EVALUATING DIFFERENTIAL GENE EXPRESSION USING RNA-SEQUENCING: A CASE STUDY IN DIET-INDUCED MOUSE MODEL ASSOCIATED WITH NON-ALCOHOLIC FATTY LIVER DISEASE (NAFLD) AND CXCL12-vs-TGFβ INDUCED FIBROBLAST TO MYOFIBROBLAST PHENOCONVERSION

A Thesis Presented

by

ARPA SAMADDER

Submitted to the Office of Graduate Studies,

University of Massachusetts Boston,

In partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

May 2019

Biology Program

EVALUATING DIFFERENTIAL GENE EXPRESSION USING RNA-SEQUENCING: A CASE STUDY IN DIET-INDUCED MOUSE MODEL ASSOCIATED WITH NON-ALCOHOLIC FATTY LIVER DISEASE (NAFLD) AND CXCL12-vs-TGFβ INDUCED FIBROBLAST TO MYOFIBROBLAST PHENOCONVERSION

A Thesis Presented

by

ARPA SAMADDER

Approved as to style and content by:

_____

Todd Riley, Assistant Professor
Committee Chair

_____

Prof. Jill Macoska, Director of CPCT, Department of Biology
Member

_____

Alexey Veraksa, Professor
Member

_____

Gregory Beck, Program Director
Biology Program

_____

Richard Kesseli, Chairperson
Biology Department

ABSTRACT

EVALUATING DIFFERENTIAL GENE EXPRESSION USING RNA-SEQUENCING: A CASE STUDY IN DIET-INDUCED MOUSE MODEL ASSOCIATED WITH NON-ALCOHOLIC FATTY LIVER DISEASE (NAFLD) AND CXCL12-vs-TGFβ INDUCED FIBROBLAST TO MYOFIBROBLAST PHENOCONVERSION

May 2019

Arpa Samadder, B.S, M.S., University of Massachusetts Boston

Directed by Professor Todd Riley

Unlike the genome, cell transcriptome is dynamic and specific for a given cell developmental stage. Transcriptomics study is crucial to understand the functional elements of the genome to divulge molecular constituents of cells. The recent development of high-throughput sequencing technologies has provided an unprecedented method to sequence RNA and it has been emerging as the preferred technology for both characterization and quantification of the cell transcripts. Using "Tailor_Pipeline" we have analyzed diet-induced mouse and stromal fibroblast RNA-Seq samples and deciphers the differentially expressed genes that were significantly up- and downregulated and associated with several metabolic immune responses that presumably associated with liver disease. Analyzing the diet-induced mice model allowed us to encapsulate the transcriptional differences between diet-induced mice that can aid in the understanding of NAFLD and consequent liver pathogenesis. Identification of genes

downregulated in metabolic processes and upregulated in immune responses indicate that mice model exhibiting liver disease. Moreover, the finding of a premalignant signature suggests that NAFLD may begin to progress towards hepatocellular carcinoma much earlier than earlier consideration.

Tissue fibrosis arises due to overgrowth, scarring of various tissues and is attributed to deposition of the extracellular matrix including collagen, influenced by the actions of several pro-fibrotic proteins that can induce myofibroblast phenoconversion. Though recent transcriptomics analysis reveals cellular identity, but its ability to provide biologically meaningful insights in fibrosis is largely unexplored. To unravel the mechanisms at the genetic level, we have considered TGFβ/TGFβR and CXCL12/CXCR4 transcriptomes in human stromal fibroblasts. Transcriptome profiling technology revealed CXCL12/CXCR4 axis is responsible for the activation of COPII vesicle formation, ubiquitination, and Golgi/ER localization/targeting. Especially, identification of CUL3 and KLHL12 are responsible for the transportation of procollagen from ER to the Golgi. Interestingly, over-expression of CUL3 and KLHL12 are highly correlated with procollagen secretion by CXCL12-treated cells, but not in TGFβ-, treated cells. Moreover, this analysis showed how activation of the CXCL12/CXCR4 axis promotes procollagen I secretion that responsible for the deposition of ECM which is a characteristic of fibrosis.

ACKNOWLEDGEMENTS

I start by thanking my supervisor, Professor Todd Robert Riley, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my M.S study. I'm eternally grateful to my mentor Professor Todd Robert Riley, for unparalleled mentoring. As my mentor, he has taught me more than I could ever give him credit for here. He has shown me, by his example, what a good scientist (and person) should be. Besides, my advisor, I would like to thank the rest of my thesis committee: Prof. Jill Macoska and Prof. Alexey Veraksa for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives. I am sincerely grateful to Prof. Macoska who provided to me a great opportunity to work in these two projects.

A special thanks to Prof. Richard Kesseli for giving me all the support, advices and motivation I could possible need.

I thank my fellow lab mates in for the stimulating discussions. I am grateful to Andrew Judell-Halfpenny for his companionship along not only the development of this thesis but throughout the entire Bioinformatics journey.

I express my deeply gratitude to my family, specially, to my parents, for all the support and comprehension. Thank you for always encourage me to give my best in every situation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Context and motivation

The chemistry Nobel Prize was awarded to Fred Sanger and Walter Gilbert, in 1980, for their crucial contribution towards the determination of base sequences in nucleic acids (Sanger et al., 1977 & Gilbert et al., 1973), since then many were involved in the developments of DNA sequencing technologies (Fiers et al., 1976). Development of these techniques conveyed a modernized approach to biological questions and, high throughput sequencing technologies such as RNA-Sequencing (RNA-Seq) has revolutionized the post genomic era, become an integral part of the biological research to access the cell transcriptome (Mardis, 2008).

Bioinformatics is an interdisciplinary field that integrates computer science and biology to research, develop, and apply computational tools to manage and process large scale of biological data (Hogeweg and Hesper, 1978). Particularly, these computational tools are suitable to analyse data generated from high-throughput sequencing platforms. Consequently, the success of next-generation sequencing (NGS) technologies is strongly related with the creation of competent computational tools to deal with the dramatic increase of data (Shendure and Ji, 2008).

Until the mid-1990s, gene expression studies were limited to measure transcription of few genes. But microarray technology changed this and allowed the study of hundreds or thousands of transcripts at a time. At that time, this technology revolutionized many areas of biology, from basic research to the understanding and treatment of human disease (Schena et al., 1995). In an analogous way, the recent availability of next-generation sequencing (NGS) analysis has opened new horizons to address the gene expression analysis, where initially NGS applications were mainly focusing on the sequencing of genomic DNA, this technology is now finding its way to be used in transcriptomics studies (Westermann et al., 2012).

An important biological aspect in recent year is to understand the complex and sophisticated mechanisms where deposition of the extracellular matrix (ECM) leads to develop progressive aging- and inflammation-associated fibrosis. Now tissue stiffness and urethral disfunction are due to the accumulation of ECM that leads to reduce tissue flexibility that lead to urinary flow block and development of lower urinary tract symptoms (LUTS). In this study, we have tried to encapsulate whether senescence-accelerated mouse prone (SAMP6) mice would also develop LUTS. Also, we have tried to see whether diet-induced obesity and type 2 Diabetes Mellitus (T2DM) have any influential role to propagate this disease. A global transcriptomics analysis can provide new insights into the disease process, leading to the identification of known and unknown transcripts, and overall gene expression regulation of different pathways and how they differ between the different samples under different diets (Gharaee-Kermani et al., 2013).

Dynamic remodeling of the extracellular matrix (ECM) deposition and gain an understanding of their role in fibrosis is a challenging factor recently. Fibrosis characterizes a contributing factor to the etiology of LUTS (Gharaee-Kermani et al., 2013). Several studies have shown that the aging prostate tissue is rich with inflammatory cells microenvironment and proteins. It is still unclear whether these inflammatory proteins, particularly CXC-type chemokines, can mediate fibroblast to myofibroblast phenoconversion and the ECM deposition which necessary for the development of prostatic tissue fibrosis (Rodríguez-Nieves et al., 2016). In human stromal fibroblast, we are trying to determine the effect of TGFβ/TGFβR and CXCL12/CXCR4 transcriptomes and find out the difference between them. In addition to this we aimed to find out any significantly differentially expressed transcripts including coding and non-coding mRNAs that may promote myofibroblast phenoconversion.

1.2 Problem formulation

This thesis has two underlying goals that are complementary to each other: the first is related to computational methodologies and the second to biologic knowledge.

- Computational goal

  Firstly, I aim to integrate available bioinformatics tools in a congruent pipeline that can process RNA-Seq data and extract reliable biological conclusions from it.

- Biological goal

  The second main goal is to use the developed tools to comprehend in which way NAFLD (Non-Alcoholic Fatty Liver Disease) is transcriptionally regulated. Furthermore, it has been well informed that tissue fibrosis is mediated by the actions of multiple pro-fibrotic proteins that can induce fibroblast to myofibroblast phenoconversion. This occurs through various signaling pathways such as Smads or MEK/Erk proteins. Apart from that the TGFβ/TGFβR and CXCL12/CXCR4 axes persuade myofibroblast phenoconversion independently through Smads and MEK/Erk proteins, respectively. To investigate these mechanisms at the genetic level, we aim is to elucidate the TGFβ/TGFβR and CXCL12/CXCR4 transcriptomes in human fibroblasts.

1.3 Thesis Outline

Apart from this introduction, this thesis is structured in six chapters.

In chapter 2 I introduce the concept of gene expression, its main regulation points and the several used approaches to have insight into this information. Particularly, in section 2.4, I give special relevance to RNA-Seq data and the current methods that are used to access the gene expression profile and extract novel biological knowledge from this type of data.

Chapter 3 presents a review on the diet-induce obesity mouse model to characterize the transcriptional landscape of NAFLD and compare it to the transcriptional signature of healthy control mice. More importantly, identification of the transcriptional signatures helps to detect of these diseases through the identification of novel markers. Moreover, in this

chapter, it is described the RNA-Seq mouse dataset that will be used in this thesis as a case study to test the developed ethodologies.

Then, in chapter 4 I describe a pipeline developed to analyze an RNA-Seq dataset. As a case study, I used the implemented pipeline to process an RNA-Seq dataset extracted from SAMP6 strain mouse fed with high fat diet (HFD) and low-fat diet (LFD). From its output I conclude whether there are any significant transcript differences between the two phenotypes, up-regulating inflammation-related processes and down-regulating metabolism related processes in HFD-fed mice compared to LFD-fed mice. This analysis is described in section 4.2. Then in section 4.3 I have discussed the detailed analysis of human fibroblast cell line where we have elucidated the TGFβ/TGFβR and CXCL12/CXCR4 transcriptomes in human fibroblasts. From the output I conclude the biological significance about the fibroblast to myofibroblast phenoconversion.

Afterwards, in chapter 5, I have described the detail analysis of two case studies done by using our recently developed pipeline known as Tailor_Pipeline.

Then, in chapter 6 I have discussed the overall journey of this analysis and how this pipeline helps to analyze the raw RNA-Seq data to extract the biological significance.

Finally, in chapter 7 I have concluded the important aspects of the analysis.

CHAPTER 2

GENE EXPRESSION

2.1 Introductory note

Eukaryotic organisms have its genetic information encoded in molecules of deoxyribonucleic acid (DNA) which are packed and organized into the cell nucleus in structures called chromosomes. The monomers of DNA are called nucleotides and they are organized in a double-stranded helix. Nucleotides are comprised by a phosphate group, a 5-carbon sugar, and a nitrogenous base. The genetic information in a DNA molecule is represented by the sequence of nucleotides containing one of four types of nucleobases: adenine (A), guanine (G), cytosine (C) and thymine (T).

Following the Watson - Crick Model (Watson and Crick, 1953), the two strands that constitute the DNA molecule are held together by hydrogen bonds that can only be established between specific pairs of nucleobases: A with T and G with C. Because of this restriction, both strains are complementary to one another and, therefore, contain the same genetic information.

2.2 The concept of gene expression

In 1958, the central dogma of molecular biology was firstly proposed by Francis Crick (Crick, 1958; Crick 1060). Particularly, central dogma states that information in nucleic acids can be transferred (Fig-1). Gene expression is the process by which a segment of DNA is copied into a ribonucleic acid (RNA) molecule which, in turn, will be used in the synthesis of functional gene products. Some RNA molecules can be the end product in themselves and some can be used as a template for the creation of other molecules such as proteins, in a process called translation. According to this distinction, RNAs can be classified as either messenger RNAs (mRNAs) or non-coding RNAs (ncRNAs). The genetic information is

transferred into an RNA molecule is designated by transcription and completed in the cell nucleus by an enzyme called RNA polymerase (RNA pol).

The RNA polymerase catalyzes and forms the phosphodiester bonds that link the nucleotides together and form the sugar-phosphate backbone. In eukaryotes, there are multiple types of RNA polymerases (RNA pol) that synthesize various types of RNA. Firstly, RNA pol I transcribe ribosomal RNAs (rRNAs) which associated with ribosomal proteins, on which mRNA is translated into protein. Secondly, RNA pol II transcribes mainly protein-coding genes (mRNAs). Finally, RNA pol III catalyzes the transcription of transfer RNAs (tRNAs), which function as adaptors selecting amino acids and holding them in place on a ribosome for their incorporation into protein. As we have seen, DNA to protein synthesis occurs in a two-step process. In the first step DNA to mRNA synthesis is called transcription. In the second step, called translation where the information in the mRNA is translated into protein.

The transcription is a process of formation of the transcript (RNA). It takes place by the usual process of complementary base pairing, catalyzed and scrutinized by the enzyme RNA polymerase. It occurs unidirectionally in which RNA chain (transcript) is synthesized from 5′ to 3′ direction. Initiation, elongation, and termination are the three steps of the gene transcription process. Initiation process begins when the RNA pol molecule binds to the upstream region of the DNA at a specialized sequence called promoter. To occur this binding, RNA polymerase requires the involvement of many accessory proteins such as transcription factors (TFs). The transcription initiation complex can be formed by the combination of the transcription factors and RNA polymerase and this complex is responsible to initiate transcription. The RNA polymerase started to synthesize mRNA by corresponding complementary bases to the original DNA strand. These must assemble on promoter along with the polymerase before the polymerase can begin transcription. Once transcription is initiated, most of the TFs are released from the DNA. Elongation involves the movement of the transcription bubble by disruption of DNA structure. The enzyme moves along the DNA

and extends the growing RNA chain. As the enzyme moves, it unwinds the DNA helix to expose a new segment of the template in single-stranded condition and added nucleotides to the 3´ end of the growing RNA chain. Finally, transcription termination occurs after RNA pol reaches a termination site. At this point, RNA pol is released from the DNA and RNA is cleaved and released from the transcriptional complex.

Simultaneously to the transcription process, the translation takes place in the cytoplasm. The mRNA molecules undergo little or no modification after synthesis by RNA polymerase in prokaryotes. In contrast, processing of eukaryotic pre-mRNA involves 5’ capping, 3′ cleavage/Polyadenylation, splicing, and RNA editing before being transported to the cytoplasm where they are translated by the ribosome. Polyadenylation is an important RNA processing step where a long chain of adenine nucleotides is added to a messenger RNA (mRNA) molecule to increase the stability of the molecule, aiding its exportation from the nucleus to the cell cytosol. In this process, a series of repeated A nucleotides – poly-(A) tail – are added to the 3′ end of the pre-mRNA molecule. First, the 3' end of the transcript is cleaved and generate a 3' hydroxyl and poly-A polymerase (PAPs) adds a chain of adenine nucleotides up to 250 residues to the RNA. The poly-A tail provides stability of the RNA molecule and prevents its degradation.

Furthermore, in the cell nucleus, the newly synthesized RNA molecules require extensive processing to become a functional RNA. In most eukaryotic genes, noncoding DNA is also found.

Such genes have split structures in which segments of coding sequence (called exons) are separated by noncoding sequences (intervening sequences, or introns). In the mRNA, the introns are then removed by splicing and yield a long RNA molecule which possess only exonic part. In a process designated by splicing, introns are removed from the mRNA molecule and then neighboring exons are stitched together. This process is exclusive of eukaryotic organisms.

The mature RNA is then selectively transported from the nucleus to the cytoplasm where mRNAs are involved in the translation process, where the information in an mRNA molecule is converted into a protein. In this process, an mRNA molecule is used as a template by a ribosome, which will match each sequence of three nucleotides (codon) on the template mRNA chain with a sequence of three complementary nucleotides (anti-codon) on a tRNA molecule. Bearing in mind that each tRNA has connected to an amino acid that its anti-codon sequence calls for, this molecule will recognize and bind to a codon at one site and to an amino acid at another site of its surface. Thus, tRNAs function as translators between nucleotide sequences in RNAs and amino acid sequences in proteins. The ribosome, as the mRNA moves through it, covalently links each amino acid to the end of the growing polypeptide chain by peptide bonds. When the translation reaches a Stop codon, denoting the end of the protein, the completed protein chain and the mRNA molecule are released, and the ribosome is dissociated into two separated subunits.

Therefore, gene expression can be seen, as a mediator that interprets the genetic information of an organism (genotype) that gives rise to an outward physical manifestation (phenotype), via gene transcription and mRNA processing.

2.3 Gene expression regulation:

Given that genes encode for proteins and proteins dictate the function of the cell and their structural proprieties. Each step of the gene expression is associated with the flow of information from DNA to RNA to protein that provides the cell with a probable control point for self-regulating that associated with its function. This allows cells to respond to maintain their cell-type specific expression patterns.

In this way a cell can regulate the amount and type of proteins that it is manufacturing by several key factors:

1. Required controlling when and how often a given set gene is transcribed;

2. To control processing of an RNA transcript;

3. To select which mRNAs are exported from the nucleus to the cytosol;

4. Degradation of certain mRNA molecules;

5. Selecting for which mRNAs are translated by ribosomes;

6. Selectively activating or inactivating proteins after they have been synthesized;

7. Controlling of mRNA degradation.

2.3.1 Transcriptional regulation

Transcriptional regulation plays a paramount role in controlling gene expression. During this process no, unnecessary intermediates are synthesized. This regulation can be executed at the promoter level by the association of TFs to the gene promoter region. As referred in section 2.2, the establishment of this connection will help to bind the RNA pol to initiate translation process. In the promoter region, nearly all genes are controlled by regulatory DNAs that may increase or decrease the activity of transcription of a certain gene. Now enhancers are sequence-specific TFs generally bind to these regulatory DNA regions and can control the switching on or off a gene, respectively. Often, the sequence specific factors and the general TFs accumulated in the promoter region and interacted via additional proteins named as co-factors. Rate of regulation of gene transcription is controlled by aiding/preventing the assembly of the general TFs and RNA pol at the promoter region (Kreimer and Pe'er, 2013). To bind the TFs and the RNA pol to the regulatory regions of the gene, the DNA chain needs to be accessible.

Hence, the activity of transcriptional regulation can be also influenced by the level of DNA packaging. DNA is usually densely packed with histones, forming a closely packed structure called chromatin. Chromatin construction allows access of condensed genomic DNA to the regulatory transcription machinery proteins, and thereby controls the efficiency to initiate the transcription initiation. (Oberdoerffer et al., 2008). After transcription initiation, the activity rate of the RNA pol II enzyme is decreased and paused on a promoter proximal position.

From this stage, depending on the type of transcription elongation factor that interacts with the RNA pol II, transcription may halt or enter elongation phase (Dvir et al., 1997).

2.3.2 Post-transcriptional regulation

During RNA synthesis, post-transcriptional regulation controls the gene expression. It contributes considerably to regulate gene expression across human tissues. The process of polyadenylation, introduced in section 2.2, influences the transcripts lifetime, protecting them from degradation and aiding their exportation to the cell cytosol.

In a similar way, modulating the capping, splicing, addition of a Poly (A) tail where a modified guanine nucleotide cap is added to the 5′ end of pre-mRNA molecules is crucial for the novel transcript to exit the cell nucleus. Therefore, both these processes are essential for the stability of the mRNA molecule into an ideal time-window. The process of splicing, also referred in section 2.2, enables the production of mature messenger from a newly made precursor messenger RNA (pre-mRNA) transcript. During RNA splicing the introns are precisely excised and the exons are ligated together. The majority of nuclear pre-mRNA is spliced constitutively; that is, only one mature mRNA species is generated from a single pre-mRNA in all tissues. In some cases, however, alternative 5′ and/ or 3′ splice sites are used during splicing, resulting in the production of more than one mRNA species from a single pre-mRNA. The production of different RNA products from a single product by changes in the usage of splicing junctions is known as alternative splicing. During this alternative splicing, the alternative 5′ and/or 3′ splice sites can result in structurally distinct mRNAs by either excluding potential exon sequences or incorporating otherwise noncoding introns sequences.

2.3.3 Translational regulation

Translation takes place in the cytoplasm. Some parts of the cytoplasm are so tightly packed with the soluble protein and cytoskeleton that ribosomes can be expected to have difficulties diffusing into them. This is usually performed by biding a repressor to the 5′

untranslated region of the mRNA, which helps to guide the ribosome to the mRNA start codon. The ribosome is, thereby, kept from finding the translation start site. When conditions change, the cell can inactivate the repressor and increase translation of the mRNA. Regulation of the rate of protein synthesis is involved by the influencing the rate-limiting steps of the translational steps. Now, this process can be accomplished by the involvement of ribosomes or initiation factors. Generally, cytoplasmic mRNAs are actively translated by ribosomes to form messenger ribonuleoprotein particles, mRNP. Translational initiation process involved by utilizing two subunits: eIF2 and eIF4E. eIF4A, eIF4G and eIF4F are other subunits, involved in the initiation of translation process. In most cells, the availability of the eIF-4E which is a cap-binding protein is the rate-limiting factor involves initiating translation. Therefore, regulation of eIF-4E levels is important to control the rate of translation.

2.3.4 Protein degradation

Once protein synthesis is complete the level of expression of that protein can be reduced by protein degradation. Cells possess specialized pathways to degrade proteins, using enzymes designated by proteases. In these pathways, proteins which lifetime must be short, or which are damaged or misfolded are marked by the attachment of a small protein called ubiquitin. Ubiquitylated proteins are then recognized and destroyed.

