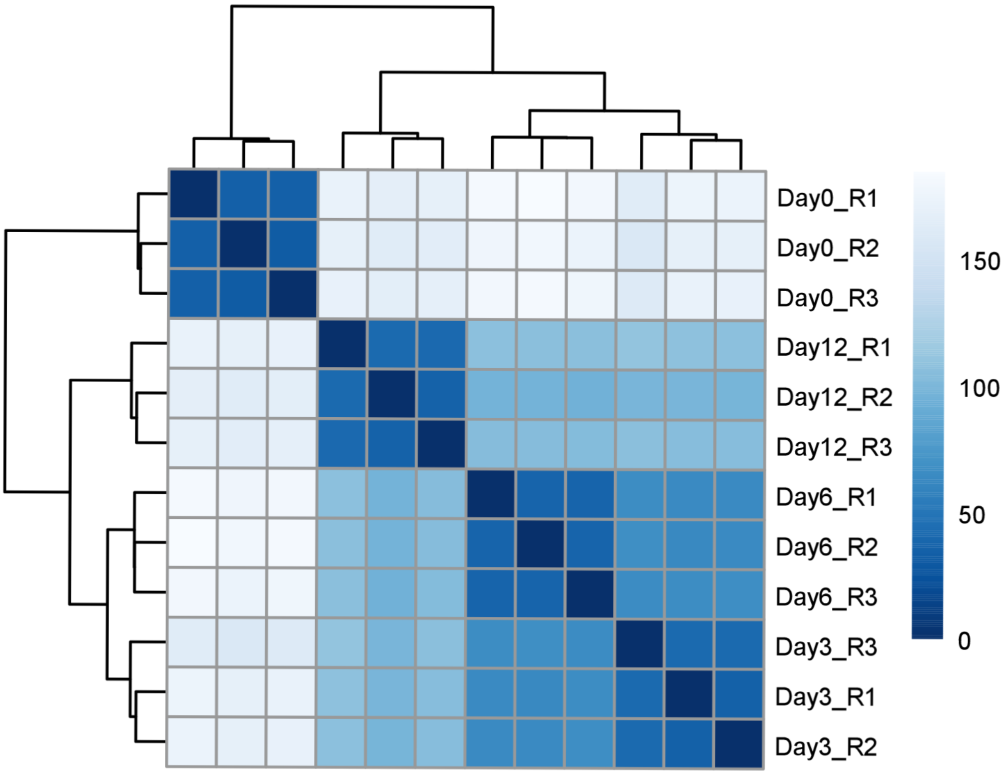

Figure 4.9 Heatmap of samples distance

Comparison **X** vs **Y**	On gene expression on in **X**& off in **Y**	Off gene expression off in **X**& on in **Y**
Day 12 vs Day 0	81	19
Day 12 vs Day 3	12	11
Day 12 vs Day 6	15	5
Day 6 vs Day 0	97	33
Day 6 vs Day 3	1	3
Day 3 vs Day 0	90	27

