COMPARATIVE ANALYSIS OF GENOMIC SEQUENCE-DEPENDENT AND SEQUENCE-INDEPENDENT APPROACHES TO IDENTIFY RNA EDITING SITES IN HUMAN PRIMARY MONOCYTES

LEONG WAI MUN

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

COMPARATIVE ANALYSIS OF GENOMIC SEQUENCE-DEPENDENT AND SEQUENCE-INDEPENDENT APPROACHES TO IDENTIFY RNA EDITING SITES IN HUMAN PRIMARY MONOCYTES

LEONG WAI MUN

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Leong Wai Mun

Matric No: SGR 160040

Name of Degree: Master of Science

Title of Thesis: Comparative Analysis of Genomic Sequence-Dependent and Sequence-Independent Approaches to Identify RNA Editing Sites in Human Primary Monocytes

Field of Study: Bioinformatics (Biological data analysis)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

COMPARATIVE ANALYSIS OF GENOMIC SEQUENCE-DEPENDENT AND SEQUENCE-INDEPENDENT APPROACHES TO IDENTIFY RNA EDITING SITES IN HUMAN PRIMARY MONOCYTES

ABSTRACT

RNA editing is an enzyme-mediated transcriptional alteration mechanism in eukaryotic cells that changes the sequences of primary RNA transcripts through nucleotide modification, insertion or deletion which leads to the diversification of gene products. Discovering RNA editing events in terms of frequency and location could provide useful information into their molecular adaptation as well as in determining their biological functions and regulation potentials at the cellular and organismic levels. The advancements of high-throughput sequencing technologies have resulted in the identification of significant numbers of RNA editing sites. Conventionally, both genomic and transcriptomic sequences are required for analysis. Recently, high-depth transcriptome sequencing approach (RNA-Seq) has enabled the identification of editing events without depending on the genomic sequences. In this study, both genomic sequence-dependent and genomic sequence-independent approaches were used to identify RNA editing sites present in human primary monocytes from a healthy individual. This will also allow comparative analysis being conducted on the reliability of both methods in discovering the editing sites. From the analysis, more editing events were detected using the genomic sequence-independent method (based on RNA sequences alone). Discrimination of RNA editing sites from genome-encoded single-nucleotide polymorphism (SNPs) is known to be one of the main challenges in identifying RNA editing sites. When we filtered the putative RNA editing sites identified through genome sequence-independent and sequence-dependent approaches with the Single Nucleotide Polymorphism Database (dbSNP), 71% and 10% of known

SNPs were found, respectively. Hence, suggesting that DNA-Seq information from the same individual may possibly reduce the chances of novel and rare genomic variants (SNPs) being interpreted as RNA editing events. Furthermore, genomic localization and distribution of RNA editing sites in healthy human primary monocytes were profiled. The results obtained showed that majority of the editing sites resided in the non-coding regions. As far as we know, this is the first study that utilized both genomic-dependent and genomic-independent sequence approaches to identify RNA editing sites using high-depth genomic and transcriptomics datasets. The pipelines described in this study would certainly be useful for RNA editing sites identification in other human cells. Moreover, our findings will also serve as a reference for future functional study of specific editing events in healthy human primary monocytes as well as comparative study for disease-states human monocytes.

Keywords: RNA editing, monocytes, transcriptome, whole genome, next generation sequencing

ANALISIS PERBANDINGAN ANTARA KAEDAH UNTUK MENGESAN PENYUNTINGAN RNA DALAM MONOSIT UTAMA MANUSIA MELALUI PENGGUNAAN URUTAN GENOM DAN TANPA MENGGUNA URUTAN GENOM

ABSTRAK

Penyuntingan RNA adalah mekanisme pengubahsuaian enzim yang diperkayakan dalam sel eukariotik yang mengubah turutan transkrip RNA primer melalui pengubahsuaian asas, pemasukan dan penghapusan nukleotida, yang dengan itu membawa kepada kepelbagaian produk gen. Penemuan mekanisma penyuntingan RNA dari segi kekerapan dan lokasi boleh memberikan maklumat berguna di dalam adaptasi molekul serta dalam menentukan potensi fungsi biologi dan peraturan mereka di peringkat selular dan organisma. Perkembangan teknologi penjujukan tahap tinggi telah menyebabkan pengenalpastian jumlah yang besar berkaitan proses penyuntingan RNA. Secara konvensional, kedua-dua turutan genomik dan transkrip diperlukan untuk analisis. Terkini, penggunaan penjujukan transkrip (RNA) yang mendalam (RNA-Seq) telah membolehkan pengenalpastian proses penyuntingan RNA tanpa bergantung kepada turutan genom. Dalam kajian ini, kedua-dua kaedah iaitu kaedah bergantung kepada turutan genom dan kaedah tidak bergantung kepada turutan genom diaplikasikan untuk mengenal pasti peristiwa penyuntingan RNA di dalam monosit utama manusia dari individu yang sihat. Ini juga akan membolehkan analisis perbandingan dijalankan ke atas kebolehpercayaan kedua-dua kaedah dalam menemukan tapak penyuntingan RNA berkenaan. Dari analisis yang telah dijalankan, lebih banyak proses penyuntingan RNA dikesan menggunakan kaedah tanpa bergantung kepada turutan genomik (iaitu berdasarkan turutan RNA sahaja). Cabaran utama dalam mengenal pasti tapak penyuntingan RNA adalah dalam membezakan tapak penyuntingan RNA yang ditemui tersebut dengan polimorfisme nukleotida tunggal (SNP) yang dikodkan oleh genom.

Apabila kami menyaring senarai tapak penyuntingan RNA yang diperolehi menggunakan kaedah tidak bergantung kepada turutan genom dan kaedah bergatung kepada turutan genom dengan Pangkalan Data Polimorfisme Nukleotida Tunggal (dbSNP), 71% dan 10% daripada SNP yang diketahui telah dijumpai. Oleh itu, kami mencadangkan bahawa maklumat turutan genom dari individu yang sama mungkin dapat mengurangkan penemuan turutan novel dan turutan yang jarang berlaku (SNP) daripada ditafsirkan sebagai peristiwa penyuntingan RNA. Selain dari itu, genom lokalisasi oleh penyuntingan RNA di dalam monosit utama manusia yang sihat juga diprofilkan. Hasilnya menunjukkan bahawa majoriti tapak penyuntingan RNA berada di kawasan bukan pengkodan. Berdasarkan pengetahuan kami, ini adalah kajian pertama yang menggunakan kedua-dua kaedah iaitu kaedah yang bergantung kepada turutan genomik dan kaedah tidak bergantung kepafa turutan genomik untuk mengenalpasti tapak penyuntingan RNA dengan menggunakan data genomik dan transkrip (RNA) yang mendalam. Aliran analisis yang dikemukakan dalam kajian ini pastinya berguna untuk pengenalpastian tapak penyuntingan RNA dalam sel-sel manusia yang lain. Selain itu, penemuan kami juga boleh menjadi suatu rujukan untuk masa depan peristiwa penyuntingan tertentu dalam kajian monosit utama manusia yang sihat serta kajian komparatif untuk penyakit-keadaan monosit manusia.

Kata kunci: Penyuntingan RNA, monosit, transkriptomik, genomik, penjujukan generasi seterusnya

ACKNOWLEDGEMENTS

First of all, my sincere gratitude to the Ministry of Education Fundamental Research Grant Scheme (FRGS) for supporting the project as well as University of Malaya and Institute of Biological and CRYSTAL for the facilities.

My sincere gratitude to my supervisors, Prof Dr Hj Amir Feisal Merican bin Hj Aljunid Merican and Dr Saharuddin Mohamad, I thank them for providing their invaluable guidance, comments and suggestions throughout the course of my master project. Most importantly, I would like to thank my beloved family for their patience and unconditional supports, my friends Dr. Hoda and Hamidah who have given me a lot of moral supports and advice during my research journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

A	:	Adenosine
ADAR	:	Adenosine deaminases acting on RNA
ALS	:	Amyotrophic lateral sclerosis
ApoB	:	Apolipoprotein B
APOBEC	:	Apolipoprotein B mRNA editing catalytic polypeptide-like
BAM	:	Binary alignment map
b	:	BLAT
Ca^{2+}	:	Calcium ion
CD14	:	Cluster of differentiation 14
CD16	:	Cluster of differentiation 16
cDNA	:	Complementary DNA
C	:	Cytosine
DARNED	:	DAtabase of RNa EDiting in humans
DNA	:	Deoxyribonucleic acid
dNTP	:	Deoxyribonucleotide
ddNTP	:	Dideoxyribonucleotide
DNA-Seq	:	DNA Sequencing
dsRNA	:	Double stranded RNA
dsRBD	:	Double stranded RNA binding domain
N	:	Frequency of variation in DNA
n	:	Frequency of variation in RNA-Seq
GATK	:	Genome analysis toolkits
GB	:	Gigabyte
GluR-B	:	Glutamate receptor 2

Q	:	Glutamine
G	:	Guanine
HSC	:	Hematopoietic stem cell
HTS	:	High throughput sequencing
kDa	:	Kilodalton
LINE	:	Long interspersed elements
lncRNA	:	Long non-coding RNA
Mb	:	Mega-basepairs
miRNA	:	MicroRNA
M	:	Millions
c	:	Minimal coverage
m	:	Minimum mapping quality
q	:	Minimum quality score
NK cell	:	Natural killer cell
NGS	:	Next generation sequencing
o	:	Output in REDItools
pre-miRNA	:	Precursor mRNA
pre-mRNA	:	Precursor mRNA
pri-mRNA	:	Primary miRNA
V	:	Reads that are not supported by DNA-Seq reads
L	:	Remove substitution in homopolymeric region of DNA-Seq
l	:	Remove substitution in homopolymeric region of RNA-Seq
RNA	:	Ribonucleic acid
RADAR	:	Rigorously annotated database of A-to-I RNA editing
RNAi	:	RNA interference
RNA-Seq	:	RNA Sequencing

SAM	:	Sequence alignment map
SINEs	:	Short interspersed elements
SMS	:	Single molecule sequencing
SNP	:	Single nucleotide polymorphism
SMRT	:	Single nucleotide real time
r	:	Substitution located within known splice junction
TB	:	Terabyte
T	:	Thymine
a	:	Truncated reads
UTR	:	Untranslated region
v	:	Variant bases
WES	:	Whole exome sequencing
WGS	:	Whole genome sequencing
WTSS	:	Whole transcriptomic shotgun sequencing
ZDD	:	Zinc-dependent cytidine or deoxycytidine deaminase domain

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Overview

Genetic contents or information of living organism are stored within deoxyribonucleic acid (DNA). Central dogma of molecular biology described the transfer and transforms action of the genetic content from DNA to RNA which will then transcript and transcribe into various gene products (Crick, 1970). Different types of cellular or gene expression give rise to different organism complexity including their morphology and behavior. However, an organism's complexity is not directly proportional to its genome size (Markov et al., 2010). For example, genome of a salamander *Ambystoma mexicanum* (the Mexican axolot) is ten times larger than human genome, yet the organism are not ten times more complex than humans (Keinath et al., 2015). According to data collected by Smith and colleagues, the team estimated the *A. mexicanum* genome is approximately 32 GB while a human genome size is approximately 6,469.66 MB in total (diploid) (Keinath et al., 2015). The expansion of organism genetic information is believed to be caused by post-transcriptional modification processes including RNA splicing, alternative splicing and editing (Nilsen & Graveley, 2010; Nishikura, 2010).

Splicing is a type of post-transcriptional process where introns from pre-mRNA are spliced out and exon are combined to form a mature mRNA. Disregarding the rule of one gene code for one polypeptide, alternative splicing is an event that allows introns to be spliced out and exon were selectively combined under different condition to produce multiple protein isoforms (Nilsen & Graveley, 2010). In contrast, RNA editing increases the genetic complexity through single nucleotide modification. It is an enzyme-mediated post- or co- transcriptional alteration process in RNA sequence without affecting the encoding DNA sequence. Various types of transcripts can be affected

through RNA editing such as mRNAs, intron RNA, exon RNA, structural RNA and regulatory RNA (Moreira et al., 2016).

In human, there are two types of well-studied RNA editing: adenosine-to-inosine (A-to-I) editing catalyzed by Adenosine Deaminase Acting on RNA (ADAR) (Bass, 1997) and cytidine-to-uridine (C-to-U) editing catalyzed by APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide) (Chester et al., 2000). These editing events are also known as canonical RNA editing events. Between these, A-to-I editing was found to be present abundantly in human compared to C-to-U editing. The first example of A-to-I editing in mammalian mRNAs was identified in transcripts encoding the GluR-2 subunit of the AMPA receptor in which a genomically-encoded glutamine codon (CAG) was altered to an arginine codon (CIG) (Melcher et al. 1995)). To date, three ADAR proteins that have been identified namely ADAR1, ADAR2 and ADAR3. These ADARs enzyme binds to double-stranded RNAs (dsRNAs) and deaminate adenosine to inosine. Inosine prefers to base pairs with cytidine, therefore, it is functionally equivalent and interpreted as guanine (G) during translation. On the contrary, C-to-U editing has been shown to be present abundantly in the plant mitochondria of higher plants (Yu et al., 1995) but were relatively less common in humans (Hamilton et al., 2010). The first report of C-to-U editing in vertebrates was the editing of mRNA encoding apolipoprotein B (apoB). Unlike ADARs, cytidine deaminase family of proteins was shown to catalyze the editing process in both RNA and DNA substrates (Conticello, 2008). The protein exists in two forms (apoB100 and apoB48) produced from a single gene and plays a key role in lipid metabolism (Teng et al., 1993). Knockout of gene encoding ADAR enzymes have shown to cause death in mice (Higuchi et al., 2000). Furthermore, disruption of RNA editing was shown to link to neurodegenerative diseases such as brain ischemia and epilepsy (Maas et al., 2006),

immune-related disorders (Mannion et al., 2014) and various human cancers (Chan et al., 2014; Han et al., 2015; Fumagalli et al., 2015).

The canonical editing events especially A-to-I editing have been catalogued from a variety of tissues mainly lung, brain, muscle, liver, kidney and heart, and appropriate databases composed of millions of annotated sites were constructed (Picardi et al., 2015). The comprehensive catalogue of A-to-I editing suggested that the occurrence of RNA editing is strongly tissue-specific (Picardi et al., 2015). Despite the advancement in RNA editing site detection in various tissues, cell level RNA sequence alteration such as RNA editing in the cells of monocytic lineage remained poorly-understood. In 2015, Sharma et al., reported that C-to-U editing was facilitated by APOBEC3A gene in macrophages as well as monocytes. However, there is little information about the A-to-I editing in monocytes, of which if present, may potentially contributes to transcriptome and proteome complexity.

In recent years, high-throughput sequencing (next-generation sequencing, NGS) has immensely aid in identification of cellular RNA editing sites. Examples of application of NGS are genomic sequencing (DNA Sequencing, DNA-Seq) and transcriptomic sequencing (RNA Sequencing, RNA-Seq). DNA-Seq is a laboratory process of obtaining sequence and arrangement of nucleotides in DNA molecules. Variations of genome sequencing includes whole genome sequencing (WGS) which involves the nucleotide order determination of the complete DNA sequence of an organism at a single time and whole exome sequencing (WES) which only involve protein-coding genes in a genome (1% of human whole genome). RNA-Seq or whole transcriptomic shotgun sequencing (WTSS) is a laboratory process of determining the nucleotide order of total RNA, which include coding mRNA and non-coding RNA in a biological sample at a given moment (Wang et al., 2009). To date, many variations of analytical

pipelines and software have been published to analyze RNA editing sites. Altogether, the identification methods can be generalized into two main categories. The first category is genome sequence-dependent approach, where the RNA-Seq data of an individual is compared with its corresponding DNA-Seq data (Ju et al., 2011; Picardi et al., 2015). Second category is genome sequence-independent approach in which RNA-Seq data alone is used to determine the RNA editing sites (Bahn et al., 2012; Peng et al., 2012; Ramaswami et al., 2013; Zhang et al., 2015). Pandey and colleagues applied high-throughput methods into a multi-omics analysis to study the landscape of genome, epigenome, transcriptome and proteome of naïve CD4+ T cells from a single individual. The study was a first attempt to identify RNA editing sites from a single, purified primary cell type under normal physiological condition using genome sequence-dependent approach (Mitchell et al., 2015). Motivated by the research, we performed in-depth bioinformatics analysis to identify and characterize the RNA editing sites in healthy human primary monocytes isolated from a healthy individual. As far as we know, the study herein is the first to report the number, distribution and genomic localization of A-to-I editing sites in human primary monocytes, although the C-to-U editing sites in monocytes had previously been reported (Sharma et al., 2015; Rayon-Estrada et al., 2017).

High sequencing depth (ie, >1000 reads per target) has shown to increase the number of predicted RNA editing sites and accuracy of RNA editing identification (Lee et al., 2013; Bahn et al., 2012). In this report, we describe the computational framework of RNA editing sites identification using genome sequence-dependent approach. Additionally, deep transcriptomic sequencing data also allowed us to compare and contrast the sensitivity and specificity between the genome sequence-dependent and independent approaches in identifying RNA editing sites. This study provides important insights into the genomic localization and distribution of A-to-I RNA editing sites in

healthy human primary monocytes. The findings will serve as a good reference for future functional A-to-I editing sites study and also comparative study for disease-state human primary monocytes.

1.2 Objectives

1. To identify and characterize A-to-I editing sites in genomic and transcriptomic sequences of healthy human primary monocytes using genome sequence-dependent.
2. To compare and contrast sensitivity and specificity of genomic sequence-dependent and sequence-independent approaches in identifying RNA editing sites in human primary monocytes.

1.3 Organization

This thesis comprises of six chapters, which are: chapter 1, Introduction, chapter 2, Literature review, chapter 3, Materials and methodology, chapter 4, Results, chapter 5, discussion and chapter 6, summary. The first chapter describes the overview of the research performed and the objectives of this study. Second chapter contains literature review of the entities related to the study, namely human immune system and human primary monocytes, high-throughput sequencing technologies, which includes first and second generation sequencing, RNA sequencing, high-depth NGS data analysis using bioinformatics, and RNA editing, discussed on A-to-I as well as C-to-U editing in detail. Chapter three, the materials and methodology chapter describes the software, hardware, parameters and research pipeline adopted in this study. Chapter four presents the results of this study and the findings are further discussed in chapter five, discussion. The last chapter summarizes the outcome of this study.