CHAPTER 3

GENE EXPRESSION ANALYSIS BY RNA-SEQUENCING DATA

3.1 Approaches for genome-wide expression analysis:

It is worth mentioning that high-throughput sequencing becoming a prime choice to measure the gene expression to get an insight about the transcriptional behavior of biological systems. Therefore, identification of differential gene is an important paradigm that is used in many areas of biology and medicine. It can be employed to identify significantly differentially expressed genes between two or more biological conditions of interest (Schena et al., 1995). To classify heterogeneous diseases such as cancer, differentially gene expression analysis plays a pivotal role (Bhattacharjee et al., 2001). This analysis is also important to understand the relation between genes profile and survival or tumor aggressiveness (Veer et al., 2002). To discover new drugs (Pagliarulo et al., 2002), diagnose diseases (Heller et al., 1997), differentially gene expression is important (Thiery et al., 2006).

Gene expression analysis can be divided into two parts namely genome-wide and target-based approach, depending on what it is anticipated to study. In the absence of any key genes of interest, the data is acquired at the biological system level. Therefore genome-wide approaches such as microarrays (Augenlicht and Kobrin, 1982) or RNA-Seq (Mortazavi A et al., 2008) technologies have emerged as a powerful technology for the detection of differential gene expression, that enables to quantify the frequency of RNA species in a certain biological system. Transcriptome profiling which is defined by the complete set of transcripts in a cell and their amount at a definite acquisition point is the main approach to measure the differential gene expression. As stated in section 2.2, knowledge of the transcriptome is very useful to provide a link between change of expression of a gene and their phenotype presented by the cell (Wang et al., 2009). On the contrary, polymerase chain reaction (qPCR) should be employed when the genes of interest are already known (Murphy

L D et al., 1990). The main purpose of this reaction is to perform the gene expression analysis. Here I have mainly focused on genome-wide GEA, particularly those using RNA-Seq technology and explain the pros and cons associated with this analysis. Furthermore, I will also include the brief RNA-Seq protocol and revisit the current methodologies which are associated to assess from the raw nucleotides sequence and their active cellular processes upon collection of the transcriptome.

To perform transcriptomics analysis an ample number of technologies have been developed over the years. Out of all these methods, High-throughput sequencing technology has endowed with an unprecedented aspect about the transcriptional landscape of an organism and becoming the paradigm to measure RNA expression levels. With the dawn of sequencing technology, it's now feasible to profile gene expression levels in every field in life sciences and becoming prevailing technique for clinical use. Generally, one of the main goals of this experiment is to identify the differential gene expression, gene isoform, post transcriptional modifications and so on to understand phenotypic variation (Rapaport et al., 2015).

Understanding the large-scale studies of gene expression levels, a microarray was a tool to detect the gene expression in the 1990s. Concurrently, the process of measuring the gene expression, microarray can provide a picturesque of transcriptional activity in a wide range of biological problems, including identification of differentially expressed genes between diseased and healthy tissue (Zhao et al., 2014). Currently, DNA microarrays are a relatively inexpensive and can afford many laboratories for transcript profiling. In microarray, a short single stranded DNA molecule, called probes, are attached to fixed locations on a solid substrate. Then RNA molecules (transcriptome) are extracted from the sample and copied into complementary DNA (cDNA) with the help of reverse transcriptase. Fluorescent dye was used to label in the cDNA. Finally, cDNA is passed over the solid support and complementary sequence will tend to hybridize. Then expression is quantified by

using a fluorescence scanner that measures the amount of fluorescence coming from each probe on the slide (Hoheisel, 2006).

In biological samples, gene expression microarray profiling endowed with precise determination of expression levels of genes in a single hybridization experiment. Identification of nearly 57000 citations by using a simple search for the term “microarray” in PubMed database shows its consequence for assaying gene expression. However, the power of this technology has several drawbacks. For instance, due to cross-hybridization, the expression measurements have high background levels. Therefore, the probe sequences must be pre-specified so that priori the sequences can be identified. Additionally, due to both background and saturation signals, the accuracy of measurement of expression is limited (Okoniewski and Miller, 2006). In order to overcome this limitation probes should be used that can differ in their hybridization properties. Otherwise it is unreliable to compare the same array between different genes (Gautier et al., 2004). Therefore, it’s crucial to maintain the experimental design to perform successful microarray experiment. In order to perform successful experiment, sometimes a major question whether the microarray experiment should be performed using the single-color or two-color to compare the relative gene expression. Until date lot of articles have been published reading this issue saying that single-color arrays are more flexible in analysis compare to the two-color. Anyway, in contrast to the microarray technology, some sequence-based methods are also important to determine the cDNA sequence. Previously, Sanger sequencing was performed to determine the cDNA sequence (Sanger et al., 2004). However, Sanger sequencing is expensive and have relatively low throughput and generally is not quantitative (Wang et al., 2009). To overcome these limitations, tag-based methods were developed namely serial analysis of gene expression (SAGE) (Velculescu et al., 1995), cap analysis of gene expression (CAGE) (Shiraki et al., 2003) and massively parallel signature sequencing (MPSS) (Brenner et al., 2000). However, SAGE based technology doesn’t measure the actual expression level of a gene. During SAGE

analysis, short tag (ten bases) has been produced, makes the analysis hard to assign a tag to a specific transcript with accuracy because these short tags can be mapped to more than one place in the reference genome, allowing an ambiguous identification of transcripts. Sometimes, same tag possesses with the two different genes and the alternatively spliced gene could have different tags at 3′ ends.

Finally, with an emergence of high-throughput sequencing technologies have overcome the limitations of both microarrays and tag-based methodologies (Church, 2006). Specially, RNA-Seq is a transcriptome profiling technology and NGS platform for differential gene expression (Mardis, 2008). Particularly, RNA-Seq technology is more reliable to arrays and employed for both mapping and quantifying transcriptome across all cell types, perturbation and states (Roberts et al., 2011). RNA-Seq technology is more persuasive and quantifies the expression of novel transcript over a wider dynamic range which is not possible to quantify in array-based technology (Marioni et al., 2008). Due to hybridization-free approach, this technique has been widely used in an integral part of microbiological research (Mardis, 2008). Additionally, RNA-Seq technology can be used to detect of gene fusion events (Maher et al., 2009), detection of single nucleotide polymorphisms (Mardis, 2008), investigation of post transcriptional RNA mutations (Garcion et al., 2004), study of alternative splicing events (Pan et al., 2008), discovery of novel transcripts (Guttman et al., 2010; Degner et al., 2009). Of course, in near future, the probable technical goal is to sequence and count entire mRNA molecules known as single-molecule sequencing which enable to quantify even single cells.

3.2 RNA-Sequencing experiment workflow

In all living organisms, RNA molecules are crucial components and several high-throughput sequencing techniques are existing to interrogate of RNA sequences on a large scale. Currently, 454 GS-FLX from Roche 454 Life Science, Genome Analyzer II from Illumina, Inc. and AB SOLiD from Applied Biosystems are believed to be the foremost

method in expression analysis. However, different technologies require different experimental procedures for sequencing study. In principle, any high-throughput sequencing technology can be used for the RNA-Seq study and Illumina's machines have already been applied for the purpose (Bennett, 2004). In RNA-Seq study, after conversion to a library of cDNA fragments with adapters attached to one or both ends and sequenced in a high-throughput way to get the short reads (Wang et al., 2009). The resulting reads are either mapped to a reference genome or de novo without genomic sequence to get the genome-scale landscape of transcriptional or post-transcriptional gene expression.

Now, RNA-Seq faces several technological challenges like other high-throughput sequencing technology such as data storage capacity, process large amounts of data. Therefore, these challenges should be overcome to reduce errors by removing low-quality reads, improvement in base-calling. Despite the challenges, RNA-Seq has facilitated us to make an unprecedented large-scale overview of the transcriptome. Keeping in mind, RNA-Seq revealed many novel transcribed regions, splicing isoform for many genes. In this thesis, I have mainly focused on the pipeline which is appropriate for NGS data generated from the Illumina platform. Additionally, I have also described the details analysis which I have performed through the pipeline. Particularly, the approach of sequencing in the Illumina machine comprises the following fundamental steps:

1) Informative RNA enrichment

An archetypal RNA-Seq experiment begins by purifying a subset of RNAs from the total RNA to analysis of transcriptome expression. Particularly, for protein coding RNAs, this enrichment analysis is carried out by selecting poly(A)+ molecules using oligo-dT associated with magnetic or cellulose beads. It is worth mentioning that prokaryotic mRNAs, poly(A)-transcripts in eukaryotic cells are frequently subject to exploration (Hrdlickova et al., 2017). Additionally, in cells the most abundant RNA is rRNA, >90% present in cells consists on rRNA (Wilhelm and Landry, 2009) and need to be depleted due to small interest in most

studies (Tariq et al., 2011). Though, varieties of selection processes have been developed recently for rRNA depletion, oligo-dT based purification of poly (A)+ RNA is the prime method that ensures to get a strong signal for the RNA population of interest.

2) RNA fragmentation

Still now, RNA fragmentation is the most commonly used technique in RNA-Seq library preparation. Before reverse transcription (RT) process, RNA samples are subjected to fragmentation process to get a certain size range. This fragmentation process happened after selection of poly (A)+ or rRNA depletion. Due to the limitations of the size, fragmentation process is mandatory in the most sequencing platforms. After purification, the larger RNAs are fragmented by using RNA hydrolysis or nebulisation. On the other hand, full length double-stranded cDNA can be fragmented by DNaseI treatment or sonication. Now cDNA fragmentation is more likely towards the 3′ end of the transcript.

3) Synthesis of double stranded cDNA

In RNA-Seq, sequencing of poly (A) RNAs is the most common application unless a very small amount of RNA is accessible. In eukaryotes, most protein-coding RNAs contain poly (A) tail. The RNA fragments are converted into cDNAs using reverse transcriptase enzyme which requires the hybridization of primers into the RNA chain. These primers can be oligo-dT or sequence of random primers. For most protein-coding RNAs (mRNAs) and many long noncoding RNAs (lncRNAs), it is not advisable to use the oligo-dT primers due it's biasness on the 3′ of the transcript. As a result, the sequencing reads will be enriched for the 3′ ends of the transcript (Wilhelm and Landry, 2009). Therefore, random primers are preferred to use which have the potential hybridization capacity to random sites of the RNA molecule. Finally, after synthesizing the first strand of cDNA, the RNA template is eliminated and generated a second cDNA by using DNA pol I and finally a double stranded cDNA molecule is generated. Therefore, poly (A) purification is a preferred method to select poly(A) + RNA.

4) Adapters ligation

In a benchmark of RNA-Seq library etiquette, a desired size of cDNAs has been generated through reverse transcription (RT) of fragmented RNAs with random hexamer primers. Before amplification and sequencing process, a required extent of cDNAs has been also generated based on the fragmentation of full-length cDNAs that are ligated to DNA adapters. During adapter ligation process, 3′ of the cDNA overhangs are switched into blunt ends by a specialized enzyme. Next a series of 3′ ligations occurred by using a truncated RNA ligase II whereas 5′ adapter ligation happened by using RNA ligase I (Hrdlickova et al., 2017). In order for ligation, the cDNA fragments to the adapter, an A base is added to the 3′ depleted end which contain a single T base over-hanged at their 3′ end. Finally, distinctive adapters are ligated to each strand of 3′ ends of the double-ended cDNA.

5) Size selection and PCR amplification

During fragmentation step, DNA molecules are divided into two different sizes. By gel extraction, a desired range of DNA length is purified to ensure that all molecules are of similar length. Furthermore, this procedure eliminates unligated adapters as well as those ligated to one another. Finally, two primers are annealed to the adaptors tail followed by amplification by PCR of the purified cDNA.

6) Cluster generation

Inside the flow cell, single-stranded DNA templates are bridged-amplified to form clonal clusters prior sequencing. During the PCR amplification, the double stranded molecules need to be denatured into single strands molecules. Subsequently, by using the high density of immobilized forward and reverse primers, the DNA templates are hybridized to a slide. Now, DNA polymerase is used to copy the templates from the hybridized primers. After the denaturation, the original templates disappear the copies immobilized on the flow cell surface. After fixation process, the immobilized copies are hybridized to adjoining primers. Then DNA polymerase copied the templates and finally formed double stranded

DNA bridges which in turn are denatured and formed two single-stranded DNA templates. Finally, using the base cleavage the reverse DNA strand is removed and the immobilized 3′-ends of the forward strand are prohibited to prevent interference in the sequencing process. This procedure is repeated to create a dense clonal cluster that contains at least 1000 molecules per cluster.

7) Sequencing-by-synthesis

High throughput sequencing has been started with the sequencing hybridization primers which added each single-stranded molecule in the clusters. Then DNA-templates are simultaneously reversing complemented by using fluorescent-labeled nucleotides. After addition of nucleotide, clusters are excited by a laser which causes fluorescence at the last integrated base. The cycle is repeated to remove the fluorescent dye and blocking group. The cycle is generated a sequence of images, containing new incorporated nucleotide where the fluorescence labeled signal of each cluster is captured. As a result, the color of the lighted spot represents a different base type. Then a sequence of nucleotide for each cluster can be obtained by combining the attaining the sequence of images. Finally, this information is saved in a text file named as FASTQ file format. FASTQ file contains a unique ID to identify the read, the sequence of nucleotides and the quality scores. Due to the ubiquitous nature in illumina, FASTQ format has become de facto for NGS analysis. Now to understand the read quality at each base is defined by the Phred score that can range from 0 to 60 on a logarithmic scale. The Phred quality score is defined as Q=-10log(e).

NGS can read sequence from both ends of a single DNA and generate “pair-end reads”. FASTQ files are always pre-processed to check the quality controls and remove any adapters in the sample preparation process. Sometimes contamination can be detected by the distribution of *k*-mers which help to detect the contamination. These all help to detect the potential pitfalls before the downstream analysis.

3.3 Quality control of sequencing reads

Next-generation sequencing (NGS) technologies have drastically broaden the area of genomic research. High-throughput sequencing technology can generate enormous amounts of data in a single sequencing run. To extract biological conclusions by analyzing acquired sequence, it is important to assess the library quality as well as the sequencing performance. Therefore, for any alignment process, the low quality of reads should be removed (Levin et al., 2010). Due to a range of artifacts generated during library preparation, NGS can be adversely affected the downstream analyses.

Until recently, to highlight the quality score of the NGS data, several software tools have been developed. Contamination with adapter sequences and biases in base composition are the primary reason to generate the low-quality base (Trivedi et al., 2014). To assess the quality of the raw reads generated by the sequencing platform is the foremost step of the quality control (QC) process and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) is a popular tool for this analysis. In NGS platform, FastQC is useful for considering the overall quality score of a sequencing run and commonly used as an initial QC checkpoint. Given a set of raw sequencing reads, the main aim of quality control (QC) is to align the reads to a reference genome and consider the quality of the alignments. The quality of alignments can be obtained by several metrics such as depth of coverage, contamination of rRNA, continuity of coverage, and GC bias (Andrews, 2010). The main purpose of performing quality control is to process raw sequence data coming from high throughput sequencing prior to aligning against a reference genome. Now to evaluate the quality of the high throughput sequencing reads, some valuable information needed to be extracted from the variation of the Phred quality score (Q score) of a sequencing platform (Yang et al., 2013). The content of bases has very little difference between different bases during sequencing. Now the number of bases added when the

sequencer is unable to produce any base call with enough confidence of reads length (Ewing et al., 1998).

Based on this type of analysis, the low sequencing quality bases should be eliminated to ensure the quality of the high throughput data. From the high throughput sequencing, the cellular activity is characterized and going to be extracted the biological conclusions.

3.4 Mapping reads

After removal of abnormal reads from the raw cDNA sequence reads, it is mandatory for the short sequenced to be mapped to a reference genome or transcriptome. The main goal of this step is to find the genomic location of each transcript sequence on a given reference genome. According to Fonseca et al. this problem can be achieved by computationally (Fonseca et al., 2001). This helps to match the reads with the reference genome and this can be challenging because sequencing generated millions of short reads that needed to be mapped to reference genomes that usually very large. So, it is important that mapping algorithms needed to be extraordinarily competent and used processors and memory in a most advantageous way. Moreover, 50% repetitive sequences present in complex organisms such as human or mouse genome, so it is another challenging aspect in next-generation sequencing that needed to be handled. Therefore, mapping tools are required to handle this multiple mapping locations (Fonseca et al., 2001).

(a) Alignment of a paired-end read to the genome. The read needs to be mapped across introns (spliced alignment)

(b) Alignment of a paired-end read to the transcriptome

Figure 3.1. Mapping of a paired-end read with different reference files: genome and transcriptome. Adapted from Trapnell and Salzberg, Figure 2 (2009).

Sometimes, aligning reads against the reference genome is slower because it considers all non-coding positions. Currently, several alignment programs are available to handle spliced alignments, including TopHat2 (Kim et al., 2013), SOAPSplice (Huang et al., 2011),

Blat (Kent, 2002) or Exonerate (Slater and Birney, 2005). On the other hand, Bowtie 2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009), MAQ (Li et al., 2008) or SOAP2 (Li et al., 2009) are specialized for aligning short reads to a reference genome.

In computational biology, Bowtie 2 is used as short-read mappers that can able to index a file and speeded up the mapping process. This mapping process allows an efficient and relatively small memory (Langmead and Salzberg, 2012). Particularly, Bowtie 2 indices are based on the Burrows-Wheeler Transform (BWT) that helps to keep the memory low (Kent, 2002). In human genome, this transformation keeps the memory to fit in 3.5 gigabytes. The alignment process is extracted raw NGS read which are likely to be matched in the genome with the Burrows-Wheeler Transform (BWT) based methodology (Burrows and Wheeler, 1994). Finally, each aligned character slender the list of possible mapping or genomic positions. Sometimes seed placement will be prioritized if Bowtie 2 cannot find a location where the read align perfectly (Langmead and Salzberg, 2012). Using the Single Instruction Multiple Data (SIMD)-accelerated dynamic programming algorithm, we can check whether sufficient numbers of alignments are examined (Slater and Birney, 2005; Trapnell and Salzberg, 2009; Langmead et al., 2009).

The output file from the mapping is a SAM format file. SAM file contains all the information of overlapped and non-overlapped reads. Particularly, overlapped reads contains the information about the genome location where the read was mapped and it's respective score (Li et al., 2009).

3.5 Expression quantification and normalization

During mapping of the RNA-Seq reads, it's needed to convert the data into a quantitative measure of gene expression. Several approaches are available now-a-days, but in this problem, the easiest approach is adding up the number of reads which lie within the location of each element (Van Verk et al., 2013; Wilhelm and Landry, 2009). It is easy to extract this information if reads were overlapped to the transcriptome otherwise gene

expression measurement can be performed by using Cufflinks package if reads were aligned to the genome (Trapnell et al., 2012). In RNA-Seq, estimation of gene/transcript expression is predominantly relying on the no. of reads that aligned to each transcript sequence. There are several algorithms available recently for transcript/gene mapping. One such algorithm known as Sailfish that mainly depends on the on *k*-mer read counting without the need for mapping (Conesa et al., 2016). However, gene expression measurement can be quantified by using HTSeq package that enable aligned reads to the genome (Anders et al., 2015). This quantification process uses GTF file that contains genome coordinates of exons and genes. Now to compare the expression levels among samples, transcript length, and total number of reads affect read count.

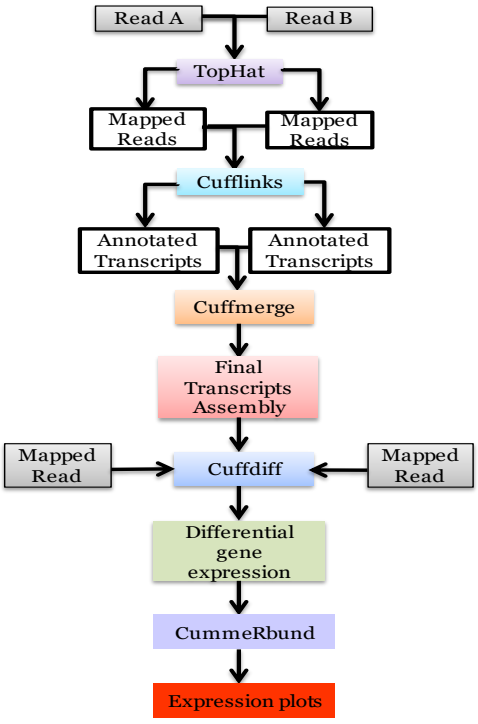

Fig 3.2. An overview of the tuxedo protocol. The assembly and characterization of expressed genes from the experimental data, statistical analysis of differential gene expression. QC study is performed in the raw RNA-Seq reads using FastQC; the filtered reads are then mapped to a reference file using Bowtie 2; from this data is measured the gene expression level and performed a differential expression test by Cuffdiff; the genes are then concatenated into GO terms using GOStats. In the end, this will give insight into the biological processes that are differentially active between the conditions in comparison. Adapted from Trapnell et al., Figure 2 (2012).

To remove this biasness, RPKM (reads per kilobase of exon model per million reads) normalizes a transcript's read by both its length and the total number of reads mapped in the

sample to remove the feature-length and library-size effects. FPKM (fragments per kilobase of exon model per million mapped reads) normalized paired-end data (Mortazavi et al., 2008). During the RNA-Seq experiment, two conditions will be examined where reads are first mapped to the genome with TopHat. These mapped reads are fed to Cufflinks, which produces one file of assembled transfrags for each replicate. Finally, Cuffdiff analysis performed to get the differential gene expression analysis. These files are visualized with CummeRbund to facilitate exploration of genes identified by Cuffdiff as differentially expressed, spliced, or transcriptionally regulated genes as it can be seen in figure 3.5.

To estimate transcript-level expression several sophisticated algorithms have been developed recently. Cufflinks approximate transcript level expression from a genome mapping by using a TopHat. Cufflinks use GTF information to identify differentially expressed transcripts. RSEM (RNA-Seq by Expectation Maximization) (Li and Dewey, 2011), eXpress (Forster et al., 2013) algorithms have been used to normalize within the sample to correct the sequencing biasness which quantify the expression from transcriptome (Finotello et al., 2014).