Table 4.6 ON/OFF genes in osteoblast dataset

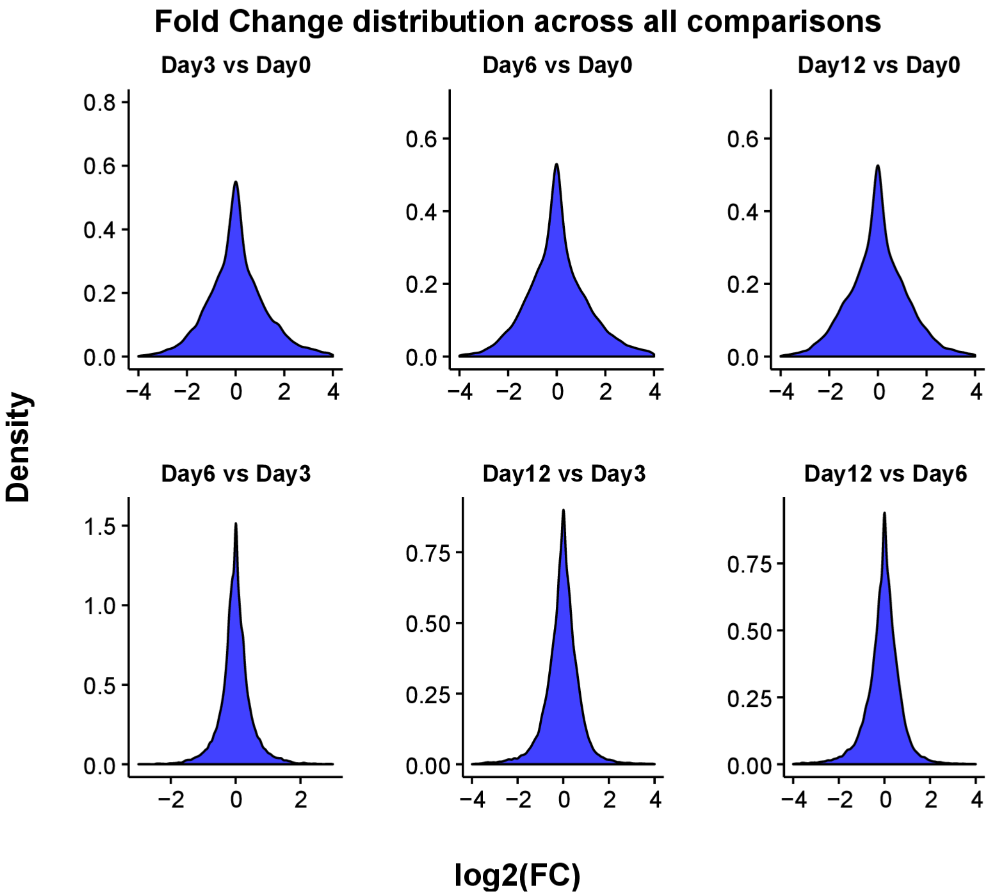

Figure 4.10 The distribution of overall fold change differences across all the comparisons

4.5 GENE ONTOLOGY ANALYSIS ENRICHMENT

4.5.1 METHOD

For clustering the differential expressed genes, we plot a heatmap that creates a similarity matrix of the values and groups the genes with similar pattern together, then highlights them in different colors. We got the normalized count of the DE genes we desire to cluster, from DESeqDataSet object(dds). Then we calculate the count mean of the replicates in each condition, we create new mean count matrix with all the genes names. The fundamental step to generate clusters heatmap, is to scale the normalized count with mean centering as the following:

$$x_{new} = \frac{x - \mu}{\sigma}$$

Where x is the value of normalized counts, μ is the mean of the rows, σ is the standard deviation, however the mean of the new row values is 0 and the standard deviation is 1. The heatmap function in R calculate the Euclidean distances and calculate the variance to cluster the genes and reorder the values according to its dendrogram. We analyzed the differentially expressed genes clusters by Ontologizer [64], using the model-based gene set analysis method "MGSA"[65] we got the top annotation-enriched GO terms of the differential gene expression group. Then we applied parent-child intersection approach [66]on the clusters, to get the enriched GO terms of each cluster. For correcting the *p*-value we used Benjamini Hochberg correction.

4.5.2 RESULTS

There are 1386 protein coding genes, differentially expressed with more stringent cutoff values (absolute fold change of fivefold or greater and adjusted *P*-value < 10^{-50}). These genes were plotted within a heatmap to identify clusters of genes according to their expression pattern. In total, nine clusters were identified by visual inspection, as illustrated in protein coding genes clusters heatmap (Figure 4.11).

4.6 DIFFERENTIAL EXONS USAGE

Alternative transcription start-sites lead to differences in the beginning of mRNA, whereas alternative splicing causes some of the exons to be skipped and not translated at all. RNA-seq offers exciting possibilities for studying the expression and regulation of isoforms on the whole genome level.

Most of the current RNA-seq methods produce short reads which do not cover full transcripts. Instead, transcripts need to be assembled from sequenced fragments. The assembly and the subsequent abundance estimation can be challenging, because isoforms typically have common or overlapping exons. Furthermore, the coverage along transcripts is not uniform because of biases introduced in sequencing and library preparation. In order to avoid uncertainties in the assembly, one approach for studying alternative isoform regulation of it is to look at differences in the usage of individual exons. RNA-seq reads can also be mapped to exons so that the differences in exon-specific counts can be compared between certain conditions, groups, or treatments.

4.6.1 METHODS

We used a statistical method to test for differential exon usage in RNA-seq data, by applying Bioconductor package DEXSeq [67], which uses generalized linear models, taking into the account the biological variability and looks for differences across conditions of the relative usage of each exon. Using HTSeq library in python, we generated a reference annotation genes model (flattened GFF file) contains one entry for each exon or exonic part, which is cut from the exon if the exon's boundary differs between transcripts. Then we got the count of reads that overlap with each of the exon counting bins defined in the flattened GFF file in each sample. The DEXSeq function normalizes these counts by the library size factor s_j, which accounts for the depth that sample j was sequenced. The number of the reads N_{ijk} overlapping counting bin (exonic part) k of gene i in sample j, follows the negative binomial (NB) distribution and modeled by GLMs, where the dispersion parameter is estimated by, firstly performing an IRLS (iteratively reweighted least square) fits for each gene, then, insert these fitted values in the log likelihood function with Smyth's Cox-Reid [68][69] term and find its maximum using Brent's line search. So, the gene expression variability is absorbed by the model parameters, while the model increase the power of the test for differential exon usage.

