Detection of Rare Events in Complex Sequencing Data

by

Yifan Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Ryan E. Mills, Chair
Assistant Professor Sue Hammoud
Professor Jun Li
Professor John V. Moran
Associate Professor Maureen Sartor

Yifan Wang

yifwang@umich.edu

ORCID: 0000-0001-8056-9755

I dedicate this thesis to my mother, Ruyuan Wang

Acknowledgements

First and foremost, I would like to thank my mentor, Dr. Ryan E. Mills, for his mentorship, support and great help through my Ph.D. studies on efficient coding, logical thinking, scientific writing and so on. In particular, I appreciate the education he provided for American culture and movie quotes, some of which are 'old-fashioned'. I would also like to thank him for introducing me to a whole different world of video games and his great tips of how to be a better gamer by carrying out a much better performance than me and explaining why he had each move. I truly appreciate Ryan's instruction on all fields he is good at.

To Dr. John V. Moran, I thank for his generous help and suggestions on both somatic SNV identification project and RNA fusion project. I truly appreciate John's support in particular in the Brain Somatic Mosaicism Network consortium.
like to thank Dr. Jeffery M. Kidd for his tremendous help and instruction on the somatic SNV identification project and his excellent suggestion on my thesis.

I would like to thank all my colleagues and friends in Mills lab. I am grateful to Dr. Weichen Zhou for his support, company, great scientific suggestions especially in the big consortium project and of course being a great main tank and teammate in Overwatch. I would like to thank Dr. Gargi Dayama for her encouragement,

comforting and bringing me to great restaurants all the time. I appreciate Dr. Tony Chun for his creativity and humor of various great ideas to mess with the others in lab including me. I truly appreciate the company from Alex Weber, especially the younger generation confidence she brought to the lab when we both did not understand Ryan's "old-fashioned" movie references and of course her excellent organizing ability in both physical lab space and storage space. I would also like to thank Marcus Sherman for the expert level python help. I truly thank Dr. Xuefang Zhao for bringing me into this fantastic lab. Finally, I am thankful to having worked with Wenjin Gu, Chen Sun, Steve Ho, Akima George, Nan Lin, Catherine Barnier and Zhengning Zhang.

I would also like to take the moment and thank all my friends. To my Michigan friends, Christina Vallianatos, Shiqing He, Yajia Zhang, Ziyong Lin, Weixi Zhao, Qiao Mu, Liang Nie, Hongjiu Zhang, Fan Zhang and Feichen Shen, I am thankful for all your help and company. To my online friends, Luxs, Vidalu, Baimu, Vertne, LuDVA, Gilgamesh, Mitsuha, LuckyDog, Ninomiya, LuAshes, Skyrainy and MasterYan, I appreciate all the time that we have spent in different games and the patience that you all have to carry me. To my favorite Overwatch League players, Guxue from Hangzhou Spark, Yveltal and Ameng from Chengdu Hunters, I appreciate their great performance and effort as well as their company when I am not able to play. To my two best friends, Jingjing Wu and Mengge Shan, I love you.

Finally, to my family, thank you all for the support. Although I am across the ocean from the family and missed so much of time to spend with you, I love you all. To my grandparents, Qingben Wang and Yanglian Jin, I truly appreciate all the time that I spent with them and all the love, proud and trust they gave me since I was a baby. To my dear mother, Ruyuan Wang, I am thankful for the love and support she always has for me in my life. It was her who have made me who I am. I would also like to thank my father, Jincheng Wang, for his support for me pursuing my education in the United States. To my parents in law, Yelin Wang and Lu Jiang, I am grateful for the warm family environment they created and all their encouragement. Last but not least, I would like to thank my beloved husband, Hao Wang, for his tremendous understanding, comfort and belief in me even when I was not confident of myself. I cannot accomplish this without him.

Table of Contents

List of Figures

List of Tables

Abstract

The dramatic development of sequencing technologies in the past several decades has made studies of low frequency events in the human cell accessible for the first time. These rare events can occur in the genome and transcriptome and include genetic variation within small populations of somatic cells as well as single molecule level RNA fusion events.

Somatic variations are the mutations that occur during cell division leading to mutations only in a portion of the cells within an individual or tissue. Such somatic variation has been established as a causal feature in various types of cancers. However, somatic mutations in normal human cells and tissues have yet to be well studied. We hence developed our own analysis pipeline to discover somatic SNVs and applied it to human postmortem brain tissues. We then performed amplicon validation in order to compare and validate the somatic SNVs identified from our pipeline. Based on our experience with somatic SNV detection in non-tumor tissues, we have developed a best practice guide to help other researchers. We conclude that it remains difficult to identify low frequency somatic SNVs from bulk sequencing data, however our approach successfully identified a conservative but accurate set of somatic SNVs for future studies.

We next shifted our focus to rare transcriptomic events and sought to identify single molecule level RNA fusion events between U6, a critical component of spliceosome, and other RNAs from five human cell lines. We developed a novel pipeline to target these specific fusion events at the RNA level and differentiate them from integrated genomic chimeras. Using this pipeline, we identified 31 individual U6/L1 fusion events that had strong support as RNA fusion candidates. Together with the biochemical and genetics experiments, we were able to support a plausible mechanism for the formation of U6/L1 pseudogenes in the human genome.

Single cell RNA sequencing further increased the sensitivity to identify rare events at RNA level. However, isoform quantification in single cell sequencing data is not well developed. We then developed Seekmer to perform a better and faster RNA isoform quantification using both bulk and single cell RNA sequencing data. This approach fills the gap between alignment-based methods and the alignment-free methods in performance and run time aspects. With the imputation module of Seekmer to collect information from other single cells with similar expression profiles, we were able to significantly improve the performance of isoform quantification from both simulated data and spike-in data.

Current sequencing technologies contain artifacts that we are unable to fully exclude using computational methods. However, we have demonstrated that with

cautious filtering and collecting extra information from other methods or other cells, we can utilize current methods to study the characteristics and possible functions of rare events in the human genome and transcriptome.

Chapter 1 Discovery of Somatic and Single Molecule Level Events from Complex Sequencing Data

Rare events in human genome and transcriptome

The Human Genome Project has sequenced 3.2 billion base pairs in the human genome (Initial sequencing and analysis of the human genome, 2001; Venter et al., 2001; Finishing the euchromatic sequence of the human genome, 2004). Together with 1000 Genomes Project (Gibbs et al., 2015) and ENCODE project (An integrated encyclopedia of DNA elements in the human genome, 2012), researchers have sequenced various human genomes and transcriptomes. With the large-scale studies on human genomics and transcriptomics in the past decades, we have gained a remarkable progress with understanding the human genome and transcriptome, as well as the discovery of correlated factors with diseases (McCarthy and MacArthur, 2017).

We have gained tremendous knowledge about the human genome and its cellular processes, including but not limited to germline genome variations, differential expression of genes, different isoform splicing in different cell types, and so forth. However, there remain many unknowns related to the human genome and transcriptome. In particular, it is becoming apparent that there exists

a multitude of rare events which do not arise in every cell in the body or even in the same tissue or cell type (Li and Williams, 2013). These biologically rare events include events that only occur in a small number of cells in human body (Poduri et al., 2013), single molecule RNA level events that only occur once per cell (Li et al., 2008), and various alternative splicing patterns and mechanisms in the same tissue for different cells (Porter, Jaamour and Iwase, 2018). There are also population level rare events, for example, rare genetic diseases that only occur in a small percentage of the population (Boycott et al., 2013). From the technical aspect, newly developed technologies such as single cell sequencing generate 'rare' cases from the general population of multiple cells. Thus, studies of different rare events remain challenging with the current technologies but can have a great impact on studies of human diversity and development as well as disease. With the development of new technologies, we have now started to improve our understanding of rare events in human genome and transcriptome.

Illumina sequencing and applications

Illumina sequencing technology

The innovation of first generation sequencing promoted the initiation of the Human Genome Project (Mardis, 2013). The completion of the Human Genome Project revealed the complexity of human genome (Initial sequencing and analysis of the human genome, 2001; Venter et al., 2001), which furthermore advocated the development of faster, lower cost and higher throughput sequencing technologies to generate larger amount of data. With the demand of

higher throughput sequencing platforms, the first high throughput-sequencing platform, as well as the first next-generation sequencing (NGS) platform, 454 sequencing, was released in 2005 (Margulies et al., 2005). Unlike the first generation sequencing methods, NGS is high throughput, generating millions of reads in parallel. Furthermore, NGS also has shorter run time and lower cost compared to the first generation sequencing methods. Over the past decade, next-generation sequencing methods continue to develop and have generated a 100-1,000 fold increase of throughput compared to the first generation sequencing methods (Kircher and Kelso, 2010).

There are two major different sequencing mechanisms involved in NGS: sequencing by synthesis (SBS) and sequencing by ligation (SBL) (Goodwin, McPherson and McCombie, 2016). In SBS NGS approaches, platforms report the DNA sequences by different fluorescent signals released or different changes in ionic concentration from the incorporation of different nucleotides onto the single stranded DNA under synthesizing with a polymerase. One base is sequenced for each signal released. In SBL methods, probes of single stranded DNAs with fluorophore are applied to the single stranded DNAs. The ligation of different complementary probes will release different fluorescence for detection. Multiple bases are sequenced for each signal released (Kircher and Kelso, 2010).

Recently, with highly successful instruments, Illumina sequencing, a method using SBS, has become the most applied NGS approach in research (Goodwin,

McPherson and McCombie, 2016). Since the sequencing technologies are utilized in the three following chapters, I will present more details about Illumina sequencing from library preparation (Figure 1.1 a), cluster amplification (Figure 1.1 b) to DNA sequencing (Figure 1.1 c) (Illumina.com, 2019) as well as limitations in each step.

In library preparation step, DNA or cDNA sample undergoes random fragmentation, followed by 5' and 3' adapter ligation (Figure 1.1 a). These adapters serve as sequencing primer binding sites, indices and regions complementary to the flow cell oligos (Illumina.com, 2019). PCR amplification is then performed using the DNA or cDNA with adapters. The PCR amplification in library preparation guarantees enough DNA material for further sequencing steps, however, possible amplification artifacts could be induced due to the error-prone DNA polymerase utilized (Zook et al., 2014) as well as the biased amplification of different regions in input DNA libraries (Rieber et al., 2013). The amplification artifacts generated in PCR amplification could affect downstream structural variation identification. Furthermore, the PCR amplification step copies the single molecule event at RNA level in the cDNA libraries for multiple copies, which is a major limitation for analysis of single molecule RNA level events.

Cluster generation, the process of isothermally amplifying each fragment molecule, is performed after PCR amplification of the ligated DNA libraries. Each lane of the flow cell contains two types of oligos. Hybridization occurs between

the first of the two types of the oligos and the DNA fragments with complementary adapters. DNA polymerases then create the complementary sequences using the DNA templates. The double stranded molecules are then denatured, and the original DNA templates are washed away. The strands are then clonal amplified through clonal amplification utilizing DNA polymerases. At the end of cluster amplification, clonal amplification of all the DNA template fragments is formed on the flow cell. After clonal amplification, the reverse strands are washed off, leaving only forward strand for future sequencing processes. The clonal amplification step ensures strong enough fluorescent signals released during sequencing procedure (Illumina.com, 2019). However, the amplification error of the polymerase, in the earlier rounds of amplification in particular, could lead to sequencing errors including single nucleotide variation, indels and CNVs in the sequencing data (Kircher, Heyn and Kelso, 2011).

With the amplified clones for each DNA fragment, sequencing begins with the extension of the first sequencing primer to produce the first read. With each cycle, complementary fluorescent-labeled nucleotide to the DNA template would be added to the template. Laser excitation will be performed after the addition of each nucleotide. The emitted fluorescence from each cluster is captured for base identification. The number of cycles determines the length of the read (Illumina.com, 2019). In each cluster, all templates in one cluster are read simultaneously. Hundreds of millions of clusters are sequenced in parallel, which provides a high throughput sequencing output. Although NGS enables much

higher throughput by sequencing great amount of DNA templates simultaneously, the high throughput approach also leads to sequencing artifacts. Image capturing of the flow cell generates one of the artifacts for example. In homopolymer regions, the sequencing artifacts could be generated by both polymerase amplification error and image capturing because of the same signal released from the same cluster in a small period of time. The generation of the same fluorescent signal could be easily mis-counted in the sequencing output data, which lead to a higher sequencing error in the homopolymer regions of the DNA templates (Ivády et al., 2018).

The sequencing process generates millions of reads. Reads from different sequencing libraries are separated based on the different indices added to DNA template fragments. Researchers then align the read sequences in each library to the reference genome for variation identification. Although the efficient high throughput sequencing approaches promoted the variation identification among different individuals, due to the limitation of relatively short read length, complex regions, including repetitive regions, large structural variations, still remain under-studied since the short reads generated could be aligned to multiple places in the genome (Pollard et al., 2018).

10X Genomics sequencing providing haplotype information

10X Genomics could generate synthetic long reads using Illumina sequencing platform. 10X Genomics linked-read sequencing is a barcoded short read

sequencing method. Each longer DNA molecule (~50kb) will be attached to a different bead inside an individual droplet. Inside each droplet, the long DNA fragments got fragmented into small pieces of DNA, attached with a barcode unique for each droplet, and then amplified. The resulting DNA library then is sequenced using standard next generation sequencing methods, for example, Illumina sequencing. After sequencing, the resulting short reads can be reunited with others containing the same barcode to reconstitute the original underlying molecule.

Longranger is a software package developed by 10X Genomics that will both align and call haplotypes using these reconstructed long molecules by using known SNPs that lie within (Figure 1.2). 10X barcoding for short reads could ensure the reads from the same long DNA fragment to be identified as from the same molecule. This could be significantly helpful to exclude PCR duplicates in certain studies. The haplotype information provided by long DNA fragments could also be useful for structural variation discovery. However, with up to 100kb fragment length, 10X Genomics is still limited if the structural variation involves is longer than the fragment length (Goodwin, McPherson and McCombie, 2016).

RNA sequencing

RNA sequencing is a simple application of next generation sequencing. Instead of genomic DNA, RNA reverse transcribed cDNA serves as the input for library preparation (Chu and Corey, 2012). With RNA sequencing, scientists have been

able to study differential gene expression in different cell types, alternative splicing, gene fusion, RNA editing and so forth. The development of RNA sequencing has facilitated a better understanding of the continuously changing human transcriptome in different tissues and in cancer cells.

RNA sequencing library preparation starts from total RNA preparation followed by a ribosomal RNA removal step. The total RNA from a sample is then processed depending on the type of RNA sequencing needed. After reverse transcription of the RNA into cDNA, library preparation and sequencing steps for RNA sequencing remains the same with regular Illumina sequencing processes (Illumina.com, 2019).

RNA sequencing library preparation before reverse transcription could be enriched for mRNA, small RNA, noncoding RNA, microRNA and total RNA (Illumina.com, 2019). Different RNA library preparation could serve for different purposes of studies, for example, downstream analysis on RNAs with noncoding features cannot be performed using mRNA sequencing library, but total RNA sequencing library include all information in the transcriptome of cells at a certain time point.

Although reverse transcription enables the application of NGS on the studies of transcriptome, the artifacts brought by reverse transcription still cannot be neglected. For example, previous studies demonstrated that the template

switching of reverse transcriptase could generate false alternative transcripts, which is a technical artifact of fusion RNAs (Cocquet et al., 2006).

Single cell genomics

The development of single cell sequencing has also improved the study of rare events in genome by providing information at single cell resolution. The technical challenges for single cell genome sequencing exist in cell isolation, whole genome amplification, and analysis of the single cell result. In cell isolation step, each single cell in the tissue needs to be efficiently separated. Four major methods with different advantages and disadvantages are listed in Table 1.1. With different request of throughput, cost and whether targeted cells needed, researchers have applied different cell isolation methods onto different studies (Table 1.1). The inefficient isolation would lead to the mixture of multiple cells for one single cell result.

Following single cell isolation, the DNA libraries from single cells need to be amplified before sequencing. It is challenging to amplify a single copy of genome without artifacts, such as amplification bias, genome loss, mutations or chimeras. Three major methods have been developed to amplify the single cell genome with their own advantages and disadvantages (Table 1.1) (Gawad, Koh and Quake, 2016). The PCR-based methods, DOP-PCR, have resulted in majority genome loss because of the uneven distribution of the common sequences utilized as primers for PCR amplification (Zhang et al., 1992). Although the

amplification product from DOP-PCR distribute relatively uniform across the amplification product, the high error rate of the thermolabile polymerases and high percent of genome loss made DOP-PCR a less preferred whole genome amplification method (Gawad, Koh and Quake, 2016). Multiple displacement amplification (MDA), contrarily, uses isothermal random priming. Although the amplification product tends to be more biased towards to the fragments amplified earlier, the low error rate and low allelic dropout make MDA a better method for single cell DNA amplification (Dean, 2001; Zhang, 2001). Recently, two hybrid methods combining both DOP-PCR and MDA have been developed. These methods utilize MDA for limited cycles, then apply PCR amplifications of the amplicons generated from the isothermal step. The hybrid methods were able to achieve an intermediate amplification error and dropout rate and a lower non-uniformity of the amplification product by combining two kinds of polymerases (Table 1.1). Although various techniques have been developed, amplification of a sequencing library from a single genome still remains biased and error-prone (Gawad, Koh and Quake, 2016).

The amplified sequencing library is then ready for sequencing for either a targeted sequencing (whole exome single cell sequencing) or whole genome sequencing. Even with all technical limitations of sequencing errors, genome drop out and biased amplification, single cell genome sequencing data still serves as a plausible method for identification of rare events at a single cell level, for example, genetic mosaicism in multicellular organisms (Lodato et al., 2015;

McConnell et al., 2017) and genetic heterogeneity in cancer (Hou et al., 2012; Xu et al., 2012).

Single cell RNA sequencing

Researchers have investigated the dynamics of transcriptome in different tissues by using RNA sequencing technology. However, the heterogeneity of transcriptome of the cells in a tissue cannot be assessed from bulk RNA sequencing. The development of single cell RNA sequencing technologies has improved our knowledge of the mosaic transcriptome at single cell level.

Different from single cell DNA library preparation, preparation of single cell RNA sequencing requires the reverse transcription and second-strand synthesis before cDNA library amplification. Researchers have established that only 10-20% of transcripts could be reverse transcribed into the cDNA library for each single cell (Islam et al., 2013). The random loss of majority of transcripts in the reverse transcription step remains an important challenge for the single cell RNA sequencing technologies (Hwang, Lee and Bang, 2018), which creates a rare event at the technical level which will be discussed in chapter 4.

For RNA reverse transcription to cDNA, there are two steps. The first step utilizes an engineered version of the Moloney murine leukemia virus reverse transcriptase for first DNA strand synthesis (Gerard, 2002; Arezi and Hogrefe, 2008). The second step is to synthesize the second strand of cDNA. Two major

approaches could be applied in the second step: poly(A) tailing (Tang et al., 2009; Sasagawa et al., 2013) or template switching mechanism (Islam et al., 2011; Ramsköld et al., 2012). The template switching approach ensures a non-biased amplification and maintains the strand information of the RNAs (Hwang, Lee and Bang, 2018). The cDNA library reverse transcribed from part of the transcriptome is then amplified using the conventional PCR or in vitro transcription. Compared to conventional PCR, in vitro transcription utilizes additional reverse transcription to make a linear amplification of the templates, which could lead to 3' coverage biases (Morris, Singh and Eberwine, 2011).

Multiple technologies for single cell RNA sequencing have been developed since 2009. There are a few major types of single cell RNA sequencing technologies widely performed, including SMART-seq/SMART-seq2 (Ramsköld et al., 2012; Picelli et al., 2014), and UMI-tag based approaches (Islam et al., 2011; Hashimshony et al., 2012; Macosko et al., 2015). While UMI-barcoded are less expensive and could improve the accuracy by removing PCR bias using barcodes, the methods could only sequence the 5' or 3' end of the transcripts (Islam et al., 2011; Hashimshony et al, 2012). Thus, SMART-seq2 (Picelli et al., 2014) with full coverage on each transcript fits better for the purpose of isoform quantification and allele specific gene expression in single cell transcriptome. Most frequently applied single cell RNA sequencing methods are compared in Table 1.3. With all the known technical artifacts of current single cell RNA sequencing technologies, researchers were yet able to apply single cell RNA

sequencing for multiple different purposes, including but not limited to, cell type identification (Ilicic et al., 2016) and cell hierarchy reconstruction (Shin et al., 2015; Habib et al., 2016).

Although Illumina sequencing has greatly promoted the studies of genome and transcriptome from population level, individual level and single cell level, the technical limitations of Illumina sequencing remain as an obstacle for rare event studies, for example, the error rate ~0.1% in average (GLENN, 2011), short read length in complex regions (Pollard et al., 2018), template switching induced false fusion transcripts (Cocquet et al., 2006), uneven amplification of genome or transcriptome majorly in single cell libraries, and so on. Better methods for analysis are required for better studies of genome ad transcriptome, in particular rare events.

Somatic variation in human genome

Somatic mosaicism in cancer and normal cells

Different from germline variations inherited from parental meiosis (Figure 1.3 a, b), somatic variations only are present in a portion of the cells within an individual or tissue. The frequency of somatic variations in a population of cells depends on the time when the mutation occurs in individual development (Figure 1.3 c, d).

One of the rare events that have not been well studied is the somatic variation in non-tumor cells. It has been known that cancer is caused by somatic mutations

(Luzzatto, 2011). Cancer is developed from the clonal expansion of a single mutated cell (Martincorena and Campbell, 2015). The experiment showing that the introduction of DNA fragments from tumor cells to normal cells could lead to malignant transformation and the identification of the mutation in the DNA fragment introduced led to the discovery of the first oncogene, whose mutation can convert the normal cells into cancer (summarized in Stratton, Campbell and Futreal, 2009). Meanwhile, scientists also revealed tumor suppressor gene, whose mutation can inactivate the function of the gene and lead to cancer (Knudson, 1971). The functional variants in the tumor suppressor gene could be either somatic or germline.

DNA damage could be brought by various risk factors including exogenous factors, endogenous factors or enzymes involved in DNA repair or genome editing (Errol et al., 2006). The insertion of certain virus could also cause mutation in oncogenes or tumor suppressor genes leading to cancer (Morales-Sánchez and Fuentes-Pananá, 2014). Somatic mutations in cells then can be obtained from unrepaired DNA damage or incorrectly repaired DNA damage.

Although a Darwinian evolution of positive selection theory has been proposed in 1975 (Cairns, 1975; Nowell, 1976), there is still no conclusive evidence showing how cancer cells progress from a single or multiple driving mutations. There are two possible explanations of how clonal expansion arises: normal cell gains a hypermutation (‘the mutator hypothesis’) (Loeb, Loeb and Anderson, 2003)

and/or the accumulation of early mutations leading to clonal expanded groups of cells acquire a ‘mutator phenotype’ which is more prone to future mutations (Tomlinson, Novelli and Bodmer, 1996). Although the detailed mechanism of how normal cell progressing to cancer cells remains unclear, scientists have demonstrated that cancer cells undergo clonal expansion and positive selections (Stratton, Campbell and Futreal, 2009) that accumulate mutations during cell division (Figure 1.4).

The somatic mosaicism in human body has been proposed 60 years ago (Szilard, 1959), however, we have been treating all cells in a human body with the same genetic content as the single embryonic cell for years for convenience because of the technical limitations. With the development of next generation sequencing (Mardis, 2011; Mardis 2017), until recently, scientists started to investigate the level of mosaicism in a human body and the possible phenotypic variations affected by the mosaicism (Shuga et al., 2010; Gottlieb et al., 2010; Gundry and Vijg, 2012; Biesecker and Spinner, 2013).

The recent studies of somatic mosaicism include single nucleotide somatic variations, copy number variations, structural variations including retrotransposons, and large-scale changes in chromosome status. Although next generation sequencing has provided plenty of data for scientists to investigate somatic mosaicism in human body, the function and phenotypic effect of most somatic variations remain unclear (Pineda-Krch and Lehtila, 2004; Rinkevich,

2004; Strassmann and Queller, 2004; Tuomi, 2004). Further studies to characterize the mechanism and functions of somatic variations are required for better understanding of how tissues with somatic variation function in human body and how they are related with different diseases including primary immune deficiencies, secondary hypertension (Beuschlein et al., 2013) and other diseases.

Somatic mutations in neurons and neurological disorders

Recent studies suggested that brain contains widespread somatic mutations, such as aneuploidy, retrotransposons, large structural variations, as part of the normal development of itself (Rehen, 2005; Muotri and Gage, 2006; Baillie et al., 2011). The long life span of neurons in human brain also provides great opportunities for accumulation of different variations in cells (McConnell et al. 2017).

In order to investigate the pattern of somatic mutations in neurons distributed in human brain, we need to understand the clonal architecture of the brains. Cortical development involves the cell division and differentiation of neuronal progenitor cells. A specific neuronal progenitor cells could transmit the somatic variations harbored in it to all its daughter cells. However, unlike clonal expansion in tumor cells, the long distance, radial migration of the daughter cells (Franco and Müller, 2013) will lead to a mixture of cells from different neuronal progenitor

cells in a nearby region of frontal cortex (Figure 1.5), which adds more difficulties for detection of somatic variations in neurons using bulk sequencing data.

Previous studies showed multiple neurological diseases correlated with somatic variations in neuronal cell. Detailed introduction of somatic variations associated neuronal diseases will be included in chapter 2.

Identification of somatic mutations in cancer and in normal tissues

Given the long history of studying cancer related somatic variations, multiple tools have been developed to discover somatic single nucleotide variations in cancer cells compare to non-cancer cells from next generation sequencing data.

Theoretically, somatic single nucleotide variants could be detected at any allele frequency with enough read depth from next generation sequencing. However, the sequencing errors and library preparation artifacts as well as the effect of other structural variations can cause tremendous false positives in somatic SNV identification. Thus, statistical models have been built to distinguish the real somatic SNVs in cancer cells from sequencing artifacts. Somatic SNV identification involves two steps: alignment of reads, and variant calling from the alignment. For read alignment, tools have been well developed, for example, BWA (Li and Durbin, 2009) and Bowtie (Langmead, 2010). There are different strategies for variant calling using next generation sequencing data. The first is to discover somatic SNVs using paired tumor and normal samples from the same

individual. The second is to discover somatic SNVs using single tumor sample. The paired sample methods identify somatic SNVs in tumor sample only and the single sample methods identify all somatic SNVs in an individual. There are various methods available for somatic SNV identification for paired samples, including MuTect, Strelka, VarScan and so on (Table 1.4). Most of the methods were built on a Bayes' rule to calculate the posterior possibility of the candidate SNV to be a true somatic SNV in cancer cells (Xu, 2018). However, the models and hard filters applied in these methods do not fit the circumstances where we need to identify somatic SNVs in non-tumor tissues without clonal expansion.

Discovery of somatic variations in normal tissues remains challenging with currently available technologies. Without clonal expansion in normal cells, the high error rate of single-cell sequencing as well as other sequencing artifacts in whole genome sequencing makes the discovery of somatic variations in normal tissues without massively accumulated somatic mutations difficult.

In chapter 2 of my thesis, I discussed the difficulties of somatic SNV identification in brain tissue using existing tools. With the validation experiment and experience learnt from manual inspection, we suggested the best practice of somatic SNV identification from human postmortem brain using various types of next generation sequencing data.

The formation of chimeric RNA in human cells and tissues

RNA fusion in cancer cells and normal cells

Fusion transcript is a phenomenon in which parts of two different genes fused into one RNA molecule. The studies of fusion RNAs started from the fusion genes at genomic level correlated with cancers. For example, chimeric genes like BCR-ABL were found in multiple cancers, including hematological cancers (Mitelman, Johansson and Mertens, 2007), prostate cancers (Tomlins, 2005), lung cancers (Soda et al., 2007), breast cancers (Guffanti et al., 2009) and so on (Berger et al., 2010; Frattini et al., 2013).

Recent studies demonstrated that chimeric RNAs can be generated not only from the transcripts of fusion genes (Figure 1.6 a), but also trans splicing (Gingeras, 2009; Li et al., 2009) and cis splicing (Zhang et al., 2012; Qin et al., 2015) (Figure 1.6 b). Scientists showed that these chimeric RNAs could also be discovered from normal human cell lines (Qin et al., 2015) as well as tissues (Carrara et al., 2013; Babiceanu et al., 2016) in addition to cancer cells.

As described above, there are two types of chimeric RNA formations; DNA fusion transcribed chimeric RNAs, and RNA level fusion from trans/cis splicing. The same chimeric RNA could be generated from different mechanism in different cells or tissues. For example, previous studies show that JAZF1-JJAZ1 and PAX3-FOXO1 fusions were generated from chromosomal translocation (DNA-level) in cancer but were generated from trans splicing (RNA-level) in normal cells (Li et al., 2008; Yuan et al., 2013). To understand the function of RNA level

fusion events, more study of mechanisms for RNA level chimeric events is required.

Although there have been studies showing the function relevance of chimeric RNAs (Li et al., 2009), the exact function of chimeric RNAs in human cells and tissues remains unclear. Future studies are required for better understanding of the function of chimeric RNAs.

RNA fusion detection methods

With the development of next generation sequencing, scientists started to investigate the chimeric RNAs in both DNA and RNA sequencing data. The RNA level fusion events could only be discovered from RNA sequencing, while the expressed fusion genes could be identified from both DNA and RNA sequencing. All the chimeric RNA detection methods rely on seeking information for paired end reads mapped to two different genes or single end reads with two split parts mapped to two different genes (Table 1.5).

However, there are still limitations for the detection of RNA fusion events. It was reported that only very few overlaps could be found among different fusion detection methods, representing high false positive rates and lack of validation of these methods (Liu et al., 2015; Kumar et al., 2016). Another limitation of all the existing fusion detection methods is reads from highly similar sequences, for example, paralogous genes, repetitive sequences are all filtered out for a better

accuracy of chimeric RNA identification. However, in our study in chapter 3, we demonstrated a new mechanism of the chimeric RNA formation between L1 and U6 sequences.

RNU6 snRNA plays an important role in RNA splicing

U6 snRNA is a key component of spliceosome. The spliceosome, an intricate machine responsible for RNA splicing, is composed of five ribonucleoprotein (RNP) subunits (U1, U2, U4, U5, U6 and their associated proteins), along with a host of associated protein co-factors (Jurica et al. 2003; Wahl et al. 2009; Will et al. 2010; Matera et al. 2014). Multiple evidence show that U6 snRNA plays a role in the catalytic center of the spliceosome (Didychuk et al. 2018). For example, crosslinking and genetics studies have showed that the strictly conserved "ACAGA-box" sequence of U6 snRNA pairs with the intron 5' splice site in the active spliceosome (Sawa et al. 1992; Sontheimer et al. 1993). Furthermore, biochemical experiments have showed that U6 snRNA is responsible for coordinating with the magnesium ions required for splicing chemistry (Yean et al. 2000; Fica et al. 2013). The fact that U6 sequence is highly conserved in evolution (Brow et al. 1988) also shows the critical role of U6 snRNAs.

U6 snRNA undergoes multiple modifications after transcription that likely contributes to its function in the spliceosome (Table 3.1). Although various studies have identified different U6 snRNA modifications, the function and timing of the modifications are still unclear (Didychuk et al. 2018). Among the multiple

modifications on U6 snRNA, the major form of U6 snRNA in humans ends in a five base polyuridine [poly(U)] tract and a terminal 2',3'-cyclic phosphate group (Lund et al. 1992) (Table 3.1). The enzyme responsible for the formation of the terminal 2',3'-cyclic phosphate group is Usix biogenesis protein 1 (Usb1) (Mroczek et al. 2012; Shchepachev et al. 2012; Hilcenko et al. 2013). Mutations in Usb1 are associated with the disease poikiloderma with neutropenia in human (Mroczek et al. 2013).

There are over 900 copies of U6 distributed in the human genome. Most of these 900 copies are U6 pseudogenes that are not expressed (Doucet et al. 2015), however, there are still at least 4 copies of active identical U6's in human genome (Domitrovich et al. 2003). The presence of multiple active copies of U6 of varying transcriptional activities in human genome has complicated the studies of each individual copy of U6. The different modification, function and localization for different copies of U6 in human genome remain unclear (Didychuk et al. 2018).

