Development of Web-Application for High-Throughput Sequencing Data and In Silico Dissection of LINE-1 Retrotransposons in Cellular Senescence

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftliche Fakultät

der Universität zu Köln

Vorgelegt von

Ankit Arora

aus Vadodara, India

Köln 2018

Berichterstatter: Prof. Dr. Peter Nürnberg
(Gutachter) Prof. Dr. Michael Nothnagel

Prüfungsvorsitzender:

Tag der mündlichen Prüfung: 24th July 2018

Abstract

Next-generation sequencing (NGS) technologies have been remarkably advantageous in opening new paths for scientific research. The large-scale generation of data has resulted in the advancement of tools for data analysis, but the requirement to implement increasing amount of tools in a systematic way is still a vast question. In the initial analysis of distant NGS datasets, we observed the need for an efficient framework for downstream analysis of NGS data. Therefore, the web application titled "RanchNGS" was developed. This particular web-framework aims to understand the downstream and integrative analysis for various sequencing data such as RNA-Seq, Ribo-seq and additionally ChIP-seq. RanchNGS benefits from an efficient analysis in a reduced time frame and without the need of any advanced programming skills. Furthermore, the role of transposable elements (TEs) in cellular senescence was revealed by studying distinct NGS technologies.

TEs are segments of DNA that have the potential to be in-motion by jumping from one location in the genome to another. Due to this reason, they are also called as "jumping genes." Long Interspersed Nuclear Element-1 (LINE-1 or L1) are the only autonomous TEs currently active in human and non-human primate genomes. Our study primarily targeted a broad extent of L1 elements. The result of excessive oncogenic stress in immortalized cells undergoing senescence were investigated. The current study focused on the transcriptional and post-transcriptional mechanisms that confer the cross-talk between L1 and host-defense machinery. It was observed that the transcriptional and post-transcriptional regulation of L1 elements are diminished as cells undergo senescence that takes precedence to the response activation of L1 elements at RNA and protein level. Moreover, we hypothesized that excessive oncogenic stress in immortalized cells could be related to L1 retrotransposition, triggering double- strand breaks and enabling cells to enter permanent cell cycle arrest.

Acknowledgements

I would like to express genuine gratitude to my supervisor Prof. Dr. Peter Nürenburg for his continuous support, suggestions and knowledge during my PhD research.

I would also like to thank Prof. Dr. Michael Nothnagel for his guidance, analytical suggestions and for reviewing my thesis. I would like to thank Dr. Karsten Heidtke for fruitful discussion during my scientific research. I would like to thank Dr. Holger Thiele, Heinrich Rohde for their assistance in implementation of my web-framework over University server.

I would like to express deeply thanks to aDDRess ITN (Marie Curie Initial Training Network).

I would like to cherish special thanks to my former colleague and friend Kerstin Neubert for her constant help and support during my PhD work.

I am highly thankful to my extraordinary friends, Vikas Bansal and Manvendra Singh. It would have never been possible to accomplish without their extreme guidance and encouragement during ups and downs of my PhD research. I am also thankful to all my colleagues and to all my friends. I would like to express immense thanks to all the people who helped me achieve my goals throughout my scientific growth.

Finally, I am extremely thankful to my parents Usha Arora, Dr. Rajendra Kumar Arora and my sister Anju for their consistent support, endeavor, love and affection.

Abbrevations

A Adenine

Alu Arthrobacter luteus

ADR Adenosine deaminase acting on RNA

APOBEC apolipoprotein B mRNA editing enzyme

A to I Adenosine to inosine

BWT Burrows wheeler transform

BAM Binary alignment map

BED Browser Extensible data

BJ Human fibroblast cell line

C Cytosine

C to U Cytosine to uracil

ChIP Chromatin Immunoprecipitation

CSS Cascading style sheets

cDNA Complementary DNA

CNV Copy number variation

DNA Deoxyribonucleic acid

DSBs Double strand breaks

EREs Endogenous retroelements

FDR false discover rate

FPKM Fragments per kilobase per million mapped reads

FASTA FAST-ALL

G Guanine

GFOLD Generalized fold change

GLM Generalized linear model

GATK Genome analysis toolkit

GTF Gene transfer format

GO Gene Ontology

HERV Human endogenous retroviruses

HS Homosapiens

HTML Hypertext Markup language

HIV-1 Human immunodeficiency virus

KRAB Kruppel associated box domain

KRAB-ZFP Kruppel associated box zinc finger protein

KAP1 KRAB associated protein 1

LINE-1 Long Interspersed nuclear elements

LTR Long terminal repeats

mRNA messenger Ribonucleic acid

MNase Micrococcal nuclease

Myrs Million years

MEME Multiple Em for Motif Elicitation

NGS Next-generation sequencing

ORFs Open reading frames

OIS Oncogene-induced senescence

Pol II RNA Polymerase II

PHP Hypertext preprocessor

rRNA Ribosomal Ribonucleic acid

RNA Ribonucleic acid

RNP Ribonucleoprotein

RPKM Reads per kilobase per million mapped reads

SINE Short Interspersed elements

SNP Single nucleotide polymorphism

SNPiR Identify SNPs in RNA-seq data

SAM Sequence alignment map

SASP Senescence associated secretory phenotype

SETDB1 set-domain protein 1

SNV Single nucleotide variants

T Thymine

TCGA The Cancer Genome Atlas

tRNA Transfer Ribonucleic acid

TERT Telomerase reverse transcriptase

TPRT Target primed reverse transcription

TSS Transcription start site

TES Transcription end site

TEs Transposable elements

TSDs Target site duplications

TC Total count

TPM Transcript per kilobase million

TRIM28 Tripartite motif-containing 28

UTR Untranslated region

UCSC University of California, Santa Cruz

UQ Upper quantile

VNTR Variable number tandem repeat

QC Quality control

List of Figures

Contents

Chapter 1

Introduction

1.1 Central dogma of biology

Deoxyribonucleic-acid (DNA) is a complex molecule that carries hereditary information in all living organisms. The structure of DNA is a double helix, which is composed of two strands that are arranged in opposite orientation. It is composed of pairs of chemicals called nitrogenous bases attached together with sugar and phosphate molecules. Jointly nitrogenous base, sugar and phosphate molecules are known as a nucleotide. Each nucleotide in DNA is made up of four nitrogenous bases i.e. adenine (A), cytosine (C), thymine (T), guanine (G). These bases are categorized into purine and pyrimidine bases. Adenine and guanine are grouped into purines and cytosine and thymine are grouped into pyrimidines [1]. The bases are linked with each other via hydrogen bonds, whereby A pairs with T and C pairs with G.

The DNA molecule combined with histone proteins are wrapped into a structure called chromosome. Humans have 23 pairs of chromosomes, in total of 46 chromosomes. Out of these, 22 pairs are known as autosomes and the remaining pair called as sex chromosomes. In it, one pair of chromosomes inherited from the paternal and the other pair of chromosomes inherited from the maternal [2]. Telomeres are an essential part of the chromosome and they are located at the ends of a linear chromosome. Genes are made up of a sequence of DNA or RNA, which encodes for a functional protein [3]. Genes are considered as transcription units that can correlate with different polypeptide chains with different functions [4]. In distinction from DNA, Ribonucleic acid (RNA) is usually single stranded. The RNA structure consists of ribose, nitrogenous bases and phosphate group. Whereas RNA nucleotide also contains adenine, cytosine and guanine, thymine is replaced by uracil (U). Mainly there are three types of RNA, ribosomal RNA (rRNA), messenger RNA (mRNA) and

transfer RNA (tRNA). According to the central dogma of molecular biology, the genetic information in DNA is transcribed to messanger RNA (mRNA) by the process of transcription. The segment of DNA transcribed into mRNA is known as a mature transcript [5]. The processed mRNA is further translated into proteins by the process of translation. The essential enzyme for maintaining gene expression that synthesizes mRNA in all eukaryotes is known as RNA polymerase II (pol II). Pol II is regarded as integral foundation of transcriptional machinery [6]. Pol II interacts with generic transcription factors during preintiation complex (PIC) and with different factors in course of elongation and termination [7]. The RNA pol II structure is well compatible for promoter binding and its initiation helps in relating with other transcription factors to create the pre-initiation complex [8].

In the process of translation the message coded on mRNA is converted into a sequence known as protein. A sequence of three bases in mRNA is referred to as a codon. The genetic information contained in the codon represents a particular amino acid. The ribosome complex plays an important role in protein synthesis. The ribosome structure is made up of protein and ribosomal RNA (rRNA) and consists of two subunits: the larger unit of ribosome is called 60-S and the smaller unit is called 40-S. The two subunits of ribosome are connected together and bind with mRNA to form a protein. Transfer RNA (tRNA) performs an essential function in carrying specific amino acids to proteins. The translation process is divided into three phases: initiation, elongation and termination. In initiation, the ribosome mobilizes the start codon AUG which is in the P-site to form the initiation complex. In elongation, at distinct time each codon is read from the mRNA and the corresponding amino acid is added to the protein chain. In termination, when a stop codon (UAG, UAA or UGA) enters the ribosome, events are generated that allow the polypeptide chain to be liberated from the tRNA [9].

1.2 Cellular senescence and DNA damage response

A range of genes involved in controlling cell growth and cell division. The cells are divided in regular fashion through mechanism of cell-cycle. The continuous progression of cell-cycle ensures intact replication of DNA by repairing defective DNA. During cell-cycle, checkpoint responses play an essential role in maintaining DNA integrity. In case of checkpoint loss, damaged DNA can be accumulated and further leads to mutations. DNA damage checkpoint responses consist of upstream protein kinases such as ATM and ATR, downstream kinases such as CHK1 and CHK2 [10]. There

are three checkpoints in the cell cycle, first is at end of the G1 phase, second one is at the transition of G2 and M phase and third one is at the M phase which generates formation of two daughter cells.

1.2.1 Cellular senescence

In 1961, Hayflick showed that normal cells have limited capacity to divide in vitro which is described as cellular senescence. There are various factors like DNA damage, oxidative stress and oncogenic stress which leads to senescence. With the onset of oncogenic stimuli, cells are unable to divide and enter a state of irreversible cell cycle arrest [11]. Cellular senescence is regulated by two tumor suppressive pathways, in particular the p53/p21 and p16INK4a/pRB pathways [12]. p16INK4a has been considered vigorous senescent biomarker in mouse and human tissues [13]. Apart from changes in gene expression, there are other vast changes which also take place during Cellular senescence namely increase in mRNA levels, secretion of numerous cytokines, chemokines, growth factors and proteases. These are referred to as senescence associated secretory phenotype (SASP) [12]. A fraction of chemokines and cytokines in SASP can induce cells of immune system. Natural killer cells, macrophages and T cells are among those allowing elimination of senescent cells. Whereas another characteristic of SASP include inciting inflammation [14]. Suppression of the cancer development and promotion of tissue repair are two important functions covered by cellular senescence and SASP [15, 16]. Expression of human telomerase reverse transcriptase (hTERT) is known to cause immortalization in human fibroblasts cells [17].

The structures located at the ends of linear chromosomes are known as telomeres, which gets shortened with each cell division. Continuous cell division without the presence of telomeres makes dysfunctional telomere. During DNA replication, incomplete replication of chromosome ends stimulates telomere shortening [18]. Further, dysfunctional telomere are able to evoke defect in DNA resulting as DNA damage response (DDR) by restraining already pursued DNA repair. Telomere shortening is considered as a cancer-driving process.

There are numerous possible causes and consequences of cellular senescence. In regard to causes, cellular senescence is described as a process of stress response and it is also known to be associated by various mechanisms like telomere shortening, potent oncogenic stimuli causes breaks in the epigenome, while certain tumor suppressor mechanism are also activated by oncogenic stimuli. In regard to consequences, typically irreversible growth arrest can restrain tumor suppression mechanism. Potential

senescent cells phenotype may also stimulate tissue repair and perhaps can also be involved in the development of cancer and ageing [12, 14, 19] as shown in (Figure 1.1).

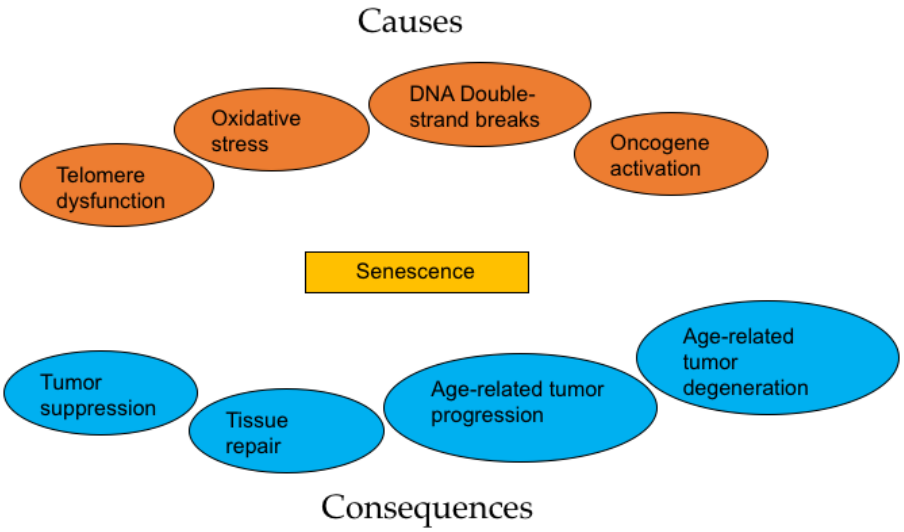


Figure 1.1: Schematic illustration of causes and consequences of cellular senescence. The ovals in orange color indicates causes of senescence and ovals in grey color indicates consequences of senescence. Figure adopted from [14] and modified.

1.2.2 DNA damage response

In human body, each cell obtains thousands of DNA lesions per day. The loss in the repair of these DNA lesions may result in the obstruction of transcription, replication, mutagenesis and cellular cytotoxicity. The network of recognizing, signalling and repairing DNA lesions are known as DNA damage response pathway [20]. Upon DNA damage, senescent cells can respond in multiple ways likewise as tumor suppression mechanism by disrupting the division and growth of cancer cells. Moreover, other strategies include tissue repair and enabling cells to aging. In order to maintain the DNA integrity, cell cycle checkpoints and DNA repair pathways are activated to prevent harmful mutations.

1.2.2.1 Sources of DNA damage response

There are numerous forms of lesions which leads to DNA damage in human genome.

1) DNA damage by cellular cytotoxicity: This type of damage belongs to the category of endogenous cellular processes. It includes: a) oxidation of bases by production of reactive oxygen species b) alkylation of bases such as formation of 7-methylguanine c) hydrolysis of bases such as deamination, depurination and depyrimidination d) mismatch of bases: it occurs due to errors caused by DNA replication. After certain threshold variety of mistakes caused by replication are no longer identified as errors. This type of mutation are referred to as base-pair substitution mutation. Base-pair substitution comprise of replacement with one purine to another or one pyrimidine to another known as transitions. In another case replacement of purine with pyrimidine are referred to as transversions. Similarly when replication errors are not rectified, they tend to evolve mutation. This type of mutation constitute of insertion, deletion, duplication, inversion and translocation where DNA fragment endure changes [21].

2) DNA damage by external agents: This type of damage belongs to the category of exogenous cellular processes. It comprise damage due to ultraviolet radiation (UV), ionizing Radiation (IR), other radiation effects including x-rays, gamma-rays. These type of radiation can damage DNA in form of single-strand breaks (SSBs) and double-strand breaks (DSBs) [22].

Consecutively with involvement of p53 tumor suppressor, DDR arrests cell cycle and cell cycle checkpoints are activated [14]. In turn, checkpoint activation halts the cell cycle and allows time for DNA repair to take place. If unrepaired DDR pathway triggers Cellular senescence or apoptosis. Stimulation of DDR pathway in senescent cells has been considered as biomarker for further studies [18]. DDR has been related to numerous health diseases. Moreover, infection of viruses along with retroviruses can also stimulate DDR [20]. Cleavage of sugar phosphate in the DNA backbone induces single strand breaks or DNA DSBs, which act as catalyst for DDR [23]. DSBs are regarded as rigid form of genotoxic stress. Telomere erosion is constrained in accelerating DSBs [24]. Activation of oncogenes in normal cells lead to cellular senescence which is also referred to as oncogene induced senescence (OIS) [23]. DSBs are constrained in the progression of defective telomeres. DSBs are subjected to affect the integrity of the genome. In which cell-cycle progression gets arrested when cells react to DSBs in-turn generating DNA damage checkpoint response. DNA damage checkpoint responses are reproduced by telomere-initiated senescence.

3) DNA damage by L1 retrotransposition: During the integration process of L1 elements, endonuclease (EN) activity encoded by ORF2 protein creates two nicks for Double stranded DNA. If the nick is made above certain threshold or if nick is at site

the damaged DNA. Then, it would lead to delay or failure of repair complex recruitment. However, L1 retrotranspsosition (Section 1.4) has the potential in inducing DNA damaging event [25, 26].

1.2.3 DNA repair mechanisms

In response to each type of DNA lesions, the cell has developed relevant repair process to overcome generated lesions. DNA damage is either repaired by direct repair system or excision repair. DNA polymerase such as polymerase-δ are involved in the repair caused by replication errors. Base-excision repair (BER) are associated with repair of damaged DNA caused by cellular metabolism such as reactive oxygen species, methylation, deamination and hydroxylation. It comprise of DNA glycosylases and AP endonuclease [27]. The repair mechanism for the damaged DNA caused by exogenous agents such UV light, IR are referred to as nucleotide excision repair (NER). The defects caused by NER includes xeroderma pigmentosum, cockayne syndrome. It occurs two NER pathways namely global genome NER (GG-NER) and transcription coupled NER (TC-NER). Interestingly, NER takes place in the cells whose genes are actively involved in the transcription on DNA. Consecutively, it inhibits elongation of RNA polymerase [27]. The selected genes involved in repair of NER consist of XPA, XPG, ERCC1-XPF, XPC-RAD23B [28]. During DNA replication, mismatches of bases occurred in the process are repaired by mismatch repair (MMR). Defective MMR has been identified in Hereditary nonpolyposis colon cancer (HNPCC). MMR assist in eliminating nucleotide mispaired by DNA polymerases and insertion/deletion loops can be found during replication or recombination [27].

DSBs are considered to be risky, if not repaired can induce genomic rearrangements. Whereas, there occurs two types of mechanisms to repair DSBs 1) non-homologous end joining (NHEJ): It involves the association of broken ends. It requires the proteins which can recognize the damaged ends and bring them together for ligating such as Ku is the well-known essential protein. 2) homologous recombination repair (HER): In it the broken ends are repaired by using one of sister chromatid or homologous chromosome as template. Two of the crucial proteins associated with homologous recombination are BRCA1 and BRCA2 [29].

1.2.4 The p53 signaling pathway

The p53 plays crucial role in the tumor suppressive DNA damage response.Therefore, p53 also known as a tumor suppressor gene [30]. p53 acts as a transcriptional regula-

tor by maintaining various types of stresses like to activate cell cycle arrest or induce apoptosis in suppression of tumor development. p53 helps in sustaining genome integrity. In unstressed or normal cells, p53 is in very unstable form. During which p53 forms complex with mouse double minute 2 (MDM2) or human double minute 2 (HDM2) to ensure degradation of p53 protein in presence of E3 ubiquitin mediated ligase [31]. Upon different forms of stress such as oxidative stress, oncogene activation, DNA DSBs, hypoxia resulting in DNA damage, p53 is activated by increasing its expression levels and DNA repair mechanism are triggered. In it p53 induces DNA repair genes,if repaired then it goes back and cell cycle continuous. If not repaired, p53 induces two different processes such as senescence and apoptosis (state of programmed cell death). Thus, p53 is also known to conserve stability of the human genome. But when p53 is inactivated, then damaged DNA progresses continuously without check, then cancerous cells are formed. p53 has been specified as one of the mutated genes found in about half of all human cancers [32]. The mutations which are found commonly in human tumor cells are missense mutations [33]. p53 has been considered best known biomarker in Cancer cells [34]. p53 has been referred to as "Gate-keepers of cells" by deciding either to eliminate damaged cells or arrest those damaged cells. Due to its well-known regulation in timely functions, p53 has been cited in more than 40,000 research papers [35]. p53, a tumor suppressor gene is involved in the suppression of retrotransposon activity [36]. In p53 mutant cells, it has been shown that DNA damage response can be caused by L1 retrotransposition [37] in turn proceeding to apoptosis [38]. DNA double strand breaks (DSBs) are initiated by L1-encoded endonuclease which might related to L1 retrotransposition [26].

1.3 Transposable elements

In 1940s, Barbara McClintock discovered Transposable elements (TEs) in maize plants.TEs are distinct pieces of DNA which move from one location in the genome to another. Therefore, they are also referred to as "jumping genes". The early theory about TEs was rejected by scientific community. They were treated as "Junk DNA". It took a long time before it got accepted. TEs have known to account for nearly half of the human genome (Figure 1.2). TEs are now being acknowledged as substantial part of human genome,exhibiting crucial role in processes along with gene expression regulation, genomic instability and genome evolution [39, 40]. TEs are well-known to regulate gene expression in mammalian genomes.Thus it's also referred as "controlling elements" [41–43]. TEs have the significant characteristic to regulate the

expression of nearby genes, whereby controlling both at transcriptional as well as post-transcriptional level [44]. TEs transcription and integration are known to be activated due to stress [45]. TEs are well integrally distributed in heterochromatin, particularly in centromeres and telomeres. Due to their entanglement in heterochromatinization, TEs constitute to play a role in gene regulation. TEs possess two attributes which distinguish them from other genetic factors. One characteristic of those attribute is changing the genetic environment, in which they are considered to be mobile. The other attribute resembles the peculiar capacity to multiply themselves during the retrotransposition process [46]. TEs are categorized into two groups: DNA transposons and retrotransposons [42, 47].

1.3.1 DNA transposons

DNA transposons constitute about 3 % of the human genome (Figure 1.2). They tend to act through a "cut and paste mechanism", and are categorized under Class II transposable elements [49]. Whereas currently they are no longer active in the human genome, but there is evidence that the last family was active back during primate evolution of 37 Mya ago [40]. Majorly, DNA transposons are classified into three evolutionary groups: anthropoid specific (40-63 Mya), primate specific (64-80 Mya) and eutherian wide (80-150 Mya) [50].

1.3.2 Retrotransposons

Retrotransposons are also known as RNA transposons. Retrotransposons, mobilize through copy and paste mechanism by use of RNA as an intermediate.They are categorized under Class I transposable elements [49, 51]. Replication of retrotransposons occurs through reverse transcription of RNA and integration of emerging cDNA into another location [52]. Retrotransposons are further categorized into two classes: Long Terminal Repeats (LTR) and non-LTR elements.

1.3.2.1 Long Terminal Repeat(LTR) retrotransposons

Retrotransposons involving LTR elements are referred to as Human Endogenous RetroViruses (HERVs). LTR retrotransposons contain Gag: group-specific antigen, Pol: polymerase, Env: envelope protein (Figure 1.3). HERVs constitute about $\tilde{8}$.3 percent of the human genome(Figure 1.2) [40, 55]. HERVs existed around 25 million years ago in our ancestors, mostly in apes and Old World monkeys. The HERVs which have been associated with other diseases comprise HERV-K, HERV-H, HERV-W [56].

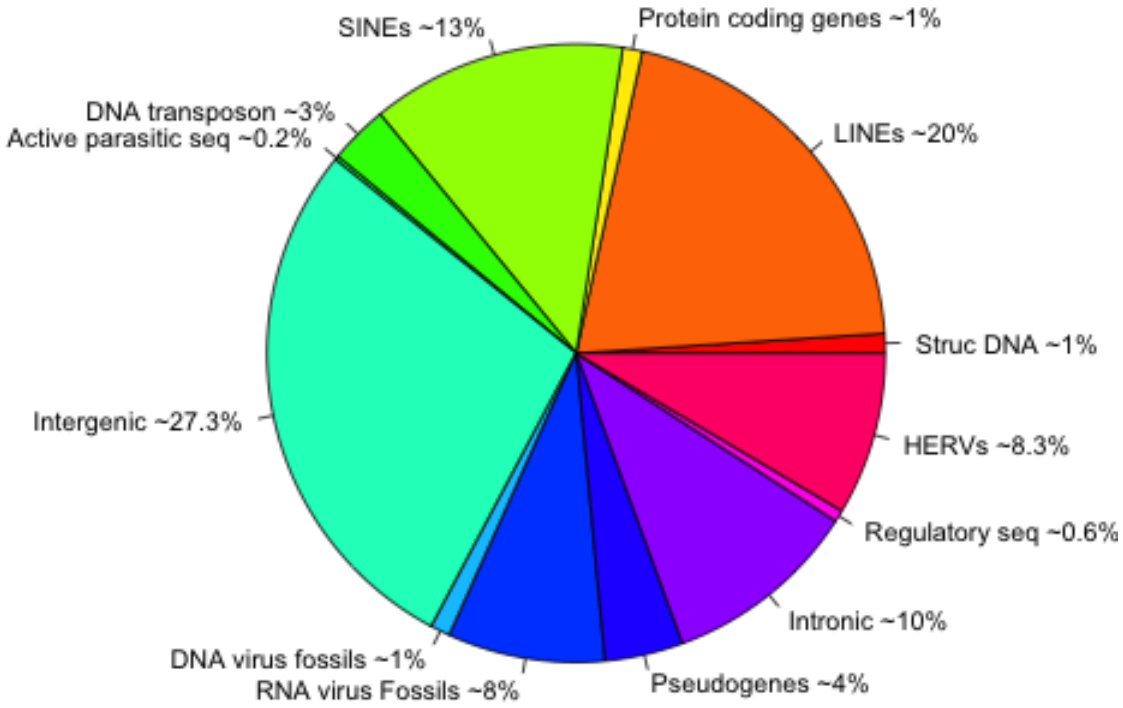


Figure 1.2: Distribution of genomic content of human genome (hg19). Figure drawn based on [48].

HERVs are considered to be the cause of autoimmune rheumatic diseases. Among them mainly are Rheumatoid Arthritis (RA) and Systematic Lupus Erythematosus (SLE) [57]. During early embryogenesis, human specific Endogenous retroelements (EREs) transcription has been repressed mainly by KRAB/KAP1 system through histone methylation, histone deacetylation and DNA methylation [58]. EREs uses sequence-specific mechanisms for the silencing of retrotransposable elements [59]. Endogenous retroviruses (ERVs) are also responsible for configuring p53 transcriptional network in a species-specific manner. Approximately 1500 human ERVs LTR ele-

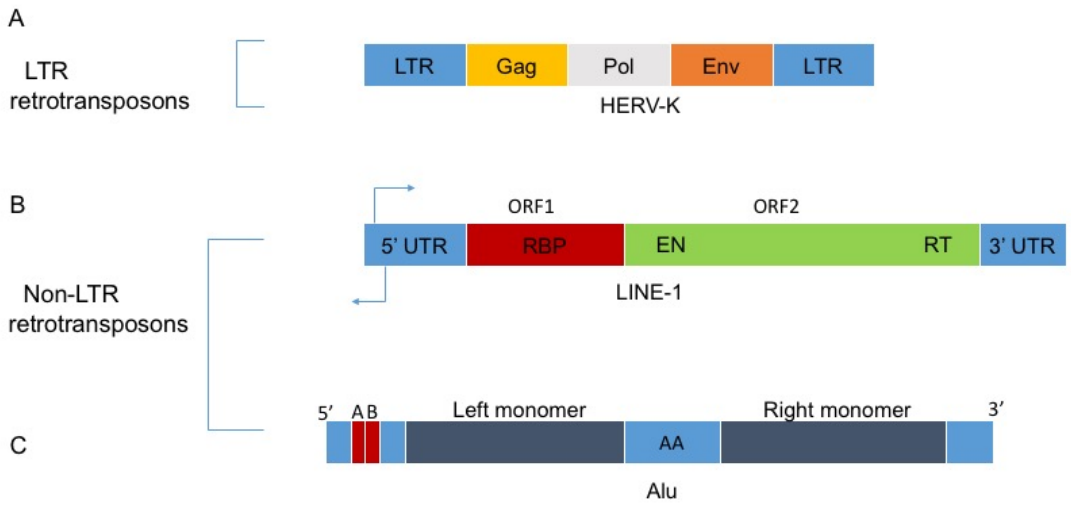


Figure 1.3: Classification of transposable element and its structure. (A) Human endogenous retroelements-K (HERV-K) belongs to class of LTR retrotransposons consists of two long terminal repeats (LTR) along with Gag,Pol and Env. (B) L1 elements are specified to class of non-LTR retrotransposons. L1 structure comprised of two open reading frames (ORFs) along 5'UTR and 3'UTR (C) Alu element also belongs to non-LTR retrotransposons. Alu structure consists of left and right monomer connected with A-rich linker zone. Figure drawn based on [53, 54].

ments are adjoined by p53 DNA binding sites. The pattern of LTR10 and MER61 families in ERV [60] have been attributable to anthropoids which include New World monkeys (squirrel monkey and marmoset) and Catarrhini (Old World monkeys i.e rhesus monkey and apes i.e. humans). The families of LTR10 and MER61 are linked to provirus HERVIP10 and HUERS-P3B respectively. Since primates, LTR elements were established in LTR10 and MER61 families. The p53 site was existent in LTR elements of LTR10 retroviruses family, but later it got extinct over a period of time. Whereas the MER61 family showed absence of p53 site but later it obtained p53 site through mutations [61]. HERV-K elements are the recent one to be shown active in various genomes of catarrhines (Old World monkeys, apes and humans). HERV-K113 is one of the prominent possibly provirus still active in humans. HERVs are considered to be a residual of ancient germ-cell infections [62].

1.3.2.2 Non-LTR retrotransposons

Non-LTR retrotransposons, as the name distinguish does not constitute of Long terminal repeats. Non-LTR retrotransposons are characterized into autonomous and non-autonomous retrotransposons [63]. Long Interspersed nuclear elements (LINE-1 or L1) are classified under autonomous retrotransposons (In detail Section 1.4). Alu and SVA elements are classified under non-autonomous retrotransposons [54]. Alu and SVA elements together constitute of about approximately 13 % of our genome [64] (Figure 1.2) . Alu and SVA elements belongs to the Short interspersed Elements (SINE) retrotransposon family [65] Active Alu elements in humans are around 300 base pairs (bp) in length [66]. Alu elements consist of enhanced regions which are gene-rich whereas L1 elements consist of enhanced regions which are gene-poor. Alu elements are also capable retrotransposition,which are attributed as master or source elements. Alu retrotransposition are also distinctly altered by APOBEC proteins. RNA polymerase III (Pol-III) plays an important role in retrotransposition of Alu elements [67]. In evolutionary structure, Alu elements belong to the class of primate specific which are $\tilde{6}5$ million year old [68]. Alu structure comprises of two monomeric sequences derived from signal recognition particle (SNP) which is part of 7SL RNA gene [40]. Left monomer of Alu structure consist of internal RNA polymerase III promoter, divergent from right monomer that consists of adenosine-rich sequence.During diverse periods of evolutionary history,different families of Alu elements have been evolved. Ancient Alu elements were identified as J subfamily, recent (Middle-aged) Alu elements are categorized into S subfamily (Sx, Sq, Sp, Sc) and youngest Alu elements subfamily are classified under Y subfamily: Ya5 and Yb8 are most prevailed in humans. Alu elements are known to have extensive influence during gene-expression, splicing and RNA-editing. Insertion of Alu elements into human genome have been indicated as a contribution in human genetic diseases [69, 70]. The SVA structure have been comprised of Alu sequence, variable number of tandem repeats (VNTR), and a sequence inferred from 3' end of HERV-K10 element (SINE-R) [54]. In mobilization, SVA elements have been pretended to benefit from L1 protein mechanism.