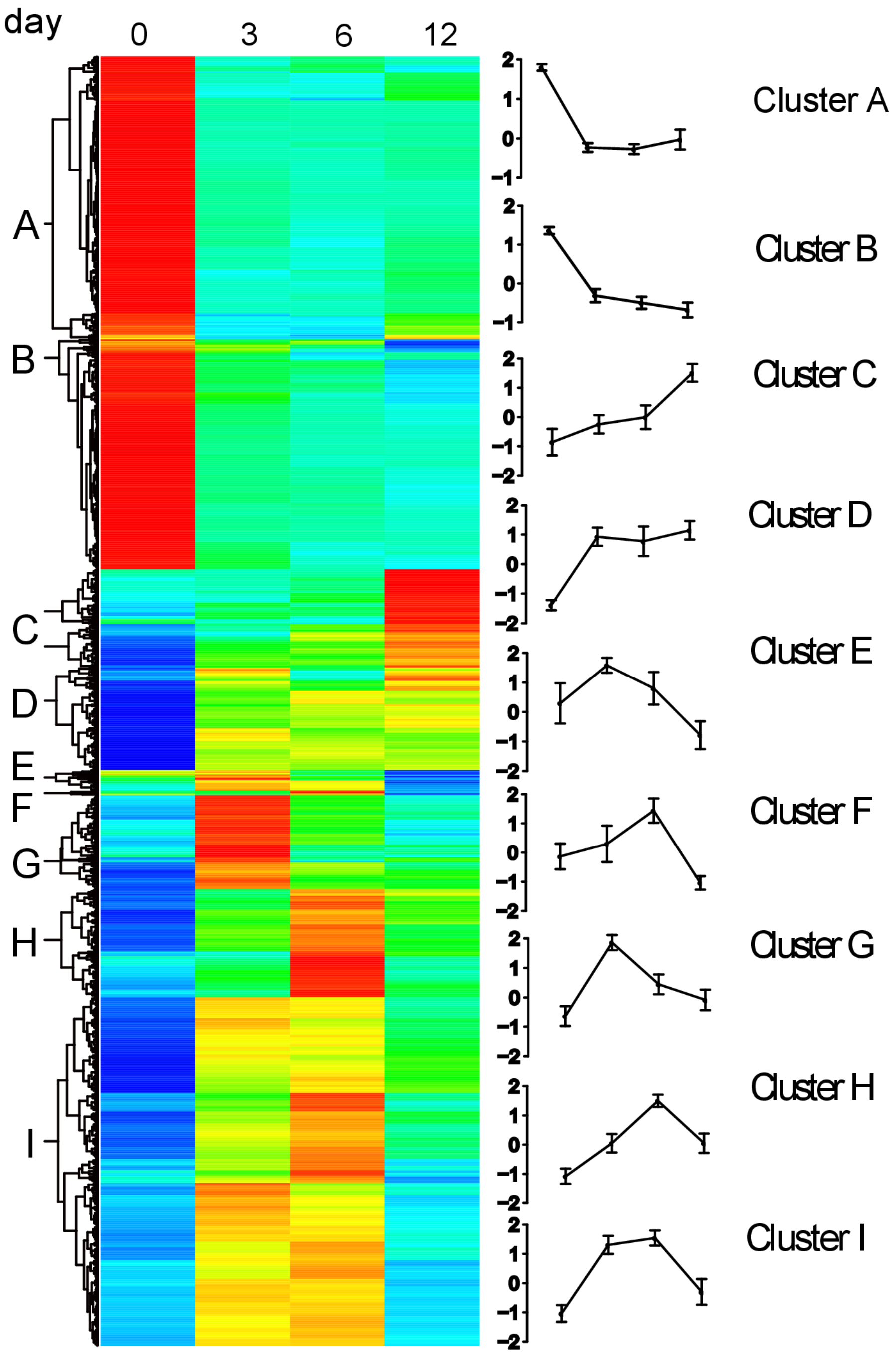

Figure 4.11 Heatmap of genes cluster with their expression patterns and significant GO terms

4.6.2 RESULTS

We used DEXseq method to test for differential exon usage in comparative RNA-Seq experiments. We mean by differential exon usage (DEU), changes in the relative usage of exons caused by the experimental condition. The relative usage of an exon is defined as:

$$\frac{\text{number of transcripts from the gene that contain this exon}}{\text{number of all transcripts from the gene}}$$

We test the differential exon usage in each comparison, applying merely the DEXSeq method with thresholds; adjusted p-value less than 0.01 and FC greater than 4, we got the following number of entries, which are exons or exonic part differentially used between the conditions, many of those exons belong to more than one transcript, and some of them overlapped with few genes. The results presented in Table 4.7

Comparisons	Adjusted p-value < 0.01	adjp < 0.01 & |logFC| > 2
Day12 vs Day0	17229	291
Day12 vs Day3	3216	71
Day12 vs Day6	2917	35
Day3 vs Day0	17283	293
Day6 vs Day0	18408	373
Day6 vs Day3	480	8

Table 4.7 Differential Exon Usage (DEU) in osteoblast comparisons; adjp <0.01, |logFC|>2

The results from DEXSeq are not adequate for further alternative splicing analysis. Therefore, the genes name and gene type associated with the entries must be provided, using Biomart data mining tool from Ensemble, I built the function code in R (Main dissertation: Code-box 3.14).

It has been found in previous studies that almost 73% of human genes are alternatively spliced [70]. To figure out the featured biological role that a gene can play when alternative transcripts expressed in different conditions, we need to determine if the exonic part that differentially used between those conditions is protein-coding or within the open reading frame ORF. To achieve this task, first I coded a function to provide the bio-type of transcripts that include the differential used exon, as it is illustrated in Code-box 3.15. The output of this function provides us with a data matrix includes the DEU, logFC, adj.p-value, Gene name, Gene type, Transcript type, Genomic start and end of the exon part, Genomic width, and genomic strand.