CHAPTER 2: LITERATURE REVIEW

2.1 The Human Immune System and Human Primary Monocytes

Human immune system is mediated by different cell types and proteins. All the cellular elements of blood, including red blood cell, platelets and white blood cells are derived from hematopoietic stem cells (HSCs) in bone marrow through the process of haematopoiesis. HSCs give rise to two types of monocytic lineages which are myeloid lineage and lymphoid lineages. The myeloid cells consist of monocytes, macrophages, neutrophils, basophils, eosinophils, erythrocytes, dendritic cells and platelets while the lymphoid cells consist of T cells, B cells and natural killer (NK) cells (Figure 2.1) (Forsbeg et al., 2006). Each cell performs a specific mechanism aimed at recognizing and/or reacting against foreign material and infection.

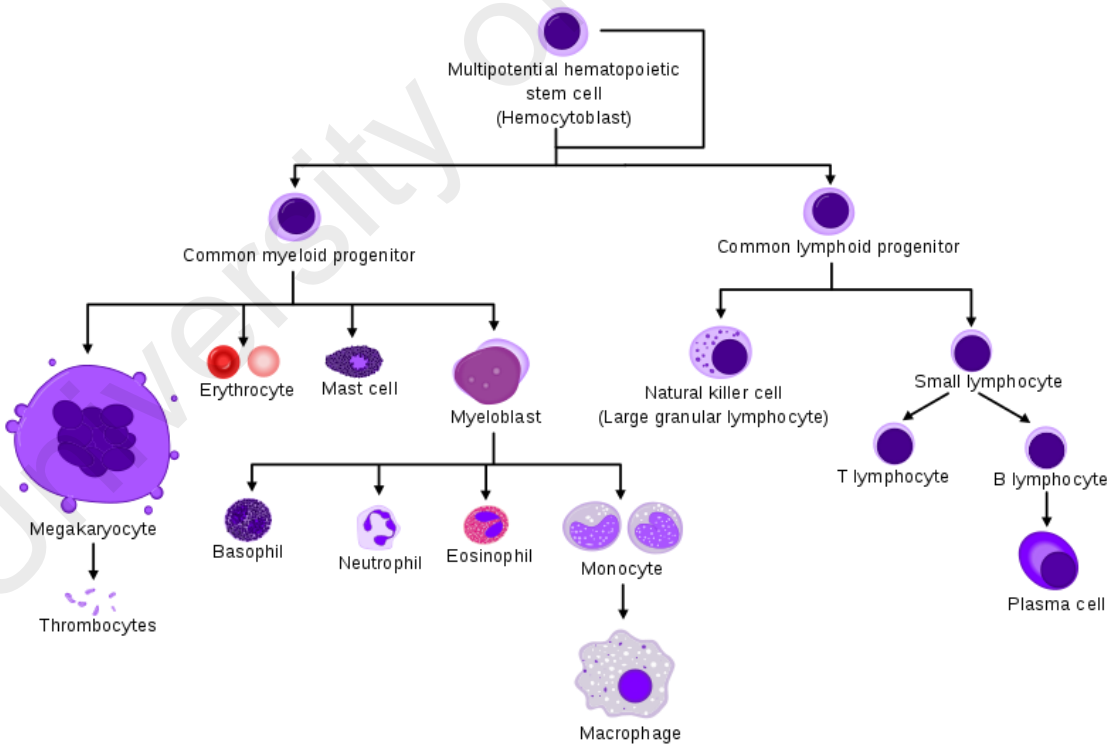

Figure 2.1: Lineage of human hematopoietic stem cells (HSCs). Adapted from A. Rad, 2009)

The human immune system is categorized into two types of system. The first system is adaptive immune system. Adaptive immune response can be further classified into two classes. The first class is the antibody response which is carried out by B cells. During the invasion of a pathogen, B cells will be activated to release antibodies to prevent the pathogen from binding to the host cells. This reaction will deactivate the foreign molecules as well as marking the invading pathogen for easier phagocytosis. Phagocytosis is a process of engulfing the invading pathogen, microbes or foreign particles followed by digestion of the engulfed materials. The second class of adaptive immune response is a cell-mediated immune response. In this response, T cells react directly against the cells that have foreign antigen on its cell surface. It will then signal macrophage to digest the phagocytosed invading pathogen. The second system of human immune system is innate immune response. This system is known to be the first line of human body's defense against foreign particles, including bacteria and viruses. The response is activated by pathogen infected cells and produces cytokines and inflammatory mediators through transcriptional and post-transcriptional mechanisms.

Adaptive immune response is activated to eliminate the pathogens (Janeway et al., 2001). Monocyte plays a significant role in responding to the pathogen in innate immune system. It is the largest leukocyte and classified into three subgroups: classical monocyte, non-classical monocytes and intermediate monocytes. These subgroups are varied in through respective chemokine receptor expression, tissue distribution and phagocytic activity. Classical monocyte is characterized by its higher level of CD14 cell surface receptor ($CD14^{++}$ $CD16^{-}$ monocyte). Non-classical monocyte is classified by its lower level of CD14 and additional co-expression of CD16 receptor ($CD14^{+}CD16^{++}$ monocyte) while intermediate monocyte is classified by its higher level expression of CD14 and lower expression of CD16 ($CD14^{++}CD16^{+}$ monocyte). A few hours after their production from bone marrow into the blood, monocyte migrates into other tissues,

such as spleen, liver, lungs and bone marrow tissues. Under inflammatory conditions, it will move into peripheral tissues and further differentiate into macrophages, which function to digest cellular debris, apoptotic cells and foreign substances. Monocytes will also differentiate into myeloid lineage dendritic cells which act as the messengers in human immune system, as well as functions to activate the naïve T cells (Mirsafian et al., 2016; Saha et al., 2011).

2.2 High-throughput sequencing technologies

2.2.1 First generation sequencing

DNA was first come upon in the late 1860s by Friedrich Miescher, a chemist from the Swiss. In the twentieth century, a Russian biochemist Phoebus Leven discovered that DNA is made up of three major components (phosphate-sugar-base) and the carbohydrate component of RNA (ribose) or DNA (deoxyribose). In 1953, James Watson and Francis Crick have proposed a three-dimensional, double-helical model of DNA structure with complementary bases held together as a pair by hydrogen bond. Although they managed to solve the structure of DNA, however, the ability to "read" or determine the order of nucleotides in biological samples was yet to be discovered. In 1970s, the first generation of DNA sequencing, Sanger sequencing was introduced (Sanger & Coulson, 1975). Sequencing is a technology that can make known the order of nucleotides within DNA or RNA molecules. Sanger sequencing or chain termination method perform sequencing by selectively incorporating chain-terminating dideoxynucleotides (ddNTPs). This was the first technique being widely adopted; hence it was known as the first-generation DNA sequencing. To perform Sanger sequencing, the DNA sample is divided into four separate containers that contain DNA polymerase, natural deoxynucleotides (dNTPs) (dATPs, dGTPs, dCTPs and dTTPs) and single type of ddNTPs (ddATPs, ddGTPs, ddCTPs or ddTTPs). By chance, the sequence elongation process is terminated due to the attachment of ddNTPs, generating sequences of

different sizes. After a few cycles, the end product will be loaded in to four different lanes and the sequences are separated through electrophoresis (Figure 2.2). In the following years, Sanger sequencing was enhanced of which fluorescent labels ddNTPs were used, allowing the reaction to be carried in a single container and the detection was improved using capillary-based electrophoresis (Ansorge et al., 1987, Kambara et al., 1988, Luckey et al., 1990).

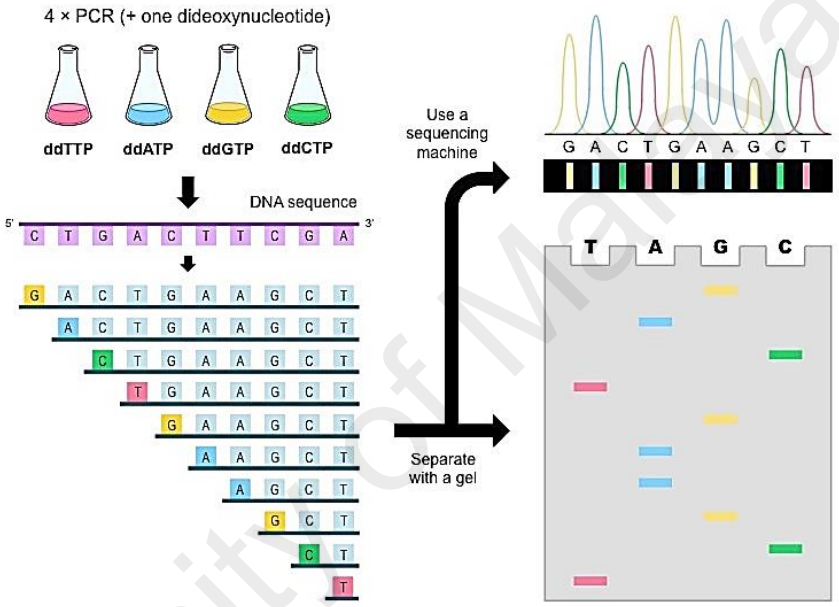

Figure 2.2: First generation sequencing or Sanger sequencing. Adapted from Gaurab Karki, 2017

2.2.2 Next-generation sequencing (NGS) technologies

Concurrent with the growth of fluorescence-based sequencing methods and DNA sequencers, which produced high-resolution image, another technique, pyrosequencing method was proposed. Instead of fluorescence emitted by ddNTPs, the alternate approach focused on the chemiluminescent detection of pyrophosphate released during incorporation of natural dNTPs (Nyrén & Lundin, 1985). Pyrosequencing approach was then licensed to a biotechnology company that was founded by Jonathan Rothberg, 454 Life Sciences where in 2005, it was successfully commercialized as second generation DNA-Seq or NGS (Voelkerding et al., 2009). There are a few NGS platforms such as

Illumina, Ion Torrent and Pacific Biosciences that are different in configurations and sequencing chemistry. However, they share a similar concept in which they perform massive parallel sequencing via amplification of DNA templates on a flow cell. Illumina platform is known to dominate the HTS market (Reuter et al., 2015). In Illumina sequencing, the DNA sample is first fragmented into fragments of 100 – 150 reads. The fragmented reads are ligated to adapter oligonucleotides and annealed on the flow cell as single-molecule DNA template. The DNA templates will then be amplified in a bridge structure on the flow cell through polymerase chain reaction (PCR), forming colonies of sequences. For sequencing, the bridge is dissociated from one end and the sequencing is initiated by DNA primers. The flow cell is flooded with DNA polymerase and 4 differently colored fluorescent reversible dye terminators/bases (dNTPs). Once a base has been attached to the strand, the sequencing processes stop, the florescence will be captured and recorded by the machine. The fluorescently-labelled terminator group will be removed from the bases and sequencing process continues (Figure 2.3) (Voelkerding et al., 2009). Illumina platform sequencing analyzed DNA sequence in base-by-base manner, which has made it a highly accurate and popularly used method.

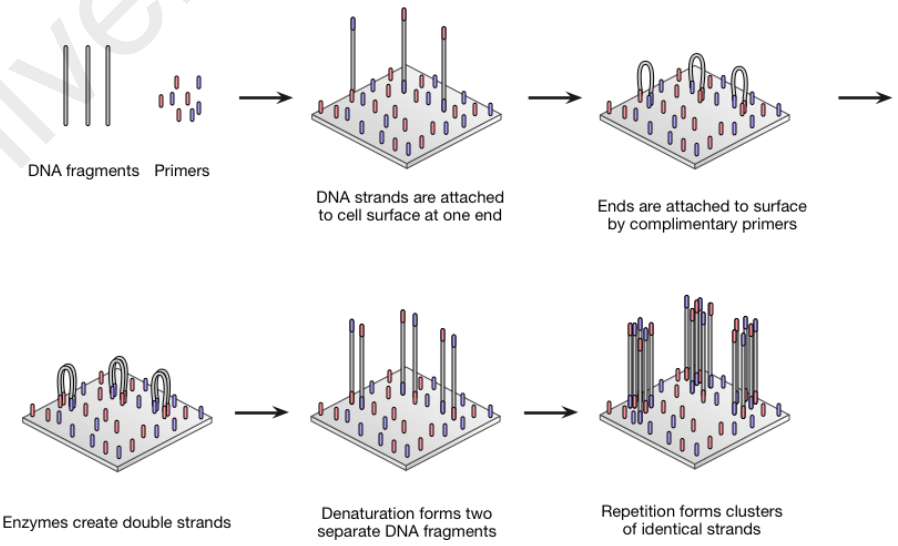

Figure 2.3: Second generation or next generation sequencing (NGS) on Illumina platform. Adapted from https://www.atdbio.com/content/58/Next-generation-sequencing.

Despite the rapid growth and acceptance of NGS, the newest generation of sequencing, single-molecule sequencing (SMS) or third generation sequencing is emerging. This platform also known as single molecule real time (SMRT) sequencing was introduced by Pacific Biosciences (Van et al., 2014). SMRT incorporates special chemicals which enable sequencing to be performed without PCR. It is also incredibly sensitive; enabling the single molecules to be sequenced in a very short amount of time (Heather & Chain, 2016).

2.2.3 RNA sequencing

The ability of next-generation sequencing methods to provide a huge amount of sequence information at a lower cost has enabled the development of whole genome sequencing, as well as whole transcriptomic shotgun sequencing (WTSS). WTSS or RNA sequencing (RNA-Seq) uses NGS to study the transcriptome of an organism. Transcriptome is the complete set of transcripts in an organism, which codes for proteins expressed at various developmental stages or physiological conditions. The RNA-Seq is a sequencing method that allows researchers to identify and quantify the transcripts that are expressed in a biological sample at a given time (Sims et al., 2014). Following RNA sample extraction, the selection of RNA is usually completed by using: oligo-dT beads which will extract the poly-Adenylated RNAs (good representation of mRNAs) or ribodepletion which ribonucleases is used to digest ribosomal RNA (rRNA) and allow the extract of total RNA (O'Neil et al., 2013). Due to lower cost, polyA-selection is the most popular choice of RNA sample selection. The extracted RNA is then converted into complementary DNA (cDNA) fragments by using random hexamer primers. Next, adapter sequences are ligated to both end of the cDNA fragment to allow hybridization of fragments into the flow cell and serve as primers to initiate the sequencing reaction. The cDNA is then sequenced by use of any high-throughput sequencing technologies. The resulting reads, namely classified into three types of reads

which are exonic reads, junction reads (Intronic reads) and polyA end-reads (Figure 2.4). Therefore, compared to DNA-Seq, RNA-Seq has several significant advantages in studying the organismal complexity: (1) not limited to transcripts that only exist is genomic sequence (Vera et al., 2008), (2) reveals precise location of exon-exon boundaries and sequence variation in the transcribed region (Morin et al., 2008) and (3) highly accurate for quantifying expression level (Nagalakshmi et al., 2008).

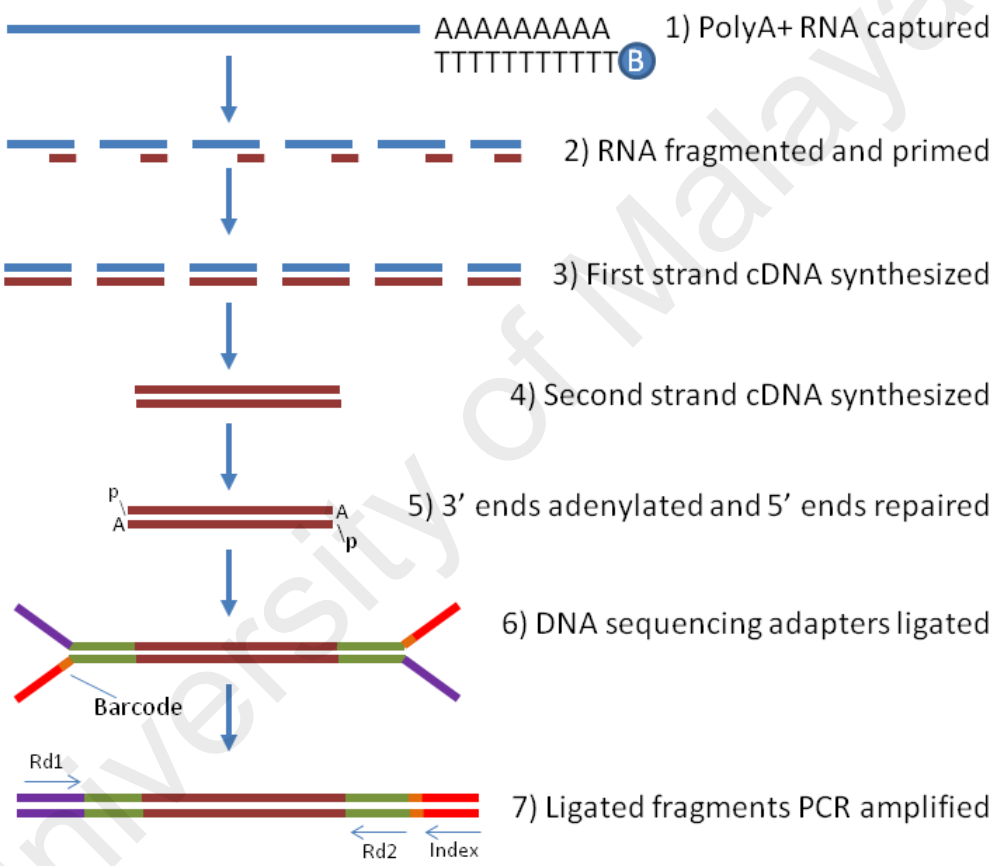

Figure 2.4: RNA sequencing with selection of poly-Adenylated RNAs. Adapted from David et al., 2013.

2.2.4 High-depth NGS data analysis using bioinformatics

Generally, sequencing coverage means the average number of short reads that are aligned and covered on the reference genome during sequencing (Figure 2.5). Sequencing depth means the average number of times a particular nucleotide is sequenced at a particular position. Redundancy of coverage is also called as depth or the depth of coverage (Sims et al., 2014) (Figure 2.5). In NGS, higher level of coverage and depth indicate the higher degree of confidence during variant discovery (Haiminen et al., 2011). Studies also showed that the ability to detect rare single nucleotide polymorphism (SNP) or RNA editing events in lower frequencies relies on sufficient sequencing depth (Chen, 2013; Ramaswami et al., 2013; Huntley et al., 2016).