1.4 Long Interspersed Nuclear Elements 1 (LINE-1 or L1)

L1 belongs to non-LTR retrotransposon family. In humans, L1 retro-transposons are currently known active and jumping along neuronal differentiation, embryonic

and cancer development. Germline and somatic tissues have exhibited presence of active L1 elements involved in affecting genome integrity [71–73]. L1s constitute of about approximately 20 % of the human genome [48, 54, 64] (Figure 1.2). A full-length active L1 element are around 6 kb in length. The human full-length L1 element structure constitute of 5'untranslated region (UTR), two open reading frames (ORFs) and 3'UTR confining with polyadenylation signal and a poly A (AATAAA) tract (Figure 1.3). 5' UTR encompasses significant RNA polymerase II (polII) as internal promoter, responsible for inducting L1 transcription. The ORF1 which consists of 40 kDa protein (ORF1p) encodes RNA binding protein [74, 75]. Whereas ORF2 consists of 150 kDa (ORF2p) protein encodes endonuclease and reverse transcriptase activities [40, 76–80]. In an average human genome contain of about 80-100 L1s, capable of retrotransposition [81].

Since around 100 million year (Myr) ago, L1 elements have been evolving and replicating in mammalian genomes.During human evolution,L1 elements has evolved into different L1 subfamilies from L1PA16 to L1PA1 (L1Hs) [PA depicts primates] with Homosapiens specific L1 element (L1Hs) being the youngest one [82–84]. Particularly in our ancestral genome, two-primate specific KZNF genes have emerged to inhibit of diverse families of non-LTR retrotransposons. ZNF91 have sustained structural changes and linked to repress SVA elements. ZNF93 have evolved to repress primate-specific L1 lineage until 12.5 Myr ago. While in L1PA3 subfamily, firstly it created a deletion of 129-bp and since then liberated from ZNF93 restriction. Although, the deletion is present in half of L1PA3 elements, in which a shorter version in L1PA3-6030 subgroup while a longer version in L1PA3-6130 subgroup. Nevertheless, intact binding site of ZNF93 is present in L1PA4-L1PA6 families [85].

1.4.1 L1 retrotransposition

L1 retrotransposons move inside the human genome through jumping. They are also known as "jumping genes". L1 jumping can result into a mutation, if its inside a gene. L1Hs insertions have emerged during human evolution and may be raised in current human population [86]. L1 insertion which are known in causing diseases are essential in understanding mechanism about change in gene expression by retrotransposons [80]. A conventional L1 integration site in target DNA consist of 5'-TTTT/AA-3' [76, 87, 88]. Insertions associated with L1 and Alu elements are involved in causing alteration in gene expression and further yields to human diseases [71] [89]. The role of L1 in tumor initiation has been discussed, through the identification of somatic L1 insertion in colo-rectal cancer (CRC) by disturbing adenomatous polyposis coli

(APC) exon [90]. The advent of sequencing, allowed us to look in detail for somatic L1 insertion and second L1 insertion was reported after 20 years in APC exon for CRC which showed insertions are responsible for initiating tumorigenesis [91]. In p53-adequate cells, L1 activity can also help in initiating DNA damage leading cells to apoptosis. In this case, p53 function to inhibit such L1 insertion by protecting human genome. L1 insertion inside an exon can tend to decrease gene expression [37]. In L1Hs family,p53 contrarily effects expression of L1 elements [36].

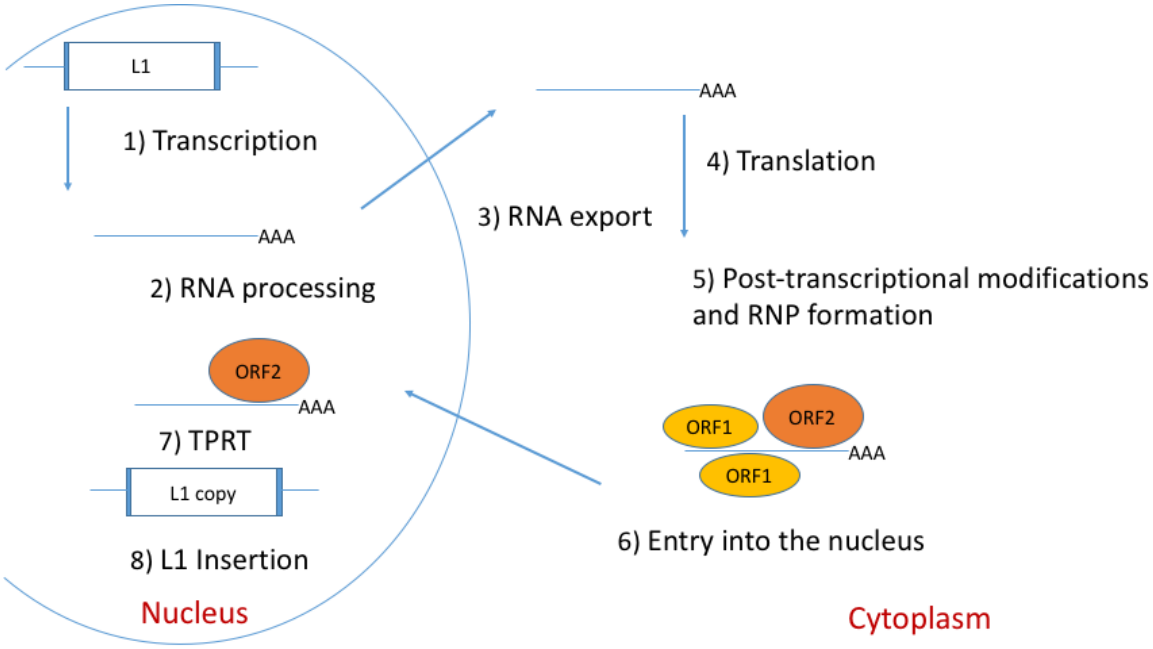


Figure 1.4: Mechanism of L1 retrotransposition.L1 element in nucleus is transcribed and exported to cytoplasm for translation. Along with L1 RNA, ORF2 (colored in orange) and two ORF1 (colored in yellow) molecule forms RNP complex. ORF2 protein linked with L1 RNA is further imported in nucleus from cytoplasm, in it L1 RNA is reverse transcribed and integrated into the new genomic site by TPRT process.Figure re-drawn based on [92] and modified.

The translation of ORF1 and ORF2 encoded proteins takes place in the cytoplasm, wherein it accumulates into ribonucleoprotein particle (RNP) encoded by L1 RNA transcripts [93–95] (Figure 1.4). RNP is considered one of the essential component of L1 retrotransposition. RNP is exported from cytoplasm into nucleus for L1 integration within human genome, occur by the process of target primed reverse transcription (TPRT) [54, 71, 79, 96–98] (Figure 1.5). Whereas RNP and TPRT carries out essential role in L1 replication [99, 100].In L1 retrotransposition process it is considered that endonuclease cuts genomic DNA releasing 3' hydroxyl, which

plays important role as primer for further reverse transcription of L1 RNA by reverse transcriptase [76, 101].

L1 retrotransposition (Figure 1.4) tend to occur during development of cells, here described as somatic mosaicism [73, 102]. L1 retrotransposition is capable of causing human diseases [76]. L1 retrotransposition tend to take place mainly in early embryonic development [103, 104]. L1 elements are also reported to be active in the brain, mainly in the hippocampal region [105]. Previous studies have shown during the initial stage of neuronal differentiation, Sox2 expression can play vital role in controlling L1 transcription and retrotransposition [106]. Although, L1 retrotransposition have demonstrated to be higher in neural progenitor cells (NPCs), derivative from human embryonic stem cells (hESCs) [107]. Since L1 retrotransposition are also involved in causing particular genetic variation in humans [108]. During TPRT, ERCC1/XPF performs an verification role at cDNA synthesis step [109]. Integration of L1 elements at DNA damage sites have shown in deriving repair of DNA by process of non-homologous end-joining (NHEJ) in human cells [97] [110]. Endonuclease activity of ORF2 in L1 element has shown its relevance with DSBs by trigerring various cellular responses inclusive of apoptosis, cellular senescence, cell-cycle checkpoints and DNA repair responses [111].

1.5 Transcriptional regulation in L1 retrotransposons

[Krüppel-associated box domain(KRAB)- associated protein-1] (KAP1) also known as TRIpartate motif-containing protein 28 (TRIM28). TRIM28/KAP1 has been linked in arrangement with different zinc finger proteins family (ZFPs) of Kruppel associated protein family (KRAB) [112–116]. TRIM28/KAP1 have also displayed interactions in gene repression [58, 87]. TRIM28/KAP1 which consist of RING-B box-coiled-coil (RBCC) motif and it's constituted of RING finger, two B box zinc fingers and coiled coil domain. TRIM28 cooperates with KRAB to be inducted into DNA [87, 117]. The two adjoining domain namely PHD (Plant homeo domain) and BR (Bromo domain) are present at C-terminus and play major role in trans-repression of target genes. BR domain interacts with Mi/NuRD complex through sumoylation and SETDB1 Histone methyltransferase that advances through trimethylation of Histone 3 on lysine 9 (HeK9) on chromatin.TRIM28/KAP1 associates with area of heterochromatin through heterochromatin particle (HP1) [114, 118]. TRIM28/KAP1 have shown essential role in DNA repair mechanisms and transcriptional regulation [87]. KAP1 acts

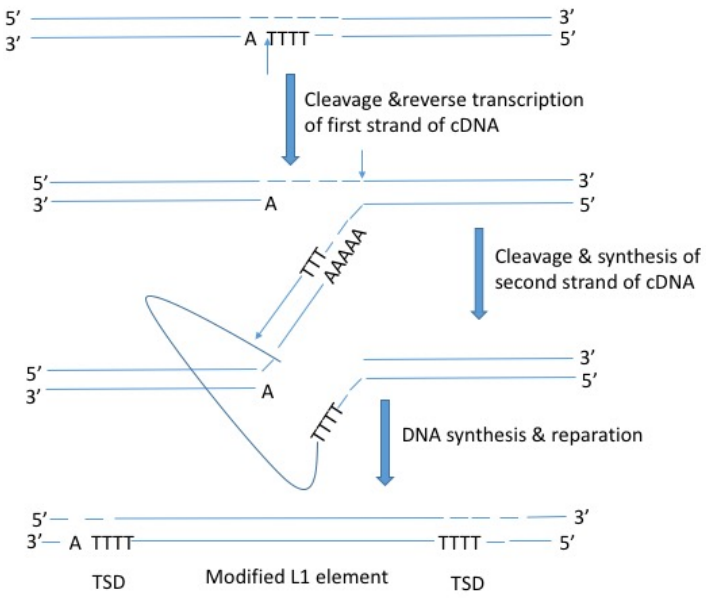


Figure 1.5: Diagrammatic representation of TPRT mechanism during L1 retrotransposition. In TPRT mechanism, L1 endonuclease cleaves first stand of target site by use of L1 RNA as template and executes reverse transcription of first strand of cDNA. The next cleavage and synthesis of second strand of cDNA occurs few bp downstream in comparison to first cleavage, with the use of first strand of cDNA as template. After completion of DNA synthesis and reparation, modified L1 element is formed associated with TSDs. Figure re-drawn based on [79]

as transcriptional repressors in early embroyonic stem cells via through DNA methylation [58]. DNA methylation has been involved in silencing of retro-transposable elements in distinct somatic cell types [119]. Most of L1 elements are highly methylated in normal somatic cells. Although hypomethylation prevails in cancerous cells, by that it increases transcription of L1 elements. DNA methylation tend to be required in suppression of L1 promoter. Therefore, DNA methylation also plays an essential role in regulation of L1 elements [120].

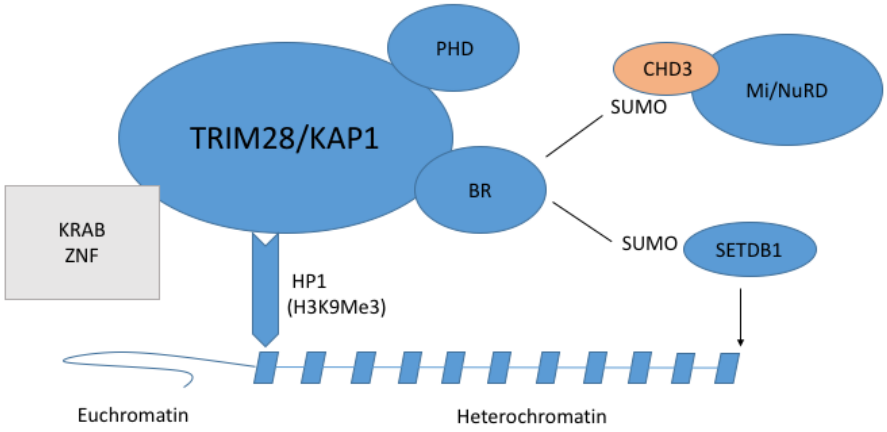


Figure 1.6: Typical illustration of transcriptional regulation mechanism by KAP1. Figure adopted from [87] and modified.

In an evolutionary arms-race of L1 element with KAP1/TRIM28, For every L1s sub-family may have been identified by KAP1 to KRAB-ZFP complex but thereafter KAP1 acquired mutation and binding got rejected. In recent L1 subfamilies, KAP1 has been recruited by sequence specific but may have acquired mutation restraining through base line expression. Youngest L1s (L1Hs) are remarkably transcribed and not yet identified by KRAB-KAP1 system. Recent finding exhibit youngest (L1Hs) lineage generates small RNAs known as Piwi-Interacting small RNAs (piRNAs) to regulate its expression through DNA methylation in human early embryonic stem cells. Furthermore, if L1 escapes the repression control from KAP1/TRIM28 transcriptionally then L1 retrotransposition has been noticed to be blocked post-transcriptionally by APOBECs proteins [58](Figure 1.7).

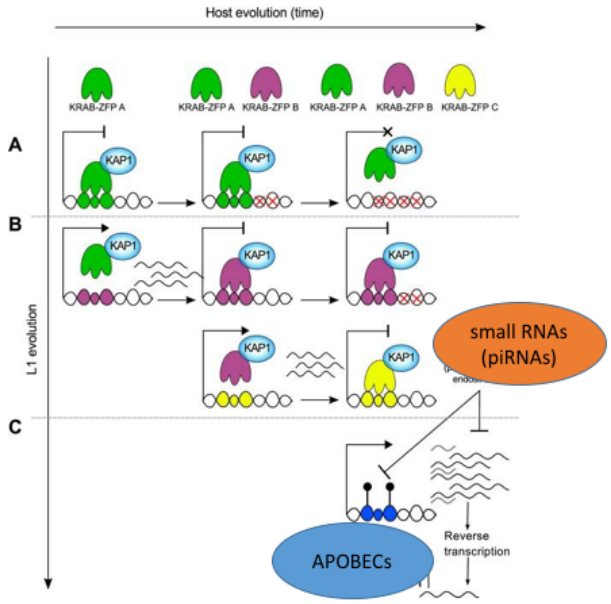


Figure 1.7: Typical illustration of evolutionary arms race between L1 elements and KRAB-ZFP. It shows the relation of L1 evolution with host evolution. Figure adopted from [58] and modified.

1.6 Post-transcriptional regulation in L1 retrotransposons

Epigenetics and post-transcriptional processes are the prerequisites for normal tissue development and gene expression [121]. Previous studies have shown the vast insight information can be gathered at RNA level.The pre-mature RNAs must undergo post-transcriptional processing to form mature RNAs. RNA molecules are involved in vast range of post-transcriptional modifications particularly 5' capping, 3'polyadenylation, RNA splicing and RNA editing[121–123].

1.6.1 Characterization of RNA editing

RNA editing has been considered one of the post-transcriptional modification process, which involves changes in the nucleotides of RNA molecule and initiates protein diversity. RNA editing has been studied a prerequisite for maintaining gene expression in mammals. RNA editing consists of two specific mechanisms like substitution editing and insertion/deletion editing. Substitution editing involves modifications of individual nucleotides. It constituted two types of substitution editing particularly A-to-I editing which include conversion from A to I nucleotide performed by ADAR

(adenosine deaminases acting on RNA) family [124, 125] of enzymes and the other one involving C-to-U editing [126, 127] which include conversion from C to U nucleotide executed by cytidine deaminase enzyme for APOBEC (apolipoprotein B mRNA editing enzyme) family of enzymes [128, 129]. ADAR enzyme family has been mostly associated in brain tissue [130]. After the editing of C-to-U, sheared protein ApoB48 is transported to the cytoplasm [127].The other type Insertion/Deletion editing constitute of insertion or deletion of RNA nucleotides. RNA editing has been related to diseases like epilepsy, brain tumor, amyotrophic lateral sclerosis (ALS) and other cancerous associated diseases [128] [129]. Activation-induced-cytidine deaminases (AID) are involved in regulation of somatic hypermutation (SHM) and class-switch recombination (CSR) activities. AID has been related in activating DNA DSBs [131]. APOBECs family play a pivotal role in innnate immune response to infection by retroviruses and also in regulation of other viruses like HIV-1, hepatitis B virus, human T cell lymphotropic virus and retrotransposons [132–136]. APOBECs illustrate major part in regulation of L1 elements (Figure 1.7). In respect to non-LTR retrotransposons such as L1 elements, at post-transcriptional phase APOBECs has been involved in inhibition of L1 retrotransposition before integration process into genome [109, 137–139]. APOBECs family of proteins has been involved in inhibition of viruses and retrotransposons [140]. By virtue of APOBEC3B entanglement in deamination of cytidines at single-stranded (ss) DNA. APOBECs is well-known to be expressed in various types of cancer cell lines. APOBECs is considered as a huge cause of mutations in different types of cancers. These mutations are accelerated by enzyme responsible for cytosine-to-uracil deamination [141]. RNA-sequencing (Whole transcriptome sequencing) has proven to be distinguished mechanism in RNA-editing analysis [142, 143]. ADAR is an constituent of adenosine deaminases acting on RNA family. The enzymes from these families are subject to deamination of adenosine(A) to inosine(I). Lately, studies have shown adequate portion of ADAR gene are commonly to L1 elements [144].

1.6.2 Chimeric transcripts

Transcriptomics profiling have made possible detection of Chimeric transcripts. Chimeric transcripts are described to form as result from fusion of two genes [145] (Figure 1.8). Gene fusion has been characterized a mechanism in which whole or parts of sequences from two or more different genes are conjugated to form chimeric transcript resulting in rearrangements of DNA or RNA. TEs likely promote gene fusions by reconciling gene duplications, straightly engaged in fusion and supporting

the access of reverse-transcriptase required in transposition [146]. Gene-fusion arching to chimeric gene or transcript are root cause of translocations in distinct types of leukaemias, malignent disorders and sarcomas [147, 148]. Chimeric transcripts are considered as potential biomarkers in deeper understanding for cancer progression and development [149]. In our study, we aimed to understand the formation between full length L1 elements and genomic regions in senescent cells distinguished from immortalized ones.

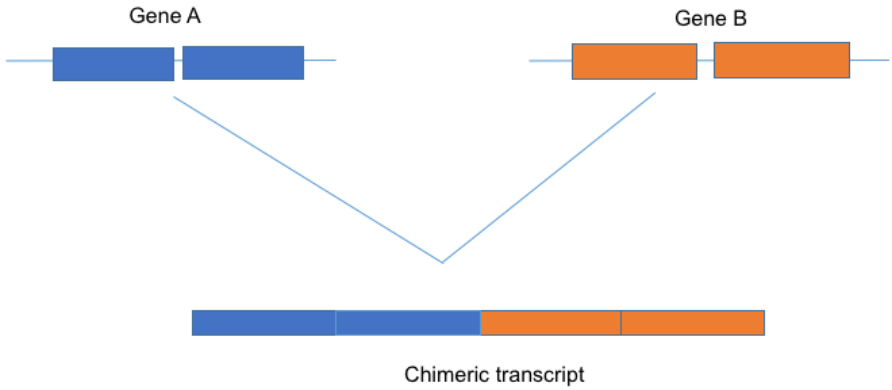


Figure 1.8: Conventional illustration of chimeric transcript formed with fusion of exons from different genes. Gene A showcased in blue color while gene B showcased in orange color.

1.7 Next-generation-sequencing

In 1977, Sanger and his colleagues developed the cutting-edge method for obtaining nucleotide sequence of DNA. This method is based on the chain termination method also known as Sanger sequencing.The first genome sequenced from this technique was from the bacteriophage ϕX174 [150]. Sanger sequencing was established on the principle of dideoxynucleotide triphosphates (ddNTPs), identical to deoxynucleotides (dNTPs) [151].In 1977 Maxam and Gilbert delivered another method for sequencing DNA and their method was based on a chemical procedure [152].Sanger sequencing is also referred to as "first generation sequencing" technology.

In 2001, the international human genome sequencing consortium (IHGSC) and celera genomics announced the draft sequence of human genome. The IHGSC planned to finish the process of human genome sequencing over a three-year period [64, 153]. In 2004, with the successful completion of the first human genome sequence under the IHGSC, marked the beginning era of sequencing technology referred to as next-generation sequencing (NGS) [154]. In the same year the national human genome research institute (NHGRI) planned a program to reduce the cost of human genome sequencing to It initiated the development of various sequencing platforms over the last decade: 454, Illumina, SOLiD, Ion torrent and Pacific Biosciences (PacBio). These platforms were also referred to as "second generation sequencing" technologies [155]. There are numerous methods involved in sequencing technologies such as template preparation, sequencing and data analysis [156]. The first NGS technology available on the market was the 454 system. This approach is comprised of the mechanism of emulsion PCR and pyrosequencing. In case of pyrosequencing approach, fragmented DNA is isolated, ligated with adapters and separated into single strands. Whereas emulsion PCR is used for clonal amplification [157, 158]. In 2005, 454 technology based on the pyrosequencing method was developed by Life sciences (now Roche) [155]. In 2006, Solexa launched the Genome Analyzer (GA) and later in 2007, it was acquired by Illumina. It is based on the mechanism of sequencing by synthesis [159]. Currently, Illumina prevails the NGS market by manufacturing a group of sequencers namely MiSeq, NextSeq 500 and HiSeq series [160].

In 2007, another NGS technology, sequencing by oligonucleotide ligation and detection (SOLiD), was developed by applied biosystems (now Life Technologies). It is based on the mechanism of sequencing by ligation. Color space coding is the main feature obtained by the sequencing from the SOLiD system [155]. For sequencing, fragments are amplified by emulsion PCR along with small magnetic beads [161].Helicos biosciences is the first group to release the sequencer HeliScope, based on the mechanism of single molecule sequencing [162]. The main advantage of this technology is that it is not influenced by any bias in library preparation or during amplification. It is mainly applied to non-amplifiable samples [163]. The various sequencing technologies adapt different approaches to attach DNA to fragments like 454 and Applied Biosytems/SOLiD use coated beads to attach DNA whereas solexa/illumina and helicos bioscicences use chips to attach DNA [164].In the late of 2010, ion torrent liberated a compact sequencer known as ion personal genome machine (PGM) [165]. It is based on semiconductor sequencing technology. It is the first sequencing technology which does not use fluorescence and camera scanning. PGM is best suitable for

small labs and clinical applications [159]. A detailed comparison of these technologies is mentioned in Table 1.1.

Sequencing Platform	Mechanism	Read Length (bp)	Run time	Data(Gb)	Advantage	Disadvantage
Sanger	Dideoxy chain termination	400~900	20 min~3hr	1.9 ~ 84 kb	Long read length	High cost, low throughput
454 GS FLX (Roche)	Pyro sequencing	700	24	0.7	Long read length, fast	High error rate
HiSeq (Illumina)	Sequencing by synthesis	50-101 (SE,PE)	3~10 days	600	High throughput	Short read assembly
SOLiDv4	Ligation	50+50	7 for SE 14 for PE (days)	120	Short read assembly	Accuracy
Helicos HeliScope	Reversible terminator	32	8	37	Non-bias of templates	High error rate
Pacific Biosciences	Real time sequencing	3000	20 min	3	Extremely long read length	Highest error rate

Table 1.1: Characterization and relation of different sequencing platforms. The data is gathered from [156, 159].SE: Single-end reads, PE: Paired-end reads, bp: base pairs Gb: Gigabit, kb: Kilobit, min: minutes.

In early 2010, pacific biosciences (PacBio) launched a sequencer based on the approach of single-molecule-rReal-tTime sequencing (SMRT) [166]. It uses estimation of enzymatic reaction in real time. Together with nanopore sequencing [167], they both are referred to as "third generation sequencing" technology [159, 168]. The major benefit of PacBio over other sequencers is mainly long reads and less sequencing runtime [155]. In PacBio, a metal film containing thousands of tens-of-nanometer holes known as zero-mode waveguides (ZMW) is used and in it sequencing reactions are carried out on silicon dioxide chips [163]. Most NGS platforms have developed fairly over recent period of time, many of them are constructed to produce short sequence reads. Nanopore sequencing developed to overcome the drawback of other techniques attaining long sequence reads over a short period of time [167]. The protein MspA has shown greater importance in developement of nanopore sequencing technology [169]. Nanopore sequencing also uses a technique based on single molecule in real time. Although this technique does not require the prerequisite of library preparation, it is possible to directly sequence the DNA or RNA. The accessible sequencer on the market which adapts this technology, the MinIONTM, was launched

by oxford nanopore sequencing technologies [155, 170]. Since the advent of sequencing of the human genome and the study of further model organisms like mouse and *c.elegans* enabled researchers to illuminate the genetic foundations of complex human diseases. The main applications of sequencing technologies include whole genome sequencing, exome sequencing, RNA-seq, ChIP-seq, Ribo-seq and several others. Due to abundant decrease in the cost of sequencing have generated enormous amount of data in the area of genomics, genetics and transcriptomics. Consequently, the field of bioinformatics was revolutionized with the demand to develop diverse tools to perform data analysis. Despite immense success in NGS, to solve biological questions in clinical diagnostics is still a big challenge. The diverse applications of sequencing technology will propel new insights based on data-driven research in many fields.

1.7.1 Chromatin immunoprecipitation sequencing (ChIP-seq)

Chromatin immunoprecipitation (ChIP) followed by high throughput sequencing (ChIP-seq) has been extensively used as an approach to identify DNA binding proteins on a genome-wide scale. It involves identification of transcription factors or histone modifications [171–176]. The ChIP-seq has been quite influential in assessing siganture of chromatin promoters,insulators and enhancers regions [177, 178]. In a ChIP-seq experiment (Figure 1.9), cells grown in tissue culture are cross-linked using formaldehyde. The cross-linked cells are adapted in buffer to isolate nuclei. Nuclei are transferred to lysis buffer and sheared chromatin [179] is obtained by sonication or enzymatic digestion by using micrococcal nuclease (MNase) [180]. In immunoprecipitation antibodies are used against a specific transcription factor or a DNA-binding protein of interest to identify a particular DNA-protein complex. Afterwards reverse crosslinking is applied to separate the protein of interest from ChIP-DNA. The library is prepared from immunoprecipitated samples and further sequenced with the help of high-throughput sequencing. The derived reads from sequencing are used for mapping to a reference genome and in addition to identify protein binding sites. In relation to a ChIP-seq sample, preparation of a control sample is essential. An input sample is obtained by crosslinking and sonication but without immunoprecipitation [181, 182] (Figure 1.9.

Before the commencement of sequencing, chromatin immunoprecipitation has been used coupled with microarray (ChIP-chip). ChIP-seq has been introduced to overcome the limitations of noise and to achieve a higher resolution than in ChIP-chip. Both techniques are involved in the functioning of gene regulation [175]. Along

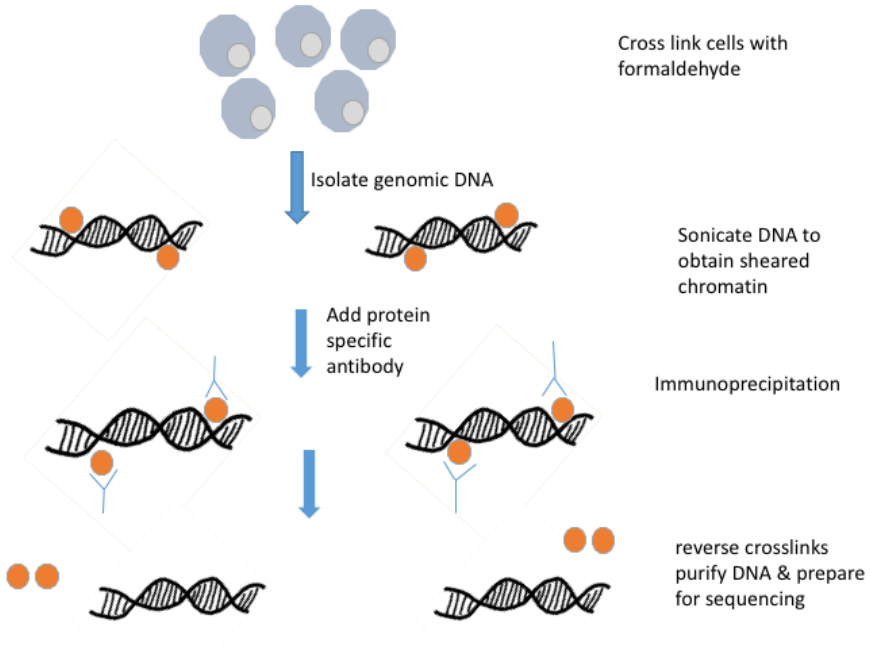


Figure 1.9: Schematic representation of Chromatin-Immunoprecipitaion followed by sequencing (ChIP-seq). Figure re-drawn based on [183]and modified.

with transcription factors, several histone modifications like histone trimethylated at lysine 4 (H3K4me3) methylation and H3K9ac acetylation are known to be enhanced near promoter regions [176]. As in regard to histone modifications, nucleosome structure shows an essential role in scattering of histone variants through the chromosome. It is always effective to compare ChIP-seq profiles of transcription factors to histone modification profiles [184].

1.7.2 Transcriptome profiling (RNA-seq)

The transcriptome is the entire set of transcripts in a cell. Transcriptomics is necessary in understanding the functional elements of the genome along with mRNAs, long non-coding RNAs and small RNAs [185, 186]. The massive amount of transcriptome data from several organisms can be obtained from the recent development of parallel cDNA sequencing (RNA-seq) [186–189]. RNA sequencing provides a robust mechanism in whole transcriptome analysis. Apart from its role in gene-expression, it comprises post-transcriptional regulation alike in alternative splicing, RNA editing, RNA degradation and translation [190]. In RNA-seq library preparation, total or long RNAs are converted into cDNA fragments by the process of reverse transcription and subsequent fragmentation. Then the sequencing adapters (grey) are added at one or both ends (Figure 1.10). After the amplification, fragments are sequenced.

Sequencing can be done at one end known as single-end sequencing or at both ends known as paired-end sequencing [185, 186]. The Primary application of the RNA-seq method was to obtain a transcriptome map in the yeast genome. RNA-seq analysis reported wide expression of the yeast genome, where 74.5 % referred to as RNA-seq tags [191].

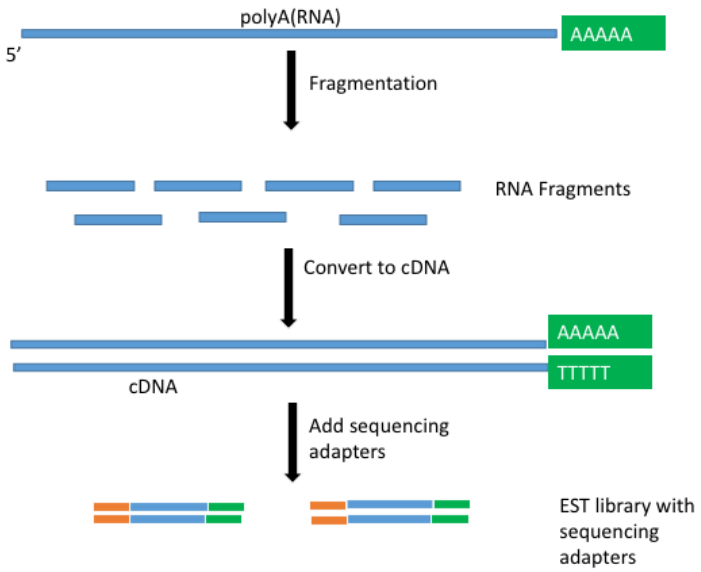


Figure 1.10: Overview of RNA-sequencing (RNA-seq) experiment. Figure re-drawn based on [185]and modified.