This data matrix is the basis of ExonORF function that define if the exonic part is within ORF. I coded ExonORF Code-box 3.16 function in the following steps:

1- Get the table exons features from BioMart in Ensemble, of all transcripts that differentially used in all comparisons. The exon feature table wouldh look like this:

Exon.Chr.Start	Exon.Chr.End	En.Transcript.ID	G.coding.start	G.coding.end	Strand
15356	15422	ENSMUST00000082423	NA	NA	-1
14145	15288	ENSMUST00000082421	14145	15288	+1

The exons that are non-coding the Genomic.coding.start and Genomic.coding.end not available, while some exons are partially within ORF so they have either Genomic coding start or end.

2- Check the exonic bin is within the range of a defined exon in Ensemble. S_E , E_E are the start and end of an exon, and S_B, E_B are the start and end of an exon part.

3- Check if this exon part is within the ORF. S_c , E_c are the start and end of coding frame.

$$S_B, E_B \in [S_E, E_E] \cap S_B, E_B \in [S_c, E_c]$$

Table 4.8 shows the number of differential used exons in each comparison, that are within the coding frame.

Comparisons	Exons within ORF (adjp < 0.01 & |logFC| > 2)
Day12 vs Day0	32
Day12 vs Day3	13
Day12 vs Day6	7
Day3 vs Day0	35
Day6 vs Day0	50
Day6 vs Day3	0

Table 4.8 number of the differential used coding exon (within ORF) in Osteoblast dataset

4.7 LONG NON-CODING RNA IDENTIFICATION

4.7.1 METHOD

We innovated an algorithm to predict the potential functions of lncRNA genes which are differentially expressed, by their correlation with protein coding genes. The algorithm is based on two concepts (Figure 4.12); the first concept is the spatial interaction of the lncRNA and the protein coding genes or what is called ***topologically associating domain*** (TAD)[71]. And the second concept is the co-expression correlation which was used in previous study for lncRNA functions characterization[72]. Then define the enriched functional terms among the protein coding genes that are significantly correlated with lncRNAs using a gene ontology tool.

Since the lncRNA expressed in lower level, so we used more tolerant cutoffs to get the significant differential expression, whereas the adjusted *p*_value less than 0.01 and the binary logarithm of fold change between the conditions is greater than 1.5 (adj.Pvalue ≤ 0.01 , FC $\geq$ 1.5). The algorithm workflow is concise in the following steps:

Mapping the reads and the default differential gene expression pipeline using DESeq2.

1. Creating a DGE data matrix for protein-coding genes and another data matrix for lncRNA.
2. For computing the correlation, we generate two matrices for lncRNA and protein-coding genes, contain normalized counts (expression values) among the different Osteoblast differentiation time point.
3. Getting the TAD annotation for mouse genome mm10.
4. Selecting the protein coding gene and lncRNA pairs based on two criteria; firstly, the protein-coding gene and the lncRNA must be within the same topological associated domain. Secondly, the absolute correlation value must be greater than 0.9, and the p_value of the correlation test less than 0.01.
5. Most of the lncRNA correlated to more than protein-coding genes. lncRNA have positive correlation with protein coding genes, when they have similar expression pattern, and negative correlation when their expression patterns are contradicted.

4.7.2 RESULTS

Applying this algorithm on osteoblast data, we got throughputs needed to be biologically verified, which was not possible due to flaws in wet-lab organization, however, according to the logical methodology we used, based on published literature [71][72], the algorithm is applicable and need RNA-seq data to confirm it.

As statistical overview of results I got from osteoblast data; there were 10760 protein coding genes differentially expressed with thresholds (adj.Pvalue ≤ 0.01 , FC $\geq$ 1.5) and within the Topological Associated Domains (TAD) , 299 of those protein coding genes are correlated with 126 lncRNA. However, we have 285 differentially expressed lncRNA with the same cutoffs, 158 of them were not correlated (Table 4.9 number of differential expressed and correlated Pro.Cod genes and lncRNA)

	DGE adjp < 0.01 & FC >1.5	Within TADs	Correlated	Not correlated
Protein coding	10775	10760	299	10461
lncRNA	285	284	126	158

Table 4.9 number of differential expressed and correlated Pro.Cod genes and lncRNA

The distinct result in our analysis, that our algorithm can define when the lncRNA is positively correlated with the protein coding gene, this guide us to the hypothesis that the lncRNA enhancing the expression of the protein coding gene, in other words, it plays positive regulation role to that Pro.Cod gene. While the lncRNA negatively correlated with the protein coding gene, it plays suppressor role.