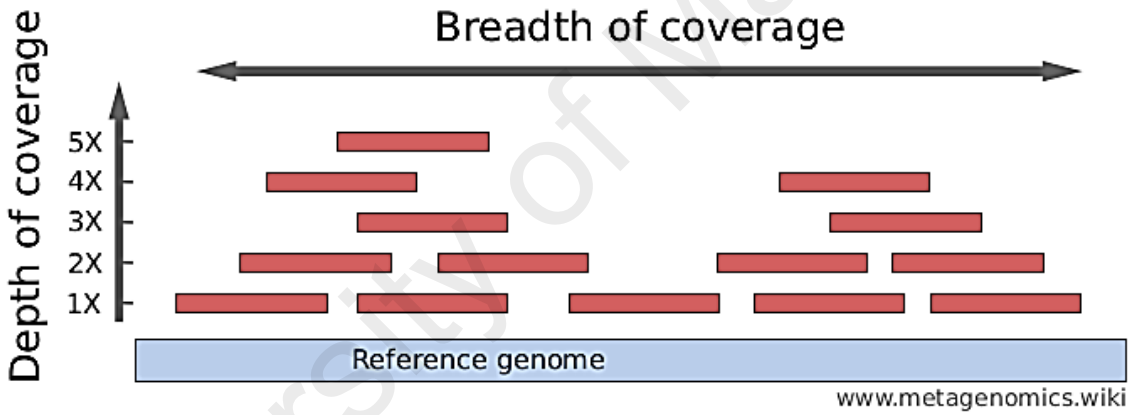

Figure 2.5: Sequencing read depth and coverage. Adapted from http://www.metagenomics.wiki/pdf/definition/coverage-read-depth

In DNA-Seq, the term "coverage" is used to describe the average times a nucleotide is being sequenced. While in RNA-Seq, the term "million reads" is used instead. Determining the coverage needed for a RNA-Seq is difficult because different transcripts are expressed at different levels. More reads will be captured from highly expressed genes while lesser reads will be captured by genes expressed at low levels. Therefore, million reads is used to describe the total amount of reads required at the end of the sequencing. To date, there was no specific guideline or scale to classify the depth of coverage of NGS data as a suitable sequencing depth and coverage is dependent on

the objectives of the study and budget. In 2015, Griffith et al. (2015) performed an optimizing cancer genome sequencing analysis and recommended to sequence whole genome sequencing to a depth of 200x to 300x and RNA-Seq to 250M reads to 300M reads for better identification of single nucleotide variants. Moreover, previous study has shown that more genes and transcripts were being identified from the transcriptomic data with sequencing depth of 200M reads compared to 100M and 50M reads (Mirsafian et al., 2017). For the identification of variants such as SNPs and RNA editing sites through NGS data analysis, sequence alignment is a crucial step. Mirsafian et al. (2017) reported that increase in sequencing depth has no effect on the sequence alignment.

Conventionally, analysis of RNA-Seq is more robust with the integration with DNA-Seq data (Griffith et al., 2015). The availability of high-depth sequencing datasets followed by rigorous computational analysis has enabled and expanded the detection of high-confidence canonical RNA editing events without the incorporation of genomic sequence. Advancement in genomic and transcriptomic sequencing has indeed revolutionized the field of bioinformatics. Bioinformatics is important in NGS workflow in overcoming the rising challenge of storage, analysis and interpretation of NGS data (Land et al., 2015). The sequencing signal generated by the manufacturer's sequencing instruments through different technologies as discussed are converted into nucleotide bases of short read data with base quality score in FASTQ format. As different bioinformatics computational software support different types of format, the FASTQ format will usually undergo format conversion for downstream bioinformatics analysis (Kulski, 2015).

2.3 RNA Editing

Flow of information from DNA to RNA leading to expression of protein and gene is described as central dogma of molecular biology (Crick, 1958). Central dogma explains a two-step process: transcription and translation. The process by which a fraction of DNA sequence in nucleus is transcribed into RNA by RNA polymerase is known as transcription. The transcriptional process follows the Watson-Crick base pairing rule and the transcribed RNA is the reverse-complement of the original DNA sequence. Transcription of gene results in the generation of precursor messenger RNA (pre-mRNA). The pre-mRNA contains exons and introns which are also known as coding and non-coding regions, respectively. Through splicing machinery, mature mRNA is produced. Spliceosome remove introns and combine or selectively combine the exons and mature mRNA is then transported to cytoplasm for translation. During the translation process, a mature mRNA is decoded by ribosome to produce a specific combination of amino acids, which then fold into a protein product.

During the post-transcriptional phase, different types of RNA processing takes place to generate the mature mRNA for protein synthesis. Typical forms of these processing include: (1) capping, a process which adds 7-methylguanylate to the 5' end; (2) polyadenylation, a process which adds poly-A tail to the 3' end of a sequence; (3) splicing which removes introns and join exons; and (4) alternative splicing which removes introns and join exons selectively (5) RNA editing which alters the RNA sequence and generates protein diversity. Organismal genetic information is widely expanded through alternative splicing and RNA editing. While alternatively splicing yields a diverse collection of mRNA by selectively combining exons, RNA editing, on the other hand, is an enzyme-mediated post- or co- transcriptional single nucleotide alteration process of the RNA sequence. The single nucleotide alteration has no effects on the encoding DNA sequence. Various types of transcripts can be targeted by RNA

editing enzymes includes mRNAs, intron RNA, exon RNA, structural RNA and regulatory RNA (Moreira et al., 2016).

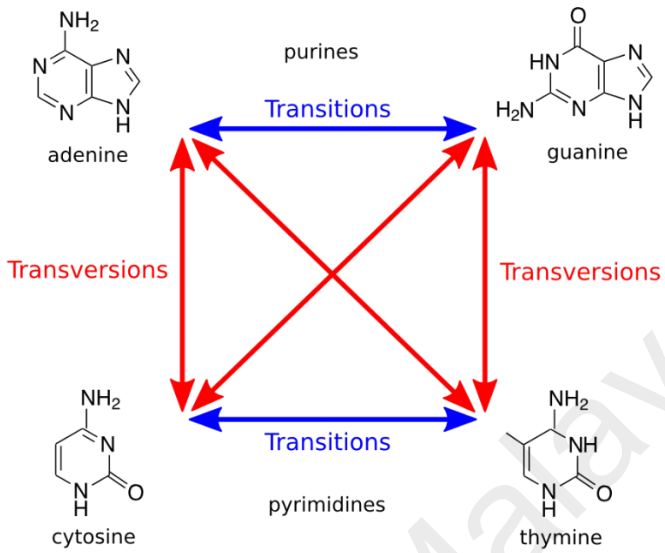

Figure 2.6: Nucleotide substitution matrix. Adapted from https://gtbinf.wordpress.com/biol-41506150/pairwise-sequence-alignment/

According to nucleotide substitution matrix (Figure 2.6), there are a total of 12 possible types of nucleotide substitutions. In humans, only adenosine-to-guanosine (A-to-G) substitution or adenosine-to-inosine (A-to-I) editing and cytidine-to-uridine (C-to-U) substitution were well-characterized. These sites were also known as canonical editing events. All other types of nucleotide changes are non-canonical RNA editing event and not associated with any known enzymatic processes. A study suggested that these non-canonical editing sites may be artifacts or errors resulting from the high-throughput sequencing as there was a lack of validation via Sanger sequencing for these editing sites (Gu et al., 2012). RNA editing events may occur in nucleus, cytosol, mitochondria and plastids of a cell. These events have been observed in eukaryotes ranging from single-celled protozoa to plants as well as mammals. The editing process was first found in the mitochondria of kinetoplastid protozoans (Simpson & Shaw, 1989).

In human, identification of RNA editing sites have been carried at cell-level including peripheral blood mononuclear cells (Chepelev, 2012), brain cell (Picardi et al., 2017) and epithelial cells (Cao et al., 2018). Besides, identification of RNA editing sites has also been extended to different type of cells in other non-human organism such as plant (Tsudzuki et al., 2001), mice (Higuchi et al., 2000), *Drosophila melanogaster* and *C. elegans* (Hoopengardner et al., 2005). To date, millions of RNA editing sites were identified. However, the biological significance of these sites still remains unknown. Scientists suggested that RNA editing may contribute into several general functions, which are described in the following sections of respective canonical RNA editing events.

2.3.1 Adenosine-to-Inosine Editing (A-to-I Editing)

The post-transcriptional modification of eukaryote transcripts has been recognized as one of the important alterations for organismal genetic information expansion. As the most extensive type of RNA editing event, A-to-I editing is a site-specific modification of RNA transcripts. It is catalyzed by the members of ADAR protein family through hydrolytic deamination of C6 position of adenine to inosine (Figure 2.7). Inosine has a preference to base pair with cytidine, therefore, is functionally equivalent and translated as guanine (G) by translational machinery (Daniel et al., 2009).

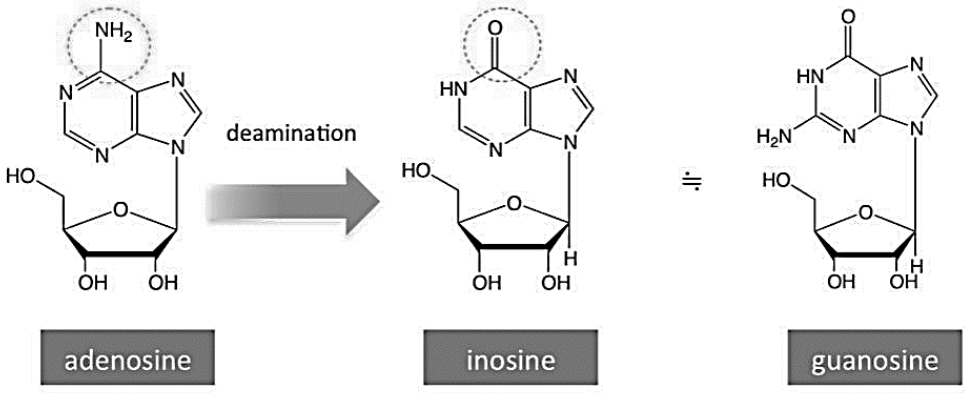

Figure 2.7: Hydrolytic deamination of C6 position of adenine to inosine. Inosine is then recognized as guanine at translation. Adapted from Yang et al., 2005.

There are a total of three types of ADAR enzymes that have been identified in mammals, namely ADAR1, ADAR2 and ADAR3 (Savva et al., 2012). While ADAR1 and ADAR2 are found in most of human tissues, ADAR3 is shown to exclusively present in the human central nervous system (Dominissini et al., 2011). As RNA editing enzyme, all the ADARs contain common domain structures: a double-stranded RNA-binding domain (dsRBD) and a conserved catalytic deaminase domain in the C-terminal region. The dsRBD consists of approximately 65 amino acids and form a highly conserved α-β-β-β-α configuration structure (Nishikura, 2010). Approximately 99% of A-to-I editing is detected in human non-coding RNAs (Athanasiadis et al., 2004). In human, A-to-I editing most frequently targets repetitive RNA sequences located within introns such as Alu elements, long and short interspersed elements (LINE and SINE) and untranslated regions (UTR), which are the 3' and 5' UTRs. With the presence of the highly conserved dsRBDs, A-to-I editing was found pervasive in Alu elements (Picardi et al., 2015, Bazak et al., 2016). This abundance is because of the double stranded RNA (dsRNA) structure formed by the widespread Alu inverted pairs (Daniel et al., 2015) (Figure 2.8). Alu elements, approximately 300 base pairs in length, are one of the SINEs found in all primates. There are approximately 1.4 million copies of Alu present in the human genome, comprising approximately 10% of its size (Lander et al., 2001). Findings suggested that A-to-I editing sites are also present in non-Alu elements such as non-Alu repetitive elements and non-repetitive elements (Li et al., 2009; Ramaswami et al., 2012).

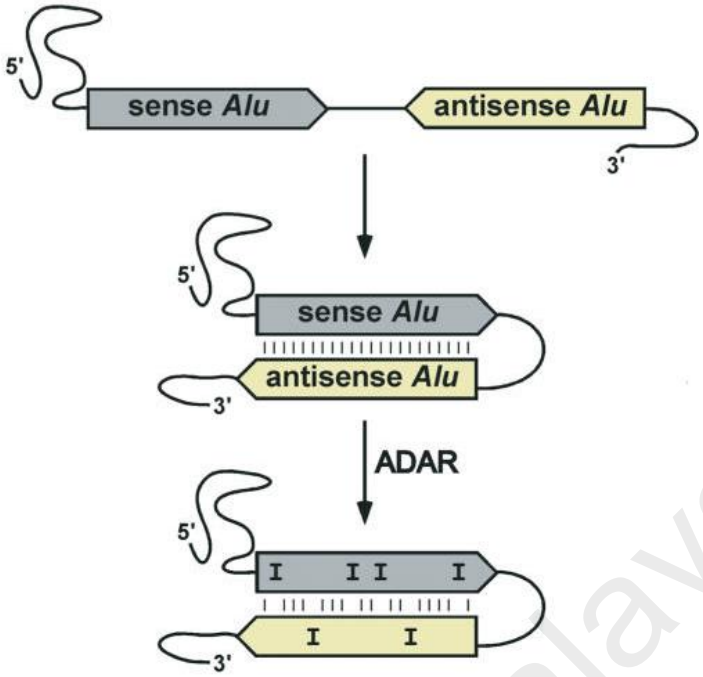

Figure 2.8: The binding of ADAR to the Alu element at intronic region of double-stranded RNA (dsRNA). Adapted from Hasler & Strub et al., 2006.

While numerous numbers of A-to-I editing sites were being discovered, significance and function of A-to-I editing remains relatively unknown. To-date, the most well reported example on the biological significance of A-to-I editing in protein-coding region was the editing in the coding region of glutamate receptor subunit GluR-B (Wright & Vissel 2012). ADAR2-mediated editing in GluR-B changed the gene-encoded glutamine (Q) codon CAG to arginine (R) codon CIG, which produced an ion channel that was impermeable to calcium ion (Ca^{2+}). A subsequent study carried out had shown that ADAR2-null mutant mice resulted in severe deficiency of A-to-I editing. The mice died a few weeks after birth, which due to neuronal death resulting from the excess influx of Ca^{2+} (Higuchi et al., 2000). Other than that, Kawahara et al., (2010) had reported that ADARs regulate the expression of micro RNA (miRNA) as well as redirect silencing targets through A-to-I editing in miRNA. A study also showed that RNA editing was interacting extensively with RNA interference (RNAi) (Bass, 2000) (Figure 2.9).

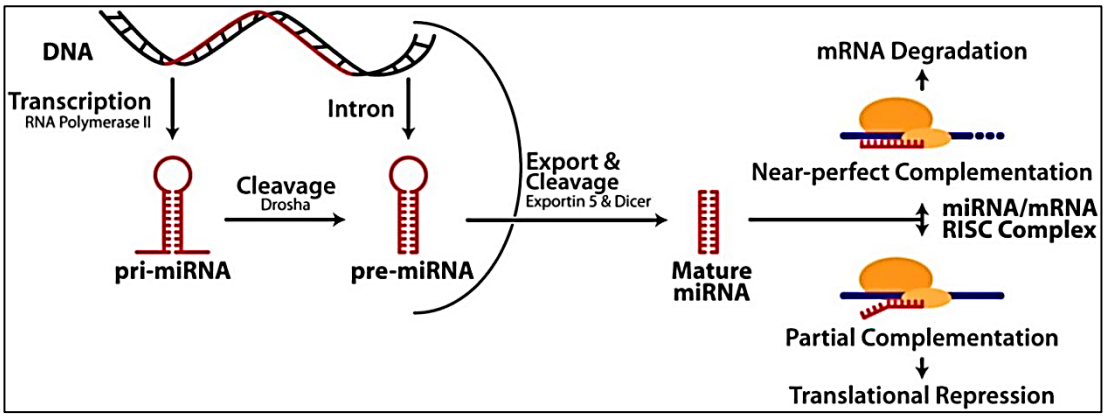

Figure 2.9: MiRNA biogenesis pathway. Adapted from Ryan et al., 2015.

Human miRNA biogenesis consists of two steps which involve nucleus and subsequent cytoplasmic cleavage events catalyzed by 2 ribonucleases III endonucleases, Drosha and Dicer (Figure 2.9). In the miRNA biogenesis pathway, miRNA gene is transcribed to primary miRNA (pri-miRNA). Then, it is processed into precursor miRNA (pre-miRNA) duplex before released as a mature miRNA. The mature miRNA guides the RNAi machinery to their target genes by forming RNA duplexes, resulting in sequence-specific mRNA degradation and translation repression (MacFarlane & Murphy, 2010). The double-stranded structure of pri-miRNA and pre-miRNA allow the binding of ADAR enzymes and lead to A-to-I editing. The editing of pri-miRNA and pre-miRNA may change the base pairing properties of miRNA, suppressing the maturation of miRNA, which regulate the expression of miRNA as well as silence the RNAi machinery (Peng et al., 2012).

Moreover, studies have also shown that dysregulation of RNA editing and abnormal ADAR activity have been linked to various types of human diseases, such as epilepsy, brain ischemia, amytrophic lateral sclerosis (ALS) (Maas et al., 2006), various human cancers (Chan et al., 2014; Han et al., 2015; Fumagalli et al., 2015) and immune-related disorders (Mannion et al., 2014). Recently, RNA editing event identification has been greatly applied in drug discovery. For example, a study show that ADAR2 editing activity exhibits tumor-suppressor capabilities and has been widely observed to be

decreased in astrocytoma tumor tissue (Slotkin et al., 2013). Identification of RNA editing sites in human primary monocyte of healthy subjects is therefore a basic prerequisite to understanding the importance of healthy RNA editing events.

2.3.2 Cytidine-to-Uridine Editing (C-to-U Editing)

Cytidine-to-uridine editing (C-to-U editing), on the other hand, is relatively less common in humans compared to A-to-I editing (Hamilton et al., 2010). It is another type of canonical RNA editing events that has been well-characterized in mammals and was shown to present abundantly in mitochondria and chloroplast of higher plants (Yu et al., 1995). This editing event occurs within highly conserved regions of amino acid sequences of mitochondrial proteins of flowering plants (Zanlungo et al., 1993). The editing event occurred in single-stranded RNA, which catalysed by APOBEC protein. Unlike ADARs, the cytidine deaminase family of protein was showed to catalyze the editing process in both RNA and DNA substrates (Conticello, 2008). The APOBEC family of proteins contain zinc-dependent cytidine or deoxycytidine deaminase domain (ZDD) that is identifiable through its primary amino acid motif. Deaminase activity of this ZDD involves hydrolytic removal of exocyclic amine at C4 position from cytidine to form uridine (Smith et al., 2012) (Figure 2.10).