During transcription, single stranded RNA is synthesized from one of the two DNA strands. In RNA-seq generation, when RNA is copied back to cDNA, it is difficult to keep which of the two DNA strands was used to retrieve mRNA. To obtain this information the strand-specific RNA-seq protocol emerged. The standard RNA-seq library preparation protocol differs from that of strand-specific RNA-seq. Several methods developed for strand-specific RNA-seq are divided into two main types. The first type uses appending different adapters in a particular direction of the 5' end and 3'end of the transcript. The second type specifies a strand by chemical modification [192].

1.7.3 Ribosomal profiling (Ribo-seq)

Gene expression is the process which allows the conversion of information from DNA to a functional form known as protein. DNA microarrays have been used extensively

to study the regulation of gene expression at genome-wide, whereas RNA sequencing at mRNA level [193]. But none of the above techniques focuses on protein synthesis. Translation is the end process of gene expression which occurs in the ribosome. It was initially discovered in bacteriophage R17, exact translational sites by RNA-ribosome complexes [194]. Previously, the complex formation of the mRNA molecule and the ribosome revealed translational regulation which is also known as polysome profiling. Polysome profiling posed difficulty in determining the exact position between ribosome and mRNA when getting translated into protein. The method encountered a limited resolution and accuracy [195]. To overcome these limitations, ribosome profiling was introduced [196, 197]. The ribosome profiling approach along with high-throughput sequencing provides the key to monitor translation. Apart from monitoring translation it also identifies translated sequences [197]. Each footprint represents the specific regions of transcriptome attaining translation. Footprints mostly comprise of coding DNA sequences (CDS) which are also defined as protein-coding sequences. The sequence in footprint fragment obtained from ribosome gives exact location and description of mRNA getting translated on the transcript. The extent of gene translation is directly proportional to amount of distinct footprint fragments in deep sequencing [198].

During the protein synthesis, ribosome profiling is considered to be a measurement for gene expression. It identifies the exact positioning of ribosome and mRNA during translation. Ribosomal profiling has been considered equivalent to mRNA abundance profiling, due to its involvement in measurement of gene expression. Experimental steps of ribosomal profiling (Figure 1.11) include generating ribosome footprints, which is done after RNAase digestion of in-vivo translating polysomes. Ribosome footprints can be specifically related to each translating ribosome. The density of it can also play an important role in replacing mRNA abundance measurement to quantify gene expression. Furthermore, these footprints are obtained and converted into a DNA library by ligation of linker succeeded by reverse transcription and circularization PCR (Figure 1.11). After which the recovered cDNA libraries are further determined through deep sequencing [195, 197, 198].

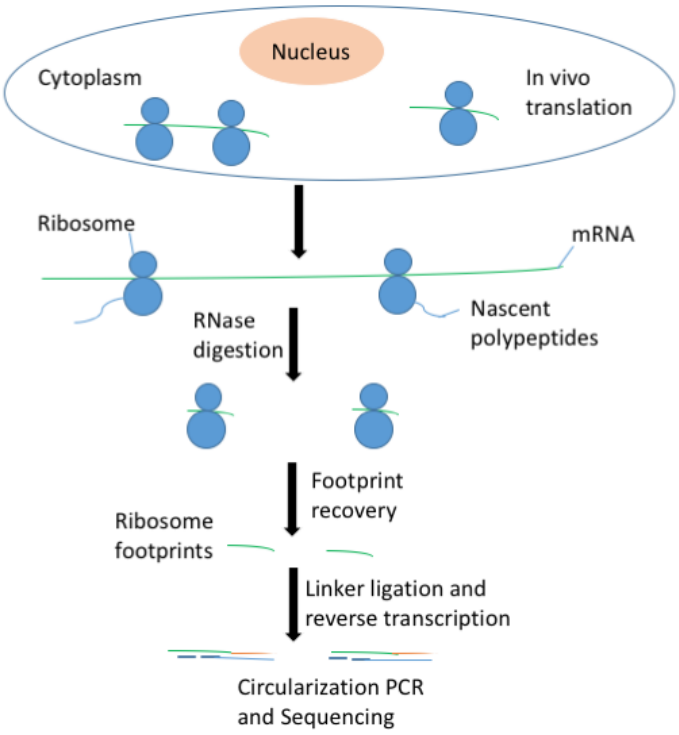


Figure 1.11: Diagrammatic illustration of ribosomal profiling experiment (Ribo-seq). Figure re-drawn based on [197] and modified.

1.8 Aims and motivation of the thesis

In comparison to Sanger sequencing, NGS technologies have proven to be faster and cost-effective. NGS technologies provide applications in diverse areas such as transcriptomics, epigenomics and genomics. An accurate downstream analysis is considered to be a critical task in NGS data analysis. In coalition to our extensive research with divergent NGS technologies, I speculate to profound a lack of platform which focuses on examining downstream and integrative analysis for NGS data. Hence, I demonstrate a web-based application for the downstream and integrative study of DNA, RNA and protein level. The distinct NGS applications applied in RanchNGS (Chapter 3 [Section 3.1]) web-interface includes exploration of transcriptomics data (RNA-seq), identification of protein-DNA binding sites (ChIP-seq) and studying translated RNA fragments (Ribo-seq). Furthermore, RanchNGS follows the principle of the central dogma of molecular biology at one platform. NGS discoveries have great influence in clinical applications by providing solutions for different diseases. The objective of this thesis focuses on to achieve and implement computational methods in

evaluating NGS datasets which can be contemplated in answering particular biological questions. I also applied RanchNGS during an elementary period of investigation in detecting function of L1 elements throughout cellular senescence by enactment with NGS datasets.

L1 elements are the only autonomous elements in the category of non-LTR retrotransposons. While other retrotransposons essentially Alu elements uses L1 machinery to be mobilized. In normal proliferative cells content of p53 is relatively high in order to restrain L1 activity [36]. p53, a tumor suppressor gene is referred as the guarding our genome due to its ability to restrain the L1 mobility, although the mechanism is vaguely understood. In our research for immortalized human primary BJ fibroblast cells (Chapter 3 [Section 3.2]), I aim to deeply understand if cells are subjected to severe oncogenic stress then what kind of changes occurs for L1 elements at transcriptional and post-transcriptional level. L1 expression can have distinct consequences on the host system like insertional mutagenesis further advancing to DNA double-strand breaks (DSBs). Insertional mutagenesis caused by active L1 or its open reading frame (ORFs) may cause adverse effects on the host system. Despite if L1 element does not land into the coding sequence of a gene during integration into the genome, it can still disturb its expression by altering the conformation of regulatory sequences around it, as manifested by the activation of proto-oncogenes. This genre of consequences due to L1 expression can percuss cells to enter cellular senescence or apoptosis via DNA damage [199, 200](Figure 1.12).

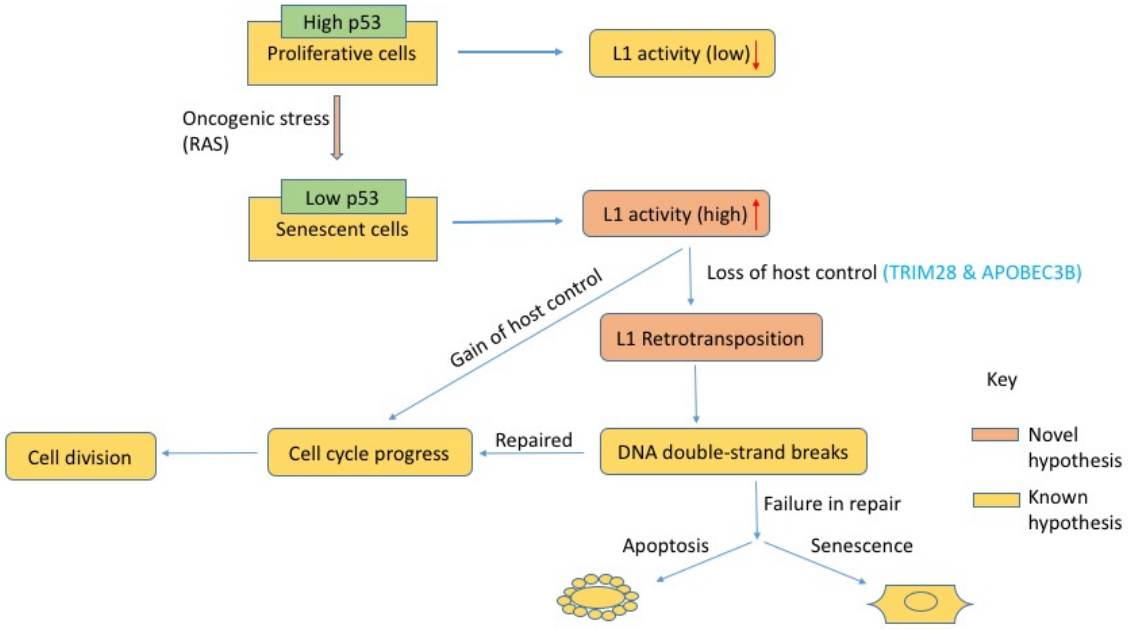


Figure 1.12: A typical diagrammatic representation of working hypothesis model.

Chapter 2

Methods

Next-generation sequencing (NGS) technologies have transformed measures in life science research to look into an extent of human diseases. Due to the rapid decline in the cost and time of sequencing. It has generated an immense amount of data and to make use of data it demands to overcome challenges in data analysis. The diverse NGS applications particularly transcriptomics study (RNA-seq), protein-DNA interactions (ChIP-seq), ribosomal profiling of mRNA fragment (Ribo-seq) and its analysis are described in this Chapter.

2.1 Alignment of sequencing reads to a reference genome

The vast development in next generation sequencing enabled the production of enormous amount of data to conduct studies that give detailed insights into a particular organism. During the analysis of next generation sequencing data, mapping of reads unto a reference genome is considered to be the most important and crucial step. The main goal of alignment is to find locations of newly sequenced reads with respect to a reference genome. Short reads obtained from a sequencing run are usually either from one end of a DNA fragment (single-end reads) or from both ends (paired-end reads). Mapping quality scores that are obtained from various algorithms for mapping of reads to a reference genome, assist in the evaluation of mapping. The mapping tools are classified into two types of algorithms either based on hash tables or Burrows-Wheeler-Transformation [201]. At present, alignment software is divided into two categories such as non-splice-aware aligner which include Bowtie [202], SOAP [203], BWA [204], MAQ [203], mrFAST [205] and SHRiMP [206]. Whereas splice-aware aligner include TopHat [207], TopHat2 [208] and STAR [209].Different alignment tools

assist in providing the same format for further downstream analysis. The sequence alignment map (SAM) format is supported by numerous tools. Binary Alignment Map (BAM) format serves as a binary form of SAM. BAM format consumes less space and easy to store in compressed file [204]

The hash table based algorithm uses the approach of seed and extend. It is based on the identification of common k-mer substrings from reads and reference genome as the first crucial step which is also known as seed detection. To further determine distinct locations of reads in a reference genome, another accurate algorithm is applied such as smith-Waterman [210] or needleman-Wunsch [211]. The aligner which apply the hash based algorithm are MAQ [203], mrFAST [205] and SOAP [203]. However, the hash-based algorithm identifies reads with errors by dividing reads into pieces [212]. The major drawback of software applying the hash-based algorithm is its dependence upon available large scale memory to index the human genome. To overcome this fast short read aligners Bowtie [202] and BWA [204] based on Burrows-Wheeler Transformation (BWT) [213] and FM-index have been developed. BWT applies a property usually used for compression, that uses reversible permutations of the characters in a string. It has a key strategy known as last-first (LF) mapping, in which the ith occurrence of character 'a' in L and the ith occurrence of character 'a' in F correlates to the same occurrence in T. The method used in the FM-index is the exact-match algorithm. The FM-index consists of an index linking the BWT with an ancillary data structure.

Splice-aware aligners like topHat [207] initially map reads to a reference genome with help of bowtie [202] and afterwards map reads to possible splice junctions. To achieve splice junctions, it first computes all sequences within the donor and acceptor sites and then uses the seed and extend approach [207] TopHat2 has been introduced to implicate development over previous version. Along with improvements in splice-alignment, TopHat2 also includes feature of read alignment across fusion breaks. TopHat2 is widely accepted and enhanced to perform well against data obtained from newer sequencing technologies producing longer reads [208]. TopHat2 [208] progresses read-alignment in three different ways 1) Transcriptome alignment 2) genome alignment 3) sliced alignment (Figure 2.1). Firstly, it performs Transcriptome alignment in which it benefits from available annotation file. Transcriptome alignment is considered highly agile in comparison to genome alignment. In case of transcriptome mapping, reads which are not present in annotation are mediated as unmapped reads. Tophat2 executes alignment of unmapped reads against genome. In it reads in extension of exons are aligned, while not in range of introns. Tophat2 uses

unmapped reads from genome alignment to further determine novel splice alignment [208] TopHat2 made known in generating sensitive and accurate alignment. STAR has been identified as highly efficient and faster in comparison to previous algorithms. STAR can also identify Chimeric (Fusion) transcripts [209].

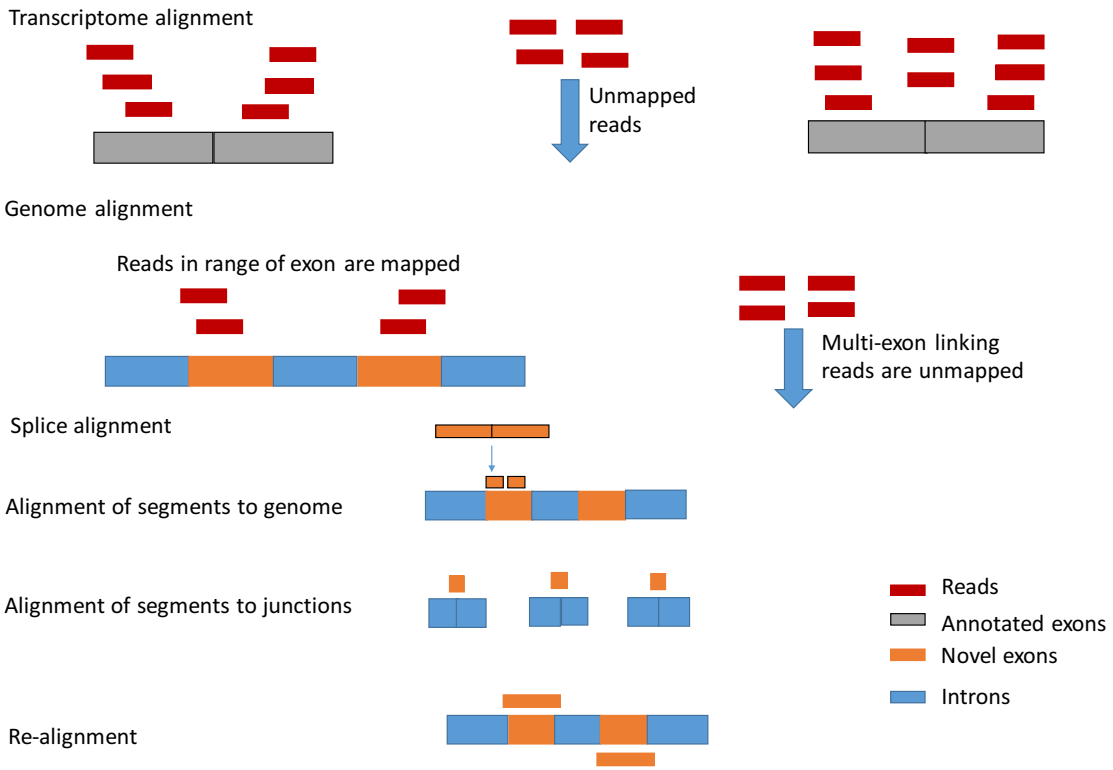


Figure 2.1: An Overview of Tophat2 workflow.Figure drawn based on[208] and modified.

2.2 ChIP-sequencing data analysis

Chromatin immunoprecipitation (ChIP-seq) in association with high-throughput parallel sequencing is a powerful technique for the identification of transcription factor binding sites and histone modification patterns. ChIP-seq is commonly known as an approach for recognizing genome-wide protein-DNA interactions. ChIP sequencing plays an essential role in epigenetics research. Short read sequences obtained from ChIP-seq are mapped onto a genome of interest as a prerequisite for detailed data analysis. The primary steps involved in the analysis are quality checking, mapping, peak calling, peak annotation and motif analysis.

2.2.1 Peak calling

After mapping of reads, estimation of regions in genome enhanced with mapped reads are referred to as proximate binding regions (peaks) [214]. Short read sequences from ChIP-seq data are also known as tags. Tags that are mapped to genomic locations are referred to as tag counts. In ChIP-seq, tag counts are retrieved from both strands of the genome, forward as well as reverse strand. The main application of the peak-calling approach is to retrieve the relation of the tag counts from both strands and identifying the true protein-DNA binding sites [215]. The genomic region where protein is bound has been represented as peaks. Model-based analysis of ChIP-seq data (MACS) [216] is a highly preferable algorithm in comparison to other peak-finding algorithms.MACS takes into account a control sample for the estimation of tag distribution. MACS determines tag distribution based on the Poisson distribution model. MACS splits forward and reverse tags from high quality peaks and aligns them to Watson and crick tag centers.The distance between these forward and reverse tags is measured as 'd'. In MACS, tags are shifted by d/2 in direction of 3'ends (Figure 2.2).

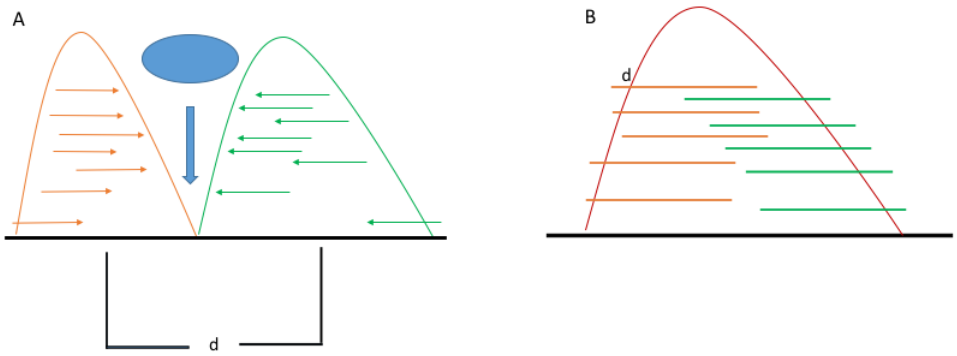


Figure 2.2: Peak modelling outline for ChIP-seq. (A) The bimodal distributions enclosing transcription factor binding site established from + and – sequence reads strand (Orange and green curves respectively) whereas distance (d) is studied as DNA fragment length formed by antibody pull down. (B) The enrichment signal of ChIP can also be attained from count of + and – sequence reads.Figure drawn based on [173].

Rather than using λ_{BG} for the whole genome, MACS applies λ_{local} for enriched region as:

$$\lambda_{local} = \max(\lambda_{BG},[\lambda_{1K}],\lambda_{5K},\lambda_{10K}\ )$$

Where λ_{1K},λ_{5K},λ_{10K} are for λ_{BG} predicted from 1kb, 5kb, 10kb windows centered at the peak location of the control sample.

The p-value in MACS [216] for each candidate peak is calculated by λ_{1K}. MACS calls candidate peaks for p-value lesser than user defined value (default:10^{-5}). The correlation between ChIP-seq tag count and local are known as fold enrichment [216, 217]. In a ChIP-seq analysis, the control sample plays an essential role in detecting enriched regions. It is usually referred to as input DNA and is prepared before Immunoprecipitation (IP). Thus, Input DNA is used as a positive control, whereas non-specific IP as IgG acts as a negative control. MACS evaluates the false discovery rate (FDR) for each candidate peak. It detects three kind of peaks such as sharp, broad and mixed. Sharp peaks are generally related to transcription factors at regulatory elements, while broad peaks are related to histone modifications [175] (Figure 2.3). In comparison to published peak-calling algorithm, MACS outperforms in peak results [216, 218, 219].

2.2.2 Peak annotation

The main aim of peak annotation is to find an association of enriched ChIP-seq peaks to various genomic parts like gene promoters, 5'UTR, 3'UTR, transcription start sites (TSS), intergenic regions, introns and coding exons [221]. The output peak file in BED format can also be uploaded for downstream analysis into the 'RanchNGS' Web-interface that have been developed in house as mentioned in (Section 3.2).

2.2.3 Gene ontology analysis

Furthermore, the peaks annotated with genes can be useful to attain different categories of gene Ontology(GO), associated with those particular genes. The GO terms are classified into biological process, Cellular components and molecular function [222]. This type of analysis can also be executed with help of 'RanchNGS'(Section 3.2).

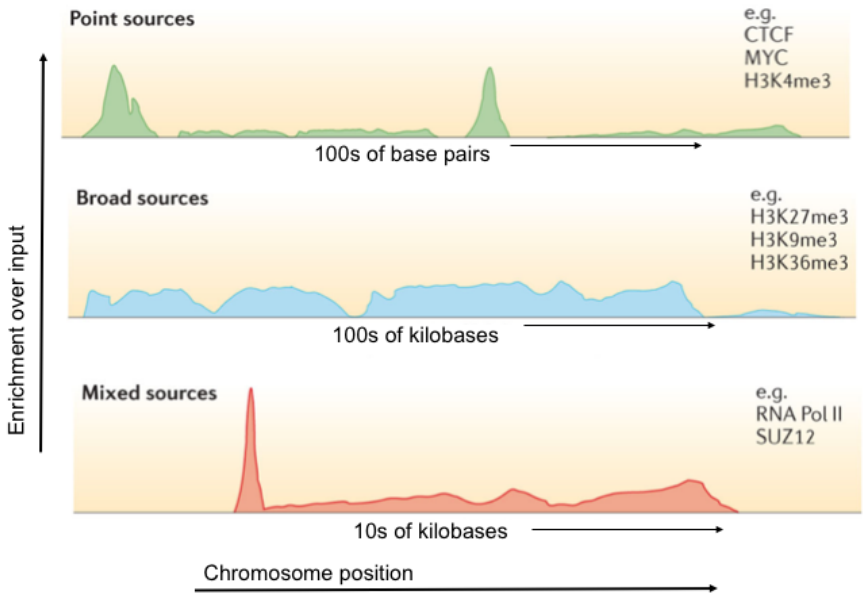


Figure 2.3: Representation of distinct types of ChIP-seq density profiles. Sharp peak profile shown in green for CTCF,H3K4me3. Mixed peak profile shown in red for RNA Pol II. Average sized broad peak profile shown in light-blue for H3K36me3, H3K9me3,H3K27me3. Figure adopted from [220] and modified.

2.2.4 Motif analysis

De novo Motif analysis helps in identifying DNA binding motifs for particular peak regions in a genome determined from a peak-calling algorithm [223]. It can also help in analyzing a motif that is found around a complex for a protein of interest. Besides comparison to previously known motif of protein from ChIP-seq experiment, it supports in validation of experiment through motif analysis [221]. MEME [224] is commonly utilized for recognition of DNA binding motifs.

2.3 RNA-sequencing data analysis

RNA sequencing (RNA-seq) enables high throughput sequencing of complementary DNA (cDNA) for in-depth study of transcriptomics. RNA-seq yields quantification of gene expression and identification of differentially expressed genes. Besides, RNA-seq is also involved in several post-transcriptional modification mechanisms like detection of splice-site junctions [225] and identification of RNA editing sites [226]. The algorithm designed for RNA-seq analysis for splice-site detection can also be used in determining gene-fusion events. In our research study we have applied all applications

of RNA-seq analysis such as gene-expression quantification, detection of differentially expressed genes and identification of RNA editing sites.

2.3.1 Gene expression quantification

Mapping of raw sequencing reads are performed either to reference genome or to transcriptome with a splice-aware aligner and non-splice-aware aligner respectively (Figure 2.4). The essentiality of normalization on mapped reads is to eradicate factors which are not determined by biological interests. The various normalization methods comprise of Total count (TC), Upper quartile (UQ), Reads per kilobase per million mapped reads (RPKM) or Fragments per kilobase per million mapped reads (FPKM). RPKM is related to single-end reads and FPKM is related to paired-end reads. RPKM and FPKM are considered as an essential element in estimation of transcript abundance [227]

$$RPKM_i = 109 * \text{Ci/l'i N}$$

In the shown formula N is the total number of mappable reads, Ci is the number of reads mapped to transcript i and l'i is the length of reads. Another important normalization method for estimation of gene expression is Transcripts per million (TPM) [228]. The main difference in normalization with TPM is to first normalize with gene length and then with sequencing depth. The sum of all TPMs in each sample are the same in case of TPM, whereas in RPKM and FPKM the sum for each sample is different. TPM is calculated as

$$TPM_i = 106 * \text{Z} * \text{Ci/l'i N}$$

Cufflinks [229] is a program that is available from the Tuxedo package. It takes mapped reads as input to provide information about expression quantification for RNA-seq data. It predicts transcript and gene abundances as RPKM and FPKM values for single-end and paired-end reads respectively [230].

2.3.2 Differential expression

In RNA-seq studies, another necessary aspect is to identify differentially expressed genes. It is attained by calculating statistical significance of differences in expression for each gene between different conditions. For instance, if we have two samples from the same patient with different conditions, one sample is from an affected part considered as treatment and another sample is from a healthy part considered as control.

We assume that both samples have different amount of RNA transcripts. The question that is often further investigated is which sample is having significantly higher or lower number of transcripts in between samples for different conditions. To address this matter, cuffdiff2 [229] was developed to evaluate expression at transcript level. Cuffdiff2 [231] calculates expression of transcripts for each condition by computing the number of their fragments. It also computes differential expression of transcripts. Cuffdiff2 is based on beta normal binomial distribution model.

Another popular method applied to the identification of differentially expressed genes is DESeq2 [232] which performs differential analysis based on two different inputs either in form of a count matrix or calculated by htseq-count at gene level (Figure 2.4). For count data, DESeq2 uses a generalized linear model (GLM) [233] to distinguish between treatment and control. DESeq2 uses negative binomial distribution [234] (it is also known as gamma-poisson distribution) to evaluate differentially expressed genes. DESeq2 applies gene ranking beyond a certain threshold of Log-fold change and p-value to identify differentially expressed genes. The p-values utilized in DESeq2 are adjusted by Benjamini-Hochberg procedure [235]. The difference between a p-value and an FDR adjusted p-value is as follows: a p-value of 0.05 signifies that 5%of tests will report into false positives and a FDR adjusted p-value of 0.05 signifies that 5% of significant tests will report into false positives. Additional open-source algorithm named GFOLD (generalized fold change) [236] for determining differentially expressed genes from RNA-seq data. The GFOLD values provides reliable log2Fold Change.

2.3.3 Post-transcriptional RNA processing

The primary transcript generally undergoes further processing after it is formed by RNA polymerase, which is described as post-transcriptional modification. Alternative splicing and RNA editing are among the important processes which take place during post-transcriptional RNA modification [190]. The primary transcript consists of three forms, namely tRNA, mRNA and rRNA that is processed into mature RNAs. Alternative splicing is the process in which exons within a pre-mRNA transcript are joined in several ways to code for multiple proteins. RNA-seq data is suitable in identifying splice junctions. Alternative splicing plays an essential role in the regulation of gene expression [237]. An alliance between splicing and RNA editing is shown by the formation of splice sites in pre-mRNA of adenosine deaminase that acts on double stranded RNA (ADAR). The Mixture of Isoforms (MISO) [225] model benefits from RNA-seq data for the computation of expression of alternatively spliced exons,

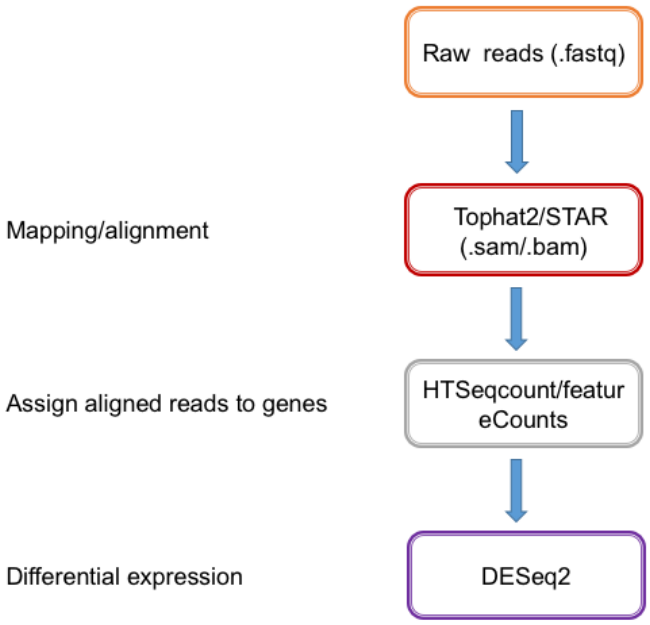


Figure 2.4: Schematic representation of RNA-seq analysis workflow.

respectively on exon or isoform level [225]. Apart from splicing, RNA editing is also a post-transcriptional process in which the information contained in the sequence of RNA is changed. Mostly two types of RNA editing enzymes are involved in the modification of RNA molecule, namely ADAR (adenosine deaminase acting on RNA) and APOBEC (apolipoprotein B mRNA editing enzyme) linked in Adenosine to inosine (A-to-I) and Cytidine to Uracil (C-to-U) respectively [238].

2.4 Ribosome-profiling data analysis

After the transcription of genomic information into RNA, it is translated into protein. Ribosomes play an important role in the translation of RNA into protein. Thus, they are also known as the protein warehouse of cells. Formerly, gene expression studies have been more focused at mRNA level. Hence none of the approaches focus at protein level. Ribosomal profiling combined with deep sequencing (Ribo-seq) is used to study in-vivo translation. It provides detailed information about protein synthesis. Ribosome profiling data provides information about protein synthesis related to the position of ribosomes on an mRNA. Whereas mRNA-seq and ribosome profiling share high resemblance with respect to data analysis.

2.4.1 Pre-processing and read mapping

The workflow for Ribo-seq data analysis includes several pre-processing steps, at first the removal of adapter sequences from the raw sequencing reads by applying cutadapt [239]. Then it involves the removal of ribosomal RNA (rRNA) sequences through fast alignment with Bowtie2 [202]. Ribosomal profiling data contains substantial amount of rRNA sequences. Elimination of rRNA sequences is considered to be necessary step to get efficient results from downstream analysis. Furthermore, unmapped reads obtained from above alignment are converted into a Fastq file with help of bedtools [240]. The attained fastq file is used to perform an alignment to the trancriptome either with Tophat2 [208] or by STAR [209].

2.4.2 Gene expression analysis

Gene expression analysis from ribosome profiling data gives insights about translational regulation at the protein level. As well as mRNA-seq data provides abundance estimation at the level of transcriptional control, ribosome profiling allows it at the level of translational control [241]. Thus, integration of mRNA-seq and ribosome profiling data will reveal gene regulation both at transcriptional and translational level Like in mRNA-seq. DESeq2 [232] is also applied for the estimation of differential expression of genes at protein level. The log2 fold-change is very effective in identifying differentially expressed genes between different conditions. Poisson distribution is the main model adopted by several algorithms for the calculation of log fold-change [242]. GFOLD [236] is a highly efficient method for recognizing up-regulation and down-regulation of genes by means of log2fold-change. Further gene enrichment analysis can be retrieved with the help of already available programs like DAVID [243] Gorilla [244] or PANTHER [245].

2.4.3 Open reading frames detection

Ribosome footprints are retrieved after the sequencing of ribosome fragments. Ribosome footprints provide information about translated open reading frames (ORFs). They define the explicit locations of start and stop codons whereas in mRNA-seq these locations are not explicitly shown. While the 5' UTR represents transcription start site (TSS), the 3'UTR represents transcription end site (TES). In ribosome profiling the area between start and stop codons consists of ORFs bound by ribosome footprints. The coverage of this area is higher in comparison to the space between start and stop codons [246](Figure 2.5).