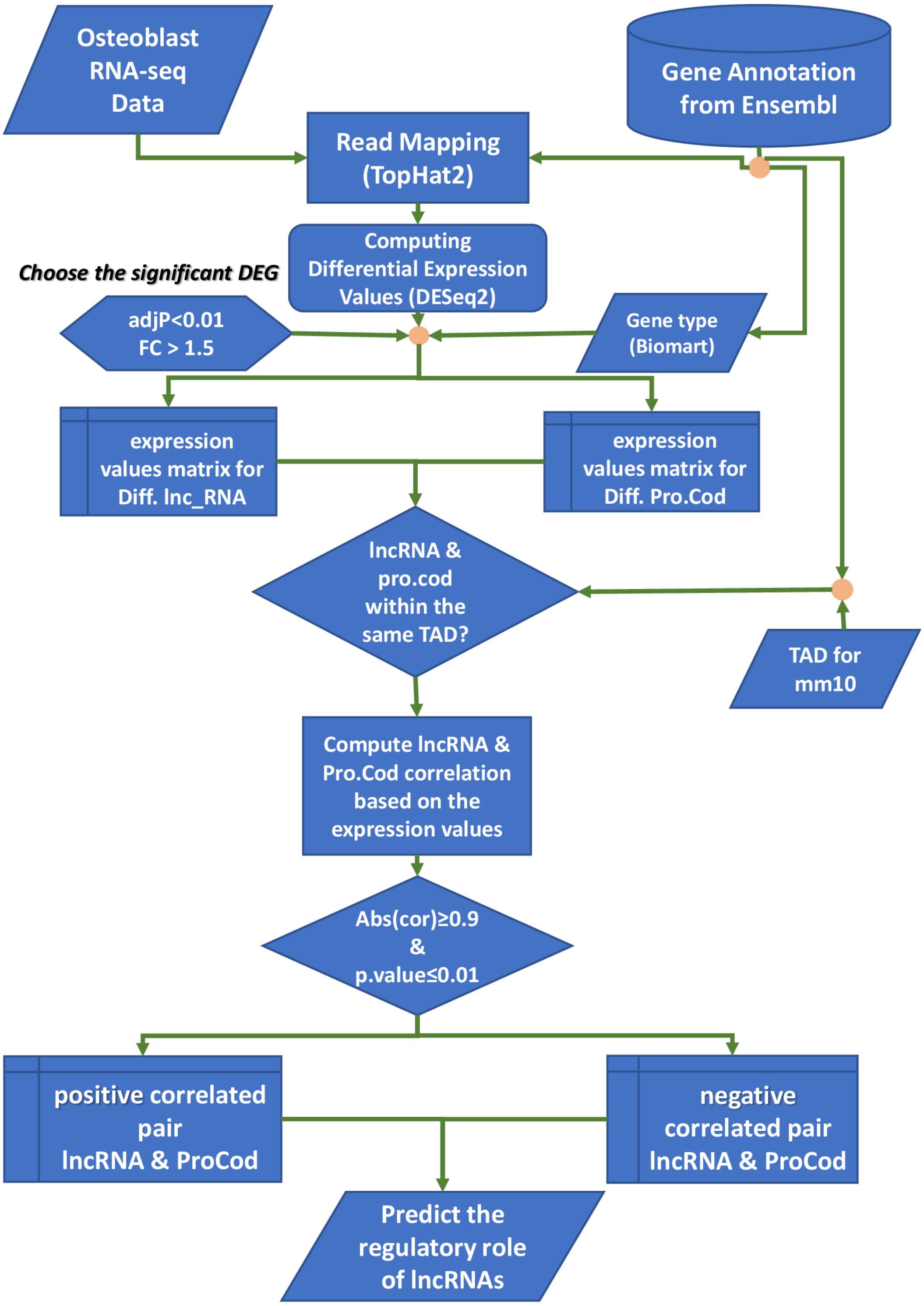
Osteoblast RNA-seq Data
Gene Annotation from Ensembl
Read Mapping (TopHat2)
Computing Differential Expression Values (DESeq2)
Choose the significant DEG
adjP<0.01 FC > 1.5
Gene type (Biomart)
expression values matrix for Diff. lnc_RNA
expression values matrix for Diff. Pro.Cod
lncRNA & pro.cod within the same TAD?
TAD for mm10
Compute lncRNA & Pro.Cod correlation based on the expression values
Abs(cor)≥0.9 & p.value≤0.01
positive correlated pair lncRNA & ProCod
negative correlated pair lncRNA & ProCod
Predict the regulatory role of lncRNAs

Figure 4.12 lncRNA correlation algorithm

Some of the lncRNA correlated with more than one Pro.Cod gene, therefore we have chosen the maximum correlated Pro.Cod gene (positively and negatively). However, for the analysis, we needed to study the full set of Pro.Cod genes that are correlated with one lncRNA. Moreover, many of lncRNA correlated positively with some Pro.Cod genes and negatively with others. We found few biological meaningful examples of such lncRNA behavior served our research on osteoblast differentiation (See our submitted paper).

	All pairs	Maximum correlated
Positive correlation	217	91
Negative correlation	131	66

Table 4.10 positive and negative correlated Pro.Cod and lncRNA pairs

5. DISCUSSION:

Despite the availability of the web-based platform for RNA-Seq data analysis as Galaxy [73], our pipeline analysis in this doctoral research still carries novel and auxiliary approaches, which provide integrity, and improve the performance of the existing tools. It provides a solution to input a set of samples into a tool, by means of a single function command. Furthermore, our pipeline provides informative results, allowing straightforward interpretation by the biologists. To discuss the novelty and the benefits of our research, I will go through the pipeline's procedures one by one as follows:

1- We coded three functions (fastqc_insert, trimo_insert, tophat_insert) to input a set of samples to FastQC, Trimmomatic and TopHat tools respectively. Despite the simplicity of the codes, they serve a good purpose by facilitating the input of samples, where the user simply needs to apply accessible paths, where the input data is stored, and where the outputs of the tools are needed to be stored. The user doesn't need to have any skills in bash script or shell command-line to run those tools. These inserting functions can be useful cores when we design an integrated web-platform for RNA-Seq data analysis, since we as the operators don't need huge data-storage to analyze the data or worry about server maintenance, and on the other hand the users don't need to upload their big fastq files elsewhere and spend hours and interrupt their network connection.