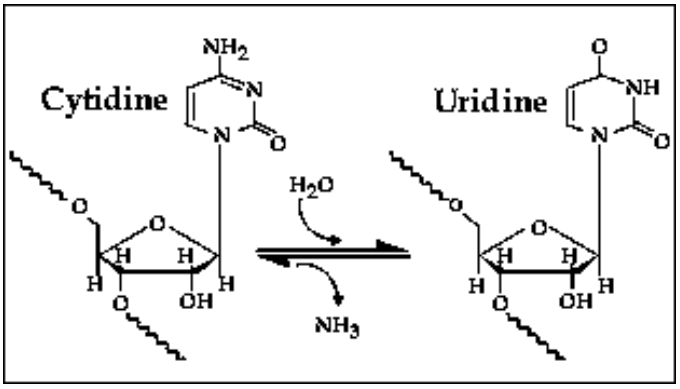

Figure 2.10: Hydrolytic deamination of C4 position in C-to-U editing. Adapted from https://www.nobelprize.org/educational/medicine/dna/a/splicing/rna_editing.html

C-to-U editing is involved in various mechanisms and functions, such as to create but cannot eliminate start and stop codon within a RNA sequence. For example, alteration of codon ACG to AUG will create a start codon which can later alter encoded amino acids and splice site. The editing event was first reported in vertebrates for the mRNA encoding apolipoprotein B (apoB). The protein exists in two forms (apoB100 and apoB48) that produced by the same gene. These proteins play a significant role in lipid metabolism (Teng et al., 1993) (Figure 2.11).

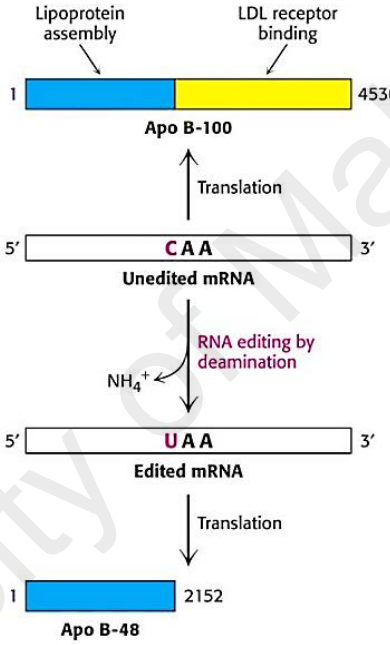

Figure 2.11: C-to-U editing of apolipoprotein B (apoB) mRNA.

Non-edited mRNA encodes a 550 –kDa protein which is the apoB100, which synthesized in the liver and functions in humans to circulate lipoprotein. During C-to-U editing, editosome hydrolytic deaminated single-stranded region of the mRNA which led to the conversion of glutamine codon (CAA) to a termination codon (UAA) at codon 2152. The truncated transcript is protected from the loss of function and encodes a 250 –kDa protein, the ApoB48. The ApoB48 functions to mediate lipid transportation solely in the small intestine (Anant et al., 2001). The ApoB RNA editing in humans is now known to be tissues-specific due to the tissue-specific expression of ApoB editing catalytic subunit (APOBEC).

CHAPTER 3: MATERIALS & METHODOLOGY

3.1 Materials

3.1.1 Transcriptomic and whole genomic dataset

The ethics approval was obtained from the Medical Research and Ethics Committee (MREC) before the study was conducted, Malaysia with the reference number NMRR-13-972-16921 before the study was conducted. The sequencing depth of the raw RNA-Seq data was 340 million (M) reads. In this study, we intended to identify RNA editing sites presented in human primary monocytes. One of the high-depth paired-end human monocyte RNA-Seq dataset from our previous study (accession number GSE80095) was used in this study (Mirsafian et al., 2017). The size of the raw transcriptomic data was 36.8 GB. The sequencing depth of the raw transcriptomic data was 340 million (M) reads. In order to carry out identification of RNA editing sites using genomic sequence-dependent approach, a set of corresponding whole genomic data from the same individual was required. Hence, whole genome sequencing was performed. Process of the whole genome sequencing was detailed in subsection 3.2.2. The sequencing generated a total of 156.35 GB of raw whole genome data. The sequencing depth of the whole genome is 1,042 M reads. The length of transcriptomic and genomic reads was 100 bp and 472 bp long, respectively. Additionally, to verify our computational pipeline, we have also obtained the high-depth paired-end RNA-Seq data of healthy human brain tissues from Gene Expression Omnibus (GEO) database with accession number GSM2745907, GSM2745917, GSM2745934, GSM2745935, GSM2745949 and GSM2745950.

3.1.2 Hardware

To carry out the analysis, Linux OS environment was necessary. In our study, we selected Ubuntu version 16.04 as the Linux distribution for our interface. The memory of our device is 125.8GB with 20 units of 2.30GHz Intel Xeon(R) CPU E5-2650 v3

processors. The operating system was of architecture 64-bit with an internal storage of 5.8TB.

3.1.3 Software

Table 3.1: The list of software used in this study.

No.	Software	Function
1.	FastQC	To identify low quality reads, sequencing biases and adaptors incorporated during library preparation.
2.	Trimmomatic	To trim or eliminate bad quality read and adaptor sequence
3.	Bowtie2	To align sequencing reads of about 50 up to 100s or 1000s bp long to reference genome
4.	HISAT2	To allow mapping of sequencing reads across exon-exon junctions on reference genome
5.	SAMtools	To manipulate the file type of high-throughput sequencing data
6.	Genome Analysis Toolkits (GATK)	Provide a variety of tools which focus on variant detection and genotyping as well as data quality assurance.
7.	Picard tools	Provide accessory tools to manipulate and process sequencing reads such as remove duplicated reads
8.	REDItools	To identify, filter and annotate RNA editing using genomic and/or transcriptomic NGS data
9.	Multiple Em for Motif Elicitation (MEME) algorithm	To identify ADAR-binding sequence motif around identified RNA editing sites

A total of eight tools were used in the analysis. For raw sequencing data quality check and trimming, FASTQC (Andrews, 2010) and Trimmomatic version 0.36 (Bolger et al., 2014) were used. Bowtie2 and HISAT2 were used to map the truncated sequencing reads. Bowtie2 is an ultrafast, memory-efficient alignment program for aligning reads of about 50 up to 100s or 1000s bp long to reference genome (Langmead et al., 2012). Due to splicing, RNA-Seq data contain large gaps which correspond to introns (Pertea et al., 2015). Thus, a program which is able to place spliced reads across intron and determine exon-intron boundaries correctly during transcriptomic data

mapping is crucial. The latter tool, HISAT2 was chosen based on its fast and sensitive alignment ability which allows mapping across exon-exon junctions (Pertea et al., 2015). HISAT2 mapped the RNA-Seq reads to the reference genome as well as known splice sites from GENCODE (version 25) and indexes named genome_snp_tran which was built using Ensembl annotated transcripts. Output of both alignment tools was given in Sequence Alignment/Map (SAM) format, is a generic alignment format that can store read alignments to reference sequences and support short and long sequencing reads (up to 128 Mbp). To parse and manipulate alignment in SAM/BAM formats, SAMtools (Version 1.3.1) was used. To manipulate the high-throughput sequencing data, Picard tools was chosen. The tool offers a variety of accessory tools such as remove duplication (MarkDuplicates) and create indexed (BuildBamIndex) allows fast retrieval of alignments in the BAM files. We adopted Genome Analysis Toolkit (GATK) version 3.7, a data analyzing and processing package developed by Broad Institute, USA to analyze the RNA-Seq and DNA-Seq data (DePristo et al., 2011). GATK offers various tools including tools that focus on variant detection and genotyping as well as sequencing data quality assurance. Putative RNA editing sites were detected using REDItools suite (Picardi et al., 2013). REDItools are python scripts developed to study RNA editing based on the next generation sequencing genomic and transcriptomic data. REDItools were selected due to its ability to facilitate the browsing of results and assist users through the annotation of predicted positions by using UCSC Genome Browser. Lastly, MEME suites (Bailey et al., 2009) were used to identify sequence motif around the identified RNA editing sites.

3.2 Methodology

3.2.1 Library preparation and whole genome sequencing

To obtain the corresponding genomic sequence of the sample, whole genome sequencing was conducted. The DNA was extracted and the sample was delivered to Malaysian Genomics Resource Centre (MGRC) for sequencing service. The DNA extraction was performed by using ENZA Blood DNA Mini Kit (Omege Bio-tek, USA). The extracted DNA sample was assessed to check the quality, quantity and integrity of the DNA. The DNA quality and quantity were measured using Nanodrop and Qubit dsDNA HS assay. The sample was also run on 1% agarose gel to determine the integrity of the DNA. Typically, at least 1 μg of high quality DNA (as measured by Qubit) with intact band seen on agarose gel is required for library preparation. DNA was fragmented using Covaris (Covaris Inc, USA) to a targeted size of 350 bp. The fragmented DNA was end-repaired, ligated to adapters and PCR-enriched using Truseq Nano DNA HT Sample Preparation Kit (Illumina, USA) according to manufacturer's protocol. The final library was quantified using Qubit DNA assay. Library size was determined using Bioanalyzer DNA Nano 6000 chip. Finally, the resulting library was sequenced on Illumina flow cell on HiSeq 2500 (Illumina, USA). The sequencing run generated an approximately 156 GB of 1,042 M raw DNA-Seq reads.

3.2.2 Pre-processing of raw sequencing transcriptomic and whole genomic reads

To perform comparative analysis between genomics sequence-dependent and sequence-independent approaches, we had developed two computational pipeline to identify RNA editing sites in our sample (Figure 3.1; Figure 3.2). Tools and filtering parameters used in both pipeline were the same to allow fair comparison. The computational pipelines were verified using 6 brain samples data downloaded from GEO database prior the analysis. The validation results were discussed in the chapter

Discussion. The analysis was initiated by inspecting the quality of the raw RNA-Seq and DNA-Seq reads using FASTQC program. Subsequently, these reads and adapters were trimmed from the raw datasets by using Trimmomatic with default parameters. Any RNA-Seq and DNA-Seq reads with average quality per base below 20 were excluded. The TruSeq adaptors in both dataset were also removed.

3.2.3 Alignment of RNA-Seq and DNA-Seq reads

Next, the cleaned DNA-Seq reads were then aligned to human reference genome (GRch38) using Bowtie2 with default parameters. To map the cleaned RNA-Seq, two short read spliced aligners, Bowtie2 and HISAT2 with default parameters were used. HISAT2 mapped the RNA-Seq reads to reference genome by referring to known splice sites from GENCODE (version 25) and indexes named genome_snp_tran, which was built using Ensembl annotated transcripts. The overall alignment of the sequences alignment will be discussed further in the Results chapter.

3.2.4 Post-processing of reads alignment

Only unique and concordant alignments in SAM format from RNA-Seq and DNA-Seq alignment were kept for further analysis. The reads were converted into binary BAM format by *samtools view* to minimize data storage and improve analysis performance. For RNA-Seq alignment, the two BAM files generated from different tools were merged into a union by samtool merge. This is to reduce the differences between the mapping of DNA-Seq and RNA-Seq due to different algorithm involved. Basic statistics of all files were calculated by *samtools flagstat*. The BAM file was then sorted by coordinate to avoid loading extra alignments into memory which will reduce the efficienct of data processing. Next, we adopted the workflow in GATK to perform reads filtering. There were four major steps in the filtering process. Firstly, duplicates were removed to mitigate biases introduced by data generation steps such as PCR

amplification within the RNA-Seq and DNA-Seq libraries using Picard tools. The duplicates-free BAM files were subsequently indexed using Picard tools. Then, only BAM file for RNA-Seq was proceeded for Split'N'Trim filtering step. This filtering was applied to RNA-Seq BAM file only to remove occurrence of artifacts in the splice junctions. Local realignment around INDELs was performed for RNA-Seq and DNA-Seq BAM files. This algorithm functions to remove possible errors that arose during initial mapping steps. Lastly, base quality scores of each base for both RNA-Seq and DNA-Seq BAM files were recalibrated in order to generate more accurate base quality score, which can improve the accuracy of the editing sites calling.

3.2.5 Identification of RNA editing sites

Putative RNA editing sites were detected using REDItools suite. To mitigate misalignment due to ambiguously mapped reads, the mapped RNA-seq was BLAT corrected by using accessory REDItools scripts (REDItoolBlatCorrection.py script) to identify a list of reads that were possibly mis-mapped. REDItoolDnaRna.py is the main script to identify the putative RNA editing sites by matching DNA-Seq and RNA-Seq data. All the genomic regions covered by RNA-Seq reads were inspected by the script to look for nucleotide changes between the reference genome and the RNA-Seq. The resulting mismatch sites were regarded as the putative RNA editing sites. To increase the sensitivity of RNA editing sites calling, we allowed the script to extract RNA and DNA positions with minimal coverage (-c) of 2 and 10, respectively, minimum quality score (-q) of 25 and minimum mapping quality (-m) of 20. In turn, to avoid false positive due to random-hexamer priming (Bass et al., 2012, Gurp et al., 2013), 6 bases at the 3' ends of each RNA-Seq reads were truncated (-a 6-0 option). We also required the sites to be supported by at least 1 variant bases (-v 1) without considering the frequency of variation in both RNA-Seq (-n 0.0) and DNA-Seq (-N 0.0). We further excluded substitutions supported by multi-mapping reads as attested by BLAT (-b),

removed substitutions in homopolymeric regions of >= 5nt in RNA-Seq and DNA-Seq (-l and –L option), and substitution located within 4nt of known spliced junction (-r and –w option). Finally, we excluded positions that were not supported by DNA-Seq reads (-V) and the outputs were stored in folder reditoolsOutput (-o option). Next, we filtered out known SNPs from dbSNP version 144 downloaded from UCSC. After removal of the known SNPs, we annotated the sites by the AnnotateTable.py script using RefSeq and RepeatMask annotations from UCSC to categorize sites in Alu and non-Alu regions. The validate the edited sites identified, we expanded our sequence analysis to DNA-Seq reads (15-nt upstream and 15-nt downstream of A-to-G sites in the Alu elements) using motif discovery tool, Multiple Em for Motif Elicitation (MEME) algorithm (Bailey et al., 2009) to find the sequence motif of ADAR-binding domain.

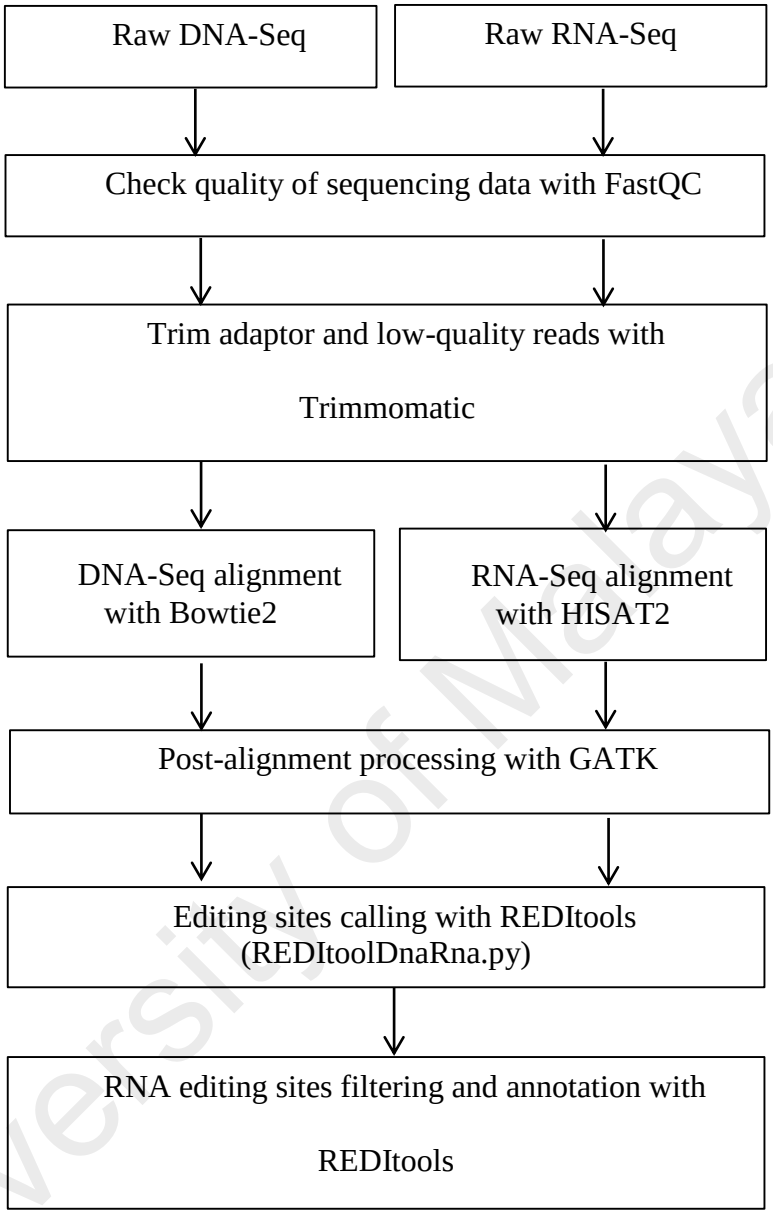

Figure 3.1: Workflow of the identification of RNA editing sites through genomic sequence-dependent approach.

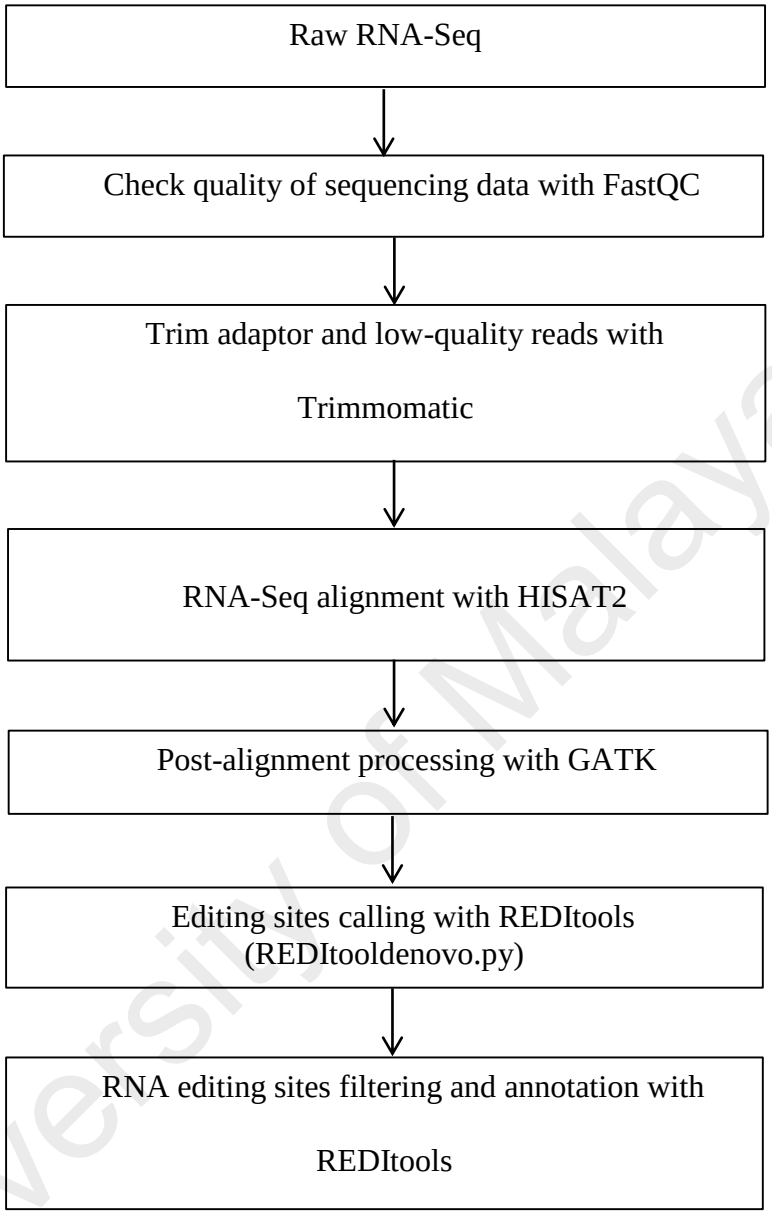

Figure 3.2: Workflow of the identification of RNA editing sites using genomics sequence-independent approach.