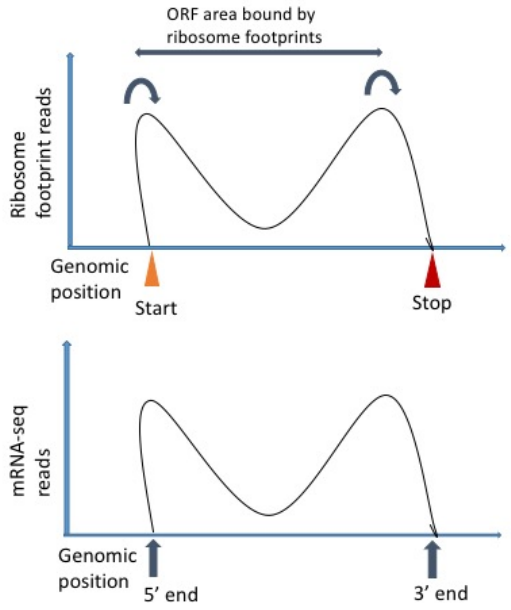


Figure 2.5: Schematic representation in comparison of Ribo-seq and mRNA-seq.Ribosome footprints shows definite arrangement between start and stop codon of a gene. Whereas mRNA-seq shows indefinite arrangement between 5' and 3' transcript boundaries. Figure drawn based on [246].

2.5 RNA-seq data preparation for senescent cells

RNA-seq data were obtained from [247] in Sequence read Archive (SRA) format.RNA-seq samples were prepared from Immortalized human primary BJ fibroblast cells (by human telomerase reverse transcriptase expression) which were cultured in Dulbecco's modified Eagle's medium. In it retroviruses were transfected obtained from Ecopack 2 cells. The samples have been processed from Normal proliferation by activation of tamoxifen-inducible RAS^{G12V} gene cultured in presence of 10-7 M 4OHT tamoxifen at early 5 days indicated as pre-senescent state and at full 14 days indicated as senescent state. Further all samples were sequenced by applying Illumina HiSeq-2000 platform. The libraries obtained from above samples were 150 bp for paired end reads and 50 bp for single end reads with different replicates (Table 2.1).

We used online available datasets for analysis of different methods such as RNA-seq, Ribo-seq data. The RNA-seq and Ribo-seq datasets for repetitive and RNA editing analysis were collected from GEO [GSE42509] under the mentioned publication [247].

Samples	Condition	Total Number of reads	Number of uniquely mapped reads
C1.rna	proliferation	26,181,799	23,718,792 (90.59%)
C2.rna	proliferation	28,243,869	25,598,250 (90.63%)
C3.rna	proliferation	49,899,303	43,422,948 (87.02%)
preS1.rna	pre-senescence	25,117,826	22,870,480 (91.05%)
preS2.rna	pre-senescence	21,619,863	19,631,893 (90.80%)
S.rna	senescence	43,395,063	36,890,213 (85.01%)

Table 2.1: Sample representation of RNA-seq reads obtained from immortalized human primary BJ fibroblast cells. C1,C2,C3:Proliferation preS1,preS2:pre-senescence S:senescence.

2.6 Analysis of RNA-seq datasets

We obtained RNA-seq samples of Immortalized human primary BJ fibroblast cells from [247]. The RNA-seq reads (single-end and paired-end) were mapped to the human (hg19) reference genome using STAR (v2.4.2a) [209]. We employed STAR due to its feature in discovering novel splice junctions and as non-canonical splice aligner in detecting chimeric transcripts and circular RNA. Mapped reads were subjected to obtain raw counts for estimation of gene expression by using featureCounts [248] from subread package (v1.4.6). To achieve this, we used gene transfer format (GTF) for human-specific genome (hg19). Further, counted reads were administered to normalization by adopting Transcript per million (TPM) method. We also applied GFOLD (v1.1.3) [236] for identifying differentially expressed genes. GFOLD is extremely useful in datasets without replicates.

2.7 Repetitive elements analysis

We performed mapping of RNA-seq reads with the human reference genome (hg19) by applying bowtie (v1.0.1) [202] short read aligner with a usage of –N 1 and –local parameter for effective alignment. FPKM values were computed by cufflinks (v2.0.2)

with default parameters [249] with use of GTF file format for specific repetitive elements (L1 & Alu) that were attained from UCSC genome browser [250]. This study mainly focuses on non-LTR retrotransposons such as the L1 and the Alu elements mentioned in (Section 3.1.3).

2.8 Data analysis workflow for detection of RNA editing sites

We designed data analysis which involves several strict filters and focuses on identification of true RNA editing sites with respect to L1 elements. We incorporated mapped RNA-seq reads to execute local realignment, base-score recalibaration and candidate-variant calling by using IndelRealigner, TableRecalibration and UnifiedGenotyper tool with parameter stand_call_conf to 0 and stand_emit_conf to 0 and output mode to EMIT_ALL_CONFIDENT_SITES from Genome Analysis toolkit (GATK)(v3.5-0) [251, 252]. In addition, we adapted few filtering steps of SNPiR [253] pipeline and obtained variants were subjected to it. The intent variants were incorporated to various filtering steps to obtain true variants by removal of false-positive variant calls. First, we filtered variants with quality up to 20. Then, the mismatches at 5' reads ends were removed in this step. Furthermore, the obtained variants were directed to filter in L1 elements (Figure 2.6) We used shell scripting and bedtools [240] to retrieve editing sites for each sample across full-length L1 elements. Motif pattern analysis for editing sites was obtained by using MEME [254] from MEME suite (v4.12.0) [224].

2.9 Identification of chimeric transcripts fused with L1 elements

We employed RNA-seq data reads to retrieve chimeric junctions. Mapping of RNA-seq reads was performed by using STAR (v2.4.0.1) [209], which also generates chimeric alignment to identify fusion transcripts. Chimeric junction file produced from STAR was used for identification of chimeric transcripts with the usage of chimSegmentMin 15 for minimum chimeric segment length and chimJunctionOverhangMin 15 for minimum overhang of a chimeric junction. Primarily, it was splited into two separate BED file for acceptor as well as donor sites. Further, splice site (donor and acceptor) for particular sample was intersected across full-length L1 elements by using shell scripting and bedtools [240]. Visualization of chimeric interactions for different conditions such as for proliferation and senescence were accomplished by using RCircos

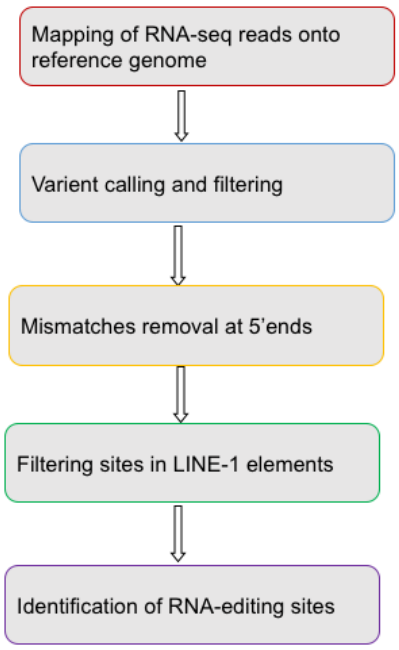


Figure 2.6: Flow-chart characterizing RNA-editing analysis.

(v1.2.0) package [255]. Motif representation for insertional event was determined by using Tomtom [256] from MEME suite (v4.12.0) [224].

2.10 Ribo-seq data preparation for senescent cells

Ribosome profiling (Ribo-seq) data were attained from [247] in Sequence read Archive (SRA) format. Ribo-seq samples were prepared from Immortalized human primary BJ fibroblast cells (by human telomerase reverse transcriptase expression) which were cultured in Dulbecco's modified Eagle's medium. In it retroviruses obtained from Ecopack 2 cells were transfected. Cells were mediated with different steps acquired from Ribo-seq protocol. The samples were processed for various conditions in which they were induced with tamoxifen-inducible by activation of RAS^{G12V} gene. Firstly, it was considered at early 5 days indicated as pre-senescent state and at full 14 days indicated as senescent state. Further all samples were sequenced by applying Illumina HiSeq-2000 platform. The libraries obtained from above samples were 35 bp for single end reads with various replicates [247].

2.11 Estimation of L1 encoded proteins

We downloaded the Repeat masker track particularly for L1 elements obtained from UCSC genome browser [250]. Then we extracted only Full length L1 elements from the Repeat masker track of L1 elements and prepared BED file for it. We obtained Ribosomal profiling (Ribo-seq) samples for Immortalized human primary BJ fibroblast cells from [247]. Firstly, adapter sequences were removed from raw fastq reads. Then obtained reads were subjected to remove ribosomal sequences with quick alignment by Tophat2 [208]. Later, unmapped reads from above alignment were converted into fastq file using bam2fastx [208]. Further, we constituted the consensus sequence for Full length L1 elements by converting BED into FASTA format with use of fastaFromBed [240]. Consensus sequence of full length L1 elements were aligned to Ribo-seq reads (single-end 50bp) with the help of bowtie (v1.0.1) [202] by using the -best parameter retrieved for different conditions varying from proliferation, pre-senescence and senescence. Normalized coverage for uniquely mapped file was obtained with help of bamCoverage [257] for each condition. Further, matrix computation was performed on the bigwig file of each condition by using computeMatrix from [257].

2.12 Statistical analysis

Wilcoxon-Mann-Whitney test was used for analysis at transcriptional and translational level. P-value$<$0.05 were considered statistically significant.

Chapter 3

Results

3.1 A computational web-based application for accelerated analysis of NGS data

3.1.1 General purpose

NGS has provided different approaches in the fields of genomics, transcriptomics and epigenomics which has resulted in solutions for complex biological problems. The applications covered in high-throughput sequencing technologies provides answers at different levels such as in DNase-seq, RNA-seq, ChIP-seq, ATAC-seq, Ribo-seq, FAIRE-seq, Methyl-seq, Hi-C, and many others [258]. The gradual decline in sequencing cost has led to a massive increase in the amount of sequencing data. To analyze sequencing data in a short period of time, it demands in carrying out appropriate solution. During the initial investigation of NGS datasets at transcriptional and translational level, I inquired insight about a framework which can exclusively relevant in-depth downstream analysis for NGS applications. Considering lack of particular framework, I developed novel web-schema that follows the central dogma process through which DNA information can be transformed into a functional product such as protein. It pursues two phases of gene expression including transcription and translation. The various applications of NGS technologies incorporated into our RanchNGS web interface are ChIP-seq [171], RNA-seq [185, 186] and Ribo-seq [195, 198] (Figure 3.1). ChIP-seq illustrates DNA-protein interactions on genome-wide scale. RNA sequencing utilizes the benefit of deep sequencing technologies to study transcriptome profiling. Transcript identification and gene expression portray two aspects of RNA sequencing. Ribosome profiling with the use of high-throughput sequencing helps to describe the ribosome footprints obtained after mRNA translation. RanchNGS is introduced as a web application used to examine NGS technologies. In

RanchNGS, RNA-seq and Ribo-seq section enables gene expression studies at transcriptional and translational level. The essential fundamentals in the development of web application include different aspects of data, algorithms, and asset of visualizations.

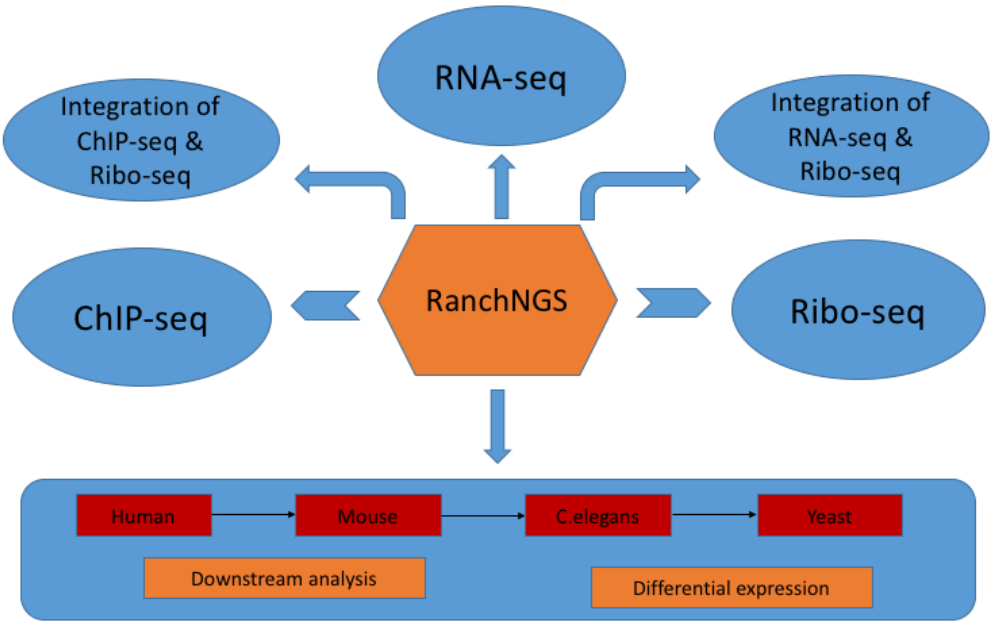


Figure 3.1: Schematic illustration of RanchNGS framework architecture.

3.1.2 Web interface framework

The RanchNGS essentially focuses on a downstream section of the NGS analysis. In case of transcriptomics methods, it adopts RNA-Seq, for epigenomics methods it adopts ChIP-seq and for ribosome profiling it adopts Ribo-seq. The starting section of RanchNGS consists of a ChIP-seq analysis followed by an RNA-seq analysis and then a Ribo-seq analysis. RanchNGS also comprises section for the integration of the ChIP-seq and RNA-seq and the integration of the RNA-seq and Ribo-seq (Figure 3.1). The benefit of integration for sequencing technique has proven to be advantageous in deeper comprehension of gene regulatory mechanisms [259].

The RanchNGS utilizes mapped reads in BAM format as input for the RNA-seq as well as for Ribo-seq analysis, whereas, for the ChIP-seq, ATAC-seq and DNA-seq analysis input are required in the BED format, which can be obtained after peak calling. In RanchNGS, data can be uploaded to a web interface for analysis in two different ways: by using an upload button for smaller files and utilizing a file transfer

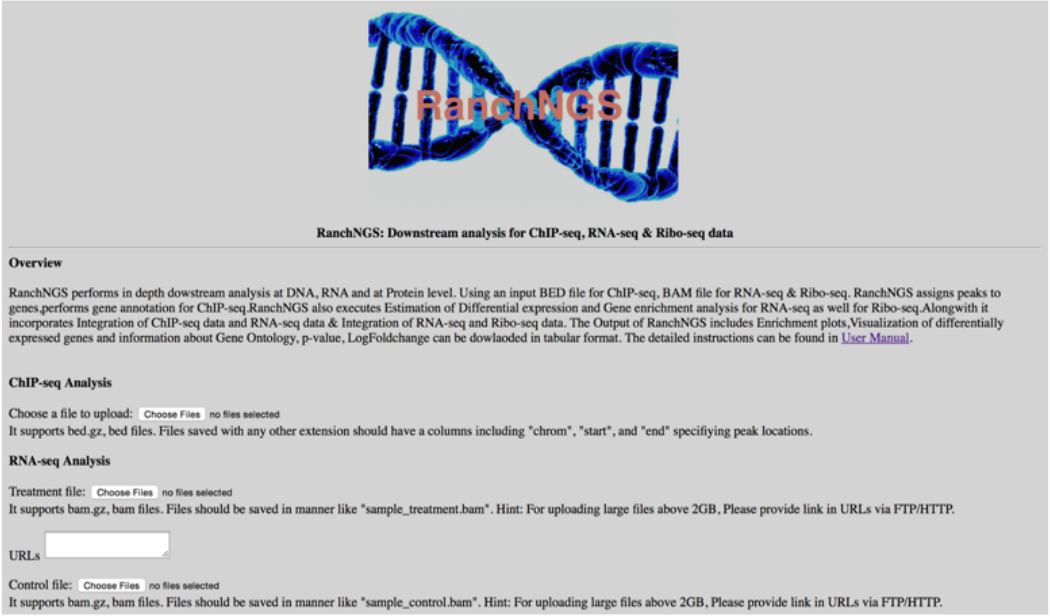


Figure 3.2: Sample snapshot of RanchNGS web-application.

protocol (FTP) or hypertext transfer protocol (HTTP) link for larger files (Figure 3.2).

In RanchNGS, ChIP-seq analysis constitutes TFTargetCaller [260] for peak annotation. The scoring method in the TFTargetCaller [260] consists of three steps, namely 1) peak-to-gene assignment 2) peak scoring and 3) integration of particular peak scores. The TFTargetCaller utilizes several methods or algorithms to identify the transcription factor (TF) targets. The primary characteristics of the closestGene method include: 1) assigning peaks to the closest gene 2) scoring peaks and 3) the summation of all peaks assigned to a given gene. The closestGene considers all peaks that are 1 Mb upstream and 1 Mb downstream of the Transcription start-site (TSS) [260]. Further annotated peaks are subjected to gene ontology analysis using bioMart [261] package and for gene enrichment, RanchNGS applies ReactomePA [262]. The ReactomePA provides pathway analysis based on Reactome pathway database. The bioMart package uses specific databases for distinct species in evaluating gene ontology. The ChIP-seq results consist of annotated peaks and lists of enriched genes that can be obtained in a tabular format. In RNA-seq, RanchNGS uses featureCounts to fulfill its requirement for counting of reads overlying with genomic features [248]. For the identification of differentially expressed genes, RanchNGS incorporates DESeq2 [232] (Section 2.3.2). Moreover, for providing gene annotation it incorporates bioMart [261] and for enriched pathway analysis, it applies ReactomePA [262]. The

RNA-seq results analyzed using RanchNGS contains annotation of differentially expressed genes. Ribosomal profiling involves counting reads with particular genomic features using featureCounts [248]. The differential gene expression for translational regulation is analyzed with DESeq2 [232]. The downstream analysis of ribosomal profiling data includes performing gene ontology using bioMart [261] and enrichment pathway analysis using ReactomePA [262] to assess the translation information.

Sequencing technique	Algorithm	Use	Version
ChIP-seq	TFTargetCaller	Peak-annotation	0.7
	BioMart	GO analysis	2.32.1
	ReactomePA	Pathway enrichment	1.20.2
RNA-seq	Featurecounts	Read-counting	1.5.2
	DeSeq2	Differential-expression	1.16.1
	BioMart	GO analysis	2.32.1
	ReactomePA	Pathway enrichment	1.20.2
Ribo-seq	Featurecounts	Read-counting	1.5.2
	DeSeq2	Differential-expression	1.16.1
	BioMart	GO analysis	2.32.1
	ReactomePA	Pathway enrichment	1.20.2

Table 3.1: Table representing algorithms implemented by RanchNGS.

3.1.3 Technical implementation

The RanchNGS is a standalone application for in-depth downstream analysis of DNA, RNA, and protein data. The development of the web application was assigned in two sections, in particular, front-end development and back-end development. The front-end implementation of the RanchNGS web interface involves the use of hypertext markup language (HTML), cascading style sheets (CSS) and javascript programming language. The back-end implementation of the web application involves the integration of various languages such as R and perl into the existing hypertext pre-processor (PHP) web programming language. The main function of the RanchNGS is to provide results in the form of PDF images, while detailed gene-related information can be obtained in tabular format, with the possibility of directly downloading it to a computer. The results and figures can be obtained accordingly without any further

user action. Distinct selective measures were taken into account to apply particular statistical R packages. The RanchNGS implements the relevant publicly available statistical R packages.

3.1.4 Example section

To inspect the efficient use of our system. I have exhibited the example segment of RanchNGS for each particular analysis with use of publicly available data. I have uploaded the applied data in the demo section of our web-interface. All the underneath figures in example section are acquired from RanchNGS.

3.1.4.1 Representation of ChIP-seq analysis

The ChIP-seq workflow involves the peak annotation using TFTargetCaller [260] which is exclusive to RanchNGS. The other task for data processing in ChIP-seq includes gene-ontology (GO) and gene-enrichment. The ChIP-seq experiment enables us to perceive broad knowledge about transcription-factors and histone-modifications. The applied demo data for ChIP-seq is from publicly available source. The results obtained from RanchNGS allows annotation of peaks to each particular gene, detail information about GO terms and enriched pathway analysis (Figure 3.3B) 3.3C). The demo datasets for ChIP-seq analysis have been collected from [263].

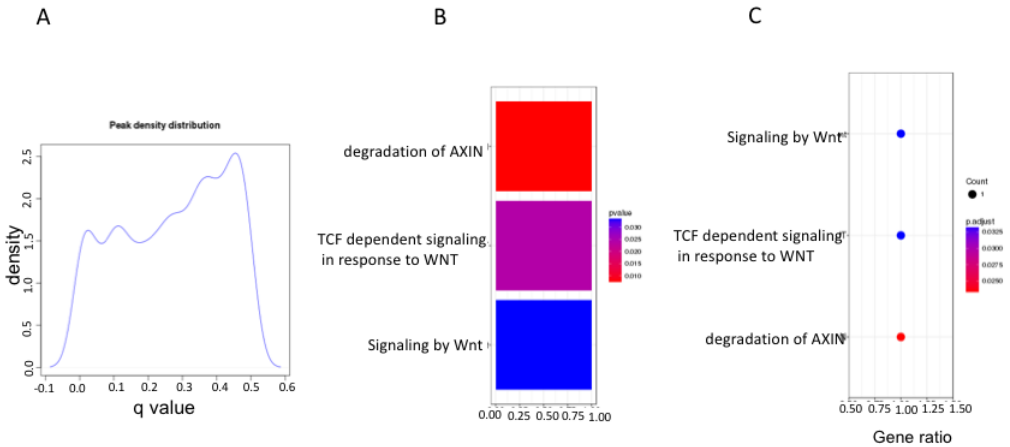


Figure 3.3: RanchNGS example segment for ChIP-seq analysis. (A) Plot interpreting density distribution for called peaks. (B) Barplot describing the enriched pathways by calculating p-value of it. (C) Dotplot representing the enriched pathways by estimation of adjusted p-value.

A

Ensembl gene ID	Entrezgene	Gene Ontology	Gene symbol
ENSMUSG00000000142	12006	Wnt signaling pathway	Axin2
ENSMUSG00000000567	20682	Signal transduction	Sox9
ENSMUSG00000000579	100040563	Nervous system development	Dynlt1c
ENSMUSG00000004018	67030	DNA-repair	Fancl

B

Ensembl gene ID	q value
ENSMUSG00000000142	0.106377896613191
ENSMUSG00000000567	0.00962686567164179
ENSMUSG00000000579	0.255610572687225
ENSMUSG00000004018	0.0467030716723549

Figure 3.4: RanchNGS example segment for ChIP-seq analysis. (A) Table obtained from RanchNGS for ChIP-seq analysis containing genesymbol, EnsemblID, EnterzgeneID and GO terms. (B) Table characterizing EnsemblID along with its calculated q-value for obtained peak.

3.1.4.2 Representation of RNA-seq analysis

The RNA-seq workflow includes the analysis of differential expression of genes for transcriptional control. RNA-seq analysis in RanchNGS benefits from identifying significantly differentially expressed genes. Genes with FC > 1 and adjusted p-value $< 5\%$ were defined as differentially expressed genes. The results in RNA-seq consists of plots and tables from downstream analysis (Figure 3.5)(Figure 3.6)(Figure 3.7)(Figure 3.8). The demo datasets for RNA-seq analysis have been collected from[247].

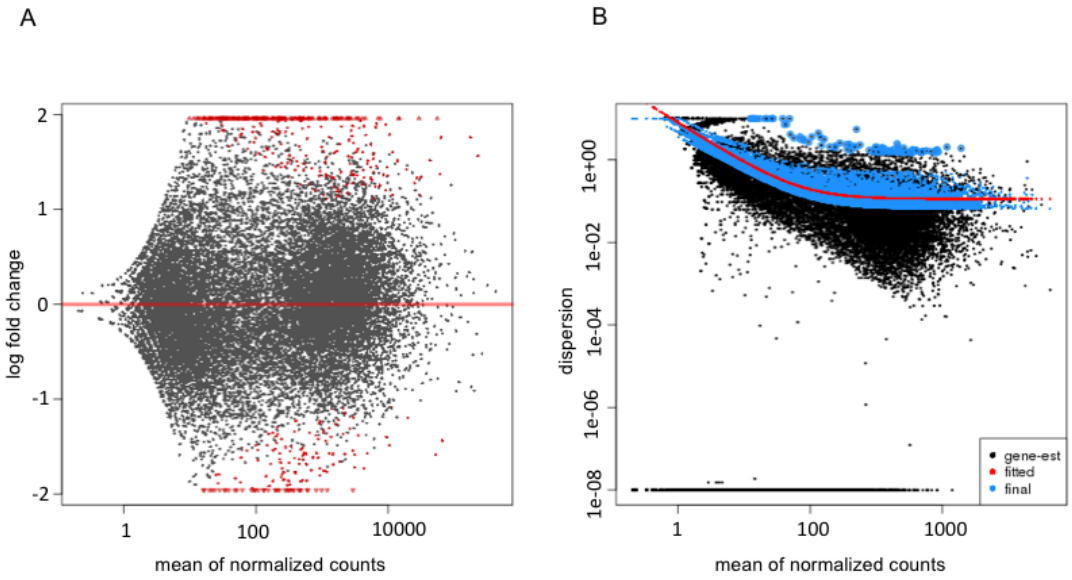


Figure 3.5: RanchNGS example segment for RNA-seq analysis (A) MA plot retrieved from RNA-seq datasets. In it, each dot equates towards particular gene. The x-axis represents average expression of normalized counts for all samples, y-axis represents log2 fold change for normalized counts between control and treatment. Genes defining adjusted p-value less than 0.1 are shown in red. (B) Dispersion plot obtained from RNA-seq datasets representing the dispersion estimates of each gene. In it, black points shows the dispersion estimates of each gene individually. The blue points exhibit true estimates of each gene. While blue points encircled are the dispersion outliers. Red line marks the distinction between two categories.

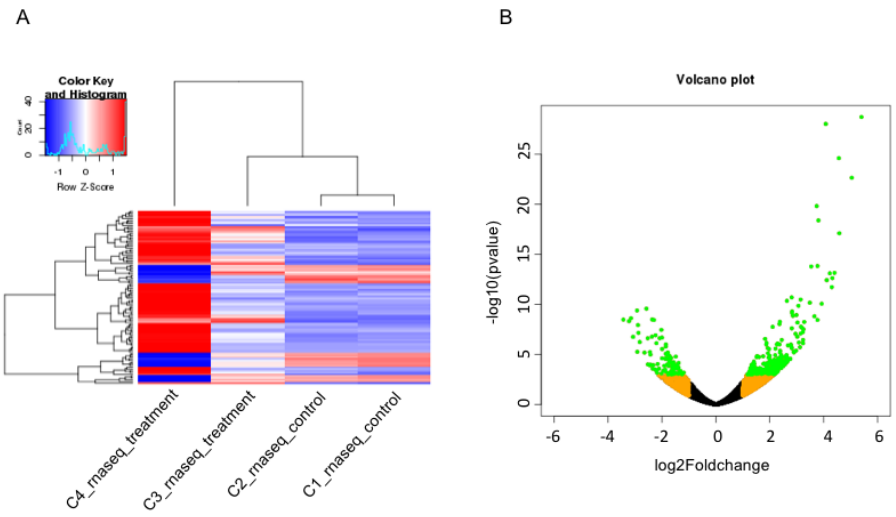


Figure 3.6: RanchNGS example segment for RNA-seq analysis (A) Heatmap obtained from demo datasets for RNA-seq experiment demonstrating top 100 variance genes for transcriptional regulation. Red color signifies the upregulation of genes and blue color signifies downregulation of genes. The dendogram on the left side shows the hierarchical clustering of the samples. (B) Volcano plot obtained from demo datasets in RNA-seq representing gene expression results. The x-axis exhibits log2 fold change and y-axis exhibits –log10 of p-value. In it orange color points represent log2FoldChange greater than 1, green points represent p adjusted value less than 0.05 and log2FoldChange greater than 1.

A

Ensembl gene ID	Entrezgene	Gene Ontology	Gene symbol
ENSG00000108821	1277	Wound healing	COL1A1
ENSG00000130513	9518	Regulation of apoptotic process	GDF15
ENSG00000126803	3306	cell differentiation	HSPA2
ENSG00000122641	3624	cell cycle arrest	INHBA

B

Ensembl gene ID	log2FoldChange	p value	p adjusted value
ENSG00000108821	-1.58499309207693	0.000414964847826351	0.0275815438429754
ENSG00000130513	2.04988825226732	0.000711194655303379	0.0410846174196231
ENSG00000126803	-1.42123958984843	0.000252924960595779	0.0191810713509089
ENSG00000122641	2.37597008911217	5.43747300394834e-09	2.75312052244031e-06

Figure 3.7: (A) Table acquired from RanchNGS for RNAseq analysis consisting of differential expression values including log2 Fold-change(log2FC), p-value and adjusted p-value. (B) Table attained from the outcome of GO determination along with gene-symbol and Enterzgene ID.

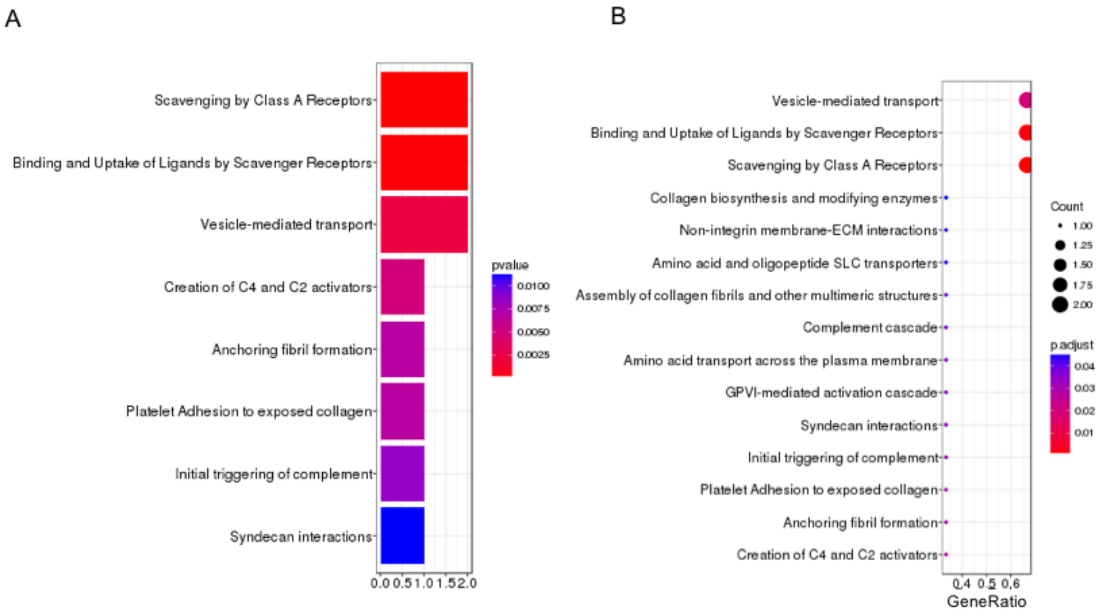


Figure 3.8: RanchNGS example section for RNA-seq analysis (A) Bar-plot obtained for enrichment results in RNA-seq datasets. Each colored bar represents enriched pathways on basis of p-value. (B) Dot plot obtained for enriched pathways from RNA-seq datasets. Each bar represents the enriched pathways on basis of adjusted p-value.

3.1.4.3 Representation of Ribo-seq analysis

Ribosomal-profiling fulfills the divergence between RNA sequencing and proteomics. The Ribo-seq workflow involves the analysis of deep sequencing of ribosomal profiling data in the measurement of expression for translational control. It assists in characterization of differentially translated genes. Quantitative analysis and visualization are quite effective in analysis of ribosome footprints (Figure 3.9) (Figure 3.10). In Volcano plot (Figure 3.9B), orange color points represent log2FC greater than 1, green points represent adjusted p-value less than 0.05 and log2FC greater than 1. The demo datasets for Ribo-seq analysis have been collected from [247].