3.6 Differential expression

After quantification and normalization, statistical testing usually performed between conditions. Due to count-based nature of RNA-Seq data, Poisson distribution provides a good fit for counts arising from technical replicates, has been performed (Marioni et al., 2008). According to Leek et al. (Langmead et al., 2010) and Smyth et al. (Robinson and Smyth, 2007) these distributions do not account for biological variability across samples. Because Poisson's distribution accounts for the variance which is associated with the biological replicates and will be prone to high false discovery rate (FDR). FDR arises from the underestimation of sampling error (Robinson and Smyth, 2008). To overcome this limitation, recently many methods were developed that measure the statistical significance in a dataset with a low number of biological replicates. Cuffdiff finds differentially expressed genes and

transcripts that are transcriptionally and post-transcriptionally regulated and groups transcripts into biologically meaningful groups (Trapnell et al., 2012). In Cuffdiff, it is assuming that the number of reads produced by each transcript is proportional to its abundance (Trapnell et al., 2012). In RNA-Seq, presence of large number of technical variabilities arises during library preparation and in the same experiment, variation of biological replicates sometimes fluctuates changes in expression. Even though it's exceptional level of accuracy, RNA-Seq has sources of bias during the gene expression analysis. However, Cuffdiff can automatically eliminate a large fraction of the bias in RNA-Seq read distribution across each transcript and improves its abundance estimates. RNA-Seq has less technical variance compared to micro-arrays (Zhao et al., 2014). During sequencing, Cuffdiff provided multiple technical or biological replicate in sequencing libraries per condition and will help how read counts vary for each gene across the replicates. These variances calculate the significance of observed changes in expression (Trapnell et al., 2012). In cuffdiff, user can fed two or more SAM/BAM files, generated from TopHat alignment, as well as a GTF file that contains transcript annotations as input. As an output file, Cuffdiff reports numerous files that contain the results of DEG analysis. User can download and can be viewed with any spreadsheet application (such as Microsoft Excel). These files contain fold change in $\log_2$ scale, P values, q value and gene/transcript name and location in the genome (Trapnell et al., 2012). During differential gene expression analysis, Cuffdiff identify genes that are differentially spliced or regulated via promoter switching. In a gene, Cuffdiff groups together isoforms that have the same TSS which are all derived from the same pre-mRNA. Cuffdiff also calculates the total expression level of a TSS group by summing up the expression levels of the isoforms. Due to the presence multiple TSSs in a gene, Cuffdiff is also looking for changes in expression between TSS in different conditions.

3.7 Pathway analysis

Being a knowledge-driven quality, biologists are facing everyday how to interpret large-scale data especially with the emergence of high-throughput technology. The limitation usually lies to understand the meaning of array of genes to divulge the underlying molecular mechanism of the phenotype. Finally, list of differentially expressed genes can be grouped into common pathways; enable to identify differentially expressed pathways. The purpose of this pathway enrichment analysis is to find ultimate possibilities of the hidden connections of molecular information (transcriptome) with the phenotype of an organism in study (Emmert-Streib and Glazko, 2011). For individual genes, the variation in gene expression depends upon a certain disease that could be only moderate or even negligible. The pathway enrichment analysis of set of genes variation can evidence differences between different phenotypes (Mootha et al., 2003) and evaluates the function of differentially expressed genes and executes different cellular pathways and this pathway knowledge available in public repositories such as the Gene Ontology (GO) (Ashburner et al., 2000) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012). The pathway information concatenates into these databases with the gene expression patterns, resulting in the transformation of the array of individual gene identification into these pathways. According to Khatri et al. pathway analysis is performed based on overrepresentation analysis (ORA) and usually provided as an input a preselected differential expressed gene list (Khatri et al., 2012; Huang et al., 2009). This preselected list of genes will be taken from the output of cuffdiff analysis which has with higher rates of under- or over-expression with a certain FDR. A test is executed to ensure if the lists of gene have any biological function in general also involved in the same cellular process (Trapnell et al., 2009). The most commonly used tests are Fisher's exact test (Evangelou et al., 2012), hypergeometric (Zeeberg et al., 2003), chi-square (Falcon, and Gentleman, 2007), or binomial distribution. An extensive list of tools that are designed to perform this type of analysis is introduced by Lempick et al. (Zhong et al., 2004).

CHAPTER 4

METHODS

4.1. Tailor pipeline:

Over the past few years, ample amounts of methods had been developed to deal with different aspects of RNA-Seq data analysis. However, it was required to combine several bespoke methods to address the needs and specificities of each problem and sometimes this combination is not a simple challenge. Therefore, special awareness must be taken to prevent erroneous biological conclusions.

With these considerations in mind, we proposed a sequence of tools (pipeline) which is suitable to perform a comparative study between RNA-Seq samples. The pipeline referred as "Tailor Pipeline" which can able to able to take from raw RNA-Seq reads, to extract the main biological processes differing between the analyzed conditions as it can be seen in figure 4.1.

Figure 4.1. Schematic representation of the Tailor_Pipeline used in the RNA-Seq data. The raw .Bcl file has been submitted in the GHPCC cluster and performed differential gene expression analysis by using TopHat, Cufflinks, and Cuffdiff. Gene set enrichment analysis performed to identify the differentially regulated BPs, CCs and MFs and visualize the most differentially expressed pathways.

The pipeline has a broad interest because sometimes biologists interested to perform a comparative analysis, on the bench, under different environments, cells are treated with different pathogenic agents. Frequently, the best-chosen approach is the sequencing of the

cell transcriptome by using high-throughput sequencing, next-generation sequencing techniques (Shendure and Ji, 2008). However, the next-generation technologies generated, routinely, a dozen of gigabytes of data. Now, to extract relevant biological information from it, the computational power is essential. For that reason, this pipeline will provide all the necessary information which is required to solve any biological problem. The details result of this pipeline described in the subsequent chapters of this thesis.

A diagram illustrating the conceived pipeline is showed in Figure 4.1. This pipeline was implemented in a Tailor_file and incorporates both publicly available tools and scripts developed by Riley lab group members to perform the biological evaluation of the RNA-Seq data.

Data analysis begins with the input of the raw read files and the reference files. Once this data is gathered, reads are processed with FastQC (Trapnell 2004). FastQC consists of Java software which provides tools to perform a QC study in raw high throughput sequencing data. The analysis performed by this tool ensures that the data is qualitatively good and there are no problems or biases in it and reported per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and Kmer content. Sometimes, sequencer and the starting library materials may create some problem which can be easily detected by a QC analysis. Some abnormalities may be resolved by trimming base pairs from the raw read. The pipeline contains a script that can trim a given number of base pairs during this analysis. The pipeline is pre-set not to trim any bases from the raw nucleotide sequence. However, this option can be modified in the Tailor_file, depending on the QC results.

Afterwards, trimming, reads will be aligned with a reference sequence using Bowtie 2, a well-established mapper (Langmead and Salzberg, 2012). In RNA-Sequencing technology Bowtie2 is a fast and memory-efficient mapping tool that is particularly suitable

for the alignment of small reads, to the respected genome reference. In our analysis we have used the human and mouse genome respectively. Previously to the mapping process, the reference file must be indexed to be used by Bowtie2. To perform this task, bowtie2-build is used. This tool constructs a Bowtie index from the set of DNA sequences in the reference file, which usually is in the FASTA format (Langmead 2010). Once the index is built, Bowtie no longer uses the original FASTA sequence. At this point, a set of options associated with the type of search performed by the Bowtie 2 algorithm need to be defined. The output of the mapping process is a Sequence Alignment/Map (SAM) file which stores the information about the read alignments against the reference sequence. Particularly, for paired-end reads two records are printed (i.e. two lines of output) describing the mapping proprieties for each comparison (Li 2009).

Typically, after mapping RNA-Seq reads to a reference genome, the number of reads that map a certain gene or transcript is measured. The read counts have been found to be roughly linearly related to the abundance of the target transcript (Mortazavi 2008). To get the gene expression information, the pipeline uses TopHat, a script integrated in the TopHat package that counts how many reads map to a certain feature (Kim 2013). Due to the nature of this study, where high-level pathway enrichment analysis was the goal, it was not relevant to consider multiple isoforms of the same gene. To perform the count of the mapping reads, a reference genome file is required that contains information about the features.

Next differential expression analysis was performed between the RNA-Seq samples to detect differentially expressed genes among the conditions in study and differential gene expression analysis has been done by using Cuffdiff (Trapnell 2012), that takes the output files from TopHat or another read aligner. This output files are two or more fragment alignment SAM/BAM files, as well as a merged.gtf from the output of Cuffmerge as input (Trapnell 2012). Cuffdiff tests the observed log fold change in transcripts expression against the null hypothesis of no change and produces several output files for changes in expression

at the transcript level, primary transcripts, and genes. To use multiple conditions, users must specify multiple replicates by feeding in the associated BAM files for each condition. For each gene, the output table contains information about the mean gene expression level, the fold change from the first to the second condition, the logarithm (to basis 2) of the fold change, the p-value for statistical significance of this change and the p-value adjusted with Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) that controls the percentage of false positives among all the rejected hypotheses (FDR).

To select the significant differentially expressed genes, the algorithm is pre-set to perform a trimming based on the raw output table by selecting only p-value less than 0.1. To understand this trimming, it was necessary to have a clear knowledge about which statistical measure the p-value translates provided the null hypothesis is true. This null hypothesis refers to a general or default position and it was rejected if the p-value is less than a significance level. In differential gene expression analysis, the null hypothesis corresponds to a scenario in which the genes were not significantly differentially expressed. This means that lower p-values was unlikely that the observed difference was occurring randomly and, thereby, the 0.1 p-value cut-off assures that only the statistically important entries will be considered for further analysis.

Lastly, the genes found to be differentially expressed are associated with GO terms using a Bioconductor package called GOStats (Gentleman and Falcon, 2013). GOStats used a hypergeometric test to relate a given gene list with the standardized controlled vocabulary in the GO database. Particularly, it consists in three structured controlled vocabularies: biological processes (BP), sets of molecular events which are essential to the functioning of integrated living units; cellular components (CC), parts of cells or its extracellular environment; and molecular functions (MF), elemental activities of a gene product at molecular level (Ashburner 2000). In the final step, it is possible to have a biological insight about the samples being compared. Based on the output table user can able to conclude the

most significantly differentially expressed active processes when the cell is subjected to different biological conditions. To perform an analysis using the Hypergeometric-based test implemented in the GOStats package, the pre-set universe is contained a genome wide annotation database for Homo sapiens, mouse and others those are mapped with Entrez Gene identifiers. The pipeline is adjusted to analyze any species RNA-Seq data. Secondly, it is necessary to define a list of genes for the analysis and this list corresponds to the collection of genes in the final table of the differential expression analysis step. For the entries generated through the hypergeometric test, it is defined a p-value cut-off of 0.1. Additionally, the test direction is set as over, so the result of this step will be a table with the over represented GO terms associated with the differentially expressed genes found in the previous pipeline step for each one of the ontologies. In other words, GOStats will identify most important ontology terms that differ between the conditions being compared in the cuffdiff analysis.

CHAPTER 5

RESULTS

Part-1: Diet induced SAMP6 mice results

5.1. Dataset Description

In this study SAMP6 strain mouse were used and colonies taken from eight females and four males. For 25 mice each SAMP6 and AKR/J were fed with high fat diet (HFD). Generally, the age of those mice is 6–8 weeks of age. High-fat diet contains fat, protein, and carbohydrates and low-fat diet (LFD) contains fat, protein, and carbohydrates. In HFD, 60% calories are coming from fat whereas in LFD, 10.2% calories are coming from fat. This feeding process continues for 6 months where mice were fed daily with fresh high, low-fat. Subsequently Mice were sacrificed, and RNA has been extracted by using TRIZOL reagent. Then, two treated, HFD and LFD, SAMP6 mice were used in biological triplicates for library preparation. All RNA samples are DNAse treated and used to create single indexed (6 base pairs) RNA-Seq libraries by using TruSeq RNA preparation kit according to the manufacturer's protocol [96]. All sequencing has been done by using the Illumina HiSeqTM 2500 instrument. HiSeqTM 2500 was used to sequence RNA-Seq libraries that have been loaded with software version 3. Then performed 51-cycle paired-end run of the single indexed RNA-Seq libraries and demultiplexing has been done by using the raw bcl base call files upon completion. Based on that RNA-Seq dataset has been created from SAMP6 background strain mice that were fed regular low-fat and high diet (Table-5.1).

High fat mice	**Low fat mice**
NK001_1_CGATGT	NK004_4_GCCAAT
NK002_2_TGACCA	NK005_5_CAGATC
NK003_3_ACAGTG	NK006_6_CCTGTA

Table 5.1. Data set of diet induced mice study. The experimental data is composed 3 biological replicates for each of the 2 different conditions. 3 high fat and 3 low-fat diet mice were used for this experiment.

The Tailor pipeline was used to process the experimental data characterized above. In the subsequent section I have described the results from each processing step, interpret them and ultimately, conclude about which are the up- and down- regulated pathways in diet induced SAMP6 mice that associated with fibrosis.

5.2. Differential Gene Expression

Overall, the quality of the HF-SAMP6-over-LF-SAMP6 RNA-Seq data was high. Nevertheless, the quality modules would generate warnings. Those inadvertences appeared on per base sequence content, on the per base GC content and on the Sequence Duplication levels sections. Particularly, the first warning topic plots out the proportion of the four DNA bases for each base position in a sequence file. These properties were not verified for any of the analyzed read files. In fact, all of them had a high variability on the first 15 bases, which pointed out to the presence of an overrepresented sequence in the library. This may be related with a problem in the library generation or can be a consequence of an abnormal sequencing process. However, the most plausible explanation was the use of random hexamer priming to introduce biases at the start of sequencing reads, as described by Hansen et al., 2010 (Hansen 2010). Based on this study, 15 bases were trimmed from the beginning of each original sequence. Given that the reads are long (90 base pairs); the loss of information inherent with this trimming is not significant. Therefore, each one of the reads is composed, after this step, by 75 base pairs.

The filtered HF-SAMP6-over-LF-SAMP6 reads were then mapped against a reference mouse genome (mm10) file. The reference file was downloaded from Ensemble website and contains the reference mouse DNA sequence in FASTA format. The reference file was indexed by bowtie-build and used by Bowtie 2 to perform the mapping of the HF-SAMP6-over-LF-SAMP6 RNA-Seq reads to the reference mouse genome. Following alignment of the RNA-Seq reads, the data need to be translated into a quantitative measure of gene expression. This task can be achieved by TopHat, which counts the number of reads that

map a given gene. To perform this, it is necessary a reference file that contains all the annotated protein coding and non-coding genes in the mouse genome release 10. This information is contained on a GTF file that was downloaded from Ensemble's website (ftp://ftp.ensembl.org/pub/release-71/gtf/homo sapiens/Homo sapiens.GRCh37.71.gtf.gz).

After running TopHat, the output file fed into Cuffquant to compute gene and transcript expression profiles and saves these profiles such that it can be analyzed in a timely manner by Cuffdiff which the last step is to determine differential gene expression (DGE). Cuffquant take the .bam mapping files made from each of the 6 biological replicates along with the merged.gtf file and generate .cxb (compressed binary file). Cuffquant reduces the computational load of quantifying gene and transcript expression of the HF-SAMP6-over-LF-SAMP6 sample especially if there are more than a handful of libraries.

Given the samples and the respective conditions that constitute the diet induced SAMP6 mice dataset, it was decided to compare SAMP6 HFD-fed to LFD-fed mice samples. Cuffdiff is a program that uses the cufflinks transcript quantification engine used to test the observed log fold change in its expression against the null hypothesis of no change and find differentially regulated genes and transcripts at the transcriptional and post-transcriptional level that share a common transcription start site.

The Cuffdiff module takes two or more fragment alignment BAM files from TopHat (such as accepted_hits.bam), as well as a reference GTF file containing transcript annotations as input. To do so, each one of the outputs from the previous step was concatenated with the table containing the information about the gene expression level in the control sample. From these concatenated tables, Cuffdiff estimates the dispersion of each gene and analyses whether there is differential expression between the defined conditions (e.g. comparison between SAMP6 HFD-fed compare to LFD-fed mice). The final throughput of this step is a table in which the entries correspond to the genes that are significantly differentially expressed among the two conditions being compared.

Our analysis included a total of 23,285 differentially expressed genes including protein coding transcripts and non-coding transcripts, lncRNAs, and microRNA from the SAMP6 HFD-fed compare to LFD-fed mice. Of these 387 genes were significantly differentially expressed between High fat and Low-fat diet mice according to the cut-off criteria (P<0.05 and |log2FC| >1.5). Now Cuffdiff result output is very large and is not possible to visualize the data. So, we have used CummeRbund to simplify the analysis and visualize the output of a differential expression analysis by using cuffdiff. CummeRbund handles the transformation of Cuffdiff data into the R statistical computing environment, making RNA-Seq expression analysis with Cuffdiff more compatible with many other advanced statistical analyses and plotting packages. CummeRbund takes the output files from cuffdiff and creates an SQLite database which describes the relationships between genes, transcripts, transcription start sites, and CDS regions.

This data can be represented in an MDS-plot, illustrates the pattern of similarities or distances among a set of objects (Figure-5.1). More specifically, each dot in the MDS-plot corresponds to a gene. In the x-axis in the plot represents the first euclidean dimension, and the y-axis the second euclidean dimension.

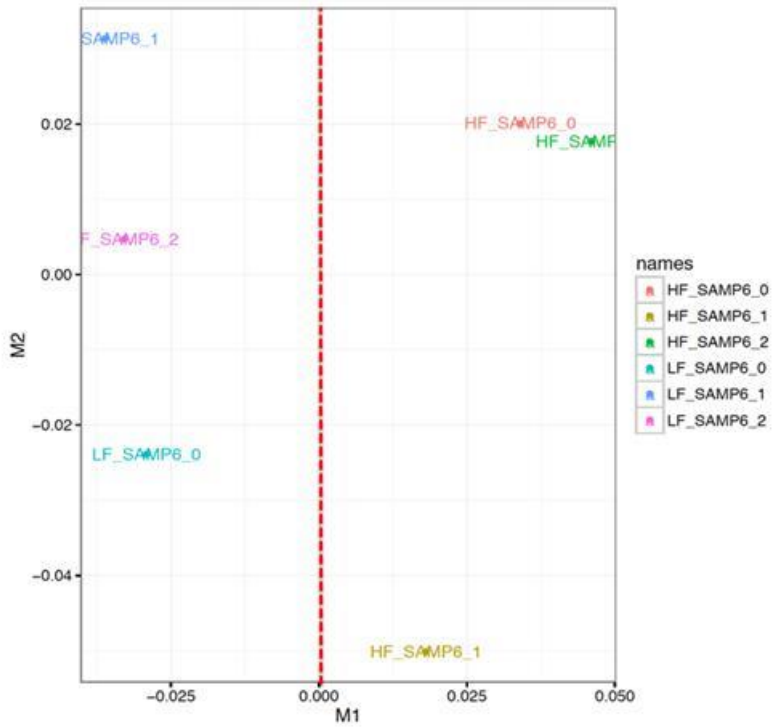

Figure 5.1. A multidimensional scaling (MDS) plot of the merged gene expression data. MDS-plots for all the high fat and the low fat diet induced to the SAMP6 mice. The figure shows a perfect separation between high fat SAMP6 and low fat SAMP6 mice. All color coded by biological replicates, with different symbols corresponding to different replicates.

Euclidean distances can be placed in a several ways. These distances can be placed in multidimensions but the standard way of representing MDS is to just plot the Euclidean distances with x-axis being *Dimension 1* and y-axis being *Dimension 2*. The dimensions are ordered based on how well samples are separated. Figure-3.2 is represented the MDS-plots for all the high fat and the low-fat diet induced to the SAMP6 mice. The closer the labels are together, the more similar the samples are. So, it is good to see that the high fat diet samples are clearly separated from the low-fat diet treated samples. In addition to this, genes classified as significantly differentially expressed with an FDR less than 0.1 in high fat and the low-fat diet induced to the SAMP6 mice are also clearly separated. By visually comparing the MDS-plots of diet induced SAMP6 mice samples lead us to check another dimension reduction technique such as Principal Component Analysis. It minimizes the dimensions and preserves the covariance of data whereas MDS minimizes dimensions, preserves distance between data

points. Figure-5.2 is represented the PCA-plot that shows a perfect separation between high-fat and low-fat diet biological replicates. In this method, the samples data points are projected onto the 2D plane in such a way so that data points are spread out in the two directions. This explain most well separation in the datapoints in the two-dimensional space. In the MDS plot, the x-axis is the direction that shows the maximum variation in the data point and is written *PC1*. The y-axis is orthogonal to the first direction which separates the data point second most in this direction and is written *PC2*. The percentage of total variance is demonstrated in the axis label which shows maximum variation.

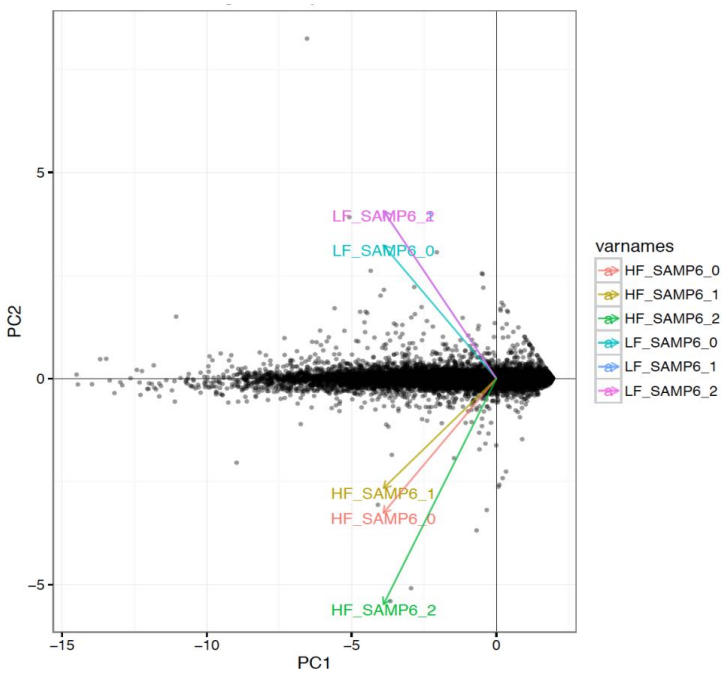

Figure 5.2. Principal component analysis. The PCA analysis was based on the gene expression patterns in high fat (HF) and low fat (LF) SAMP6 mice with induction of different diet. Analysis showed a perfect separation between two different diets. In the figure, LF_SAMP6_1 overlapped with LF_SAMP6_2.

However, it is important to keep always in mind that this analysis was performed with any number of biological replicates. In fact, as stated before, it is important to note that the biological conclusions extracted from the described methodologies must be interpreted with care. In fact, differences in library construction and variability intrinsic to the biological samples can greatly influence the number of false positives. It is imperative to have

biological replicates in the RNA-Seq dataset since these are essential in the measurement of the sample's intrinsic variability. Therefore, the absence of replicates is reducing the power of DE inference among RNA-Seq samples. After performing this analysis, we have decided to perform an additional layer of analysis that enables to see the functions of those differentially expressed genes.