CHAPTER 4: RESULTS

With the advancement of NGS technologies, the identification of RNA editing sites has been growing rapidly, as seen by querying the numbers of RNA editing sites in the human genome to the RNA editing database (Ramaswami & Li, 2016). Various computational pipelines and software have been designed and introduced to facilitate the identification process of these sites. Through literature review, two approaches were found to be commonly used by researchers in order to identify these editing sites: genome sequence-dependent and genome sequence-independent approaches. Hence, this study served two objectives as outlined in Chapter 1 (page 5): (1) to identify and characterize A-to-I editing sites in genomic and transcriptomic sequences of healthy human primary monocytes using genome sequence-dependent and (2) to compare and contrast sensitivity and specificity of genome sequence-dependent and sequence-independent approaches in identifying RNA editing sites in human primary monocytes.

To answer the two main objectives, this results chapter was grouped into two sections. The first section (Section 4.1 with subsections of 4.1.1 to 4.1.4) describes the identification and characterization of RNA editing sites in human primary monocytes through computational data analysis of genomic and transcriptomic datasets. In this section, the quality of the datasets used, total number of edited sites identified under different types of filtering parameters, type of edited sites and distribution of the canonical A-to-I edited sites are presented in detail. In the second section (Section 4.2 with subsections of 4.2.1 and 4.2.2), the experiment as described above was repeated using genome sequence-independent approach. The sensitivity and specificity of both approaches in identifying edited sites in human primary monocytes were compared using the available RNA editing sites database, REDIportal.

4.1 Genome sequence-dependent approach in identifying RNA editing sites

The objective of this section outlined under objective 1 was to identify and characterize A-to-I editing sites in human primary monocytes through genome sequence-dependent approach. To the best of our knowledge, this study is the first to characterize the A-to-I canonical RNA editing sites presented in healthy human primary monocytes. Motivated by the study of Mitchell et al. (2015) using single and purified cell type isolated from single individual, we utilized a high-depth RNA-Seq raw data from Mirsafian et al. (2017) and performed whole genome sequencing to obtain the corresponding DNA information of the same healthy individual. Significant evolution of computing technologies such as super computer with high speed and capacity RAMs can handle huge volume of data which enabled the assembly and determination of variants in human genome.

4.1.1 Pre-processing of genomic and transcriptomic data

A total of 1,042,444,134 raw DNA-Seq reads with insert size range of 350 were generated using Illumina HiSeq 2000 platform. Sequencing reads quality is typically the first step in analyzing next generation sequencing data (Figure 3.1). For both DNA-Seq and RNA-Seq dataset, adapter and low quality reads were filtered to obtain an optimal quality score of 20 or higher at each base (Figure 4.1 to Figure 4.4). The reason for large number of low quality reads at the end of sequencing was due to the degradation of sequencing chemistry with the increase of read length. High sequencing coverage and the availability of reference genome during sequence alignment aid in overcoming this error. The filtered reads were exported as fastq files for further bioinformatics data analysis.

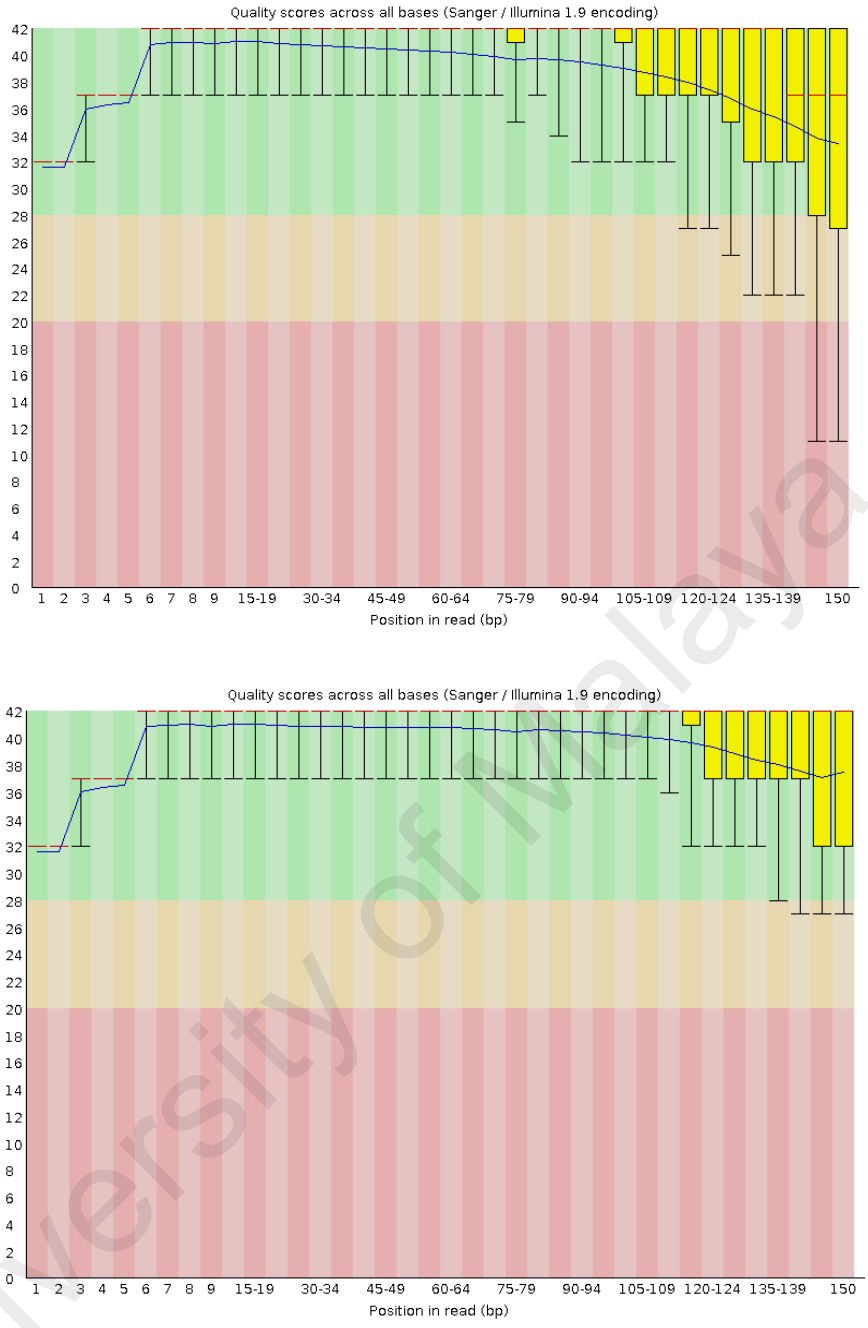

Figure 4.1: Quality control of the generated forward strand of DNA-Seq data before and after trimming process. Upper panel: before trimming; lower panel after trimming

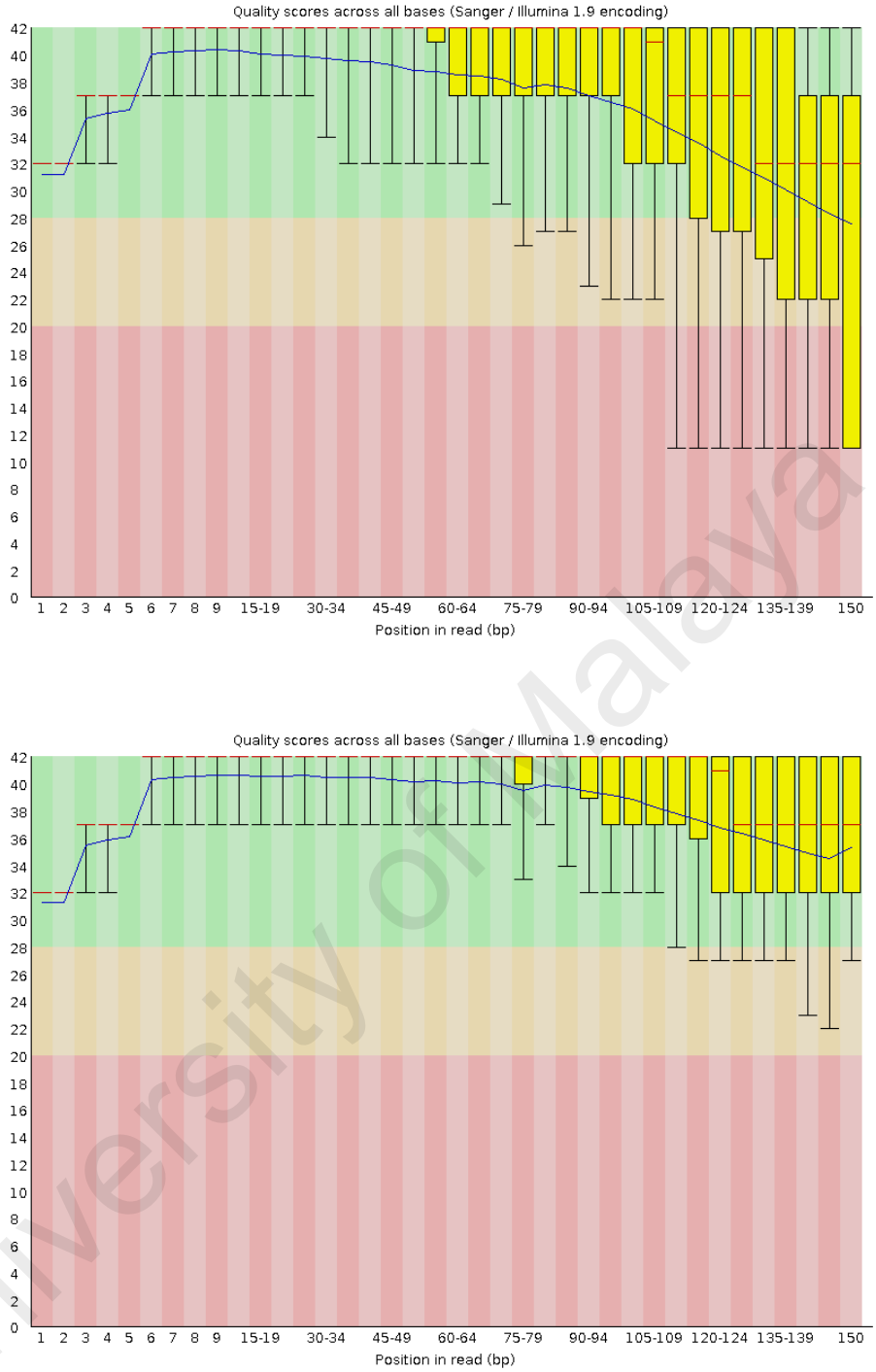

Figure 4.2: Quality control of the generated reverse strand of DNA-Seq data before and after trimming process. Upper panel: before trimming; lower panel after trimming

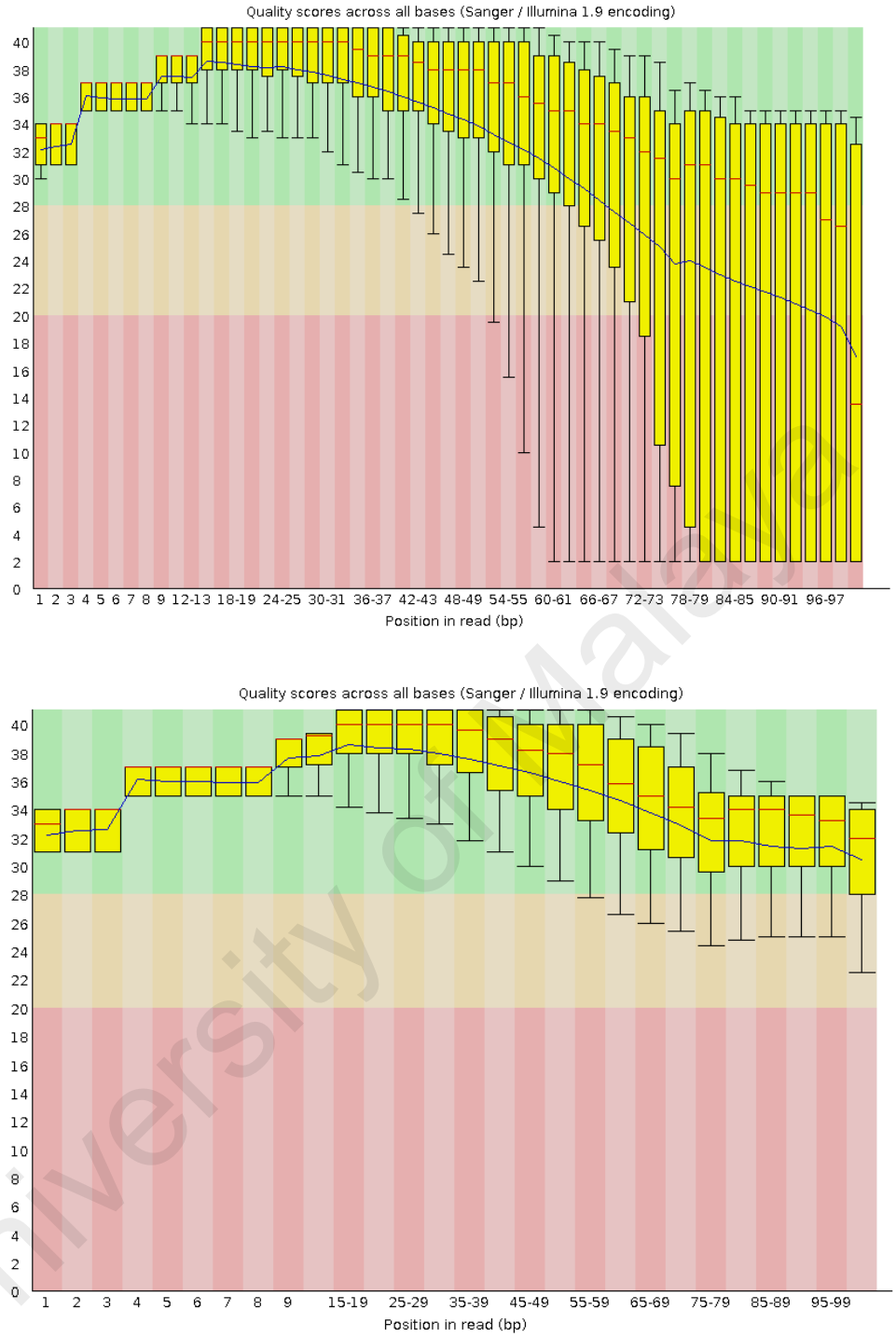

Figure 4.3: Quality control of the generated forward strand of RNA-Seq data before and after trimming process. Upper panel: before trimming; lower panel after trimming

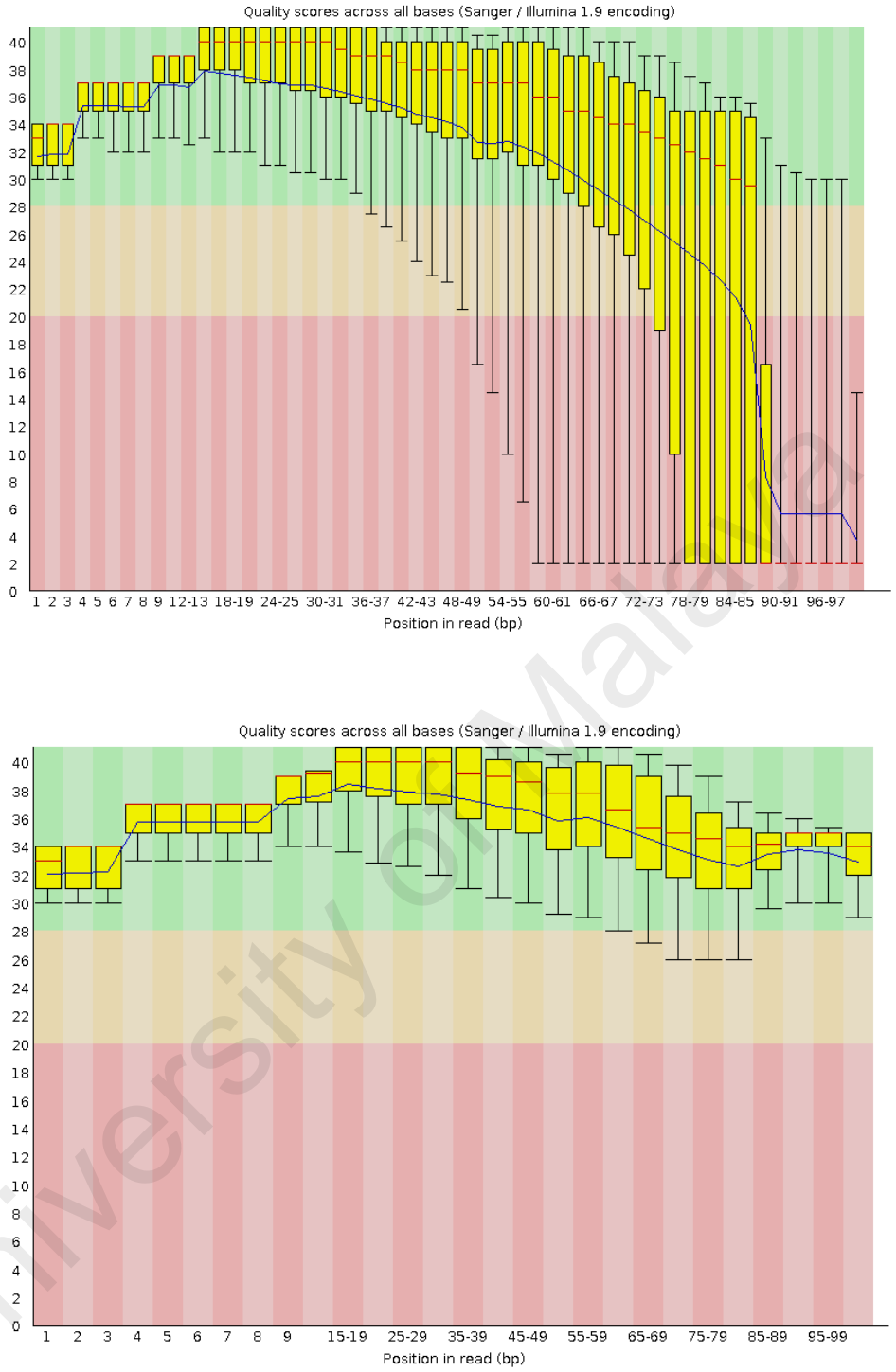

Figure 4.4: Quality control of the generated reverse strand of RNA-Seq data before and after trimming process. Upper panel: before trimming; lower panel after trimming