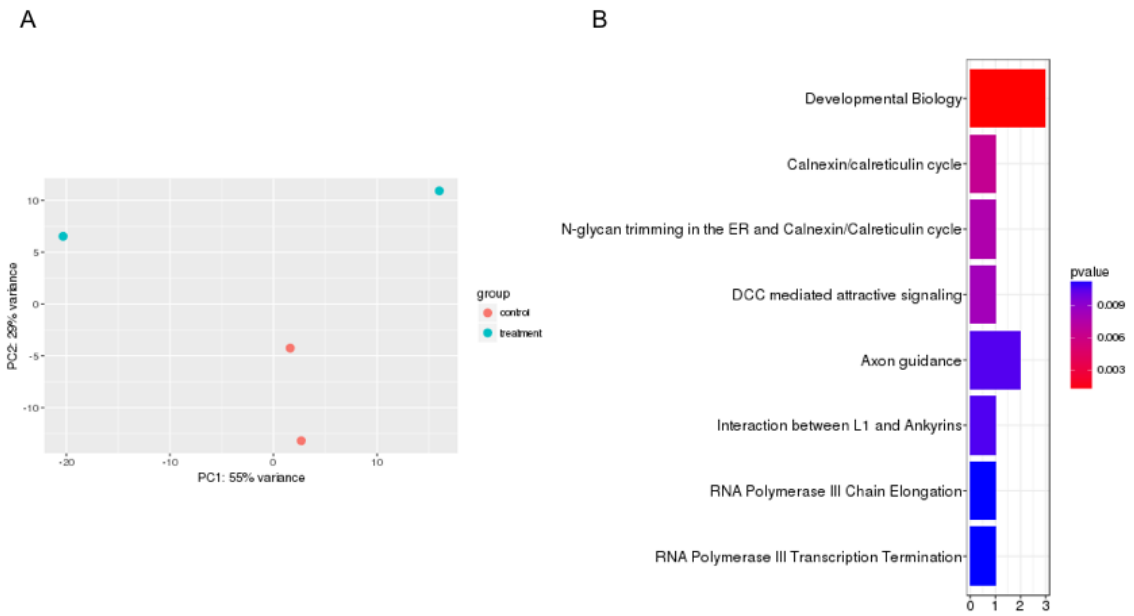


Figure 3.9: RanchNGS example section for Ribo-seq analysis. (A) Principal components analysis (PCA) plot inspecting the relationship between different conditions (control and treatment) in Ribo-seq datasets. (B) Barplot obtained for enriched pathways dissection in Ribo-seq datasets. Each colored bar represents pathways on basis of adjusted p-value.

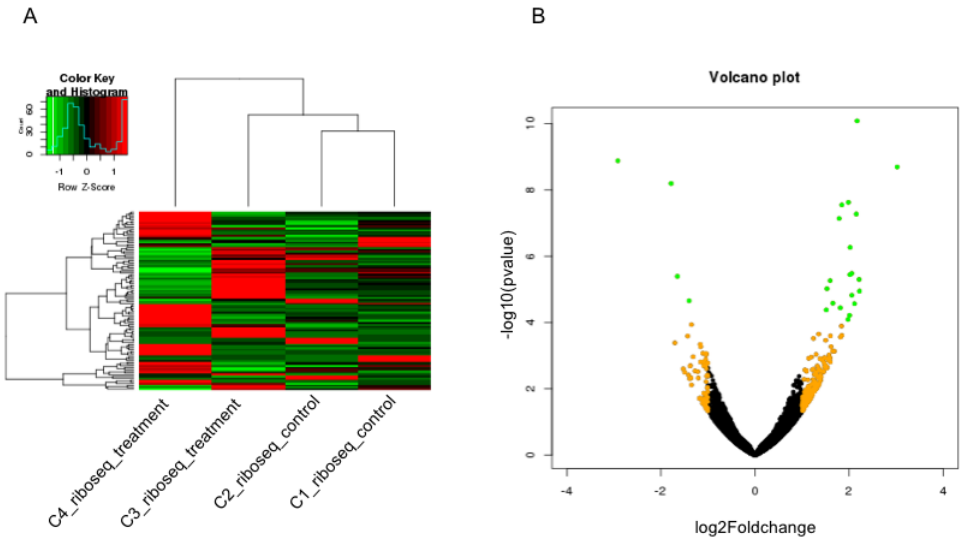


Figure 3.10: RanchNGS example section for Ribo-seq analysis. (A) Heatmap obtained from demo datasets in Ribo-seq experiment demonstrating top 100 variance genes for translational regulation. The dendogram on the left side shows the hierarchical clustering of the samples. (B) Volcano plot obtained from demo datasets in Ribo-seq samples. The x-axis exhibits log2 fold change and y-axis exhibits –log10 of p-value.

3.1.4.4 Illustration of integration for RNA-seq and Ribo-seq

The integration of RNA-seq and Ribo-seq can be useful in knowing insight about genes involved in transcriptional as well as translational control (Figure 3.11).

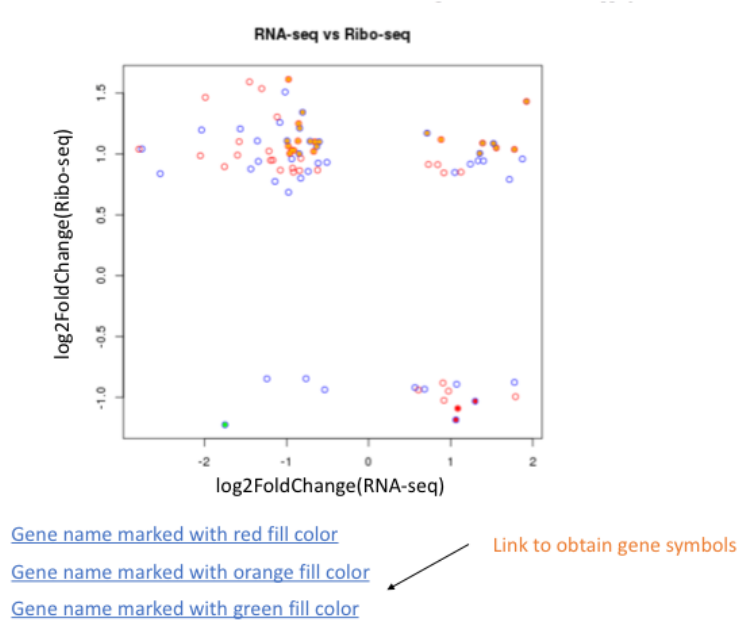


Figure 3.11: RanchNGS example section of integration for RNA-seq and Ribo-seq data. The x-axis depicts log2FC for RNA-seq data and y-axis depicts log2FC for Ribo-seq data. Plot representing transcriptional and translational association. Dot marked with red fill shows genes getting upregulated in RNA-seq are getting down-regulated in Ribo-seq. Dot marked with green fill represents genes getting down-regulated in RNA-seq as well as in Ribo-seq.

3.1.5 Characteristics of RanchNGS

The aim in designing RanchNGS web-tool was to make more user-friendly for the "biologists." RanchNGS is available as an open-source application. Moreover, RanchNGS requires from robust hardware resources to achieve its task. It aids in a wide range of NGS applications and contributes in automation of downstream analysis. RanchNGS carries out analysis in the background and results can be obtained directly on completion of respective analysis. In addition, RanchNGS is not confined to user's operating system (Windows, macOS, linux). I have also applied the above web-framework in an investigation for transcriptome analysis in the vicinity of L1 elements (Section 3.2).

3.1.6 Comparison of RanchNGS to other web-interfaces

The characteristics of RanchNGS differs from other available web-interfaces in respect to different criteria. The measure of computational analysis in RanchNGS is quite uncomplicated to available web-interfaces. Among other platforms, Chip-enrich [264] solely performs Chip-seq analysis whereas DEB [265] focuses particularly on RNA-seq analysis. Galaxy platform however, performs tasks closely related to RanchNGS application as it also executes different types of analysis in one platform. In identifying ribosome footprints, RiboProfiling (a Bioconductor package) focuses on the quantification at the transcript level whereas, RanchNGS focuses on quantification at the gene level. In contrast to other web-interfaces, our system supports reduced computational time in analyzing either small or larger input files in comparison to galaxy web-framework. Our system therefore enables users to obtain analysis results in a reduced period. A novel approach in form of web-schema, RanchNGS system provides the opportunity in performing integrative analysis of NGS techniques. An overview of attributes of RanchNGS in comparison to other resources are mentioned in (Table 3.2).

	RanchNGS	**Chip-enrich**	**DEB**	**Galaxy**
Applications	ChIP-seq, RNA-seq, Ribo-seq	Chip-seq	RNA-seq	Universal framework
Integration	Applied (ChIP-seq and RNA-seq)	None	None	None
Translation efficiency	Applied (RNA-seq and Ribo-seq)	None	None	None
Chromatin accessibility	ATAC-seq, DNase-seq	None	None	None
Purpose of web-framework	End user passage with analysis results	Start of workflows	Start of workflows	Start of workflows

Table 3.2: Comparison of RanchNGS features to available web-interface systems.

3.2 In silico dissection of L1 retrotransposons in cellular senescence

3.2.1 General purpose

L1 elements are regarded as a major player in evolution, capable of causing diseases to its host. In the human genome, p53 binding sites have been identified within L1 elements [36]. During the evolution of L1 elements, p53 have been shown to regulate L1 activity by preventing L1 insertion. Hence, p53 is also referred to as guardian of the genome [37]. In our study, I attempted to understand the consequences of p53 over-activation in immortalized cells transforming into senescent cells. To address this, in-house developed web-framework (Section 3.1) for data characterization was used during the initial evaluation. An additional frontier of defense in L1 elements has been regulated by TRIM28 (also known as KAP1) at the transcriptional level and mediated through APOBECs proteins at post-transcriptional level. Herein, I re-analyzed RNA-seq and Ribo-seq datasets to confer the cross-talk of L1 and host-defense machinery, while cells undergo senescent. APOBEC3B proteins were administered in RNA editing mechanism by converting C to U base [127]. Moreover, I inspected *in-silico* RNA editing pattern from APOBEC3B protein from immortalized cells undergoing senescence.

L1 elements are responsible for the modification in chromosomal rearrangement which could result in tumorigenesis [92]. If activity of L1 elements are not restrained by host then it would result in L1 retrotransposition. Although L1 retrotransposition activity can also be further associated with double-strand breaks (DSBs) during integration into genome [38, 199]. To investigate if L1 is undergoing retrotransposition the genome was examined for new insertional events, by means of chimeric transcripts. Furthermore, our hypothesis was based on to understand the role of L1 retrotransposons in a probable mechanism of immortalized cells undergoing senescence, since they are incapable of undergoing apoptosis (Figure **??**).

3.2.2 Data characterization of senescent cells with cells of its origin

RNA-seq reads were mapped to the human genome (hg19) using STAR [209] and further subjected to quantification and expression analysis [Table 2.1]. I used our in-house developed RanchNGS web-framework (Section 3.1) to obtain a diverse genome-wide response in senescent cells along with cells of its origin (Figure 3.12). Primarily,

I obtained a scatter-plot for log2foldchange and average of counts, which were normalized by size factor. The red dot in a scatter-plot represents FDR of less than 5 % (Figure 3.12A). The PCA plot was plotted to visualize the variation between different samples. It can be adopted as an approach in gathering quality control of the data (Figure 3.12B). In addition, I also acquired dispersion plot which provides the dispersion estimates for each gene. The genes represented by dots that are circled blue have higher gene-estimates‘. The red line in the plot depicts samples are moving toward final estimate (Figure 3.12C).

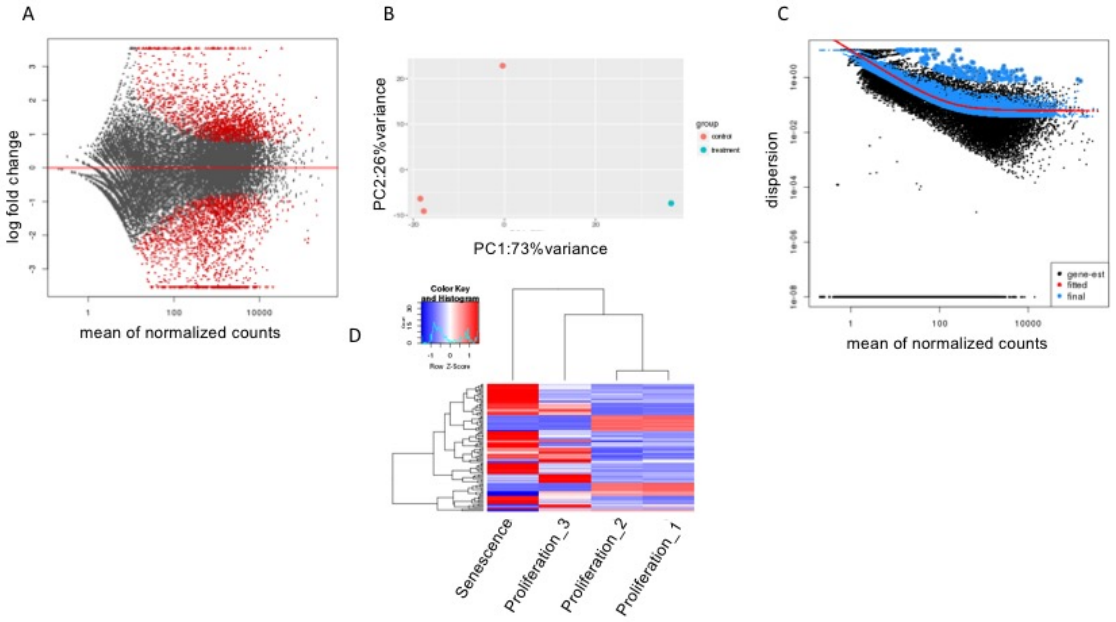


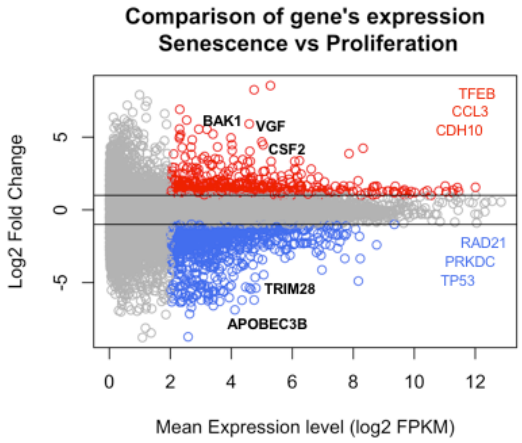
Figure 3.12: Data characterization of senescent cells with its cells of origin in immortalized human primary BJ fibroblast cells. (A) MA plot shows the dependency between the samples in terms of log fold change and mean of normalized counts. Dots displayed in red implies if the adjusted p-value is less than 0.05 (B) PCA plot displays the sample-to-sample distances in terms of variation between expression analysis samples (C) Dispersion plot represents the dispersion estimates for each gene. (D) Heatmap exhibits Z-score (Spearman and Pearson correlation used for columns and rows clustering) of top variance genes over different samples. In it, blue color illustrate down-regulation of genes. Whereas red color illustrate up-regulation of genes. The above images are obtained from our developed ”RanchNGS” web-interface.

3.2.3 Gene expression transition during cellular senescence

The expression levels of genes in senescence and proliferation were compared. Genes with Fold-change (FC) $>$ 1 and False-discovery rate (FDR) $<$ 5% were defined as

differentially expressed genes. Out of 24,000 genes, 2874 genes were differentially expressed. I observed that genes such as *BAK1* (BCL2 antagonist/killer 1), *VGF* (VGF nerve growth factor inducible), *CSF2* (Colony stimulating factor 2), *TFEB* (Transcription factor EB), *CCL3* (C-C motif chemokine ligand 3) and CDH10 (Cadherin 10) were upregulated in senescence while *TRIM28*, *APOBEC3B*, *RAD21* (RAD21 Cohesion complex component), *PRKDC* (Protein kinase DNA-activated catalytic polypeptide) and *TP53* (Tumor protein p53) were downregulated in senescence (Figure 3.13A). Following GO analysis, cell cycle arrest term with a significantly high p-value (p-value = 9.47e-59) in comparison to double-strand break repair followed by DNA damage response, p53 signal transduction (p-value = 3.08e-06) and double-strand break repair (p-value = 5.48e-14) were observed (Figure 3.13B) using differentially expressed gene-list analysis in PANTHER.

A



B

Categorization of 2874 differentially expressed genes

Gene Ontology	Observed/ Expected	p-value	FDR
Double strand break repair	3.55	5.48E-14	1.20E-11
DNA integrity checkpoint	3.61	8.45E-14	1.82E-11
DNA damage response, p53 signal transduction	3.5	3.08E-06	1.50E-04
DNA conformation change	3.86	1.12E-25	8.32E-23
cell cycle arrest	2.68	9.47E-59	1.47E-54

Figure 3.13: Overview of downstream analysis for differentially expressed genes. (A) Comparison of gene's expression profile for senescence vs prolilferation. Genes mentioned in blue color are getting downregulated whereas genes mentioned in red color are getting upregulated. (B) GO enrichment analysis of differentially expressed genes in between senescence and proliferation.

3.2.4 Down regulation of TRIM28 and APOBEC3B during cellular senescence

TRIM28 (TRIpartate motif-containing protein 28), which is also described as *KAP1*(KRAB-associated protein 1) is a co-factor and it recruits ZFPs. KRABs are known to exhibit various functions in Transcription and DNA repair [87, 266]. The *APOBEC3B* (apolipoprotein B mRNA editing enzyme 3) proteins are known to constrain replication of endogenous retroviruses [267] Previous studies have demonstrated the relationship between L1 expression and KRAB-KAP1 system [58]. The APOBECs family of proteins have been shown to inhibit L1 retrotransposition [268–271]. In this current study for immortalized human primary BJ fibroblast cells, the results enlighten for the first time about the expression of *TRIM28* was reduced in the pre-senescent and senescent cells (decrease of 2-fold) when compared to proliferative cells (Figure 3.14A). Moreover, the *APOBEC3B* was not expressed in senescent cells but only expressed in the proliferative cells (Figure 3.14B).

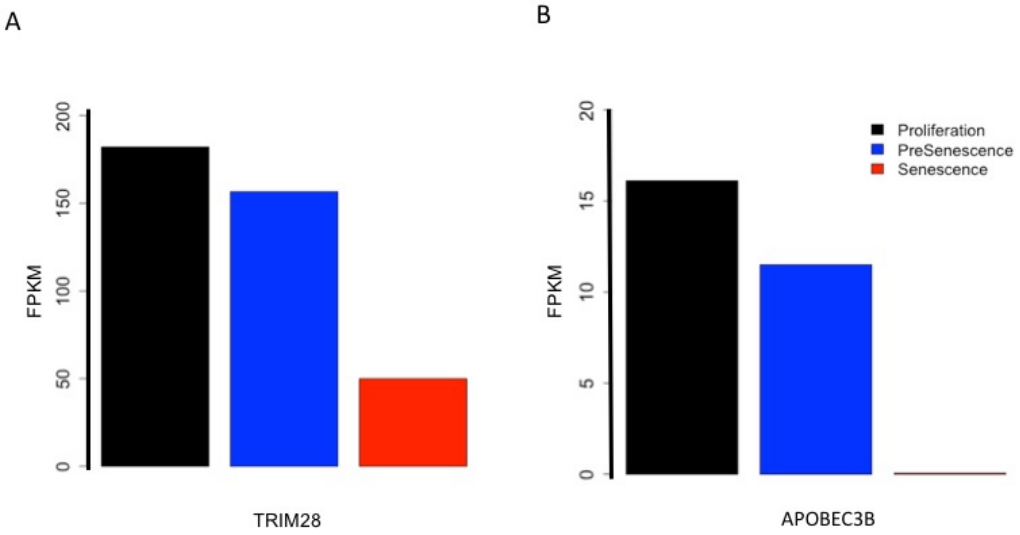


Figure 3.14: (A) Downregulation of TRIM28/KAP1 determined during transition from proliferative to senescent condition. (B) Downregulation of APOBEC3B measured during transition from proliferative to senescent condition.

3.2.5 Up regulation of retrotransposable elements during cellular senescence

To further investigate the role of L1 and Alu elements in the course of cells undergoing cellular senescence, RNA-seq data was used for analysis of retrotransposable elements. Mapped RNA-seq reads were obtained after mapping to the human reference genome (hg19) using bowtie (v0.9.6) with specific parameters. Moreover, FPKM values were estimated using Cufflinks (2.0.2). Expression analysis for conditions such as proliferation, pre-senescence and senescence was carried out from mapped RNA-seq reads for L1 and Alu elements. The expression profile of L1 elements associated with different cell conditions showed significantly upregulated expression levels for cells in senescence when compared to those proliferating (p-value = 2.2e-16) (Figure 3.15). However, some genes belonging to L1 subfamily, such as L1PA5 and L1PA7, were upregulated in proliferating cells (Figure A.2). To confirm the reactivation of TEs during cellular senescence, the same analysis was performed for Alu element subfamilies (mainly Alu J, Alu S, and Alu Y) forming part of SINEs class of retrotransposable elements. Expression of AluJ subfamily was significantly upregulated (p-value = 2.2e-16) during senescence cells in comparison to proliferating cells(Figure 3.16). Moreover, the gene expression of AluS and AluY subfamilies also showed significant upregulation in senescence cells (p-value = 2.2e-16). Alu and L1 retrotransposition are both involved in gene insertions [68]. Due to the dependence of Alu's transposition elements on L1 ORF2p, Alu elements are known to cause a number of diseases as compared to L1 by means of insertions [101].

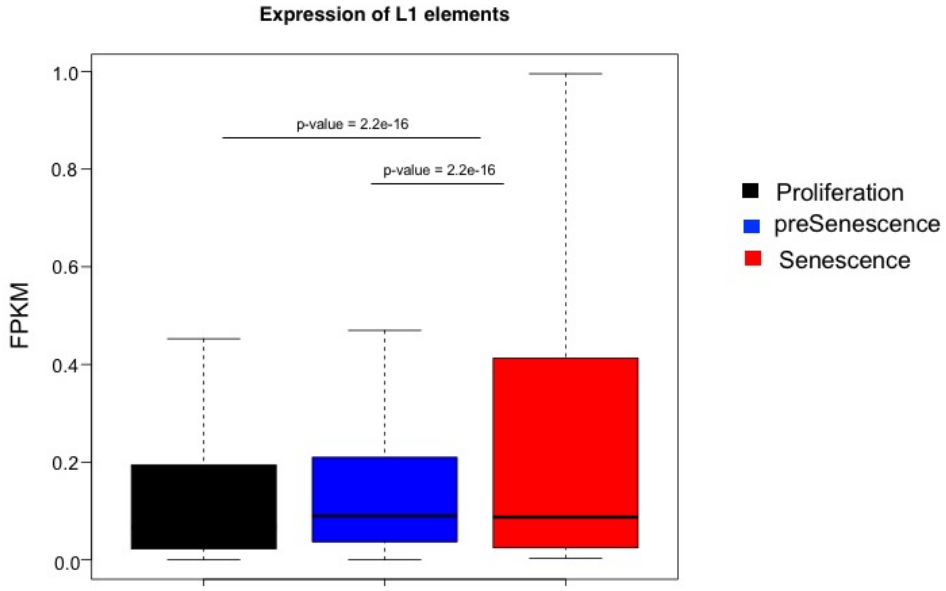


Figure 3.15: Expression analysis for L1 elements. Gene expression profile of L1 elements showed significant up-regulation in senescence condition as compare to other conditions (p-value = 2.2e-16). The statistical significance of expression difference was calculated using Wilcoxon-Mann-Whitney test

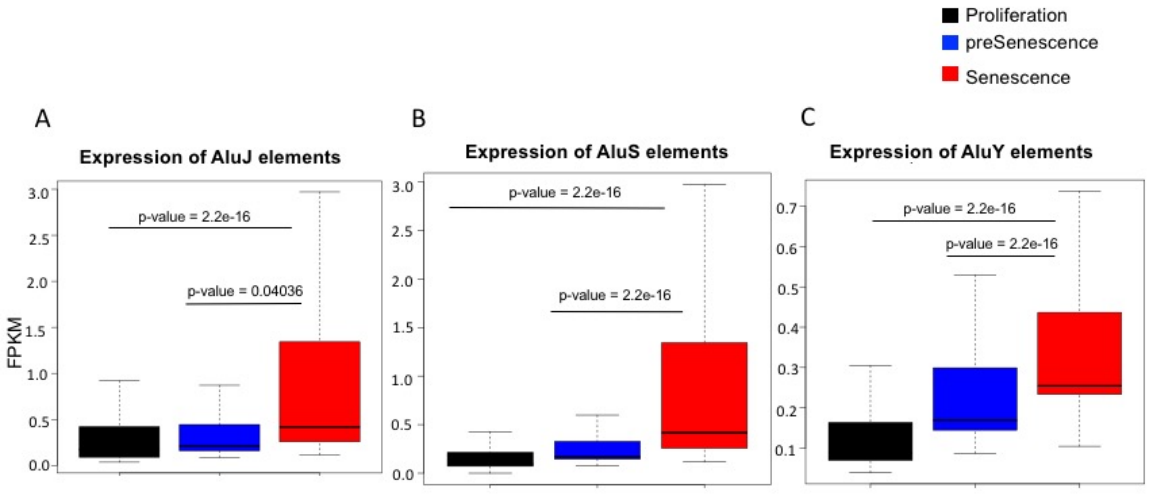


Figure 3.16: Expression analysis in different families of Alu elements. (A) In AluJ family of elements it exhibited significant up-regulation in senescent cells. (B) In AluS family of elements it showed significant upregulation of expression in senescent cells. (C) For AluY family of elements also showed significant upregulation in expression for senescent cells (p-value = 2.2e-16). The p-value was calculated using Wilcoxon-Mann-Whitney test

3.2.6 Confirmation of L1 encoded proteins at translational level

To further corroborate L1 activity during cellular senescence, ribosomal profiling (Ribo-seq) data obtained from immortalized human primary BJ fibroblast cells in different conditions was re-analyzed. L1 elements consist of two open reading frame proteins (ORF1p and ORF2p). ORF1p encodes RNA binding protein, which has a function in nucleic acid Chaperone activity. ORF2p consists of endonuclease and reverse transcriptase activities that are a required for L1 retrotransposition [272–274]. Before mapping, adapter and ribosomal RNA (rRNA) sequences were removed to ensure efficient downstream analysis. After mapping, normalized coverage by using bamCoverage [257] was attained. In addition, average normalized coverage in locus of L1 elements for each particular condition was plotted. Interestingly, significant translation of L1 open reading frames (ORFs) in ribo-seq repertoire in senescent cells was observed when compared to immortalized ones (p-value = 2.2e-16). (Figure 3.17).

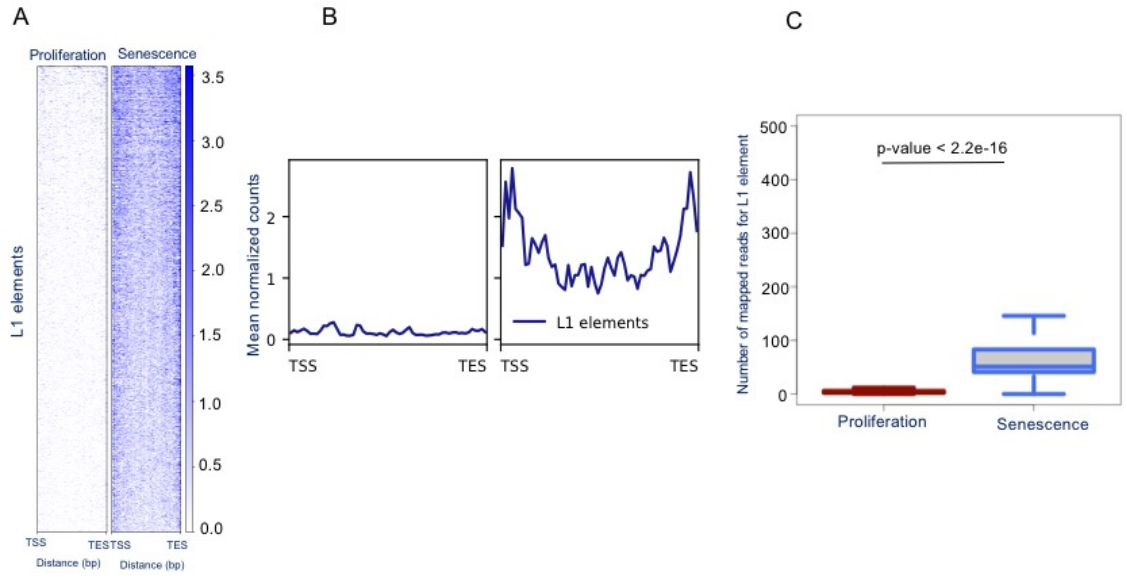


Figure 3.17: Comparative overview of L1 coding transcripts in ribo-seq repository. (A) Heatmap exhibiting the enrichment in senescence as compare to proliferation. (B) Coverage-plot demonstrating the enrichment in senescence as compare to proliferation. (C) Identification of significant level of mapped reads for L1-coding transcripts in ribo-seq repertoire (p-value = 2.2e-16). The statistical significance for L1 coding transcripts was calculated using Wilcoxon-Mann-Whitney test

3.2.7 RNA-editing role from APOBEC3B protein in regulation of L1 retrotransposition control

In contemplation of APOBECs group of proteins performing RNA editing function aimed at inhibiting L1 retrotransposition [270], profiling of RNA-seq data was used for detecting RNA editing sites. In RNA editing analysis, firstly the variant calling approach was applied for the identification of SNPs from RNA-seq data obtained from immortalized human primary fibroblast cells. Subsequently, mismatches at read ends were filtered to avoid false positives and sequencing error and this was followed by filtration based on full-length of L1 elements to determine RNA editing sites. Upon further analysis of RNA-seq data, 56020 editing sites in proliferative and 48325 editing sites in senescent cells were obtained before filtrating repetitive elements. 326 editing sites for proliferative and 316 editing sites for senescent were obtained following filtration of repetitive elements. In RNA editing analysis, we observed the differential pattern of RNA editing for *ADAR* as well as for *APOBEC3B* in proliferative and senescent cells in loci of full-length L1 elements (Figure 3.18). The RNA editing pattern for ADAR and APOBEC3B in loci L1 elements which are not capable of jumping was further investigated (Figure 3.18A and C). The distribution of differential RNA editing for the youngest L1 subfamilies (L1Hs, L1PA2, L1PA3) which are capable of retrotransposition in the cells undergoing senescence for *ADAR* (Figure 3.18B) as positive control against RNA editing pattern for *APOBEC3B* gene was observed. Interestingly, significant differential RNA editing for *APOBEC3B* gene in loci of youngest L1 lineages (p-value = 0.02403) for cells undergoing senescence was also observed. (Figure 3.18D). Furthermore, *in-silico* motif pattern which were executed as a way of validating RNA editing in L1 elements using MEME-suite was also established. The RNA editing analysis demonstrated an association of APOBECs or ADAR RNA editing in L1 elements as shown in (Figure 3.20).

During the past 40 miliion years (Myrs), L1 subfamilies have evolved in a prominent way along primate lineage initiating in humans lineage [84, 275]. The rise in the number of RNA editing sites in the senescent state for the middle-aged L1 subfamilies as compare to the youngest L1 subfamilies was also revealed (Figure 3.19). Moreover, a pattern in the number of RNA editing sites in proliferative state for L1 element subfamilies was also established (Figure A.3).