Long INterspersed Element-1 (LINE-1 or L1) is a critical component of human genome

Long INterspersed Element-1 (LINE-1 or L1) and L1-derived sequences account for ~17% of human genomic DNA (Lander et al. 2001). L1 is the only known human autonomous non-Long Terminal Repeat (non-LTR) retrotransposon, which means that L1 can move its own sequence around in genome. L1 'jumps'

in the genome through a ‘copy and paste’ mechanism termed retrotransposition. Most of the L1-derived sequences in the genome cannot move to new genomic locations because of 5’-trunction, insertions and/or deletions (indels) or single nucleotide variations in the sequences (Grimaldi et al, 1983; Scott et al. 1987; Lander et al., 2001). Among all L1-derived sequences in human genome, there is ~80-100 retrotransposition-competent L1’s present in an average human genome (Sassaman et al. 1997; Brouha et al. 2003; Beck et al. 2010).

As the only self-autonomous transposable element in human genome, the mobilization of L1 and L1 mediated mobile elements is a driving force in the dynamic nature of the human genome in evolution. The diversity of L1 insertions has also created extensive diversity in different individuals and populations (Ewing et al. 2010; Huang et al. 2010). The insertion of L1’s could also lead to various diseases. To date, scientists have identified over 130 diseases related to the pathogenic retrotransposition events mediated by L1 (Hancks et al. 2011; Kazazian and Moran, 2017). Germline and somatic L1 insertions gained during development is a mutagenesis highly correlated with various diseases, including neurological diseases (Richardson et al. 2014) and various cancers (Iskow et al. 2010; Burns et al. 2017; Scott et al. 2017).

Full length L1s are ~6000 base pairs long, containing a 5’UTR, two open reading frames (ORF1, ORF2) separated by a 63-base inter-ORF spacer, and a 3’UTR with a polyadenosine-rich (poly(A)) tract (Scott et al. 1987; Dombroski et al.

1991). The full length L1 with all elements above is required for efficient L1 retrotransposition (Feng et al. 1996; Moran et al. 1996; Doucet et al. 2015). Human L1 ORF1 encodes a ~40kDa RNA binding protein (ORFp1). ORFp1 is a nucleic acid binding protein and also has nucleic acid chaperone activity (Hohjoh et al, 1996; Moran et al., 1996; Kolosha et al, 1997; Martin and Bushman, 2001; Basame et al., 2006; Januszyk et al., 2007; Khazina et al, 2009). Human L1 ORF2 encodes a ~150kDa protein (ORFp2) (Ergun et al., 2004; Doucet et al., 2010). ORFp2 contains enzymatic activities responsible for both endonuclease (EN) and reverse transcription (RT) (Feng et al., 1996; Mathias et al., 1991).

U6/L1 chimeric sequences are present in the human genome

L1-encoded proteins can mobilize other RNAs (for example, Short INterspersed Element (SINE) (Dewannieux et al., 2003; Hancks et al., 2011; Raiz et al., 2011), non-coding RNAs (Buzdin et al., 2003; Buzdin et al., 2002; Garcia-Perez et al., 2007; Gilbert et al., 2005) and messenger RNAs (Esnault et al., 2000; Wei et al., 2001) in the human genome. Previous studies suggest that ~35% of full length U6's in genome are chimeric sequences with L1's among the 161 full length U6 pseudogenes in human genome (Buzdin et al. 2003). The mechanism of the formation of U6/L1 chimeric sequences in the genome, however, remains unclear.

Although RNA-sequencing has been well developed for years, the computational analysis for discovery of chimeric sequences between repetitive elements is lacking. RNA fusion detection methods, like STAR fusion (Haas et al. 2017) or

Tophat-Fusion (Kim et al. 2011), were designed for detection of fusion genes in RNA-seq data. However, repetitive sequences like L1 or U6 are excluded from the analysis for existing methods because of the low mapping quality for reads mapped to multiple places in reference sequence.

In chapter 3 of my thesis, I have developed a computational method to identify supportive reads for U6/L1 chimeric transcripts as well as for U6 and all other RNA chimeric transcripts. The supportive reads we identified from RNA sequencing data provided evidence for the mechanism of formation of U6/L1 pseudogenes in the human genome.

RNA sequencing and single cell RNA sequencing

DNA transcription and RNA modifications

In cells, the information goes from DNA transcribes to RNA then translated to protein (Figure 1.7) (CRICK, 1970). RNA, which serves as the bridge between DNA and protein, undergoes tremendous regulation and modification in every single cell (Gilbert, 1986). Thus, understanding the dynamics and modification of human transcriptome is important.

RNA includes multiple different categories, including messenger RNA, transfer RNA, ribosomal RNA and so on (Clancy, 2008). Messenger RNA (mRNA) carries the information of the protein. mRNA, as a bridge between DNA and protein, plays a critical role in cells. Previous studies demonstrated that mRNA

undergoes various modifications after transcribed from DNA (Gilbert, Bell and Schaening, 2016). Among these modifications, RNA splicing, which allows the same DNA sequence to generate different RNA sequences, is a regulator of development and tissue identity (Baralle and Giudice, 2017).

RNA sequencing quantification methods

With the development of next generation sequencing, we could now quantify the dynamics of isoform expression in different tissues and in different single cells. Multiple isoform quantification tools have been developed over the past decades using next generation sequencing data. There are two major kinds of methods: alignment-based methods and alignment-free methods (Chandramohan et al., 2013; Teng et al., 2016). For alignment-based methods, reads from RNA sequencing were aligned to a reference genome or transcriptome first. Then the tools will quantify the isoform expression based on the number of reads mapped to each isoform. For alignment-free methods, *kmer* searching was performed instead of read alignment. These methods seek *kmers* with certain length from the reads in the reference. With matched reference found, a pseudo-alignment process would be performed to identify if the reads mapped to the isoform/contig. The read count is then utilized to calculate the isoform expression as the alignment-based methods.

There are both advantages and disadvantages for both types of methods. Alignment-based methods are more accurate since accurate alignment is

performed, which could exclude any artifacts from small sequencing errors, SNPs or indels. However, since these methods need to perform alignment before isoform quantification, the run time for these methods are much longer compared to alignment-free methods. Alignment-free methods skipped the alignment step, which significantly improves the speed of algorithms. However, since only exact kmer searching was performed, the performance for alignment-free methods is not as accurate as alignment-based methods.

In chapter 4, we developed an isoform quantification tool, Seekmer, which combines the characters of alignment-based methods and alignment-free methods. We were able to fill the gap between alignment-based and alignment-free methods in accuracy without too much sacrifice on run time.

Single cell RNA sequencing quantification methods

Bulk RNA sequencing is only able to assess the gene expression and isoform dynamics in a population of cells. With the development of single cell sequencing technologies, we now could investigate the cell-to-cell isoform profile difference among different single cells. Although experimental methods for scRNA-seq have been developed well to capture the dynamic of each single cell transcriptome, the bioinformatics tools for analyzing scRNA-seq data remains limited and undeveloped (Hwang, Lee and Bang, 2018). The random dropout of many genes (Wagner, Regev and Yosef, 2016) and biased amplification of certain genes (Bacher and Kendziorski, 2016) (Table 1.2) are the two major challenges for

most of scRNA-seq analysis tools. With relatively small amount if reads in scRNA-seq data compared to bulk RNA-seq data, the normalization models applied to bulk RNA-seq data does not perform well in scRNA-seq.

In chapter 4, we further developed Seekmer with an imputation function to obtain information from cells with similar expression profiles to better quantify the isoform expression in single cells.

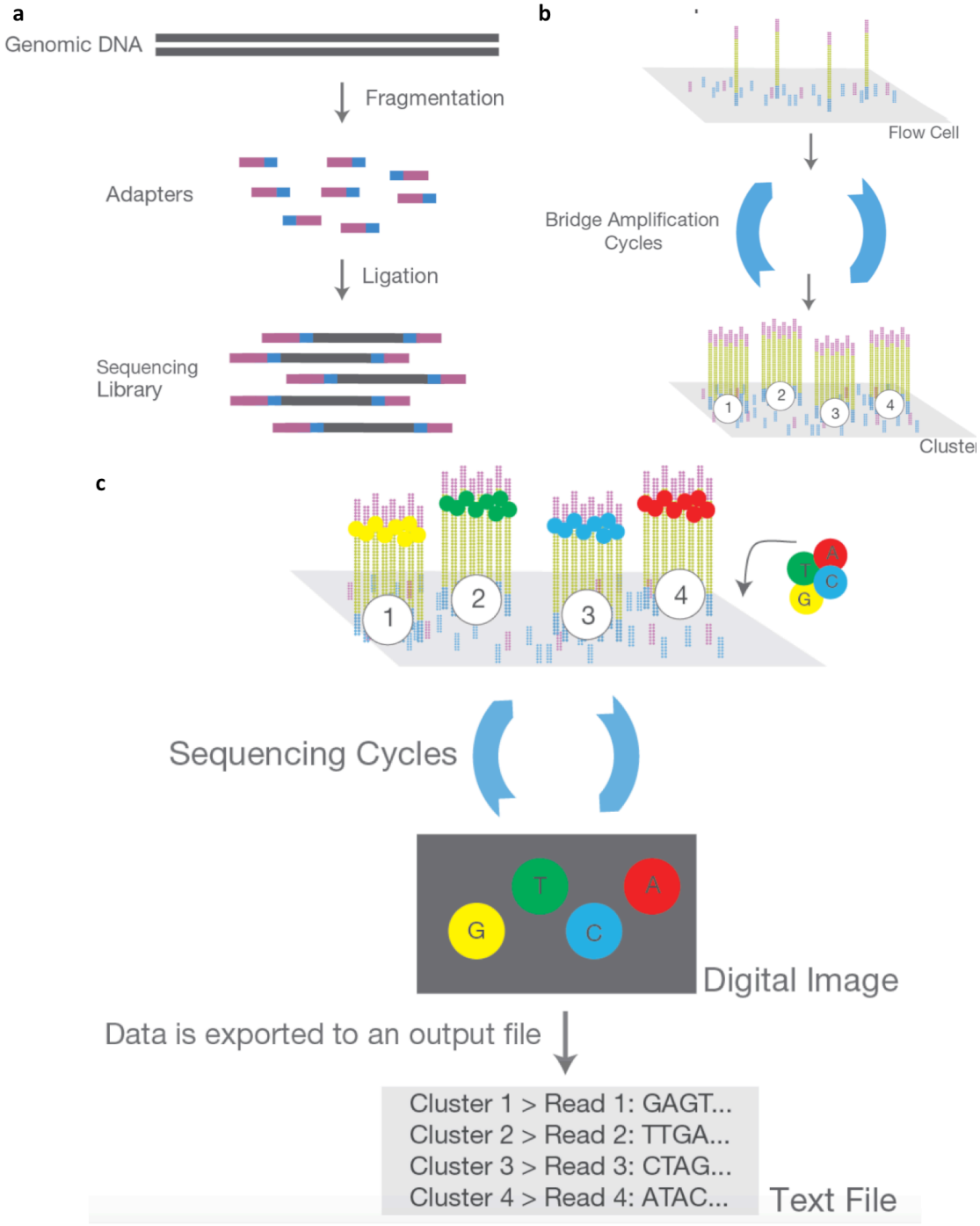

a
Genomic DNA
Fragmentation
Adapters
Ligation
Sequencing
Library
b
Flow Cell
Bridge Amplification
Cycles
1
2
3
4
c
Sequencing Cycles
T
A
G
C
Digital Image
Data is exported to an output file
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
Text File

Figure 1.1: Illumina sequencing technology.
Illumina NGS include 3 steps:
(a) Library preparation: NGS library is prepared by fragmenting a gDNA sample and ligation specialized adapters to both fragment ends.
(b) Cluster amplification: Library is loaded to a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.
(*c) Sequencing:* Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.

*Adapted from: Illumina.com. (2019). [online] Available at: https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf [Accessed 24 Apr. 2019].

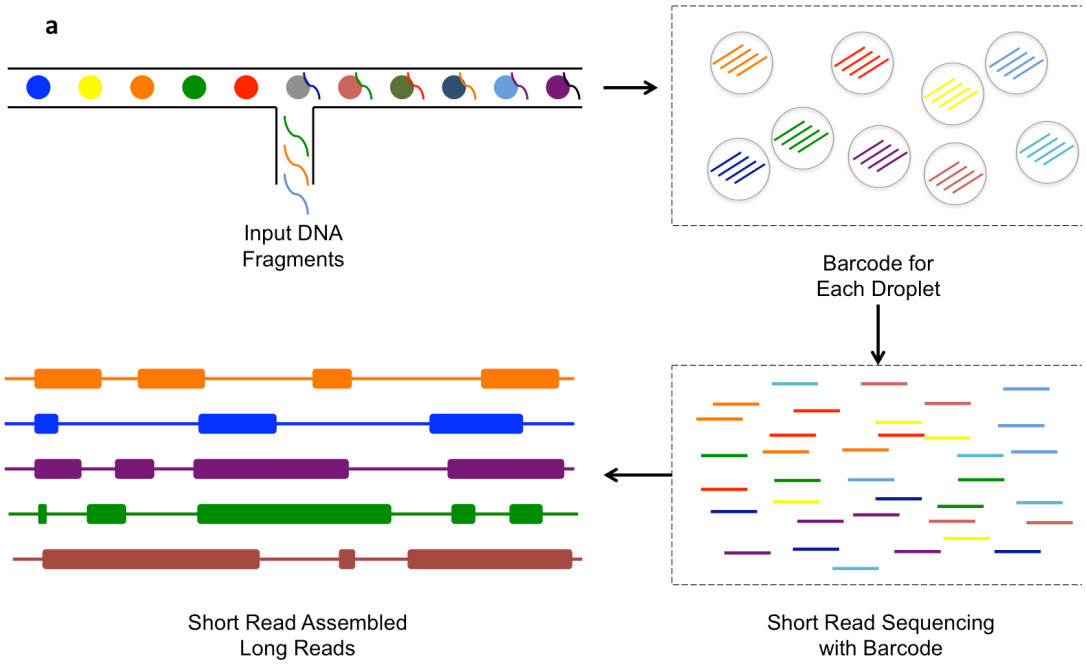


Figure 1.2: Schematic illustration of 10X barcoding work flow.

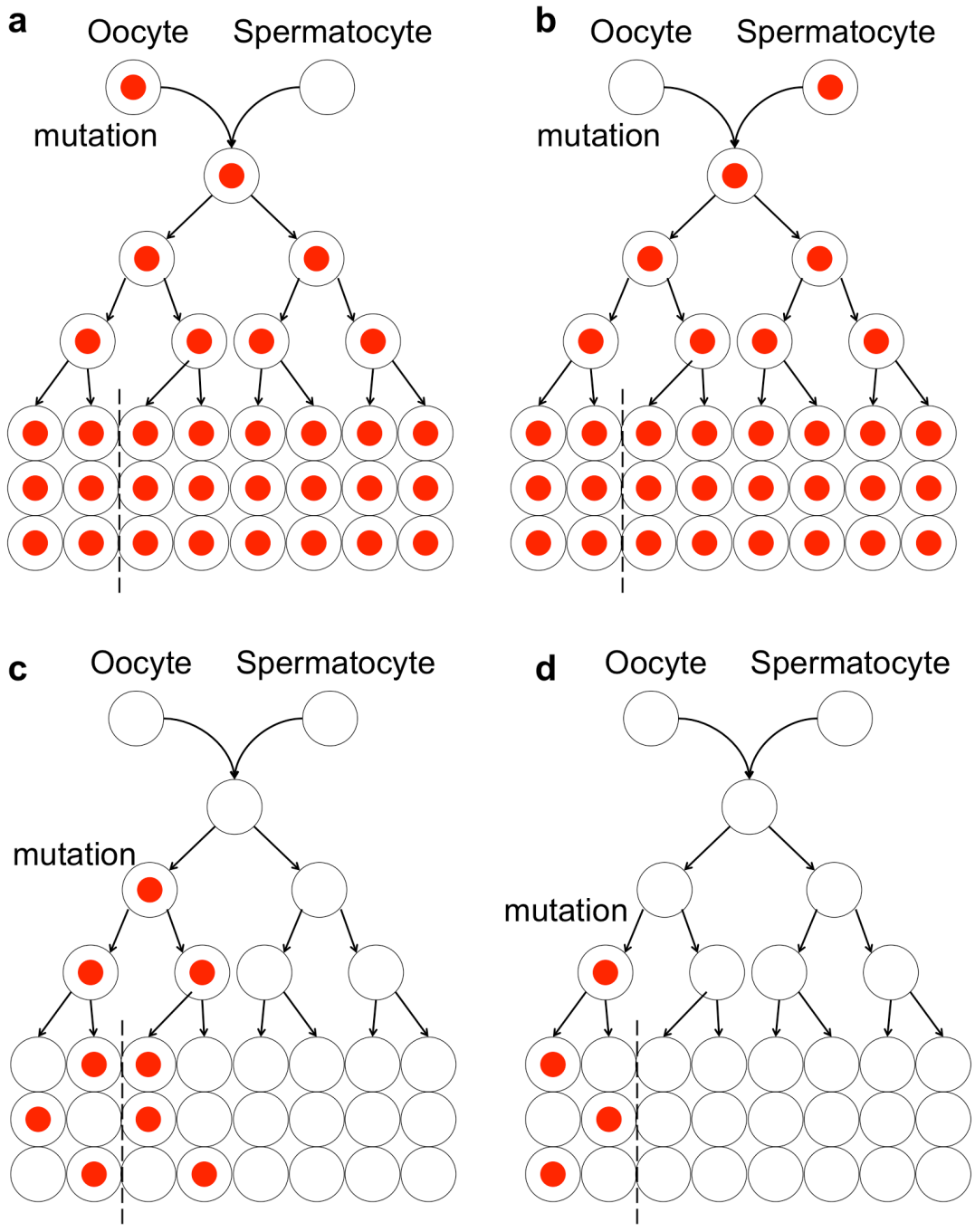

a
Oocyte
Spermatocyte
mutation
b
Oocyte
Spermatocyte
mutation
c
Oocyte
Spermatocyte
mutation
d
Oocyte
Spermatocyte
mutation

Figure 1.3: Schematic illustrations of somatic mutations in human cells and development.
a) Germline mutation from mother's side. Mutation in all cells in the body.
b) Germline mutation from father's side. Mutation in all cells in the body.
c) Somatic mutation happened early in development. Mutation in half of multiple tissues.
d) Somatic mutation happened late in development. Mutation in half of one tissue.

*Adapted from: Poduri, A., Evrony, G., Cai, X. and Walsh, C. (2013). Somatic Mutation, Genomic Variation, and Neurological Disease. Science, 341(6141), p.1237758.

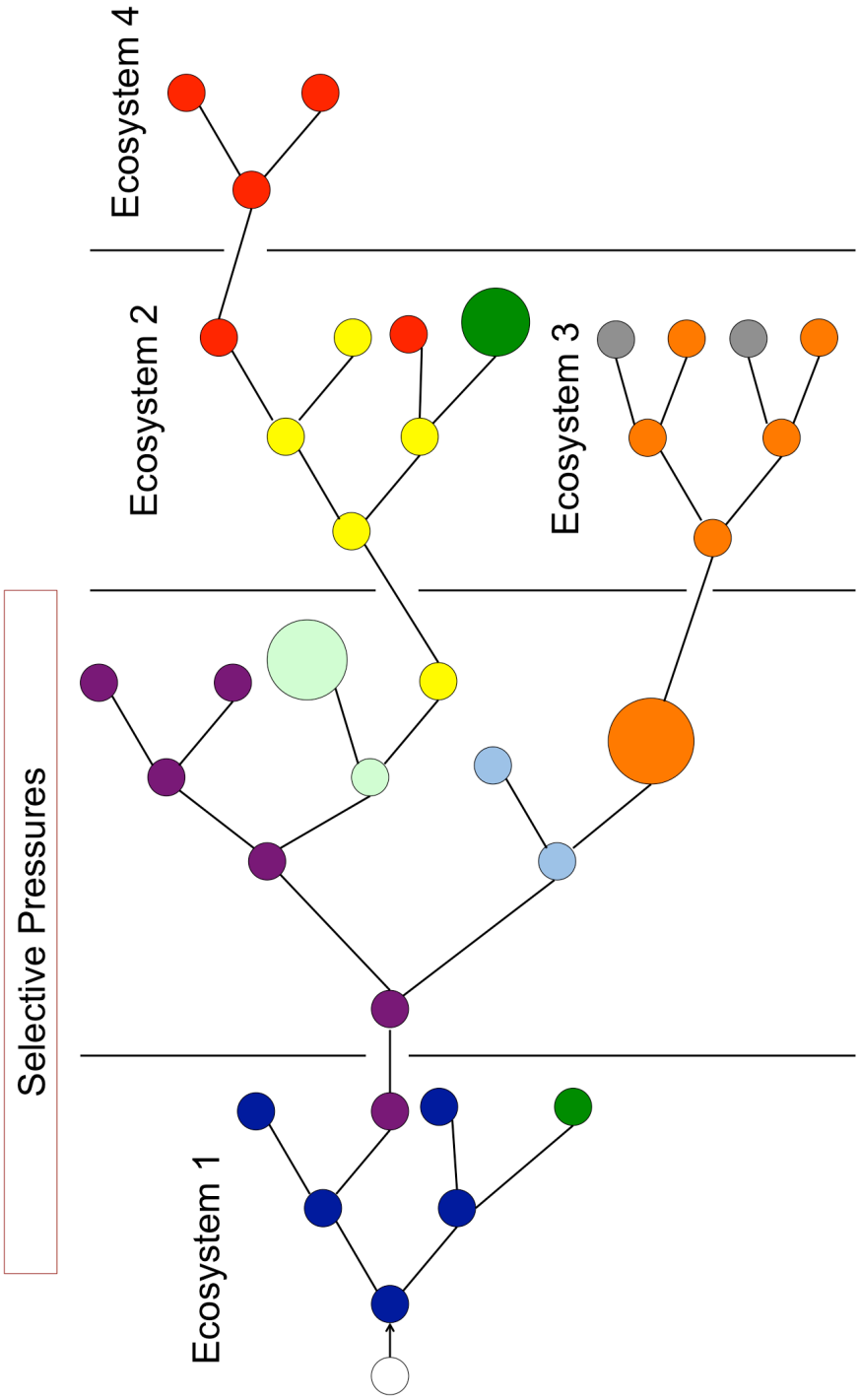


Figure 1.4: Clonal expansion in tumor cells.
*Adapted from: Greaves, M. and Maley, C. (2012). Clonal evolution in cancer. Nature, 481(7381), pp.306-313.

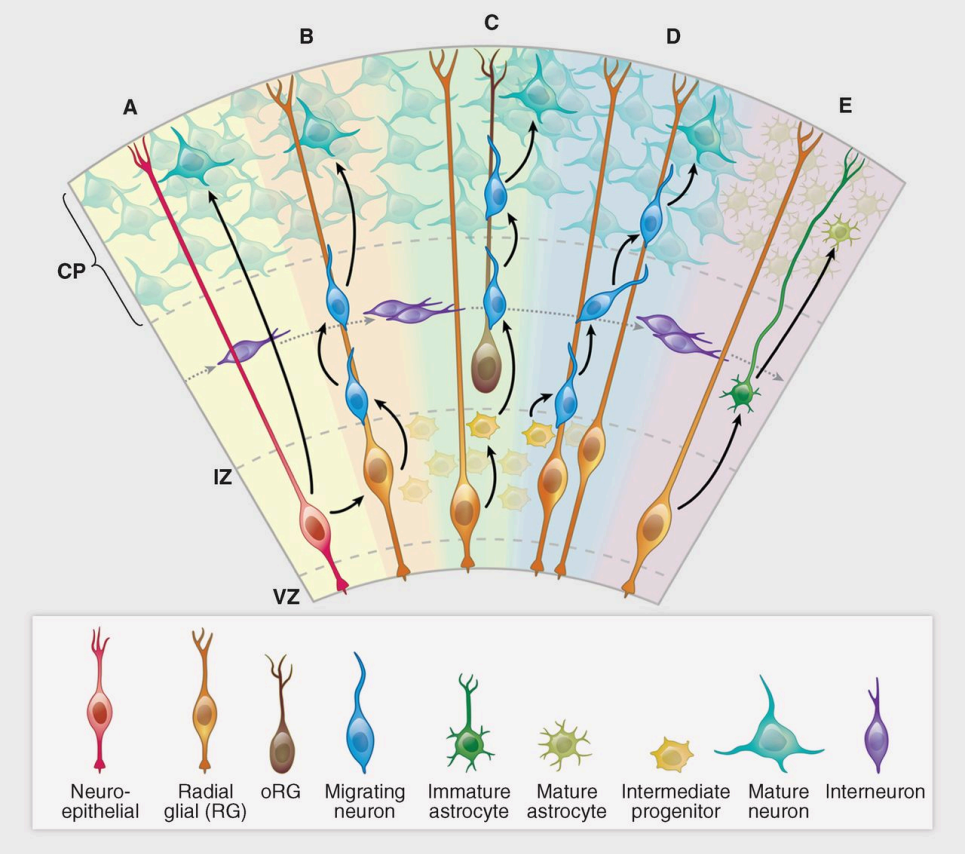


Figure 1.5: Cortical development—origins of pyramidal neurons and astrocytes in the cerebral cortex.
(A) A neuroepithelial cell (red) at the ventricular zone serves as progenitor for both a pyramidal neuron (green-blue) as well as a radial glial cell (gold). (B) A newly differentiated neuron (blue) migrates along a radial glial process. (C) Neurons (blue) continue to migrate as intermediate progenitor cells (small yellow) form. (D) Intermediate progenitor cells begin to generate neurons (blue). (E) The progenitor cells in the ventricular zone begin to give rise to astrocytes (dark green). Interneurons (purple) generated elsewhere migrate tangentially. CP, cortical plate; IZ, intermediate zone; VZ, ventricular zone. The VZ early in development has a thickness of ~10 cell bodies (50 to 100 μm). The CP ranges in thickness from two to three cell bodies at the earliest stages of development, eventually forming a mature cerebral cortex that is 2 to 4 mm thick.

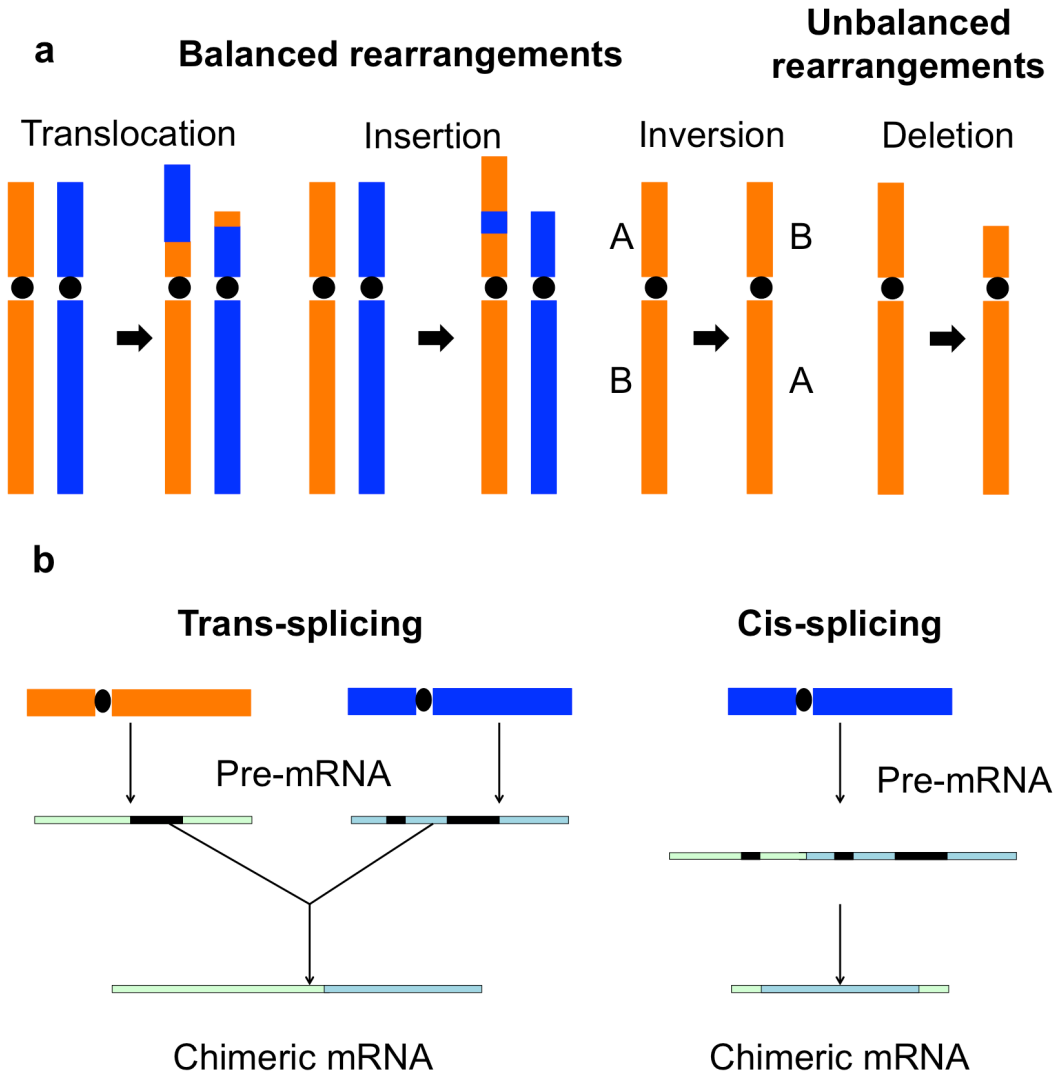


Figure 1.6: Chimeric RNA formation mechanisms.
a) RNA fusions formed at DNA level. At DNA level, fusions form by 'unbalanced' or 'balanced' chromosome rearrangements.
b) RNA fusions formed at RNA level. At RNA level, fusions form by either cis or trans splicing in neighboring genes.

*Adapted from Kumar, S., Razzaq, S., Vo, A., Gautam, M. and Li, H. (2016). Identifying fusion transcripts using next generation sequencing. Wiley Interdisciplinary Reviews: RNA, 7(6), pp.811-823.

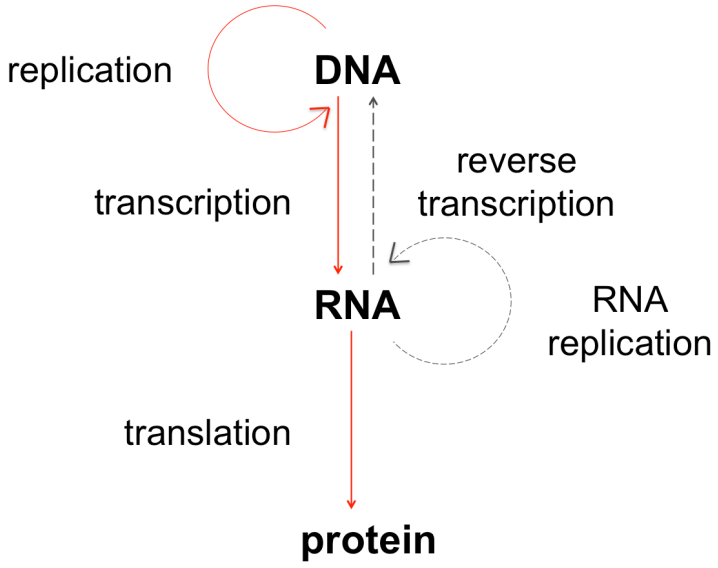


Figure 1.7: Central dogma of information flow in biological systems.

*Adapted from CRICK, F. (1970). Central Dogma of Molecular Biology. Nature, 227(5258), pp.561-563.

Table 1.1 Advantages and disadvantages of common single-cell isolation methods

Method	Unbiased (randomized) or biased (targeted)	Throughput	Cost	Manual or automatic isolation process	Refs
Micromanipulation	Unbiased	Low-throughput	Low	Mainly manually	Kurimoto et al., 2007; Choi et al., 2010; Reizel et al., 2011; Shlush et al., 2012; Zong et al., 2012
Fluorescence-activated cell sorting	Either biased or unbiased	High- throughput	High	Automatic	Dalerba et al., 2011
Laser-capture microdissection	Unbiased	Low- throughput	High	Manually	Bhattacherjee et al., 2004; Frumkin et al., 2008; Yachida et al., 2010
Microfluidics	Unbiased	High- throughput	High	Automatic	Fan et al., 2010; White et al., 2011; Lecault et al., 2012; Wang et al., 2012

Adapted from: Shapiro, E., Biezuner, T. and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics, 14(9), pp.618-630.