4.1.2 Sequence alignment of genomic and transcriptomic data of human

Discrimination of RNA editing sites from SNPs and technical artifacts caused by sequencing or read-mapping errors have always been the greatest challenge during RNA editing sites identification. Hence, accurate mapping was crucial in RNA editing sites identification. Accuracy of mapping is described as the proportion of reads that map to the correct location against human reference genome (Schbath et al., 2012). Aligning large numbers of reads to the reference genome was indeed a challenge for the mapping tools or software. In a comparative study between mapping tools by Schbath et al. (2012), Bowtie was observed to have excellent reads sensitivity and reasonable execution time with default parameters. Hence, for DNA sequencing, we generated 1,042 M reads using HiSeq2500 (Illumina) and aligned the read to human reference genome (version hg38) by using Bowtie2. Of all the reads generated during DNA sequencing, we obtained a high overall alignment percentage of 99.15%, of which approximately 84% aligned concordantly or uniquely to the hg38, approximately 2% aligned discordantly and the remaining 15% were observed to align to more than one position on the hg38 (Table 4.1). In contrast to DNA sequence alignment, RNA sequences have two additional challenges. Firstly, RNA reads consist of exons and introns of different length. Hence, RNA sequence alignment tools must be able to handle the splice junctions. Secondly, GENCODE reported that there were over 14,000 pseudogenes present in the human genome. Some or all of the introns may be removed due to the presence of pseudogenes and may cause RNA read to map incorrectly (Pei et al., 2012). Therefore in this study, we used multiple tools, Bowtie2 and HISAT2 to remap the raw RNA-seq reads (See Methods). HISAT2 is a tool that allow mapping across exon-exon junction. Of all the RNA sequencing reads aligned to hg38, the results showed that HISAT2 obtained a higher overall alignment percentage compared to Bowtie2 (Table 4.1). Moreover, HISAT2 also obtained a noticeably higher number of

reads that aligned concordantly and discordantly to hg38 compared to Bowtie2 (Table 4.1). Our results proved that tool that allow the mapping of RNA-Seq reads across splice junction indeed improve the accuracy of mapping.

Table 4.1: Mapping percentages of DNA-Seq and RNA-Seq of healthy human primary monocyte through different sequence aligner

	DNA-Seq	**RNA-Seq**	
Tools	**Bowtie2**	**Bowtie2**	**HISAT2**
Aligned concordantly	431,430,781 (83.55%)	45,187,499 (43.27%)	73,445,927 (70.32%)
Aligned > 1 times	75,901,378 (14.70%)	22,026,587 (21.09%)	23,821,460 (22.81%)
Aligned discordantly	9,059,193 (1.75%)	37,228,384 (35.64%)	7,175,083 (6.87%)
Overall alignment	99.15%	85.21%	97.03%

As reported previously (Bahn et al., 2012), only uniquely mapped reads were considered in RNA editing events identification. Therefore, in this study, we extracted the reads that aligned concordantly to hg38 to perform further analysis. As outlined in the GATK Best Practices (Van et al., 2013), post-alignment processing of aligned reads would improve the accuracy of variant calling. In agreement with Tian et al. (2016), the study suggested that local realignment around known INDELs and BQSR reduced erroneous calls of variants, although the extent of benefits of post-alignment processing were yet to be determined. The post-alignment processing in this study included the removal of duplication, local realignment around known INDELs and base quality scare recalibration (BQSR) (see Material and Methodology chapter).

4.1.3 Identification and characterization of RNA editing sites

To identify RNA editing sites in human primary monocytes, we adopted a computational framework based on the popularly used transcriptomic (HISAT2) and genomic (Bowtie2) mapping tools, variants calling best practice (Van et al., 2013) and RNA editing sites filtering criteria (Bahn et al., 2012; Bass et al., 2012) (Figure 3.1). REDItools package was selected as RNA editing sites detector. REDItoolDnaRna.py, the accessory scripts of REDItools detects putative RNA editing sites by comparing the pre-aligned DNA-Seq and RNA-Seq reads in the standard BAM format to look for nucleotide changes. DNA-Seq reads functions to support the presence of RNA editing site and exclude potential SNPs or somatic mutations. The scripts explore genomic position sites by site and returns a table containing the coverage depth, mean quality score, frequency of variation and all the possible types of nucleotide substitution or mismatch such as A-to-G, A-to-C, A-to-T, G-to-A and etc. Then, the RNA editing sites can be filtered according to read coverage, base quality score, mapping quality, number of bases supporting the variation, type of mismatch and frequency (Picardi et al., 2013). Under standard sequencing data quality filters (coverage, mapping and base quality) for variant calling, we identified a total of 108,475 putative RNA editing sites. To distinguish RNA editing sites from known or rare SNPs, we annotated our list of putative RNA editing sites with dbSNP144 and 53% of the sites were found to be known SNPs while approximately 47% of the remaining sites were RNA editing sites which is 51,484 (Table 4.2). The RNA editing sites were comprised of 22% of A-to-G changes (which is also A-to-I editing), 21% of T-to-C changes, 8% C-to-T changes, 8% G-to-A changes and 8 other possible nucleotide changes which range from 4% to 6% (Figure 4.5).

To further reduce false positive, we integrated additional strict filters (BLAT correction, removal of first six bases of reads, homopolymeric regions longer than five

residues and intronic sites in the first four bases from known splice sites) in REDItools to call for RNA editing sites from the same dataset. The total number of putative RNA editing sites identified was 57,108. Of the putative sites identified under strict filtering parameters, 78% were annotated as known SNPs while the remaining 22% were RNA editing sites which is 12,421 (Table 4.2). Of the 22% RNA editing sites identified, 35% were A-to-G changes, 34% were T-to-C changes, 7% were C-to-T changes, 7% G-to-A changes and the remaining portions were shared by 8 other possible types of nucleotide changes which range from 1% to 3% (Figure 4.5). In either of the filtering parameters applied, A-to-G changes represented the largest category of RNA editing sites in human primary monocytes (Figure 4.5). On top of that, T-to-C changes also demonstrated a higher number of substitutions compared to other types of nucleotide changes (Figure 4.5). These sites were believed to originate from A-to-I editing events. Bazak et al. (2014) reported that the editing sites could appear as A-to-G changes if the transcription initiated from the reference strand or T-to-C changes if transcription initiated from the reverse strand5. Consistently, Bahn et al. (2012), reported a decrease in T-to-C changes upon the knocked down of ADAR enzymes in human. This showed that significant proportion of these sites may be produced by the canonical A-to-I RNA editing. Overall, with a total of 69% of A-to-I canonical sites (A-to-G and T-to-C changes) and 14% C-to-U canonical sites (C-to-T and G-to-A changes) were identified in our healthy subject under strict filters, our finding supports the existing knowledge that the A-to-I editing is the primary type of RNA editing events in human (Picardi et al., 2015; Zinshteyn and Nishikura, 2009; Li and Church, 2013). By comparing the canonical sites identified under both filters, the proportion of A-to-I editing showed an increment of 26% and while C-to-U editing showed a slight decrements of 2% after the integration of strict filters (Figure 4.5), respectively. While other types of changes were also identified, previous studies showed that these non-canonical editing events are unlikely

to be real and with at least 85%-98% of these changes were originated from technical artifacts (Piskol et al., 2013; Kleinman et al., 2012).

Table 4.2: Number of putative RNA editing, RNA editing sites and known SNPs identified in human primary monocyte under standard quality and strict filters using genome sequence-dependent approach.

	Standard quality filters	**Strict filters**
Putative RNA editing sites	108,475	57,182
RNA editing sites	51,484 (47%)	12,421 (22%)
Known SNPs in dbSNPs144	56,991 (53%)	44,767 (78%)

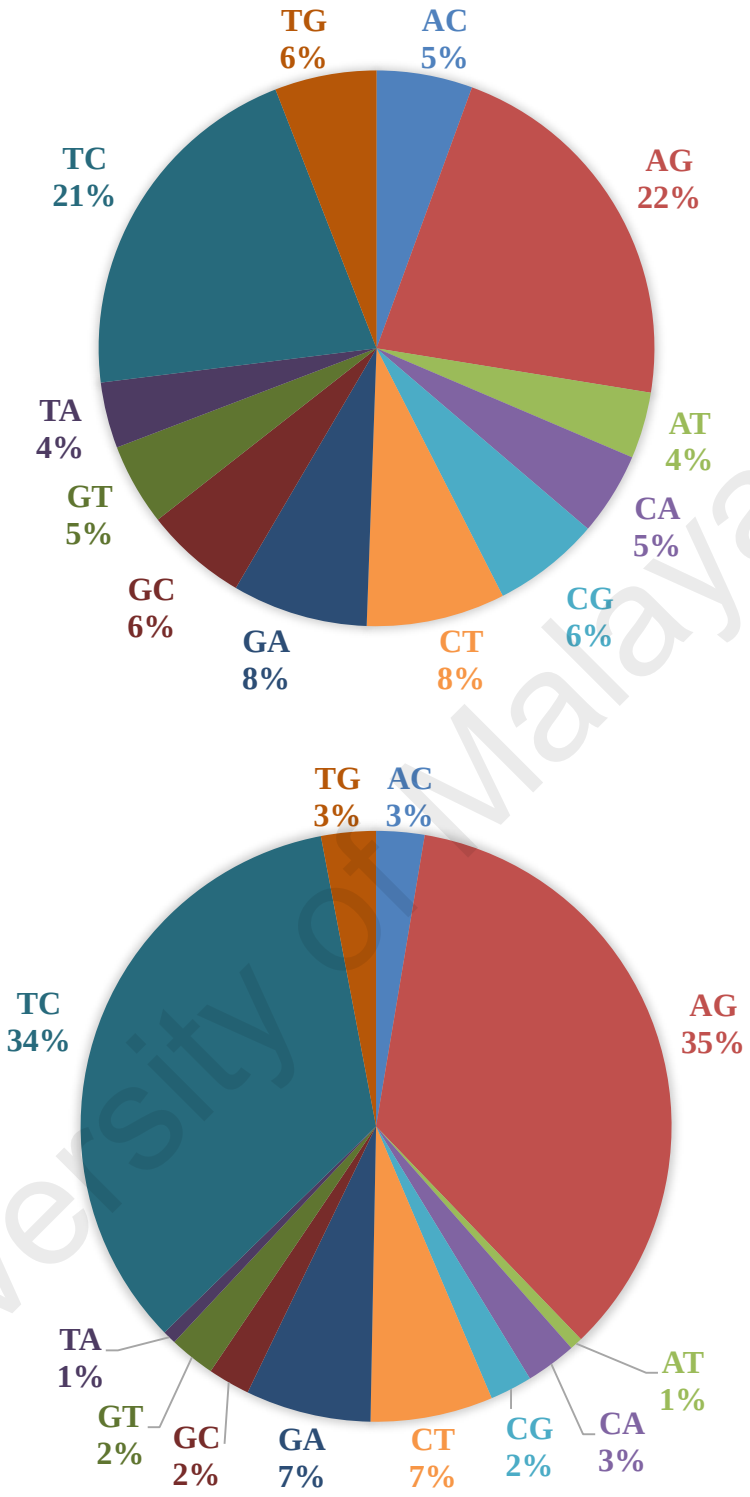

Figure 4.5: Genome-wide presence of canonical and non-canonical editing sites in healthy human primary monocytes using genome sequence-dependent approach. Upper panel: Under standard quality filters; lower panel: under strict filters.

4.1.4 Characterization of A-to-I editing sites of human primary monocytes

We then further classified the canonical A-to-I RNA editing sites identified under strict filters using genome sequence-dependent approach. Repetitive elements especially Alu elements were known to be frequently targeted by ADAR enzyme which then led to A-to-I deamination (Kim et al., 2004). Alu elements are widespread repetitive elements in human genome and are abundantly interspersed within introns and untranslated regions (UTR). By chance, adjacent inverted Alu elements form long stable stem-loop structures due to their complementary sequences in the opposite orientations, which are the favorable editing substrate (Bass, 2002; Deininger, 2011). While millions of editing sites were identified from different studies (Bazak et al., 2014; Bass, 2002; Ulbricht & Emeson, 2014) the function of Alu-editing has yet to be elucidated. To understand the location and distribution of canonical editing sites in repetitive elements of human primary monocyte, we annotated our results with RepBase (Bao et al., 2015), a library of known repeats using accessory script from REDItools. Out of a total of 8,632 A-to-I editing sites (4,370 A-to-G changes and 4,262 T-to-C changes) identified in our subject, the A-to-I editing sites were showed to enrich in the repetitive elements (67%) which 56% in repetitive Alu elements and 12% in repetitive non-Alu elements while the remaining 32% were found in non-repetitive elements (Figure 4.6). A study has shown that editing sites in non-Alu editing events were induced by inverted Alu repeats and these elements usually located hundreds of nucleotides away from Alu editing sites (Daniel et al., 2014). While our observations were in-line with the existing knowledge of which majorities of the edited sites in higher eukaryotes resided in repetitive sequence regions (Bazak et al., 2014; Picardi et al., 2015), there was a considerable number of canonical RNA editing sites identified in non-repetitive elements. Identification of RNA editing in non-repetitive elements has known to be challenging

(Bass et al., 2012), therefore a more careful examination need to be carried out at RNA editing sites resided at non-repetitive elements to further reduce false positives.

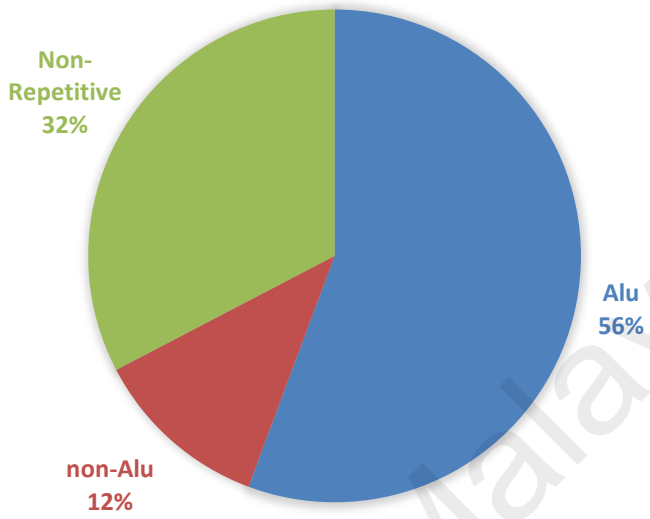

Figure 4.6: Distribution of canonical RNA editing sites (A-to-G and T-to-C changes) in human primary monocytes in repetitive Alu elements, repetitive non-Alu elements and non-repetitive non-Alu elements.

To understand the genomic localization of the canonical editing sites identified in our subject, we annotated our catalogue of canonical editing sites with NCBI Reference Sequence (RefSeq) database downloaded from the UCSC genome browser by using accessory script in REDItools. In our dataset, 50% (4,302/8,632 sites) of the sites were in the intron while only 0.6% (53/8,632 sites) of the A-to-I editing located in exons (Figure 4.7). In agreement with a recent study in human, RNA editing sites identified from lung tissues of three individuals by using genome sequence-dependent approach has also revealed the enrichment of edited sites in the intronic regions (Goldstein et al., 2017). On the other hand, approximately 2% (183/8,632 sites) of the canonical RNA editing sites identified in our study was in the 3' untranslated regions (3' UTR) and very few editing sites, 3% (44/8,632 sites) were in 5' UTR (Figure 4.7). Even though there were large number of sites being identified, only 3.4% (293/8,632 sites) were present in the coding sequences (CDS). The remaining 44% (3,757/8,632) RNA editing sites were classified as unknown events (Figure 4.7).

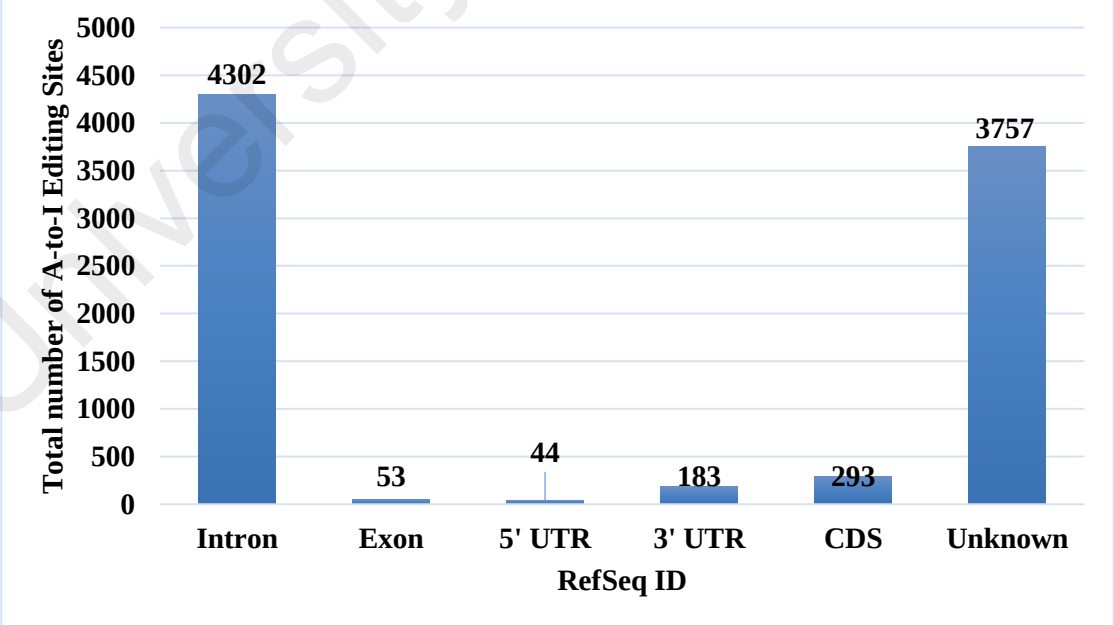

Figure 4.7: The genomic localization of canonical RNA editing sites (A-to-G and T-to-C changes) in human primary monocytes.