5.3. Gene Ontology enrichment – GOStats

Using the pre-set options defined in the tailor pipeline, GOStats identified the statistically significant ontology terms that differ between SAMP6 HFD-fed compare to LFD-fed mice. In the following subsections a summary of the obtained results is described and compared with what it was expected, having into account that urinary voiding dysfunction was more severe in SAMP6 and was associated with pronounced prostatic and urethral tissue fibrosis. The X-axis represent the $-\log_{10}$ (p-value) and the Y-axis represent the biological, cellular or the molecular processes.

5.3. A Biological Processes:

SAMP6 HFD-fed compare to LFD-fed mice analysis provided significant ontologies that described operations or sets of molecular events pertinent to the functioning of fibrosis that are associated with diet induced high fat mice. In all, 13 biological processes were highly over-represented in our gene list, with p-values < 0.05 and fold-enrichment values of >2-fold (Table - 5.2).

Table – 5.2 Summary of the BP ontology terms and respective p-value associated with the set of differentially expressed genes from SAMP6 HFD-fed compare to the LFD-fed analysis.

GOBPID	Odds Ratio	Term	$-\log_{10}$(P-value)
GO:0009605	Inf	response to external stimulus	3
GO:0009611	45.863	response to wounding	3
GO:0001907	1836.556	killing by symbiont of host cells	3
GO:0032571	1836.556	response to vitamin K	3
GO:0044004	1836.556	disruption by symbiont of host cell	3
GO:0051919	1836.556	positive regulation of fibrinolysis	3

GO:0007598	1377.333	blood coagulation, extrinsic pathway	3
GO:0065008	Inf	regulation of biological quality	2.69897
GO:0016485	51.987	protein processing	2.69897
GO:0080134	32.604	regulation of response to stress	2.69897
GO:0051604	48.202	protein maturation	2.69897
GO:0017187	500.636	peptidyl-glutamic acid carboxylation	2.522879
GO:0018214	500.636	protein carboxylation	2.522879
GO:0051818	500.636	disruption of cells of other organism involved in symbiotic interaction	2.522879
GO:0051883	500.636	killing of cells in other organism involved in symbiotic interaction	2.522879
GO:0006508	28.997	Proteolysis	2.522879
GO:0006828	458.889	manganese ion transport	2.522879
GO:0010640	458.889	regulation of platelet-derived growth factor receptor signaling pathway	2.522879
GO:0051917	458.889	regulation of fibrinolysis	2.522879
GO:0006957	393.286	complement activation, alternative pathway	2.39794
GO:0007597	305.815	blood coagulation, intrinsic pathway	2.30103
GO:0031639	275.2	plasminogen activation	2.30103
GO:0030194	239.261	positive regulation of blood coagulation	2.221849
GO:0050927	239.261	positive regulation of positive chemotaxis	2.221849
GO:1900048	239.261	positive regulation of hemostasis	2.221849
GO:0050926	229.278	regulation of positive chemotaxis	2.221849
GO:0050820	220.093	positive regulation of coagulation	2.221849
GO:0007596	28.787	blood coagulation	2.221849
GO:0042730	211.615	Fibrinolysis	2.154902
GO:0007599	28.521	Hemostasis	2.154902
GO:0050817	28.521	Coagulation	2.154902
GO:0030449	196.476	regulation of complement activation	2.154902
GO:0018200	189.69	peptidyl-glutamic acid modification	2.154902
GO:0019835	189.69	Cytolysis	2.154902
GO:0031640	189.69	killing of cells of other organism	2.154902
GO:0044364	189.69	disruption of cells of other organism	2.154902
GO:2000257	183.356	regulation of protein activation cascade	2.154902

The differential biological processes analysis shows that “GO:0007598”, “GO:0017187” and “GO:0051818” are top most over-represented. The associated processes are blood coagulation, extrinsic pathway, peptidyl-glutamic acid carboxylation and disruption of cells of other organism involved in symbiotic interaction.

SAMP6 HFD-fed compare to LFD-fed mice has led us to provide deregulation of wound healing response, coagulation is responsible for the tissue fibrosis. Over the last

decade, coagulation signaling coordinate inflammation and tissue repair through the generation of fibrin and activation of proteinase-activated receptors (PARs) (Kryczka and Boncela, 2017). Coagulation cascade promote hemostasis and limit blood loss in response to tissue injury which will help to promote tissue fibrosis. Therefore, targeting the PARs will be a potential approach to limit fibrosis.

From this evaluation it is possible to conclude that the main differences between SAMP6 HFD-fed compare to LFD-fed mice is high fat diet induced SAMP6 mice associated with blood coagulation, extrinsic pathway which is a key factor for the tissue fibrinolysis which is concomitant to tissue fibrosis in high fat diet mice.

5.3. B Cellular Components:

Cellular components ontology associated with parts of a cell or its extracellular environment. Comparing SAMP6 HFD-fed to LFD-fed analysis, “GO:0005579” and “GO:0005615” were over-represented cellular components (Table - 5.3).

Table - 5.3 Summary of the CC ontology terms and respective p-value associated with the set of differentially expressed genes from SAMP6 HFD-fed compare to the LFD-fed analysis.

GOCCID	Odds Ratio	Cellular Components	$-\log_{10}$(P-value)
GO:0005615	37.326	extracellular space	2.69897
GO:0005579	986.167	membrane attack complex	2.69897
GO:0046930	369.604	pore complex	2.39794
GO:0005886	Inf	plasma membrane	2.221849

The associated terms were membrane attack complex and extracellular space. The $-\log_{10}$ (P-Value) is highly enriched for the term “membrane attack complex”. Remembering that BP “blood coagulation” is highly overrepresented for the SAMP6 HFD-fed compare to the LFD-fed analysis. But in the case of cellular components, membrane attack complex is highly over represented. Evidence suggested that formation of membrane attack complex direct associated with accumulation of fibrosis and complement activation may be responsible for the profibrotic response that occurs in the tubulointerstitial compartment (Abe 2004).

Stimulation of proximal tubular epithelial cells with membrane attack complex increased the mRNA concentrations of collagen type IV and its intracellular chaperone such as Heat Shock Protein 47 (HSP47).

5.3.C Molecular Functions:

The significant MF were described the elemental activities of a gene product at the molecular level. Only one GO term is associated with the molecular function in this analysis. The GO term "GO: 0004252" is associated with serine-type endopeptidase activity (Table - 5.4).

Table - 5.4 Summary of the MF ontology terms and respective p-value associated with the set of differentially expressed genes from SAMP6 HFD-fed compare to the LFD-fed mice analysis.

GOMFID	Odds Ratio	Molecular Functions	$-\log_{10}$(P-value)
GO:0070679	1038.000	inositol 1,4,5 trisphosphate binding	2.69897
GO:0004175	76.749	endopeptidase activity	2.69897
GO:0015279	830.300	store-operated calcium channel activity	2.69897
GO:0070011	55.297	peptidase activity, acting on L-amino acid peptides	2.39794
GO:0008233	53.387	peptidase activity	2.39794

In has been reported, protease may target many substrates which activate cell migration and fibrosis which support statistically (Kryczka and Boncela, 2017). Furthermore, serine-type endopeptidase activity is also described to be related with the mesenchymal transition and fibrosis. On the other hand, this activity term evidences the regulation of cell junction decomposition and ECM degradation. This may help to liberate sequestered growth factors such as TGFβ or VEGF that increases leukocytes infiltration and prolong inflammation. Finally, these proteases target many substrates and thus inflicting changes in distinct biological processes which correlated with cell migration and fibrosis (Kryczka and Boncela, 2017).

5.4. Pathway Analysis:

Finally, the list of differentially expressed genes can be grouped into common pathways. This analysis identifies differentially active pathways and, ultimately, possibilities the connection of molecular information (transcriptome) with the SAMP6 HFD-fed compare to the LFD-fed mice analysis. Gene Ontology analysis reveals high fat diet SAMP6 mice involved in blood coagulation process which plays pivotal roles in orchestrating inflammatory response. In addition to this high fat diet SAMP6 mice engaged in membrane attack complex which is directly associated with accumulation of fibrosis. This insinuates us to check the significant pathway which may associate with fibrosis.

Tailor pipeline identified 39 significantly differentially expressed pathways. Pathway enrichment analysis evaluates the significantly differentially expressed genes that concatenate into cellular pathways. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a public repository that contains pathway information (Kanehisa 2012). This is performed by relating the pathway information into these databases with the gene expression patterns, resulting in the transformation of the list of individual genes into a set of pathways. Previously report suggested up-regulation of a pre-fibrotic pathway namely the “ECM-Receptor Interaction” has been associated with fibrosis (Ekstedt 2006).

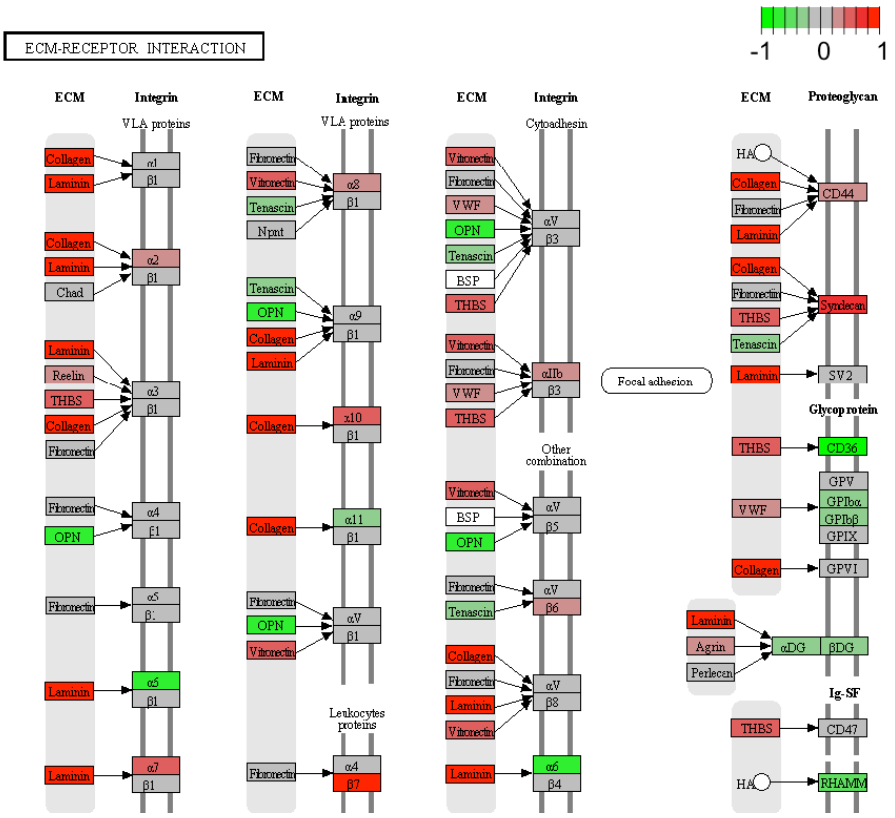

Figure 5.3. KEGG pathway of mm10 illustration of ECM receptor pathway. **ECM receptor interaction** showed differential expression of specific genes in this pathway. Genes significantly up-regulated consequent to high fat diet treatment in red, up-regulated consequent to low fat diet treatment in green, not differentially expressed in gray (arbitrary scale indicates extent of differential expression).

The ECM plays a central role to maintain the normal function of different tissues. It has been widely comprehended, over-expression of extracellular matrix (ECM) proteins have been associated with fibrosis (Almanza, D., 2018). In our analysis it can be shown collagen is up-regulated which presumably participates in the development of tissue fibrosis (Figure-5.3). Though diet induced SAMP6 mice did not show any fibrotic livers, may be this pathway insinuate the early stage of tumorigenesis. Apart from the extra cellular matrix pathway we have observed an unusual pathway known as the "peroxisome proliferator-activated receptor pathway" is highly over-expressed in our analysis (Figure-5.4).

Peroxisome proliferator-activated receptors (PPARs) are ligand-activated transcription factors of nuclear hormone receptor super family composed of three members namely PPAR-α, PPAR-δ, and PPAR-γ and play an essential role in metabolism by heterodimerization with

the retinoid X receptor (RXR) that bind to the specific regions on the DNA of the target genes. From the signaling pathway it can be visualized that PPAR-γ is targeted many genes such as ME1, ACBP, FABP1, LPL, ACO, CYP4A1, Thiolase B. Their main function is to promote lipogenesis, cholesterol metabolism, fatty acid transport, fatty acid oxidation. Due to down regulation of PPAR-γ in SAMP6 HFD-fed compare to the LFD-fed mice analysis, all the downstream target genes become down expressed and will not be able to transport fatty acid. Finally, oxidation of fatty acid and the metabolism of cholesterol become inhibited. Now chronic imbalances in lipid metabolism are often associated with obesity, Type-2 diabetes (T2D) and chronic liver disease. The common cause of chronic liver disease is the nonalcoholic fatty liver disease (NAFLD) disease which is responsible to accumulate the white adipose tissue in the liver (Ekstedt 2006). In this analysis we can visualize that oxidation of fatty acids are down regulated, suggested us to speculate that blood levels of triglycerides and free fatty acids are chronically elevated and excess fatty acids are derived from the extracellular source such as diet. Finally, it is reported that in liver, chronically elevated fat deposits result in NAFLD, which can lead to steatohepatitis (NASH) and, eventually, to non-reversible hepatic cirrhosis (Ekstedt 2006).

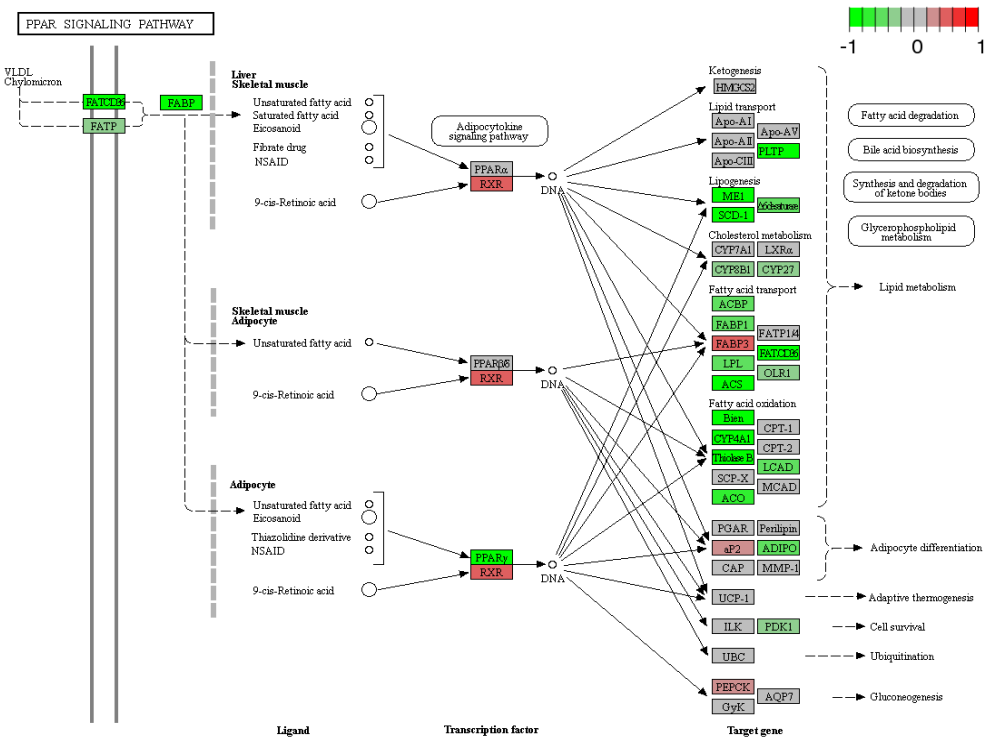

Figure 5.4. KEGG pathway of mm10 illustration of peroxisome proliferator-activated receptor (PPAR) pathway. PPAR pathway showed differential expression of specific genes in this pathway. Genes significantly up-regulated consequent to high fat diet treatment in red, up-regulated consequent to low fat diet treatment in green, not differentially expressed in gray (arbitrary scale indicates extent of differential expression).

It has been reported that NAFLD can develop a worst prognosis like cirrhosis and hepatocellular carcinoma (Yu 2016). In hepatocellular carcinoma, several different biomarkers have documented recently, and our aim is to find any of these known biomarkers are presented in our diet induced mice samples. After performing differential gene expression analysis by cuffdiff, we have identified genes/transcripts that are significantly differentially expressed in hepatocellular carcinoma. It is worth mentioning that several growth factors such as EGF1, EGF2 are upregulated in our diet induced SAMP6 mice model system and we are suspecting these may provide a pre-malignant signature. Because, EGFR plays an important role in cell growth and concurrently lead to the development of transformation by increasing the transcriptional activity. However, this premalignant signature has been implicated in cancer considering this may play an important role for uncontrolled cell growth and proliferation which is a characteristic feature of cancer (Grandhi 2016).

Part-II. Stromal Fibroblast Cell line results

5.5. Dataset Description

In this analysis N1 cells were used and these cells were derived from a stromal fibroblast cell line. These cells were expressed fibroblastic markers such as vimentin and calponin. These cells were demonstrated proliferation and secretion profiles which was somewhat similar with aging primary prostate fibroblasts (Rodríguez-Nieves 2016). Fibroblast cells were treated with human CXCL12 and human TGFβ and Trizol is used to extract RNA. Isolated RNA from N1 cells treated with CXCl12 or TGFβ were used to prepare libraries. All RNA samples are DNAse treated and used to create single indexed (6 base pairs) RNA-Seq libraries by using TruSeq RNA preparation kit according to the manufacturer's protocol. All sequencing has been done by using the Illumina HiSeqTM 2500 instrument. HiSeqTM 2500 was used to sequence RNA-Seq libraries that have been loaded with software version 3. Then performed 51-cycle paired-end run of the single indexed RNA-Seq libraries and demultiplexing has been done by using the raw bcl base call files upon completion. The experimental data is composed 3 biological replicates for each of the 2 different conditions mentioned such as CXCL12-vs-Control, TGFβ- vs- Control, and CXCL12-vs-TGFβ.

5.6. Differential Gene Expression

The pipeline described above was used to process the experimental data characterized above. In the subsequent section results from each processing step, interpretation and ultimately, conclusion has been made to decipher which are the up- and down- regulated pathways in CXCL12-vs-TGFβ induced fibroblast to myofibroblast phenoconversion. The .Bcl basecall files generated by Illumina HiSeq2500 were converted to FASTQ format using Tailor_Pipeline. This conversion was done with the help of bcl2fastq tool. To generate a single FASTQ file for each biological replicate, we have used the default parameter that split the files after 4 million reads to the reference human genome. This alignment file will be used

in the further downstream analysis of the stromal fibroblast data. In this analysis we have mainly focused on the Cuffdiff that enables to perform the differential gene expression analysis for the quantification of the transcripts for the samples. In computational biology it's important to know the function the genes/transcripts which is differentially expressed in the sample and gene ontology analysis is the prime important for this field and we have used GOStats R package to analyze the result considering that we have kept p-value of ≤0.05. In cell line data analysis, we have considered the pathview package that associated with the enrichment analysis. This enrichment analysis used to elucidate and visualized the top up-regulated cellular pathways, or down-regulated in CXCL12, or TGFβ treated cells by considering the cutoff q value ≤0.05. Transcriptomics analysis using human N1 cells which were derived from a stromal fibroblast demonstrated secretion and proliferation profile which was consistent with aging primary prostate fibroblasts (Patalano 2018). Moreover, RNA-Seq analysis of stromal fibroblast data revealed total of 10,633 transcripts were induced by CXCL12 or TGFβ compared to controls.

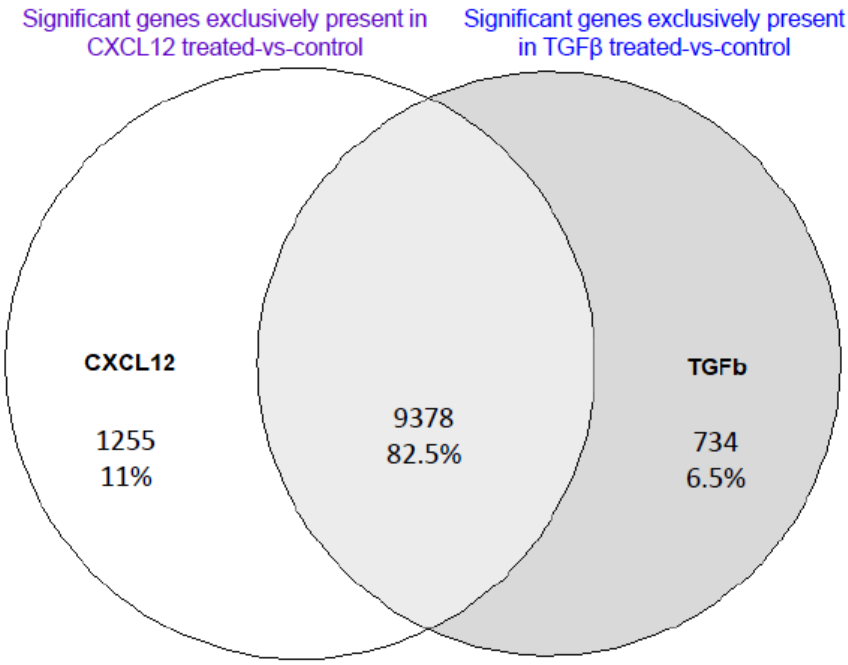

Figure 5.5. Venn Diagram of CXCL12- and TGFβ-Induced Transcriptomes. Cuffdiff analysis of gene transcripts differentially expressed consequent to CXCL12- or TGFβ –treatment is expressed as a Venn diagram. Of the total 9,378 transcripts induced by CXCL12 and TGFβ, 734 (6.4%) were differently expressed consequent to TGFβ treatment only, 1255 (11%) by CXCL12 treatment only.

These studies utilized human N1 cells which were derived from a stromal nodule of benign prostatic hyperplasia, exhibit a fibroblastic morphology, and demonstrate secretion and proliferation profiles consistent with aging primary prostate fibroblasts. Analysis of RNA-Seq data revealed that a total of 10,633 transcripts were induced by CXCL12 and TGFβ compared to vehicle controls. Of these, 9378 (82.5%) transcripts were significantly differentially expressed by CXCL12 and TGFβ, 734 (6.5%) were differently expressed consequent to TGFβ treatment only, 1255 (11%) by CXCL12 treatment only as it can be seen in figure 5.5.