.

Table 1.2 Features of different single cell amplification methods

	PCR-based (DOP-PCR)	**Isothermal** (MDA)	**Hybrid** (MALBAC or PicoPLEX)
False-negative rate (coverage and allelic dropout)	High	Low	Intermediate
Non-uniformity	Low	High	Low
False-positive rate (amplicon error rate)	High	Low	Intermediate

Adapted from: Gawad, C., Koh, W. and Quake, S. (2016). Single-cell genome sequencing: current state of the science. Nature Reviews Genetics, 17(3), pp.175-188.

Table 1.3 Comparison of scRNA-seq library preparation methods

Platform	Smart-seq	MARS-seq	CEL-seq	Drop-seq
Region	Full-length	3' end	3' end	3' end
Target read depth (per cell)	10^6	10^4-10^5	10^4-10^5	10^4-10^5
UMI	None	Yes	Yes	Yes
Amplification	PCR	IVT	IVT	PCR
Feature	Isoform analysis	FACS sorting Multiplex barcoding	Linear amplification	Emulsion low cost

Adapted from: Hwang, B., Lee, J. and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine, 50(8).

Table 1.4 Tumor-normal somatic SNV identification tools.

Variant caller	Type of variant	Single-sample mode	Type of core algorithm
BAYSIC (Cantarel et al., 2014)	SNV	No	Machine learning
CaVEMan (Jones et al., 2016)	SNV	No	Joint genotype analysis
deepSNV (Gerstung et al., 2012)	SNV	No	Allele frequency analysis
EBCall (Shiraishi et al., 2013)	SNV, indel	No	Allele frequency analysis
FaSD-somatic (Wang et al., 2014)	SNV	Yes	Joint genotype analysis
FreeBayes (Garrison and Marth, 2012.)	SNV, indel	Yes	Haplotype analysis
HapMuC (Usuyama et al., 2014)	SNV, indel	Yes	Haplotype analysis
JointSNVMix2 (Roth et al., 2012)	SNV	No	Joint genotype analysis
LocHap (Sengupta et al., 2015)	SNV, indel	No	Haplotype analysis
LoFreq (Wilm et al., 2012)	SNV, indel	Yes	Allele frequency analysis
LoLoPicker (Carrot-Zhang and Majewski, 2017)	SNV	No	Allele frequency analysis
MutationSeq (Ding et al., 2011)	SNV	No	Machine learning
MuSE (Fan et al., 2016)	SNV	No	Markov chain model
Mutect (Cibulskis et al., 2013)	SNV	Yes	Allele frequency analysis
SAMtools (Li, 2011)	SNV, indel	Yes	Joint genotype analysis
Platypus (Rimmer et al., 2014)	SNV, indel, SV	Yes	Haplotype analysis
qSNP (Kassahn et al., 2013)	SNV	No	Heuristic threshold
RADIA (Radenbaugh et al., 2014)	SNV	No	Heuristic threshold
Seurat (Christoforides et al., 2013)	SNV, indel, SV	No	Joint genotype analysis
Shimmer (Hansen et al., 2013)	SNV, indel	No	Heuristic threshold
SNooper (Spinella et al., 2016)	SNV, indel	Yes	Machine learning
SNVSniffer (Liu et al., 2016)	SNV, indel	Yes	Joint genotype analysis
SOAPsnv (SOAPsnv)	SNV	No	Heuristic threshold
SomaticSeq (Fang et al., 2015)	SNV	No	Machine learning
SomaticSniper (Larson et al., 2011)	SNV	No	Joint genotype analysis
Strelka (Saunders et al., 2012)	SNV, indel	No	Allele frequency analysis
TVC (TVC)	SNV, indel, SV	Yes	Ion Torrent specific
VarDict (Lai et al., 2016)	SNV, indel, SV	Yes	Heuristic threshold
VarScan2 (Koboldt et al., 2012)	SNV, indel	Yes	Heuristic threshold
Virmid (Kim et al., 2013)	SNV	No	Joint genotype analysis

Adapted from: Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and Structural Biotechnology Journal, 16, pp.15-24.

Table 1.5 Features of RNA fusion detection tools

Tools	Anchor Length Filter	Read Through transcript filter	Supported reads filter	PCR artifact filter	Homology based filter	Alignment tool
Bellerophontes	N	Y	Y	Y	Y	TopHat
BreakFusion	N	N	N	N	N	BWA/BLAT
ChimeraScan	10	Y	4	N	N	Bowtie/BWA
EricScript	N	Y	3/1	Y	Y	BWA/BLAT
FusionAnalyser	Y	Y	Y	N	Y	BWA
FusionCatcher	10	Y	3/1	N	Y	Bowtie/STAR/BLAT/Bowtie2
FusionFinder	N	Y	N	N	Y	Bowtie
FusionHunter	10	Y	3/1	Y	Y	Bowtie
FusionMap	Y	Y	Y	Y	Y	GSNAP
FusionQ	10	N	3/1	N	Y	Bowtie
FusionSeq	N	Y	Y	Y	Y	ELAND
JAFFA	N	Y	3/1	N	Y	Bowtie/BLAT
MapSplice	N	N	N	N	N	Bowtie
deFuse	10	Y	3/1	N	Y	Bowtie/BLAT
SOAPFuse	10	N	3/1	N	N	Soap2/BWA/BLAT
TopHat-Fusion	10	Y	3/1	N	Y	Bowtie
PRADA	N	N	N	N	N	BWA/BLAST
ShortFuse	N	N	Y	N	N	Bowtie
SnowShoes-FTD	N	Y	2/N	Y	Y	Bowtie/BWA

Adapted from: Kumar, S., Razzaq, S., Vo, A., Gautam, M. and Li, H. (2016). Identifying fusion transcripts using next generation sequencing. Wiley Interdisciplinary Reviews: RNA, 7(6), pp.811-823.

Reference

10X Genomics. Longranger Version 2.2.2. Pleasanton, CA. 2018.

Abate, F., Acquaviva, A., Paciello, G., Foti, C., Ficarra, E., Ferrarini, A., Delledonne, M., Iacobucci, I., Soverini, S., Martinelli, G. and Macii, E. (2012). Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. Bioinformatics, 28(16), pp.2114-2121.

Adams, C. and Kron, S. (1997). Method for performing amplification of nucleic acid with two primers bound to a single solid support. US5641658.

An integrated encyclopedia of DNA elements in the human genome. (2012). Nature, 489(7414), pp.57-74.

Arezi, B. and Hogrefe, H. (2008). Novel mutations in Moloney Murine Leukemia Virus reverse transcriptase increase thermostability through tighter binding to template-primer. Nucleic Acids Research, 37(2), pp.473-481.

Asmann, Y., Hossain, A., Necela, B., Middha, S., Kalari, K., Sun, Z., Chai, H., Williamson, D., Radisky, D., Schroth, G., Kocher, J., Perez, E. and Thompson, E. (2011). A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Research, 39(15), pp.e100-e100.

Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y., Lazar, I. and Li, H. (2016). Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Research, 44(6), pp.2859-2872.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biology, 17(1).

Baillie, J., Barnett, M., Upton, K., Gerhardt, D., Richmond, T., De Sapio, F., Brennan, P., Rizzu, P., Smith, S., Fell, M., Talbot, R., Gustincich, S., Freeman, T., Mattick, J., Hume, D., Heutink, P., Carninci, P., Jeddeloh, J. and Faulkner, G. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature, 479(7374), pp.534-537.

Baralle, F. and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nature Reviews Molecular Cell Biology, 18(7), pp.437-451.

Basame, S., Wai-lun Li, P., Howard, G., Branciforte, D., Keller, D. and Martin, S. (2006). Spatial Assembly and RNA Binding Stoichiometry of a LINE-1 Protein Essential for Retrotransposition. Journal of Molecular Biology, 357(2), pp.351-357.

Beck, C., Collier, P., Macfarlane, C., Malig, M., Kidd, J., Eichler, E., Badge, R. and Moran, J. (2010). LINE-1 Retrotransposition Activity in Human Genomes. Cell, 141(7), pp.1159-1170.

Benelli, M., Pescucci, C., Marseglia, G., Severgnini, M., Torricelli, F. and Magi, A. (2012). Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics, 28(24), pp.3232-3239.

Berger, M., Levin, J., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L., Robinson, J., Verhaak, R., Sougnez, C., Onofrio, R., Ziaugra, L., Cibulskis, K., Laine, E., Barretina, J., Winckler, W., Fisher, D., Getz, G., Meyerson, M., Jaffe, D., Gabriel, S., Lander, E., Dummer, R., Gnirke, A., Nusbaum, C. and Garraway, L. (2010). Integrative analysis of the melanoma transcriptome. Genome Research, 20(4), pp.413-427.

Beuschlein, F., Boulkroun, S., Osswald, A., Wieland, T., Nielsen, H., Lichtenauer, U., Penton, D., Schack, V., Amar, L., Fischer, E., Walther, A., Tauber, P., Schwarzmayr, T., Diener, S., Graf, E., Allolio, B., Samson-Couterie, B., Benecke, A., Quinkler, M., Fallo, F., Plouin, P., Mantero, F., Meitinger, T., Mulatero, P., Jeunemaitre, X., Warth, R., Vilsen, B., Zennaro, M., Strom, T. and Reincke, M. (2013). Somatic mutations in ATP1A1 and ATP2B3 lead to aldosterone-producing adenomas and secondary hypertension. Nature Genetics, 45(4), pp.440-444.

Bhattacherjee, V., Mukhopadhyay, P., Singh, S., Roberts, E., Hackmiller, R., Greene, R. and Pisano, M. (2004). Laser capture microdissection of fluorescently labeled embryonic cranial neural crest cells. genesis, 39(1), pp.58-64.

Biesecker, L. and Spinner, N. (2013). A genomic view of mosaicism and human disease. Nature Reviews Genetics, 14(5), pp.307-320.

Brouha, B., Schustak, J., Badge, R., Lutz-Prigge, S., Farley, A., Moran, J. and Kazazian, H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proceedings of the National Academy of Sciences, 100(9), pp.5280-5285.

Brow, D. and Guthrie, C. (1988). Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. Nature, 334(6179), pp.213-218.

Boycott, K., Vanstone, M., Bulman, D. and MacKenzie, A. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature Reviews Genetics, 14(10), pp.681-691.

Burns, K. (2017). Transposable elements in cancer. Nature Reviews Cancer, 17(7), pp.415-424.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y. and Sverdlov, E. (2002). A New Family of Chimeric Retrotranscripts Formed by a Full Copy of U6 Small Nuclear RNA Fused to the 3′ Terminus of L1. Genomics, 80(4), pp.402-406.

Buzdin, A. (2003). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Research, 31(15), pp.4385-4390.

Cairns, J. (1975). Mutation selection and the natural history of cancer. Nature, 255(5505), pp.197-200.

Cantarel, B., Weaver, D., McNeill, N., Zhang, J., Mackey, A. and Reese, J. (2014). BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics, 15(1), p.104.

Carrara, M., Beccuti, M., Cavallo, F., Donatelli, S., Lazzarato, F., Cordero, F. and Calogero, R. (2013). State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues?. BMC Bioinformatics, 14(S7).

Carrot-Zhang, J. and Majewski, J. (2017). LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. Oncotarget, 8(23).

Chandramohan, R., Po-Yen Wu, Phan, J. and Wang, M. (2013). Benchmarking RNA-Seq quantification tools. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Chen, K., Wallis, J., Kandoth, C., Kalicki-Veizer, J., Mungall, K., Mungall, A., Jones, S., Marra, M., Ley, T., Mardis, E., Wilson, R., Weinstein, J. and Ding, L. (2012). BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. Bioinformatics, 28(14), pp.1923-1924.

Choi, J., Ogunniyi, A., Du, M., Du, M., Kretschmann, M., Eberhardt, J. and Love, J. (2010). Development and optimization of a process for automated recovery of

single cells identified by microengraving. Biotechnology Progress, 26(3), pp.888-895.

Christoforides, A., Carpten, J., Weiss, G., Demeure, M., Von Hoff, D. and Craig, D. (2013). Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. BMC Genomics, 14(1), p.302.

Chu, Y. and Corey, D. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. Nucleic Acid Therapeutics, 22(4), pp.271-274.

Cibulskis, K., Lawrence, M., Carter, S., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology, 31(3), pp.213-219.

Clancy, S. (2008) RNA Functions. Nature Education 1(1):102.

Cocquet, J., Chong, A., Zhang, G. and Veitia, R. (2006). Reverse transcriptase template switching and false alternative transcripts. Genomics, 88(1), pp.127-131.

CRICK, F. (1970). Central Dogma of Molecular Biology. Nature, 227(5258), pp.561-563.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P., Rothenberg, M., Leyrat, A., Sim, S., Okamoto, J., Johnston, D., Qian, D., Zabala, M., Bueno, J., Neff, N., Wang, J., Shelton, A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. and Quake, S. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nature Biotechnology, 29(12), pp.1120-1127.

Dean, F. (2001). Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. Genome Research, 11(6), pp.1095-1099.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nature Genetics, 35(1), pp.41-48.

Didychuk, A., Butcher, S. and Brow, D. (2018). The life of U6 small nuclear RNA, from cradle to grave. RNA, 24(4), pp.437-460.

Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M., Condon, A., Aparicio, S. and Shah, S. (2011). Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. Bioinformatics, 28(2), pp.167-175.

Dombroski, B., Mathias, S., Nanthakumar, E., Scott, A. and Kazazian, H. (1991). Isolation of an active human transposable element. Science, 254(5039), pp.1805-1808.

Domitrovich, A. (2003). Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. Nucleic Acids Research, 31(9), pp.2344-2352.

Doucet, A., Hulme, A., Sahinovic, E., Kulpa, D., Moldovan, J., Kopera, H., Athanikar, J., Hasnaoui, M., Bucheton, A., Moran, J. and Gilbert, N. (2010). Characterization of LINE-1 Ribonucleoprotein Particles. PLoS Genetics, 6(10), p.e1001150.

Doucet, A., Droc, G., Siol, O., Audoux, J. and Gilbert, N. (2015). U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. Molecular Biology and Evolution, 32(7), pp.1815-1832.

Esnault, C., Maestre, J. and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nature Genetics, 24(4), pp.363-367.

Ergün, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Strätling, W. and Schumann, G. (2004). Cell Type-specific Expression of LINE-1 Open Reading Frames 1 and 2 in Fetal and Adult Human Tissues. Journal of Biological Chemistry, 279(26), pp.27753-27763.

Errol, C., Roger, A., Wolfram, S., Graham, C., Tom, E. and Richard, D. (2006). DNA Repair and Mutagenesis, Second Edition.

Ewing, A. and Kazazian, H. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Research, 20(9), pp.1262-1270.

Fan, H., Wang, J., Potanina, A. and Quake, S. (2010). Whole-genome molecular haplotyping of single cells. Nature Biotechnology, 29(1), pp.51-57.

Fan, Y., Xi, L., Hughes, D., Zhang, J., Zhang, J., Futreal, P., Wheeler, D. and Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biology, 17(1).

Fang, L., Afshar, P., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J., Gibeling, G., Barr, S., Asadi, N., Gerstein, M., Koboldt, D., Wang, W., Wong, W. and Lam, H. (2015). An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biology, 16(1).

Feng, Q., Moran, J., Kazazian, H. and Boeke, J. (1996). Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. Cell, 87(5), pp.905-916.

Finishing the euchromatic sequence of the human genome. (2004). Nature, 431(7011), pp.931-945.

Fica, S., Tuttle, N., Novak, T., Li, N., Lu, J., Koodathingal, P., Dai, Q., Staley, J. and Piccirilli, J. (2013). RNA catalyses nuclear pre-mRNA splicing. Nature, 503(7475), pp.229-234.

Francis, R., Thompson-Wicking, K., Carter, K., Anderson, D., Kees, U. and Beesley, A. (2012). FusionFinder: A Software Tool to Identify Expressed Gene Fusion Candidates from RNA-Seq Data. PLoS ONE, 7(6), p.e39987.

Franco, S. and Müller, U. (2013). Shaping Our Minds: Stem and Progenitor Cell Diversity in the Mammalian Neocortex. Neuron, 77(1), pp.19-34.

Frattini, V., Trifonov, V., Chan, J., Castano, A., Lia, M., Abate, F., Keir, S., Ji, A., Zoppoli, P., Niola, F., Danussi, C., Dolgalev, I., Porrati, P., Pellegatta, S., Heguy, A., Gupta, G., Pisapia, D., Canoll, P., Bruce, J., McLendon, R., Yan, H., Aldape, K., Finocchiaro, G., Mikkelsen, T., Privé, G., Bigner, D., Lasorella, A., Rabadan, R. and Iavarone, A. (2013). The integrated landscape of driver genomic alterations in glioblastoma. Nature Genetics, 45(10), pp.1141-1149.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Harmelin, A., Rechavi, G. and Shapiro, E. (2008). Amplification of multiple genomic loci from single cells isolated by laser micro-dissection of tissues. BMC Biotechnology, 8(1), p.17.

Garcia-Perez, J., Doucet, A., Bucheton, A., Moran, J. and Gilbert, N. (2007). Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome Research, 17(5), pp.602-611.

Garrison E., Marth G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907.

Gawad, C., Koh, W. and Quake, S. (2016). Single-cell genome sequencing: current state of the science. Nature Reviews Genetics, 17(3), pp.175-188.

Ge, H., Liu, K., Juan, T., Fang, F., Newman, M. and Hoeck, W. (2011). FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics, 27(14), pp.1922-1928.

Gerard, G. (2002). The role of template-primer in protection of reverse transcriptase from thermal inactivation. Nucleic Acids Research, 30(14), pp.3118-3129.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H. and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nature Communications, 3(1).

Gibbs, R., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E., Altshuler, D., Gabriel, S., Gupta, N., Gharani, N., Toji, L., Gerry, N., Resch, A., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M., Amstislavskiy, V., Borodina, T., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M., Mardis, E., Wilson, R., Fulton, L., Fulton, R., Sherry, S., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G., Durbin, R., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J., Davies, C., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C., Kong, Y., Marcketta, A., Gibbs, R., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G., Garrison, E., Kural, D., Lee, W., Fung Leong, W., Stromberg, M., Ward, A., Wu, J., Zhang, M., Daly, M., DePristo, M., Handsaker, R., Altshuler, D., Banks, E., Bhatia, G., del Angel, G., Gabriel, S., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E., McCarroll, S., Nemesh, J., Poplin, R., Yoon, S., Lihm, J., Makarov, V., Clark, A., Gottipati, S., Keinan, A., Rodriguez-Flores, J., Korbel, J., Rausch, T., Fritz, M., Stütz, A., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W., Ritchie, G., Smith, R., Zerbino, D., Zheng-Bradley, X., Sabeti, P., Shlyakhter, I., Schaffner, S., Vitti, J., Cooper, D., Ball, E., Stenson, P., Bentley, D., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E., Batzer, M., Konkel, M., Walker, J., MacArthur, D., Lek, M., Sudbrak, R., Amstislavskiy, V., Herwig, R., Mardis, E., Ding, L., Koboldt, D., Larson, D., Ye, K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Frederick Willems, T., Simpson, J., Shriver, M., Rosenfeld, J., Bustamante, C., Montgomery, S., De La Vega, F., Byrnes, J., Carroll, A., DeGorter, M., Lacroute, P., Maples, B., Martin, A., Moreno-Estrada, A., Shringarpure, S., Zakharia, F., Halperin, E., Baran, Y., Lee, C., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F., Craig, D., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A.,

Sinari, S., Squire, K., Sherry, S., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K., Burchard, E., Hernandez, R., Gignoux, C., Haussler, D., Katzman, S., James Kent, W., Howie, B., Ruiz-Linares, A., Dermitzakis, E., Devine, S., Abecasis, G., Min Kang, H., Kidd, J., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Kate Wing, M., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y., Shi, X., Quitadamo, A., Lunter, G., McVean, G., Marchini, J., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D., Oleksyk, T., Fu, Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Eichler, E., Browning, B., Browning, S., Hormozdiari, F., Sudmant, P., Khurana, E., Durbin, R., Hurles, M., Tyler-Smith, C., Albers, C., Ayub, Q., Balasubramaniam, S., Chen, Y., Colonna, V., Danecek, P., Jostins, L., Keane, T., McCarthy, S., Walter, K., Xue, Y., Gerstein, M., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y., Harmanci, A., Jin, M., Lee, D., Liu, J., Jasmine Mu, X., Zhang, J., Zhang, Y., Li, Y., Luo, R., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G., Garrison, E., Kural, D., Lee, W., Ward, A., Wu, J., Zhang, M., McCarroll, S., Handsaker, R., Altshuler, D., Banks, E., del Angel, G., Genovese, G., Hartl, C., Li, H., Kashin, S., Nemesh, J., Shakir, K., Yoon, S., Lihm, J., Makarov, V., Degenhardt, J., Korbel, J., Fritz, M., Meiers, S., Raeder, B., Rausch, T., Stütz, A., Flicek, P., Paolo Casale, F., Clarke, L., Smith, R., Stegle, O., Zheng-Bradley, X., Bentley, D., Barnes, B., Keira Cheetham, R., Eberle, M., Humphray, S., Kahn, S., Murray, L., Shaw, R., Lameijer, E., Batzer, M., Konkel, M., Walker, J., Ding, L., Hall, I., Ye, K., Lacroute, P., Lee, C., Cerveira, E., Malhotra, A., Hwang, J., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Craig, D., Homer, N., Church, D., Xiao, C., Sebat, J., Antaki, D., Bafna, V., Michaelson, J., Ye, K., Devine, S., Gardner, E., Abecasis, G., Kidd, J., Mills, R., Dayama, G., Emery, S., Jun, G., Shi, X., Quitadamo, A., Lunter, G., McVean, G., Chen, K., Fan, X., Chong, Z., Chen, T., Witherspoon, D., Xing, J., Eichler, E., Chaisson, M., Hormozdiari, F., Huddleston, J., Malig, M., Nelson, B., Sudmant, P., Parrish, N., Khurana, E., Hurles, M., Blackburne, B., Lindsay, S., Ning, Z., Walter, K., Zhang, Y., Gerstein, M., Abyzov, A., Chen, J., Clarke, D., Lam, H., Jasmine Mu, X., Sisu, C., Zhang, J., Zhang, Y., Gibbs, R., Yu, F., Bainbridge, M., Challis, D., Evani, U., Kovar, C., Lu, J., Muzny, D., Nagaswamy, U., Reid, J., Sabo, A., Yu, J., Guo, X., Li, W., Li, Y., Wu, R., Marth, G., Garrison, E., Fung Leong, W., Ward, A., del Angel, G., DePristo, M., Gabriel, S., Gupta, N., Hartl, C., Poplin, R., Clark, A., Rodriguez-Flores, J., Flicek, P., Clarke, L., Smith, R., Zheng-Bradley, X., MacArthur, D., Mardis, E., Fulton, R., Koboldt, D., Gravel, S., Bustamante, C., Craig, D., Christoforides, A., Homer, N., Izatt, T., Sherry, S., Xiao, C., Dermitzakis, E., Abecasis, G., Min Kang, H., McVean, G., Gerstein, M., Balasubramanian, S., Habegger, L., Yu, H., Flicek, P., Clarke, L., Cunningham, F., Dunham, I., Zerbino, D., Zheng-Bradley, X., Lage, K., Berg Jespersen, J., Horn, H., Montgomery, S., DeGorter, M., Khurana, E., Tyler-Smith, C., Chen, Y., Colonna, V., Xue, Y., Gerstein, M., Balasubramanian, S., Fu, Y., Kim, D., Auton, A., Marcketta, A., Desalle, R., Narechania, A., Wilson Sayres, M., Garrison, E., Handsaker, R., Kashin, S., McCarroll, S., Rodriguez-Flores, J., Flicek, P., Clarke,

L., Zheng-Bradley, X., Erlich, Y., Gymrek, M., Frederick Willems, T., Bustamante, C., Mendez, F., David Poznik, G., Underhill, P., Lee, C., Cerveira, E., Malhotra, A., Romanovitch, M., Zhang, C., Abecasis, G., Coin, L., Shao, H., Mittelman, D., Tyler-Smith, C., Ayub, Q., Banerjee, R., Cerezo, M., Chen, Y., Fitzgerald, T., Louzada, S., Massaia, A., McCarthy, S., Ritchie, G., Xue, Y., Yang, F., Gibbs, R., Kovar, C., Kalra, D., Hale, W., Muzny, D., Reid, J., Wang, J., Dan, X., Guo, X., Li, G., Li, Y., Ye, C., Zheng, X., Altshuler, D., Flicek, P., Clarke, L., Zheng-Bradley, X., Bentley, D., Cox, A., Humphray, S., Kahn, S., Sudbrak, R., Albrecht, M., Lienhard, M., Larson, D., Craig, D., Izatt, T., Kurdoglu, A., Sherry, S., Xiao, C., Haussler, D., Abecasis, G., McVean, G., Durbin, R., Balasubramaniam, S., Keane, T., McCarthy, S., Stalker, J., Bodmer, W., Bedoya, G., Ruiz-Linares, A., Cai, Z., Gao, Y., Chu, J., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J., Oleksyk, T., Barnes, K., Mathias, R., Hennis, A., Watson, H., McKenzie, C., Qadri, F., LaRocque, R., Sabeti, P., Zhu, J., Deng, X., Sabeti, P., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Jallow, M., Sisay Joof, F., Corrah, T., Rockett, K., Kwiatkowski, D., Kooner, J., Tịnh Hiê`n, T., Dunstan, S., Thuy Hang, N., Fonnie, R., Garry, R., Kanneh, L., Moses, L., Sabeti, P., Schieffelin, J., Grant, D., Gallo, C., Poletti, G., Saleheen, D. and Rasheed, A. (2015). A global reference for human genetic variation. Nature, 526(7571), pp.68-74.

Gilbert, N., Lutz, S., Morrish, T. and Moran, J. (2005). Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. Molecular and Cellular Biology, 25(17), pp.7780-7795.

Gilbert, W. (1986). Origin of life: The RNA world. Nature, 319(6055), pp.618-618.

Gilbert, W., Bell, T. and Schaening, C. (2016). Messenger RNA modifications: Form, distribution, and function. Science, 352(6292), pp.1408-1412.

Gingeras, T. (2009). Implications of chimaeric non-co-linear transcripts. Nature, 461(7261), pp.206-211.

Gleeson, J. (2000). Classical lissencephaly and double cortex (subcortical band heterotopia): LIS1 and doublecortin. Current Opinion in Neurology, 13(2), pp.121-125.

GLENN, T. (2011). Field guide to next-generation DNA sequencers. Molecular Ecology Resources, 11(5), pp.759-769.

Goodwin, S., McPherson, J. and McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), pp.333-351.

Gottlieb, B., Beitel, L., Alvarado, C. and Trifiro, M. (2010). Selection and mutation in the “new” genetics: an emerging hypothesis. Human Genetics, 127(5), pp.491-501.

Greaves, M. and Maley, C. (2012). Clonal evolution in cancer. Nature, 481(7381), pp.306-313.

Grimaldi, G. and Singer, M. (1983). Members of the KpnI family of long interspersed repeated sequences join and interrupt α-satellite in the monkey genome. Nucleic Acids Research, 11(2), pp.321-338.

Guffanti, A., Iacono, M., Pelucchi, P., Kim, N., Soldà, G., Croft, L., Taft, R., Rizzi, E., Askarian-Amiri, M., Bonnal, R., Callari, M., Mignone, F., Pesole, G., Bertalot, G., Bernardi, L., Albertini, A., Lee, C., Mattick, J., Zucchi, I. and De Bellis, G. (2009). A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics, 10(1), p.163.

Gundry, M. and Vijg, J. (2012). Direct mutation analysis by high-throughput sequencing: From germline to low-abundant, somatic variants. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 729(1-2), pp.1-15.

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J., Hession, C., Zhang, F. and Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science, 353(6302), pp.925-928.

Hancks, D., Goodier, J., Mandal, P., Cheung, L. and Kazazian, H. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. Human Molecular Genetics, 20(17), pp.3386-3400.

Hansen, N., Gartner, J., Mei, L., Samuels, Y. and Mullikin, J. (2013). Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics, 29(12), pp.1498-1503.

Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports, 2(3), pp.666-673.

Hilcenko, C., Simpson, P., Finch, A., Bowler, F., Churcher, M., Jin, L., Packman, L., Shlien, A., Campbell, P., Kirwan, M., Dokal, I. and Warren, A. (2012). Aberrant 3' oligoadenylation of spliceosomal U6 small nuclear RNA in poikiloderma with neutropenia. Blood, 121(6), pp.1028-1038.

Hohjoh, H. and Singer, M. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. The EMBO Journal, 15(3), pp.630-639.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X. and Wang, J. (2012). Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. Cell, 148(5), pp.873-885.

Huang, C., Schneider, A., Lu, Y., Niranjan, T., Shen, P., Robinson, M., Steranka, J., Valle, D., Civin, C., Wang, T., Wheelan, S., Ji, H., Boeke, J. and Burns, K. (2010). Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome. Cell, 141(7), pp.1171-1182.

Hwang, B., Lee, J. and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine, 50(8).

Ilicic, T., Kim, J., Kolodziejczyk, A., Bagger, F., McCarthy, D., Marioni, J. and Teichmann, S. (2016). Classification of low quality cells from single-cell RNA-seq data. Genome Biology, 17(1).

Illumina.com. (2019). [online] Available at: https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf [Accessed 24 Apr. 2019].

Initial sequencing and analysis of the human genome. (2001). Nature, 409(6822), pp.860-921.

Iskow, R., McCabe, M., Mills, R., Torene, S., Pittard, W., Neuwald, A., Van Meir, E., Vertino, P. and Devine, S. (2010). Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. Cell, 141(7), pp.1253-1261.

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J., Lonnerberg, P. and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research, 21(7), pp.1160-1167.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson, S. (2013). Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods, 11(2), pp.163-166.

Ivády, G., Madar, L., Dzsudzsák, E., Koczok, K., Kappelmayer, J., Krulisova, V., Macek, M., Horváth, A. and Balogh, I. (2018). Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. BMC Genomics, 19(1).

Iyer, M., Chinnaiyan, A. and Maher, C. (2011). ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics, 27(20), pp.2903-2904.

Januszyk, K., Li, P., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J., Martin, S. and Clubb, R. (2007). Identification and Solution Structure of a Highly Conserved C-terminal Domain within ORF1p Required for Retrotransposition of Long Interspersed Nuclear Element-1. Journal of Biological Chemistry, 282(34), pp.24893-24904.

Jones, D., Raine, K., Davies, H., Tarpey, P., Butler, A., Teague, J., Nik-Zainal, S. and Campbell, P. (2016). cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. Current Protocols in Bioinformatics, 56(1), pp.15.10.1-15.10.18.

Jurica MS, Moore MJ. (2003). Pre-mRNA splicing: awash in a sea of proteins. Molecular cell, 12(1):5–14.

Kassahn, K., Holmes, O., Nones, K., Patch, A., Miller, D., Christ, A., Harliwong, I., Bruxner, T., Xu, Q., Anderson, M., Wood, S., Leonard, C., Taylor, D., Newell, F., Song, S., Idrisoglu, S., Nourse, C., Nourbakhsh, E., Manning, S., Wani, S., Steptoe, A., Pajic, M., Cowley, M., Pinese, M., Chang, D., Gill, A., Johns, A., Wu, J., Wilson, P., Fink, L., Biankin, A., Waddell, N., Grimmond, S. and Pearson, J. (2013). Somatic Point Mutation Calling in Low Cellularity Tumors. PLoS ONE, 8(11), p.e74380.

Kazazian, H. and Moran, J. (2017). Mobile DNA in Health and Disease. New England Journal of Medicine, 377(4), pp.361-370.

Khazina, E. and Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proceedings of the National Academy of Sciences, 106(3), pp.731-736.

Kircher, M., Heyn, P. and Kelso, J. (2011). Addressing challenges in the production and analysis of illumina sequencing data. BMC Genomics, 12(1).