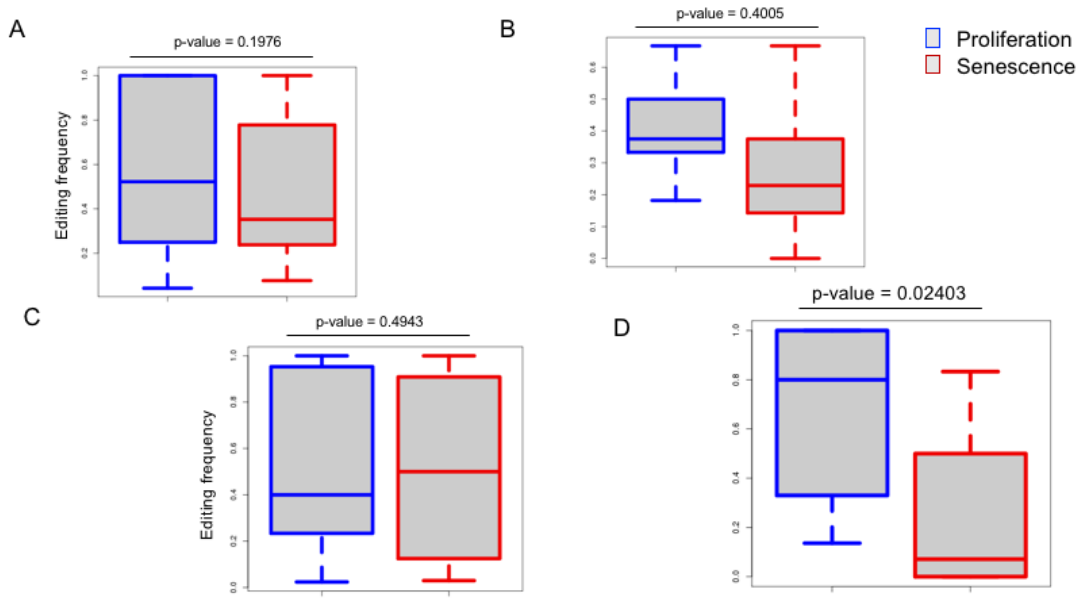


Figure 3.18: Differential RNA editing profiling of *ADAR* and *APOBEC3B* on loci of L1 families in the cells sustaining senescence. The blue-bordered color implies distribution of proliferative cells, whereas the red-bordered color implies distribution of senescent cells. (A) Differential RNA editing pattern for ADAR gene in loci of ancient and middle-aged L1 subfamilies. (B) Differential RNA editing for ADAR gene in loci of L1 elements capable of transposition. (C) Differential RNA editing for *APOBEC3B* gene in loci of ancient and middle-aged L1 subfamilies. (D) Differential RNA editing for *APOBEC3B* in loci of youngest L1 elements (L1Hs,L1PA2 and L1PA3) which are capable of transposition (p-value = 0.02403). The statistical significance in editing frequency was calculated using Wilcoxon-Mann-Whitney test.

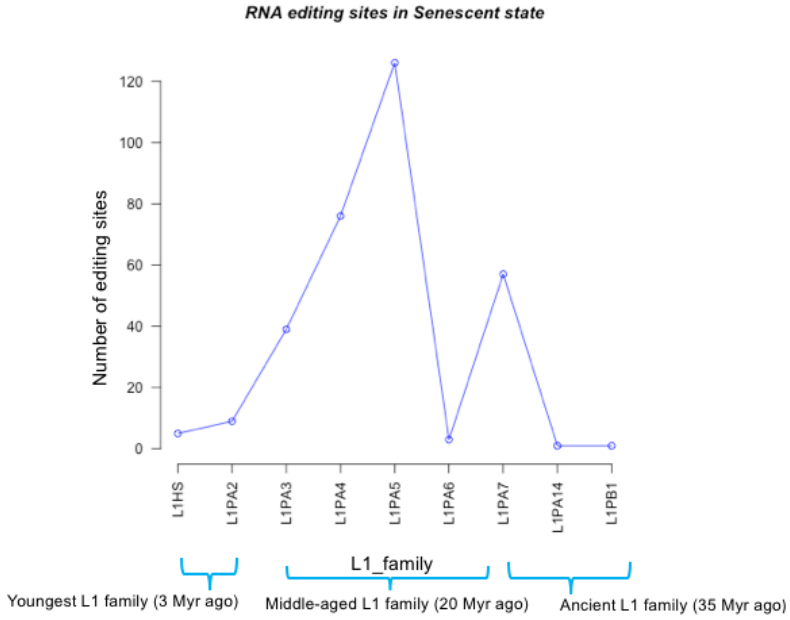


Figure 3.19: Distribution of RNA editing sites for entire L1 families. The first quarter represents youngest L1 subfamilies which are 3 Myr ago. The second quarter represents middle-aged L1 subfamilies which are 20 Myr ago. The third quarter represents ancient L1 subfamilies which are 35 Myr ago.

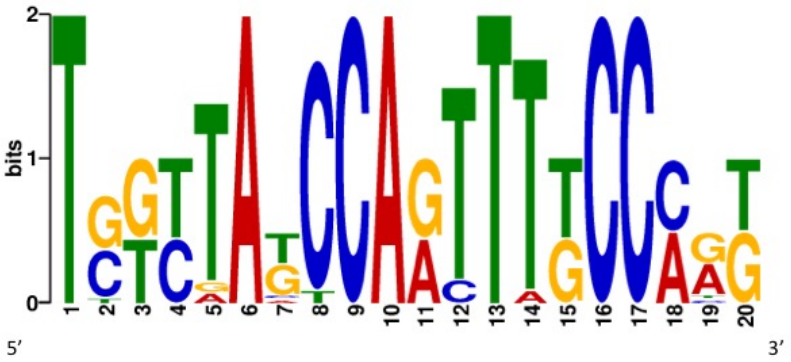


Figure 3.20: *De novo* motif analysis for RNA editing sites in loci of L1 elements. It specifies top hit retrieved for motif distribution.

3.2.8 Specification of insertional events using chimeric transcripts design

Considering that L1 is involved in retrotransposition, then the genome is likely to have new insertions. Herein, a recently developed novel approach to predict new insertional events was adopted. By using our in-house developed pipeline, a diverse pattern of chimeric transcript between the full-length L1 elements and genomic regions in senescent cells compared with proliferative ones was observed. Chimeric transcripts are known to form as a result of gene fusion [276]. In L1 elements chimeras are formed during L1 integration mechanism [277]. Circos plot representing the pattern for proliferative cells from acceptor sites showed 0 chimeric transcripts (Figure 3.21A). Moreover, the pattern for cells entering senescent state demonstrated 4 chimeric transcripts (Figure 3.21B). Noticeably, chimeric transcripts are enriched in senescent cells as compare to proliferative cells. The link lines in circos plot (Figure 3.21) for the fusion transcripts shows the relationship between two chromosomes for a specific condition in loci of L1 elements. The chimeric transcript analysis was also performed for proliferative and senescent cells from donor sites (Figure A.1). The circos plots for donor as well as for acceptor splice sites showed a different distribution pattern in immortalized human primary BJ fibroblast cell-type undergoing senescence (Figure 3.21)(Figure A.1). In our study, *de novo* a motif pattern from the above chimeric transcripts formed between junction of full-length L1 elements and specified particular condition from immortalized human primary BJ fibroblast cells was observed. L1 mediated insertion can cause disruption in target gene expression and function. During L1 retrotransposition, ORF2p enclosing endonuclease activity is liable for creating a nick on DNA at an insertion site. However, the motif pattern from our chimeric RNAs for acceptor sites were compared with consensus sequence specified for target site duplications (TSDs) acquired from The Cancer Genome Atlas [278]. Moreover, the consensus sequence of chimeric RNAs for proliferative and senescent cells, uniquely for donor splice sites were also compared with consensus sequence of TSDs (Figure A.4). Strikingly, in juxtaposition of motif comparison shows the presence of splice site with (p-value = 3.10e-02) and (E-value = 3.10e-02), thus predicting new insertional event in cells undergoing senescence (Figure 3.22).

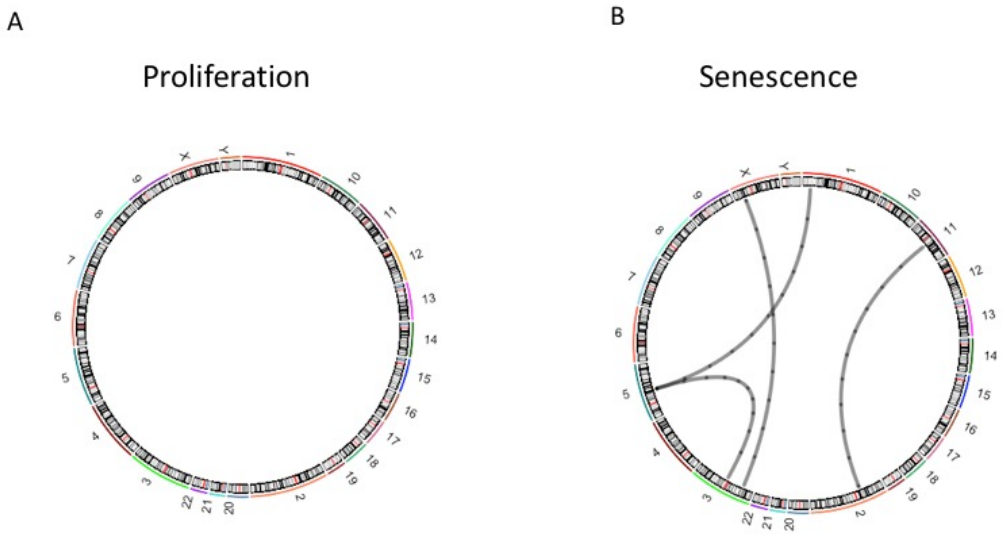


Figure 3.21: Chimeric RNA transcript representation in immortalized human primary BJ fibroblast cells for L1 elements capable of jumping (A) Circos plot illustration of chromosome interactions for proliferative cells exhibited 0 chimeric transcripts. (B) Analytical circos plot representation of chromosome interactions for senescent cells exhibited 4 chimeric transcripts.

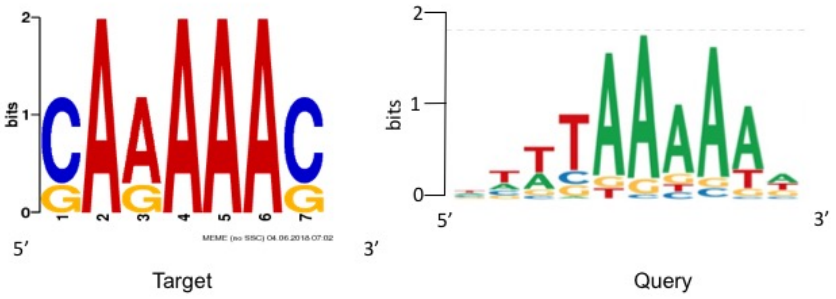


Figure 3.22: *De novo* motif analysis for insertional event of L1 elements from chimeric acceptor sites. Target represents consensus sequence of chimeric RNAs found in senescent cells. While query motif pattern displayed on left side represents consensus sequence of TSDs. The statistical significance of motif comparison was calculated using Tomtom (p-value = 3.10e-02).

Chapter 4

Discussion

4.1 Web-based framework for high-throughput sequencing data analytics

Downstream data analysis plays a crucial role in retrieving detailed information from any type of data. Therefore, in order to achieve this task, our novel web-interface linked to applied Next-generation sequencing (NGS) applications, titled "RanchNGS", for downstream and integrative analysis of ChIP-seq, RNA-seq and Ribo-seq data (Section 3.1), will be of distinct importance (Figure 3.1). Chromatin immunoprecipitation combined with extensive next generation sequencing (ChIP-seq) technology has been associated with genome-wide profiling of protein-DNA interactions of TFs and histone modifications. In cases of ChIP-seq data analysis, after mapping of sequenced reads onto a reference genome, enriched regions are determined using available peak detection methods. The developed web-framework is based on downstream analysis of ChIP-seq data, including peak annotation and gene enrichment. In particular, RanchNGS implements the TFTargetCaller [260] method for peak annotation and respective R packages like bioMart for gene ontology and ReactomePA for gene enrichment analysis. RNA-seq allows for an elaborative study of the transcriptome. In cases of RNA-seq data, the analysis also includes an estimation of differential gene expression and is effective in determining novel gene networks. In addition, for RNA-seq, RanchNGS enables downstream data analysis for the evaluation of ontology and analysis of enrichment of differentially expressed genes.

The translational process is an essential and coordinated step in gene expression. To attain translation, the ribosome profiling approach was proposed. It consists of deep sequencing of ribosome protected mRNA fragments. Ribosomal profiling has been previously used in identifying translated ORFs [246]. Therefore, an integra-

tion of RNA-seq and Ribo-seq data can be effective in investigating whether a transcribed gene will be translated or not. Publicly available datasets were used in our web-interface to inspect the outcome of Chip-seq, RNA-seq and Ribo-seq analysis. RanchNGS results are displayed as different enrichment plots showing visualization of differentially expressed genes and tabular information about p-values and adjusted p-values, FC, EntrezID, EnsemblID, and GO-terms for further investigation. We have demonstrated vital representation for every section with usage of demo datasets (Section 3.1.4). RanchNGS was constructed to carry out integration of different data types. RanchNGS allows for the analysis of diverse species, particularly *human*, *mouse*, *c.elegans* and *yeast*. RanchNGS uses different programming languages in its static framework for downstream analysis. During the development of the RanchNGS web-interface, HTML and CSS have been used to build a convenient web-based user interface. PHP along with the perl language has been integrated into the web application to allow the user to upload data and to enable interaction with the server for analysis. RanchNGS has the ability to upload the files in two forms; either by using the upload button or by providing the URL through the FTP/HTTP link. RanchNGS has been developed to also ensure the user's data security and data privacy. The main characteristics of RanchNGS in comparison to available NGS web applications is that it performs the downstream analysis for each respective section by focusing on the information directed from DNA to RNA and to form protein in a more direct way. RanchNGS is quite easy to access and use by biological researchers. It is available as open-source, along with its user manual.

NGS development has administered many fundamental aspects of genomics research. NGS technologies consist of a vast number of NGS applications which I have discussed in my introduction section of the thesis such as ChIP-seq, RNA-seq and Ribo-seq [279]. Before the onset of sequencing, chromatin immunoprecipitation in combination with microarrays techniques (ChIP-chip) [280] were widely used in determining protein-DNA interactions on a genome-wide scale range. Chromatin open assay techniques, namely DNase-seq, FAIRE-seq and genome-wide DNaseI footprinting, are currently considered as substitutes for chromatin immunoprecipitation [281]. Another method, named chromosome conformation capture (3C), determines areas of DNA carried forward by the nucleus, also known as looping [180]. In gene expression studies, RNA-seq serves to gain in-depth knowledge about whole transcriptomes. RNA-seq does not have any of the drawbacks of microarray technology. To remove technical bias, an appropriate normalization method must be applied prior to the

comparison of expression levels between different conditions [282]. The primary applications of RNA-seq include detection of splicing, gene-fusions and RNA editing [283]. RNA attached to ribosomes are represented as ribosome footprinting and sequencing performed on short fragments of RNA is referred to as ribosome profiling (Ribo-seq) [195, 198, 284].

4.2 Regulation of host-defence genes against L1 retrotransposons at the transcriptional and post-transcriptional level

I have re-analysed RNA-seq and Ribo-seq datasets from immortalized human primary BJ fibroblast cells to confirm the presence of cross-talk between the L1 elements and its host-defence machinery while the cells undergo senescence (Section 3.2). In brief, to obtain the transcriptome-wide gene expression response for our datasets, I used the in-house developed "RanchNGS" web-framework. During transcriptome-wide investigation in our datasets, genes like *BAK1* related to induce apoptosis, *CDH10* linked to colorectal cancer, *CSF2* related to AML, *CCL3* related to HIV-1 were upregulated during transition from proliferation to senescence (Figure 3.13A). Consequently, in GO analysis, terms like cell cycle arrest, DNA damage response and DNA double-strand break repair were retrieved with highly significant p-values (Figure 3.13B). Since TRIM28 and *APOBEC3B* play a crucial role in the regulation of L1 retrotransposons and in our datasets it showed downregulation in cells undergoing senescence (Figure 4.1A). *TRIM28* and *APOBEC3B* genes are essential constituent in regulation of host. In every L1 subfamily, *TRIM28* recruits in particular the KRAB-ZFP protein to repress the L1 element. Timely mutation over the course of the repression mechanism makes the L1 sequence inactive, which results in no requirement for KRAB-ZFP binding. L1 can then acquire necessary mutations to get it to be transcribed. To inhibit expression of the L1 element, the host forms another copy of KRAB-ZFP to repress the newly transcribed L1 elements. This arms race then continues during which *TRIM28* has been involved in repressing the L1 elements by recruiting the histone methlytransferase SETDB1 complex that controls trimethylation of histone H3 on lysine 9 on chromatin [87, 114]. *TRIM28* are also known to play an essential role in DNA repair. While in DDR, *TRIM28* gets phosphorylated upon ATM activation [266]. If L1 breaks the cage of the transcriptional repression machinery, then PIWI-interacting RNA (piRNAs) are able to degrade the L1 elements. Moreover, apart from DNA methylation, piRNAs are well-known to control L1 elements in germ cells

[285]. Recent findings have shown that piRNAs regulates L1Hs in early embryonic stem cells [58]. Moreover, post-transcriptionally, another line of defence mechanisms involved in the silencing of the L1 element is illustrated by the APOBECs group of RNA editing proteins that have demonstrated an interaction with the L1 ORF1 protein [271].

4.3 Activation of L1 elements during cellular senescence

Previous studies have shown that *p53* tends to activate or repress L1 transcription, due to the presence of *p53* DNA binding sites inside L1 elements [37] and *p53* has been proven to be involved in inhibition of L1 mobility [36]. Therefore,*p53* is also known as a guardian of the genome. The different types of stress that lead to cellular senescence include telomere erosion, oxidative stress, oncogene activation and DSBs [12, 14]. All other forms of stress, namely chemotherapy, IR, and activated oncogenes may share a common mediator of DNA damage in stimulating senescence [286, 287]. Upon oncogenic stress induction the RAS^{G12V} gene causes diversification during cellular senescence (Figure1.12).

TEs were analysed for the presence of a class of non-LTR retrotransposons to authenticate the transformation during senescence. In the case of L1 retrotransposons, significant upregulation of L1 transcription in the pre-senescent state that progressed until the senescent state was observed (Figure 3.15). Another class of non-LTR retrotransposons, namely Alu elements, are non-autonomous elements and are known to use L1 machinery to be mobilized (Figure 4.1B). L1 elements provide endonuclease and reverse transcriptase activities for the mobilization of Alu elements. I noticed a significant transcriptional pattern for the younger Alu subfamily through the pre-senescent state that progressed towards the senescent state. In combination, these findings indicate that L1 activation might be a consequence of the deactivation of host-defence genes, which control them. To further validate this hypothesis about the acquired transcriptional L1 activity during senescence, translational activity in the L1-encoded proteins was investigated (Figure 4.1C). Interestingly, I detected a significant number of L1-encoded transcripts in the Ribo-seq repository for immortalized cells undergoing senescence (Figure 3.17). L1 ORF1 encodes RNA binding protein and is considered to be a tumor-specific antigen. ORF1 has been linked to *p53* mutation in lung cancers [278, 288]. ORF2 encodes for both endonuclease and reverse transcriptase activities, and both are necessary for L1 retrotransposition [272,

289–291], and the endonuclease activity of L1 ORF2 is responsible for initiating DSBs and causing DNA damage response (DDR) stimulation [292, 293].

4.4 RNA editing regulation in L1 mediated machinery

In our study, at the post-transcriptional level, the *APOBEC3B* gene showed diminishing regulation over L1 elements during cellular senescence. Cytidine deaminase is known for performing RNA editing in the APOBECs protein family from cytidine (C) to uridine (U) [140, 294] in order to safeguard our host genome against external pathogens. Furthermore, I investigated the *in-silico* RNA editing pattern for ADAR as well as for APOBEC3B enzymes in loci of ancient and middle-aged L1 subfamilies, and I noticed a significant differential pattern of RNA editing in the youngest L1 subfamily for proliferative cells undergoing senescence in the loci of full-length L1 elements (Figure 4.1D). The youngest L1 lineages that are capable of retrotransposition are L1Hs, L1PA2, and L1PA3. RNA editing has been found to be a phenomenon necessary in organisms for regulation of gene expression [295]. RNA editing includes two types of single-base modifications like cytidine (C) to uridine (U) and adenosine (A) to inosine (I) deaminations [128]. Moreover, there are two types of RNA editing enzymes, namely cytidine deaminase and adenosine deaminase, which play a crucial role in recoding RNA sequence modifications [238]. Primarily, adenosine deaminases acting on RNA (ADAR) are associated with the conversion of adenosine (A) to inosines (I) in double-stranded RNAs [296]. In contrast, there are eight orthologues of the APOBECs family in humans [297]. The APOBEC3G gene has been observed to act in defence against retroviral activity by HIV-1 strains [132, 298] and APOBEC3B is established to be upregulated in around half of human cancers [299]. Interestingly, I observed *de novo* a motif pattern associated with a differential RNA editing event in the full-length L1 elements. This might indicate an affiliation to APOBECs or the ADAR gene. Certainly, the specified line of evidence including a significant pattern of L1-encoded transcripts and the observed differential RNA editing pattern in cells undergoing cellular senescence suggest that it might be a result of the deactivation of the *APOBEC3B* gene. The APOBEC3B gene component of the APOBECs protein family has displayed critical functions in inhibition of L1 retrotransposons during primate evolution [298].

4.5 Blueprint of insertional events in L1 elements by utilizing chimeric transcripts formation

L1 retrotransposons tend to move inside the genome through a "copy and paste" mechanism. In the course of this event, the L1 element gets shoved out of its original position and integrates into a new position [92]. At the time of integration, the L1 element can interrupt the function of host genes in numerous ways. For example, an L1 element inserted in the coding part of a gene can disrupt the protein coding function and if inserted in an upstream segment of a gene it can disturb the regulatory sequences [151]. The L1 retrotransposition plays an important role in causing DNA damage in the human genome, which has not been thoroughly studied. Herein, I adopted a novel approach to predict L1 insertional events in association with genomic regions within the process of L1 retrotransposition (Figure 4.1E). By using our in-house pipeline, I detected a chimeric RNA transcript formed between the full-length L1 element and genomic regions. In the L1 retrotransposition machinery, the L1 RNA are transcribed and exported to the cytoplasm. In the cytoplasm, they get translated into two ORFs proteins and contribute to the formation of ribonucleoprotein (RNP) complexes. ORF1p encodes for RNA binding protein and ORF2p encodes for endonuclease as well for reverse transcriptase activities. The endonuclease (EN) activity of ORF2 creates gaps at L1 insertion sites generating DSBs [292, 300]. The reverse transcriptase activity (RT) is also responsible for establishing the L1 retrotransposition [301]. During the process of target primed-reverse transcription, the endonuclease function of ORF2p cleaves the 3'OH of the insertion site for priming and then the RT function extends the strand to be used as a template. Rectification of the double-stranded break requires the host DNA repair proteins. Failure of DNA repair leads to insertional mutagenesis [102, 278]. In somatic cells, retrotranspositions in Alu elements occur very few times [278, 302, 303]. In the course of germline projection, retrotransposition in Alu elements is higher than in L1 [291]. In L1, overexpression of the ORF1 protein has been considered to be an indication of human cancers [291]. L1 retrotransposition contribute about $\tilde{0}.3\%$ of human mutations [40]. L1 retrotransposition has been known to cause disease mutations in humans through insertional mutagenesis [108]. There are numerous consequences of L1 mediated insertions, which include disturbances in target gene expression and function [79]. It total, 124 insertions have been identified as causing genetic disease from germline-mediated L1 retrotransposition [304].However, L1 insertions in somatic cells have also been detected in colon cancer [91]. L1s retrotransposons can be a source of insertions and

deletions during the retrotransposition process [92]. In our investigation, the ratio of chimeric transcripts for cells undergoing senescence were higher in comparison to proliferative ones (Figure A.4). Uniquely, motifs retrieved from *de novo* junction of L1 and genomic sequences were enriched in comparison to existing TSDs with a significant p-value, suggesting a new insertional event in cells undergoing senescence. Since L1 activity can contribute to DSBs during integration into the genome, these novel insights provide consequences of oncogenic stress in proliferative cells that might be related to L1 retrotransposition, thereby triggering DSBs and enabling cells to enter senescence.

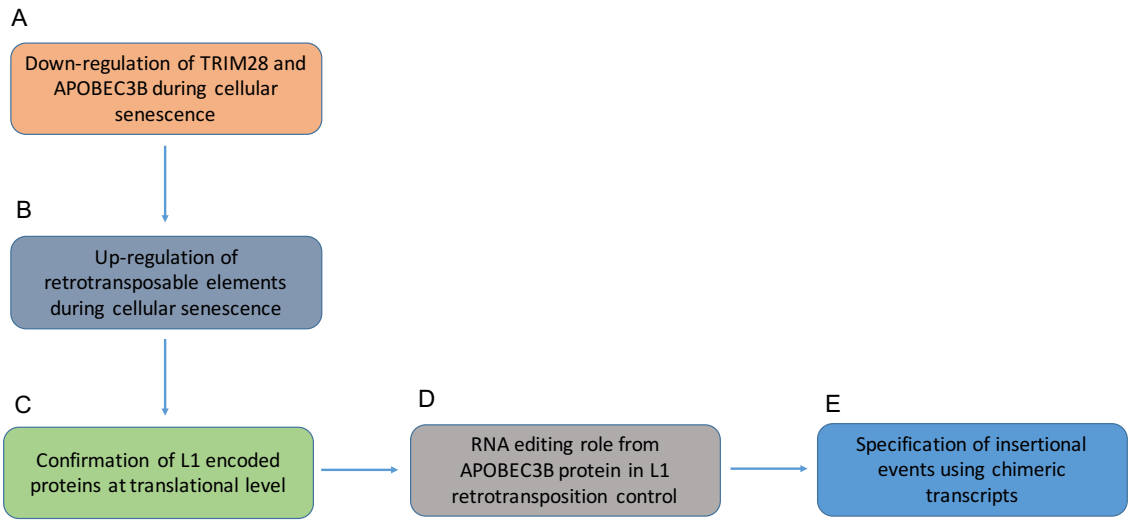


Figure 4.1: A schematic representation of working development model.

4.6 Outlook

In this thesis, NGS technologies have been widely studied, and these include ChIP-seq, RNA-seq, and Ribo-seq. DNA methylation is considered to be one of the mainstream methods of epigenetic modification, and others include transcription, chromatin structure, embryonic development, and genome imprinting [305]. Gene expression is studied as one of the essential mechanisms in transcriptomics. The RNA-seq method is used to measure the expression of genes, while the Ribo-seq method is used to determine the translational abundance of gene products. RNA-seq is not only confined to measuring gene expression but is also applicable in mutation analysis, fusion tran-

script and alternative splicing detection [306]. Alternative splicing has been linked to several inherited diseases [307].

TEs activity exhibits an interconnection between cancer and aging. In the area of cancer research, L1 elements insertions have proven to be of high significance. The first somatic L1 insertion was detected in the APC gene. In association with aging, L1 activity has demonstrated a diverse role. It has been reported that the extent of L1 retrotransposition in cancer increases vastly with age. In another study corroborating these findings, in humans, TEs epigenetic silencing was shown to decrease with age [308]. A comprehensive association between TEs and aging is a wide and promising area of research.

4.7 Future work

Due to a huge reduction in the cost of genome sequencing, the number of NGS applications has increased. NGS techniques have been used in all types of species and are beneficial to any component of the genome. NGS techniques have helped in bridging the gap between the research and diagnostics field [309]. Sequencing technologies help pathologists in providing personalized medicine [310]. Moreover, diverse NGS applications have been executed in clinical diagnostics [311, 312] and NGS has been considered an effective tool in discovering causes and solutions for rare diseases. A huge shift has been witnessed in transcriptomics sequencing from "bulk" RNA-seq to "single"-cell RNA-seq (scRNA-seq). scRNA-seq has the potential to disclose complete new cell types. The drawback of using bulk RNA-seq in analysing heterogeneous systems such as early embryonic development and brain system can be prevented by using scRNA-seq. Cloud computing services have been adopted by data researchers to provide benefits in medical innovations. However, the drawbacks from second generation technologies prevail during the development into third generation technology. The third generation sequencing technology such as the PacBio RS system (Pacific biosciences) and nanopore sequencing technology provide longer read lengths of about 5500–8500 bp [313]. Long reads have exhibited better characterization in studying the interaction between structural nuclear variants (SNV) and copy number variant (CNV) in virus populations. Several NGS applications have generated enormous amount of data, thus posing challenges in the storage of such big data sets. Collectively, the above mentioned aspects constitute the extensive requirement for computational and data storage infrastructure [314]. Hence, with the

advanced growth in NGS technologies, it is necessary to address major challenges of data storage, security, data analysis and integration.

In the past era, TEs have provided beneficial outcomes for the scientific community in terms of health and disease. It has been proven TEs are more than just "junk" elements. In relation to cross-species analysis, it has shown the evolutionary modifications between the human and non-human primate genome. In addition to the extensive development of the sequencing technology, longer read sequencing will be highly advantageous in examining repetitive regions. In a recently published research article, by using whole-genome sequencing with the support from the PacBio sequencer, they were able to detect candidates for somatic L1 insertions in human colon cancer [91]. Longer reads are quite essential to characterize the interactions between SNV and CNV in emerging viral populations. Furthermore, there are numerous TEs families that are currently active and capable of transposition inside human genomes such as HERV-K, L1, Alu, and SVA elements. Previous studies have reported a vast number of diseases linked to active non-LTR retrotransposons such as haemophilia, leukaemia, breast and colon cancer [39]. A Contempo research published recently showed the involvement of a distinct set of genes in L1 regulation by using CRISPR-Cas9, an approach for genome-wide DNA editing [315]. In future, an investigation is required to retrieve L1-mediated retrotransposition events in germline and somatic cells [54].

Furthermore, in future, it would be intriguing to study the role of TEs in cellular senescence at the scRNA-seq level using longer sequence reads from PacBio.

Summary

The advancement in the field of high-throughput sequencing technologies has fundamentally transformed biological research. Due to reduction in the cost of sequencing, the number of sequenced genomes from human and diverse other species has been raised considerably. The establishment of sequencing technology has constituted vast amount of manifold NGS applications. During the exploration of NGS datasets at transcriptional and translational level, we pursued enlightenment about a framework which can entirely pertinent to NGS applications. Therefore, we aim to create a software framework, which is entirely pertinent for analysis and integration of data from several NGS applications. These include characterization of transcriptomes by RNA-seq, determination of ribosome profiling by Ribo-seq and genome-wide identification of protein-DNA interactions by ChIP-seq. Hence, I developed a web application (RanchNGS) for downstream and integrative analysis of the corresponding data. Besides particular individual analysis, RanchNGS also attains to illuminate biological coherances by means of data integration. Currently RanchNGS can be used to analyse data from *human*, *mouse*, *c.elegans* and *yeast*. I have utilized publicly available tools in my web interface. The biggest challenges of NGS applications are data storage and analysis. RanchNGS is highly efficient and particularly beneficial for the life-science community, because it can be used without any software-coding skills. In this thesis I have focused on the development of computational programs for detailed analysis of data obtained from RNA-seq and Ribo-seq experiments. I have also demonstrated the performance of the entire established RanchNGS web-framework at length using datasets from experiments about cellular aging (senescence).

Aberrant expression of oncogenes can induce cellular responses like senescence or apoptosis. Oncogenic stress has been known to play a vital role in the induction of DNA damage response by restraining a mechanism of oncogene-induced senescence (OIS). Herein, I sought to understand the result of excessive oncogenic stress by targeting p53 in immortalized cells. p53 has been known to safeguard the activity of L1 elements in the genome by preventing L1 insertions. L1 activity is known to

be constrained transcriptionally by *TRIM28*. While in case of evasion, APOBECs group of enzymes regulate L1 activity by RNA editing machinery. In novel finding, downregulation of *TRIM28* and *APOBEC3B* during senescent state was observed indicating depletion of regulation in L1 elements. Interestingly, L1 elements were up-regulated in cells undergoing senescence. Possibly, reactivation of L1 elements might be a result of deactivation of host-defense genes (*TRIM28* and *APOBEC3B*) which control regulation of L1 elements. To eliminate the possible illusion of false alarm at transcriptional level, I inspected the translational pattern of L1 ORFs in Ribo-seq domain. In relation to post-transcriptional analysis, I detected a significant pattern of L1 coding transcripts in the cells undergoing senescence. Probably, this observation might be explained due to the deactivation of RNA editing enzymes. Furthermore, I inspected in silico RNA editing events for ADAR as well as APOBEC3B enzymes in ancient and middle-aged L1 subfamilies. A significant differential RNA editing pattern was observed through the action of APOBEC3B enzyme between proliferative and senescent cells in loci of youngest L1 elements which are capable of jumping. If the nick in DNA is above a certain threshold, L1 elements generate insertions and constitute double-strand breaks (DSBs) during integration into the genome. I detected the insertional event by means of chimeric transcripts between L1 elements and genomic regions. Curiously, the motif for target site duplications (TSDs) was enriched at the de novo junction of L1 elements and genomic sequence indicated new insertions in senescent cells. Interestingly, our results suggest that excessive oncogenic stress in immortalized cells might be linked to L1 jumping, thereby inducing DSBs and enabling cells to enter into cellular senescence.