A study by Ramaswami and Li (2016) highlighted the overlap of editing sites identified between several studies is quite low suggesting that each study is querying a portion of the total of editing collection. To obtain an insight on the common sites between our results with other study, we intersected the results with editing sites downloaded from REDIportal (Picardi et al., 2017). REDIportal is currently known to be the largest non-redundant collection of A-to-I editing sites. The portal yielded a total of 4,668,508 non-redundant A-to-I editing sites by merging RADAR (Ramaswami & Li, 2013) with all the RNA editing sites identified in ATLAS project (Picardi et al., 2015). ATLAS is the largest single collection of human editing events in six human tissues (brain, kidney, liver, lung, heart, muscle) from three individuals. Out of the 8,634 canonical A-to-I editing sites identified in our dataset, 0.8% (70/8,632 sites) of canonical RNA editing sites were found intersecting with REDIportal while 99.2% (8,564/8,632 sites) were found to be novel to REDIportal (Figure 4.8).

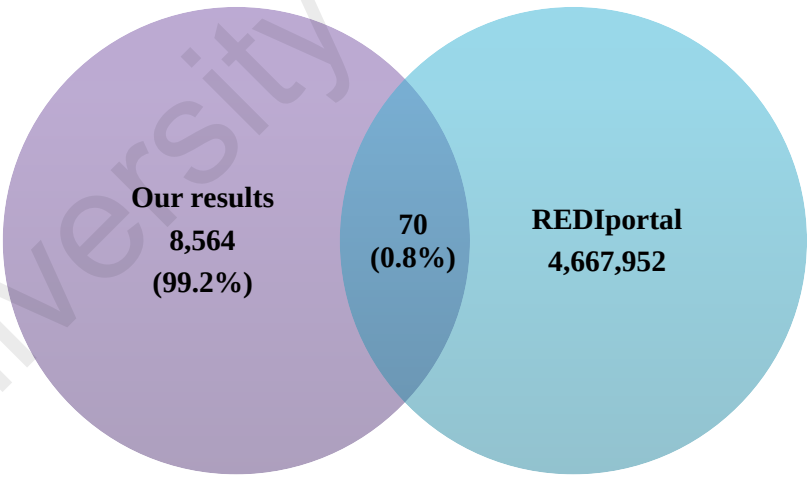

Figure 4.8: Venn diagram of canonical RNA editing sites (A-to-G and T-to-C changes) identified in our study and REDIportal.

In the ATLAS project, almost 97% of the A-to-I editing sites documented in the project were edited at frequency value of 0.73 – 0.96 (Picardi et al., 2015). In this study, the usage of high depth sequencing data enable us to detect RNA editing sites in lower frequencies (Chen, 2013; Huntley et al., 2011; Lee et al., 2013; Ramaswami et al., 2013). Therefore, approximately 90% (7,670/8,564 sites) of the novel canonical editing sites identified were edited at frequency value of < 0.7 while only 10% (894/8,564 sites) of the sites had editing frequency value of ⩾ 0.7 (Figure 4.9). Of these, approximately 4% of the novel editing sites was showed to have a frequency level of 100% (completely edited).

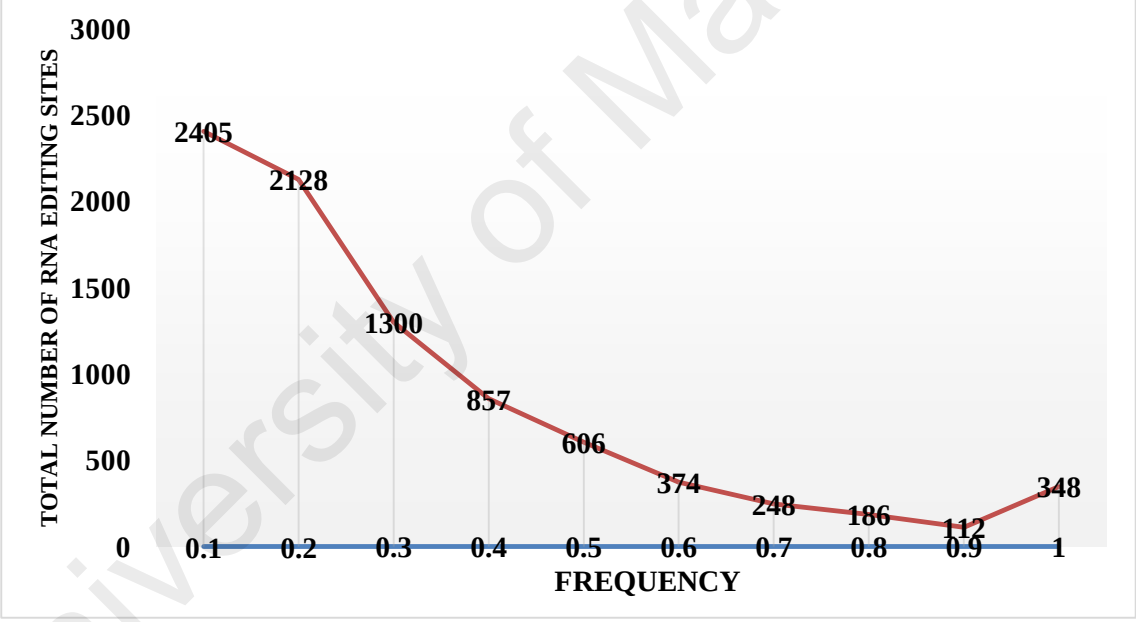

Figure 4.9: Total number of canonical RNA editing sites (A-to-G and T-to-C changes) identified at different frequency values.

To validate the large amount of newly identified RNA editing sites through in silico experiment, A-to-I editing sites that were resided in Alu element were subjected to further analysis. According to Bahn et al. (2012), A-to-I editing were mediated by ADAR that recognize specific sequence motifs around the RNA editing sites. We expanded our sequence analysis to DNA-Seq reads of these sites (15-nt upstream and 15-nt downstream of A-to-G sites in the Alu elements) using motif discovery tool, Multiple Em for Motif Elicitation (MEME) algorithm (Bailey et al., 2009) to find the sequence motif. Previous studies showed that the A-to-G sites were observed to have a depletion of nucleotide G at the -1 position (upstream) and preference for G at the +1 position (downstream) (Bazak et al., 2014; Lehmann & Bass, 2000; Eggington et al., 2011). As a result, our editing sites (present at position 16) were in accordance with the known sequence motif of ADAR which the one base upstream of the edited sites was enriched by nucleotide C while the one base downstream of the edited sites was enriched by nucleotide G (Figure 4.10).

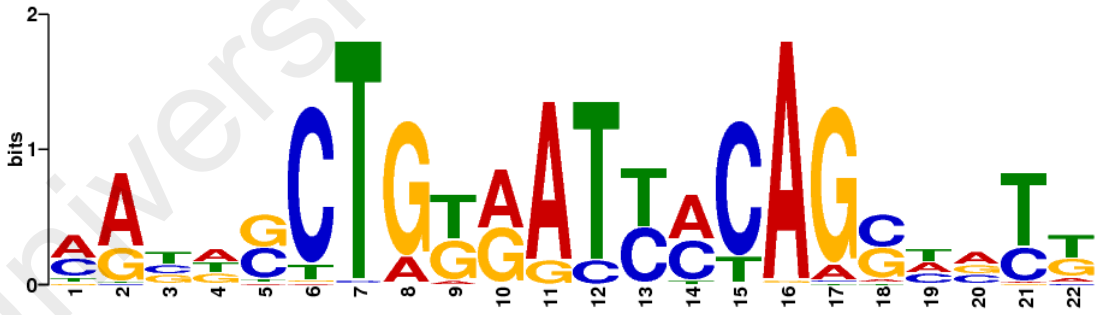

Figure 4.10: Sequence motif of A-to-I editing sites (editing sites present at position 16).The motif showed to have a depletion of G at the -1 position (enriched by C) and preference for G at the +1 position.

In summary, there were a total of 8,632 canonical A-to-I editing sites identified using the genome sequence-dependent approach with strict filters. Majority of these sites were located at the intronic regions and repetitive Alu sequences of human primary monocytes. High-depth genomic and transcriptomic datasets also revealed editing sites

that were edited at lower frequencies. The results of section 4.2 were further elaborated and discussed in the Discussion chapter (Chapter 5).

4.2 Genome sequence-independent approach in identifying RNA editing sites

In the previous section, we have successfully identified and characterized the RNA editing sites presented in healthy human primary monocytes. Conventionally, genomic and transcriptomic sequences are required to analyze the event. Recently, high-depth RNA-Seq has enabled the identification of editing events without the incorporation of genomic sequences. Therefore, in this section, we developed a genome sequence-independent computational pipeline for the identification of RNA editing sites based on the widely used bioinformatics tools. We also compared and contrasted the RNA editing sites identified through both approaches as outlined in objective 2 (see page 5). Identification of RNA editing sites using genome sequence-independent approach

To perform genome sequence-independent analysis, we used only the RNA-Seq data from genome sequence-dependent analysis without incorporation of the DNA-Seq data to the same pipeline (Figure 3.2). REDItooldenovo.py is one of the accessory scripts from REDItools. The script calculates the distribution of expected bases and observed bases at all genomic positions. Significant positions passing the false positive discovery rate were output as putative RNA editing sites. Similar to genome sequence-dependent approach, we initiated the search by using standard quality filters followed by strict filters. Under standard sequencing data quality filters, we identified a total of 94,961 putative RNA editing sites. After filtered out approximately 75% of known SNPs present in dbSNP144, the number of RNA editing sites identified was 22,988 (Table 4.3).

Table 4.3: Number of putative RNA editing, RNA editing sites and known SNPs identified in human primary monocyte under standard quality and strict filters using genome sequence-independent approach.

	Standard quality filters	**Strict filters**
Putative RNA editing sites	57,182	89,855
RNA editing sites	12,421 (22%)	22,127 (25%)
Known SNPs in dbSNPs144	44,767 (78%)	67,728 (75%)

The RNA editing sites were comprised of 78% of A-to-I editing sites (A-to-G and T-to-C changes), 11% C-to-U editing sites (C-to-T changes and G-to-A changes) and 21% of non-canonical RNA editing sites (Figure 4.11). Under strict filters circumstance, the total number of RNA editing sites identified was 22,167 with 79% of A-to-I editing sites, 11% C-to-U editing sites (C-to-T changes and G-to-A changes) and 20% of non-canonical RNA editing sites (Figure 4.11). The above observation showed the output of RNA editing sites identification though genome sequence-independent approach using single set of high depth RNA-Seq data were in-line with the existing knowledge. To enhance the sensitivity in RNA editing sites identification, we applied the similar pipeline on multiple human primary monocytes RNA-Seq data and retrieved the recurring RNA editing sites (Appendix A). By using the above method, we found that the majority of RNA editing sites in human primary monocytes were indeed A-to-G and T-to-C changes which indicative A-to-I editing followed by C-to-U changes at 11% (Figure 4.12). RNA editing is a tissue-specific event and human brain tissues were shown to be one of the greatly edited tissues (Li & Church, 2013). Hence, in order to validate the pipeline used in the above analysis, we downloaded the RNA-Seq data of human brain tissues from GEO database and called RNA editing sites by using the same pipeline (Appendix B). The result showed that 95% of the RNA editing sites identified

were canonical editing sites (88% A-to-I editing and 6% C-to-U editing) (Figure 4.12).

The above observation supported that pipeline were reliable.

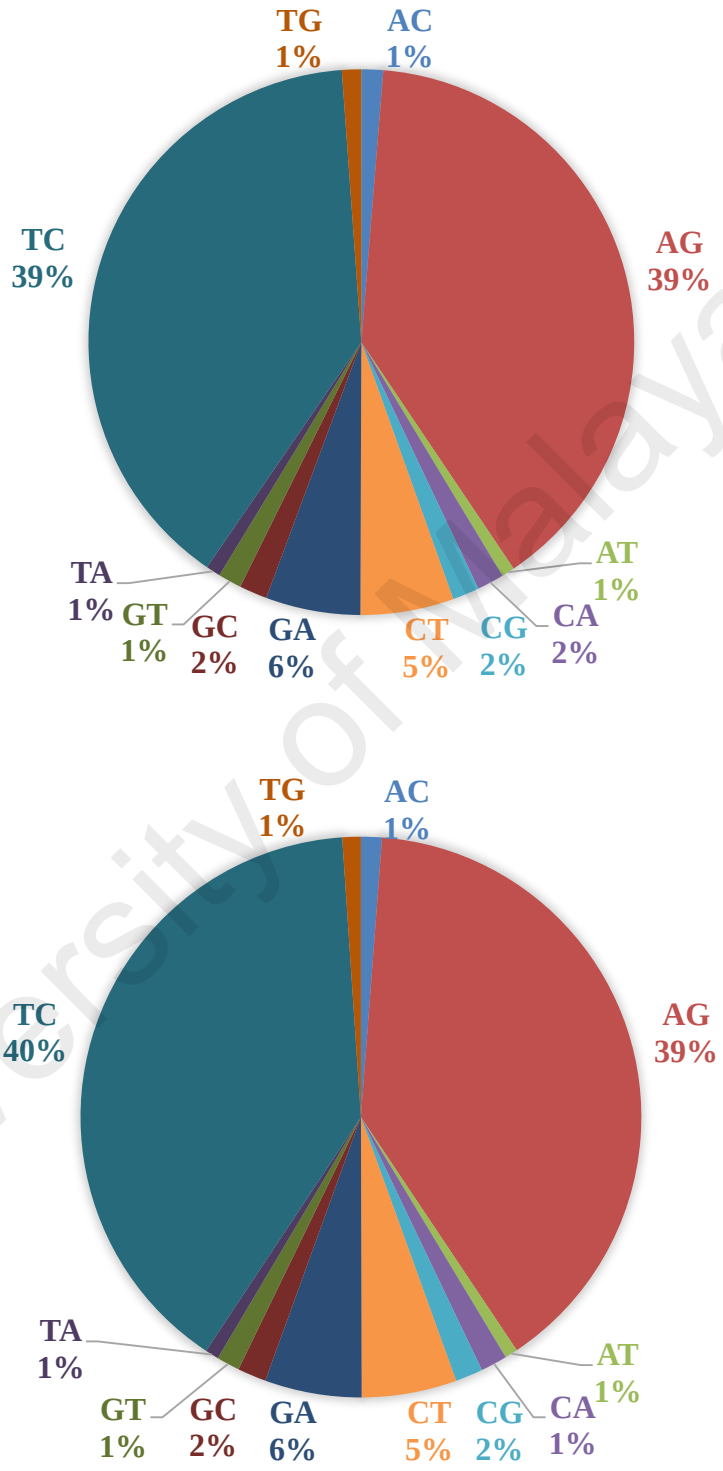

Figure 4.11: Genome-wide presence of canonical and non-canonical editing sites in healthy human primary monocytes using genome sequence-independent approach. Upper panel: Under standard quality filters; lower panel: under strict filters.

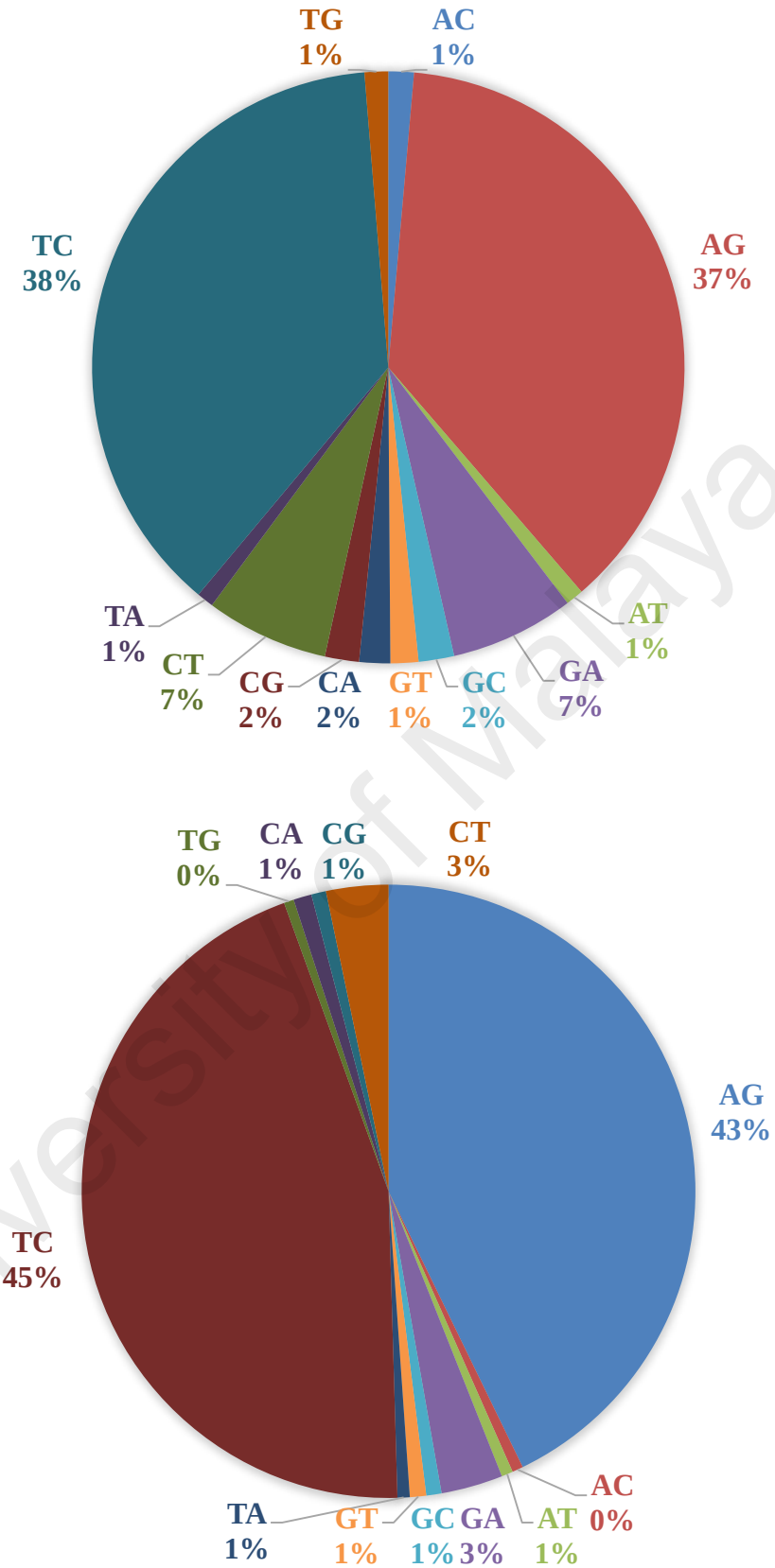

Figure 4.12: Pie charts of genome-wide presence of canonical and non-canonical editing events of multiple samples. Upper panel: healthy human primary monocytes. Lower panel: healthy human brain tissues

CHAPTER 5: DISCUSSION

Genome sequence-dependent and sequence-independent has been widely used in different studies and project. To date, there was no specific guideline on the type of approach need to be used as it is dependent on the goal of the study and budget. In this section, the sensitivity and specificity of RNA editing sites identified were compared and contrasted by overlapping the results from both approaches with each other as well as overlapped with the available RNA editing sites database, REDIportal. To the best of our knowledge, this is the first comparison on the approaches in identification of RNA editing in human cells.