Recent transcriptomics analysis reveals an astounding number of non-coding RNAs (ncRNAs) present in human genome. These ncRNAs have lack the capacity to code for a protein. Therefore, to date, these ncRNAs are as a “dark matter” and “junks” of human genome. Yet, over the past decade, several studies have shown these ncRNAs have numeral biological functions. However, it is still in debate, whether, ncRNAs transcription reflects accurate biology or offshoot of a leaky transcriptional system. Now, it is a broad question how we can able to interpret the biological meaning of transcription that distinguishes a gene that is simply transcribed.

Depending on the type of ncRNAs, transcription can occur by incorporating three RNA polymerases namely RNA Pol I, RNA Pol II and RNA Pol III. ncRNAs can be classified into two categories such as small ncRNAs and long ncRNAs depending upon the size. Recently, long noncoding RNAs (lncRNAs) are emerging aspect of modern biology, especially their role in human diseases have come into our attention. Many lncRNAs with tumor-suppressor or oncogenic functions in cancer have been discovered. However, genome-wide transcriptomics study mediated by high-throughput sequencing technique has revolutionized the genomics study and the pipeline identified lncRNAs that are significantly differentially expressed in stromal fibroblast cell line and have their role in tissue fibrosis.

We used RNA-Seq dataset that was acquired from N1 cells treated with CXCL12 or TGFβ. The experimental data is composed 3 biological replicates for each of the 2 different conditions. The datasets were sequenced by using paired-end sequencing on an illumina Hiseq-2500.

5.7. Prediction of fibrosis associated lncRNAs:

In this study, transcripts were reconstructed by using the genome guided methods. Current transcriptomics study falls into two categories based on the availability of genome: genome-guided and genome independent de-novo assembly (Garber et al., 2011). Also, we have determined the coding potential of lincRNAs was proposed in this study. First, TopHat was used to map the RNA-Seq reads in each sample to the human GRCh38 reference genome and 85.9% of the total RNA-Seq reads in each sample have been successfully mapped to the reference genome. Then, Cufflinks was subsequently used to assemble these aligned reads into transcripts based on the known gene annotation (Trapnell 2004), and the assembled transcripts were annotated and grouped into different categories using the Cuffcompare program from the Cufflinks package.

The fundamental aim of differential expression analysis is to identify genes that change in abundance between different experimental conditions. In this study, we used Cuffdiff (a module in Cufflinks package), to detect significantly differentially expressed lncRNAs between the CXCL12-vs-TGFβ induced stromal fibroblast tissues (Trapnell 2012), where the false discovery rate (FDR) was set to be 0.05.

High-throughput sequencing followed by bioinformatics analysis is a main stream of detecting the lncRNA. They have recently gained an attention due to their widespread involvement in disease. Based on the differentially expressed lncRNAs, we have performed cluster analysis to see the variation of expression between CXCL12 and TGFβ compared to control as it can be seen in figure 5.6.

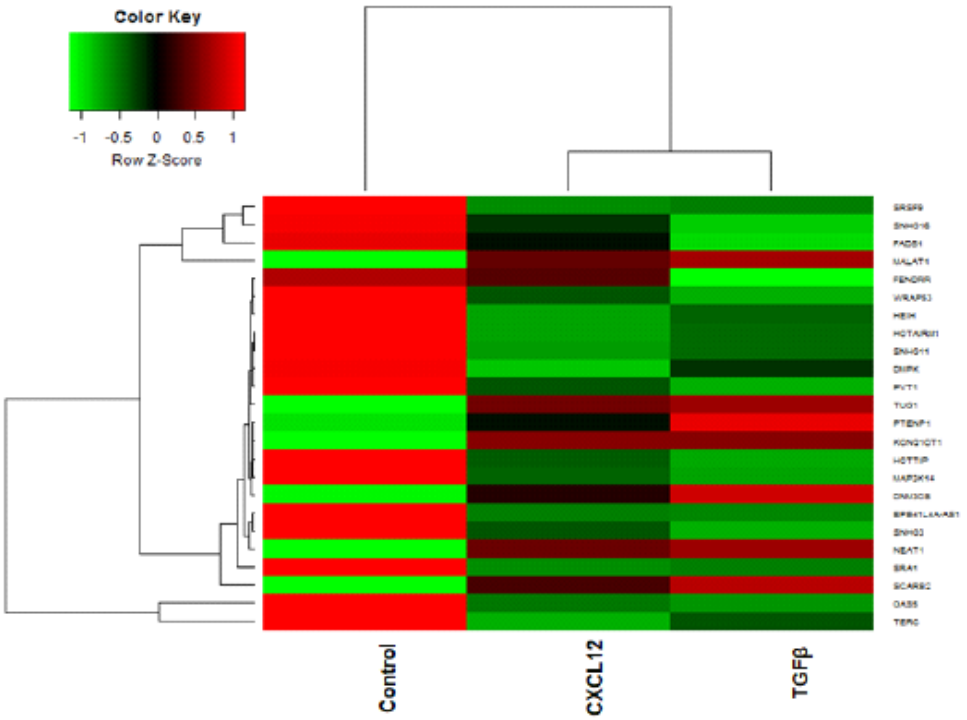

Figure 5.6. Unsupervised hierarchical clustering of differentially expressed lncRNAs. Cluster analysis created heat map based on differentially expressed lncRNA genes generated by RNA-Seq data from our own sequencing. It's showing the distinct expression pattern of lncRNAs in CXCL12 treated and TGFβ treated cells compared to normal. Each column represents a specimen and each row represents a gene. Red color indicates genes that were upregulated and green color indicates genes that were downregulated. Black indicates genes whose expression is unchanged in CXCL12 treated and TGFβ treated compared to control.

From the analysis the expression levels of the significantly differentially expressed lncRNAs have altered. Based on that, we have identified lncRNAs namely MALAT1, NEAT1, TUG1, PTENP1, Kcnq1ot1, DNM3OS, Scarb2, SRSF9, SNHG16, FADS1, WRAP53, HEIH, HEIH, HOTAIRM1, SNHG11, DMPK, PVT1, MAP3K14, SNHG3, SRA1, GAS5, TERC that were upregulated in our analysis. Of these differentially expressed lncRNAs, we have observed 7 lncRNAs namely MALAT1, NEAT1, TUG1, PTENP1, Kcnq1ot1, DNM3OS and Scarb2 are upregulated in both CXCL12 and TGFβ induced N1 cell line. MALAT1 involved promoting tumor growth and metastasis and regulate alternative splicing and cell cycle regulation that may associated with Prostate cancer (Cheetham 2013). MALAT1 increased HCC cell migration; tumor metastasis and recurrence through Wnt/TCF/β-catenin and Hippo/yes-associated protein (YAP) signaling pathways (Nordin et al 2014; Wang et al. 2014). NEAT1

has proven to be a transcriptional regulator for numerous genes; few of them are involved in liver and prostate cancer progression and promoted cell proliferation and invasion (Chakravarty 2014). TUG1 has also been suggested to be significantly associated with hepatocellular carcinoma and promoted cell growth and apoptosis (Mehra and Chauhan 2017). Finally, PTENP1 is a highly homologous processed pseudogene of the tumor suppressor gene PTEN that itself exerts a tumor suppressive function by acting as a decoy for PTEN targeting miRNAs (Poliseno 2010). Chen et al, reported the main function of PTENP1 is to repress tumorigenic properties of HCC. The role of lncRNA potassium voltage-gated channel subfamily Q member 1 opposite strand/antisense transcript 1 (KCNQ1OT1) is remain elusive in the context of myofibroblast phenoconversion. It is transcribed from intron 10 of the maternally expressed Kncq1 (KvLQT1) gene from a CpG island that is the imprinting control region (IC2) (Smilinich 1999). It has been reported that lncKCNQ1OT1 has been associated with diverse array of functions. Of these, one of the most important function is the involvement of cell proliferation. Because KCNQ1OT1 promotes cell proliferation through the upregulation of SMAD4 which is upregulated in both CXCL12 and TGFβ treated cell line. Previously, it was reported that TGFβ promotes the myofibroblast phenoconversion through SMAD dependent pathway. Furthermore, the results indicated that an increase level of KCNQ1OT1 may correspondingly regulate SMAD4 expression levels. So, we may suspect the lncRNA KCNQ1OT1 may promote fibroblast to myofibroblast phenoconversion through SMAD dependent pathway. DNM3OS, a gene that is transcribed into a non-coding RNA (ncRNA), contains three micro RNAs (miRNAs), miR-199a, miR-199a*, and miR-214, whose functions remain unknown. It has been reported the long non-coding RNA DNM3OS as a critical downstream effector of TGF-β-induced myofibroblast activation via SMAD dependent pathway. However, in the context of stromal fibroblast cell line, it remains unknown their function and the mechanism through which they promote myofibroblast phenoconversion. Hence, it may provide a novel paradigm for the treatment of

fibrosis. Finally, the function of lncRNA Scarb2 in cell proliferation is still elusive and need to be further investigated. The differentially expressed lncRNAs have been documented in the table 5.5.

Table 5.5:

Gene Name	Location	CXCL12 treated-vs control		TGFβ treated-vs control	
		Fold change	Q-Value	Fold change	Q-Value
MALAT1	11q13.1	1.995986551	0.000074448	2.229654234	0.000081672
NEAT1	11q13.1	1.568616847	0.000074448	1.686939226	0.000081672
TUG1	22q12.2	2.635305103	0.000074448	2.894210791	0.000081672
PTENP1	9p13.3	1.390510848	0.0040057	1.922231374	0.000081672
Kcnq1ot1	11p15.5	2.31493229	0.0000820	2.2448359	0.0000885787
DNM3OS	1q24.3	1.92294833	0.0000820508	2.458008662	0.0000885787
Scarb2	4q21.1	1.6579724	0.0000820	1.888684862	0.0000885787
SRSF9	12q24.31	0.571988878	0.0000820508	0.596421843	0.0000885787
SNHG16	17q25.1	0.780392598	0.0000820508	0.635598665	0.0000885787
FADS1	11q12.2	0.828545597	0.0452502	0.637077592	0.00017404
WRAP53	17p13.1	0.59680023	0.0000820508	0.464709164	0.0000885787
HEIH	5q35.3	0.540972064	0.0000820508	0.629099011	0.0000885787
HOTAIRM1	7p15.2	0.606565774	0.0000820508	0.67005666	0.000258076
SNHG11	20q11.23	0.70463791	0.000315805	0.738212463	0.00183138
DMPK	19q13.32	0.623637679	0.0000820508	0.78306013	0.0000885787
PVT1	8q24.21	0.669973065	0.0000820508	0.532798897	0.0000885787
MAP3K14	17q21.31	0.671974783214	0.00120927	1.076055134	0.0032499
SNHG3	1p35.3	0.653230926	0.00120927	0.531131868	0.0327341
SRA1	5q31.3	0.604459674	0.0000820508	0.610607832	0.0000885787

GAS5	1q25.1	0.73045991	0.0000820508	0.691657636	0.0000885787
TERC	3q26.2	0.412029631	0.0000820508	0.563245583	0.0000885787

In CXCL12 treated cell, COL1A1 and smooth muscle α-actin (αSMA) is upregulated but not TGFβ treatment. It's an open area of research whether silencing of MALAT1 reduces the mRNA levels of smooth muscle α-actin (α-SMA) and collagen type I, α1. However, to the best of our knowledge, there have been no published studies to demonstrate the specificity of these genes in fibroblast tissue. Thus, the results of this study can open a new theoretic insight into the identification of fibrosis specific genes.

5.8. Prediction of fibrosis associated miRNAs:

Identification of lncRNA by using the Tailor pipeline, our next question is whether any other short noncoding RNA that may differentially expressed to promote myofibroblast phenoconversion. Specially, microRNAs (miRNAs), play a crucial role in tumorigenesis and attenuates TGF-β signaling to stimulate angiogenesis and tumor growth (Bartel 2009; Suzuki and Miyazono 2011). They also play a significant role in tumor suppression in cancers (Hwang and Mendell 2006). Thus, we may hypothesize that miRNAs might be the important regulators to proliferate cancer cell and to promote myofibroblast phenoconversion.

By RNA-Seq, we found 9 miRNAs namely miR100HG, miR143HG, miR17HG, miR210HG, miR22HG, miR4435-2HG, miR663A, miR663AHG and miR-let7BHG that had significantly differentially expressed based on q-value < 0.05 and considering they may have a role in promoting myofibroblast phenoconversion. Our bioinformatics analysis revealed differentially expressed miRNAs, we have performed cluster analysis to see the variation of expression between CXCL12 and TGFβ compared to control as it can be seen in figure 5.6. From the hierarchical analysis it can be revealed miRNAs expression have been altered. Of note, the two most upregulated miRNAs are miR100HG and miR22HG (Fig. 5.7) that may interest in this analysis and assuming their possible role is to promote myofibroblast

phenoconversion. Though it's suspected but not validated yet their positive role in myofibroblast phenoconversion. The detailed list of miRNAs has been documented in the table 5.6. MIR100HG is a polycistronic miRNA host gene, which encodes miR-100, let-7a-2, and miR-125b-1 within its third intron, involved in cell proliferation and differentiation (Emmrich et al., 2014). Previous reports showed that miRNAs played an essential role in fibrosis, while the mechanism was not clear and needed

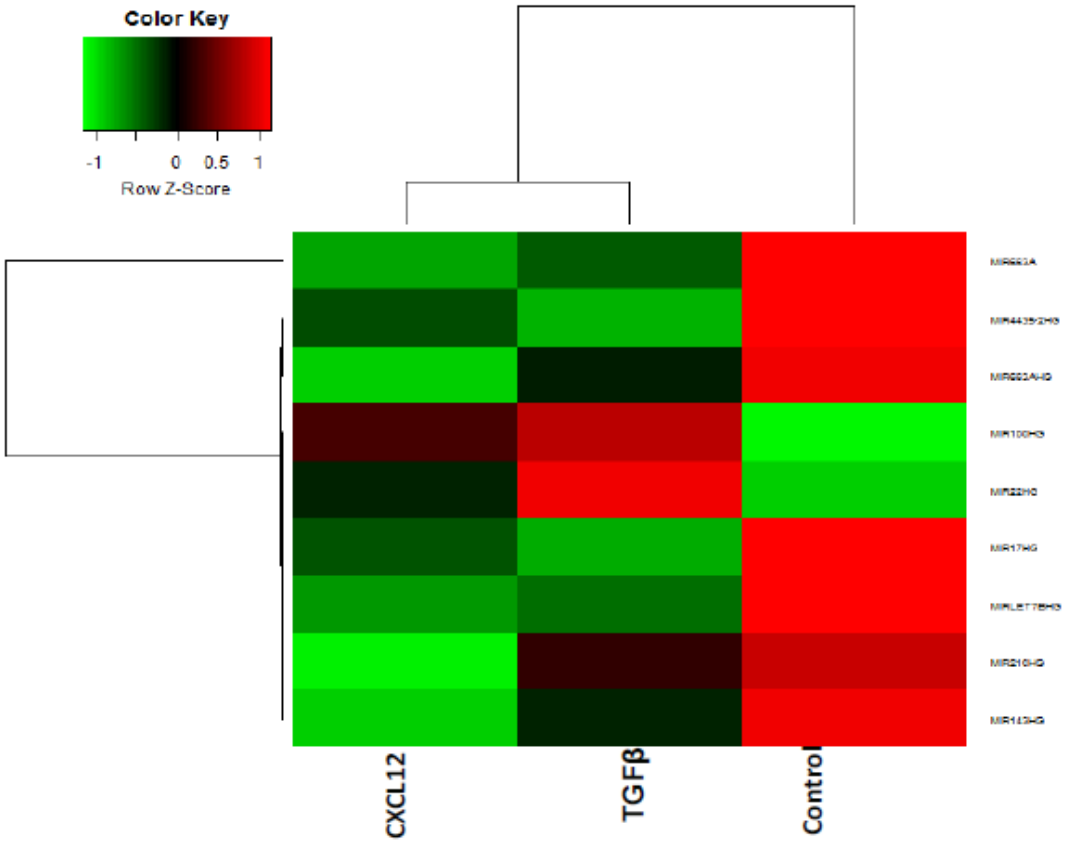

Figure 5.7. Unsupervised hierarchical clustering of differentially expressed miRNAs. Cluster analysis created heat map based on differentially expressed miRNA genes generated by RNA-Seq data from our own sequencing. It's showing the distinct expression pattern of lncRNAs in CXCL12 treated and TGFβ treated cells compared to normal. Each column represents a specimen and each row represents a gene. Red color indicates genes that were upregulated and green color indicates genes that were downregulated. Black indicates genes whose expression is unchanged in CXCL12 treated and TGFβ treated compared to control.

more elucidation. The detailed information of upregulated miRNAs is documented in table 5.6. Furthermore, over-expression of miR-100HG and miR-22HG in this stromal fibroblast cell line upon inducing the TGFβ highlighting a miRNA-mediated regulatory network potentially important for cellular proliferation.

Table 5.6:

Gene Name	Control (FPKM value)	CXCL12 treated –vs control (FPKM value)	TGFβ treated -vs control (FPKM value)	Fold Change	Q-Value
MIR100HG	5.97311	14.3705	17.3948	2.66289118	0.00007806
MIR143HG	1.74695	0.776572	1.13852	0.549678074	0.01771129
MIR17HG	3.634215	2.72922	2.48242	0.716698003	0.00099742
MIR210HG	5.14924	2.57128	4.27341	0.667055105	0.00813042
MIR22HG	9.454085	11.4505	14.5692	1.378657307	0.00306516
MIR4435-2HG	130.665	96.4805	86.2446	0.698804625	0.00007806
MIR663A	35614.05	11488.3	16110.3	0.388444969	0.00007806
MIR663AHG	88.0577	33.4411	55.4406	0.506454721	0.00007806
MIRLET7BHG	3.408655	2.1428	2.26701	0.647233958	0.00007806

However, Recent data suggested miR22HG upregulated and located in 17p13.3, a chromosomal region that is frequently deleted, hypermethylated in hepatocellular carcinoma [Zhang et al., 2018]. Previously, it was reported that miR22HG expressed significantly lower in HCC and an associated with the prognosis of patients with HCC. In the contrary, in our work we have noticed a significant increase in the expression of miR22HG and presumably it may promote myofibroblast phenoconversion.

5.8. Revisit of aminoacyl tRNA synthetase

Genes of ARS being housekeeping, for long their connection to diseases remained unsuspected. Their expressions vary dynamically from cell line to cell line and under stress conditions. Besides their major canonical role in translation, they are involved in pathways of cell signaling, cell survival, metabolisms of amino acids, stress response programs, regulations of enzyme synthesis and apoptosis. Many consider them to be hotspots of the regulation system (Ibba, M. & Söll, D. 2001).

In human there are 37 ARS genes, distinguished into two distinct sets based on their locations, either in cytoplasm (designated with single letter amino acid code followed by RS) or in mitochondria (has a '2' suffix). There are 17 cytoplasmic ARS (including the bifunctional glutamyl-prolyl-tRNA synthetase, EPRS, in charge for aminoacylation of tRNAGlu and tRNAPro), 18 mitochondrial, and 2 dual-localized, GARS and KARS, present in cytoplasm as well as in mitochondria (Yao, P. & Fox, P. L. 2013).

Mammalian ARS interacts with multifunctional proteins (AIMPs) by catalyzing the ligation of amino acids to their cognate tRNAs. Along with catalytic activity domains, ARS has other motifs to interact with diverse regulatory factors. These structural convolutions are linked to functional flexibility, notably to oncogenic pathways of apoptosis, angiogenesis, cell growth, cell proliferation, signal transduction and many more (Park, S. G. et al.,2008). The deviations of ARS/ARS2 gene expressions presumably meet the differential protein needs of cancer cells, driving the malignancy.

To observe the common transcription profile of the ARS/ARS2 genes, we performed a large-scale RNA-seq analysis on all the considered datasets. RNA-Seq analysis of stromal fibroblast cell line, using 37 ARS/ARS2 gene expression signatures clearly pointed to the differential expression of ARS/ARS2 genes. Following different quality control and normalization procedures, differentially expressed genes (DEGs) were identified initially using a fold change cutoff of >2. The differential expressions of ARS/ARS2 in leukemia were visualized from the clusterogram, Fig. 5.8.

Observation of large-scale alteration of ARS/ARS2 gene-expressions indicated that enhanced statistical analysis needed to be applied to identify more robust and reliable signatures. We integrated different statistical approaches to achieve superior result. P values of genes were calculated across samples to identify gene expression signatures that differentiated cancer tissues from normal tissues. Further, we computed FDR (False Discover Rate), to sharpen the significance of our result. Our RNA-Seq analysis, based on q-value (FDR test), showed that

IARS, IARS2, EPRS, LARS, NARS, TARS, WARS2, in stromal fibroblast cell line, were upregulated. To verify these anomalous expressions, we have seen that two AARS/AARS2 namely NARS and WARS2 are upregulated in both the CXCL12 and TGFβ induced fibroblast cell line.