2- The quality control tool FastQC returns two sets of files for each sample (one for the forward reads and the other for reverse reads), this means for a small experiment, it will produce a range of 20 output sets. Retrieving quality information from the FastQC files will be a burden to go through the plain text files for each sample. Therefore, we coded two functions; "BasicStatistic" to retrieve the length of reads and the number of reads, before and after trimming. And "QualityScore" to calculate the mean quality of all bases in a fastq file. Furthermore, we coded "read_qual" function to get average quality in each base along reads positions, and to plot the mean quality of bases across the reads positions.

3- The coverage uniformity of the gene features is one of the concerns when using libraries based on polyadenylated RNA. Although there are tools to plot graphs and heatmap represent the coverage along the gene body asRSeQC[74], our method to check the coverage uniformity gives a detailed vision of the coverage distribution along the gene's features, by plotting the density of mean of base coverage after separating the first exons in 5′UTR, the last exon in 3′URT, and the middle exons.

4- In the differential gene expression pipeline, we defined the significant differential expressed genes using the statistical model of DESeq [61]. However, we improved the default pipeline, the purpose being to obtain informative outputs from the differential gene expression matrix. We proposed a procedure to set the suitable cutoffs of the *P*-value and logarithm fold change as described in the "Setting Thresholds:" section. Apart from this, we defined the biotype of each significant differentially expressed gene, so we could do the gene enrichment for the protein-coding genes and separate the long non-coding RNA for our further analysis. Furthermore, based on the differential gene expression matrix, we proposed the analyzing procedure for the multiple conditions experiment, we called it "ON/OFF genes". This procedure suggests defining the genes that are completely silent during a biological condition while it is expressed in the others. Using this analysis, we can get a set of genes that are uniquely regulated between biological conditions.

5- For alternative splicing analysis, we improved the performance of the differential exon usage tool DEXSeq[67], by defining whether the differential used exon (or part of the exon) is a coding exon within the ORF. Followed by the comparison of the domains of the transcripts that contain the differentially expressed exon, to determine the functions or the products that are affected by the alternative splicing of the gene.

6- long non-coding RNA (lncRNA) species have been identified whose loci locate both within and between protein coding genes. While lncRNAs remain the most enigmatic ncRNA species in terms of function, there is now much effort centered on their functional characterization and their molecular mechanisms in different cell types[75]. However, our method is concentrated on finding the potential interaction between the lncRNA and the protein-coding genes, by finding the expression correlation between the lncRNA and protein coding genes that are within the same Topological Associated Domain (TAD). Although we interduce our method as a novel approach, it is based on existing and approved researches. TAD is a known genome architecture, it is a self-interacting genomic region, meaning that DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD [71]. And defining the gene ontology terms of lncRNA by finding the expression correlation with protein-coding genes [72]. However, LncRNA2Function tool defines the co-expression without taking in consideration the topological associated domains. Furthermore, this database is available for GO terms in human genome, where our method can be applicable to any RNA-seq experiment.

REFERENCES

[1] M. L. Metzker, “Sequencing technologies - the next generation.,” *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2010.

[2] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics.,” *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, 2009.

[3] S. N. Peirson and J. N. Butler, “RNA Extraction From Mammalian Tissues BT - Circadian Rhythms: Methods and Protocols,” E. Rosato, Ed. Totowa, NJ: Humana Press, 2007, pp. 315–327.

[4] Korpelainen Eija and Tuimala Jarno, *RNA-seq Data Analysis: A Practical Approach - Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss, Garry Wong - Google Books*. CRC Press, 2015.

[5] S. R. Gallagher and P. R. Desjardins, “Quantitation of DNA and RNA with absorption and fluorescence spectroscopy.,” *Curr. Protoc. Mol. Biol.*, vol. Appendix 3, p. Appendix 3D, 2006.

[6] A. Masotti and T. Preckel, “Analysis of small RNAs with the Agilent 2100 Bioanalyzer,” *Nat. Methods*, vol. 3, no. 8, p. 62507, 2006.

[7] J. Pease and R. Sooknanan, “A rapid, directional RNA-seq library preparation workflow for Illumina[reg] sequencing,” *Nat Meth*, vol. 9, no. 3, Mar. 2012.

[8] S. Goodwin, J. D. McPherson, and W. R. McCombie, “Coming of age: ten years of next-generation sequencing technologies,” *Nat Rev Genet*, vol. 17, no. 6, pp. 333–351, 2016.

[9] E. R. Mardis, “Next-Generation DNA Sequencing Methods,” *Annu. Rev. Genomics Hum. Genet.*, vol. 9, no. 1, pp. 387–402, 2008.