As shown in Figure 4.8, a low intersected percentage was observed between our findings and REDIportal using genome sequence-dependent approach suggesting several possibilities. Firstly, we assumed the low intersection may be due to the novel A-to-I sites identified in this study were edited at a low frequency (Figure 4.9). There were approximately 4% of the edited were completely edited. Few year ago, Li et al. (2009) reported RNA editing sites in seven tissues of a single individual with frequency level from 2% to 100% 55. A later study by Zhu et al. (2012) has also reported RNA editing sites expressed in three normal human brain samples with frequency level of 5% to 95%. The study suggested that to accurately measure the frequency of RNA editing, which may range from 0% to 100%, a much higher coverage is required to compensate for sequencing error 56. In our study, we used a set of high-depth NGS data (>300 M) and it has been well described that accuracy is increased with by coverage in the NGS platform (Harismendy et al., 2009). Secondly, cell-level variation in post-transcriptional modification such as A-to-I RNA editing was remains poorly understand. Human tissues composed of a variety of cell classes and subtypes. The patterns and extent of RNA editing may be different among cells. In a recent study, RNA editing events were

showed to distribute differently among different types of brain cells (Picardi et al., 2017). Moreover, A-to-I editing events were known to be strong tissue-dependent events which brain tissues have been reported as the most edited human tissue (Picardi et al., 2015). Hence, we assumed human primary monocyte may have lower dependency for A-to-I RNA editing to take place. Besides, the lack of investigation of RNA editing sites at the single cell-type of monocytic lineage may contribute to the low overlapped percentages with the database. To validate the novel canonical sites identified by in silico method, editing sites that resided in the Alu elements were subjected to further analysis. As a result, our finding showed that the A-to-G sites were observed to have a depletion of nucleotide G at the -1 position (upstream) and preference for G at the +1 position (downstream). This result supports the idea that the A-to-G sites found do contain mostly genuine editing sites but the functional validation of the identified motif shall be the subject of future study. In agreement with Pandey and colleagues, by using a set of deeply-sequenced DNA-Seq and RNA-Seq reads, our results were showed to in line with the existing knowledge which majority of the RNA editing sites in healthy human primary monocytes were A-to-I editing and these sites were enriched in repetitive elements and intronic region. However, further studies are deemed necessary to be carried using more samples in order to derive statistically significant conclusion.

Overall, both approaches used in our findings showed that human primary monocytes were enriched with A-to-I editing with a noticeable proportion of C-to-U editing sites. This suggested that human primary monocytes may behave in a unique way and further analysis is indeed needed in order to establish a conclusion. As shown in Table 5.1, genome sequence-independent was more commonly used by researchers to identify RNA editing sites. This may largely due the hassle in handling DNA-Seq data. Even with the advancement of sequencing which resulted in a drop of sequencing cost, the cost of DNA-Seq is still relatively more expensive than RNA-Seq. Additionally,

whole genome is classified as large data and more computational time, hardware memory and storage is required in order to perform RNA editing sites identification.

Despite the difficulties in performing the analysis, Table 4.2 and 4.3 revealed that genome sequence-independent approach identified more RNA editing sites compared to genome sequence-dependent approach under strict filtering parameters. Our observation was in agreement with Ramaswami et al. (2013) who also reported the presences of more edited sites were identified using RNA-Seq data alone (Ramaswami et al., 2013). We suggested that the DNA-Seq information from the same individual have possibly reduced the novel and rare genomic variants (SNPs) as well as somatic mutations being interpreted as RNA editing events causing the total number of RNA editing sites obtained to be lower. Researchers are recommended to take all these issues into consideration while designing related experiments.

Table 5.1: Pros and cons of genome sequence-dependent and sequence-independent methods.

Genome sequence-dependent	**Aspects**	**Genome sequence-independent**
Less popular	Popularity	More popular
Higher	Sequencing cost	Lower
Longer	Computational time	Shorter
More disk space is required	Data storage	Lesser disk space is required
Lesser	Number of RNA editing sites identified	More
Lesser	Number of SNPs	More

CHAPTER 6: CONCLUSION

In this thesis, methodology of the identification of RNA editing sites using genome sequence-dependent and sequence-independent were discussed. With the availability of whole transcriptome, we sequenced whole genome of the corresponding individual to perform RNA editing sites identification under different filtering parameters. The objectives of this study are to identify and characterize RNA editing sites in human primary monocytes, as well as compare the approaches used in the study.

As a result, we have successfully identified a total of 12,429 A-to-I editing sites using genome sequence-dependent approach under strict filtering parameters in human primary monocytes. The A-to-I editing sites were enriched in intronic regions as well as repetitive elements especially the Alu elements. The use of high-depth RNA sequencing and DNA sequencing datasets revealed large number of novel A-to-I RNA editing sites with 95% of them were edited at lower frequency. This study also demonstrated that genome sequence-independent approach identified more RNA editing sites compared to genome sequence-dependent approach. While more sites were being identified using sequence-independent method, the method also showed to contain more false positives (SNPs). Additionally, we have listed the pros and cons of both approaches in chapter Discussion which are beneficial in assisting researchers during experimental design.

In conclusion, the findings have provided sufficient evidence in order to fulfill all the objectives of the study. It has been shown that RNA editing sites were found in human primary monocyte and were enriched in the intronic repetitive elements. Comparison between both approaches also showed that genome sequence-dependent yielded more confident results, however sequence-independent approach is more favorable by researches as it is more cost effective. Our study therefore provides a novel insight to

the distribution of A-to-I RNA editing in healthy human primary monocyte at cellular-level.

To the best of our knowledge, this is the first study which identified and characterized RNA editing sites in healthy human primary monocytes. Due to sample-type limitation, we unable to provide the effects of canonical RNA editing to disease-state human primary monocyte in this report. In future, functional relevance of A-to-I editing events in healthy monocytes as well as in the disease-state should be explored. Profiling the RNA editing sites in both healthy and disease samples will be a novel approach to understand the possible influence of RNA editing towards human disease. Further studies are deemed necessary to be carried using more samples in order to derive statistically significant conclusions.

REFERENCES

Anant, S., & Davidson, N. O. (2001). Molecular mechanisms of apolipoprotein B mRNA editing. *Current Opinion in Lipidology, 12(2)*, 159-165.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research, 37(suppl_2)*, W202-W208.

Bahn, J. H., Lee, J. H., Li, G., Greer, C., Peng, G., & Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Research, 22(1)*, 142-150.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA, 6(1)*, 11.

Bass, B. L. (1997). RNA editing and hypermutation by adenosine deamination. *Trends in Biochemical Sciences, 22(5)*, 157-162.

Bass, B. L. (2002). RNA editing by adenosine deaminases that act on RNA. *Annual Review of Biochemistry, 71(1)*, 817-846.

Bass, B., Hundley, H., Li, J. B., Peng, Z., Pickrell, J., Xiao, X. G., & Yang, L. (2012). The difficult calls in RNA editing. *Nature Biotechnology, 30(12)*, 1207.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., ... Levanon, E. Y. (2014). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Research, 24(3)*, 365-376.

Benne, R., Van Den Burg, J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H., & Tromp, M. C. (1986). Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell, 46(6)*, 819-826.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30(15)*, 2114-2120.

Chan, T. H. M., Lin, C. H., Qi, L., Fei, J., Li, Y., Yong, K. J., ... Yuan, Y. F. (2013). A disrupted RNA editing balance mediated by ADARs (Adenosine DeAminases that act on RNA) in human hepatocellular carcinoma. *Gut*, gutjnl-2012.

Chen, L. (2013). Characterization and comparison of human nuclear and cytosolic editomes. *Proceedings of the National Academy of Sciences, 110(29)*, E2741-E2747.

Chen, S. H., Habib, G., Yang, C. Y., Gu, Z. W., Lee, B. R., Weng, S. A., ... Rosseneu, M. (1987). Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science, 238(4825)*, 363-366.

Chester, A., Scott, J., Anant, S., & Navaratnam, N. (2000). RNA editing: cytidine to uridine conversion in apolipoprotein B mRNA. *Biochimica ET Biophysica Acta (BBA)-Gene Structure and Expression, 1494(1-2)*, 1-13.

Conticello, S. G. (2008). The AID/APOBEC family of nucleic acid mutators. *Genome Biology, 9(6)*, 229.

Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258), 561-563.

Daniel, C., Silberberg, G., Behm, M., and Öhman, M. (2014). Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome Biology, 15(2)*, R28.

Deininger, P. (2011). Alu elements: know the SINEs. Genome Biology, 12(12), 236.

Forsberg, E. C., Bhattacharya, D., and Weissman, I. L. (2006). Hematopoietic stem cells. *Stem Cell Reviews, 2(1)*, 23-30.

Freyer, R., Kiefer-Meyer, M. C., & Kössel, H. (1997). Occurrence of plastid RNA editing in all major lineages of land plants. *Proceedings of the National Academy of Sciences, 94(12)*, 6285-6290.

Fumagalli, D., Gacquer, D., Rothé, F., Lefort, A., Libert, F., Brown, D., ... Larsimont, D. (2015). Principles governing A-to-I RNA editing in the breast cancer transcriptome. *Cell Reports, 13(2)*, 277-289.

Goldstein, B., Agranat-Tamir, L., Light, D., Zgayer, O. B. N., Fishman, A., &Lamm, A. T. (2017). A-to-I RNA editing promotes developmental stage–specific gene and lncRNA expression. *Genome Research, 27(3)*, 462-470.

Griffith, M., Miller, C. A., Griffith, O. L., Krysiak, K., Skidmore, Z. L., Ramu, A., ... Demeter, R. T. (2015). Optimizing cancer genome sequencing and analysis. *Cell Systems, 1(3)*, 210-223.

Hamilton, C. E., Papavasiliou, F. N., & Rosenberg, B. R. (2010). Diverse functions for DNA and RNA editing in the immune system. *RNA Biology, 7(2)*, 220-228.

Han, L., Diao, L., Yu, S., Xu, X., Li, J., Zhang, R., ...Li, J. (2015). The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Cell, 28(4)*, 515-528.

Higuchi, M., Maas, S., Single, F. N., Hartner, J., Rozov, A., Burnashev, N., ... Seeburg, P. H. (2000). Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature, 406(6791)*, 78.

Huntley, M. A., Lou, M., Goldstein, L. D., Lawrence, M., Dijkgraaf, G. J., Kaminker, J. S., & Gentleman, R. (2016). Complex regulation of ADAR-mediated RNA-editing across tissues. *BMC Genomics, 17(1)*, 61.

Ju, Y. S., Kim, J. I., Kim, S., Hong, D., Park, H., Shin, J. Y., ...Park, S. S. (2011). Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nature Genetics, 43(8)*, 745.

Keinath, M. C., Timoshevskiy, V. A., Timoshevskaya, N. Y., Tsonis, P. A., Voss, S. R., & Smith, J. J. (2015). Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. *Scientific Reports, 5*, 16413.

Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S., & Gabriel, A. (2004). Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Research, 14(9)*, 1719-1725.

Kiran, A. M., O'mahony, J. J., Sanjeev, K., & Baranov, P. V. (2012). Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Research, 41(D1)*, D258-D261.

Kleinman, C. L., Adoue, V., & Majewski, J. (2012). RNA editing of protein sequences: a rare event in human transcriptomes. Rna, 18(9), 1586-1596.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9(4)*, 357.

Lee, J. H., Ang, J. K., & Xiao, X. (2013). Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA, 19(6)*, 725-732.

Maas, S., Kawahara, Y., Tamburro, K. M., & Nishikura, K. (2006). A-to-I RNA editing and human disease. *RNA Biology, 3(1)*, 1-9.

Mannion, N. M., Greenwood, S. M., Young, R., Cox, S., Brindle, J., Read, D., ...Jantsch, M. F. (2014). The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. *Cell Reports, 9(4)*, 1482-1494.

Markov, A. V., Anisimov, V. A., & Korotayev, A. V. (2010). Relationship between genome size and organismal complexity in the lineage leading from prokaryotes to mammals. *Paleontological Journal, 44(4)*, 363-373.

Melcher, T., Maas, S., Higuchi, M., Keller, W., & Seeburg, P. H. (1995). Editing of α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid receptor GluR-B pre-mRNA in vitro reveals site-selective adenosine to inosine conversion. *Journal of Biological Chemistry, 270(15)*, 8566-8570.

Mirsafian, H., Ripen, A. M., Leong, W. M., Manaharan, T., Mohamad, S. B., & Merican, A. F. (2017). Transcriptome landscape of human primary monocytes at different sequencing depth. *Genomics, 109(5)*, 463-470.

Mitchell, C. J., Getnet, D., Kim, M. S., Manda, S. S., Kumar, P., Huang, T. C., ...Wu, X. (2015). A multi-omic analysis of human naïve CD4+ T cells. *BMC Systems Biology, 9(1)*, 75.

Moreira, S., Valach, M., Aoulad-Aissa, M., Otto, C., & Burger, G. (2016). Novel modes of RNA editing in mitochondria. *Nucleic Acids Research, 44(10)*, 4907-4919.

O'Neil, D., Glowatz, H., & Schlumpberger, M. (2013). Ribosomal RNA Depletion for Efficient Use of RNA‐Seq Capacity. *Current Protocols In Molecular Biology*, 4-19.

Peng, Z., Cheng, Y., Tan, B. C. M., Kang, L., Tian, Z., Zhu, Y., ... Guo, J. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature Biotechnology, 30(3),* 253.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols, 11(9),* 1650.

Picardi, E., & Pesole, G. (2013). REDItools: high-throughput RNA editing detection made easy. *Bioinformatics, 29(14),* 1813-1814.

Picardi, E., D'Erchia, A. M., Lo Giudice, C., & Pesole, G. (2016). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Research, 45(D1),* D750-D757.

Picardi, E., Horner, D. S., & Pesole, G. (2017). Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA, 23(6),* 860-865.

Picardi, E., Manzari, C., Mastropasqua, F., Aiello, I., D'Erchia, A. M., & Pesole, G. (2015). Profiling RNA editing in human tissues: towards the inosinome Atlas. *Scientific Reports, 5,* 14941.

Piskol, R., Ramaswami, G., & Li, J. B. (2013). Reliable identification of genomic variants from RNA-seq data. *The American Journal of Human Genetics, 93(4),* 641-651.

Ramaswami, G., & Li, J. B. (2013). RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research, 42(D1),* D109-D113.

Ramaswami, G., & Li, J. B. (2016). Identification of human RNA editing sites: A historical perspective. *Methods, 107,* 42-47.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P., O'connell, M. A., & Li, J. B. (2013). Identifying RNA editing sites using RNA sequencing data alone. *Nature Methods, 10(2),* 128.

Rayon-Estrada, V., Harjanto, D., Hamilton, C. E., Berchiche, Y. A., Gantman, E. C., Sakmar, T. P., ... Papavasiliou, F. N. (2017). Epitranscriptomic profiling across cell types reveals associations between APOBEC1-mediated RNA editing, gene expression outcomes, and cellular function. *Proceedings of the National Academy of Sciences,* 201714227.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, 74(12),* 5463-5467.

Sharma, S., Patnaik, S. K., Taggart, R. T., Kannisto, E. D., Enriquez, S. M., Gollnick, P., & Baysal, B. E. (2015). APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nature Communications, 6,* 6881.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology, 26(10),* 1135.

Simpson, L., & Shaw, J. (1989). RNA editing and the mitochondrial cryptogenes of kinetoplastid protozoa. *Cell, 57(3),* 355-366.

Smith, H. C., Bennett, R. P., Kizilyer, A., McDougall, W. M., & Prohaska, K. M. (2012, May). Functions and regulation of the APOBEC family of proteins. *In Seminars in Cell & Developmental Biology (Vol. 23, No. 3, Pp. 258-268).* Academic Press.

Soundararajan, R., Stearns, T. M., Griswold, A. J., Mehta, A., Czachor, A., Fukumoto, J., ... Kolliputi, N. (2015). Detection of canonical A-to-G editing events at 3′ UTRs and microRNA target sites in human lungs using next-generation sequencing. *Oncotarget, 6(34),* 35726.

Tang, W., Fei, Y., & Page, M. (2012). Biological significance of RNA editing in cells. *Molecular Biotechnology, 52(1),* 91-100.

Teng, B., Burant, C. F., & Davidson, N. O. (1993). Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science, 260(5115),* 1816-1819.

Tian, S., Yan, H., Kalmbach, M., & Slager, S. L. (2016). Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics, 17(1),* 403.

Ulbricht, R. J., & Emeson, R. B. (2014). One hundred million adenosine‐to‐inosine RNA editing sites: Hearing through the noise. *Bioessays, 36(8),* 730-735.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy‐Moonshine, A., ... Banks, E. (2013). From FastQ data to high‐confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics,* 11-10.

van Gurp, T. P., McIntyre, L. M., & Verhoeven, K. J. (2013). Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS One, 8(12),* e85583.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10(1),* 57.

Yu, W., & Schuster, W. (1995). Evidence for a site-specific cytidine deamination reaction involved in C to U RNA editing of plant mitochondria. *Journal of Biological Chemistry, 270(31),* 18227-18233.

Zanlungo, S., Bégu, D., Quiñones, V., Araya, A., & Jordana, X. (1993). RNA editing of apocytochrome b (cob) transcripts in mitochondria from two genera of plants. *Current Genetics, 24(4),* 344-348.

Zhang, Q., & Xiao, X. (2015). Genome sequence–independent identification of RNA editing sites. *Nature Methods, 12(4),* 347.

Zinshteyn, B., & Nishikura, K. (2009). Adenosine‐to‐inosine RNA editing. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *1(2)*, 202-209.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Manuscript published

Leong, W. M., Ripen, A. M., Mirsafian, H., Mohamad, S. B., & Merican, A. F. (2018). Transcriptogenomics identification and characterization of RNA editing sites in human primary monocytes using high-depth next generation sequencing data. Genomics. DOI: 10.1016/j.ygeno.2018.05.019 (Q2, Impact factor: 3.327)