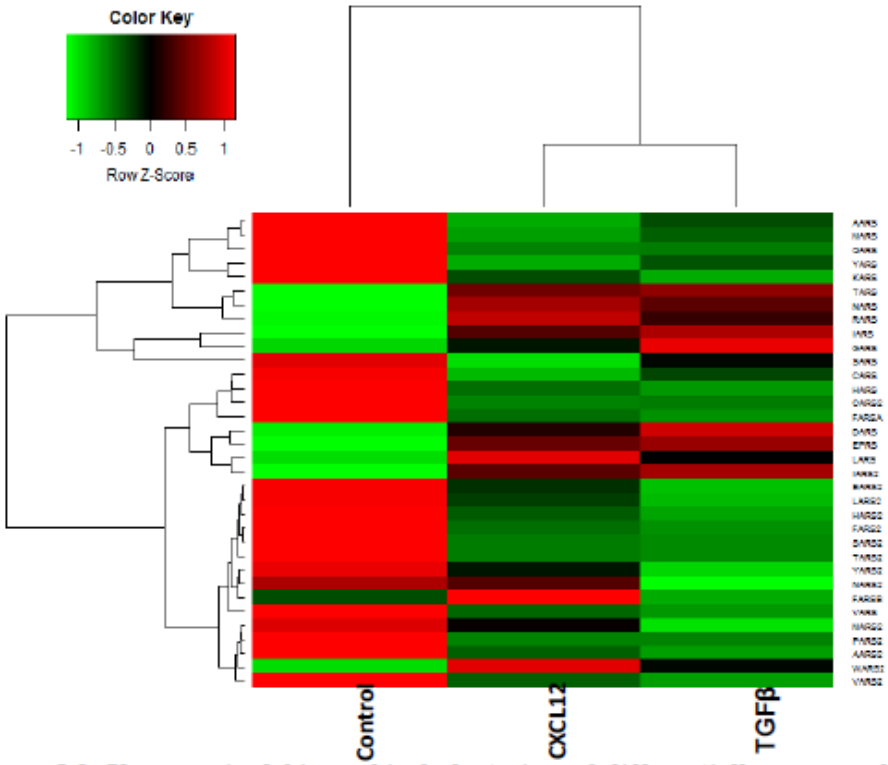

Figure 5.8. Unsupervised hierarchical clustering of differentially expressed AARSs. Cluster analysis created heat map based on differentially expressed miRNA genes generated by RNA-Seq data from our own sequencing. It's showing the distinct expression pattern of AARS/AARS2 genes in CXCL12 treated and TGFβ treated cells compared to normal. Each column represents a specimen and each row represents a gene. Red color indicates genes that were upregulated and green color indicates genes that were downregulated. Black indicates genes whose expression is unchanged in CXCL12 treated and TGFβ treated compared to control.

As we observed varied rates of differentiation of ARS/ARS2 genes, it was assumed they took part in biological processes and molecular functions other than just in translation. Hence, we performed gene ontology (GO) analysis to determine the distribution of ARS/ARS2 genes in cell, the biological processes they took part in and the molecular functions they performed. From GO database we figured the precise scattering of ARS/ARS2 in and around a cell. Interestingly, they were not only restricted to cytosol and mitochondria but also dispersed in their surroundings. Remarkably, NARS, have the potential to promote tumor metastasis and

growth and WARS2 was involved in angiogenesis followed by migration and proliferation. Though what that implies functionally, still needs to be interpreted.

5.9. Gene Ontology Analysis

After performing the differential gene expression analysis, our next aim is to perform the gene ontology analysis to see any significantly upregulated ontology that was associated with the fibroblast to myofibroblast phenoconversion. Based on that in our developed tool, we used some condition in the gene ontology step to get the over represented GO-terms. In overrepresented GO terms, if the condition becomes TRUE, the hypergeometric test performed by using the conditional algorithm to estimate for each biological term whether they are statistically overrepresented at the specified p-value cutoff where it finds all child terms are significant. Calculation of log odds ratio (LR) revealed GO biological process (table 5.7) was profoundly weighted towards DNA synthesis in the TGFβ-mediated treatment whereas this was less evident in the CXCL12-mediated signature, where protein synthesis, protein metabolism, protein modification and ER to Golgi vesicle-mediated transport are more prominent.

Table – 5.7 Summary of the BP ontology terms and respective p-value associated with the set of differentially expressed genes from CXCL12-over-control analysis.

GOBPID	Odds Ratio	Term	$-\log_{10}$(p-value)
GO:0000183	24.82936394	chromatin silencing at rDNA	7.274905479
GO:1990542	19.06785988	mitochondrial transmembrane transport	5.411168274
GO:0007080	18.45824575	mitotic metaphase plate congression	5.217527376
GO:0051204	16.14479465	protein insertion into mitochondrial membrane	4.480172006
GO:0032508	15.8948578	DNA duplex unwinding	8.517126416
GO:1903747	14.45812097	regulation of establishment of protein localization to mitochondrion	3.943095149
GO:1900740	14.41086604	positive regulation of protein insertion into mitochondrial membrane involved in apoptotic signaling pathway	3.931814138
GO:1901522	14.41086604	positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus	3.931814138

GO:0051436	13.31509217	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	6.896196279
GO:0008625	13.29756425	extrinsic apoptotic signaling pathway via death domain receptors	3.580044252
GO:0060444	13.25546539	branching involved in mammary gland duct morphogenesis	3.568636236
GO:0010972	12.67793051	negative regulation of G2/M transition of mitotic cell cycle	3.388276692
GO:0006418	12.41262554	tRNA aminoacylation for protein translation	6.341988603
GO:0006302	11.90967462	double-strand break repair	6.022733788
GO:0006120	11.83303051	mitochondrial electron transport, NADH to ubiquinone	5.982966661
GO:0006595	11.52319145	polyamine metabolic process	3.029653124
GO:0007004	11.52319145	telomere maintenance via telomerase	3.029653124
GO:0031571	10.41741799	mitotic G1 DNA damage checkpoint	9.838631998
GO:0019068	10.38302752	virion assembly	2.677987561
GO:0043153	10.36889313	entrainment of circadian clock by photoperiod	2.674484337
GO:0006415	10.02806271	translational termination	11.56703071
GO:0033044	9.897565234	regulation of chromosome organization	2.524910197
GO:0000097	9.791909169	sulfur amino acid biosynthetic process	2.498393078
GO:0006895	9.791909169	Golgi to endosome transport	2.498393078
GO:0006995	9.791909169	cellular response to nitrogen starvation	2.498393078
GO:0032981	9.791909169	mitochondrial respiratory chain complex I assembly	2.498393078
GO:0043984	9.791909169	histone H4-K16 acetylation	2.498393078
GO:0072401	9.689339469	signal transduction involved in DNA integrity checkpoint	8.966576245
GO:0006399	9.492567568	tRNA metabolic process	6.649751982
GO:0072413	9.39830346	signal transduction involved in mitotic cell cycle checkpoint	8.617982957
GO:1902402	9.39830346	signal transduction involved in mitotic DNA damage checkpoint	8.617982957
GO:2001020	9.289625916	regulation of response to DNA damage stimulus	4.431798276
GO:0006361	9.227594114	transcription initiation from RNA polymerase I promoter	4.401209493
GO:0009226	9.215035299	nucleotide-sugar biosynthetic process	2.32339783
GO:0046755	9.215035299	viral budding	2.32339783
GO:0050687	9.215035299	negative regulation of defense response to virus	2.32339783
GO:1902590	9.215035299	multi-organism organelle organization	2.32339783
GO:1902400	9.107379013	intracellular signal transduction involved in G1 DNA damage checkpoint	8.272458743
GO:0006900	8.959219858	membrane budding	4.238072162
GO:0043038	8.660193246	amino acid activation	5.91721463

GO:1900101	8.647345302	regulation of endoplasmic reticulum unfolded protein response	2.151995729
GO:0007096	8.638271487	regulation of exit from mitosis	2.149721447
GO:0008535	8.638271487	respiratory chain complex IV assembly	2.149721447
GO:0010664	8.638271487	negative regulation of striated muscle cell apoptotic process	2.149721447
GO:0042772	8.638271487	DNA damage response, signal transduction resulting in transcription	2.149721447
GO:0048194	8.638271487	Golgi vesicle budding	2.149721447
GO:0051571	8.638271487	positive regulation of histone H3-K4 methylation	2.149721447
GO:0051788	8.638271487	response to misfolded protein	2.149721447
GO:0060055	8.638271487	angiogenesis involved in wound healing	2.149721447
GO:0006614	8.369598373	SRP-dependent cotranslational protein targeting to membrane	12.55284197
GO:0006363	8.360110803	termination of RNA polymerase I transcription	3.886056648
GO:0000184	8.128249567	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	13.69250396
GO:0002042	8.071060172	cell migration involved in sprouting angiogenesis	3.714442691
GO:0072599	7.981565268	establishment of protein localization to endoplasmic reticulum	13.35457773
GO:0006413	7.963422108	translational initiation	11.69680394
GO:0000723	7.918843642	telomere maintenance	5.258060922
GO:0007569	7.807743979	cell aging	3.557520231
GO:0034644	7.799597855	cellular response to UV	6.73754891
GO:0048199	7.78206475	vesicle targeting, to, from or within Golgi	3.54515514
GO:2000785	7.78206475	regulation of autophagosome assembly	3.54515514
GO:0051437	7.750186943	positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition	8.230622674
GO:0006414	7.678671896	translational elongation	11.13489603
GO:0019083	7.586166812	viral transcription	12.42945706
GO:0070124	7.525527831	mitochondrial translational initiation	9.381951903
GO:0070125	7.525527831	mitochondrial translational elongation	9.381951903
GO:0050434	7.509285851	positive regulation of viral transcription	6.402304814
GO:0006298	7.493124523	mismatch repair	3.375717904
GO:0033014	7.493124523	tetrapyrrole biosynthetic process	3.375717904
GO:0072583	7.493124523	clathrin-mediated endocytosis	3.375717904
GO:1901028	7.308158062	regulation of mitochondrial outer membrane permeabilization involved in apoptotic signaling pathway	4.732828272
GO:2001022	7.217753519	positive regulation of response to DNA damage stimulus	3.214670165
GO:0000722	7.204239473	telomere maintenance via recombination	3.208309351
GO:0046685	7.204239473	response to arsenic-containing substance	3.208309351

GO:0061615	7.204239473	glycolytic process through fructose-6-phosphate	3.208309351
GO:0061621	7.204239473	canonical glycolysis	3.208309351
GO:1904292	7.204239473	regulation of ERAD pathway	3.208309351
GO:0051188	6.943963027	cofactor biosynthetic process	3.055517328

On the other hand, cellular component exhibited respiratory chain complexes, protein synthesis and degradation, and cell division were predominant in the TGFβ-mediated signature, whereas cellular signaling and Cul4-RING E3 ubiquitin ligase complex, clathrin vesicle coat, ECM component binding were prevalent in the CXCL12-mediated signature (table 5.8).

Table - 5.8 Summaries of the CC ontology terms and respective p-value associated with the set of differentially expressed genes from CXCL12-over-control analysis.

GOCCID	Odds Ratio	Term	$-\log_{10}$(p-value)
GO:0080008	14.94099	Cul4-RING E3 ubiquitin ligase complex	4.031517
GO:0055029	13.77231	nuclear DNA-directed RNA polymerase complex	3.664996
GO:0030125	13.69341	clathrin vesicle coat	3.645314
GO:0005680	13.06978	anaphase-promoting complex	3.453179
GO:0000502	12.81038	proteasome complex	6.473661
GO:0030904	12.44627	retromer complex	3.26184
GO:0005838	12.44627	proteasome regulatory particle	3.26184
GO:0022625	11.23328	cytosolic large ribosomal subunit	8.095284
GO:0015934	10.01159	large ribosomal subunit	4.785156
GO:0034719	9.953378	SMN-Sm protein complex	2.506006
GO:0005685	9.953378	U1 snRNP	2.506006
GO:0070469	9.368798	respiratory chain	2.329249
GO:0032040	9.341876	small-subunit processome	4.393619
GO:0005762	9.341876	mitochondrial large ribosomal subunit	4.393619
GO:0034045	9.330439	pre-autophagosomal structure membrane	2.319935
GO:0030117	8.749517	membrane coat	5.88941
GO:0030014	8.707614	CCR4-NOT complex	2.135281
GO:0005849	8.707614	mRNA cleavage factor complex	2.135281
GO:0042645	8.101777	mitochondrial nucleoid	5.329754
GO:0030140	8.09333	trans-Golgi network transport vesicle	3.664145
GO:0022627	7.684893	cytosolic small ribosomal subunit	4.970616
GO:0005876	7.483634	spindle microtubule	6.261219
GO:0005689	7.469399	U12-type spliceosomal complex	3.305304
GO:0030686	6.845697	90S preribosome	2.951187
GO:0000313	6.659933	organellar ribosome	7.767004
GO:0000784	6.45774	nuclear chromosome, telomeric region	7.431798

GO:0005801	5.919117	cis-Golgi network	4.524329
GO:1902555	5.910571	endoribonuclease complex	2.430589
GO:0030137	5.910571	COPI-coated vesicle	2.430589
GO:1990391	5.602764	DNA repair complex	3.234427
GO:0036452	5.598976	ESCRT complex	2.260372
GO:0005839	5.598976	proteasome core complex	2.260372
GO:0000159	5.598976	protein phosphatase type 2A complex	2.260372
GO:0016604	5.066767	nuclear body	3.602025
GO:0030529	5.02695	ribonucleoprotein complex	16.29158
GO:0098803	4.99055	respiratory chain complex	5.847712
GO:0032154	4.9847	cleavage furrow	4.308035
GO:0071339	4.978867	MLL1 complex	2.73702
GO:0098687	4.949023	chromosomal region	9.453457
GO:0016592	4.825659	mediator complex	3.366154
GO:0044452	4.811433	nucleolar part	5.527244
GO:0015935	4.78507	small ribosomal subunit	2.583606
GO:0043601	4.770978	nuclear replisome	2.574413
GO:0031519	4.734597	PcG protein complex	3.985908
GO:1990234	4.649523	transferase complex	10.45469
GO:0005763	4.563127	mitochondrial small ribosomal subunit	2.413596
GO:0030880	4.549874	RNA polymerase complex	10.1707
GO:0010494	4.5135	cytoplasmic stress granule	3.046508
GO:0070603	4.507754	SWI/SNF superfamily-type complex	6.262013
GO:0005840	4.37867	Ribosome	11.86646
GO:0030660	4.36177	Golgi-associated vesicle membrane	4.127844
GO:0097546	4.355314	ciliary base	2.254707
GO:0032993	4.309986	protein-DNA complex	5.228413
GO:0035097	4.29852	histone methyltransferase complex	5.79588
GO:0005925	4.216674	focal adhesion	28.15677
GO:0005657	4.209384	replication fork	5.037631
GO:0030055	4.143092	cell-substrate junction	28.29073
GO:0032153	4.095622	cell division site	4.277366
GO:0005654	4.072825	Nucleoplasm	183.6421
GO:0000118	3.974436	histone deacetylase complex	4.583359
GO:0005730	3.841616	Nucleolus	49.5867
GO:0000151	3.797252	ubiquitin ligase complex	15.45469
GO:0016591	3.793665	DNA-directed RNA polymerase II, holoenzyme	6.9914
GO:0016607	3.744944	nuclear speck	12.49349
GO:0012507	3.736606	ER to Golgi transport vesicle membrane	3.216913
GO:0016363	3.703721	nuclear matrix	6.69897
GO:0005901	3.687077	Caveola	5.35164
GO:0070013	3.674888	intracellular organelle lumen	207.567
GO:0005643	3.617216	nuclear pore	4.754487
GO:0031974	3.598013	membrane-enclosed lumen	222.266

GO:0031463	3.577708	Cul3-RING ubiquitin ligase complex	2.127909
GO:0005819	3.49422	Spindle	6.39794
GO:0097525	3.490056	spliceosomal snRNP complex	2.461535
GO:0044424	3.486763	intracellular part	2.911095
GO:0016605	3.479006	PML body	5.982967
GO:0030496	3.409893	Midbody	7.954677
GO:0000922	3.390406	spindle pole	7.510042
GO:0072686	3.348538	mitotic spindle	3.418038
GO:0005813	3.274238	Centrosome	20.87615
GO:0098798	3.26026	mitochondrial protein complex	2.195246
GO:0000228	3.186168	nuclear chromosome	25.06198
GO:0005778	3.184547	peroxisomal membrane	3.456195
GO:0005912	3.175634	adherens junction	24.30892
GO:1902554	3.173936	serine/threonine protein kinase complex	4.08302
GO:0005741	3.130656	mitochondrial outer membrane	8.247184
GO:0000788	3.112128	nuclear nucleosome	2.360233
GO:0008305	3.111162	integrin complex	2.032217
GO:0005759	3.088013	mitochondrial matrix	16.54363
GO:0000785	3.046311	Chromatin	4.928118
GO:0000777	3.043233	condensed chromosome kinetochore	5.228413
GO:0032592	3.035933	integral component of mitochondrial membrane	2.867686
GO:0031902	2.948891	late endosome membrane	5.44855
GO:0005694	2.932461	Chromosome	12.46344
GO:0008180	2.904121	COP9 signalosome	2.089693
GO:0005782	2.904121	peroxisomal matrix	2.089693
GO:0005740	2.899519	mitochondrial envelope	30.49214
GO:0000786	2.862159	Nucleosome	3.610763
GO:0031201	2.8371	SNARE complex	2.789035
GO:0005776	2.829458	Autophagosome	3.780958
GO:0030027	2.818115	Lamellipodium	7.818156
GO:0005905	2.804605	coated pit	3.470975
GO:0005881	2.803116	cytoplasmic microtubule	2.97492
GO:0031965	2.80047	nuclear membrane	10.00305
GO:0005623	2.777139	Cell	97.72584
GO:0019867	2.772263	outer membrane	8.223299
GO:1903293	2.767649	phosphatase complex	2.660285
GO:0005637	2.767649	nuclear inner membrane	2.660285
GO:0005765	2.738271	lysosomal membrane	11.24642
GO:0000792	2.733042	Heterochromatin	3.527382
GO:0045121	2.70816	membrane raft	11.17783
GO:0044438	2.660447	microbody part	3.768717
GO:0001726	2.625921	Ruffle	3.664777

Molecular function analysis revealed that CXCL12 and TGFβ related genes encoded proteins involved in DNA/RNA synthesis and regulation; protein synthesis and degradation and ubiquitination (table 5.9).

Table – 5.9 Summary of the MF ontology terms and respective p-value associated with the set of differentially expressed genes from CXCL12-over-control analysis.

GOMFID	Odds Ratio	Term	$-\log_{10}$(p-value)
GO:0034593	10.951	phosphatidylinositol bisphosphate phosphatase activity	3
GO:0070064	10.373	proline-rich region binding	2.69897
GO:0008175	9.796	tRNA methyltransferase activity	2.522879
GO:0017025	9.796	TBP-class protein binding	2.522879
GO:0019787	8.769	ubiquitin-like protein transferase activity	2.154902
GO:0008353	8.642	RNA polymerase II carboxy-terminal domain kinase activity	2.154902
GO:0010485	8.642	H4 histone acetyltransferase activity	2.154902
GO:0004709	7.207	MAP kinase kinase kinase activity	3
GO:0004298	5.474	threonine-type endopeptidase activity	2.221849
GO:0004708	5.474	MAP kinase kinase activity	2.221849
GO:0019200	5.186	carbohydrate kinase activity	2.09691
GO:0016538	4.804	cyclin-dependent protein serine/threonine kinase regulator activity	2.69897
GO:0051721	4.804	protein phosphatase 2A binding	2.69897
GO:0004712	4.618	protein serine/threonine/tyrosine kinase activity	2.522879
GO:0008200	4.614	ion channel inhibitor activity	3
GO:0031369	4.419	translation initiation factor binding	2.30103
GO:0051539	4.038	4 iron, 4 sulfur cluster binding	3
GO:0017048	3.93	Rho GTPase binding	3
GO:0043022	3.806	ribosome binding	3
GO:0016896	3.747	exoribonuclease activity, producing 5'-phosphomonoesters	2.30103
GO:0031492	3.747	nucleosomal DNA binding	2.30103
GO:0019213	3.691	deacetylase activity	2.69897
GO:0008536	3.458	Ran GTPase binding	2.045757
GO:0061631	3.458	ubiquitin conjugating enzyme activity	2.045757
GO:0003755	3.364	peptidyl-prolyl cis-trans isomerase activity	2.69897
GO:0016627	3.234	oxidoreductase activity, acting on the CH-CH group of donors	2.221849
GO:0050681	3.172	androgen receptor binding	2.522879
GO:0019003	3.101	GDP binding	3
GO:0030374	3.029	ligand-dependent nuclear receptor transcription coactivator activity	3

GO:0019900	2.785	kinase binding	3
GO:0008094	2.748	DNA-dependent ATPase activity	2.39794
GO:0048365	2.718	Rac GTPase binding	2.154902
GO:0005080	2.628	protein kinase C binding	2.39794
GO:0003727	2.623	single-stranded RNA binding	3
GO:0004860	2.623	protein kinase inhibitor activity	3
GO:0004722	2.356	protein serine/threonine phosphatase activity	2.522879
GO:0019903	2.349	protein phosphatase binding	3
GO:0001104	2.278	RNA polymerase II transcription cofactor activity	3
GO:0061733	2.211	peptide-lysine-N-acetyltransferase activity	2.154902
GO:0043566	2.2	structure-specific DNA binding	3
GO:0017137	2.181	Rab GTPase binding	3
GO:0003729	2.068	mRNA binding	3
GO:0002020	2.009	protease binding	3

5.9. Pathway enrichment analysis

In modern molecular biology, identification of associations between an input set of gene and annotated gene sets (e.g., pathways) is an important problem. Tailor pipeline identified 39 differentially expressed pathways and Ubiquitin-mediated proteolysis is the most significantly differentially expressed pathway (Fig. 5.9). ECM deposition is the common feature of fibrotic disease which interrupts the normal structure of the affected organs and leading to their dysfunction and failure. Degradation of protein via the ubiquitin-proteasome system is the significantly differentially expressed pathway that controls many critical cellular functions including cell-cycle progression, cell growth, and differentiation (Chen and Dou 2010). Anomalous alterations of expression of genes associated with proteasome pathway dysregulated cellular homeostasis and development of cancers, fibrosis, and neurodegenerative disorders, etc. Although the ubiquitin-mediated proteolysis mainly investigated in the field of cancers, recent transcriptomics of stromal fibroblast cell line data analysis revealed ubiquitin-mediated proteolysis may provide a rational basis for the discovery of novel therapy for fibrotic diseases and the consisted genes of this pathway are the part of multi-subunit RING-finger type 3 Cullin-RBX E3. Cullin proteins are molecular

scaffolds and essentially responsible for the assembly of Cullin-RING ubiquitin ligases (CRLs) that leads to ubiquitin-mediated proteolysis by facilitating the covalent attachment of ubiquitin group to target proteins. SEC31 is the target substrate of the Cullin-RING ubiquitin ligase complex.