Kim, D. and Salzberg, S. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biology, 12(8), p.R72.

Kim, S., Jeong, K., Bhutani, K., Lee, J., Patel, A., Scott, E., Nam, H., Lee, H., Gleeson, J. and Bafna, V. (2013). Virmid: accurate detection of somatic mutations with sample impurity inference. Genome Biology, 14(8), p.R90.

Kinsella, M., Harismendy, O., Nakano, M., Frazer, K. and Bafna, V. (2011). Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics, 27(8), pp.1068-1075.

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing - concepts and limitations. BioEssays, 32(6), pp.524-536.

Knudson, A. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. Proceedings of the National Academy of Sciences, 68(4), pp.820-823.

Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L. and Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research, 22(3), pp.568-576.

Kolosha, V. and Martin, S. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proceedings of the National Academy of Sciences, 94(19), pp.10155-10160.

Kumar, S., Razzaq, S., Vo, A., Gautam, M. and Li, H. (2016). Identifying fusion transcripts using next generation sequencing. Wiley Interdisciplinary Reviews: RNA, 7(6), pp.811-823.

Kumar, S., Vo, A., Qin, F. and Li, H. (2016). Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Scientific Reports, 6(1).

Kurimoto, K., Yabuta, Y., Ohinata, Y. and Saitou, M. (2007). Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. Nature Protocols, 2(3), pp.739-752.

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J. and Dry, J. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Research, 44(11), pp.e108-e108.

Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. Nature, 409(6822), pp.860-921.

Langmead, B. (2010). Aligning Short Sequencing Reads with Bowtie. Current Protocols in Bioinformatics.

Larson, D., Harris, C., Chen, K., Koboldt, D., Abbott, T., Dooling, D., Ley, T., Mardis, E., Wilson, R. and Ding, L. (2011). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics, 28(3), pp.311-317.

Lecault, V., White, A., Singhal, A. and Hansen, C. (2012). Microfluidic single cell analysis: from promise to practice. Current Opinion in Chemical Biology, 16(3-4), pp.381-390.

Li, C. and Williams, S. (2013). Human Somatic Variation: It's Not Just for Cancer Anymore. Current Genetic Medicine Reports, 1(4), pp.212-218.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), pp.1754-1760.

Li, H., Wang, J., Mor, G. and Sklar, J. (2008). A Neoplastic Gene Fusion Mimics Trans-Splicing of RNAs in Normal Human Cells. Science, 321(5894), pp.1357-1361.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), pp.2987-2993.

Li, Y., Chien, J., Smith, D. and Ma, J. (2011). FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics, 27(12), pp.1708-1710.

Liu, S., Tsai, W., Ding, Y., Chen, R., Fang, Z., Huo, Z., Kim, S., Ma, T., Chang, T., Priedigkeit, N., Lee, A., Luo, J., Wang, H., Chung, I. and Tseng, G. (2015). Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Research, 44(5), pp.e47-e47.

Liu, Y., Loewer, M., Aluru, S. and Schmidt, B. (2016). SNVSniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations. BMC Systems Biology, 10(S2).

Loeb, L., Loeb, K. and Anderson, J. (2003). Multiple mutations and cancer. Proceedings of the National Academy of Sciences, 100(3), pp.776-781.

Lodato, M., Woodworth, M., Lee, S., Evrony, G., Mehta, B., Karger, A., Lee, S., Chittenden, T., D'Gama, A., Cai, X., Luquette, L., Lee, E., Park, P. and Walsh, C. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science, 350(6256), pp.94-98.

Lund, E. and Dahlberg, J. (1992). Cyclic 2',3'-phosphates and nontemplated nucleotides at the 3' end of spliceosomal U6 small nuclear RNA's. Science, 255(5042), pp.327-330.

Luzzatto, L. (2011). Somatic mutations in cancer development. Environmental Health, 10(Suppl 1), p.S12.

Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A. and McCarroll, S. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell, 161(5), pp.1202-1214.

Mardis, E. (2011). A decade's perspective on DNA sequencing technology. Nature, 470(7333), pp.198-203.

Mardis, E. R. Next-generation sequencing platforms. Annu. Rev. Anal. Chem. (Palo Alto Calif.) 6, 287–303 (2013).

Mardis, E. (2017). DNA sequencing technologies: 2006–2016. Nature Protocols, 12(2), pp.213-218.

Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., Dewell, S., Du, L., Fierro, J., Gomes, X., Godwin, B., He, W., Helgesen, S., Ho, C., Irzyk, G., Jando, S., Alenquer, M., Jarvie, T., Jirage, K., Kim, J., Knight, J., Lanza, J., Leamon, J., Lefkowitz, S., Lei, M., Li, J., Lohman, K., Lu, H., Makhijani, V., McDade, K., McKenna, M., Myers, E., Nickerson, E., Nobile, J., Plant, R., Puc, B., Ronan, M., Roth, G., Sarkis, G., Simons, J., Simpson, J., Srinivasan, M., Tartaro, K., Tomasz, A., Vogt, K., Volkmer, G., Wang, S., Wang, Y., Weiner, M., Yu, P., Begley, R. and Rothberg, J. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), pp.376-380.

Martin, S. and Bushman, F. (2001). Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. Molecular and Cellular Biology, 21(2), pp.467-475.

Martincorena, I. and Campbell, P. (2015). Somatic mutation in cancer and normal cells. Science, 349(6255), pp.1483-1489.

Matera, A. and Wang, Z. (2014). A day in the life of the spliceosome. Nature Reviews Molecular Cell Biology, 15(2), pp.108-121.

Mathias, S., Scott, A., Kazazian, H., Boeke, J. and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science, 254(5039), pp.1808-1810.

McCarthy, M. and MacArthur, D. (2017). Human disease genomics: from variants to biology. Genome Biology, 18(1).

McConnell, M., Moran, J., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J., Fasching, L., Flasch, D., Freed, D., Ganz, J., Jaffe, A., Kwan, K., Kwon, M., Lodato, M., Mills, R., Paquola, A., Rodin, R., Rosenbluh, C., Sestan, N., Sherman, M., Shin, J., Song, S., Straub, R., Thorpe, J., Weinberger, D., Urban, A., Zhou, B., Gage, F., Lehner, T., Senthil, G., Walsh, C., Chess, A., Courchesne, E., Gleeson, J., Kidd, J., Park, P., Pevsner, J. and Vaccarino, F. (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science, 356(6336), p.eaal1641.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., Pacheco, M., Marra, M., Hirst, M., Nielsen, T., Sahinalp, S., Huntsman, D. and Shah, S. (2011). deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. PLoS Computational Biology, 7(5), p.e1001138.

McPherson, A., Wu, C., Hajirasouliha, I., Hormozdiari, F., Hach, F., Lapuk, A., Volik, S., Shah, S., Collins, C. and Sahinalp, S. (2011). Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. Bioinformatics, 27(11), pp.1481-1488.

McPherson, A., Wu, C., Wyatt, A., Shah, S., Collins, C. and Sahinalp, S. (2012). nFuse: Discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Research, 22(11), pp.2250-2261.

Mitelman, F., Johansson, B. and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. Nature Reviews Cancer, 7(4), pp.233-245.

Mirzaa, G., Conti, V., Timms, A., Smyser, C., Ahmed, S., Carter, M., Barnett, S., Hufnagel, R., Goldstein, A., Narumi-Kishimoto, Y., Olds, C., Collins, S., Johnston, K., Deleuze, J., Nitschké, P., Friend, K., Harris, C., Goetsch, A., Martin, B., Boyle, E., Parrini, E., Mei, D., Tattini, L., Slavotinek, A., Blair, E., Barnett, C., Shendure, J., Chelly, J., Dobyns, W. and Guerrini, R. (2015). Characterisation of mutations of the phosphoinositide-3-kinase regulatory subunit, PIK3R2, in perisylvian polymicrogyria: a next-generation sequencing study. The Lancet Neurology, 14(12), pp.1182-1195.

Mirzaa, G., Campbell, C., Solovieff, N., Goold, C., Jansen, L., Menon, S., Timms, A., Conti, V., Biag, J., Olds, C., Boyle, E., Collins, S., Ishak, G., Poliachik, S., Girisha, K., Yeung, K., Chung, B., Rahikkala, E., Gunter, S., McDaniel, S., Macmurdo, C., Bernstein, J., Martin, B., Leary, R., Mahan, S., Liu, S., Weaver, M., Dorschner, M., Jhangiani, S., Muzny, D., Boerwinkle, E., Gibbs, R., Lupski,

J., Shendure, J., Saneto, R., Novotny, E., Wilson, C., Sellers, W., Morrissey, M., Hevner, R., Ojemann, J., Guerrini, R., Murphy, L., Winckler, W. and Dobyns, W. (2016). Association ofMTORMutations With Developmental Brain Disorders, Including Megalencephaly, Focal Cortical Dysplasia, and Pigmentary Mosaicism. JAMA Neurology, 73(7), p.836.

Morales-Sánchez, A. and Fuentes-Pananá, E. (2014). Human Viruses and Cancer. Viruses, 6(10), pp.4047-4079.

Moran, J., Holmes, S., Naas, T., DeBerardinis, R., Boeke, J. and Kazazian, H. (1996). High Frequency Retrotransposition in Cultured Mammalian Cells. Cell, 87(5), pp.917-927.

Morris, J., Singh, J. and Eberwine, J. (2011). Transcriptome Analysis of Single Cells. Journal of Visualized Experiments, (50).

Mroczek, S., Krwawicz, J., Kutner, J., Lazniewski, M., Kucinski, I., Ginalski, K. and Dziembowski, A. (2012). C16orf57, a gene mutated in poikiloderma with neutropenia, encodes a putative phosphodiesterase responsible for the U6 snRNA 3' end modification. Genes & Development, 26(17), pp.1911-1925.

Mroczek, S. and Dziembowski, A. (2013). U6 RNA biogenesis and disease association. Wiley Interdisciplinary Reviews: RNA, 4(5), pp.581-592.

Muotri, A. and Gage, F. (2006). Generation of neuronal variability and complexity. Nature, 441(7097), pp.1087-1093.

Nowell, P. (1976). The clonal evolution of tumor cell populations. Science, 194(4260), pp.23-28.

Piazza, R., Pirola, A., Spinelli, R., Valletta, S., Redaelli, S., Magistroni, V. and Gambacorti-Passerini, C. (2012). FusionAnalyser: a new graphical, event-driven tool for fusion rearrangements discovery. Nucleic Acids Research, 40(16), pp.e123-e123.

Picelli, S., Faridani, O., Björklund, Å., Winberg, G., Sagasser, S. and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols, 9(1), pp.171-181.

Pineda-Krch, M. and Lehtila, K. (2004). Costs and benefits of genetic heterogeneity within organisms. Journal of Evolutionary Biology, 17(6), pp.1167-1177.

Poduri, A., Evrony, G., Cai, X. and Walsh, C. (2013). Somatic Mutation, Genomic Variation, and Neurological Disease. Science, 341(6141), p.1237758.

Pollard, M., Gurdasani, D., Mentzer, A., Porter, T. and Sandhu, M. (2018). Long reads: their purpose and place. Human Molecular Genetics, 27(R2), pp.R234-R241.

Porter, R., Jaamour, F. and Iwase, S. (2018). Neuron-specific alternative splicing of transcriptional machineries: Implications for neurodevelopmental disorders. Molecular and Cellular Neuroscience, 87, pp.35-45.

Qin, F., Song, Z., Babiceanu, M., Song, Y., Facemire, L., Singh, R., Adli, M. and Li, H. (2015). Discovery of CTCF-Sensitive Cis-Spliced Fusion RNAs between Adjacent Genes in Human Prostate Cells. PLOS Genetics, 11(2), p.e1005001.

Radenbaugh, A., Ma, S., Ewing, A., Stuart, J., Collisson, E., Zhu, J. and Haussler, D. (2014). RADIA: RNA and DNA Integrated Analysis for Somatic Mutation Detection. PLoS ONE, 9(11), p.e111516.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Löwer, J., Strätling, W., Löwer, R. and Schumann, G. (2011). The non-autonomous retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery. Nucleic Acids Research, 40(4), pp.1666-1683.

Ramsköld, D., Luo, S., Wang, Y., Li, R., Deng, Q., Faridani, O., Daniels, G., Khrebtukova, I., Loring, J., Laurent, L., Schroth, G. and Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology, 30(8), pp.777-782.

Rehen, S. (2005). Constitutional Aneuploidy in the Normal Human Brain. Journal of Neuroscience, 25(9), pp.2176-2180.

Reizel, Y., Chapal-Ilani, N., Adar, R., Itzkovitz, S., Elbaz, J., Maruvka, Y., Segev, E., Shlush, L., Dekel, N. and Shapiro, E. (2011). Colon Stem Cell and Crypt Dynamics Exposed by Cell Lineage Reconstruction. PLoS Genetics, 7(7), p.e1002192.

Richardson, S., Morell, S. and Faulkner, G. (2014). L1 Retrotransposons and Somatic Mosaicism in the Brain. Annual Review of Genetics, 48(1), pp.1-27.

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., Jäger, N., Kool, M., Taylor, M., Lichter, P., Pfister, S., Wolf, S., Brors, B. and Eils, R. (2013). Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. PLoS ONE, 8(6), p.e66621.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S., Wilkie, A., McVean, G. and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics, 46(8), pp.912-918.

Rinkevich, B. (2004). Will two walk together, except they have agreed? Amos 3:3. Journal of Evolutionary Biology, 17(6), pp.1178-1179.

Rivière, J., Mirzaa, G., O'Roak, B., Beddaoui, M., Alcantara, D., Conway, R., St-Onge, J., Schwartzentruber, J., Gripp, K., Nikkel, S., Worthylake, T., Sullivan, C., Ward, T., Butler, H., Kramer, N., Albrecht, B., Armour, C., Armstrong, L., Caluseriu, O., Cytrynbaum, C., Drolet, B., Innes, A., Lauzon, J., Lin, A., Mancini, G., Meschino, W., Reggin, J., Saggar, A., Lerman-Sagie, T., Uyanik, G., Weksberg, R., Zirn, B., Beaulieu, C., Majewski, J., Bulman, D., O'Driscoll, M., Shendure, J., Graham, J., Boycott, K. and Dobyns, W. (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nature Genetics, 44(8), pp.934-940.

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., Marra, M., Aparicio, S. and Shah, S. (2012). JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics, 28(7), pp.907-913.

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K., Imai, T. and Ueda, H. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biology, 14(4).

Sassaman, D., Dombroski, B., Moran, J., Kimberland, M., Naas, T., DeBerardinis, R., Gabriel, A., Swergold, G. and Kazazian, H. (1997). Many human L1 elements are capable of retrotransposition. Nature Genetics, 16(1), pp.37-43.

Saunders, C., Wong, W., Swamy, S., Becq, J., Murray, L. and Cheetham, R. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics, 28(14), pp.1811-1817.

Sawa, H. and Abelson, J. (1992). Evidence for a base-pairing interaction between U6 small nuclear RNA and 5' splice site during the splicing reaction in yeast. Proceedings of the National Academy of Sciences, 89(23), pp.11269-11273.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D., Rozowsky, J., Tewari, A., Kitabayashi, N., Moss, B., Chee, M., Demichelis, F., Rubin, M. and Gerstein, M. (2010). FusionSeq: a modular framework for finding gene fusions by analyzing Paired-End RNA-Sequencing data. Genome Biology, 11(10), p.R104.

Scott, A., Schmeckpeper, B., Abdelrazik, M., Comey, C., O'Hara, B., Rossiter, J., Cooley, T., Heath, P., Smith, K. and Margolet, L. (1987). Origin of the human L1

elements: Proposed progenitor genes deduced from a consensus DNA sequence. Genomics, 1(2), pp.113-125.

Scott, E. and Devine, S. (2017). The Role of Somatic L1 Retrotransposition in Human Cancers. Viruses, 9(6), p.131.

Sengupta, S., Gulukota, K., Zhu, Y., Ober, C., Naughton, K., Wentworth-Sheilds, W. and Ji, Y. (2015). Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. Nucleic Acids Research, 44(3), pp.e25-e25.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics, 14(9), pp.618-630.

Shchepachev, V., Wischnewski, H., Missiaglia, E., Soneson, C. and Azzalin, C. (2012). Mpn1, Mutated in Poikiloderma with Neutropenia Protein 1, Is a Conserved 3′-to-5′ RNA Exonuclease Processing U6 Small Nuclear RNA. Cell Reports, 2(4), pp.855-865.

Shin, J., Berg, D., Zhu, Y., Shin, J., Song, J., Bonaguidi, M., Enikolopov, G., Nauen, D., Christian, K., Ming, G. and Song, H. (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell, 17(3), pp.360-372.

Shiraishi, Y., Sato, Y., Chiba, K., Okuno, Y., Nagata, Y., Yoshida, K., Shiba, N., Hayashi, Y., Kume, H., Homma, Y., Sanada, M., Ogawa, S. and Miyano, S. (2013). An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Research, 41(7), pp.e89-e89.

Shirley, M., Tang, H., Gallione, C., Baugher, J., Frelin, L., Cohen, B., North, P., Marchuk, D., Comi, A. and Pevsner, J. (2013). Sturge–Weber Syndrome and Port-Wine Stains Caused by Somatic Mutation in GNAQ. New England Journal of Medicine, 368(21), pp.1971-1979.

Shlush, L., Chapal-Ilani, N., Adar, R., Pery, N., Maruvka, Y., Spiro, A., Shouval, R., Rowe, J., Tzukerman, M., Bercovich, D., Izraeli, S., Marcucci, G., Bloomfield, C., Zuckerman, T., Skorecki, K. and Shapiro, E. (2012). Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. Blood, 120(3), pp.603-612.

Shuga, J., Zeng, Y., Novak, R., Mathies, R., Hainaut, P. and Smith, M. (2010). Selected technologies for measuring acquired genetic damage in humans. Environmental and Molecular Mutagenesis, 51(8-9), pp.851-870.

Sicca, F., Kelemen, A., Genton, P., Das, S., Mei, D., Moro, F., Dobyns, W. and Guerrini, R. (2003). Mosaic mutations of the LIS1 gene cause subcortical band heterotopia. Neurology, 61(8), pp.1042-1046.

SOAPsnv. http://soap.genomics.org.cn/SOAPsnv.html.

Soda, M., Choi, Y., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y. and Mano, H. (2007). Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. Nature, 448(7153), pp.561-566.

Sontheimer, E. and Steitz, J. (1993). The U5 and U6 small nuclear RNAs as active site components of the spliceosome. Science, 262(5142), pp.1989-1996.

Spinella, J., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., Ouimet, M., Healy, J. and Sinnett, D. (2016). SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC Genomics, 17(1).

Strassmann, J. and Queller, D. (2004). Genetic conflicts and intercellular heterogeneity. Journal of Evolutionary Biology, 17(6), pp.1189-1191.

Stratton, M., Campbell, P. and Futreal, P. (2009). The cancer genome. Nature, 458(7239), pp.719-724.

Supper, J., Gugenmus, C., Wollnik, J., Drueke, T., Scherf, M., Hahn, A., Grote, K., Bretschneider, N., Klocke, B., Zinser, C., Cartharius, K. and Seifert, M. (2013). Detecting and visualizing gene fusions. Methods, 59(1), pp.S24-S28.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B., Siddiqui, A., Lao, K. and Surani, M. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods, 6(5), pp.377-382.

Teng, M., Love, M., Davis, C., Djebali, S., Dobin, A., Graveley, B., Li, S., Mason, C., Olson, S., Pervouchine, D., Sloan, C., Wei, X., Zhan, L. and Irizarry, R. (2016). A benchmark for RNA-seq quantification pipelines. Genome Biology, 17(1).

Tomlins, S. (2005). Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. Science, 310(5748), pp.644-648.

Tomlinson, I., Novelli, M. and Bodmer, W. (1996). The mutation rate and cancer. Proceedings of the National Academy of Sciences, 93(25), pp.14800-14803.

Troutt, A., McHeyzer-Williams, M., Pulendran, B. and Nossal, G. (1992). Ligation-anchored PCR: a simple amplification technique with single-sided specificity. Proceedings of the National Academy of Sciences, 89(20), pp.9823-9825.

Tuomi, J. (2004). Genetic heterogeneity within organisms and the evolution of individuality. Journal of Evolutionary Biology, 17(6), pp.1182-1183.

TVC. https://github.com/iontorrent/TS.

Usuyama, N., Shiraishi, Y., Sato, Y., Kume, H., Homma, Y., Ogawa, S., Miyano, S. and Imoto, S. (2014). HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. Bioinformatics, 30(23), pp.3302-3309.

Van Raamsdonk, C., Bezrookove, V., Green, G., Bauer, J., Gaugler, L., O'Brien, J., Simpson, E., Barsh, G. and Bastian, B. (2008). Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. Nature, 457(7229), pp.599-602.

Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., Yandell, M., Evans, C., Holt, R., Gocayne, J., Amanatides, P., Ballew, R., Huson, D., Wortman, J., Zhang, Q., Kodira, C., Zheng, X., Chen, L., Skupski, M., Subramanian, G., Thomas, P., Zhang, J., Gabor Miklos, G., Nelson, C., Broder, S., Clark, A., Nadeau, J., McKusick, V., Zinder, N., Levine, A., Roberts, R., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T., Higgins, M., Ji, R., Ke, Z., Ketchum, K., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G., Milshina, N., Moore, H., Naik, A., Narayan, V., Neelam, B., Nusskern, D., Rusch, D., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-

Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J., Guigó, R., Campbell, M., Sjolander, K., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001). The Sequence of the Human Genome. Science, 291(5507), pp.1304-1351.

Wagner, A., Regev, A. and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), pp.1145-1160.

Wahl, M., Will, C. and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell, 136(4), pp.701-718.

Wang, J., Fan, H., Behr, B. and Quake, S. (2012). Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell, 150(2), pp.402-412.

Wang, K., Singh, D., Zeng, Z., Coleman, S., Huang, Y., Savich, G., He, X., Mieczkowski, P., Grimm, S., Perou, C., MacLeod, J., Chiang, D., Prins, J. and Liu, J. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, 38(18), pp.e178-e178.

Wang, W., Wang, P., Xu, F., Luo, R., Wong, M., Lam, T. and Wang, J. (2014). FaSD-somatic: a fast and accurate somatic SNV detection algorithm for cancer genome sequencing data. Bioinformatics, 30(17), pp.2498-2500.

Wei, W., Gilbert, N., Ooi, S., Lawler, J., Ostertag, E., Kazazian, H., Boeke, J. and Moran, J. (2001). Human L1 Retrotransposition: cis Preference versus trans Complementation. Molecular and Cellular Biology, 21(4), pp.1429-1439.

White, A., VanInsberghe, M., Petriv, O., Hamidi, M., Sikorski, D., Marra, M., Piret, J., Aparicio, S. and Hansen, C. (2011). High-throughput microfluidic single-cell RT-qPCR. Proceedings of the National Academy of Sciences, 108(34), pp.13999-14004.

Will, C. and Luhrmann, R. (2010). Spliceosome Structure and Function. Cold Spring Harbor Perspectives in Biology, 3(7), pp.a003707-a003707.

Wilm, A., Aw, P., Bertrand, D., Yeo, G., Ong, S., Wong, C., Khor, C., Petric, R., Hibberd, M. and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research, 40(22), pp.11189-11201.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and Structural Biotechnology Journal, 16, pp.15-24.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y. and Wang, J. (2012). Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. Cell, 148(5), pp.886-895.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R., Eshleman, J., Nowak, M., Velculescu, V., Kinzler, K., Vogelstein, B. and Iacobuzio-Donahue, C. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature, 467(7319), pp.1114-1117.

Yean, S., Wuenschell, G., Termini, J. and Lin, R. (2000). Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. Nature, 408(6814), pp.881-884.

Yorukoglu, D., Hach, F., Swanson, L., Collins, C., Birol, I. and Sahinalp, S. (2012). Dissect: detection and characterization of novel structural alterations in transcribed sequences. Bioinformatics, 28(12), pp.i179-i187.

Yuan, H., Qin, F., Movassagh, M., Park, H., Golden, W., Xie, Z., Zhang, P., Sklar, J. and Li, H. (2013). A Chimeric RNA Characteristic of Rhabdomyosarcoma in Normal Myogenesis Process. Cancer Discovery, 3(12), pp.1394-1403.

Zhang, D. (2001). Ramification amplification: A novel isothermal DNA amplification method. Molecular Diagnosis, 6(2), pp.141-150.

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W. and Arnheim, N. (1992). Whole genome amplification from a single cell: implications for genetic analysis. Proceedings of the National Academy of Sciences, 89(13), pp.5847-5851.

Zhang, Y., Gong, M., Yuan, H., Park, H., Frierson, H. and Li, H. (2012). Chimeric Transcript Generated by cis-Splicing of Adjacent Genes Regulates Prostate Cancer Cell Proliferation. Cancer Discovery, 2(7), pp.598-607.

Zong, C., Lu, S., Chapman, A. and Xie, X. (2012). Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. Science, 338(6114), pp.1622-1626.

Zook, J., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W. and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology, 32(3), pp.246-251.

Chapter 2 Somatic Single Nucleotide Variants Identification in Non-Tumor Samples

This chapter presents a major portion of the computational analysis conducted as part of the Brain Somatic Mosaicism Network common experiment to determine an accurate and scalable method for identifying somatic SNVs from non-tumor tissues. I performed the analysis on all Michigan SNV identification as well as validation data analysis from all 5 different institutes, and I had a large contribution to the overall determination and application of the final best practice methodology.

Introduction

Mechanisms of somatic single nucleotide variations

The human body reaches a steady state of ~10^{14} cells in adulthood (McConnell et al. 2017). During cell division and growth in later development stages, the failure of repair DNA damage during replication, transcription and cellular metabolism (Leibeling, Laspe and Emmert, 2006; WILSONIII and BOHR, 2007; Kanaar, Wyman and Rothstein, 2008) leads to diverse variation within the genomes of individual cells in a monozygotic individual.

In particular, the human neuronal population has an extraordinary diversity compared to other cell types (Muotri and Gage, 2006). The adult human brain contains 86 billion neurons (Azevedo et al., 2009; Herculano-Houzel, 2009) derived from neural stem cells and progenitor cells (NPCs). The tens of billions of cell divisions needed for the generation of neurons have potentially accumulated various somatic mutations in different cell lineages. Furthermore, as one of the longest-living cell types in human body, somatic mutations accumulated in NPCs as well as post-mitotic neurons can have a dramatic effect on neuronal development and function (Muotri and Gage, 2006; Bushman and Chun, 2013).

Previous studies have showed that multiple neurological diseases are associated with somatic variations in neuronal cell. The level of mosaicism of PIK3R2 gene is correlated with developmental brain disorders ranging from BPP (Bilateral perisylvian polymicrogyria) with a normal head size to the MPPH (Megalencephaly-Polymicrogyria-Polydactyly-Hydrocephalus) syndrome (Rivière et al., 2012; Mirzaa et al., 2015; Mirzaa et al., 2016). Scientists discovered that a somatic activating mutation in GNAQ is related with Sturge-Weber syndrome and port-wine stains from whole genome sequencing data of 97 individuals (Shirley et al., 2013). Previous studies also showed that somatic mutations in GNAQ in cells of a later development stage are correlated with uveal melanoma and blue nevi (Van Raamsdonk et al., 2008). There are also heritable neurological diseases where germline mutations exhibit a milder phenotype when somatic mutations occur in the same gene. For example, somatic mutations in LIS1 or DCX genes

can lead to gross disruptions of neuronal migration, whereas germline mutations in LIS1 or DCX result in lissencephaly (Gleeson 2000; Sicca et al., 2003).

Methods to detect somatic single nucleotide variations

Like brain tissue, somatic SNVs in normal tissue have a much lower candidate allele frequency compared to the typical mutational burden found in tumor samples (Lee. 2016). Thus, it is important to balance the sensitivity and accuracy of the methods applied for SNV identification in non-tumor tissues. To discover low allele frequency somatic SNVs in non-tumor tissues, both high depth whole genome/exome sequencing with different platforms and single-cell sequencing data have been suggested as potential technologies (McConnell et al. 2017).

For high coverage whole genome/exome sequencing, traditional Sanger sequencing could not be applied because the cost for Sanger sequencing is extremely high for high coverage over the genome (Genome.gov, 2019) and it could not detect mosaic SNVs with less than 17% candidate allele frequency (Jamuar et al., 2014). Previous studies showed that a read depth >1000X could detect a somatic SNV site with allele frequency of 1% with >90% probability (Shirley et al., 2013) with Illumina whole genome sequencing. However, amplification bias and certain artifacts in different methods could still bring false positives to the discovery.

Most of the existing computational methods for somatic SNV detection were developed for cancer genomics, for example, MuTect (Cibulskis et al. 2013), Strelka (Saunders et al. 2012), and VarScan (Koboldt et al., 2009). Previous studies have reported somatic SNVs with 1%-10% candidate allele frequency in paired intractable focal epilepsy brain and blood samples (Lim et al., 2015; Nakashima et al., 2015). However, since these methods were all designed for cancer somatic SNV identification, the sensitivity of these methods significantly drops when the candidate allele frequency is less than 8%.

Validation of somatic mutations

There are three common validation methods that are used to validate somatic SNVs: targeted DNA capture followed by high coverage sequencing, high coverage amplicon sequencing, and droplet digital PCR (ddPCR) (McConnell et al. 2017). Targeted DNA capture followed by sequencing with higher than 100X sequencing depth can validate somatic SNVs that are present in ~1% of the cells. High coverage amplicon sequencing with sequencing depth larger than 1000X could further validate somatic SNVs present in ~0.1% of the cells. Droplet digital PCR (ddPCR) has the potential to reach the highest sensitivity for validation of somatic SNVs (Hindson et al., 2011) by partitioning a DNA sample into large numbers of individual droplets containing one copy of template DNA but is also the most labor intensive. Thus, when we need to validate hundreds of somatic SNVs, ddPCR is not as realistic as amplicon sequencing. However, for sites that are difficult to validate with amplicon sequencing data, for example, similar

amplification bias in both original WGS sequencing and amplicon sequencing, we then can apply ddPCR for a more accurate validation.

Schizophrenia and somatic SNVs

Genetics constitutes a crucial risk factor for Schizophrenia. Family studies have demonstrated a higher rate of schizophrenia in relatives with schizophrenia compared to the general population (Kendler and Zerbin-Rüdin, 1996; Schulz, 1933). Multiple large-scale studies have identified several common or rare de novo genetic variations associated with schizophrenia phenotypes and have proposed several candidate genes as potentially involved with its pathogenicity (Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014; Fromer et al., 2014; Cai et al., 2015). However, these genes are not sufficient to explain the majority of sporadic schizophrenia cases (Lee, 2016). Existing genetics experiments, including linkage analysis, GWAS, and next generation sequencing have demonstrated that schizophrenia is a very complex, heterogeneous, and polygenic disease (Meehl, 1962; Badner and Gershon, 2002; Lewis et al., 2003; Ng et al., 2008; Purcell et al., 2009; Genome-wide association study identifies five new schizophrenia loci, 2011). Given that previous schizophrenic analysis has primarily focused on searching for germline SNVs from blood samples and exploring the influence of somatic SNVs to neurological diseases, our goal is to explore the possible somatic SNVs related to schizophrenia (Insel, 2013).

Method and Materials

Common experiment and schizophrenic brain samples and cell culture

Pulverized, frozen brain tissue and dural fibroblasts (FIBRO) from a deceased, male individual (5154) without known mental health disease were received from Lieber Brain Institute (Baltimore, MD) along with pulverized, frozen dorsolateral prefrontal cortex (DLPFC) and hippocampus (HIPPO) tissue from 42 deceased, male individuals, 21 diagnosed with schizophrenia and 21 without known mental health disease. Samples from individuals with schizophrenia were paired with samples from individuals without mental health disease based on age and ethnicity to create 21 schizophrenia/control sample pairs. We also received dural fibroblasts (FIBRO) from 8 pairs that are a subset of the 21 pairs.

Dural fibroblast were cultured in DMEM (Gibco/ Life Technologies) supplemented with 10% FBS (Gibco/ Life Technologies), 2% GlutaMAX (Gibco/ Life Technologies), and 1% Antibiotic, Antimycotic ((Gibco/ Life Technologies). Cells were cultured 3-10 weeks and passaged when they reached 85%-95% confluence.