Zusammenfassung

Die Fortschritte im Bereich der Hochdurchsatz-Sequenzierverfahren haben die biologische Forschung grundlegend gewandelt. Infolge der Reduzierung der Sequenzierungskosten hat sich die Zahl der sequenzierten Genome der humanen und verschiedener anderer Spezies deutlich erhöht. Die Etablierung der Sequenziertechnologie hat eine enorme Anzahl vielfältiger NGS-Anwendungen hervorgebracht. Bei der Analyse von NGS-Datensätzen auf transkriptioneller und translationaler Ebene suchten wir nach einem universellen Rahmen, der für viele verschiedene NGS-Anwendungen genutzt werden kann. Wir wollten ein Software-Framework erstellen, das für die Auswertung und Integration von Daten aus verschiedenen NGS-Anwendungen von großem Nutzen ist. Dazu zählen die Charakterisierung von Transkriptomen mit RNA-Seq, die Bestimmung von Ribosomen-Profilen mit Ribo-Seq und die genomweite Identifikation von Protein-DNA-Interaktionen mit ChIP-Seq. Demzufolge habe ich eine Web-Applikation (RanchNGS) für die Verarbeitung und integrative Analyse entsprechender Daten entwickelt. Abgesehen von spezifischen einzelnen Analysen zielt RanchNGS auch darauf ab, biologische Zusammenhänge durch Integration von Daten zu erhellen. Derzeit kann RanchNGS zur Analyse von Daten aus *Mensch*, *Maus*, *C.elegans* und *Hefe* verwendet werden. Ich habe in meinem Web-Interface öffentlich verfügbare Programme verwendet. Die größten Herausforderungen der NGS-Anwendungen sind die Daten-Speicherung und –Analyse. RanchNGS arbeitet sehr effizient und ist besonders nützlich für die Life-Science-Gemeinde, da es auch ohne Programmierfähigkeiten genutzt werden kann. In dieser Doktorarbeit habe ich mich auf die Entwicklung von Rechenprogrammen für die detaillierte Analyse von Daten konzentriert, die aus RNA-Seq- und Ribo-Seq-Experimenten gewonnen wurden. Die Leistungsfähigkeit des umfangreichen RanchNGS-Web-Frameworks habe ich ausführlich anhand von Datensätzen aus Experimenten zur zellulären Alterung (Seneszenz) demonstriert.

Die anomale Expression von Onkogenen kann zelluläre Reaktionen wie Seneszenz oder Apoptose hervorrufen. Es ist bekannt, dass onkogener Stress eine entschei-

dende Rolle bei der Induktion von Reaktionen auf Schäden der DNA durch Unterdrückung von onkogen-induzierter Seneszenz (OIS) spielt. Hier habe ich versucht, die Folgen von exzessivem onkogenen Stress auf das Target p53 in immortalisierten Zellen zu verstehen. Es ist bekannt, dass p53 die Aktivität von L1-Elementen im Genom kontrolliert, indem es Insertionen von L1 unterbindet. Man weiß, dass die Aktivität von L1 transkriptionell durch TRIM28 gehemmt wird, wohingegen im Fall der Umgehung, die APOBEC-Enzymgruppe die L1-Aktivität durch den RNA-Editierungsmechanismus steuert. Als neuen Befund konnte ich eine Herabregulierung von *TRIM28* und *APOBEC3B* im Seneszenzzustand beobachten, was auf einen Verlust der Regulation der L1-Elemente hinweist. Interessanterweise werden L1-Elemente in Zellen, die die Seneszenz durchlaufen, hochreguliert. Wahrscheinlich ist die Reaktivierung der L1-Elemente die Folge der Deaktivierung von Genen der Wirtsabwehr (*TRIM28* und *APOBEC3B*), die die Regulation von L1-Elementen kontrollieren. Um eventuell irreführende Befunde auf Transkriptionsebene zu vermeiden, habe ich zusätzlich die Translationsmuster von L1-ORFs durch den Einsatz von Ribo-Seq überprüft. Bei der post-transkriptionalen Analyse habe ich in den Zellen, die Seneszenz durchlaufen, ein signifikantes Muster von L1-kodierenden Transkripten entdeckt. Diese Beobachtung lässt sich durch die Deaktivierung von RNA-editierenden Enzymen erklären. Des Weiteren habe ich in silico die durch die Enzyme ADAR und APOBEC3B hervorgerufenen RNA-editierenden Ereignisse in sehr alten und mittelalten L1-Subfamilien untersucht. Ein deutlich unterschiedliches APOBEC3B-abhängiges RNA-Editierungsmuster konnte zwischen proliferierenden und in Seneszenz befindlichen Zellen an den Loci der jüngsten L1-Elemente, die zum Springen fähig sind, beobachtet werden. Wenn Einzelstrangbrücke in der DNA einen gewissen Schwellenwert übersteigen, generieren L1-Elemente Insertionen und erzeugen während ihrer Integration in das Genom Doppelstrangbrüche (DSBs). Ich habe solche Insertionsereignisse durch den Nachweis chimärer Transkripte zwischen L1-Elementen und genomischen Regionen gefunden. Interessanterweise war das Motiv für Target Site- Duplikationen (TSDs) an der de novo-Verbindung von L1-Elementen und genomischer Sequenz angereichert, was auf neue Insertionen in den Seneszenz-Zellen hinweist. Unsere Ergebnisse lassen darauf schließen, dass exzessiver onkogener Stress in immortalisierten Zellen mit dem Springen von L1-Elementen einhergeht, wodurch Doppelstrangbrüche verursacht und die Zellen befähigt werden, in die zelluläre Seneszenz überzugehen.

Bibliography

[1] James D Watson and Francis HC Crick. "Genetical implications of the structure of deoxyribonucleic acid". In: *Nature* 171.4361 (1953), pp. 964–967.

[2] Torbjörn Caspersson et al. "Identification of human chromosomes by DNA-binding fluorescent agents". In: *Chromosoma* 30.2 (1970), pp. 215–227.

[3] Mark B Gerstein et al. "What is a gene, post-ENCODE? History and updated definition". In: *Genome research* 17.6 (2007), pp. 669–681.

[4] Walter Gilbert. "Why genes in pieces?" In: *Nature* 271.5645 (1978), pp. 501–501.

[5] Thomas R Gingeras. "Origin of phenotypes: genes and transcripts". In: *Genome research* 17.6 (2007), pp. 682–690.

[6] Patrick Cramer et al. "Architecture of RNA polymerase II and implications for the transcription mechanism". In: *Science* 288.5466 (2000), pp. 640–649.

[7] Steven Hahn. "Structure and mechanism of the RNA polymerase II transcription machinery". In: *Nature Structural and Molecular Biology* 11.5 (2004), p. 394.

[8] Francisco J Asturias and John L Craighead. "RNA polymerase II at initiation". In: *Proceedings of the National Academy of Sciences* 100.12 (2003), pp. 6893–6895.

[9] Joachim Frank et al. "The process of mRNA–tRNA translocation". In: *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 19671–19678.

[10] Fabrizio d'Adda di Fagagna et al. "A DNA damage checkpoint response in telomere-initiated senescence". In: *Nature* 426.6963 (2003), pp. 194–198.

[11] Leonard Hayflick. "The limited in vitro lifetime of human diploid cell strains". In: *Experimental cell research* 37.3 (1965), pp. 614–636.

[12] Judith Campisi and Fabrizio d'Adda di Fagagna. "Cellular senescence: when bad things happen to good cells". In: *Nature reviews. Molecular cell biology* 8.9 (2007), p. 729.

[13] Marco Demaria et al. "Cellular senescence promotes adverse effects of chemotherapy and cancer relapse". In: *Cancer discovery* 7.2 (2017), pp. 165–176.

[14] Judith Campisi. "Aging, cellular senescence, and cancer". In: *Annual review of physiology* 75 (2013), pp. 685–705.

[15] Francis Rodier and Judith Campisi. “Four faces of cellular senescence”. In: *The Journal of cell biology* (2011), jcb–201009094.

[16] Jean-Philippe Coppé et al. “The senescence-associated secretory phenotype: the dark side of tumor suppression”. In: *Annual Review of Pathological Mechanical Disease* 5 (2010), pp. 99–118.

[17] Manuel Collado, Maria A Blasco, and Manuel Serrano. “Cellular senescence in cancer and aging”. In: *Cell* 130.2 (2007), pp. 223–233.

[18] Zhenkun Lou and Junjie Chen. “Cellular senescence and DNA repair”. In: *Experimental cell research* 312.14 (2006), pp. 2641–2646.

[19] Jan M Van Deursen. “The role of senescent cells in ageing”. In: *Nature* 509.7501 (2014), pp. 439–446.

[20] Stephen P Jackson and Jiri Bartek. “The DNA-damage response in human biology and disease”. In: *Nature* 461.7267 (2009), pp. 1071–1078.

[21] Leslie Pray. “DNA replication and causes of mutation”. In: *Nature education* 1.1 (2008), p. 214.

[22] Anuja Mehta and James E Haber. “Sources of DNA double-strand breaks and models of recombinational DNA repair”. In: *Cold Spring Harbor perspectives in biology* 6.9 (2014), a016428.

[23] Fabrizio D’adda Di Fagagna. “Living on a break: cellular senescence as a DNA-damage response”. In: *Nature reviews. Cancer* 8.7 (2008), p. 512.

[24] Véronique Gire et al. “DNA damage checkpoint kinase Chk2 triggers replicative senescence”. In: *The EMBO journal* 23.13 (2004), pp. 2554–2563.

[25] Evan A Farkash and Eline T Luning Prak. “DNA damage and L1 retrotransposition”. In: *BioMed Research International* 2006 (2006).

[26] S Mehdi Belgnaoui et al. “Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells”. In: *Cancer cell international* 6.1 (2006), p. 13.

[27] Jan HJ Hoeijmakers. “Genome maintenance mechanisms for preventing cancer”. In: *nature* 411.6835 (2001), p. 366.

[28] Orlando D Schärer. “Nucleotide excision repair in eukaryotes”. In: *Cold Spring Harbor perspectives in biology* 5.10 (2013), a012609.

[29] Anthony J Davis and David J Chen. “DNA double strand break repair via non-homologous end-joining”. In: *Translational cancer research* 2.3 (2013), p. 130.

[30] Arnold J Levine, Jamil Momand, and Cathy A Finlay. “The p53 tumour suppressor gene”. In: *Nature* 351.6326 (1991), p. 453.

[31] Subhasree Nag et al. “The MDM2-p53 pathway revisited”. In: *Journal of biomedical research* 27.4 (2013), p. 254.

[32] Toshinori Ozaki and Akira Nakagawara. “Role of p53 in cell death and human cancers”. In: *Cancers* 3.1 (2011), pp. 994–1013.

[33] Scott E Kern et al. “Identification of p53 as a sequence-specific DNA-binding protein”. In: *Science* 252.5013 (1991), pp. 1708–1712.

[34] Noa Rivlin et al. “Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis”. In: *Genes & cancer* 2.4 (2011), pp. 466–474.

[35] Elie Dolgin. “THE GREATEST HITS OF THE HUMAN GENOME”. In: *Nature* 551.7681 (2017), pp. 427–431.

[36] Annika Wylie et al. “p53 genes function to restrain mobile elements”. In: *Genes & development* 30.1 (2016), pp. 64–77.

[37] CR Harris et al. “p53 responsive elements in human retrotransposons”. In: *Oncogene* 28.44 (2009), p. 3857.

[38] Abdelali Haoudi et al. “Retrotransposition-competent human LINE-1 induces apoptosis in cancer cells with intact p53”. In: *BioMed Research International* 2004.4 (2004), pp. 185–194.

[39] Selvam Ayarpadikannan and Heui-Soo Kim. “The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases”. In: *Genomics & informatics* 12.3 (2014), pp. 98–104.

[40] Richard Cordaux and Mark A Batzer. “The impact of retrotransposons on human genome evolution”. In: *Nature reviews. Genetics* 10.10 (2009), p. 691.

[41] Barbara McClintock. *The significance of responses of the genome to challenge.* 1993.

[42] Susan R Wessler. *Transposable elements and the evolution of eukaryotic genomes.* 2006.

[43] Michael Cowley and Rebecca J Oakey. “Transposable elements re-wire and fine-tune the transcriptome”. In: *PLoS genetics* 9.1 (2013), e1003234.

[44] Cédric Feschotte. “The contribution of transposable elements to the evolution of regulatory networks”. In: *Nature reviews. Genetics* 9.5 (2008), p. 397.

[45] Henry L Levin and John V Moran. “Dynamic interactions between transposable elements and their hosts”. In: *Nature reviews. Genetics* 12.9 (2011), p. 615.

[46] Aurélie Hua-Van et al. “The struggle for life of the genome’s selfish architects”. In: *Biology direct* 6.1 (2011), p. 19.

[47] David J Finnegan. “Transposable elements”. In: *Current opinion in genetics & development* 2.6 (1992), pp. 861–867.

[48] AP Jason de Koning et al. “Repetitive elements may comprise over two-thirds of the human genome”. In: *PLoS genetics* 7.12 (2011), e1002384.

[49] Thomas Wicker et al. “A unified classification system for eukaryotic transposable elements”. In: *Nature Reviews Genetics* 8.12 (2007), pp. 973–982.

[50] John K Pace and Cédric Feschotte. "The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage". In: *Genome research* 17.4 (2007), pp. 422–432.

[51] David J. Finnegan. "Eukaryotic transposable elements and genome evolution". In: *Trends in Genetics* 5.Supplement C (1989), pp. 103–107. ISSN: 0168-9525. DOI: https://doi.org/10.1016/0168-9525(89)90039-5. URL: http://www.sciencedirect.com/science/article/pii/0168952589900395.

[52] Ericka R Havecker, Xiang Gao, and Daniel F Voytas. "The diversity of LTR retrotransposons". In: *Genome biology* 5.6 (2004), p. 225.

[53] Guia Guffanti et al. "Transposable elements and psychiatric disorders". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 165.3 (2014), pp. 201–216.

[54] Christine R Beck et al. "LINE-1 elements in structural variation and disease". In: *Annual review of genomics and human genetics* 12 (2011), pp. 187–215.

[55] Matthew T Reilly et al. "The role of transposable elements in health and diseases of the central nervous system". In: *Journal of Neuroscience* 33.45 (2013), pp. 17577–17586.

[56] David Moyes, David J Griffiths, and Patrick J Venables. "Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease". In: *Trends in genetics* 23.7 (2007), pp. 326–333.

[57] Nicola Tugnet et al. "Human endogenous retroviruses (HERVs) and autoimmune rheumatic disease: is there a link?" In: *The open rheumatology journal* 7 (2013), p. 13.

[58] Nathaly Castro-Diaz et al. "Evolutionally dynamic L1 regulation in embryonic stem cells". In: *Genes & development* 28.13 (2014), pp. 1397–1409.

[59] Helen M Rowe and Didier Trono. "Dynamic control of endogenous retroviruses during development". In: *Virology* 411.2 (2011), pp. 273–287.

[60] Cédric Feschotte and Clément Gilbert. "Endogenous viruses: insights into viral evolution and impact on host biology". In: *Nature Reviews Genetics* 13.4 (2012), pp. 283–296.

[61] Ting Wang et al. "Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53". In: *Proceedings of the National Academy of Sciences* 104.47 (2007), pp. 18613–18618.

[62] Norbert Bannert and Reinhard Kurth. "Retroelements and the human genome: new perspectives on an old relation". In: *Proceedings of the National Academy of Sciences* 101.suppl 2 (2004), pp. 14572–14579.

[63] Eline T Luning Prak and Haig H Kazazian Jr. "Mobile elements and the human genome". In: *Nature reviews. Genetics* 1.2 (2000), p. 134.

[64] IHGSC IHGSC. "Initial sequencing and analysis of the humane genome". In: *Nature* 409 (2001), pp. 520–562.

[65] Alan M Weiner. “SINEs and LINEs: the art of biting the hand that feeds you”. In: *Current opinion in cell biology* 14.3 (2002), pp. 343–350.

[66] Abdel-Halim Salem et al. “Alu elements and hominid phylogenetics”. In: *Proceedings of the National Academy of Sciences* 100.22 (2003), pp. 12787–12791.

[67] Andrew J Oler et al. “Alu expression in human cell lines and their retrotranspositional potential”. In: *Mobile DNA* 3.1 (2012), p. 11.

[68] Jinchuan Xing et al. “Mobile DNA elements in primate and human evolution”. In: *American journal of physical anthropology* 134.S45 (2007), pp. 2–19.

[69] Prescott Deininger. “Alu elements: know the SINEs”. In: *Genome biology* 12.12 (2011), p. 236.

[70] Victoria P Belancio, Dale J Hedges, and Prescott Deininger. “Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health”. In: *Genome research* 18.3 (2008), pp. 343–358.

[71] Haig H Kazazian Jr. “Mobile elements and disease”. In: *Current Opinion in Genetics & Development* 8.3 (1998), pp. 343–350. ISSN: 0959-437X. DOI: https://doi.org/10.1016/S0959-437X(98)80092-0. URL: http://www.sciencedirect.com/science/article/pii/S0959437X98800920.

[72] Stéphane Boissinot et al. “The insertional history of an active family of L1 retrotransposons in humans”. In: *Genome research* 14.7 (2004), pp. 1221–1231.

[73] Tatjana Singer et al. “LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes?” In: *Trends in Neurosciences* 33.8 (2010), pp. 345–354. ISSN: 0166-2236. DOI: https://doi.org/10.1016/j.tins.2010.04.001. URL: http://www.sciencedirect.com/science/article/pii/S0166223610000536.

[74] Hirohiko Hohjoh and Maxine F Singer. *Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon.* 1997.

[75] Hirohiko Hohjoh and Maxine F Singer. “Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon”. In: *The EMBO journal* 16.19 (1997), pp. 6034–6043.

[76] Qinghua Feng et al. “Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition”. In: *Cell* 87.5 (1996), pp. 905–916.

[77] SL Mathias et al. “Reverse transcriptase encoded by a human transposable element”. In: *Science* 254.5039 (1991), pp. 1808–1810. ISSN: 0036-8075. DOI: 10.1126/science.1722352. eprint: http://science.sciencemag.org/content/254/5039/1808.full.pdf. URL: http://science.sciencemag.org/content/254/5039/1808.

[78] Ryan E Mills et al. “Which transposable elements are active in the human genome”. In: *Trends in genetics* 23.4 (2007), pp. 183–191.

[79] Wenyong Ding et al. “L1 elements, processed pseudogenes and retrogenes in mammalian genomes”. In: *IUBMB life* 58.12 (2006), pp. 677–685.

[80] Dustin C Hancks and Haig H Kazazian. “Active human retrotransposons: variation and disease”. In: *Current opinion in genetics & development* 22.3 (2012), pp. 191–203.

[81] Brook Brouha et al. “Hot L1s account for the bulk of retrotransposition in the human population”. In: *Proceedings of the National Academy of Sciences* 100.9 (2003), pp. 5280–5285.

[82] Stéphane Boissinot, Ali Entezam, and Anthony V Furano. “Selection against deleterious LINE-1-containing loci in the human lineage”. In: *Molecular biology and evolution* 18.6 (2001), pp. 926–935.

[83] S Boissinot and AV Furano. “The recent evolution of human L1 retrotransposons”. In: *Cytogenetic and genome research* 110.1-4 (2005), pp. 402–406.

[84] Hameed Khan, Arian Smit, and Stéphane Boissinot. “Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates”. In: *Genome research* 16.1 (2006), pp. 78–87.

[85] Frank MJ Jacobs et al. “An evolutionary arms race between KRAB zinc finger genes 91/93 and SVA/L1 retrotransposons”. In: *Nature* 516.7530 (2014), p. 242.

[86] Fang-miin Sheen et al. “Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition”. In: *Genome research* 10.10 (2000), pp. 1496–1508.

[87] Heeyoun Bunch and Stuart K Calderwood. “TRIM28 as a novel transcriptional elongation factor”. In: *BMC molecular biology* 16.1 (2015), p. 14.

[88] Daria V Babushok and Haig H Kazazian. “Progress in understanding the biology of the human mutagen LINE-1”. In: *Human mutation* 28.6 (2007), pp. 527–539.

[89] Mark A Batzer and Prescott L Deininger. “Alu repeats and human genomic diversity”. In: *Nature reviews genetics* 3.5 (2002), pp. 370–379.

[90] Yoshio Miki et al. “Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer”. In: *Cancer research* 52.3 (1992), pp. 643–645.

[91] Emma C Scott et al. “A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer”. In: *Genome research* 26.6 (2016), pp. 745–755.

[92] Eric M Ostertag, Blair B Madison, and Hiroki Kano. “Mutagenesis in rodents using the L1 retrotransposon”. In: *Genome biology* 8.1 (2007), S16.

[93] Wei Wei et al. “Human L1 retrotransposition: cispreference versus trans complementation”. In: *Molecular and cellular biology* 21.4 (2001), pp. 1429–1439.

[94] Deanna A Kulpa and John V Moran. "Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition". In: *Human molecular genetics* 14.21 (2005), pp. 3237–3248.

[95] Aurélien J Doucet et al. "Characterization of LINE-1 ribonucleoprotein particles". In: *PLoS genetics* 6.10 (2010), e1001150.

[96] Gregory J Cost et al. "Human L1 element target-primed reverse transcription in vitro". In: *The EMBO journal* 21.21 (2002), pp. 5899–5910.

[97] Tammy A Morrish et al. "DNA repair mediated by endonuclease-independent LINE-1 retrotransposition". In: *Nature genetics* 31.2 (2002), p. 159.

[98] Shawn M Christensen, Junqiang Ye, and Thomas H Eickbush. "RNA from the 5 end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site". In: *Proceedings of the National Academy of Sciences* 103.47 (2006), pp. 17602–17607.

[99] Hirohiko Hohjoh and Mxine F Singer. "Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA." In: *The EMBO journal* 15.3 (1996), p. 630.

[100] Javier G Pizarro and Gaël Cristofari. "Post-transcriptional control of LINE-1 retrotransposition by cellular host factors in somatic cells". In: *Frontiers in cell and developmental biology* 4 (2016).

[101] Prescott L Deininger et al. "Mobile elements and mammalian genome evolution". In: *Current opinion in genetics & development* 13.6 (2003), pp. 651–658.

[102] Shuji Kubo et al. "L1 retrotransposition in nondividing and primary human somatic cells". In: *Proceedings of the National Academy of Sciences* 103.21 (2006), pp. 8036–8041.

[103] Jose L Garcia-Perez et al. "LINE-1 retrotransposition in human embryonic stem cells". In: *Human molecular genetics* 16.13 (2007), pp. 1569–1577.

[104] Hiroki Kano et al. "L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism". In: *Genes & development* 23.11 (2009), pp. 1303–1312.

[105] J Kenneth Baillie et al. "Somatic retrotransposition alters the genetic landscape of the human brain". In: *Nature* 479.7374 (2011), p. 534.

[106] Alysson R Muotri et al. "Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition". In: *nature* 435.7044 (2005), p. 903.

[107] Nicole G Coufal et al. "L1 retrotransposition in human neural progenitor cells". In: *Nature* 460.7259 (2009), p. 1127.

[108] Christine R Beck et al. "LINE-1 retrotransposition activity in human genomes". In: *Cell* 141.7 (2010), pp. 1159–1170.

[109] Stephen L Gasior, Astrid M Roy-Engel, and Prescott L Deininger. "ERCC1/XPF limits L1 retrotransposition". In: *DNA repair* 7.6 (2008), pp. 983–989.

[110] Tammy A Morrish et al. "Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres". In: *Nature* 446.7132 (2007), p. 208.

[111] Victoria P Belancio, Prescott L Deininger, and Astrid M Roy-Engel. "LINE dancing in the human genome: transposable elements and disease". In: *Genome medicine* 1.10 (2009), p. 97.

[112] Josh R Friedman et al. "KAP-1, a novel corepressor for the highly conserved KRAB repression domain." In: *Genes & development* 10.16 (1996), pp. 2067–2078.

[113] Daniel Wolf and Stephen P Goff. "TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells". In: *Cell* 131.1 (2007), pp. 46–57.

[114] Helen M Rowe et al. "KAP1 controls endogenous retroviruses in embryonic stem cells". In: *Nature* 463.7278 (2010), p. 237.

[115] Gernot Wolf et al. "The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses". In: *Genes & development* 29.5 (2015), pp. 538–554.

[116] Andrew Paul Hutchins and Duanqing Pei. "Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs". In: *Science bulletin* 60.20 (2015), pp. 1722–1733.

[117] Hongzhuang Peng, Irina Feldman, and Frank J Rauscher. "Hetero-oligomerization among the TIF family of RBCC/TRIM domain-containing nuclear cofactors: a potential mechanism for regulating the switch between coactivation and corepression". In: *Journal of molecular biology* 320.3 (2002), pp. 629–644.

[118] Toshiyuki Matsui et al. "Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET". In: *Nature* 464.7290 (2010), p. 927.

[119] Liana Fasching et al. "TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells". In: *Cell reports* 10.1 (2015), pp. 20–28.

[120] Christine Steinhoff and WA Schulz. "Transcriptional regulation of the human LINE-1 retrotransposon L1. 2B". In: *Molecular Genetics and Genomics* 270.5 (2004), pp. 394–402.

[121] Dan Dominissini et al. "Adenosine-to-inosine RNA editing meets cancer". In: *Carcinogenesis* 32.11 (2011), pp. 1569–1577.

[122] Susan M Rueter, T Renee Dawson, and Ronald B Emeson. "Regulation of alternative splicing by RNA editing". In: *Nature* 399.6731 (1999), p. 75.

[123] Jurga Laurencikiene et al. "RNA editing and alternative splicing: the importance of co-transcriptional coordination". In: *EMBO reports* 7.3 (2006), pp. 303–307.

[124] Erez Y Levanon et al. "Systematic identification of abundant A-to-I editing sites in the human transcriptome". In: *arXiv preprint q-bio/0411045* (2004).

[125] Erez Y Levanon et al. "Evolutionarily conserved human targets of adenosine to inosine RNA editing". In: *Nucleic acids research* 33.4 (2005), pp. 1162–1168.

[126] Prem Mohini Sharma et al. "RNA editing in the Wilms' tumor susceptibility gene, WT1." In: *Genes & development* 8.6 (1994), pp. 720–731.

[127] Valerie Blanc and Nicholas O Davidson. "C-to-U RNA editing: mechanisms leading to genetic diversity". In: *Journal of Biological Chemistry* 278.3 (2003), pp. 1395–1398.

[128] Wei Tang, Yongjun Fei, and Michael Page. "Biological significance of RNA editing in cells". In: *Molecular biotechnology* 52.1 (2012), pp. 91–100.

[129] Eddie Park et al. "RNA editing in the human ENCODE RNA-seq data". In: *Genome research* 22.9 (2012), pp. 1626–1633.

[130] Louis Valente and Kazuko Nishikura. "ADAR gene family and A-to-I RNA editing: diverse roles in posttranscriptional gene regulation". In: *Progress in nucleic acid research and molecular biology* 79 (2005), pp. 299–338.

[131] Linda Bross et al. "DNA double-strand breaks". In: *Journal of Experimental Medicine* 195.9 (2002), pp. 1187–1192.

[132] Ann M Sheehy et al. "Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein". In: *Nature* 418.6898 (2002), p. 646.

[133] Sara L Sawyer, Michael Emerman, and Harmit S Malik. "Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G". In: *PLoS biology* 2.9 (2004), e275.

[134] Hal P Bogerd et al. "APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells". In: *Nucleic acids research* 34.1 (2006), pp. 89–95.

[135] Renato S Aguiar and B Matija Peterlin. "APOBEC3 proteins and reverse transcription". In: *Virus research* 134.1 (2008), pp. 74–85.

[136] Spyridon Stavrou and Susan R Ross. "APOBEC3 proteins in viral immunity". In: *The Journal of Immunology* 195.10 (2015), pp. 4565–4570.

[137] Mark D Stenglein and Reuben S Harris. "APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism". In: *Journal of Biological Chemistry* 281.25 (2006), pp. 16837–16841.

[138] Hal P Bogerd et al. "Cellular inhibitors of long interspersed element 1 and Alu retrotransposition". In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8780–8785.

[139] Atsushi Koito and Terumasa Ikeda. "Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases". In: *Frontiers in microbiology* 4 (2013).

[140] Shraddha Sharma et al. "APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages". In: *Nature communications* 6 (2015).

[141] Michael B Burns, Nuri A Temiz, and Reuben S Harris. "Evidence for APOBEC3B mutagenesis in multiple human cancers". In: *Nature genetics* 45.9 (2013), pp. 977–983.

[142] Zhiyu Peng et al. "Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome". In: *Nature biotechnology* 30.3 (2012), pp. 253–260.

[143] Jae Hoon Bahn et al. "Accurate identification of A-to-I RNA editing in human by transcriptome sequencing". In: *Genome research* 22.1 (2012), pp. 142–150.

[144] Elisa Orecchini, Loredana Frassinelli, and Alessandro Michienzi. "Restricting retrotransposons: ADAR1 is another guardian of the human genome". In: *RNA biology* 14.11 (2017), pp. 1485–1491.

[145] Genıs Parra et al. "Tandem chimerism as a means to increase protein complexity in the human genome". In: *Genome research* 16.1 (2006), pp. 37–44.

[146] Anna Williford and Esther Betrán. "Gene Fusion". In: *eLS* ().

[147] Felix Mitelman, Bertil Johansson, and Fredrik Mertens. "The impact of translocations and gene fusions on cancer causation". In: *Nature Reviews Cancer* 7.4 (2007), p. 233.

[148] Matteo Benelli et al. "Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript". In: *Bioinformatics* 28.24 (2012), pp. 3232–3239.

[149] Scott A Tomlins et al. "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer". In: *science* 310.5748 (2005), pp. 644–648.

[150] Frederick Sanger, Steven Nicklen, and Alan R Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[151] Laura Bonetta. "Genome sequencing in the fast lane". In: *Nature Methods* 3.2 (2006), pp. 141–147.

[152] Allan M Maxam and Walter Gilbert. "A new method for sequencing DNA". In: *Proceedings of the National Academy of Sciences* 74.2 (1977), pp. 560–564.

[153] J Craig Venter et al. "The sequence of the human genome". In: *science* 291.5507 (2001), pp. 1304–1351.

[154] International Human Genome Sequencing Consortium et al. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (2004), pp. 931–945.

[155] Erwin L Van Dijk et al. "Ten years of next-generation sequencing technology". In: *Trends in genetics* 30.9 (2014), pp. 418–426.

[156] Michael L Metzker. "Sequencing technologies–the next generation". In: *Nature reviews. Genetics* 11.1 (2010), p. 31.

[157] José F Siqueira, Ashraf F Fouad, and Isabela N Roˆcas. "Pyrosequencing as a tool for better understanding of human microbiomes". In: *Journal of oral microbiology* 4.1 (2012), p. 10743.

[158] Jay Shendure and Hanlee Ji. "Next-generation DNA sequencing". In: *Nature biotechnology* 26.10 (2008), p. 1135.

[159] Lin Liu et al. "Comparison of next-generation sequencing systems". In: *BioMed Research International* 2012 (2012).

[160] Jason A Reuter, Damek V Spacek, and Michael P Snyder. "High-throughput sequencing technologies". In: *Molecular cell* 58.4 (2015), pp. 586–597.

[161] Elaine R Mardis. "Next-generation DNA sequencing methods". In: *Annu. Rev. Genomics Hum. Genet.* 9 (2008), pp. 387–402.

[162] Patrice Milos. "Helicos BioSciences". In: (2008).

[163] Martin Kircher and Janet Kelso. "High-throughput DNA sequencing–concepts and limitations". In: *Bioessays* 32.6 (2010), pp. 524–536.