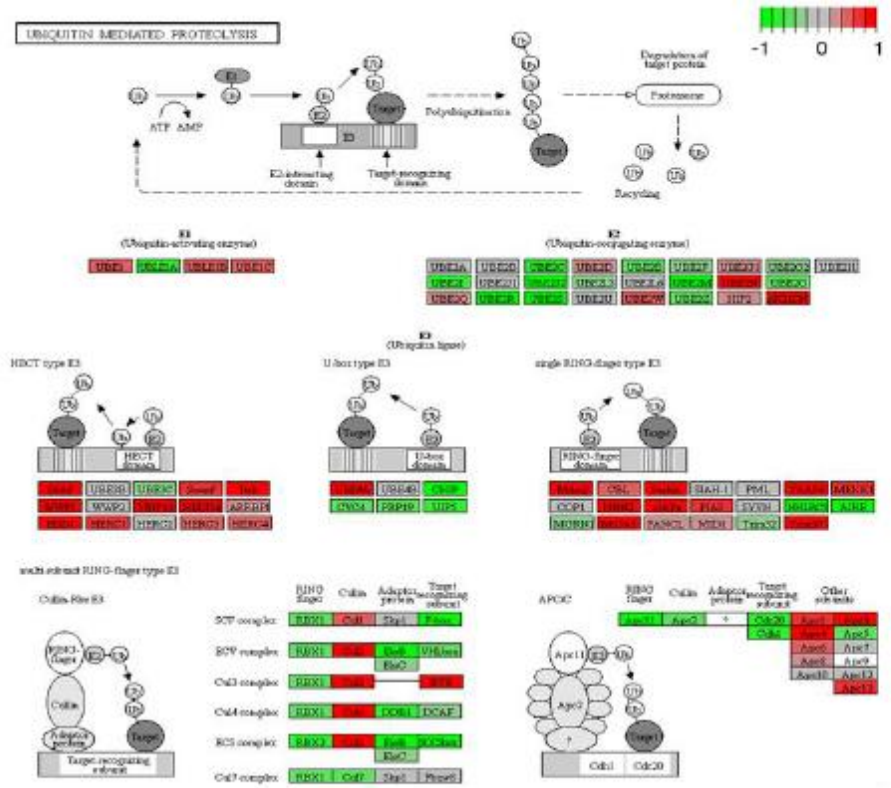

Figure 5.9. CXCL12 Transcriptionally Up-Regulates Cullin-RING Ubiquitin Ligases. KEGG Pathway hsa04120 illustration of multi-subunit Cullin-RBX E3 ubiquitin ligase structure and of differential expression of specific genes in this complex. Genes significantly up-regulated consequent to CXCL12 treatment in red, up-regulated consequent to TGFβ treatment in green, not differentially expressed in gray (arbitrary scale indicates extent of differential expression).

SEC31 monoubiquitination by CUL3-KLHL12 is necessary for the oversize COPII vesicle formation (Patalano 2018) and genes encoding the CUL3, KLHL12, and SEC23 proteins were differentially expressed by CXCL12- compared to TGFβ- treated cells. Other genes such as SCAP that preferentially up-regulated by the CXCL12/CXCR4 axis and associated with COPII vesicle-mediated ER-to-Golgi protein secretion (Patalano 2018); finally, it may be concluded these group of genes have a prospective role of CXCL12/CXCR4-mediated initiation of COPII vesicle formation and fibroblast to myofibroblast phenoconversion.

CHAPTER 6

DISCUSSION

There were many contributors that are involved in the metabolic disease and NAFLD disease progression. These factors may need to understand that can aid to diagnosis of these diseases. Previous studies suggested SAMP6 mice are associated with type 2 diabetes mellitus (Gharaee-Kermani 2013). In our current work, we have mainly elucidated the transcriptional regulation of genes that are associated in NAFLD. Now to initiate the NAFLD and the other associated diseases, inflammation play a pivotal role. In HFD-fed mice compared to LFD-fed mice analysis, several genes were overexpressed, and it's expected to find some significantly differentially expressed genes that were associated with metabolic syndrome-induced inflammation of liver. In addition to this, we have also observed several downregulated genes that might expect to play liver organ dysfunction. On the other hand, cell signaling pathway identification is also a main target in our analysis to know how HFD fed mice associated with liver disease. In our analysis, HFD-fed mice, the significant "PPAR-Gamma signaling pathway" was the top up-regulated pathway. Accumulation of excess white adipose tissue (WAT) can lead to develop inflammation, metabolic syndrome, type 2 diabetes, and NAFLD (Jung and Choi 2014). In SAMP6 mice accumulation of excess WAT or fat has been observed compare to the low-fat diet mice. This can be an explained due to the direct effect of excess WAT that may contribute the inflammation aspect. Previously it has been reported that under the same dietary conditions SAMP6 strain mice are able to progress several disease concerns (Brenner 2000). As previously reported result from our research group, these SAMP6 mice espoused type II diabetes which is a risk factor to emerge NAFLD. Taken together these data it may be concluded that SAMP6 mice model which has been used in this study, is able to instigate the metabolic syndrome disease which may develop to NAFLD. Therefore, we have showed, for the first-time alteration of immune-response, downregulation of metabolic processes that allowed us to study the unique transcriptional response to

NAFLD, which can aid to understand further knowledge in liver disease and cancer. In this analysis, we have failed to observe alteration of growth factors, heat shock proteins whereas elevation of collagen has been observed. This insinuates to the development of early stages of tumorigenesis and hepatocellular carcinoma which is a unique signature of NAFLD. Hepatocellular carcinoma and cirrhosis are strongly associated with NAFLD if it's not treated (Cholankeril 2017). Though current studies revealed several biomarkers have been identified that were associated with Hepatocellular Carcinoma, we intended to observe any significantly differentially expressed genes that may act as a significant biomarker in HFD-fed mice to LFD-fed mice.

Recent studies showed that deposition of collagen, extracellular matrix (ECM) are associated with fibrosis that can contribute to the etiology of LUTS. It is reported microenvironment of aging prostate tissue contained ample amount of inflammatory proteins particularly CXC-type chemokines (Rodriguez-Nieves 2013) whether these proteins can mediate fibroblast to myofibroblast phenoconversion is still under suspicion. It is well known that CXCL12 and TGFβ are inflammatory cytokines and achieved diverse cellular functions such cellular proliferation and differentiation (Huang 2009). TGFβ is well known pathogenic effector of fibrosis and it acts as a driving factor to promote fibroblast to myofibroblast phenoconversion, and ECM deposition (Rodriguez-Nieves 2013). However, in prostate stroma several C-X-C type chemokines such as CXCL1, CXCL2, CXCL3, CXCL8, and CXCL12 are altered and secreted and associated with benign hyperplasia (Gharaee-Kermani 2012). Previously, it was reported that, MAP Kinase signaling pathway activated by both CXCl12 and TGFβ. Previous report suggested TGFβ promoted fibroblast to myofibroblast phenoconversion in a Smad-dependent manner whereas CXCL12/CXCR4 achieved this phenoconversion by transactivating EGFR and promoting downstream MAPK signaling (Rodríguez-Nieves 2016). As a result, activation of these signaling cascades promoted the activation of the COL1A1 and COL1A2 genes and finally involved in the production of

procollagen protein. Accumulation of the extra cellular matrix deposition is a characteristic feature of fibroblast to myofibroblast phenoconversion and hallmark of tissue fibrosis (Gharaee-Kermani 2012). It is important to understand the underlying mechanism how CXCL12 and TGFβ concurrently promote fibrosis through transactivating of collagen. Treated the stromal fibroblasts cells with both CXCL12 and the TGFβ, followed by perform the transcriptomics analysis reveal interesting result which may be a remarkable feature for the myofibroblast phenoconversion. The remarkable distinguishing feature is an increased expression of ubiquitination/polyubiquitination with activation of the CXCL12, compared to TGFβ. Previously report suggested CXCL12 specifically activates the transcriptional response in the human prostate and epithelial cells. This activated transcriptional signal promotes cellular proliferation of stromal prostate epithelial cells which concurrently activates genes encoding proteins that promotes cellular proliferation (Begley 2008). In this analysis we have observed several CUL proteins that were over-expressed and in human these proteins may play an important role to promote CXCL12/CXCR4-mediated cellular proliferation and myofibroblast phenoconversion. In ubiquitin mediate proteolysis we found Cul4A is upregulated upon the treatment with CXCL12. In addition to Cul4A, CUL1 is upregulated in this analysis. So, it is still an unexplored area of research whether these CUL proteins may initiate to promote CXCL12/ CXCR4-mediated cellular proliferation and myofibroblast phenoconversion. Another molecular mechanism found in this analysis is several miRNAs are regulated by both CXCL12 and TGFβ which, in turn, inhibit the translation of mRNAs. Consistent with our analysis, we are trying to find out whether both similar and dissimilar subsets of miRNAs are activated by TGFβ/TGFβR axis compared to the CXCL12/CXCR4 axis because microRNA plays a crucial role in controlling cell migration and invasion (Baranwal and Alahari 2010). Alteration of miRNAs expression is widely altered in cancer, suggesting that deregulations of miRNAs are deeply associated in the development of tumor and cancer progression (Liu 2011). We found miR100HG,

miR22HG; miR210HG, miR663A, and miR663AHG are upregulated in CXCL12/CXCR4 axis compared to the TGFβ/TGFβR axis. Previous report suggested deregulation of miRNAs including miR-15, miR-16 have been associated with cancer progression (Liu 2011). So, further studies are required to decipher whether they promote fibroblast to myofibroblast phenoconversion. Recent studies from this analysis demonstrated significant amount of long non-coding RNAs (lncRNAs) which are upregulated in CXCL12/CXCR4 axis compared to TGFβ/TGFβR axis. LncRNAs are more than 200 nucleotides in length that have deficiency of protein-coding capacity (Spurlock 2016). LncRNAs regulated fibrosis by deposition of ECM that concomitantly stimulates the accumulation of collagen and glycosaminoglycans (Zhang 2018). LncRNAs are a functional and stable part of a genome and plays important biological roles such as cellular-, structural- processes that direct towards the complexity of an organism. Based on the stromal fibroblast cell line analysis, we are trying to find out whether any lncRNAs that may regulate fibroblast to myofibroblast phenoconversion. Our analysis demonstrated several lncRNAs including MALAT1, NEAT1, TUG1, PTENP1, Kcnq1ot1, DNM3OS and Scarb2 are upregulated in both TGFβ/TGFβR axis and CXCL12/CXCR4 axis. So, it's still an unexplored area of research whether lncRNAs play a pivotal role in fibroblast to myofibroblast phenoconversion.

Recently, research on miRNAs have been increasingly rapidly. Several studies have demonstrated that certain miRNAs are specifically correlated with certain cancer and the different expression level of miRNAs presumably function as an indicator for cancer metastasis and prognosis. The function of the lncRNA hostgenes MIR22HG and MIR100HG within this ncRNA ensemble remained elusive. Given the large-scale regulation of miRNAs in stromal fibroblast, it may possible these miRNAs are directly linked to myofibroblast phenoconversion. Notably, upregulation of miR100HG and miR22HG, may involve cell proliferation, migration, and invasion which holds true for the assumption, these miRNAs play an important role in fibroblast to myofibroblast phenoconversion. Thus, expression

patterns of miR22HG and miR100HG transcripts implicate an independent, yet unknown function in the context of myofibroblast phenoconversion. Therefore, loss-and gain function need to be performed to elucidate the role of these miRNAs in myofibroblast phenoconversion.

Over the past decade, role of AARS/AARS2 have been overlooked due to their prime function as a protein translation. However, recent high throughput sequencing provided a platform to revisit their role. Due to their pleiotropic role in protein translational regulation, cell signaling and amino acid metabolism, dysregulation of ARS genes has been associated with tumorigenesis. In this report, we investigated ARS gene expression in human stromal fibroblast to decipher their role in myofibroblast phenoconversion. RNA-Seq analyses of 9 datasets from stromal fibroblast cell line indicates anomalous expression of ARS in human fibroblast cell line. Aberrant expression of AARS genes shows upregulation of several AARS genes such as IARS, IARS2, EPRS, LARS, NARS, TARS, WARS2. The ARS/ARS2 genes arose early in evolution, and perchance, because of their presence from the beginning, these genes have been available for adaptation and recruitment to emerging cell signaling pathways, even those related to cancer. This functional flexibility allows ARS/ARS2 genes to play role in pathways other than protein synthesis. Clearly, the increase of ARS/ARS2 gene expression support increased protein synthesis in cancer cells and drives cell transformation. NARS is also involved in differentiation, presumably contributing to carcinogenesis. NARS, a class II ARS, identified as an up-regulated protein in this study. Our findings demonstrate that NARS is involved in cell proliferation, differentiation. Moreover, the current study demonstrated a novel role of NARS in promoting the migration ability of stromal fibroblast. On the other hand, WARS2 which is a mitochondrial aminoacyl tRNA synthetase gene involved in angiogenesis. Angiogenesis is an important factor playing a pivotal role in cancer cell metastasis and proliferation. Though these ARS/ARS2 genes are often considered as housekeeping genes recent evidence and our study clearly shows that their basal level of

expression to carry out normal physiological processes are often perturbed in disease condition. Our findings indicate that increase of the ARS/ARS2 genes must benefit fibroblast cells in some way favoring their survival and proliferation. How these genes promote myofibroblast phenoconversion are working in concert, if any, opens up a new arena of investigation.

CHAPTER 7

CONCLUSION

Emerging technological advances in genomics augmented an enormous amount of data at unprecedented high resolution (Khatri 2012). High-throughput sequencing of RNA allowed us to provide simultaneous measurement of RNAs sequence and expression at whole cellular level (Wang 2009). With the introduction of these new technologies, new bioinformatic approaches are required to analyze gigantic amount of data. In this thesis we have developed a pipeline for the analysis of RNA-Seq data and made contributions to the understanding of diet induced mouse model that are associated with the non-alcoholic fatty liver disease and the fibroblast to myofibroblast phenoconversion by using the human stromal fibroblast cell line.

Widespread genome-wide transcriptome study reconciled by high throughput sequencing technique has revolutionized the study of genetics at unprecedented resolution. Recent research divulged that an enigmatic amount of regulatory coding and non-coding RNAs encoded in human transcriptome (Tripathi 2017). Previous report suggested many unmentionables technology has been developed and categorized these non-coding RNAs as dark matter" and "junks". To debunk that idea, RNA-seq is an experimental technique that has been revolutionized and widely being used for studying non-coding RNAs recently due to its physiological and pathological significance.

First, we have implemented a complete pipeline to analyze RNA-Seq data. This pipeline begins by performing a data quality assessment, next it aligns the cleaned reads to a reference genome, measures the data gene expression level, tests for differential expression and, finally, concatenates this data into GO terms to find out significant ontology terms that has been associated with the biological problems. The outcome of this pipeline is a table that contains the differentially active cellular process between the RNA-Seq samples being processed. This enables the user to draw patterns for cataloguing gene function from high-volume data.

Subsequently, we have included a step that can able to investigate if a certain biological pathway that is significantly differentially expressed on a given RNA-Seq dataset. In RNA-Seq, the major problem in determining the conclusions is due to the low number of replicate samples. Lower number of biological replicates in RNA-Seq dataset provides a poor statistical significance. To overcome this problem, tailor pipeline incorporated cuffdiff step that perform a differential gene expression analysis. Gene ontology analysis is important to find the certain biological processes and molecular functions of the differentially expressed genes and it's a common approach for the gene set enrichment analysis. The motivation behind the introduction of Gene Ontology (GO) has grown to be the largest resource of its type which infers functional relationship of the differential gene. In tailor pipeline we have added this gene ontology step that will provide the functionality of known and newly discovered genes. To detect an association between set of input gene and sets of an annotated gene is a prime interest in molecular biology. To overcome this problem, we have included a pathview step in the pipeline to identify the differentially regulated pathways. It maps and delivers user data on relevant pathway graphs based on the array of gene interest. Pathway analysis is useful for the validation of the conclusions extracted from user biological problems. It's hopeful this complete package of pipeline can be useful not only for bioinformaticians but also for biologists in the future detect novel gene and their target pathway associated with any biological phenotype.

To evaluate the developed tools, we have studied two biological problems such as a diet induced SAMP6 mouse RNA-Seq dataset and the stromal fibroblast cell line dataset to study the myofibroblast phenoconversion. In the diet induced SAMP6 mice system transcriptome was collected from population of cells infected with high fat diet and low-fat diet. On the other hand, transcriptomics analysis performed on stromal fibroblast cell line data set which is characterized by the induction of CXCL12 and TGFβ.

Regarding the analysis of these datasets with the developed pipeline, it was possible to extract biological meaningful conclusions. To initiate metabolic syndrome, fat, high blood pressure, and elevated glucose levels are the key factors to promote metabolic syndrome in diet induced SAMP6 mice model system which concurrently initiate to develop diabetes, heart disease and finally cancer. Understand the transcriptional landscape is an important factor that can able to diagnose of these diseases. Until date, several studies reported that development of metabolic syndrome has been shown to be very closely associated with lack of physical activity and consequently it provides a tendency to rise of obesity rates among adults. Often NAFLD highly associated with the development of metabolic syndrome that can lead to liver dysfunction, cirrhosis of liver, and hepatocellular carcinoma. Previous studies showed SAMP6 mice can develop type-2 diabetes, a key factor to reduce the quality of life and health of the mice. We are trying to investigate how NAFLD affect the transcriptional landscape in liver pathophysiology. Transcriptomics analysis of HFD-fed mice showed many genes were up-regulated when compared to LFD-fed mice and associated with inflammations. It insinuates us to find any immune-related genes in our dataset that might be correlated between metabolic syndrome and inflammation which is not previously been stated. Additionally, this analysis showed some down-regulated genes associated with metabolic processes, which was able to point towards the fatty liver organ dysfunction. On the other hand, in HFD-fed mice, the significant up-regulated "PPAR-gamma signaling" pathway was the top up-regulated pathway in our study. Emergence of next generation sequencing technology showed HFD induced SAMP6 mice showed liver enlargement with accumulation of fat which conclude our mice might suffer from NAFLD. Several biological processes involved in including inflammation, metabolism, cellular stress responses, and ECM deposition have allowed us to scrutinize this exceptional transcriptional rejoinder to NAFLD, which can support in further understanding this disease. In our study, we have failed to find any cancerous or fibrotic phenotype of SAMP6 mice upon treatment with high fat

diet. However, growth factors such as EGF1, EGF2, and heat shock proteins, and collagen such as COL1, COL3 so on which have been overexpressed in this study and suspecting they are associated early stages of tumorigenesis and hepatocellular carcinoma. Finally, we were able to find the transcriptional association and the hallmark that are associated with NAFLD and early stages of tumorigenesis. Finally, molecular fibrosis signature associated with NAFLD disease increases our understanding towards the cellular response in mice model which is a novel approach towards the better understanding of translational application of the human fibrosis processes.

To strengthen the reliability of the tailor pipeline, a new dataset, with more robust information, has been processed by using the developed pipeline described above. In the new dataset we have aimed tissue fibrosis which is reconciled by the associations of several pro-fibrotic proteins that induce fibroblast to myofibroblast phenoconversion. Previous report suggested fibroblast to myofibroblast phenoconversion occurs through Smads and MEK/Erk proteins independently. In this study, we have treated the stromal fibroblast cell line with TGFβ/TGFβR and CXCL12/CXCR4. Previously, several reports suggested TGF-β1 promoted the transcription of both αSMA and COL1, which is coupled to myofibroblast phenoconversion. We therefore aimed whether CXC-type chemokines upregulated the level of αSMA and COL1 expression. Transcriptomics analysis reveals several upregulated transcripts COL1A1 and COL1A2 genes and resulted in increased levels of procollagen production, characteristic of myofibroblast phenoconversion. This analysis divulged unreported pathway name as ubiquitin mediated proteolysis, activates COPII-mediated vesicle formation responsible for transportation of large cargo complex, from the endoplasmic reticulum (ER) to the Golgi apparatus. Therefore, induction of CXCL12/CXCR4 facilitates the procollagen secretion and initiates ECM deposition which is a characteristic of tissue fibrosis. Several upregulated transcripts reported in this analysis such as CUL3 and KLHL12 are promoted in increased level of procollagen secretion,

transported from the ER to the Golgi in prostatic fibroblast. Increased level of procollagen promotes ECM deposition, hallmark of tissue fibrosis. Earlier transcriptomics analysis identified protein-coding genes only. Recently emerging technological innovation upfront multifarious capability identified uncharacterized ncRNAs, figuring out its biological significance. Tailor pipeline enables us to identify 15 differentially expressed ncRNAs in the stromal fibroblast analysis. It is noteworthy MALAT1, NEAT1, TUG1, PTENP1, Kcnq1ot1, DNM3OS and Scarb2 that were significantly differentially expressed in CXCL12/CXCR4 axis and the TGFβ/TGFβR axis, insinuating us to perform further research to decode their role in fibroblast to myofibroblast phenoconversion. In conclusion, the results of this study further highlight the pivotal roles played by ncRNAs in mediating changes in gene expression and cell functions occurring during pulmonary fibrosis. In particular, our results identified these lncRNAs as a new determinant of prostatic fibrosis and mechanistically ascribed its profibrotic effect to the regulation of myofibroblast phenoconversion leading to CXCL12 and TGF-β-dependent activation of stromal fibroblasts. We thus anticipate this analysis may represent a new effective therapeutic option to treat fibrosis in the future. Recent report suggested that over expression of MALAT1 may contribute to the development of fibrosis in non-alcoholic steatohepatitis (NASH) in liver through mechanisms involving inflammatory C-X-C motif chemokine ligand 5 (CXCL5) (Leti 2017). It may be concluded that potential consequence of myofibroblast phenoconversion may be associated with impaired smooth muscle activity, disrupted smooth muscle function and consequently deposition of ECM.

To regulate gene expression at the post-transcriptional and translational level, miRNAs play an important role (Morris et al. 2004). Based on gene ontology and literature mining, revealed their involvement to regulate cellular proliferation and cellular growth. In this study, miR22HG and miR100HG are presented strong evidence these miRNAs expressed significantly. However, their role in the context of myofibroblast phenoconversion and accumulation of ECM is still an open area of research and whether under-expression of

MIR22HG and MIR100HG have the optimal specificity and sensitivity for liver cancer diagnosis also needs future confirmation.

From this analysis it is clear that not all ARS/ARS2 genes are altered in cancer rather that they are cancer specific. This is presumably due to the codon bias for the oncogenes specific for a cancer. WARS2, though suspected, has never been implicated in fibrosis earlier. We provide here direct evidence of anomalous WARS2 and NARS expression in myofibroblast phenoconversion. In general, our study collectively implies that genes like AARS/AARS2 which are often designated as housekeeping are dysregulated in disease condition and plays an important role in cancer cell survival/proliferation.