Common experiment and schizophrenic genomic DNA isolation and sequencing

Genomic DNA (gDNA) was extracted from 5154 brain tissue, DFPLC and HIPPO from 10 pairs, and FIBRO from 5154 and 8 pairs using the MagAttract HMW DNA Kit (Qiagen, Germantown, MD).

Length of gDNA was determined by standard electrophoresis in 0.4% agarose or pulse field gel electrophoresis in 1% agarose and 0.5 x TBE for 16 hours at 6 V/cm and 1200 angle with initial switch time 1 second and final switch time-6 seconds. For 5154 brain gDNA, 5154 FIBRO, and DFPLC from 10 pairs, an aliquot containing 1-5 µg of gDNA from the elution with the longest gDNA was sent to HudsonAlpha Discovery (Huntsville, AL) for linked read sequencing using 10x Genomics technology (Pleasanton, CA). Long Ranger v2.2 (10x Genomics) was used to align reads and then call and phase SNPs in order to obtain haplotype information for each read.

Whole exome sequencing was done on all extracted gDNA. Duplicate libraries were made for each sample by shearing 75-200ng of gDNA to 350bp. Libraries were purified with 0.65x SPRIselect beads (Beckman Coulter), quantitated by Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific, Carlsbad, CA), a 50ng aliquot was removed, and the remaining 400-800ng was used for exome target enrichment. Target enrichment was done using SeqCap EZ Exome Probes v3.0 (Roche Sequencing Solutions, Pleasanton, CA) according to manufacturer's protocol with the exception of a 72-hour incubation for hybridization and 12-16 cycles of post-capture LM-PCR to amplify captured DNA. Quantity was measured by Qubit™ dsDNA HS Assay Kit and target enrichment was determined by abundance of control targets in post capture libraries relative to these targets in pre capture libraries as outline in SeqCap_EZ_UGuide_v5.4 (Roche Sequencing Solutions).

Common process of sequencing data to alignment files for the consortium

We applied a common process pipeline to align and pre-process all the whole exome and whole genome sequencing raw data in fastq format. We utilized bwa v0.7.16a (Li and Durbin, 2009) for read alignment to human reference genome hg19. We then transferred sam files to bam files and merged output bam files from different lanes using sambamba v0.6.7 (Tarasov et al., 2015). Picard v2.12.1 (Broadinstitute.github.io, 2019) was applied to mark duplicates in the bam files. We applied GATK v3.7.0 (McKenna et al., 2010) to perform indel realignment and base recalibration to generate the bam file for downstream somatic SNV identification for all institutions.

Mosaic SNV identification and filtering pipeline

Candidate variants from paired brain and dural fibroblast samples were initially called with MuTect and Strelka with the default parameters. Concurrently, candidate variants from single (unpaired) brain samples were called with GATK Haplotype Caller with ploidy=5 parameter. Candidate mosaic SNV sites were then filtered out based on multiple quality filters. We first filtered out variants that overlapped with repetitive regions or low mappability regions, including regions covered by UCSC RepeatMasker simple repeats, Segmental Duplication, Simple Repeat tracks, regions not covered by 1000 Genome mappable segments as well as coverage not within +/- 3 standard deviations of mean coverage. We excluded common variants in gnomAD (Lek et al., 2016) with a population allele

frequency larger than 0.1%. During the process of counting alleles at candidate positions, reads with mapping quality lower than the 90 percentile of control sites (sites with high confidence), with more than 3% mismatches, as well as candidate site base quality lower than 20, were not considered. After counting for different alleles at candidate positions, sites with a candidate allele frequency larger than 0.01 in the control sample NA12878, from Genome in a Bottle Project (Genome in a bottle—a human DNA standard, 2015), were filtered out. We applied a Fisher Exact test to exclude the sites whose alternative alleles are enriched on one strand compared to the other. We then filtered out sites with a known indel in +/- 5 base pair region. Finally, we applied an allele frequency cutoff at 0.03 to exclude extremely low frequency events.

After removing the low-quality sites, we applied a binomial test with false-discovery protection using the Benjamini-Hochberg procedure (Freed et al., 2016.) to filter out potential heterozygous germline sites. We also used the haplotype information from the 10X sequencing data from the common experiment brain to further filter out false positive sites (Figure 2.5 a). We further filtered out candidate site clusters within 100 base pair distance.

Filtering false positive mosaic SNV sites using haplotype information

We have filtered the false positive mosaic SNV sites by using haplotype information provided by 10X Genomics “linked-read” sequencing data. 10X Genomics “linked-read” sequencing data is a barcoded short read sequencing

method. Each longer DNA molecule (~50kb) will be attached to a different bead then form individual droplet. Inside each droplet, the long DNA fragments are fragmented into small pieces of DNA and then attached with the unique barcode within each droplet and then subsequently amplified. The DNA library then is sequenced using standard next generation sequencing methods, for example, Illumina sequencing. Using the short reads with the same barcodes from the same molecule, we then applied the LongRanger pipeline (10X Genomics, LongRanger. 2018) to align and call haplotypes using reconstructed long molecules as well as SNPs in the long molecules (Figure 2.5 a).

There are two categories of false positive mosaic SNV sites that we could filter out using haplotype information. The first category is when the candidate allele is observed on both haplotypes. This indicates that the mosaic SNV discovered is likely a sequencing artifact because the mutation of a position in a diploid genome is highly unlikely to occur on both of the chromosomes at the same position. The second category is when a candidate allele represents the vast majority of the observed alleles within a haplotype (>90%). In this case, this SNV is more likely germline SNV rather than a mosaic SNV because a mosaic event is not expected to take over all cells for one chromosome (Figure 2.5 b).

Amplicon validation of mosaic SNVs data generation

Putative mosaic SNVs were validated by high throughput sequencing of amplicons that contain the SNV and then calculating relative abundance of reads

containing the alternate allele. Primers were designed using Primer 3 software (http://bioinfo.ut.ee/primer3-0.4.0/) with 300-400bp of genomic sequencing surrounding the SNV as the input. Since the read length of the amplification product is 300bp, we were able to gain an overlapped region between the paired reads. With the overlapped regions containing the somatic SNV candidates, we were able to sequence each candidate site twice and increase the accuracy of sequencing result by excluding reads with non-concordant bases from the pair ended reads at the candidate SNV positions. If possible, SNPs known to be heterozygous in our samples were included in the genomic sequence used as input. Primers were tested in silico (Kent et al.) to confirm they uniquely target the correct region of the genome. Phusion® High-Fidelity DNA Polymerase (New England Biolabs) was used according to manufacturer's instructions for amplification and primers were cycled under varying conditions to determine optimal PCR mix and annealing temperature. To generate amplicons for sequencing, either NA12878 or gDNA from 5154 brain tissue was used as template. PCR product was purified with 0.7x SPRIselect beads (Beckman Coulter) and 10% of product was visualized on agarose gel to confirm only one amplicon the correct size was present. If the size is the same as the primer designed in the electrophoresis gel, we then performed MiSeq to sequence the targeted amplified genetic fragment. If the size is incorrect, we designed a second batch of primers to obtain unique amplification. If none of the primers worked, the candidate was flagged with 'primer not designed'. (Figure 2.8 a) Protocols and reagents from NEBNext® Ultra™ DNA Library Prep Kit for

Illumina® (New England Biolabs) were used for end repair, dA-tailing, and to ligate NextFlex adapters (Perkin Elmer, Waltham, MA) onto amplicons. After ligation, reactions were purified with 0.7x SPRIselect beads (Beckman Coulter) and PCR enrichment of adapter-ligated DNA was done for 10 cycles using NEBNext® Ultra™ DNA Library Prep Kit (New England Biolabs). Amplified libraries were purified with 0.7x SPRIselect beads and sequenced with MiSeq Reagent Kit v3, 600 cycle PE on MiSeq sequencer (Illumina, San Diego, CA).

In the validation experiment, we have also applied our validation pipeline to amplicon sequencing data generated by other four institutions. Two of the four institutions applied a similar amplicon sequencing method as ours, except that instead of running electrophoresis gel to ensure the size of the amplicon, they designed the primers, amplified the targeted regions, then directly sent the amplification material for sequencing without a further quality check of the DNA material. Although the amplicon sequencing data generated this way is of lesser quality as the data generated by our approach, it still represents a high coverage data set covering candidate mosaic SNV sites for a valid validation. There is one institution which also applied amplicon sequencing as the previous two but only generated single ended sequencing data. The last institution applied ion torrent sequencing with amplified targeted regions containing mosaic SNV sites for validation instead of amplicon sequencing.

Amplicon sequencing data analysis pipeline

After performing MiSeq, pair-ended sequencing data was assembled into to a single read by using PEAR (Zhang et al., 2013.). The assembly of pair ended reads with overlap can increase the accuracy by sequencing the candidate position twice, and only when the bases at the candidate positions are concordant between the paired reads, the reads were counted in the following steps for further filtering. When the paired-end reads of a single fragment are combined, the non-concordant bases between the two reads are set to N with a base quality of 0. We then applied bwa mem (Li et al., 2009.) for read alignment to hg19, followed by application of the Genome Analysis Toolkit (McKenna et al., 2010.) for indel realignment (Figure 2.4 a).

After this pre-processing, we applied a series of filters to evaluate the mosaic SNVs (Figure 2.4 b). A lower limit of 200 reads covering the candidate position was established as a minimum requirement; sites with less than 200 reads covering in the amplicon sequencing were marked as ‘read not enough’ since the data is too sparse to make any conclusion at these sites. We then compared the candidate allele frequency of the brain sample to the candidate allele frequency of the negative control sample. Given the hypothesis that the same mosaic SNV event should not take place in two different individuals, we do not expect the candidate allele called from the brain sample to also be present in the negative control sample. By applying both a hard cutoff on the candidate allele frequency for NA12878 and a skellam test comparing the candidate allele frequency of the

two samples, we exclude possible false positive candidates caused by biased sequencing error in certain genomic context.

Skellam test:

$$p(k;\mu_1;\mu_2) = \Pr\{K = k\} = e^{-(\mu_1+\mu_2)}\left(\frac{\mu_1}{\mu_2}\right)^{\frac{k}{2}} Ik(2\sqrt{\mu_1\mu_2})$$

where μ1 is the coverage of brain sample at candidate position; μ_2 is the coverage of NA12878 at candidate position; k is the difference between the alternative counts of brain and NA12878.

We also established an empirical error model to exclude the biased sequencing error because of the artifacts by DNA amplification, library preparation and sequencing processed. This error model was derived for different amplicon sequencing libraries by assessing the mismatch rate (second allele frequency) of the rest of positions in the overlapped region between the pair ended reads except for the candidate position (Figure 2.8 d). We then take the 95 percentile of mismatch rate distribution for each kind of base change as the cutoff for sequencing error for different kinds of base changes in for candidate positions. In addition, we also applied a candidate allele frequency filter at 0.4 to exclude possible germline SNVs.

Results

Mosaic SNVs identified by existing methods for cancer samples

Previous study developed various methods to identify mosaic SNVs in tumor compared to normal samples. We thus initially applied MuTect (Cibulskis et al.,

2013), Strelka (Saunders et al., 2012) with default parameters and GATK (McKenna et al., 2010) with ploidy 5 to discover mosaic SNVs in pulverized brain tissue compared to dural fibroblast tissue from the same individual using the exome captured sequencing data. The coverage of our exome-captured sequencing data was ~250X (Figure 2.1 b). In total, the three methods discovered 249,030 mosaic SNVs across the exome, with only 5 overlapping sites among the three methods. We further filtered this candidate site list down by limiting the mosaic SNV candidates to sites only inside the exome capture targeted regions. This results in 23,215 mosaic SNV candidates combining all three existing methods, however, there was no overlap among the three methods. This lack of agreement among the three different methods within the targeted regions indicates a lack of specificity of the methods when identifying somatic SNVs in non-tumor normal tissues (Figure 2.2 a).

Mosaic SNVs identified by new mosaic SNV identification methods

We next applied our mosaic SNV identification and filtering pipeline to the whole exome and whole genome data generated by other institutes using the same pulverized brain tissue. In total, we identified 1148 mosaic SNV sites from 8 pairs of brain-fibroblast WGS data and 2 pairs of brain-fibroblast WES data (Figure 2.6 b) with different coverage and library preparation techniques (Table 2.1) using our filtering method (Figure 2.4). The candidate allele frequency for the sites identified was primarily less than 0.05 (Figure 2.6 a). 43 discovered mosaic SNV

sites takes place in more than 10% of the cells, among which 8 sites are present in more than 20% of brain cells.

Among the 1148 mosaic SNV sites, 36% of the sites were cytosine (C) to adenine (A) changes, 26% are adenine (A) to thymine (T) changes, and cytosine (C) to guanine (G) takes the least percent of the whole set (Figure 2.6 b). We also observed that base changes in the middle of tri-nucleotide with the same three bases were a frequent occurrence across all mosaic sites discovered (Figure 2.6 c). These sites in the middle of a homopolymer region are more susceptible to DNA mutations (Denver et al. 2005). However, they are also enriched in sequencing error prone regions. Because of the amplification nature of the sequencing methods, these possible sequencing artifacts inside the homopolymer regions are difficult to distinguish from the true mosaic SNV events.

We then compared the mosaic SNVs identified from the different aliquots of the same brain sample in all four WGS datasets. Only 4 out of 1148 sites were discovered from all 4 datasets. The two datasets with the highest coverage (Table 2.1) have the most overlapped sites between each other. However, the set with the lowest coverage reported the highest number of mosaic SNVs (n=455). We then inspected the candidate allele count of the mosaic SNVs (Figure 2.6 e). For the two WGS datasets with lower coverage, the candidate allele counts are mostly less than 5 reads. Compared to the two samples with lower coverage, the candidate allele counts for the candidate sites called from

the two higher coverage samples are much higher. This shows the low specificity of both identification methods and filtering pipeline when the sequencing data coverage is low.

Comparison of mosaic SNVs identified by other different filtering pipelines

There are in total six groups which have made mosaic SNV identification using the same bam files from the common processing pipeline (See Methods). We performed a similar analysis of base changes and genomic context of mosaic SNV candidates by using the candidates from all six groups. The result shows that all pipeline have a similar bias towards the homopolymer regions as well as the cytosine (C) to adenine (A) and adenine (A) to thymine (T) changes (Figure 2.7 a, b). There is limited overlap among the mosaic SNV calls from different groups using different identification and filtering strategies. Only three sites overlap were identified by five groups, and in total 10 sites in common for four groups among the 1298 mosaic SNV candidates identified in total by all six groups. With different sensitivity and accuracy trade-off and different number of libraries used, different pipelines had dramatic different call sets of mosaic SNVs for the same sample (Figure 2.7 c).

Selection of Mosaic SNV sites for Validation experiment

We collectively decided to select 400 / 1298 sites for validation by splitting all sites into 4 categories (Table 2.2). They are absolute singletons, data source singletons, approach singletons and multi-calls. Absolute singletons were defined

when the mosaic SNVs were identified by only one method and have supportive evidence from only one sequencing library among the six. The supportive evidence was measured by the candidate allele count in each library. If there was one high quality read supporting this mosaic SNV site in a library, we could define it as with supportive evidence from this library. Then data source singletons were sites discovered by multiple methods, but only supported by one sequencing library. Approach singletons are then the sites identified by only one approach but have supportive evidence from multiple data sources. Multi-calls are the sites identified by multiple approaches with supportive evidence from multiple data sources. In total, all six approaches identified 1298 sites, among which 45 sites were multi-calls (Table 2.3). These calls are either true positives or artifacts that none of the approaches were able to exclude.

As our pipeline traded for better sensitivity over accuracy, this resulting in our having the most candidate mosaic SNVs compared to any other groups. Due to this and the limited number of validation experiment that could be carried out, we decided to include the 181 sites identified by all other five approaches and randomly select 219 sites from the 1114 unique sites identified by our method for validation. The 219 sites were selected based on the ratio of the four categories described above.

Mosaic SNV validation result, manual inspection and best practice of mosaic SNV identification from non-tumor tissue

With 2 primary methods of validation (see Methods), we conducted validation experiments on the 400 sites with 100 sites distributed each institution, with 20 sites from each cohort replicated by another group (Table 2.6). We filtered the 400 sites based on the validation result using the pipeline we developed (see Methods). In total, we identified 86 true positive mosaic SNV candidates, among which 12 candidates were validated by two institutions. 19 candidates have non-concordant decisions between the two validation institutes. We then manually inspected the true positive sites as well as the sites with non-concordant result from the validation data of two institutions.

With the manual inspection of all PASS sites as well as the non-concordant sites, we identified 42 PASS sites as false positive because of candidates in homopolymer, structural variation, CNV, biased sequencing error or biased amplification of one haplotype (Figure 2.10; Figure 2.11; Figure 2.12; Figure 13), and we were able to rescue 1 non-concordant sites as PASS. There are 11 candidate sites where we could not decide if they are true candidate or false positive sites. We have sent a subset of these sites for droplet digital PCR for further validation (Table 2.6).

From the experience of validation result and manual inspection, we have analyzed the efficiency of different filters. We have summarized the best practice to discover somatic SNV sites from non-cancer tissues from both single sample and paired sample data (Figure 2.14).

Discussion

We initially attempted to identify mosaic SNV in brain tissue compared to dural fibroblast tissue in our WES data using existing mosaic SNV identification tools, MuTect (Cibulskis et al., 2013), Strelka (Saunders et al., 2012) and GATK (McKenna et al., 2010) (ploidy=5). However, out result showed that the overlap among the three existing methods was only 5 sites which were removed after only considering regions targeted by the whole exome capturing kit. Given the non-concordance of the three existing methods, we concluded that existing mosaic SNV identification tools designed for cancer tissues do not fit the purpose of identifying somatic SNVs in non-cancer tissues, i.e. brain tissue and developed our own approach.

We applied our filters from the raw candidates of MuTect (Cibulskis et al., 2013), Strelka (Saunders et al., 2012) and GATK (McKenna et al., 2010) (ploidy=5). With our extra quality and information filters, we were able to identify 1298 candidates from 4 pairs of WGS data and 2 pairs of WES data. We analyzed the 1298 candidates from different perspectives (enriched homopolymer calls, few overlaps among different libraries from the same sample and low alternative allele count from certain libraries (Figure 2.5)) and showed that there are still possible false positive calls from the pipeline. We also compared out somatic SNV calls with the calls from other institutes. The relatively few overlaps among

different institutes also represent possible false positives from this version of pipeline.

In order to validate the candidates identified by different institutes, in particular the low allele frequency sites, we performed extremely high coverage amplicon sequencing for 400 sites from both the brain 5154 sample and NA12878 as a negative control. We then applied our analysis pipeline to filter out the false positive candidates from high coverage sequencing data. By comparing the candidate allele frequency with an out-group negative control NA12878, we were trying to exclude the biased amplification error with high sequencing depth in amplicon sequencing.

We manually inspected all candidate sites passed the validation analysis pipeline as well as the candidate sites with different decision between two institutes to ensure the validation status of the final somatic SNV status. From the manual inspection, we were able to discover multiple cases where the validation analysis did not exclude certain false positive sites, for example, somatic SNV sites inside homopolymer regions, somatic SNVs in structural variations, somatic SNVs in CNVs, germline SNV in a biased amplified region or sequencing error that was specific to brain sample not detected in NA12878.

For somatic SNV candidates in homopolymer regions, we excluded the sites as false positive sites (Figure 2.12). However, it is difficult to decide whether a

somatic SNV site in homopolymer region is a sequencing artifact or a true somatic SNV event. Homopolymer regions have always been considered as regions with high sequencing error because of the principals used for detection of next generation sequencing (Hyman et al. 1988; Ronaghi 1998; Metzker et al. 2010). However, homopolymer regions in genome are highly prone to duplication error due to DNA replication polymerase slippage (Denver et al. 2005). Thus, even with an out-group negative control, we still could not decide if a somatic SNV in homopolymer region is a true somatic SNV event or a false positive site. To make our call set more accurate, we have excluded the sites in homopolymer regions as false positives but future work in this area is needed as some of these could represent *bona fide* somatic variation.

Somatic SNV candidates in possible structural variation regions or CNV regions are also difficult to exclude from the true positive sites due to aberrant read depth in such regions (Figure 2.13). Thus, we added a filter in our best practice to exclude the candidate sites inside structural variation or CNV regions identified using other structural variation and copy number variation identification tools. However, SV/CNV callers themselves still have a measurable false negative rate, and thus we still suggest manual inspection to add additional support to the true positive sites.

During our manual inspection, we also discovered the importance of the haplotype information provided by 10X Genomics data. With the 10X haplotype

information, we were able to exclude both sequencing artifacts (Figure 2.10) and biased amplified germline variants (Figure 2.11). For sites with alternative alleles on both of the haplotype, these were identified as sequencing artifacts (Figure 2.10) since somatic mutation should only affect one of the two chromosomes in a cell. For sites with alternative allele taking over all reads of one haplotype, although it may only have a 30% allele fraction, it was still identified as a germline mutation since somatic mutation should only occur on one of the two chromosomes (Figure 2.11). These sites commonly have relatively high alternative allele frequency and cannot be excluded from any other data. Thus, the haplotype information is critical to exclude false positive sites due to amplification bias. The amplification bias here could be caused by structural variation, copy number variation or other reasons that we did not observe. When 10X Genomics data is not available, haplotype information from paired end whole genome sequencing data could also be applied to exclude false positive sites.

From the different methods applied by other institutes, we also observed the importance of a panel of normal samples as negative control and single cell data to identification of somatic SNVs. Since we do not expect that the same SNVs could take place in multiple other individuals, a panel of normal samples could serve as a better negative control than only NA12878 used in our pipeline. Single cell data could also provide insights to exclude negative control sites. A candidate somatic SNV should not present in majority of cells if it is a true somatic SNV site.

In our best practice approach to identify somatic SNVs in non-tumor tissue, we suggest two possible kinds of input. If only one target sample is sequenced, we would suggest using GATK with ploidy=2-10 to obtain the raw candidates as input for further filtering pipeline. If the targeted sample is sequenced with a paired sample from the same individual, we would suggest applying MuTect, Strelka and GATK with ploidy=2-10 to generate raw candidates since MuTect and Strelka have been demonstrated to have higher sensitivity at low allele frequency regions with a paired sample. This would be followed by the application of germline filters including common SNPs from gnomAD and a binomial test that could exclude possible germline events from GATK in particular. We then exclude the sites within complex regions including repetitive regions, structural variants, copy number variants, indel and sites that are not in 1000 Genomes Project confident regions of human genome. After excluding possible germline SNVs and sites at less confidence regions, we then suggest applying a series of quality filters including base quality, mapping quality and percent mismatch of reads, together with strand bias Fisher Exact test and candidates as multi-allelic sites, NA12878 allele frequency to exclude possible false positive candidates brought by sequencing artifacts. Here, if a panel of unrelated normal samples is available, the complex quality filters could be replaced with the panel of normal sites. The hypothesis here is that a true somatic SNV site in one individual should not be identified in another unrelated sample. In the end, if single cell data or 10X genomics sequencing data is

available, we could utilize the extra information provided by these two different sequencing libraries to exclude possible false positives as described above in methods and results.

With the summarized best practice above and in Figure 2.14, we will apply the best practice to the Schizophrenic samples. We have 10 Schizophrenic brain samples (DLPFC and Hippocampus) with age-matched 10 neurotypical brain samples (DLPFC and Hippocampus). We also have 8 Schizophrenic fibroblast samples with age-matched 8 neurotypical fibroblast samples. We will apply our best practice using the whole exome sequencing and 10X Genomics sequencing of these samples. From the comparison between Schizophrenic samples and neurotypical samples, we expect that we could discover the somatic SNVs associated with Schizophrenia. However, further validations both in vitro and in vivo are required to confirm if any somatic SNVs identified is the possible cause of Schizophrenia.

Our study still has some limitations. For the somatic SNV candidates in homopolymer regions, we excluded them for a higher accuracy. Although the homopolymer regions are highly prone to sequencing errors, they are also the regions where mutations could occur because of the DNA polymerase slippage during DNA duplication. Because of the limitation of current next generation sequencing methods, we have not been able to make the decision for all candidates in homopolymer regions.

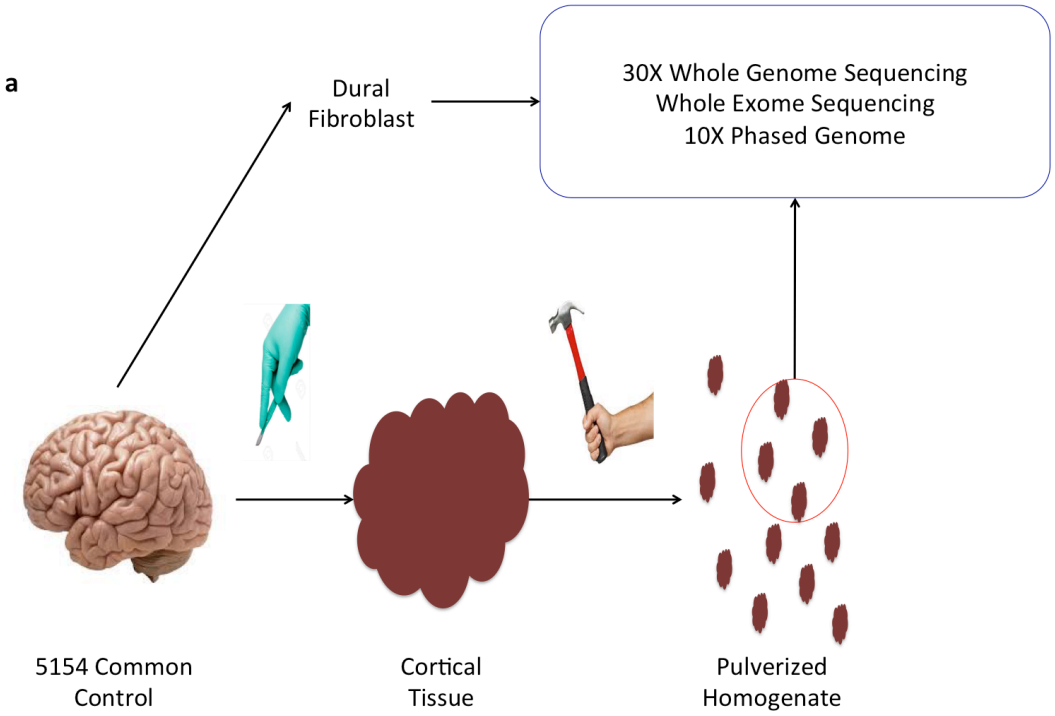
Another possible limitation is that methods for discovery of structural variations or copy number variants from next generation sequencing data is not perfect. Even with the filters of possible structural variants and copy number variants in the best practice pipeline, false positive somatic SNV candidates in structural variants or copy number variants regions could still not be excluded from the raw sites because of the inaccuracy of structural variants and copy number variants discovery.

Possible limitations for our experiment to identify Schizophrenia related somatic SNVs include region of the brain and likely intergenic functional somatic SNVs. If the somatic SNV related to Schizophrenia is not in the brain region that we sequenced, we will not be able to discover any Schizophrenia related somatic SNVs. Furthermore, although human exome contains most of the disease related mutations, it is still possible that the somatic SNVs correlated with Schizophrenia are located in intergenic region and performs an important function in brain development.

Conclusion and Future Remarks

In sum, we have demonstrated that the available somatic SNV discovery tools designed for tumor does not fit the purpose of identification of somatic SNVs in non-tumor tissue without clonal expansion with an allele frequency as low as 1% depending on the total coverage. From the analysis of our validation result, we

were able to summarize the best practice to identify somatic SNVs in non-tumor tissue and the methods for validation. We would then apply the best practice on the Schizophrenic samples and discover possible somatic SNVs related to Schizophrenia.


a
Dural
Fibroblast
30X Whole Genome Sequencing
Whole Exome Sequencing
10X Phased Genome
5154 Common
Control
Cortical
Tissue
Pulverized
Homogenate

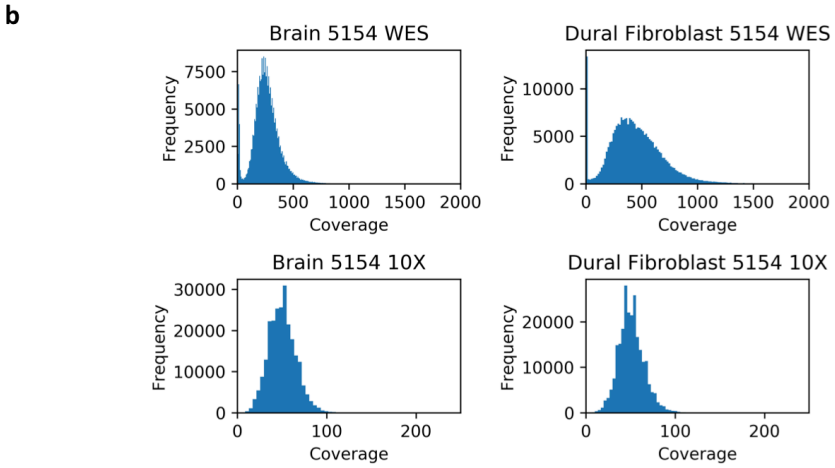
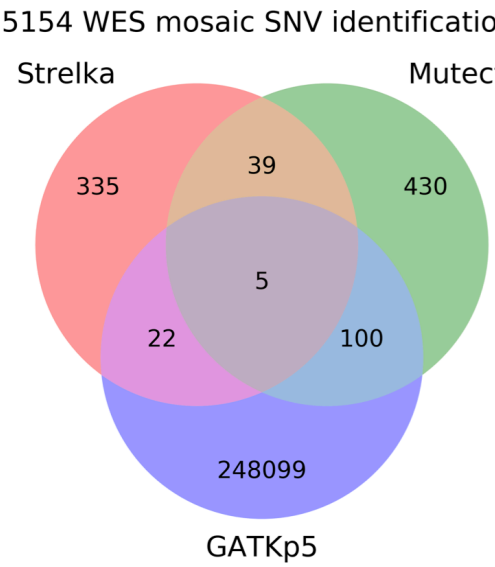

b
Brain 5154 WES
Dural Fibroblast 5154 WES
Brain 5154 10X
Dural Fibroblast 5154 10X
Frequency
Coverage

Figure 2.1: Common experiment design, 10X sequencing and data description.

a) Common control experiment design: Our collaborator from Lieber Institute dissected both cortical tissue and dual fibroblast from the same common control sample postmortem brain 5154. They then pulverized the cortical tissue sample from brain 5154 into homogenate and distributed the homogenate to 6 different institutions. Meanwhile, our collaborators in Lieber Institute cultured the dural fibroblast cells, then distributed the cultured dural fibroblast tissue to 6 institutions together with the homogenate. After we received the cortical tissue homogenate with the dural fibroblast, we prepared the sequencing libraries for these tissues using: 1) 10X Genomics Whole Genome Sequencing library preparation kit; 2) SeqCap EZ Exome Capture kit; 3) Gentra Puregen Whole Genome Sequencing library preparation kit to generate the libraries for 10X sequencing, whole exome sequencing as well as whole genome sequencing.

b) Data description and data summary: Histograms of coverage for WES and 10X libraries used for mosaic SNV identification. The pulverize brain sample has a coverage with mean coverage around 300X, and around 70X for 10X data. The dural fibroblast has coverage with around 500X, and round 70X for 10X data. By this much of depth, the lowest limit of mosaic SNV allele frequency could be identified from the brain WES data is theoretically 0.33% on average if there is one high quality read supporting the mosaic SNV event.

a


Brain 5154 WES mosaic SNV identification result
Strelka
Mutect
335
39
430
5
22
100
248099
GATKp5

b

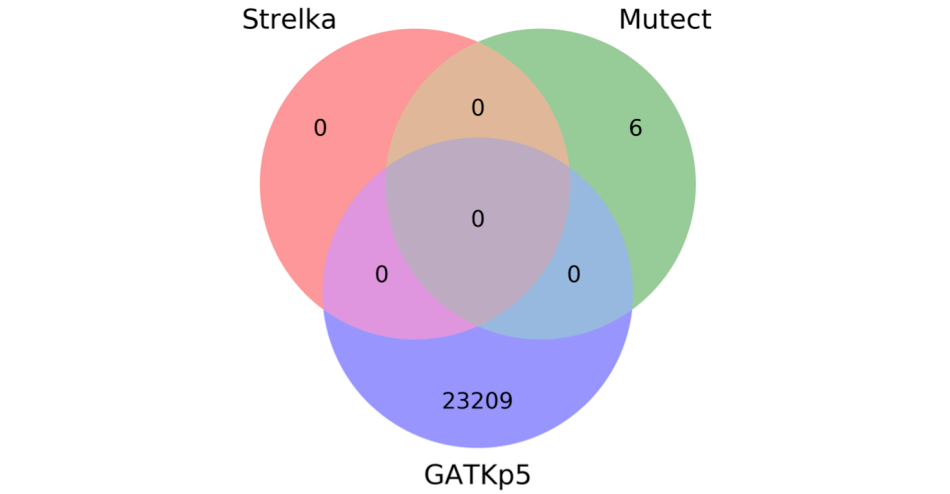

Brain 5154 WES mosaic SNV identification result - in targeted exome
Strelka
Mutect
0
0
6
0
0
0
23209
GATKp5

Figure 2.2: Existing methods on identifying mosaic SNVs.
a) Mosaic SNV identification using existing methods: Somatic SNV candidates were identified using MuTect, Strelka (pair sample method comparing brain and fibroblast) and GATK with ploidy parameter 5 (single sample method considering only brain) using the WES data generated in Michigan. A few overlapped sites were identified from the WES data.
b) Mosaic SNV identified by the three existing methods fall inside the targeted regions of the exome capturing kit (high confident regions): No overlapped sites were found among the three methods.