[164] Mihai Pop and Steven L Salzberg. "Bioinformatics challenges of new sequencing technology". In: *Trends in Genetics* 24.3 (2008), pp. 142–149.

[165] Michael A Quail et al. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers". In: *BMC genomics* 13.1 (2012), p. 341.

[166] John Eid et al. "Real-time DNA sequencing from single polymerase molecules". In: *Science* 323.5910 (2009), pp. 133–138.

[167] Meni Wanunu. "Nanopores: A journey towards DNA sequencing". In: *Physics of life reviews* 9.2 (2012), pp. 125–158.

[168] Anthony Rhoads and Kin Fai Au. "PacBio sequencing and its applications". In: *Genomics, proteomics & bioinformatics* 13.5 (2015), pp. 278–289.

[169] Ian M Derrington et al. "Nanopore DNA sequencing with MspA". In: *Proceedings of the National Academy of Sciences* 107.37 (2010), pp. 16060–16065.

[170] Andrew H Laszlo et al. "Decoding long nanopore sequencing reads of natural DNA". In: *Nature biotechnology* 32.8 (2014), pp. 829–833.

[171] David S Johnson et al. "Genome-wide mapping of in vivo protein-DNA interactions". In: *Science* 316.5830 (2007), pp. 1497–1502.

[172] Stephen G Landt et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". In: *Genome research* 22.9 (2012), pp. 1813–1831.

[173] Hyunjin Shin et al. "Computational methodology for ChIP-seq analysis". In: *Quantitative biology* 1.1 (2013), pp. 54–70.

[174] A Gordon Robertson et al. "Genome-wide relationship between histone H3 lysine 4 mono-and tri-methylation and transcription factor binding". In: *Genome research* 18.12 (2008), pp. 1906–1917.

[175] Peter J Park. "ChIP-seq: advantages and challenges of a maturing technology". In: *Nature reviews. Genetics* 10.10 (2009), p. 669.

[176] Artem Barski et al. “High-resolution profiling of histone methylations in the human genome”. In: *Cell* 129.4 (2007), pp. 823–837.

[177] Nathaniel D Heintzman et al. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”. In: *Nature genetics* 39.3 (2007), p. 311.

[178] Yin Shen et al. “A map of the cis-regulatory sequences in the mouse genome”. In: *Nature* 488.7409 (2012), p. 116.

[179] Bing Li, Michael Carey, and Jerry L Workman. “The role of chromatin during transcription”. In: *Cell* 128.4 (2007), pp. 707–719.

[180] Artem Barski and Keji Zhao. “Genomic location analysis by ChIP-Seq”. In: *Journal of cellular biochemistry* 107.1 (2009), pp. 11–18.

[181] Debasish Raha, Miyoung Hong, and Michael Snyder. “ChIP-Seq: A method for global identification of regulatory elements in the genome”. In: *Current protocols in molecular biology* (2010), pp. 21–19.

[182] Peggy Farnham. “Insights from genomic profiling of transcription factors”. In: *Nature reviews. Genetics* 10.9 (2009), p. 605.

[183] Elaine R Mardis. “ChIP-seq: welcome to the new frontier”. In: *Nature methods* 4.8 (2007), pp. 613–614.

[184] Christoph D Schmid and Philipp Bucher. “ChIP-Seq data reveal nucleosome architecture of human promoters”. In: *Cell* 131.5 (2007), pp. 831–832.

[185] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[186] Ali Mortazavi et al. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature methods* 5.7 (2008), pp. 621–628.

[187] Manfred G Grabherr et al. “Full-length transcriptome assembly from RNA-Seq data without a reference genome”. In: *Nature biotechnology* 29.7 (2011), pp. 644–652.

[188] Olena Morozova, Martin Hirst, and Marco A Marra. “Applications of new sequencing technologies for transcriptome analysis”. In: *Annual review of genomics and human genetics* 10 (2009), pp. 135–151.

[189] Ryan Lister et al. “Highly integrated single-base resolution maps of the epigenome in Arabidopsis”. In: *Cell* 133.3 (2008), pp. 523–536.

[190] Samuel Marguerat and Jürg Bähler. “RNA-seq: from technology to biology”. In: *Cellular and molecular life sciences* 67.4 (2010), pp. 569–579.

[191] Ugrappa Nagalakshmi et al. “The transcriptional landscape of the yeast genome defined by RNA sequencing”. In: *Science* 320.5881 (2008), pp. 1344–1349.

[192] Joshua Z Levin et al. “Comprehensive comparative analysis of strand-specific RNA sequencing methods”. In: *Nature methods* 7.9 (2010), pp. 709–715.

[193] Patrick O Brown and David Botstein. "Exploring the new world of the genome with DNA microarrays." In: *Nature genetics* 21 (1999).

[194] Joan Argetsinger Steitz. "Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA". In: *Nature* 224.5223 (1969), pp. 957–964.

[195] Nicholas T Ingolia et al. "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling". In: *science* 324.5924 (2009), pp. 218–223.

[196] Yoav Arava et al. "Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae". In: *Proceedings of the National Academy of Sciences* 100.7 (2003), pp. 3889–3894.

[197] Nicholas T Ingolia. "Ribosome profiling: new views of translation, from single codons to genome scale". In: *Nature reviews. Genetics* 15.3 (2014), p. 205.

[198] Nicholas T Ingolia et al. "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments". In: *Nature protocols* 7.8 (2012), p. 1534.

[199] Nicholas A Wallace, Victoria P Belancio, and Prescott L Deininger. "L1 mobile element expression causes multiple types of toxicity". In: *Gene* 419.1 (2008), pp. 75–81.

[200] Georges St Laurent III, Neil Hammell, and Timothy A McCaffrey. "A LINE-1 component to human aging: do LINE elements exact a longevity cost for evolutionary advantage?" In: *Mechanisms of ageing and development* 131.5 (2010), pp. 299–305.

[201] Sophie Schbath et al. "Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis". In: *Journal of Computational Biology* 19.6 (2012), pp. 796–813.

[202] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome biology* 10.3 (2009), R25.

[203] Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". In: *Genome research* 18.11 (2008), pp. 1851–1858.

[204] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[205] Can Alkan et al. "Personalized copy number and segmental duplication maps using next-generation sequencing". In: *Nature genetics* 41.10 (2009), pp. 1061–1067.

[206] Stephen M Rumble et al. "SHRiMP: accurate mapping of short color-space reads". In: *PLoS computational biology* 5.5 (2009), e1000386.

[207] Cole Trapnell, Lior Pachter, and Steven L Salzberg. "TopHat: discovering splice junctions with RNA-Seq". In: *Bioinformatics* 25.9 (2009), pp. 1105–1111.

[208] Daehwan Kim et al. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome biology* 14.4 (2013), R36.

[209] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[210] Temple F Smith and Michael S Waterman. “Identification of common molecular subsequences”. In: *Journal of molecular biology* 147.1 (1981), pp. 195–197.

[211] Saul B Needleman and Christian D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of molecular biology* 48.3 (1970), pp. 443–453.

[212] Jing Shang et al. “Evaluation and comparison of multiple aligners for next-generation sequencing data analysis”. In: *BioMed research international* 2014 (2014).

[213] Michael Burrows and David J Wheeler. “A block-sorting lossless data compression algorithm”. In: (1994).

[214] Raja Jothi et al. “Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data”. In: *Nucleic acids research* 36.16 (2008), pp. 5221–5231.

[215] Jialin Xu and Yu Zhang. “A generalized linear model for peak calling in ChIP-Seq data”. In: *Journal of Computational Biology* 19.6 (2012), pp. 826–838.

[216] Yong Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9 (2008), R137.

[217] Jianxing Feng, Tao Liu, and Yong Zhang. “Using MACS to identify peaks from ChIP-Seq data”. In: *Current protocols in bioinformatics* (2011), pp. 2–14.

[218] Anthony P Fejes et al. “FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology”. In: *Bioinformatics* 24.15 (2008), pp. 1729–1730.

[219] Anton Valouev et al. “Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data”. In: *Nature methods* 5.9 (2008), pp. 829–834.

[220] David Sims et al. “Sequencing depth and coverage: key considerations in genomic analyses”. In: *Nature Reviews Genetics* 15.2 (2014), pp. 121–132.

[221] Timothy Bailey et al. “Practical guidelines for the comprehensive analysis of ChIP-seq data”. In: *PLoS computational biology* 9.11 (2013), e1003326.

[222] Michael Ashburner et al. “Gene Ontology: tool for the unification of biology”. In: *Nature genetics* 25.1 (2000), pp. 25–29.

[223] Wenxiu Ma and Wing Hung Wong. “3 The Analysis of ChIP-Seq Data”. In: *Methods in enzymology* 497 (2011), p. 51.

[224] Timothy L Bailey et al. “MEME SUITE: tools for motif discovery and searching”. In: *Nucleic acids research* 37.suppl_2 (2009), W202–W208.

[225] Yarden Katz et al. “Analysis and design of RNA sequencing experiments for identifying isoform regulation”. In: *Nature methods* 7.12 (2010), pp. 1009–1015.

[226] Jin Billy Li et al. “Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing”. In: *Science* 324.5931 (2009), pp. 1210–1213.

[227] Marie-Agnès Dillies et al. “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”. In: *Briefings in bioinformatics* 14.6 (2013), pp. 671–683.

[228] Günter P Wagner, Koryu Kin, and Vincent J Lynch. “Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples”. In: *Theory in biosciences* 131.4 (2012), pp. 281–285.

[229] Cole Trapnell et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. In: *Nature protocols* 7.3 (2012), p. 562.

[230] In Seok Yang and Sangwoo Kim. “Analysis of whole transcriptome sequencing data: workflow and software”. In: *Genomics & informatics* 13.4 (2015), pp. 119–125.

[231] Cole Trapnell et al. “Differential analysis of gene regulation at transcript resolution with RNA-seq”. In: *Nature biotechnology* 31.1 (2013), pp. 46–53.

[232] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), p. 550.

[233] Heather Turner. “Introduction to generalized linear models”. In: *Rapport technique, Vienna University of Economics and Business* (2008).

[234] Dominique Lord, Byung-Jung Park, and Poisson-Gamma Model. “Negative Binomial Regression Models and Estimation Methods”. In: *Probability Density and Likelihood Functions* (2012), pp. 1–15.

[235] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[236] Jianxing Feng et al. “GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data”. In: *Bioinformatics* 28.21 (2012), pp. 2782–2788.

[237] Christopher WJ Smith and Juan Valcárcel. “Alternative pre-mRNA splicing: the logic of combinatorial control”. In: *Trends in biochemical sciences* 25.8 (2000), pp. 381–388.

[238] Keren L Witkin et al. “RNA editing, epitranscriptomics, and processing in cancer progression”. In: *Cancer biology & therapy* 16.1 (2015), pp. 21–27.

[239] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet. journal* 17.1 (2011), pp–10.

[240] Aaron R Quinlan and Ira M Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[241] Nicholas T Ingolia. "Ribosome footprint profiling of translation throughout the genome". In: *Cell* 165.1 (2016), pp. 22–33.

[242] Hui Jiang and Wing Hung Wong. "Statistical inferences for isoform expression in RNA-Seq". In: *Bioinformatics* 25.8 (2009), pp. 1026–1032.

[243] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". In: *Nucleic acids research* 37.1 (2008), pp. 1–13.

[244] Eran Eden et al. "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". In: *BMC bioinformatics* 10.1 (2009), p. 48.

[245] Huaiyu Mi et al. "Large-scale gene function analysis with the PANTHER classification system". In: *Nature protocols* 8.8 (2013), p. 1551.

[246] Gloria A Brar and Jonathan S Weissman. "Ribosome profiling reveals the what, when, where, and how of protein synthesis". In: *Nature reviews. Molecular cell biology* 16.11 (2015), p. 651.

[247] Fabricio Loayza-Puch et al. "p53 induces transcriptional and translational programs to suppress cell proliferation and growth". In: *Genome biology* 14.4 (2013), R32.

[248] Yang Liao, Gordon K Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (2013), pp. 923–930.

[249] Cole Trapnell et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature biotechnology* 28.5 (2010), pp. 511–515.

[250] W James Kent et al. "The human genome browser at UCSC". In: *Genome research* 12.6 (2002), pp. 996–1006.

[251] Aaron McKenna et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome research* 20.9 (2010), pp. 1297–1303.

[252] Mark A DePristo et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nature genetics* 43.5 (2011), pp. 491–498.

[253] Robert Piskol, Gokul Ramaswami, and Jin Billy Li. "Reliable identification of genomic variants from RNA-seq data". In: *The American Journal of Human Genetics* 93.4 (2013), pp. 641–651.

[254] Timothy L Bailey, Charles Elkan, et al. "Fitting a mixture model by expectation maximization to discover motifs in bipolymers". In: (1994).

[255] Hongen Zhang, Paul Meltzer, and Sean Davis. "RCircos: an R package for Circos 2D track plots". In: *BMC bioinformatics* 14.1 (2013), p. 244.

[256] Shobhit Gupta et al. “Quantifying similarity between motifs”. In: *Genome biology* 8.2 (2007), R24.

[257] Fidel Ramırez et al. “deepTools: a flexible platform for exploring deep-sequencing data”. In: *Nucleic acids research* 42.W1 (2014), W187–W191.

[258] Jay Shendure and Erez Lieberman Aiden. “The expanding scope of DNA sequencing”. In: *Nature biotechnology* 30.11 (2012), pp. 1084–1094.

[259] Claudia Angelini and Valerio Costa. “Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems”. In: *Frontiers in cell and developmental biology* 2 (2014).

[260] Weronika Sikora-Wohlfeld et al. “Assessing computational methods for transcription factor target gene identification based on ChIP-seq data”. In: *PLoS computational biology* 9.11 (2013), e1003342.

[261] Steffen Durinck et al. “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt”. In: *Nature protocols* 4.8 (2009), p. 1184.

[262] Guangchuang Yu and Qing-Yu He. “ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization”. In: *Molecular BioSystems* 12.2 (2016), pp. 477–479.

[263] Georgia Chatzinikolaou et al. “ERCC1–XPF cooperates with CTCF and cohesin to facilitate the developmental silencing of imprinted genes”. In: *Nature cell biology* 19.5 (2017), p. 421.

[264] Ryan P Welch et al. “ChIP-Enrich: gene set enrichment testing for ChIP-seq data”. In: *Nucleic acids research* 42.13 (2014), e105–e105.

[265] Ji Qiang Yao and Fahong Yu. “DEB: A web interface for RNA-seq digital gene expression analysis”. In: *Bioinformation* 7.1 (2011), p. 44.

[266] David E White et al. “KAP1, a novel substrate for PIKK family members, colocalizes with numerous damage response factors at DNA lesions”. In: *Cancer research* 66.24 (2006), pp. 11594–11599.

[267] Bryan R Cullen. “Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors”. In: *Journal of virology* 80.3 (2006), pp. 1067–1076.

[268] Masanobu Kinomoto et al. “All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition”. In: *Nucleic acids research* 35.9 (2007), pp. 2955–2964.

[269] Cécile Esnault et al. “APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses”. In: *Nature* 433.7024 (2005), pp. 430–434.

[270] Heide Muckenfuss et al. “APOBEC3 proteins inhibit human LINE-1 retrotransposition”. In: *Journal of Biological Chemistry* 281.31 (2006), pp. 22161–22172.

[271] Nika Lovšin and B Matija Peterlin. "APOBEC3 Proteins Inhibit LINE-1 Retrotransposition in the Absence of ORF1p Binding". In: *Annals of the New York Academy of Sciences* 1178.1 (2009), pp. 268–275.

[272] Jeffrey S Han, Suzanne T Szak, and Jef D Boeke. "Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes". In: *Nature* 429.6989 (2004), pp. 268–274.

[273] Sandra L Martin. "Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1". In: *RNA biology* 7.6 (2010), pp. 706–711.

[274] Daniel Ardeljan et al. "The human long interspersed element-1 retrotransposon: an emerging biomarker of neoplasia". In: *Clinical Chemistry* 63.4 (2017), pp. 816–822.

[275] Jungnam Lee et al. "Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons". In: *Gene* 390.1 (2007), pp. 18–27.

[276] Mihaela Babiceanu et al. "Recurrent chimeric fusion RNAs in non-cancer tissues and cells". In: *Nucleic acids research* 44.6 (2016), pp. 2859–2872.

[277] Anton Buzdin et al. "The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination". In: *Nucleic acids research* 31.15 (2003), pp. 4385–4390.

[278] Kathleen H Burns. "Transposable elements in cancer". In: *Nature Reviews Cancer* (2017).

[279] Francesca Telese et al. ""Seq-ing" insights into the epigenetics of neuronal gene regulation". In: *Neuron* 77.4 (2013), pp. 606–623.

[280] Michael J Buck and Jason D Lieb. "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments". In: *Genomics* 83.3 (2004), pp. 349–360.

[281] Terrence S Furey. "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions". In: *Nature reviews. Genetics* 13.12 (2012), p. 840.

[282] Kai-Oliver Mutz et al. "Transcriptome analysis using next-generation sequencing". In: *Current opinion in biotechnology* 24.1 (2013), pp. 22–30.

[283] Cliff Meldrum, Maria A Doyle, and Richard W Tothill. "Next-generation sequencing for cancer diagnostics: a practical perspective". In: *The Clinical Biochemist Reviews* 32.4 (2011), p. 177.

[284] Nicholas T Ingolia. "Genome-wide translational profiling by ribosome footprinting". In: *Methods in enzymology* 470 (2010), pp. 119–142.

[285] Dubravka Pezic et al. "piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells". In: *Genes & development* 28.13 (2014), pp. 1410–1428.

[286] Michael T Hemann and Masashi Narita. "Oncogenes and senescence: breaking down in the fast lane". In: *Genes & development* 21.1 (2007), pp. 1–5.

[287] WJ Mooi and DS Peeper. "Oncogene-induced cell senescence—halting on the road to cancer". In: *New England Journal of Medicine* 355.10 (2006), pp. 1037–1046.

[288] Nemanja Rodić and Kathleen H Burns. "Long interspersed element–1 (LINE-1): passenger or driver in human neoplasms?" In: *PLoS genetics* 9.3 (2013), e1003402.

[289] Sandra L Martin. "The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition". In: *BioMed Research International* 2006 (2006).

[290] John L Goodier et al. "LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex". In: *Molecular and cellular biology* 27.18 (2007), pp. 6469–6483.

[291] Nemanja Rodić et al. "Long interspersed element-1 protein expression is a hallmark of many human cancers". In: *The American journal of pathology* 184.5 (2014), pp. 1280–1286.

[292] Stephen L Gasior et al. "The human LINE-1 retrotransposon creates DNA double-strand breaks". In: *Journal of molecular biology* 357.5 (2006), pp. 1383–1393.

[293] Yasuo Ariumi. "Guardian of the human genome: host defense mechanisms against LINE-1 retrotransposition". In: *Frontiers in chemistry* 4 (2016).

[294] BaBie Teng, Charles F Burant, and Nicholas O Davidson. "Molecular cloning of an apolipoprotein B messenger RNA editing protein". In: *Science* 260.5115 (1993), pp. 1816–1819.

[295] Mingyao Li, Isabel X Wang, and Vivian G Cheung. "Response to comments on "Widespread RNA and DNA sequence differences in the human transcriptome"". In: *science* 335.6074 (2012), pp. 1302–1302.

[296] Elisa Orecchini et al. "ADAR1 restricts LINE-1 retrotransposition". In: *Nucleic acids research* 45.1 (2016), pp. 155–168.

[297] Silvestro G Conticello et al. "Evolution of the AID/APOBEC family of polynucleotide (deoxy) cytidine deaminases". In: *Molecular biology and evolution* 22.2 (2004), pp. 367–377.

[298] Richard N McLaughlin Jr et al. "Conservation and innovation of APOBEC3A restriction functions during primate evolution". In: *Molecular biology and evolution* 33.8 (2016), pp. 1889–1901.

[299] Charles Swanton et al. "APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity". In: *Cancer discovery* 5.7 (2015), pp. 704–712.

[300] John L Goodier, Ling E Cheung, and Haig H Kazazian Jr. "Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition". In: *Nucleic acids research* 41.15 (2013), pp. 7401–7419.

[301] Stephen L Mathias, Alan F Scott, et al. "Reverse transcriptase encoded by a human transposable element". In: *Science* 254.5039 (1991), p. 1808.

[302] J Pattamadilok et al. "LINE-1 hypomethylation level as a potential prognostic factor for epithelial ovarian cancer". In: *International Journal of Gynecological Cancer* 18.4 (2008), pp. 711–717.

[303] Eunjung Lee et al. "Landscape of somatic retrotransposition in human cancers". In: *Science* 337.6097 (2012), pp. 967–971.

[304] Dustin C Hancks and Haig H Kazazian. "Roles for retrotransposon insertions in human disease". In: *Mobile DNA* 7.1 (2016), p. 9.

[305] Keith D Robertson. "DNA methylation and human disease". In: *Nature reviews. Genetics* 6.8 (2005), p. 597.

[306] Zhifu Sun et al. "Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations". In: *Briefings in bioinformatics* (2016), bbw069.

[307] Thomas A Cooper, Lili Wan, and Gideon Dreyfuss. "RNA and disease". In: *Cell* 136.4 (2009), pp. 777–793.

[308] Lavanya Rishishwar et al. "Population and clinical genetics of human transposable elements in the (post) genomic era". In: *Mobile genetic elements* 7.1 (2017), pp. 1–20.

[309] Katja Lohmann and Christine Klein. "Next generation sequencing and the future of genetic diagnosis". In: *Neurotherapeutics* 11.4 (2014), pp. 699–707.

[310] Jay Shendure et al. "Advanced sequencing technologies: methods and goals". In: *Nature reviews. Genetics* 5.5 (2004), p. 335.

[311] Karen S Frese, Hugo A Katus, and Benjamin Meder. "Next-generation sequencing: from understanding biology to personalized medicine". In: *Biology* 2.1 (2013), pp. 378–398.

[312] R Matthew Ward et al. "Big data challenges and opportunities in high-throughput sequencing". In: *Systems Biomedicine* 1.1 (2013), pp. 29–34.

[313] Yaran Yang, Bingbing Xie, and Jiangwei Yan. "Application of next-generation sequencing technology in forensic science". In: *Genomics, proteomics & bioinformatics* 12.5 (2014), pp. 190–197.

[314] Elaine R Mardis. *The challenges of big data*. 2016.

[315] Nian Liu et al. "Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators". In: *Nature* 553.7687 (2018), p. 228.

Appendix

In-Silico Dissection of L1 retrotransposons in Cellular Senescence

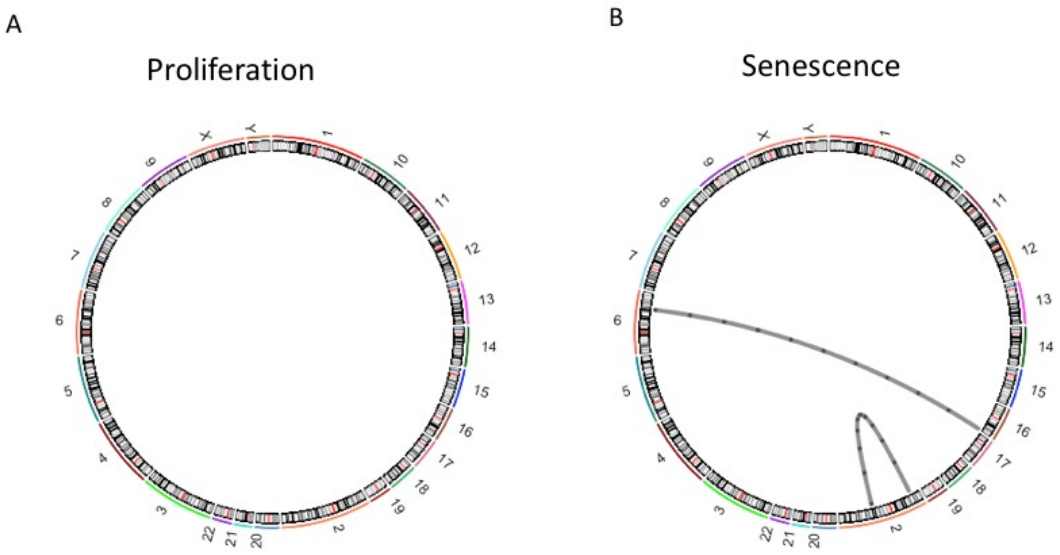


Figure A.1: Chimeric transcript representation obtained from chimeric donor sites for L1 elements capable of jumping. (A) Circos plot illustration of chromosome interactions for proliferative cells exhibited 0 chimeric transcripts. (B) Analytical circos plot representation of chromosome interactions for senescent cells exhibited 2 chimeric transcripts.

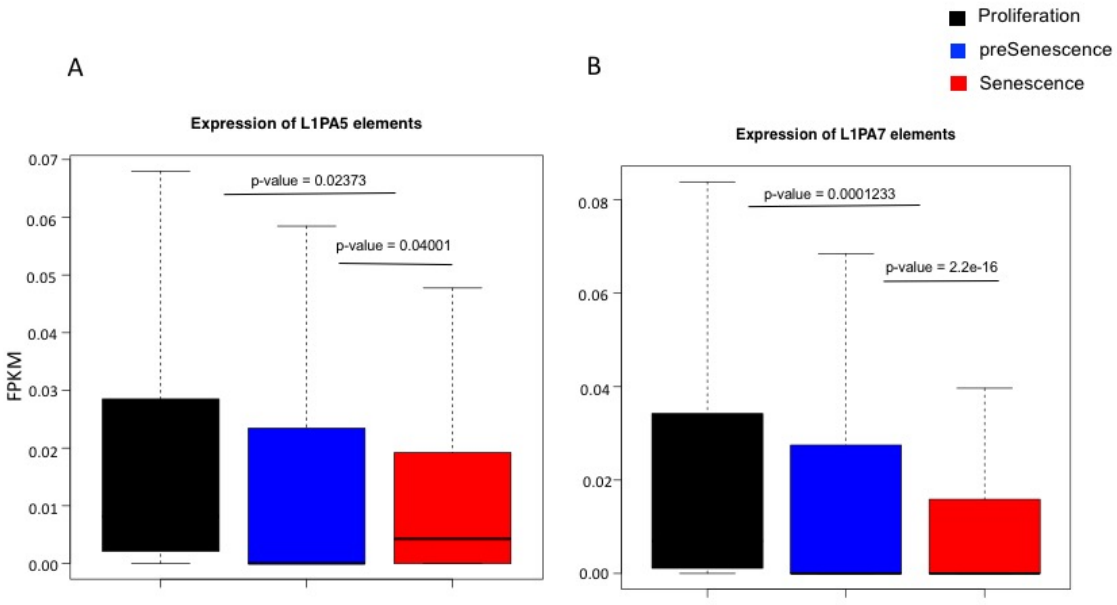


Figure A.2: Expression analysis in diverse families of L1 elements.(A) In L1PA5 subfamily of elements it showed up-regulation of expression in proliferative as compare to other conditions. (B) For L1PA7 subfamily of elements it also displayed up-regulation in proliferative state as compare to other conditions.

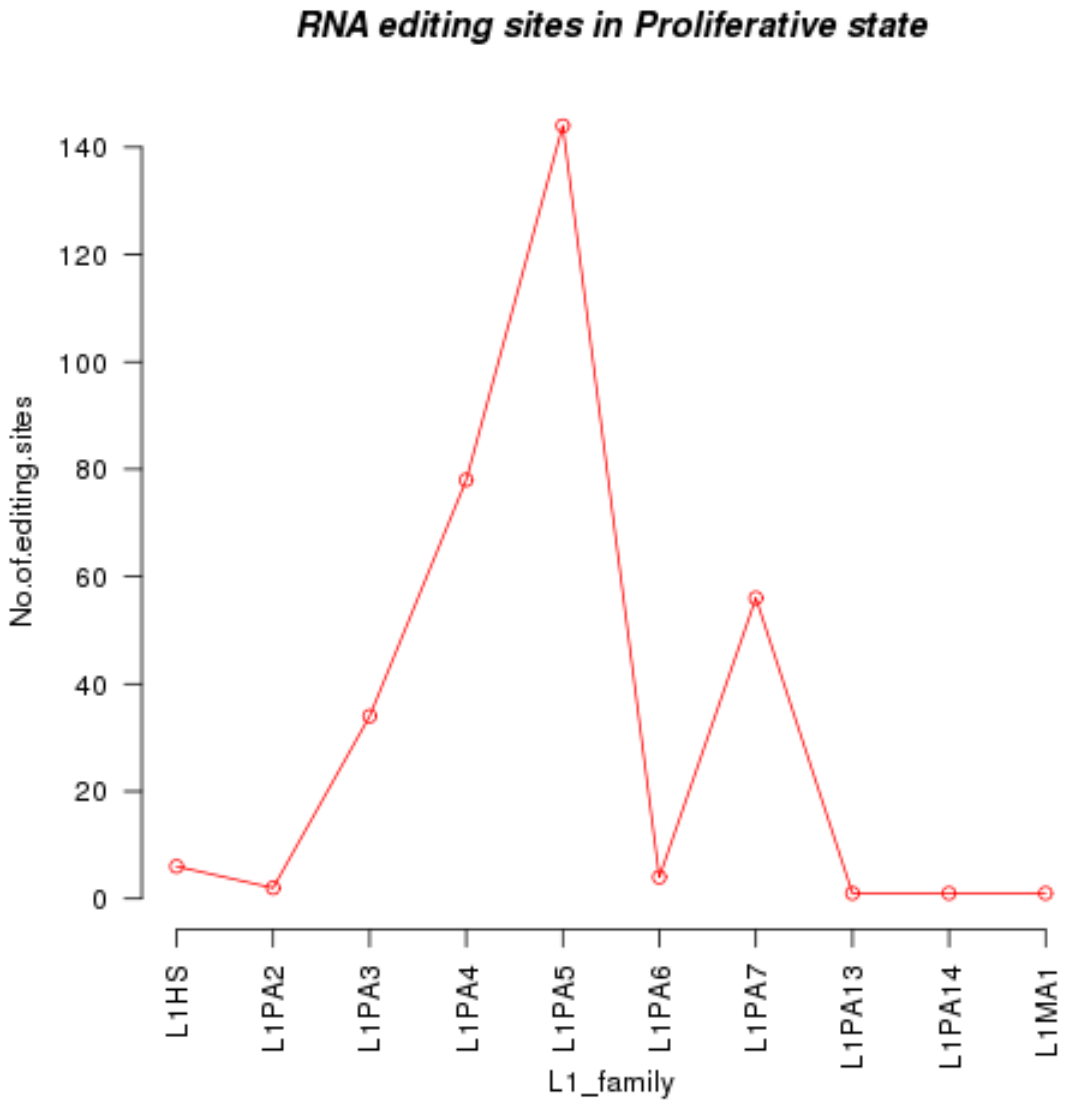


Figure A.3: RNA editing sites distribution in proliferative cells. The quarter of LIHs and L1PA2 refers to youngest L1 family. The second quarter from L1PA3 to L1PA5 refers to middle-aged L1 family. The third quarter from L1PA6 to L1MA1 refers to ancient L1 family.

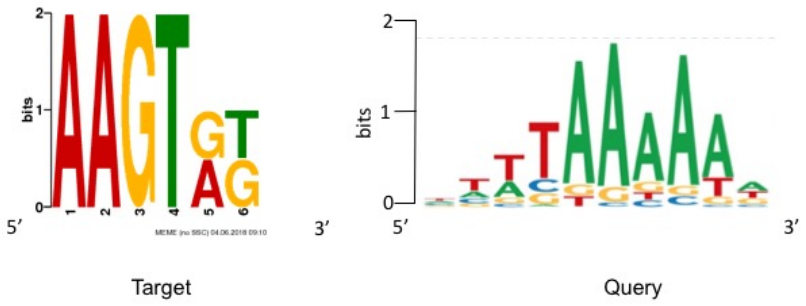


Figure A.4: *De novo* motif analysis for insertional events from Chimeric donor sites. Target represents motif pattern of donor insertional event for senescent cells. While, Query represents consensus sequence of TSDs.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit- einschließlich Tabellen und Abbildungen-, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt.Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Peter Nürnberg betreut worden.

Köln, 2018

Ankit Arora