The pipeline described in this thesis will provide a new arena in the field of genomics research. With the rapid advancement of sequencing technology coupled with augmented knowledge of the role of genomics in human disease, speeded up for the diagnosis for patients. We believe, the increasing 'mainstreaming' of whole genome sequencing is important of genomics research for many clinicians. Hope tailor pipeline will endow with a genomics research and its clinical applications, including its contribution to personalized medicine.

REFERENCE

1. Abe, K., Li. K., et al., The membrane attack complex, C5b, up regulates collagen gene expression in renal tubular epithelial cells. ClinExp Immunol. 136, 1, 60-66, Apr 2004.
2. Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., et al., Essential cell biology, 3rd ed. Garland Science, 2009.
3. Almanza, D., Gharaee-Kermani, M., Zhilin-Roth, A., et al., Nonalcoholic Fatty Liver Disease Demonstrates a Pre-fibrotic and Premalignant Molecular Signature. Dig Dis Sci, 5, Dec 2018.
4. Anders, S., Theodor Pyl, P., Huber, W. "HTSeq — A Python framework to work with high-throughput sequencing data Bioinformatics 2014.
5. Andrews, S., "FastQC: Duplicate sequences," http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3AnalysisModules/9 DuplicateSequences.html.
6. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25, 25–29, 2000.
7. Augenlicht, L. H. and Kobrin, D. "Cloning and screening of sequences expressed in a mouse colon tumor." Cancer research, 42, 1088–1093, Mar 1982.
8. Baranwal, S and Alahari, S. K. miRNA control of tumor cell invasion and metastasis. Int J Cancer. 126, 1283-1290, Mar 2010.
9. Begley, L. A., et al., The inflammatory microenvironment of the aging prostate facilitates cellular proliferation and hypertrophy. Cytokine, 43, 194-199, Aug 2008.
10. Benjamini, Y and Hochberg, Y. "Controlling the false discovery rate: a practical and powerful approach to multiple testing" J. Roy. Statist. Soc. Ser. B, 57, 289–300, 1995.
11. Bennett, S. "Solexa Ltd." Pharmacogenomics, 5, 433–438, Jun 2004.
12. Bhattacharjee, W. G., Richards, J., Staunton, C., Li, S., Monti, P., Vasa, C., et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." Proceedings of the National Academy of Sciences of the United States of America, 98, 13 790–795, Nov 2001.
13. Carlson M org.Hs.eg.db: Genome wide annotation for Human. R package version
14. Brenner, D. A. et al., New aspects of hepatic fibrosis. Journal of Hepatology. 32, 32–38, 2000.

15. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., at al., “Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.” Nature biotechnology, 18, 630–634, Jun 2000.
16. “Central dogma of molecular biology,” Nature, 227, 561–563, 1970.
17. Chakravarty, D., Sboner, A., Nair, S. S., Giannopoulou, E., Li, R., Hennig, S., et al., The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nat Commun. 5, 5383, Nov 2014.
18. Cheetham, S. W., Gruhl, F., Mattick, J. S and Dinger, M. E. Long noncoding RNAs and the genetics of cancer. Br J Cancer. 108, 2419-2425, Jun 2013.
19. Chen, D and Dou, Q. P. The ubiquitin-proteasome system as a prospective molecular target for cancer treatment and prevention. Curr Protein Pept Sci. 11, 459-470, Sep 2010.
20. Cholankeril, G., Patel, R., Khurana, S and Satapathy, S. K. Hepatocellular carcinoma in nonalcoholic steatohepatitis: Current knowledge and implications for management. World J Hepatol. 9, 533-543, Apr 2017.
21. Church, G. M. “Genomes for All,” Scientific American, 294, 46–54, Jan 2006.
22. Conesa. A, Madrigal. P, et al., “A survey of best practices for RNA-seq data analysis.” Genome Biol, 13, Jan 2016.
23. Crick, F. “On protein synthesis,” Symp. Soc. Exp. Biol, 12, 138–163, 1958.
24. Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y. et al., “Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.” Bioinformatics, 25, 3207–3212, Dec 2009.
25. Dvir, A., Conaway, R. C and Conaway, J. W. “A role for TFIIH in controlling the activity of early RNA polymerase II elongation complexes,” Proceedings of the National Academy of Sciences, 94, 9006-9010, Aug 1997.
26. Ekstedt, M., et al., Long-term follow-up of patients with NAFLD and elevated liver enzymes. Hepatology. 44, 865-873, Oct 2006.
27. Emmert-Streib, F and Glazko, G. V. “Pathway analysis of expression data: deciphering functional building blocks of complex diseases.” PLoS computational biology, 7, e1002053, May 2011.
28. Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L and Dudbridge, F. “Comparison of methods for competitive tests of pathway analysis.” PloS one, 7, e41018, Jan 2012.

29. Ewing, B., Hillier, L., Wendl, M. C and Green, P. “Base-calling of automated sequencer traces using phred. I. Accuracy assessment.” Genome research, 8, 175–185, Mar 1998.
30. F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors.” Proceedings of the National Academy of Sciences of the United States of America, vol. 74, no. 12, pp. 5463–7, Dec. 1977.
31. Falcon, S and Gentleman, R. “Using GOStats to test gene lists for GO term association,” Bioinformatics, 23, 257–258, Jan 2007.
32. Fiers, W., Contreras, R., Duerinck, F., G. Haegeman, D., Iserentant, J., Merregaert, W., et al., “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene,” Nature, 260, 500–507, Apr. 1976.
33. Finotello, F., Lavezzo, E, et al., “Reducing bias in RNA sequencing data: a novel approach to compute counts.” BMC Bioinformatics, 15, Jan 2014.
34. Fonseca, N. A., Rung, J., Brazma, A., Marioni, J. C. “Tools for mapping high-throughput sequencing data.” Bioinformatics, 28, 3169–3177, Dec 2012.
35. Forster, S.C., Finkel, A.M. et al., “RNA-eXpress annotates novel transcript features in RNA-seq data.” Bioinformatics, 29, 810-812, Mar 2013.
36. Garcion, E., Wallace, B., Pelletier, L and Wion, D. “RNA mutagenesis and sporadic prion diseases.” Journal of theoretical biology, 230, 271–274, Sep 2004.
37. Gautier, L., Cope, L., Bolstad, B. M and Irizarry, R. A. “affy–analysis of Affymetrix GeneChip data at the probe level.” Bioinformatics, 20, 307–315, Feb 2004.
38. Gentleman, R and Falcon, S. “Package GOstats: Reference manual,” 2013, http://www. bioconductor.org/packages/2.14/bioc/manuals/GOstats/man/GOstats.pdf.
39. Gharaee-Kermani, M., et al., CXC-type chemokines promote myofibroblast phenoconversion and prostatic fibrosis. PLoS One. 7, e49278, 2012.
40. Gharaee-Kermani, M., et al., Obesity-induced diabetes and lower urinary tract fibrosis promote urinary voiding dysfunction in a mouse model. Prostate. 73, 1123-1133, Jul 2013.
41. Gharaee-Kermani, M., Rodriguez-Nieves, J.A., Mehra, R.,Vezina, C. A., Sarma, A.V., Macoska, J.A. “Obesity induced diabetes and lower urinary tract fibrosis promote urinary voiding dysfunction in a mouse model,” Prostate, 10, 1123-1133, Jul 2013.

42. Gilbert, W and Maxam, A. “The nucleotide sequence of the lac operator.” Proceedings of the National Academy of Sciences of the United States of America, 70, 3581–3584, Dec. 1973.
43. Grandhi, M. S., et al., Hepatocellular carcinoma: From diagnosis to treatment. Surg Oncol. 25, 74–85, 2016.
44. GTF file,” ftp://ftp.ensembl.org/pub/release-71/gtf/homo sapiens/Homo sapiens.GRCh37.71.gtf.gz.
45. Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., et al.,“Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.” Nature biotechnology, 28, 503–510, May 2010.
46. Hansen, K. D., Brenner, S. E and Dudoit, S. “Biases in Illumina transcriptome sequencing caused by random hexamer priming.” Nucleic acids research, 38, e131, Jul 2010.
47. Heller, R.A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J. et al., “Discovery and analysis of inflammatory disease-related genes using cDNA microarrays.” Proceedings of the National Academy of Sciences of the United States of America, 94, 2150–2155, Mar 1997.
48. Hogeweg, P and Hesper, B. “Interactive instruction on population interactions,” Computers in Biology and Medicine, 8, 319–327, Jan. 1978.
49. Hoheisel, J. D. “Microarray technology: beyond transcript profiling and genotype analysis.” Nature reviews. Genetics, 7, 200–210, Mar 2006.
50. Hrdlickova, R., Toloue, M and Tian, B. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA. 8, Jan 2017.
51. Huang, D. W., Sherman, B. T and Lempicki, R. A. “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.” Nucleic acids research, 37, 1–13, Jan. 2009.
52. Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T. W., et al., “SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data.” Frontiers in genetics, 2, 46, Jan 2011. I.sequencing, “Data sheet: TruSeq RNA and DNA sample preparation kits v2,” http://res.illumina.com/documents/products/datasheets/datasheet_truseq_sample_prep kits.pdf.

53. Ibba, M and Söll, D. The renaissance of aminoacyl-tRNA synthesis. EMBO Rep, 2, 382-387, May 2001.
54. Illumina, TruSeq Stranded Total RNA Sample Preparation Guide (2012), pp. 1–162.
55. Jung, U. J and Choi, M.S. Obesity and its metabolic complications: the role of adipokines and the relationship between obesity, inflammation, insulin resistance, dyslipidemia and nonalcoholic fatty liver disease. Int J Mol Sci. 15, 6184-6223, Apr 2014.
56. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M and Tanabe, M. "KEGG for integration and interpretation of large-scale molecular data sets." Nucleic acids research, 40, D109–D114, Jan 2012.
57. Kent, W. J. "BLAT–the BLAST-like alignment tool." Genome research, 12, 656–664, Apr 2002.
58. Khatri, P., Sirota, M and Butte, A. J. "Ten years of pathway analysis: current approaches and outstanding challenges." PLoS computational biology, 8, e1002375, Jan 2012.
59. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R and Salzberg, S. L. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome biology, 14, R36, Apr 2013.
60. Kreimer, A and I, Pe'er. "Variants in exons and in transcription factors affect gene expression in trans." Genome biology, 14, R71, Jul 2013.
61. Kryczka, J and Boncela, J. Proteases Revisited: Roles and Therapeutic Implications in Fibrosis. Mediators Inflamm. 2570154, May 2017.
62. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al.,"Initial sequencing and analysis of the human genome." Nature, 409, 860–921, Feb 2001.
63. Langmead, B and Salzberg, S. L. "Fast gapped-read alignment with Bowtie 2." Nature methods, 9, 357–359, Apr 2012.
64. Langmead, B., et al., Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics. Chapter 11: Unit 11.7. doi: 10.1002/0471250953.bi1107s32. Dec 2010
65. Langmead, B., Hansen, K. D and Leek, J. T "Cloud-scale RNA-sequencing differential expression analysis with Myrna." Genome biology, 11, R83, Jan 2010.

66. Langmead, B., Trapnell, C., Pop, M and Salzberg, S. L. “Ultrafast and memory-efficient alignmen of short DNA sequences to the human genome.” Genome biology, 10, R25, Jan 2009.
67. Leti, F., et al., Altered expression of MALAT1 lncRNA in nonalcoholic steatohepatitis fibrosis regulates CXCL5 in hepatic stellate cells. Transl Res. 190, 25-39, Dec 2017.
68. Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., et al., “Comprehensive comparative analysis of strand-specific RNA sequencing methods.” Nature methods, 7, 709–715, Sep 2010.
69. Li, B and Dewey, C.N. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.” BMC Bioinformatics, 12, Aug 2011.
70. Li, H and Durbin, R. “Fast and accurate short read alignment with Burrows-Wheeler transform.” Bioinformatics, 25, 1754–1760, Jul 2009.
71. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al., “The Sequence Alignment/Map format and SAMtools.” Bioinformatics, 25, 2078–2079, Aug 2009.
72. Li, H., Ruan, J and Durbin, R. “Mapping short DNA sequencing reads and calling variants using mapping quality scores.” Genome research, 18, 1851–1858, Nov 2008.
73. Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K et al.,“SOAP2: an improved ultrafast tool for short read alignment.” Bioinformatics (Oxford, England), 25, 1966–1967, Aug 2009.
74. Liu, J., Zheng, M., Tang, Y. L., Liang, X. H and Yang, Q. MicroRNAs, an active and versatile group in cancers. Int J Oral Sci. 3, 165-175, Oct 2011.
75. M. D. Robinson and G. K. Smyth, “Moderated statistical tests for assessing differences in tag abundance.” Bioinformatics, vol. 23, no. 21, pp. 2881–7, Nov. 2007.
76. Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X. et al., “Transcriptome sequencing to detect gene fusions in cancer.” Nature, 458, 97–101, Mar 2009.
77. Mardis, E. R. “The impact of next-generation sequencing technology on genetics.” Trends in genetics, 24, 133–141, Mar. 2008.
78. Mardis, E. R. “The impact of next-generation sequencing technology on genetics.” Trends in genetics, 24, 133–141, Mar 2008.

79. Mardis, E. R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 9, 387-402, 2008.
80. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M and Gilad, Y. “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.” Genome research, 18, 1509–1517, Sep 2008.
81. Mehra, M and Chauhan, R. Biomark Cancer. 9, 1179299X17737301, Nov 2017.
82. Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., et al., “PGC-1alpha-responsive genes involved in oxidative phosphorylation are co-ordinately downregulated in human diabetes.” Nature genetics, 34, 267–273, Jul 2003.
83. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., Wold, B. “Mapping and quantifying mammalian transcriptomes by RNA-Seq.” Nature methods, 5, 621–628, Jul 2008.
84. Murphy, L. D., Herzog, C. E., Rudick, J. B., Fojo, A. T and Bates, S. E. “Use of the polymerase chain reaction in the quantitation of mdr-1 gene expression.” Biochemistry, 29, 351–356, Nov 1990.
85. Nordin, M., Bergman, D., Halje, M., Engstrom, W., Ward, A. Epigenetic regulation of the Igf2/H19 gene cluster. Cell Prolif. 47, 189–199, Jun 2014.
86. Oberdoerffer, P., Michan, S., McVay, M., Mostoslavsky, R., Vann, J., Park, S.-K., et al., “SIRT1 redistribution on chromatin promotes genomic stability but alters gene expression during aging.” Cell, 135, 907–918, Nov 2008.
87. Okoniewski, M. J and Miller, C. J. “Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations.” BMC bioinformatics, 7, 276, Jan 2006.
88. Pagliarulo, V., Datar, R. H and Cote, R. J. “Role of genetic and expression profiling in pharmacogenomics: the changing face of patient management.” Current issues in molecular biology, 4, 101–110, Oct 2002.
89. Pan, Q., Shai, O., Lee, L. J., Frey, B. J and Blencowe, B. J. “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.” Nature genetics, 40, 1413–1415, Dec 2008.
90. Park, S. G., Schimmel, P. and Kim, S. Aminoacyl tRNA synthetases and their connections to disease. Proc Natl Acad Sci USA, 105, 11043-11049, Aug 2008.
91. Patalano, S., et al., CXCL12/CXCR4-Mediated Procollagen Secretion Is Coupled to Cullin-RING Ubiquitin Ligase Activation. Sci Rep, 8, 3499, Feb 2018.

92. Poliseno, L., et al., A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 465, 1033-1038, Jun 2010.
93. Rapaport, F., Khanin, R., et al., Erratum to: Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 16, 261, Nov 2015.
94. Roberts, A., Pimentel, H., Trapnell, C., Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 87, 2325-2329, Sep 2011.
95. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al., Integrative Genomics Viewer. Nature Biotechnology29, 24–26, 2011.
96. Robinson, M. D., Smyth, G. K. "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." Biostatistics, 9, 321–332, Apr 2008.
97. Rodriguez-Nieves, J. A and Macoska, J. A. Prostatic fibrosis, lower urinary tract symptoms, and BPH. Nat Rev Urol. 10, 546-550, Sep 2013.
98. Rodríguez-Nieves, J. A., Patalano, S. C., et al., CXCL12/CXCR4 Axis Activation Mediates Prostate Myofibroblast Phenoconversion through Non-Canonical EGFR/MEK/ERK Signaling. PLoS One. 11, e01594902016, Jul 2016.
99. Rodríguez-Nieves, J.A., Patalano, S.C. et al., "CXCL12/CXCR4 Axis Activation Mediates Prostate Myofibroblast Phenoconversion through Non-Canonical EGFR/MEK/ERK Signaling." PLoS One. 11, e0159490, Jul 2016.
100. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., et al., "Nucleotide sequence of bacteriophage _X174 DNA," Nature, 265, 687–695, Feb. 1977.
101. Schena, M., Shalon, D., Davis, R. W and Brown, P. O. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," Science, 270, 467–470, Oct 1995.
102. Schena, M., Shalon, D., Davis, R. W and Brown, P. O. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," Science, 270, 467–470, Oct 1995.
103. Shendure, J and Ji, H. "Next-generation DNA sequencing." Nature biotechnology, 26, 1135–1145, Oct. 2008.
104. Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., et al., "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." Proceedings of the National Academy of Sciences of the United States of America, 100, 776–781, Dec 2003.

105. Slater, G. S and Birney, E. “Automated generation of heuristics for biological sequence comparison.” BMC bioinformatics, 6, 31, Jan 2005.
106. “Specification sheet: cBot - Fully automated clonal cluster generation for Illumina sequencing,” http://res.illumina.com/documents/products/datasheets/datasheet cbot.pdf.
107. Spurlock, C. F., Crooke, P. S and Aune, T. M. Biogenesis and Transcriptional Regulation of Long Noncoding RNAs in the Human Immune System. J Immunol. 197, 4509-4517, Dec 2016.
108. Tariq, M. A., Kim, H. J., Jejelowo, O and Pourmand, N. “Whole-transcriptome RNAseq analysis from minute amount of total RNA.” Nucleic acids research, 39, e120, Oct 2011.
109. “Technology spotlight: Illumina sequencing technology - Highest data accuracy, simple workflow and a broad range of applications,”
110. http://res.illumina.com/documents/products/techspotlights/techspotlight sequencing.pdf.
111. Thiery, J. P., Sastre-Garau, X., Vincent-Salomon, B., Sigal-Zafrani, X., Pierga, J. Y., Decraene, C. et al., “Challenges in the stratification of breast tumors for tailored therapies.” Bulletin du cancer, 93, E81–E89, Aug 2006.
112. Trapnell, C and Salzberg, S. L. “How to map billions of short reads onto genomes.” Nature biotechnology, 27, 455–457, May 2009.
113. Trapnell, C., Pachter, L and Salzberg, S. TopHat: discovering splice junctions with RNA-Seq. 25, 1105-1111, May 2009.
114. Trapnell. C, Roberts. A, et al., “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.” Nat Protoc, 7, 562-578, Mar 2012.
115. Trapnell. C., et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 28, 5, 511-515, May 2004.
116. Tripathi, R., et al., Unraveling long non-coding RNAs through analysis of high-throughput RNA-sequencing data. Noncoding RNA Res. 2, 111-118, Jun 2017.
117. Trivedi, U.H., Cézard, T., et al., Quality control of next-generation sequencing data without a reference. Front Genet. 5, 111, May 2014.
118. Van Verk, M. C., Hickman, R., Pieterse, C. M., Van Wees, S. C. “RNA-Seq: revelation of the messengers.” Trends in plant science, 18, 175–179, Apr 2013.

119. Van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M. et al., "Gene expression profiling predicts clinical outcome of breast cancer." Nature, 415, 530–536, Jan 2002.
120. Velculescu, V. E., Zhang, L., Vogelstein, B and Kinzler, K. W. "Serial analysis of gene expression." Science, 270, 484–487, Oct 1995.
121. Wang, J., Wang, H., Zhang, Y., et al. Mutual inhibition between YAP and SRSF1 maintains long non-coding RNA, Malat1-induced tumorigenesis in liver cancer. Cell Signal. 26, 1048–1059, May 2014.
122. Wang, Z., Gerstein, M and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10, 57-63, Jan 2009.
123. Wang, Z., Gerstein, M. and Snyder, M. "RNA-Seq: a revolutionary tool for transcriptomics," Nature Reviews Genetics, 10, 57–63, Jan 2009.
124. Watson, J and Crick, F. "A Structure for Deoxyribose Nucleic Acid," Nature, 171, 737–738, Apr 1953.
125. Westermann, J., Gorski, S. A and Vogel, J. "Dual RNA-seq of pathogen and host," Nat Rev Microbiol, 10, 618–630, Sep 2012.
126. Wheeler, D. J and Burrows, M. "A block-sorting lossless data compression algorithm," Digital SRC Reports 124, 1994.
127. Wilhelm, B. T and Landry, J. R. "RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing." Methods, 48, 249–257, Jul 2009.
128. Yao, P and Fox, P.L. Aminoacyl-tRNA synthetases in medicine and disease. EMBO Mol Med, 5, 332-343, Mar 2013.
129. Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., et al., "HTQC: a fast quality control toolkit for Illumina sequencing data." BMC bioinformatics, 14, 33, Jan 2013.
130. Yu, J., Marsh, S., Hu, J., Feng, W., Wu, C. The Pathogenesis of Nonalcoholic Fatty Liver Disease: Interplay between Diet, Gut Microbiota, and Genetic Background. Gastroenterology Research and Practice. 1–13, 2016:2862173. doi: 10.1155/2016/2862173 2016.
131. Zeeberg, B., Feng, W., Wang, G., Wang, M., Fojo, A., Sunshine, M., et al., "GoMiner: a resource for biological interpretation of genomic and proteomic data," Genome Biology, 4, R28, 2003.
132. Zhang, D.Y., Zou, X.J., Cao, C.H., et al., Identification and Functional Characterization of Long Non-coding RNA MIR22HG as a Tumor Suppressor for Hepatocellular Carcinoma. Theranostics, 8, 3751-3765, Jun 2018.

133. Zhang. Y, et al., Critical effects of long non-coding RNA on fibrosis diseases. Exp Mol Med. 19, 50, 1, e428, Jan 2018.
134. Zhao, S., Fung-Leung, W. P., et al., “Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells.” PLoS One, 9, Jan 2014.
135. Zhao, S., Fung-Leung, W.P., et al., Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 16, 9, Jan 2014.
136. Zhong, S., Storch, K. F., Lipan, O., Kao, M. C., Weitz, C. J., Wong, W. H, “GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space.” Applied bioinformatics, 3, 261–264, Jan 2004.