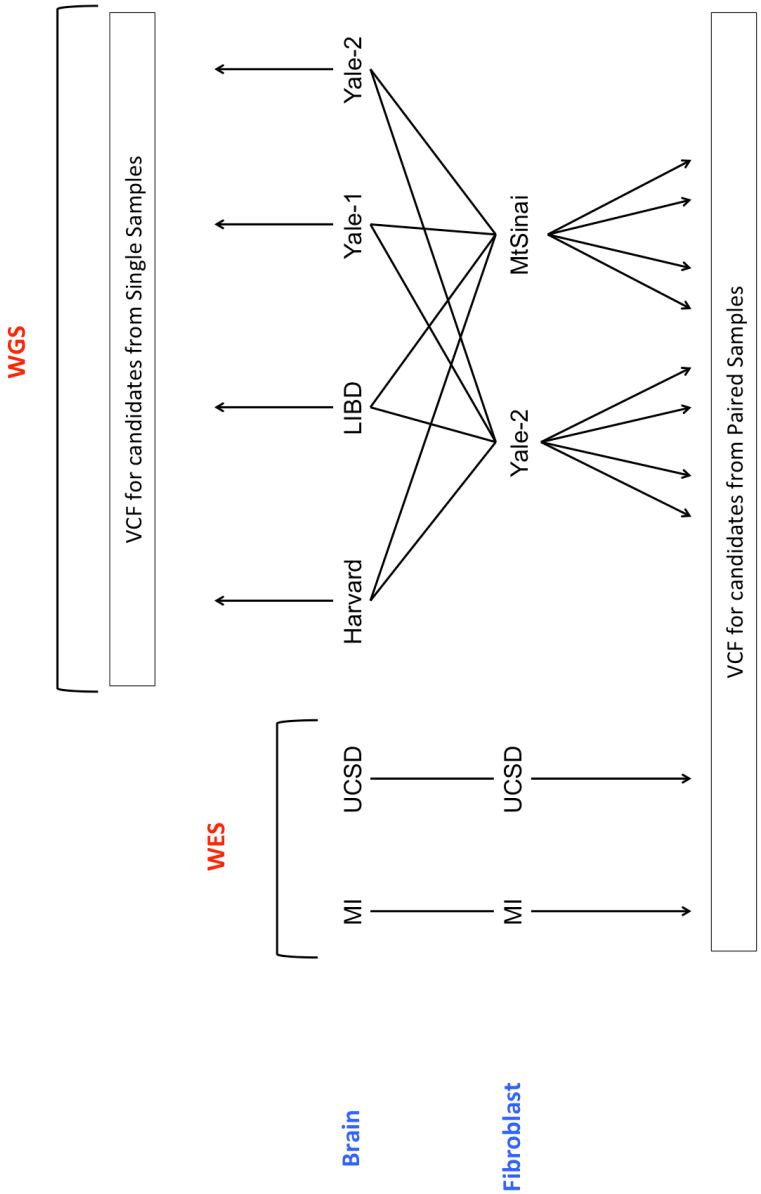


Figure 2.3: Mosaic SNV Identification common experiment design.
Paired mosaic SNV identification using consortium brain and dural fibroblast sequencing data. Each WGS brain data is paired with 2 different dural fibroblast data. Each WES brain data is paired with the other WES dural fibroblast data because the different exome capture kit used. Mosaic SNV candidates were then discovered by using our customized identification pipeline.

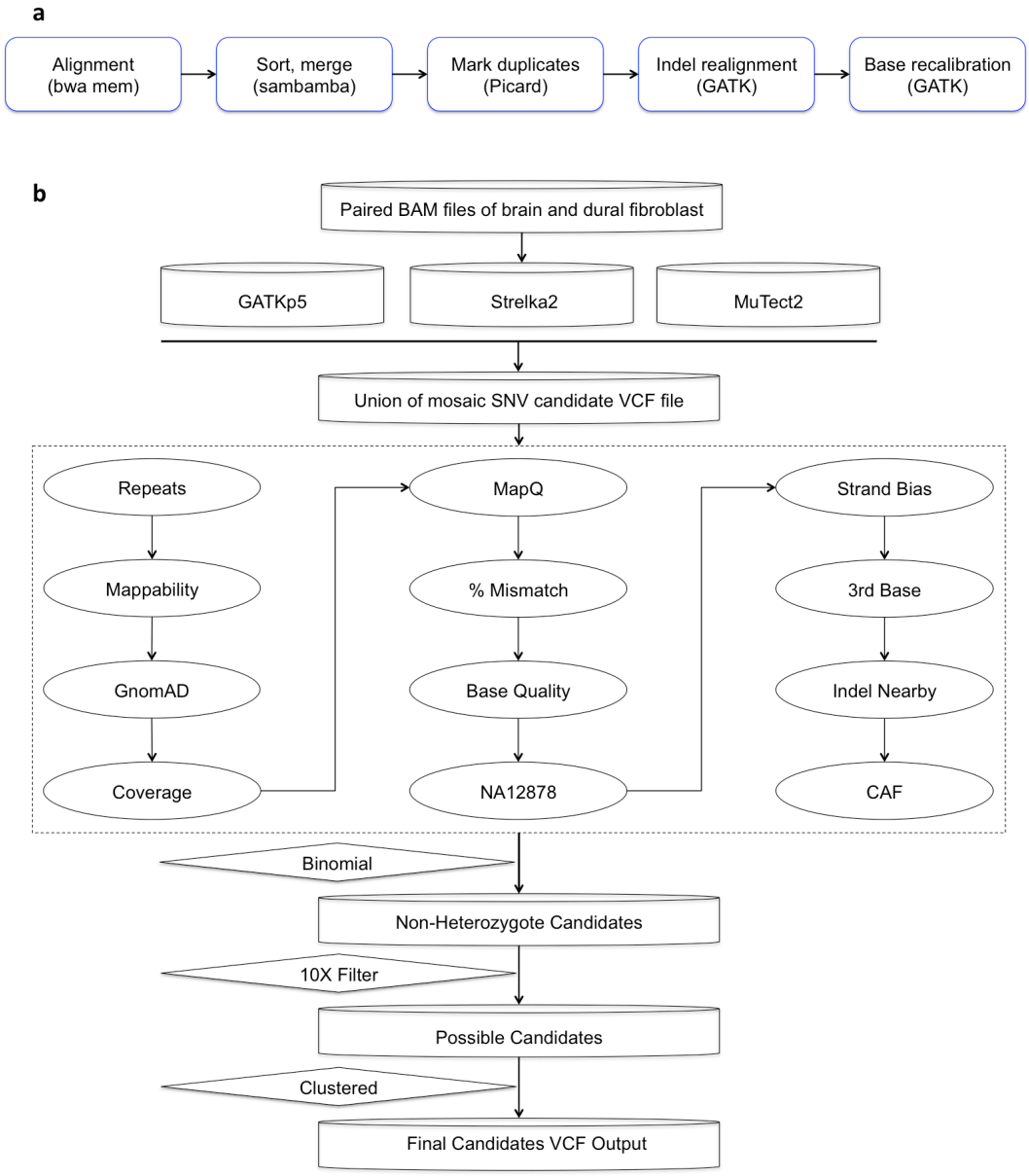
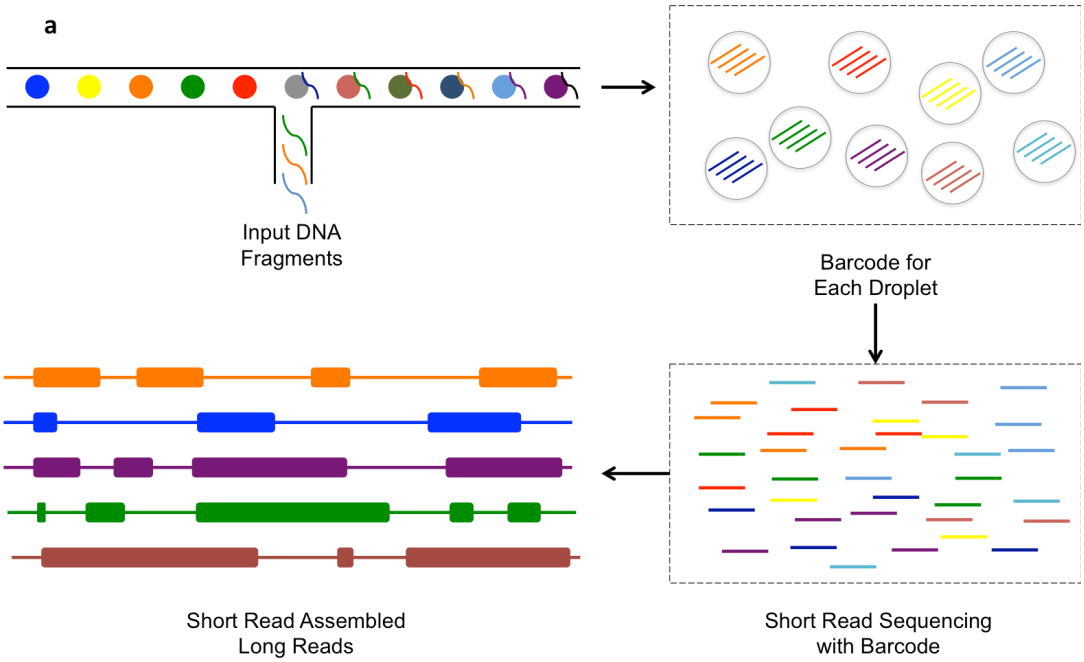
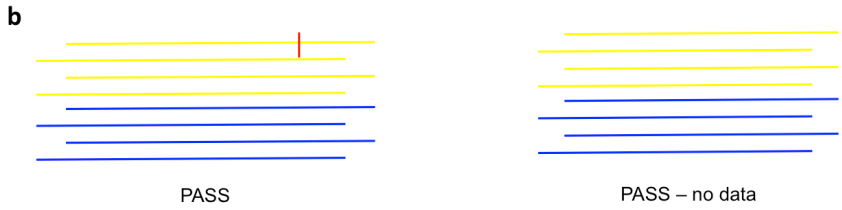

a
Alignment (bwa mem)
Sort, merge (sambamba)
Mark duplicates (Picard)
Indel realignment (GATK)
Base recalibration (GATK)
b
Paired BAM files of brain and dural fibroblast
GATKp5
Strelka2
MuTect2
Union of mosaic SNV candidate VCF file
Repeats
Mappability
GnomAD
Coverage
MapQ
% Mismatch
Base Quality
NA12878
Strand Bias
3rd Base
Indel Nearby
CAF
Binomial
Non-Heterozygote Candidates
10X Filter
Possible Candidates
Clustered
Final Candidates VCF Output

Figure 2.4: Common process pipeline and mosaic SNV Identification and filtering pipeline, input files, formats and outputs.

a) Common process pipeline of data generated from different institutions: Fastq files from Illumina sequencing using different library preparation methods were aligned to hg19 decoy 5 genome using bwa mem version 0.7.16a. We then applied sambamba v0.6.7 to sort and merge the bam files for the different libraries from the same sample. Picard v2.12.1 was then applied to mark replication duplicates from the library preparation step. We applied GATK v3.7-0 for indel realignment and base recalibration.

b) Mosaic SNV identification pipeline: We applied MuTect, Strelka and GATKp5 to identify the raw mosaic SNV candidates as input for downstream filtering steps. Candidates inside repetitive regions, segment duplication regions and self-chain regions with more than 90% identity are not considered. Only sites in mappable regions of the genome defined by 1000 Genome Project with P-base are considered for further filtering steps. After filtering candidates for high confidence regions in the genome, we filtered out the sites with more than 0.1% of population allele frequency from gnomAD project in all population. Candidates in abnormal coverage regions (outside 3 standard deviations from mean coverage of each sample) are filtered out. Then we take 95% percentile of map quality, percent mismatch and base quality as cutoff for high quality sites. The candidate allele frequency is then compared with the candidate allele frequency in a control sample, NA12878 from the Genome in a Bottle (GIAB) project with 300X coverage. We then used Fisher Exact test to exclude the false positive sites caused by strand bias in sequencing libraries. We compared the candidate allele frequency to the allele frequency of the third base at this position. The high allele frequency of the third base represents high sequencing artifacts at this position. We also exclude the sites with an indel present in upstream/downstream 5 base pair because of the alignment artifacts brought by this indel. We then set our cutoff for candidate allele frequency as 0.03 to eliminate low allele frequency candidates. We applied binomial test to exclude the germline events. We utilized the haplotype information from 10X data to filter out false positive sites caused by sequencing error or germline events. At the end, we filtered out sites which are within 100 base pair between each other.


a
Input DNA
Fragments
Barcode for
Each Droplet
Short Read Assembled
Long Reads
Short Read Sequencing
with Barcode


b
PASS
PASS – no data


Fail – sequencing error
Fail - germline

Figure 2.5: Common process pipeline and mosaic SNV Identification and filtering pipeline, input files, formats and outputs.
a) 10X Genomics sequencing with barcode provided haplotype information from short reads: Instead of fragmenting genome into small pieces, 10X technique takes longer fragments of DNA (~40kb) as input. Each long DNA fragment would be attached to a bead then fragmented into smaller pieces for Illumina sequencing inside each bead. Unique barcode is then added to all small fragments inside each bead. All the shorter fragments are then sequenced using Illumina pair ended sequencing. This way, we could then build a long fragment based on the barcode from short reads. With the SNPs on each long fragment, we could then phase the sequencing result into two haplotypes.
b) 10X haplotype information helps with excluding false positive mosaic SNV sites: The haplotype information is applied for filtering false positive sites. An ideal mosaic SNV site should be presented only at one haplotype and not take the major part of this haplotype. If the candidate allele is on both haplotypes, this site is a sequencing artifact since mosaic single nucleotide mutation is barely possible to take place on both chromosomes. If the candidate allele takes the major counts on one haplotype (>90%), this site is more possible to be an under-amplified germline event or in a copy number / structural variant region.

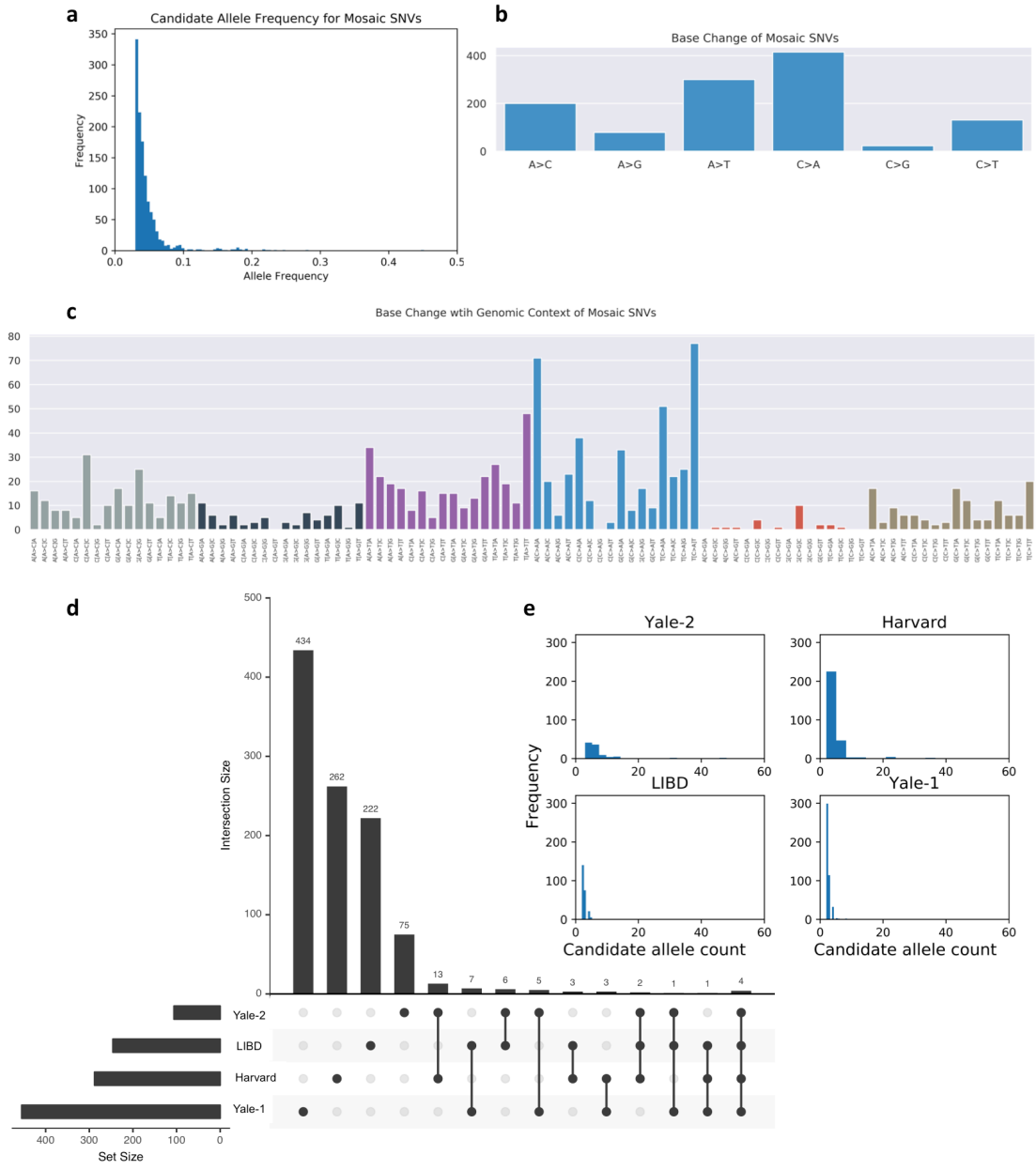

a
Candidate Allele Frequency for Mosaic SNVs
Frequency
Allele Frequency
b
Base Change of Mosaic SNVs
A>C
A>G
A>T
C>A
C>G
C>T
c
Base Change wtih Genomic Context of Mosaic SNVs
d
Intersection Size
Set Size
Yale-2
LIBD
Harvard
Yale-1
e
Yale-2
Harvard
LIBD
Yale-1
Frequency
Candidate allele count
Candidate allele count

Figure 2.6: Summary of mosaic SNV identified from the filtering pipeline.
a) Candidate allele frequency distribution of mosaic SNV identified: The histogram of candidate allele frequency identified from our pipeline shows that we were able to exclude the germline events effectively. However, for the low allele frequency events, it is difficult to exclude the false positives. Downstream validation analysis is necessary to evaluate this result.
b) Base change bar plot of mosaic SNVs identified: The mosaic SNV base changes we identified are enriched with C to A and A to T changes. Only a few C to G changes were identified.
c) Base change with genomic sequence context bar plot of SNVs identified: Most of the mosaic SNV candidates identified are in the middle of tri-nucleotide homopolymer regions. This shows a possible false positive calling because of the confidence for sequencing inside a homopolymer region is low. Future validation would show if these sites are true positives or not.
d) Upset plot of SNVs identified in different libraries prepared by different institutions: Only 4 common sites were identified from all four libraries made from the same tissue. We discovered the most mosaic SNVs from the sample with the lowest coverage. This shows the allele frequency filter may not be as effective as an allele count cutoff.
e) Candidate allele count identified from different WGS brain bam files: This figure shows most sites called from Yale-1 library only have less than 4 read counts supporting the mosaic SNV event.

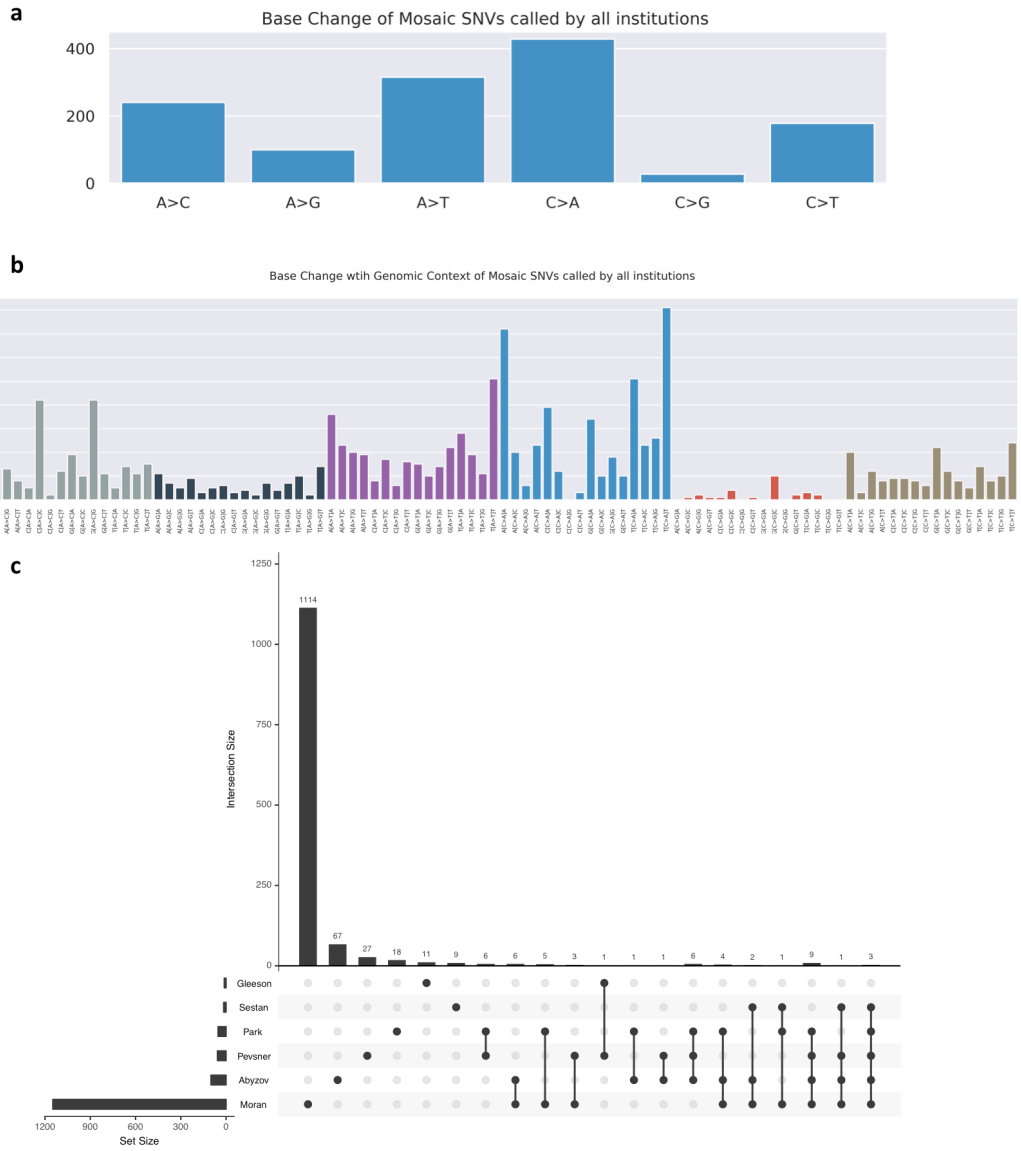
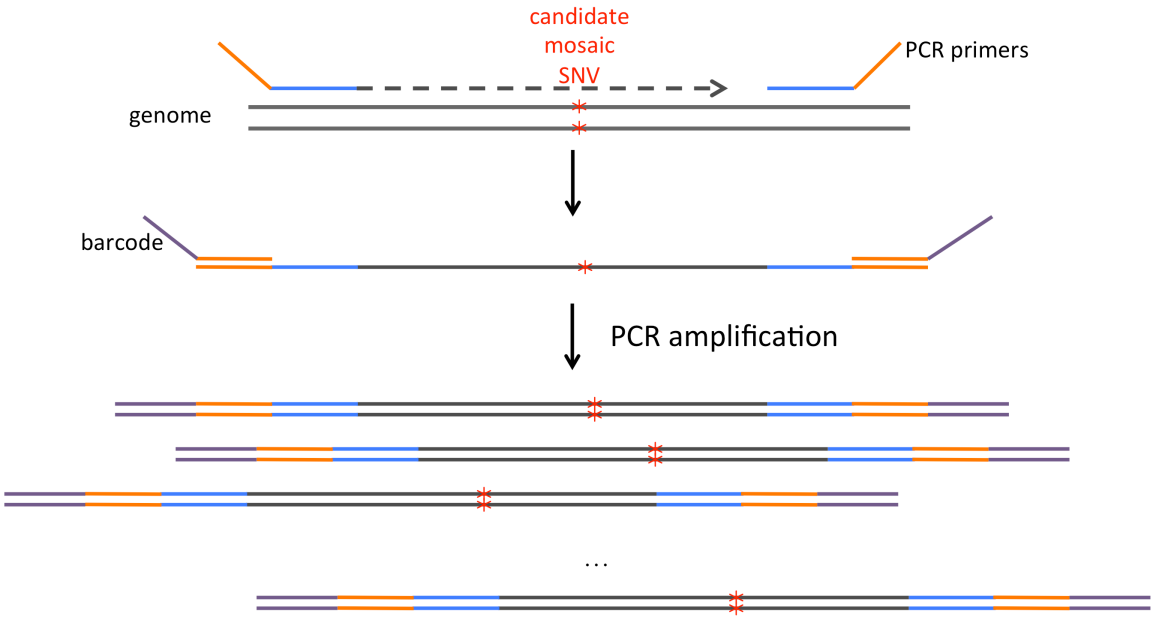
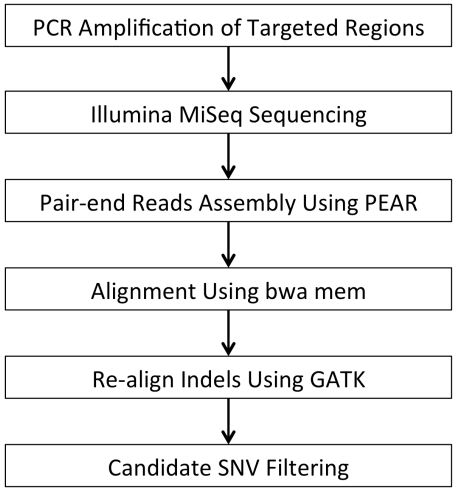

a
Base Change of Mosaic SNVs called by all institutions
400
200
0
A>C
A>G
A>T
C>A
C>G
C>T
b
Base Change wtih Genomic Context of Mosaic SNVs called by all institutions
c
Intersection Size
1250
1000
750
500
250
0
1114
67
27
18
11
9
6
6
5
3
1
1
1
6
4
2
1
9
1
3
Gleeson
Sestan
Park
Pevsner
Abyzov
Moran
1200
900
600
300
0
Set Size

Figure 2.7: Summary of mosaic SNV identified from all six institutions using different identification pipelines.
a) Base change bar plot of mosaic SNVs identified: Combined together, all six methods identified C to A, A to T and A to C base changes the most. C to G changes is under-represented.
b) Base change with genomic sequence context bar plot of SNVs identified: Combining all sites discovered from different methods, candidate mosaic SNV candidates in tri-nucleotide homopolymer regions are still enriched.
c) Upset plot of SNVs identified by different institutions using different filtering strategies: There is no common mosaic SNVs discovered by all six pipelines, and only three common mosaic SNVs identified from 5 of the methods. The few overlapped sites among the six different methods require downstream validation experiment in order to evaluate the performance of each different methods.

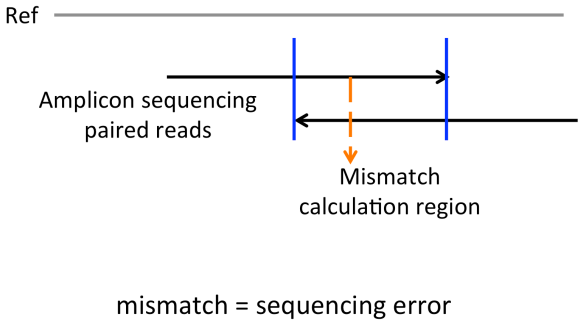
a



b



d



c

Validation Criteria	Cut-off
Amplicon sequencing coverage	200
NA12878 candidate allele frequency	0.01
Allele count difference between brain 5154 and NA12878 using Skellam test	p-value: 0.01
Empirical model for sequencing error	customized
candidate allele frequency	0.4-0.6

Figure 2.8: Validation of candidate mosaic SNVs using amplicon sequencing.

a) Amplicon sequencing of targeted regions in genome. We designed primers to amplify regions ~300bp around the candidate mosaic SNV sites. After the designed DNA fragment is amplified, a barcode would be added to each unique DNA fragment. After massive amplification, we then use MiSeq to sequence the DNA fragment we amplified. The massive amplification in this process could guarantee a high sequencing depth covering the candidate site so that even low allele frequency mosaic SNV events could be validated.

b) Amplicon sequencing data pre-processing pipeline. We applied PEAR v0.9.10 to assemble the pair end reads with overlap between each other to single end reads for lower sequencing error by taking the consensus base of the two ends if available. We then used bwa v0.7.16a to align the reads to hg19d5 genome. We used GATK to re-align the indels around the candidate regions. We then applied further analysis to filter out false positive mosaic SNV sites from the amplicon sequencing data.

c) Mosaic SNV validation filters using amplicon sequencing data. The filters we applied include: 1) number of reads covering the candidate sites (200); 2) NA12878 candidate allele frequency at the same position with the same amplicon sequencing method; 3) Candidate allele frequency difference between brain 5154 amplicon sequencing and NA12878 using Skellam test; 4) empirical sequencing error model 95 percentile as the cutoff for sequencing error; 5) exclude germline events by setting a cutoff between 0.4 and 0.6.

d) *Illustration of empirical model of sequencing error for each different amplicon sequencing libraries.* The empirical model was built based on the base change frequency of all the other sites in the amplified regions. For pair end reads which overlap with each other, the empirical model was built on the overlapped regions only. For single end reads, the empirical model was built on all positions sequenced except the primer regions. The error rate for each different category of base changes for different sequencing libraries could represent the overall sequencing error of this library.