[10] A. Diekstra *et al.*, “Translating sanger-based routine DNA diagnostics into generic massive parallel ion semiconductor sequencing,” *Clin. Chem.*, vol. 61, no. 1, pp. 154–162, 2015.

[11] A. Rhoads and K. F. Au, “PacBio Sequencing and Its Applications,” *Genomics, Proteomics and Bioinformatics*, vol. 13, no. 5. pp. 278–289, 2015.

[12] K. J. Travers, C. S. Chin, D. R. Rank, J. S. Eid, and S. W. Turner, “A flexible and efficient template format for circular consensus sequencing and SNP detection,” *Nucleic Acids Res.*, vol. 38, no. 15, p. e159, 2010.

[13] A. Mccarthy, “Third Generation DNA Sequencing: Pacific Biosciences’ Single Molecule Real Time Technology,” *Chem. Biol.*, vol. 17, pp. 675–676, 2010.

[14] J. Eid *et al.*, “Real-time DNA sequencing from single polymerase molecules.,” *Science*, vol. 323, no. 5910, pp. 133–8, 2009.

[15] S. Koren and A. M. Phillippy, “One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly,” *Current Opinion in Microbiology*, vol. 23. pp. 110–120, 2015.

[16] Y. Feng, Y. Zhang, C. Ying, D. Wang, and C. Du, “Nanopore-based fourth-generation DNA sequencing technology,” *Genomics, Proteomics and Bioinformatics*, vol. 13, no. 1. pp. 4–16, 2015.

[17] P. Svoboda and A. Di Cara, “Hairpin RNA: A secondary structure of primary importance,” *Cellular and Molecular Life Sciences*, vol. 63, no. 7–8. pp. 901–918, 2006.

[18] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat. Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[19] “gene_structure.” [Online]. Available: http://disi.unitn.it/~teso/courses/sciprog/_images/genestructure.jpg. [Accessed: 02-Sep-2017].

[20] M. Chen *et al.*, “Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome.,” *Sci. Rep.*, vol. 5, no. November, p. 18019, 2015.

[21] C. Trapnell *et al.*, “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation,” *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511–515, 2010.

[22] J. Peltonen, V. Aarnio, L. Heikkinen, M. Lakso, and G. Wong, “Chronic ethanol exposure increases cytochrome P-450 and decreases activated in blocked unfolded protein response gene family transcripts in caenorhabditis elegans,” *J. Biochem. Mol. Toxicol.*, vol. 27, no. 3, pp. 219–228, 2013.

[23] A. C. Nica and E. T. Dermitzakis, “Expression quantitative trait loci: present and future.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 368, no. 1620, p. 20120362, 2013.

[24] F. Tang *et al.*, “mRNA-Seq whole-transcriptome analysis of a single cell,” *Nat. Methods*, vol. 6, no. 5, pp. 377–382, 2009.

[25] T. Hashimshony, F. Wagner, N. Sher, and I. Yanai, “CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification,” *Cell Rep.*, vol. 2, no. 3, pp. 666–673, 2012.

[26] H. Edgren *et al.*, “Identification of fusion genes in breast cancer by paired-end RNA-sequencing.,” *Genome Biol.*, vol. 12, no. 1, p. R6, 2011.

[27] H. Edgren *et al.*, “Identification of fusion genes in breast cancer by paired-end RNA-sequencing,” *Genome Biol.*, vol. 12, no. 1, 2011.

[28] R. Fani, M. Brilli, M. Fondi, and P. Lió, “The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case,” *BMC Evol. Biol.*, vol. 7, no. 2, 2006.

[29] E. M. Quinn *et al.*, “Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data,” *PLoS One*, vol. 8, no. 3, 2013.

[30] A. Djari *et al.*, “Gene-based single nucleotide polymorphism discovery in bovine muscle using next-generation transcriptomic sequencing,” *BMC Genomics*, vol. 14, p. 17, 2013.

[31] N. E. Ilott and C. P. Ponting, “Predicting long non-coding RNAs using RNA sequencing,” *Methods*, vol. 63, no. 1, pp. 50–59, 2013.

[32] M. A. Faghihi *et al.*, “Expression of a noncoding RNA is elevated in Alzheimer’s disease and drives rapid feed-forward regulation of β-secretase,” *Nat. Med.*, vol. 14, no. 7, pp. 723–730, 2008.

[33] J. S. Mattick and I. V. Makunin, “Non-coding RNA.,” *Human molecular genetics*, vol. 15 Spec No. 2006.

[34] J. Srinivasan *et al.*, “The draft genome and transcriptome of Panagrellus redivivus are shaped by the harsh demands of a free-living lifestyle,” *Genetics*, vol. 193, no. 4, pp. 1279–1295, 2013.

[35] T. R. Mercer *et al.*, “Targeted RNA sequencing reveals the deep complexity of the human transcriptome,” *Nat. Biotechnol.*, vol. 30, no. 1, pp. 99–104, 2011.

[36] T. Komori, “Regulation of osteoblast differentiation by transcription factors,” *Journal of Cellular Biochemistry*, vol. 99, no. 5. pp. 1233–1239, 2006.

[37] G. D. Roodman, “Cell biology of the osteoclast,” *Experimental Hematology*, vol. 27, no. 8. pp. 1229–1241, 1999.