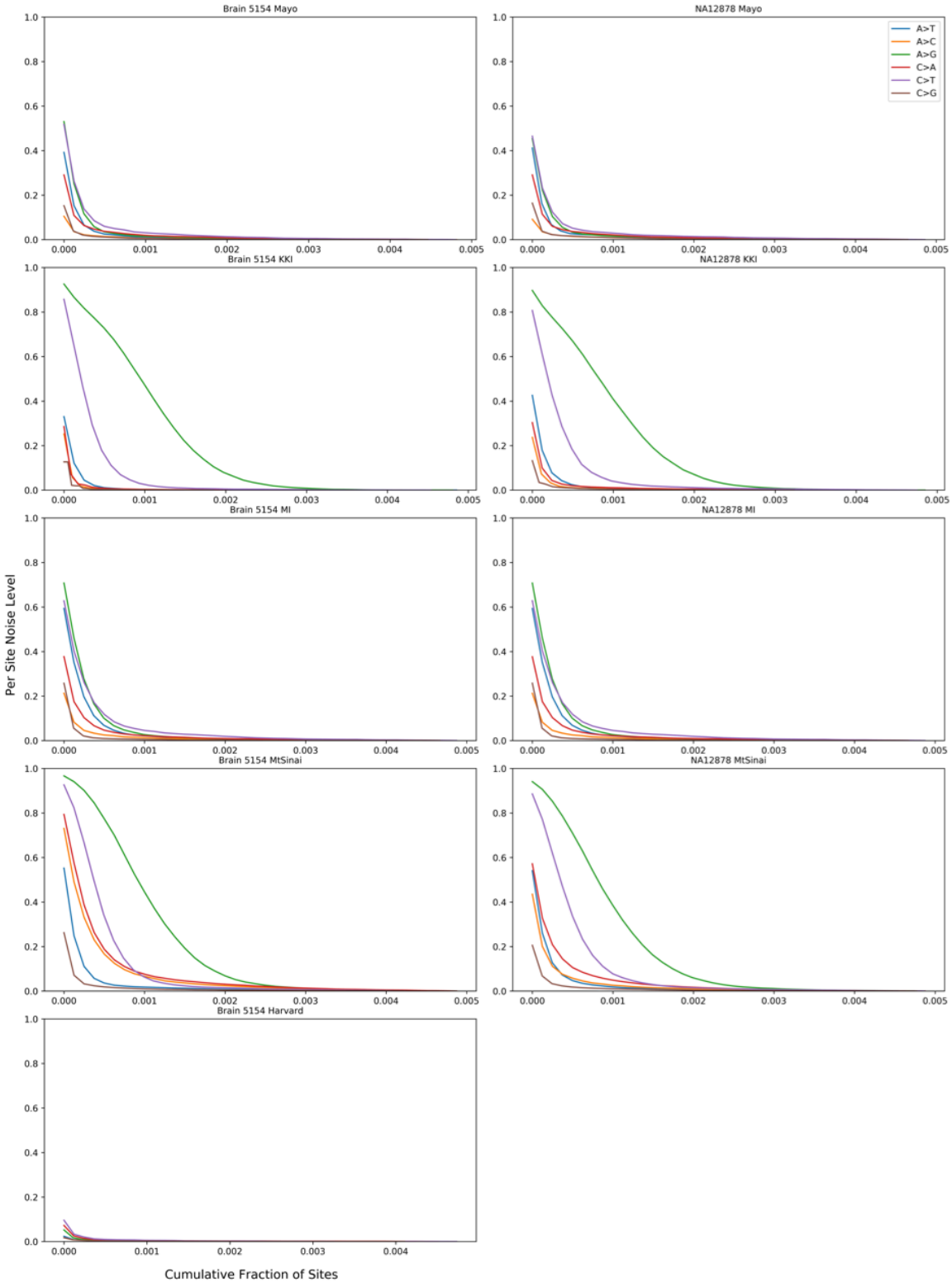


Figure 2.9: Sequencing error cumulative curve for each different amplicon sequencing libraries.

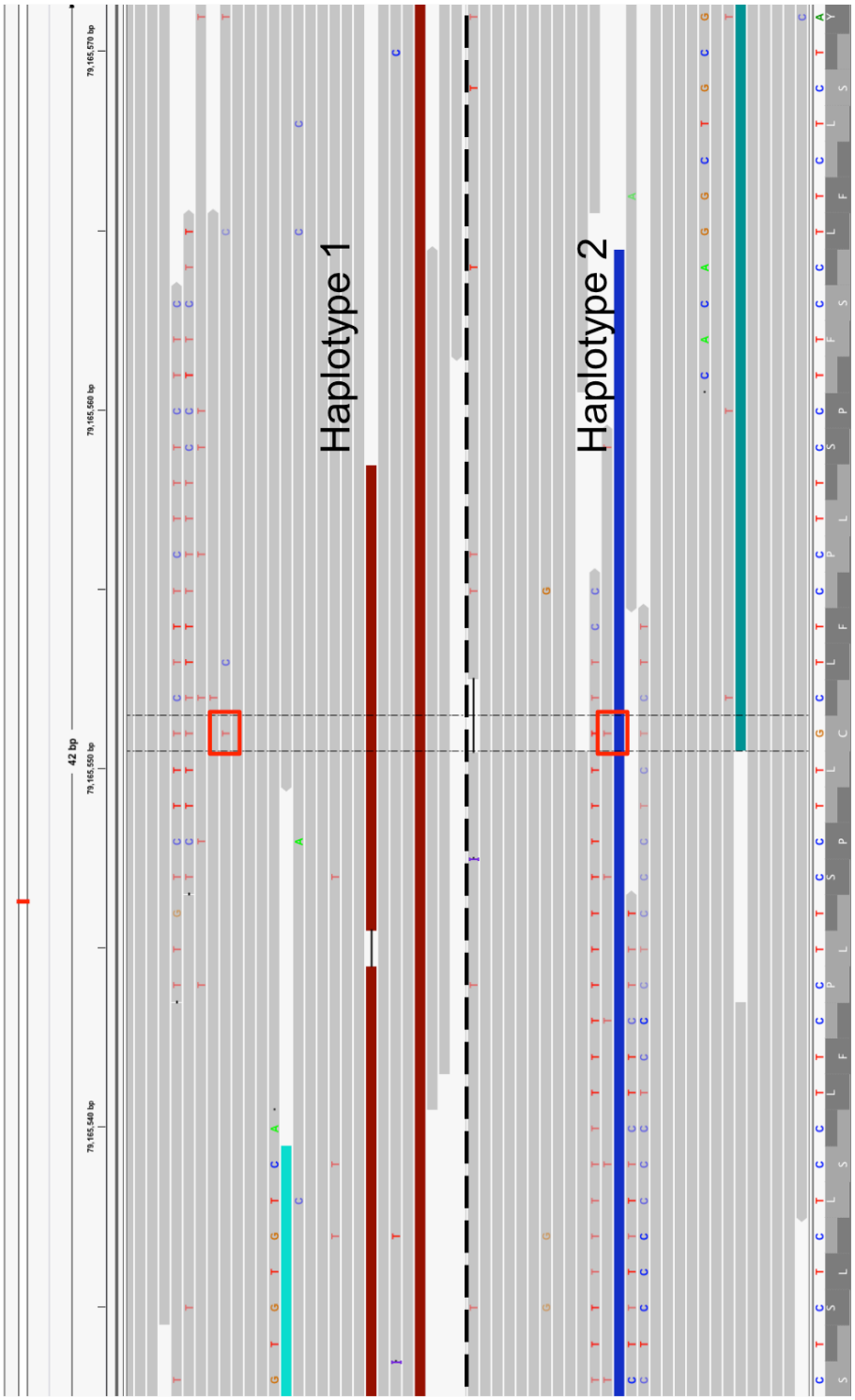

Haplotype 1
Haplotype 2
42 bp
79,165,570 bp
79,165,560 bp
79,165,550 bp
79,165,540 bp

Figure 2.10: Examples of manual inspection of mosaic SNVs (10X)
Somatic SNV candidate with manual inspection on chr3: 79165551. The allele frequency of this cite in brain WGS data is 0.066, and in NA12878 is 0. This candidate has a relatively low allele frequency. However, the alternative allele is on both of the haplotypes from the 10X data. Thus, this is a false positive candidate. This IGV screen shot presents the 10X Genomics data of this candidate. The black dashed line splits the reads belonging to two haplotypes. Two red boxes show the alternative allele present in both of the haplotypes.

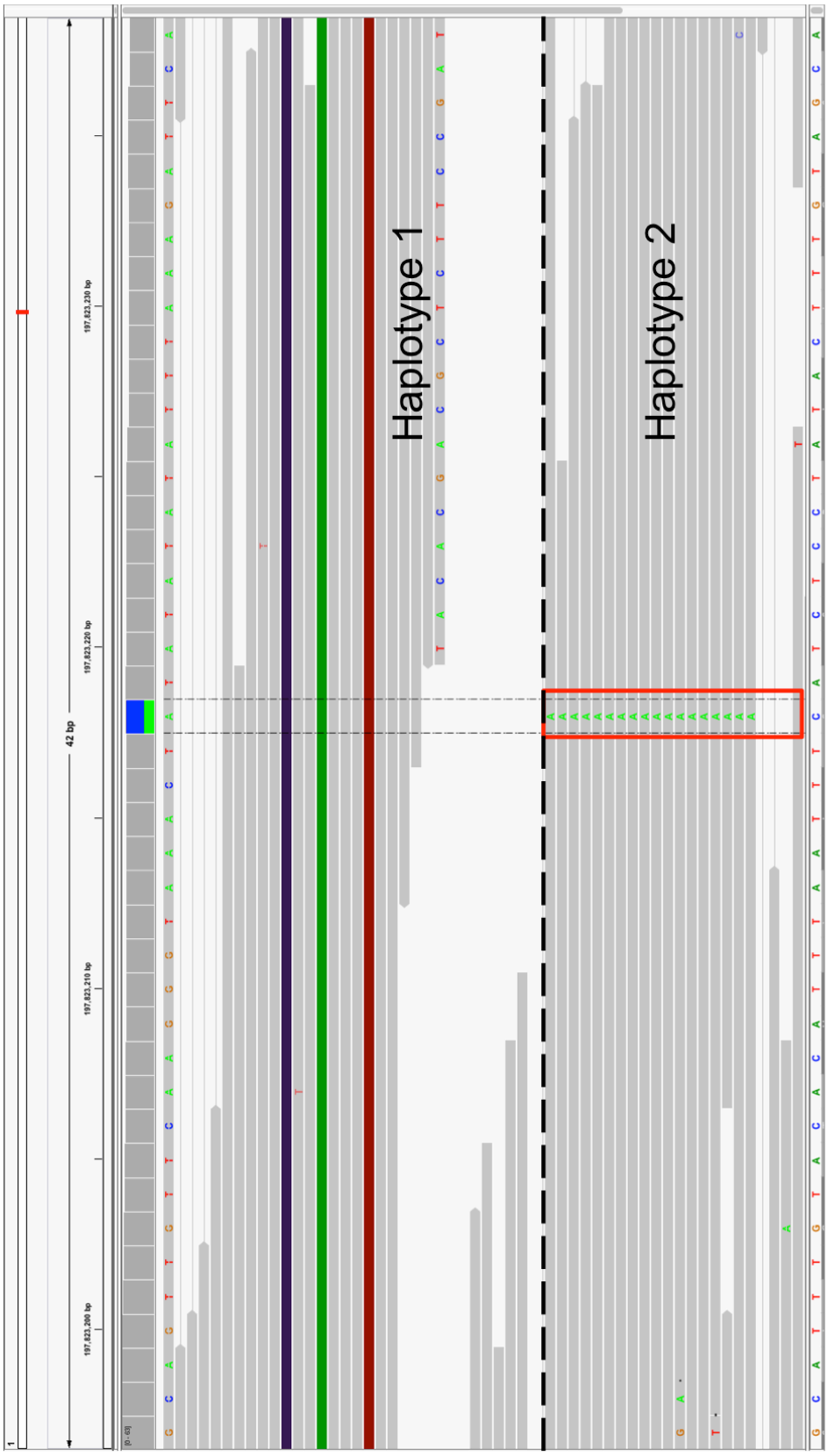

Haplotype 1
Haplotype 2

Figure 2.11: Examples of manual inspection of mosaic SNVs (10X)
Somatic SNV candidate with manual inspection on chr1: 197823218. The allele frequency of this cite in brain WGS data is 0.271, and in NA12878 is 0. This candidate has a high allele frequency reasonable somatic SNV candidate. However, from the 10X haplotype information, we observed that all alternative alleles on this site are on one haplotype only and take all of the reads in this haplotype. Thus, we identified this site as a germline SNV instead of a somatic SNV. This IGV screen shot presents the 10X Genomics data of this candidate. The black dashed line splits the reads belonging to two haplotypes. The red box shows the alternative allele takes the majority of haplotype 2, showing that this is a germline event.


91,717,490 bp
91,717,500 bp
91,717,510 bp
91,717,520 bp
42 bp

Figure 2.12: Examples of manual inspection of mosaic SNVs (homopolymer)
Somatic SNV candidate with manual inspection on chr4: 91717508. The allele frequency of this cite in brain WGS data is 0.08, and in NA12878 is 0. However, this is a low allele frequency candidate inside a string of polyT region. This site is a possible false positive candidate. The screen shot shows the candidate position in the brain alignment files with the highest coverage. This base change alters an ‘A’ in a string of ‘T’ to ‘T’.

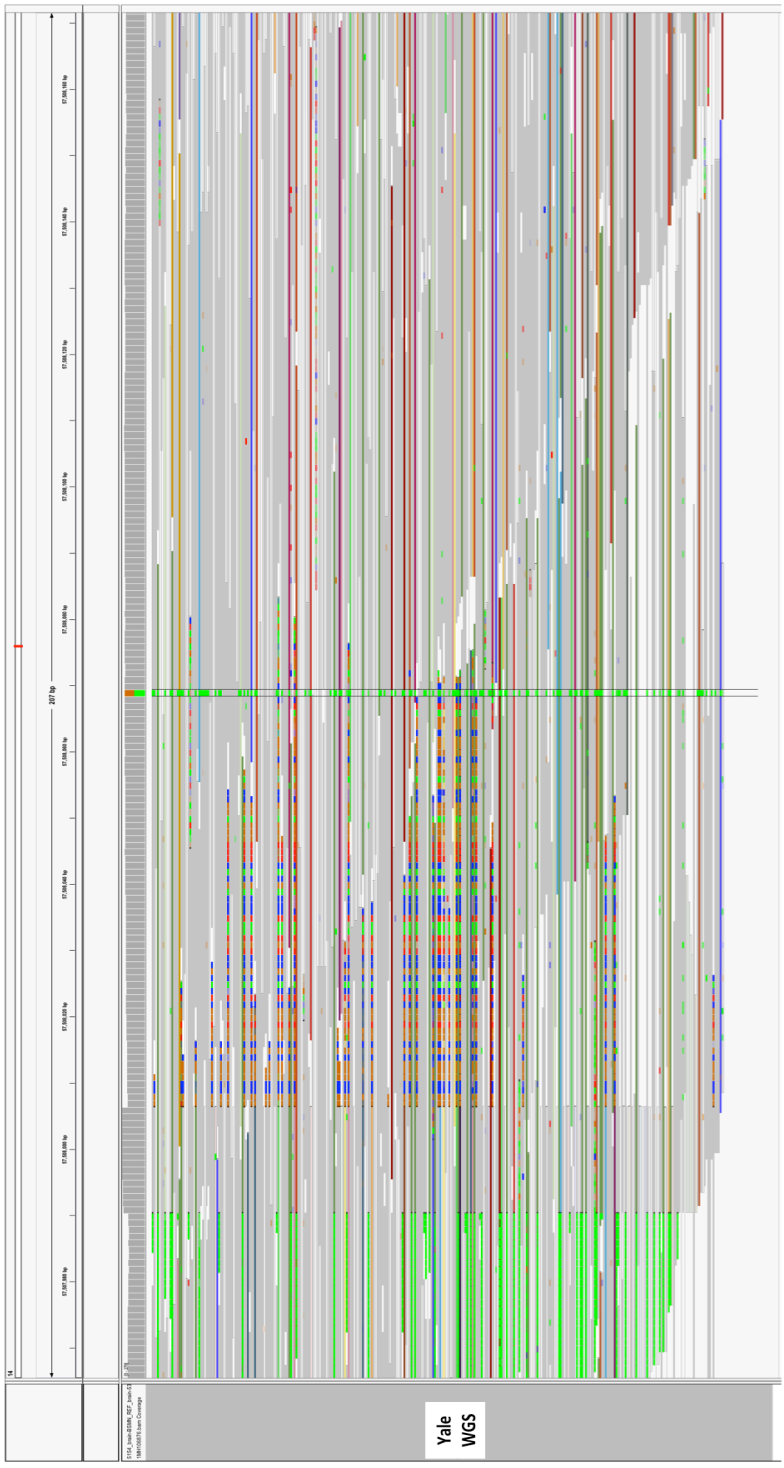

Yale
WGS

Figure 2.13: Examples of manual inspection of mosaic SNVs (SV)
Somatic SNV candidate with manual inspection on chr14: 57508069. The allele frequency of this cite in brain WGS data is 0.293, and in NA12878 is 0. This site is a reasonable candidate from all other aspects. However, from the colored clipped reads in the IGV screenshot, we could observe that the structural variation around this region, which caused a false positive call as a somatic SNV at this position. The grey part of reads shows a perfect match with the reference. The same colored pattern shows the existence of an insertion in this region.

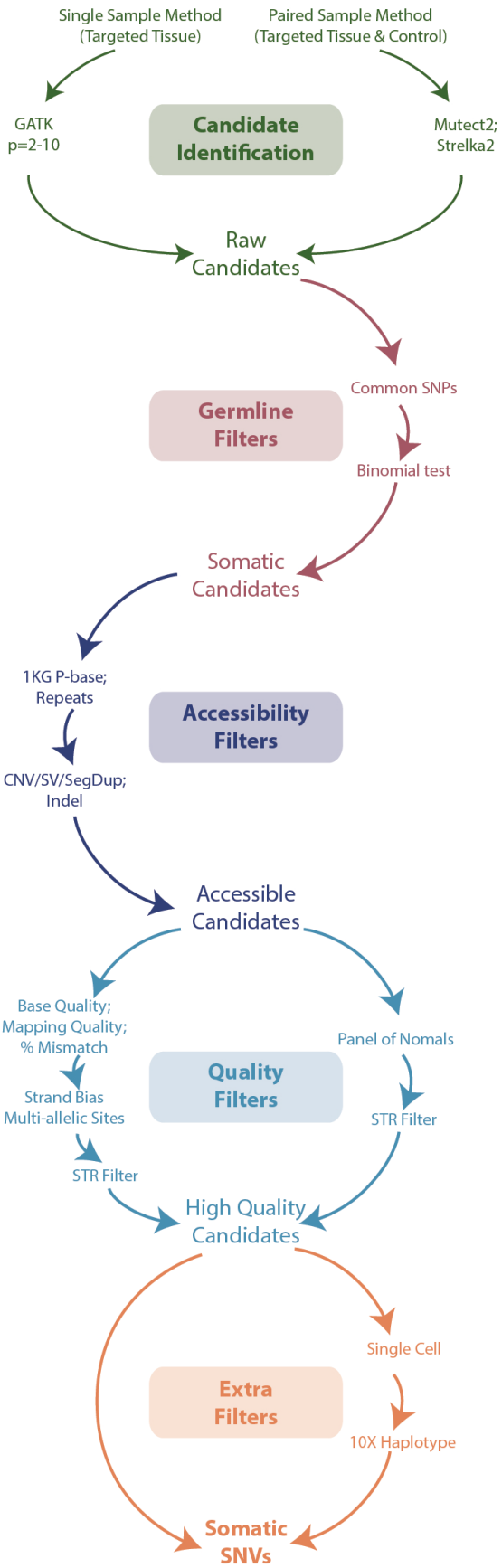

Single Sample Method
(Targeted Tissue)
Paired Sample Method
(Targeted Tissue & Control)
GATK
p=2-10
Candidate Identification
Mutect2;
Strelka2
Raw Candidates
Common SNPs
Germline Filters
Binomial test
Somatic Candidates
1KG P-base;
Repeats
Accessibility Filters
CNV/SV/SegDup;
Indel
Accessible Candidates
Base Quality;
Mapping Quality;
% Mismatch
Panel of Nomals
Quality Filters
Strand Bias
Multi-allelic Sites
STR Filter
STR Filter
High Quality Candidates
Single Cell
Extra Filters
10X Haplotype
Somatic SNVs

Figure 2.14: Best practice for identification of mosaic SNVs.
The best practice that we suggest applying in order to identify somatic SNVs in non-tumor tissue includes two possible kinds of input. If only one target sample is sequenced, we would suggest utilizing GATK with ploidy=2-10 to obtain the raw candidates as input for further filtering pipeline. If the targeted sample is sequenced with a paired sample from the same individual, we would suggest applying Mutect, Strelka and GATK with ploidy=2-10 as raw candidate since Mutect and Strelka have been demonstrated to have higher sensitivity at low allele frequency regions with a paired sample. We then would like to apply the germline filters including common SNPs from gnomAD and binomial test that could exclude possible germline events from GATK in particular. After excluding possible germline SNVs, we then suggest applying a series of quality filters including base quality, mapping quality and percent mismatch of reads, together with strand bias Fisher Exact test and candidates as multi-allelic sites, NA12878 allele frequency to exclude possible false positive candidates brought by sequencing artifacts. Here, if a panel of unrelated normal samples is available, the complex quality filters could be replaced with the panel of normal sites. The hypothesis here is that a true somatic SNV site in one individual should not be identified in another unrelated sample. In the end, if single cell data or 10X genomics sequencing data is available, we could utilize the extra information provided by these two different sequencing libraries to exclude possible false positives as described above in methods and results.

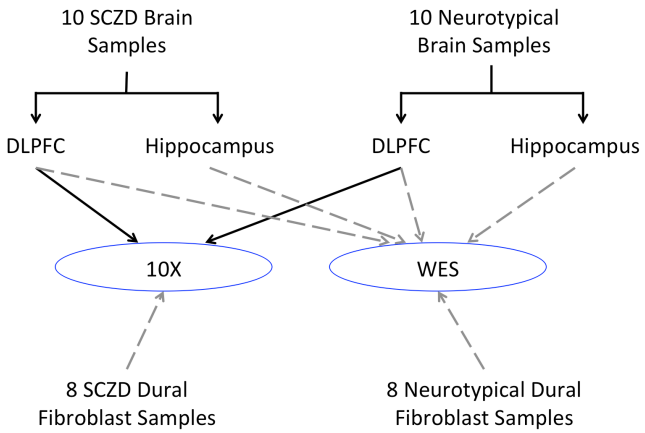


Figure 2.15: Schizophrenia related mosaic SNV identification experiment design.

Table2.1: Consortium common experiment data summary.

Data Source	Tissue	Library Type	Coverage	Standard Deviation
UM	Brain	WES	352.38	218.44
UM	Dural Fibroblast	WES	447.82	250.60
UCSD	Brain	WES	432.89	251.65
UCSD	Dural Fibroblast	WES	145.70	85.38
Yale-1	Brain	WGS	74.59	13.72
Harvard	Brain	WGS	189.00	33.87
LIBD	Brain	WGS	90.44	13.77
Yale-2	Brain	WGS	236.27	33.45
Yale-2	Dural Fibroblast	WGS	197.41	33.15
Mt. Sinai	Dural Fibroblast	WGS	19.99	5.5

Table 2.2: Categories of mosaic SNVs identified from different institutions.

Categories	Definition
Absolute Singletons	Identified by one approach, supportive evidence from one data source
Data Source Singletons	Identified by multiple approaches, supportive evidence from one data source
Approach Singletons	Identified by one approach, supportive evidence from multiple data sources
Multi-calls	Identified by multiple approaches, supportive evidence from multiple data sources

Table 2.3: Summary of categories of SNVs identified from different institutions.

	Number of Total Candidates	Number of Absolute Singletons	Number of Data Source Singletons	Number of Approach Singletons	Number of Multi-calls
Abyzov	100	4	2	63	31
Gleeson	12	0	0	11	1
Moran	1148	135	3	979	31
Park	53	5	2	13	33
Pevsner	57	3	1	24	29
Sestan	16	0	0	9	7
Total	1298	148	4	1101	45

Table2.4: Amplicon sequencing library specific sequencing error cutoffs.

	A>T & T>A	A>C & T>G	A>G & T>C	C>A & G>T	C>T & G>A	C>G & G>C
MI	0.000764	0.000408	0.000927	0.000671	0.00132	0.000248
Mayo	0.000440	0.000139	0.000511	0.000609	0.00106	0.000151
KKI	0.000373	0.000248	0.00229	0.000249	0.000989	0.000125
MtSinai	0.000611	0.00199	0.00251	0.00276	0.00138	0.000269
Harvard	0	0	0.000126	0.000130	0.000142	0

Table 2.5: Mosaic SNV candidate sites for ddPCR.

chrm	pos	ref	alt	comments	category
8	2210830	C	G	amplicon seq shows real, possible CNV	ambiguous
1	197823218	C	A	fail 10X, but all other evidence show true positive	ambiguous
14	38720782	C	T	low CAF in WGS data, and messy in amplicon seq	ambiguous
3	59577834	G	A	possible indel	ambiguous
6	47565225	G	C	possible SV around	ambiguous
18	39944439	G	A	messy region, third allele frequency relatively high	ambiguous
18	12614010	A	G	high CAF in NA12878	FP
9	100100349	G	T	amplicon seq looks ok, but found in other 17 samples	FP
14	35629120	G	A	fail 10X, all other evidence show a weak true positive	FP
5	58367359	A	T	indel downstream close by	FP
14	35673152	A	T	fail 10X, all other evidence show a weak true positive	FP
4	91717508	A	T	A in a poly T string	FP
3	71303452	G	A	Relatively low quality in amplicon seq	PASS
7	112009378	G	A	Relatively low quality in amplicon seq	PASS
15	46636659	C	G	strong pass, relatively low CAF	PASS
3	150634574	G	A	strong pass, relatively low CAF	PASS
4	3857654	C	T	strong pass, high CAF	PASS
14	72870518	A	G	strong pass, high CAF	PASS

Table 2.6: Amplicon validation for 400 mosaic SNVs.

chrm	pos	ref	alt	status1	brain_fq1	status2	brain_fq2	Decision
1	185567840	G	A	PASS	0.0063	Not Assigned	Not Assigned	PASS
1	197823218	C	A	PASS	0.2802	Not Assigned	Not Assigned	PASS
1	205408125	C	T	PASS	0.0096	Not Assigned	Not Assigned	PASS
2	5199906	C	T	PASS	0.0197	Not Assigned	Not Assigned	PASS
2	86485588	C	T	PASS	0.0234	Not Assigned	Not Assigned	PASS
3	10935924	G	C	PASS	0.0089	Not Assigned	Not Assigned	PASS
3	24256930	C	T	PASS	0.0292	Not Assigned	Not Assigned	PASS
3	66745914	C	T	PASS	0.0565	PASS	0.0571	PASS
3	71303452	G	A	PASS	0.1156	Not Assigned	Not Assigned	PASS
3	150634574	G	A	PASS	0.0128	PASS	0.0206	PASS
3	178541243	C	T	PASS	0.0380	Not Assigned	Not Assigned	PASS
4	3857654	C	T	PASS	0.2832	PASS	0.2731	PASS
4	45920691	C	T	PASS	0.0018	Not Assigned	Not Assigned	PASS
5	113236393	C	A	PASS	0.0053	Not Assigned	Not Assigned	PASS
5	116767707	G	A	PASS	0.0001	Not Assigned	Not Assigned	PASS
6	83968771	G	C	PASS	0.0224	Not Assigned	Not Assigned	PASS
6	96086198	A	C	PASS	0.0001	Not Assigned	Not Assigned	PASS
6	100611507	C	A	PASS	0.0094	Not Assigned	Not Assigned	PASS
6	116873861	A	T	PASS	0.0154	Not Assigned	Not Assigned	PASS
7	54489183	C	A	PASS	0.2346	Not Assigned	Not Assigned	PASS
7	80158107	C	T	PASS	0.0049	PASS	0.0178	PASS
7	90545084	G	A	PASS	0.0665	Not Assigned	Not Assigned	PASS
7	110664350	C	T	PASS	0.0125	PASS	0.0153	PASS
7	112009378	G	A	PASS	0.0101	Not Assigned	Not Assigned	PASS
7	112461481	G	T	PASS	0.0002	Not Assigned	Not Assigned	PASS
7	148726847	C	T	PASS	0.0224	Not Assigned	Not Assigned	PASS
8	91413335	G	A	PASS	0.0238	Not Assigned	Not Assigned	PASS
8	103281483	C	T	PASS	0.0116	Not Assigned	Not Assigned	PASS
8	113973702	A	G	PASS	0.1799	Not Assigned	Not Assigned	PASS
8	126390601	G	A	PASS	0.0312	PASS	0.0441	PASS
10	1917663	C	T	PASS	0.0235	Not Assigned	Not Assigned	PASS
10	92165549	C	T	PASS	0.0289	Not Assigned	Not Assigned	PASS
10	107663441	T	C	PASS	0.0468	Not Assigned	Not Assigned	PASS
11	74361284	T	C	PASS	0.0182	Not Assigned	Not Assigned	PASS
11	97180935	G	A	PASS	0.0371	PASS	0.0410	PASS
11	120386729	C	T	failseqerr	0.0001	PASS	0.0247	PASS
12	65791007	C	T	PASS	0.0122	Not Assigned	Not Assigned	PASS
14	21125803	G	A	PASS	0.0059	Not Assigned	Not Assigned	PASS
14	49010001	A	G	PASS	0.0324	Not Assigned	Not Assigned	PASS
15	46636659	C	G	PASS	0.0183	PASS	0.0183	PASS

16	64683305	G	A	PASS	0.0246	Not Assigned	Not Assigned	PASS
17	44885146	G	A	PASS	0.0230	Not Assigned	Not Assigned	PASS
18	5481334	A	G	PASS	0.2126	Not Assigned	Not Assigned	PASS
19	9493288	G	A	PASS	0.0055	Not Assigned	Not Assigned	PASS
1	34260647	A	G	readsnotenough	0.0000	Not Sequenced	Not Sequenced	notenoughdata
1	51679303	T	G	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
1	179464991	A	T	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
1	191297715	C	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
1	226036743	T	G	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
2	43452819	C	T	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
2	230714651	C	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
3	192544607	A	T	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
5	103990701	T	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
6	10920813	G	A	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
6	17416384	T	C	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
6	76180939	T	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
7	71907794	A	G	readsnotenough	0.1972	Not Sequenced	Not Sequenced	notenoughdata
7	114539432	T	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
7	150344734	A	C	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
7	157688045	T	A	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
9	91408737	T	A	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
9	127255300	A	G	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
10	46970573	C	G	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
10	79430930	T	G	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
10	87963307	A	C	readsnotenough	0.0000	readsnotenough	0	notenoughdata
11	61111724	A	G	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
11	68688607	A	T	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
11	83548887	G	T	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
12	34510345	T	G	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
12	127840500	C	T	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
14	92255639	G	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
14	105816410	G	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
16	597627	C	G	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
16	19193644	T	G	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
16	47043306	G	T	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
17	38114199	T	G	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
19	58695924	A	C	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
20	3656538	A	C	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
21	26267966	T	A	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
X	1027504	T	G	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
X	6234721	A	T	readsnotenough	0.0000	Not Assigned	Not Assigned	notenoughdata
X	107770019	A	T	Not Sequenced	Not Sequenced	Not Assigned	Not Assigned	notenoughdata
2	81089746	T	C	failNA12878skellam	0.0002	failseqerr	0	not-decided