[38] K. K. Papachroni, D. N. Karatzas, K. A. Papavassiliou, E. K. Basdra, and A. G. Papavassiliou, “Mechanotransduction in osteoblast regulation and bone disease,” *Trends in Molecular Medicine*, vol. 15, no. 5. pp. 208–216, 2009.

[39] V. Kumar, A. K. Abbas, J. C. Aster, and N. Fausto, *Robbins & Cotran pathologic basis of disease*, 8th ed. Philadelphia: Saunders, 2009.

[40] A. Rutkovskiy, K.-O. Stensløkken, and I. J. Vaage, “Osteoblast Differentiation at a Glance.,” *Med. Sci. Monit. Basic Res.*, vol. 22, pp. 95–106, 2016.

[41] T. Komori *et al.*, “Targeted Disruption of Cbfa1 Results in a Complete Lack of Bone Formation owing to Maturational Arrest of Osteoblasts,” *Cell*, vol. 89, no. 5, pp. 755–764, 1997.

[42] F. Long, “Building strong bones: molecular regulation of the osteoblast lineage,” *Nat. Rev. Mol. Cell*

Biol., vol. 13, no. 1, pp. 27–38, 2011.

[43] D. L. Black, "Mechanisms of Alternative Pre-Messenger RNA Splicing," *Annu. Rev. Biochem.*, vol. 72, no. 1, pp. 291–336, 2003.

[44] J. F. Cáceres and A. R. Kornblihtt, "Alternative splicing: Multiple control mechanisms and involvement in human disease," *Trends in Genetics*, vol. 18, no. 4. pp. 186–193, 2002.

[45] "Alternative RNA splicing." [Online]. Available: http://www.exonhit.com/technology/alternative-rna-splicing. [Accessed: 21-Sep-2017].

[46] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, 2009.

[47] Simon Andrews, "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data." [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed: 10-Aug-2017].

[48] L. Wang, S. Wang, and W. Li, "RSeQC: quality control of RNA-seq experiments.," *Bioinformatics*, vol. 28, no. 16, pp. 2184–5, 2012.

[49] "Trimmomatic Manual: V0.30."

[50] S. Anders, P. T. Pyl, and W. Huber, "HTSeq-A Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[51] I. Illumina, "Sequencing Analysis Software User Guide," *Illumina Inc, San Diego, CA, USA.*, 2008.

[52] J.-W. Li *et al.*, "The NGS WikiBook: a dynamic collaborative online training effort with long-term sustainability," *Brief. Bioinform.*, vol. 14, 2013.

[53] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," vol. 30, no. 15, pp. 2114–2120, 2014.

[54] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome."

[55] C. Trapnell *et al.*, "Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms."

[56] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data."

[57] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, 2009.

[58] "Bowtie 2: fast and sensitive read alignment." [Online]. Available: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml. [Accessed: 28-Aug-2017].

[59] "iGenomes." [Online]. Available: https://support.illumina.com/sequencing/sequencing_software/igenome.html. [Accessed: 28-Aug-2017].

[60] C. et al Trapnell, "TopHat2 Manual." [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp120.

[61] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.," *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.

[62] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *J. R. Stat. Soc. Ser. A*, vol. 135, no. 3, p. 370, 1972.

[63] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1905.

[64] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson, "Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data exploration," *Bioinformatics*, vol. 24, no. 14, pp. 1650–1651, 2008.

[65] S. Bauer, J. Gagneur, and P. N. Robinson, "Going Bayesian: Model-based gene set analysis of genome-scale data," *Nucleic Acids Res.*, vol. 38, no. 11, pp. 3523–3532, 2010.

[66] S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron, "Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis," *Bioinformatics*, vol. 23, no. 22, pp. 3024–3031, 2007.

[67] S. Anders, A. Reyes, and W. Huber, "Detecting diferential usage of exons from RNA-seq data," *Genome Res*, vol. 22, no. 10, pp. 2008–2017, 2012.

[68] D. R. Cox and N. Reid, "Parameter orthogonality and approximate conditional inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 49. pp. 1–39, 1987.

[69] G. K. Smyth and A. P. Verbyla, "A Conditional Likelihood Approach to Residual Maximum Likelihood

Estimation in Generalized Linear Models," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 3, pp. 565–572, 1996.

[70] J. M. Johnson, "Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays," *Science (80-.).*, vol. 302, no. 5653, pp. 2141–2144, 2003.

[71] A. Pombo and N. Dillon, "Three-dimensional genome architecture: players and mechanisms," *Nat. Rev. Mol. Cell Biol.*, vol. 16, no. 4, pp. 245–257, 2015.

[72] Q. Jiang *et al.*, "LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data," *BMC Genomics*, vol. 16, no. Suppl 3, p. S2, 2015.

[73] "Galaxy." [Online]. Available: https://usegalaxy.org/. [Accessed: 11-Oct-2017].

[74] L. Wang, S. Wang, and W. Li, "RSeQC: quality control of RNA-seq experiments," *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185, Aug. 2012.

[75] M. Cabili *et al.*, "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses," *Genes Dev.*, vol. 25, no. 18, pp. 1915–1927, 2011.

LIST OF FIGURES

LIST OF TABLES