3	59577834	G	A	PASS	0.0039	readsnotenough	0.0105	not-decided
4	188330530	A	G	PASS	0.0018	failNA12878skellam	0.0017	not-decided
6	47565225	G	C	PASS	0.0066	Not Assigned	Not Assigned	not-decided
9	100100349	G	T	PASS	0.0084	Not Sequenced	Not Sequenced	not-decided
14	38720782	C	T	PASS	0.0112	Not Assigned	Not Assigned	not-decided
14	72870518	A	G	readsnotenough	0.2523	Not Assigned	Not Assigned	not-decided
14	97512875	A	C	PASS	0.0012	failNA12878skellam	0	not-decided
17	71395357	G	A	PASS	0.0002	Not Assigned	Not Assigned	not-decided
18	12614010	A	G	PASS	0.0098	Not Assigned	Not Assigned	not-decided
18	39944439	G	A	PASS	0.0098	Not Assigned	Not Assigned	not-decided
6	135199781	A	G	PASS	0.0647	Not Assigned	Not Assigned	germline
10	47066579	A	G	PASS	0.0320	Not Assigned	Not Assigned	germline
14	73148541	C	G	PASS	0.3635	Not Assigned	Not Assigned	germline
18	5499876	C	T	PASS	0.0172	Not Sequenced	Not Sequenced	germline
8	2210830	C	G	PASS	0.0486	Not Sequenced	Not Sequenced	false_positive
10	485315	G	A	PASS	0.0005	PASS	0.0073	false_positive
17	44198245	T	C	PASS	0.3268	Not Assigned	Not Assigned	false_positive
1	10798144	T	G	PASS	0.0001	Not Assigned	Not Assigned	false_positive
1	10871477	C	T	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
1	16778457	A	G	failseqerr	0.0021	Not Assigned	Not Assigned	false_positive
1	20799282	A	C	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
1	29788830	A	C	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
1	33562483	T	C	failNA12878skellam	0.0008	Not Assigned	Not Assigned	false_positive
1	35366547	T	G	failNA12878skellam	0.0010	Not Assigned	Not Assigned	false_positive
1	39021614	T	G	failseqerr	0.0004	failNA12878skellam	0	false_positive
1	54653802	T	G	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
1	71214596	A	T	failNA12878	0.0313	Not Assigned	Not Assigned	false_positive
1	72604782	C	A	failNA12878skellam	0.0030	Not Assigned	Not Assigned	false_positive
1	78401682	G	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
1	86383228	A	T	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
1	86383232	A	T	failNA12878skellam	0.0004	Not Sequenced	Not Sequenced	false_positive
1	107319979	T	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
1	108865995	A	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
1	147381393	C	A	failNA12878skellam	0.0018	Not Assigned	Not Assigned	false_positive
1	157532721	T	G	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
1	157636883	G	T	failNA12878	0.1412	Not Assigned	Not Assigned	false_positive
1	160334919	A	C	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
1	188615921	T	A	failNA12878skellam	0.0002	failNA12878skellam	0.0001	false_positive
1	198161499	A	T	failNA12878skellam	0.0023	Not Assigned	Not Assigned	false_positive
1	217196186	T	G	failNA12878skellam	0.0009	failNA12878skellam	0	false_positive
1	220384078	A	G	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
1	226549266	C	T	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
1	230186624	G	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive

1	232822257	A	G	failNA12878skellam	0.0005	Not Assigned	Not Assigned	false_positive
1	239503530	C	A	failseqerr	0.0000	PASS	0.0104	false_positive
1	249103953	A	C	failNA12878skellam	0.0004	Not Assigned	Not Assigned	false_positive
2	3479966	A	G	failfreq	0.5079	Not Assigned	Not Assigned	false_positive
2	20196958	C	T	Not Sequenced	Not Sequenced	failNA12878skellam	0.0007	false_positive
2	20562971	T	C	failNA12878skellam	0.0008	Not Assigned	Not Assigned	false_positive
2	26943588	T	G	PASS	0.1841	Not Assigned	Not Assigned	false_positive
2	37280594	A	T	failseqerr	0.0000	readsnotenough	0	false_positive
2	39629939	C	A	failNA12878skellam	0.0006	Not Assigned	Not Assigned	false_positive
2	44910354	C	A	PASS	0.0024	Not Assigned	Not Assigned	false_positive
2	51248705	C	T	failNA12878skellam	0.0050	Not Assigned	Not Assigned	false_positive
2	63123790	G	T	failseqerr	0.0004	failseqerr	0	false_positive
2	80526502	A	C	PASS	0.0038	Not Assigned	Not Assigned	false_positive
2	82511802	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
2	113048297	G	C	failNA12878	0.2733	Not Assigned	Not Assigned	false_positive
2	122678719	C	T	PASS	0.0022	Not Assigned	Not Assigned	false_positive
2	135736455	A	C	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
2	163408355	T	G	failseqerr	0.0005	failNA12878skellam	0	false_positive
2	165433304	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
2	200122611	T	G	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
2	202107935	G	T	failseqerr	0.0000	readsnotenough	0	false_positive
2	222030970	A	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
2	222270513	G	T	failNA12878skellam	0.0093	Not Assigned	Not Assigned	false_positive
2	231669920	A	C	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
2	233445561	T	G	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
2	235859132	A	C	failNA12878skellam	0.0000	failNA12878skellam	0.0003	false_positive
3	1083092	C	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
3	6091420	G	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
3	13298985	A	C	failNA12878skellam	0.0001	failseqerr	0	false_positive
3	13762684	G	T	failNA12878skellam	0.0004	Not Assigned	Not Assigned	false_positive
3	26395998	A	C	failNA12878skellam	0.0000	failNA12878skellam	0.0031	false_positive
3	30897421	C	A	PASS	0.0015	failNA12878skellam	0.0009	false_positive
3	33291623	T	A	failNA12878skellam	0.0005	Not Assigned	Not Assigned	false_positive
3	38609519	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
3	41513485	T	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
3	51231319	A	T	readsnotenough	0.0000	failNA12878skellam	0.0006	false_positive
3	54098845	T	C	failfreq	0.5075	Not Assigned	Not Assigned	false_positive
3	58155054	T	G	failNA12878skellam	0.0033	failNA12878skellam	0.0010	false_positive
3	59889823	C	T	PASS	0.2261	PASS	0.0467	false_positive
3	79165551	G	T	PASS	0.0201	Not Assigned	Not Assigned	false_positive
3	83749180	C	T	failNA12878skellam	0.0005	Not Assigned	Not Assigned	false_positive
3	94050943	A	C	failNA12878	0.2759	failNA12878skellam	0.0005	false_positive
3	126864130	A	C	PASS	0.0033	failNA12878skellam	0	false_positive

3	153160150	G	T	failNA12878skellam	0.0000	failNA12878skellam	0.0016	false_positive
3	175290702	A	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
3	181631534	C	T	failNA12878skellam	0.0008	Not Assigned	Not Assigned	false_positive
3	185839534	T	G	failNA12878skellam	0.0011	Not Assigned	Not Assigned	false_positive
4	962222	G	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
4	10076371	T	G	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
4	29265883	T	C	failfreq	0.4875	failfreq	0.5044	false_positive
4	56435894	T	G	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
4	86171870	A	T	PASS	0.0012	Not Assigned	Not Assigned	false_positive
4	86293704	A	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
4	90877885	G	T	failNA12878skellam	0.0015	Not Assigned	Not Assigned	false_positive
4	91717508	A	T	PASS	0.0100	failNA12878	0.0189	false_positive
4	102289690	C	A	failseqerr	0.0000	readsnotenough	0	false_positive
4	113184134	T	G	failseqerr	0.0000	failNA12878skellam	0	false_positive
4	115934226	T	G	failNA12878	0.0137	Not Assigned	Not Assigned	false_positive
4	134759331	A	T	PASS	0.0005	Not Assigned	Not Assigned	false_positive
4	142366461	C	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
4	153129393	A	C	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
4	153418587	T	A	failseqerr	0.0000	readsnotenough	0.0258	false_positive
4	167193426	G	T	failNA12878skellam	0.0004	readsnotenough	0	false_positive
4	174856826	T	A	PASS	0.0000	failNA12878skellam	0.0001	false_positive
5	2628236	T	G	PASS	0.0000	Not Assigned	Not Assigned	false_positive
5	3559828	G	A	failNA12878skellam	0.0002	failNA12878skellam	0.0006	false_positive
5	27862789	C	A	PASS	0.0397	failNA12878skellam	0	false_positive
5	38515736	T	G	failNA12878skellam	0.0036	Not Assigned	Not Assigned	false_positive
5	50971730	G	T	failNA12878skellam	0.0004	Not Assigned	Not Assigned	false_positive
5	58255846	A	G	failfreq	0.5086	readsnotenough	0	false_positive
5	58367359	A	T	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
5	67767155	C	T	PASS	0.0001	Not Assigned	Not Assigned	false_positive
5	77570181	T	A	failNA12878skellam	0.0002	failNA12878skellam	0	false_positive
5	108964023	G	T	failseqerr	0.0004	Not Assigned	Not Assigned	false_positive
5	126215731	T	G	failNA12878skellam	0.0005	Not Assigned	Not Assigned	false_positive
5	138406395	C	A	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
5	147821878	A	C	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
5	160807871	T	A	failNA12878skellam	0.0018	Not Assigned	Not Assigned	false_positive
5	165664363	T	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
5	167143878	G	T	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
5	169417100	T	C	failseqerr	0.0018	Not Assigned	Not Assigned	false_positive
6	6605071	G	A	failNA12878skellam	0.0041	failNA12878skellam	0.0025	false_positive
6	16943118	T	A	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
6	18811308	C	T	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
6	35755658	A	G	failNA12878	0.3321	Not Assigned	Not Assigned	false_positive
6	35765151	T	C	failNA12878	0.3455	Not Assigned	Not Assigned	false_positive

6	35765739	T	C	failNA12878	0.2926	failNA12878	0.3326	false_positive
6	63374625	G	A	failNA12878skellam	0.0031	PASS	0.0054	false_positive
6	94486269	A	C	failNA12878skellam	0.0000	failNA12878skellam	0	false_positive
6	99224229	C	A	PASS	0.0016	failNA12878skellam	0	false_positive
6	102546118	G	T	failNA12878skellam	0.0001	failseqerr	0	false_positive
6	118360881	A	T	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
6	161263429	A	T	PASS	0.0092	Not Assigned	Not Assigned	false_positive
7	7555467	A	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
7	14823911	G	T	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
7	17008453	A	G	failNA12878skellam	0.0018	Not Assigned	Not Assigned	false_positive
7	29185157	T	C	failseqerr	0.0000	PASS	0.0172	false_positive
7	32920141	C	A	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
7	50629380	A	C	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
7	71176992	C	T	failNA12878skellam	0.0013	Not Assigned	Not Assigned	false_positive
7	85338850	G	T	failNA12878skellam	0.0013	Not Assigned	Not Assigned	false_positive
7	87519130	A	G	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
7	113842146	G	A	failseqerr	0.0000	failNA12878skellam	0.0006	false_positive
7	125142865	A	T	failNA12878skellam	0.0015	Not Assigned	Not Assigned	false_positive
7	134177108	C	T	failNA12878skellam	0.0011	Not Assigned	Not Assigned	false_positive
7	136349550	G	T	failNA12878skellam	0.0001	failNA12878skellam	0.0003	false_positive
7	137000098	G	A	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
7	144833026	A	T	Not Sequenced	Not Sequenced	failNA12878skellam	0	false_positive
7	150438324	C	A	failNA12878skellam	0.0023	Not Assigned	Not Assigned	false_positive
7	150545388	A	C	failNA12878	0.0817	Not Assigned	Not Assigned	false_positive
7	157705850	G	T	failNA12878skellam	0.0010	Not Assigned	Not Assigned	false_positive
8	6264017	G	T	failseqerr	0.0000	failNA12878skellam	0	false_positive
8	8369109	T	G	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
8	14316984	C	A	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
8	19571965	A	C	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
8	23856152	A	T	failNA12878	0.3603	Not Assigned	Not Assigned	false_positive
8	25641757	C	A	PASS	0.0003	Not Assigned	Not Assigned	false_positive
8	35065083	C	A	failseqerr	0.0018	Not Assigned	Not Assigned	false_positive
8	38160548	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
8	38801168	G	T	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
8	65817572	C	A	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
8	83340122	T	A	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
8	84839623	A	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
8	109397075	C	A	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
8	131324707	A	T	failseqerr	0.0004	Not Assigned	Not Assigned	false_positive
8	134322326	C	A	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
8	134972322	A	T	failNA12878	0.4986	Not Assigned	Not Assigned	false_positive
8	140655717	A	C	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
9	16825053	C	T	PASS	0.0202	Not Assigned	Not Assigned	false_positive

9	30336489	C	T	failseqerr	0.0001	Not Assigned	Not Assigned	false_positive
9	35357548	G	T	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
9	86540378	G	A	failNA12878skellam	0.0004	Not Assigned	Not Assigned	false_positive
9	102532443	G	A	failNA12878skellam	0.0008	failNA12878skellam	0	false_positive
9	112093683	T	G	failNA12878skellam	0.0006	Not Assigned	Not Assigned	false_positive
9	138485889	G	A	failNA12878skellam	0.0013	failNA12878skellam	0.0009	false_positive
10	23922480	A	C	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
10	36938391	C	T	failNA12878skellam	0.0005	failNA12878skellam	0.0004	false_positive
10	37538722	G	A	failseqerr	0.0000	failNA12878skellam	0.0007	false_positive
10	43959293	G	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
10	54987087	T	A	failNA12878skellam	0.0005	Not Assigned	Not Assigned	false_positive
10	66110752	G	A	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
10	68356850	C	A	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
10	71162274	A	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
10	87257726	A	T	failseqerr	0.0000	failNA12878skellam	0.0002	false_positive
10	116088338	A	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
10	126137142	A	C	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
10	126349585	T	A	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
11	200035	G	A	failfreq	0.5004	Not Assigned	Not Assigned	false_positive
11	16092704	T	C	failNA12878skellam	0.0028	Not Assigned	Not Assigned	false_positive
11	24813517	C	T	failNA12878	0.0119	Not Assigned	Not Assigned	false_positive
11	31419808	A	T	PASS	0.0000	Not Assigned	Not Assigned	false_positive
11	46752346	A	G	PASS	0.0205	Not Assigned	Not Assigned	false_positive
11	55447282	G	T	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
11	64078628	A	G	failfreq	0.4883	failfreq	0.4780	false_positive
11	64326584	G	C	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
11	71100137	A	C	PASS	0.0000	failNA12878skellam	0.0007	false_positive
11	83545514	G	C	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
11	115225623	A	G	failNA12878skellam	0.0012	Not Assigned	Not Assigned	false_positive
11	116460961	A	C	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
12	1446527	A	T	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
12	12959688	T	A	PASS	0.0064	Not Assigned	Not Assigned	false_positive
12	14693158	T	A	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
12	23981687	C	A	failNA12878skellam	0.0009	failNA12878skellam	0.0004	false_positive
12	25260574	G	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
12	46299033	G	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
12	72122667	C	A	failNA12878skellam	0.0028	Not Assigned	Not Assigned	false_positive
12	93969049	A	C	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
12	95971001	A	C	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
12	99108593	A	T	failNA12878skellam	0.0036	PASS	0	false_positive
12	100445423	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
12	101724340	A	G	readsnotenough	0.0000	failNA12878skellam	0.0002	false_positive
12	103675636	T	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive

12	118255566	T	C	failNA12878skellam	0.0057	Not Assigned	Not Assigned	false_positive
12	120079309	T	C	failNA12878skellam	0.0004	Not Assigned	Not Assigned	false_positive
12	124496153	C	T	failseqerr	0.0000	failNA12878	0.2711	false_positive
13	37244385	G	T	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
13	48663332	A	C	failNA12878skellam	0.0008	failNA12878skellam	0.0019	false_positive
13	49876270	T	G	readsnotenough	0.0000	failseqerr	0.0016	false_positive
13	79778665	T	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
13	96820946	C	A	failNA12878skellam	0.0053	Not Assigned	Not Assigned	false_positive
13	99006094	G	T	failNA12878skellam	0.0005	Not Assigned	Not Assigned	false_positive
13	107086032	A	C	failNA12878	0.0354	Not Assigned	Not Assigned	false_positive
13	112570956	T	C	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
13	113231211	T	A	failNA12878	0.0194	Not Assigned	Not Assigned	false_positive
14	20216484	A	C	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
14	34493324	G	T	PASS	0.0116	failNA12878	0.0368	false_positive
14	35413420	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
14	35629120	G	A	PASS	0.3290	Not Assigned	Not Assigned	false_positive
14	35673152	A	T	PASS	0.3407	Not Assigned	Not Assigned	false_positive
14	38680450	A	C	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
14	43752350	T	G	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
14	43980974	C	T	failNA12878skellam	0.0006	readsnotenough	0	false_positive
14	53517817	C	A	failNA12878skellam	0.0057	PASS	0.0141	false_positive
14	57508069	G	A	PASS	0.1054	Not Assigned	Not Assigned	false_positive
14	64460551	A	T	PASS	0.0151	Not Assigned	Not Assigned	false_positive
14	66263885	C	A	failseqerr	0.0001	Not Assigned	Not Assigned	false_positive
14	67356563	T	A	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
14	68284524	A	G	failNA12878skellam	0.0002	readsnotenough	0	false_positive
14	94184622	A	C	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
14	104250100	A	C	failNA12878skellam	0.0005	Not Assigned	Not Assigned	false_positive
15	23931814	C	T	failNA12878skellam	0.0008	Not Assigned	Not Assigned	false_positive
15	42801562	C	A	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
15	48599623	C	A	failseqerr	0.0000	failNA12878skellam	0	false_positive
15	56931486	C	A	failNA12878skellam	0.0000	failNA12878skellam	0.0014	false_positive
15	81002072	A	C	failNA12878skellam	0.0019	failNA12878skellam	0.0002	false_positive
15	83574009	T	A	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
15	85150677	A	C	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
15	89819945	T	G	failNA12878skellam	0.0000	failNA12878skellam	0	false_positive
15	91611998	C	T	failNA12878skellam	0.0051	Not Assigned	Not Assigned	false_positive
15	98304382	A	G	failNA12878skellam	0.0027	failNA12878skellam	0.0063	false_positive
16	148328	C	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
16	1588998	G	A	Not Sequenced	Not Sequenced	failNA12878	0.0012	false_positive
16	16730657	C	T	failNA12878	0.1567	Not Assigned	Not Assigned	false_positive
16	18040558	C	T	failfreq	0.4808	Not Assigned	Not Assigned	false_positive
16	27475614	T	G	PASS	0.0000	Not Assigned	Not Assigned	false_positive

16	32586271	A	C	failNA12878	0.3238	Not Assigned	Not Assigned	false_positive
16	47556462	A	T	failNA12878skellam	0.0059	Not Assigned	Not Assigned	false_positive
16	49040091	A	C	failseqerr	0.0004	failseqerr	0	false_positive
16	50836894	G	C	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
16	58906982	A	T	failNA12878skellam	0.0014	Not Sequenced	Not Sequenced	false_positive
16	62089393	A	T	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
16	68469512	T	G	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
16	69295259	G	T	failNA12878skellam	0.0000	PASS	0.0197	false_positive
16	84502596	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
16	87017311	T	G	failNA12878skellam	0.0009	Not Assigned	Not Assigned	false_positive
17	486647	C	T	failNA12878	0.0362	Not Assigned	Not Assigned	false_positive
17	486651	G	A	failNA12878skellam	0.0009	Not Assigned	Not Assigned	false_positive
17	14952264	A	T	failfreq	0.5055	Not Assigned	Not Assigned	false_positive
17	32760113	G	A	failNA12878skellam	0.0001	failNA12878skellam	0.0004	false_positive
17	50495972	C	A	failNA12878skellam	0.0004	Not Assigned	Not Assigned	false_positive
17	52466053	T	G	failNA12878skellam	0.0001	PASS	0.0064	false_positive
17	54006407	C	A	failseqerr	0.0004	Not Assigned	Not Assigned	false_positive
17	61091217	G	T	failNA12878skellam	0.0049	Not Assigned	Not Assigned	false_positive
18	5952670	G	A	failfreq	0.5072	failfreq	0.4905	false_positive
18	45925102	A	C	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
18	66870912	G	A	failseqerr	0.0005	Not Assigned	Not Assigned	false_positive
19	792183	T	G	failNA12878skellam	0.0004	Not Assigned	Not Assigned	false_positive
19	3345510	T	C	failseqerr	0.0001	Not Assigned	Not Assigned	false_positive
19	6591373	A	C	PASS	0.0027	Not Assigned	Not Assigned	false_positive
19	11505881	T	C	failNA12878	0.0349	Not Assigned	Not Assigned	false_positive
19	18415232	T	G	failNA12878	0.0130	Not Assigned	Not Assigned	false_positive
19	42528285	A	C	PASS	0.0157	Not Assigned	Not Assigned	false_positive
19	46284797	T	G	PASS	0.0000	PASS	0.0023	false_positive
19	50673439	A	C	readsnotenough	0.0000	failNA12878skellam	0.0004	false_positive
19	52338664	T	A	PASS	0.0017	PASS	0.0051	false_positive
19	54984467	T	G	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
20	18293788	G	T	failNA12878skellam	0.0006	Not Assigned	Not Assigned	false_positive
20	19699823	T	G	failseqerr	0.0000	failNA12878skellam	0.0003	false_positive
20	23318936	C	A	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
20	33791826	T	A	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
20	51021056	G	T	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
20	59241417	T	A	PASS	0.0000	Not Assigned	Not Assigned	false_positive
21	18187206	C	A	failNA12878skellam	0.0001	failseqerr	0.0005	false_positive
21	47089164	G	T	failNA12878skellam	0.0015	Not Assigned	Not Assigned	false_positive
22	34819068	A	T	failNA12878skellam	0.0003	Not Assigned	Not Assigned	false_positive
22	36682920	C	T	failNA12878skellam	0.0009	Not Assigned	Not Assigned	false_positive
22	49743408	T	G	failNA12878	0.0529	Not Assigned	Not Assigned	false_positive
X	12811469	C	T	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive

X	39971300	T	G	failNA12878	0.2565	PASS	0.0038	false_positive
X	57032203	G	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
X	68836605	C	G	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
X	86254255	C	A	failNA12878skellam	0.0040	Not Assigned	Not Assigned	false_positive
X	95400821	T	A	failNA12878skellam	0.0014	Not Assigned	Not Assigned	false_positive
X	108506261	G	T	failNA12878skellam	0.0009	Not Assigned	Not Assigned	false_positive
X	121831997	G	T	failseqerr	0.0010	Not Assigned	Not Assigned	false_positive
X	130905070	G	C	failseqerr	0.0000	failNA12878skellam	0	false_positive
X	134606031	C	A	failNA12878skellam	0.0002	Not Assigned	Not Assigned	false_positive
X	138338550	A	T	failNA12878skellam	0.0003	failNA12878skellam	0	false_positive
X	138628753	C	A	failNA12878skellam	0.0000	Not Assigned	Not Assigned	false_positive
X	150725216	A	T	failNA12878	0.2802	Not Assigned	Not Assigned	false_positive
X	154555643	C	A	failNA12878skellam	0.0001	Not Assigned	Not Assigned	false_positive
Y	14756883	T	A	failNA12878skellam	0.0010	Not Assigned	Not Assigned	false_positive
Y	16642956	A	T	failseqerr	0.0000	Not Assigned	Not Assigned	false_positive
Y	18833875	G	T	Not Sequenced	Not Sequenced	PASS	0.0038	false_positive

Reference

10X Genomics. LongRanger Version 2.2.2. Pleasanton, CA. 2018.

Azevedo, F., Carvalho, L., Grinberg, L., Farfel, J., Ferretti, R., Leite, R., Filho, W., Lent, R. and Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. The Journal of Comparative Neurology, 513(5), pp.532-541.

Badner, J. and Gershon, E. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. Molecular Psychiatry, 7(4), pp.405-411.

Broadinstitute.github.io. (2019). Picard Tools - By Broad Institute. [online] Available at: http://broadinstitute.github.io/picard/ [Accessed 28 Apr. 2019].

Bushman, D. and Chun, J. (2013). The genomically mosaic brain: Aneuploidy and more in neural diversity and disease. Seminars in Cell & Developmental Biology, 24(4), pp.357-369.

Cai, N., Bigdeli, T., Kretzschmar, W., Li, Y., Liang, J., Song, L., Hu, J., Li, Q., Jin, W., Hu, Z., Wang, G., Wang, L., Qian, P., Liu, Y., Jiang, T., Lu, Y., Zhang, X., Yin, Y., Li, Y., Xu, X., Gao, J., Reimers, M., Webb, T., Riley, B., Bacanu, S., Peterson, R., Chen, Y., Zhong, H., Liu, Z., Wang, G., Sun, J., Sang, H., Jiang, G., Zhou, X., Li, Y., Li, Y., Zhang, W., Wang, X., Fang, X., Pan, R., Miao, G., Zhang, Q., Hu, J., Yu, F., Du, B., Sang, W., Li, K., Chen, G., Cai, M., Yang, L., Yang, D., Ha, B., Hong, X., Deng, H., Li, G., Li, K., Song, Y., Gao, S., Zhang, J., Gan, Z., Meng, H., Pan, J., Gao, C., Zhang, K., Sun, N., Li, Y., Niu, Q., Zhang, Y., Liu, T., Hu, C., Zhang, Z., Lv, L., Dong, J., Wang, X., Tao, M., Wang, X., Xia, J., Rong, H., He, Q., Liu, T., Huang, G., Mei, Q., Shen, Z., Liu, Y., Shen, J., Tian, T., Liu, X., Wu, W., Gu, D., Fu, G., Shi, J., Chen, Y., Gan, X., Liu, L., Wang, L., Yang, F., Cong, E., Marchini, J., Yang, H., Wang, J., Shi, S., Mott, R., Xu, Q., Wang, J., Kendler, K. and Flint, J. (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature, 523(7562), pp.588-591.

Cibulskis, K., Lawrence, M., Carter, S., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology, 31(3), pp.213-219.

Denver, D., Feinberg, S., Estes, S., Thomas, W. and Lynch, M. (2005). Mutation Rates, Spectra and Hotspots in Mismatch Repair-Deficient Caenorhabditis elegans. Genetics, 170(1), pp.107-113.

Freed, D. and Pevsner, J. (2016). The Contribution of Mosaic Variants to Autism Spectrum Disorder. PLOS Genetics, 12(9), p.e1006245.
Fromer, M., Pocklington, A., Kavanagh, D., Williams, H., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D., Carrera, N., Humphreys, I., Johnson, J., Roussos, P., Barker, D., Banks, E., Milanova, V., Grant, S., Hannon, E., Rose, S., Chambert, K., Mahajan, M., Scolnick, E., Moran, J., Kirov, G., Palotie, A., McCarroll, S., Holmans, P., Sklar, P., Owen, M., Purcell, S. and O'Donovan, M. (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature, 506(7487), pp.179-184.

Genome.gov. (2019). DNA Sequencing Costs: Data | NHGRI. [online] Available at: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data [Accessed 27 Apr. 2019].

Genome in a bottle—a human DNA standard. (2015). Nature Biotechnology, 33(7), pp.675-675.

Genome-wide association study identifies five new schizophrenia loci. (2011). Nature Genetics, 43(10), pp.969-976.

Gleeson, J. (2000). Classical lissencephaly and double cortex (subcortical band heterotopia): LIS1 and doublecortin. Current Opinion in Neurology, 13(2), pp.121-125.

Hindson, B., Ness, K., Masquelier, D., Belgrader, P., Heredia, N., Makarewicz, A., Bright, I., Lucero, M., Hiddessen, A., Legler, T., Kitano, T., Hodel, M., Petersen, J., Wyatt, P., Steenblock, E., Shah, P., Bousse, L., Troup, C., Mellen, J., Wittmann, D., Erndt, N., Cauley, T., Koehler, R., So, A., Dube, S., Rose, K., Montesclaros, L., Wang, S., Stumbo, D., Hodges, S., Romine, S., Milanovich, F., White, H., Regan, J., Karlin-Neumann, G., Hindson, C., Saxonov, S. and Colston, B. (2011). High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Analytical Chemistry, 83(22), pp.8604-8610.

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. Frontiers in Human Neuroscience, 3.

Hyman, E. (1988). A new method of sequencing DNA. Analytical Biochemistry, 174(2), pp.423-436.

Insel, T. (2013). Brain somatic mutations: the dark matter of psychiatric genetics?. Molecular Psychiatry, 19(2), pp.156-158.

Ivády, G., Madar, L., Dzsudzsák, E., Koczok, K., Kappelmayer, J., Krulisova, V., Macek, M., Horváth, A. and Balogh, I. (2018). Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. BMC Genomics, 19(1).

Jamuar, S., Lam, A., Kircher, M., D'Gama, A., Wang, J., Barry, B., Zhang, X., Hill, R., Partlow, J., Rozzo, A., Servattalab, S., Mehta, B., Topcu, M., Amrom, D., Andermann, E., Dan, B., Parrini, E., Guerrini, R., Scheffer, I., Berkovic, S., Leventer, R., Shen, Y., Wu, B., Barkovich, A., Sahin, M., Chang, B., Bamshad, M., Nickerson, D., Shendure, J., Poduri, A., Yu, T. and Walsh, C. (2014). Somatic Mutations in Cerebral Cortical Malformations. New England Journal of Medicine, 371(8), pp.733-743.

Kanaar, R., Wyman, C. and Rothstein, R. (2008). Quality control of DNA break metabolism: in the 'end', it's a good thing. The EMBO Journal, 27(4), pp.581-588.

Kendler, K. and Zerbin-Rüdin, E. (1996). Abstract and review of "studien über vererbung und entstehung geistiger störungen. I. Zur vererbung und neuentstehung der dementia praecox." (Studies on the inheritance and origin of mental illness: I. To the problem of the inheritance and primary origin of dementia praecox.). American Journal of Medical Genetics, 67(4), pp.338-342.

Kent WJ. http://genome.ucsc.edu/cgi-bin/hgPcr?hgsid=720560077_NwAor74VxpSPpJ6CgWxT4N7wUpbQ

Koboldt, D., Chen, K., Wylie, T., Larson, D., McLellan, M., Mardis, E., Weinstock, G., Wilson, R. and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics, 25(17), pp.2283-2285.

Lee, J. (2016). Somatic mutations in disorders with disrupted brain connectivity. Experimental & Molecular Medicine, 48(6), pp.e239-e239.

Leibeling, D., Laspe, P. and Emmert, S. (2006). Nucleotide excision repair and cancer. Journal of Molecular Histology, 37(5-7), pp.225-238.

Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., Cummings, B., Tukiainen, T., Birnbaum, D., Kosmicki, J., Duncan, L., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M., Moonshine, A., Natarajan, P., Orozco, L., Peloso, G., Poplin, R., Rivas, M., Ruano-Rubio, V., Rose, S., Ruderfer, D., Shakir, K., Stenson, P., Stevens, C., Thomas, B., Tiao, G., Tusie-Luna, M., Weisburd, B., Won, H., Yu, D., Altshuler, D., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J., Gabriel, S., Getz, G., Glatt, S., Hultman, C., Kathiresan, S., Laakso, M.,

McCarroll, S., McCarthy, M., McGovern, D., McPherson, R., Neale, B., Palotie, A., Purcell, S., Saleheen, D., Scharf, J., Sklar, P., Sullivan, P., Tuomilehto, J., Tsuang, M., Watkins, H., Wilson, J., Daly, M. and MacArthur, D. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616), pp.285-291.

Lewis, C., Levinson, D., Wise, L., DeLisi, L., Straub, R., Hovatta, I., Williams, N., Schwab, S., Pulver, A., Faraone, S., Brzustowicz, L., Kaufmann, C., Garver, D., Gurling, H., Lindholm, E., Coon, H., Moises, H., Byerley, W., Shaw, S., Mesen, A., Sherrington, R., O'Neill, F., Walsh, D., Kendler, K., Ekelund, J., Paunio, T., Lönnqvist, J., Peltonen, L., O'Donovan, M., Owen, M., Wildenauer, D., Maier, W., Nestadt, G., Blouin, J., Antonarakis, S., Mowry, B., Silverman, J., Crowe, R., Cloninger, C., Tsuang, M., Malaspina, D., Harkavy-Friedman, J., Svrakic, D., Bassett, A., Holcomb, J., Kalsi, G., McQuillin, A., Brynjolfson, J., Sigmundsson, T., Petursson, H., Jazin, E., Zoëga, T. and Helgason, T. (2003). Genome Scan Meta-Analysis of Schizophrenia and Bipolar Disorder, Part II: Schizophrenia. The American Journal of Human Genetics, 73(1), pp.34-48.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), pp.1754-1760.

Lim, J., Kim, W., Kang, H., Kim, S., Park, A., Park, E., Cho, Y., Kim, S., Kim, H., Kim, J., Kim, J., Rhee, H., Kang, S., Kim, H., Kim, D., Kim, D. and Lee, J. (2015). Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. Nature Medicine, 21(4), pp.395-400.

McConnell, M., Moran, J., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J., Fasching, L., Flasch, D., Freed, D., Ganz, J., Jaffe, A., Kwan, K., Kwon, M., Lodato, M., Mills, R., Paquola, A., Rodin, R., Rosenbluh, C., Sestan, N., Sherman, M., Shin, J., Song, S., Straub, R., Thorpe, J., Weinberger, D., Urban, A., Zhou, B., Gage, F., Lehner, T., Senthil, G., Walsh, C., Chess, A., Courchesne, E., Gleeson, J., Kidd, J., Park, P., Pevsner, J. and Vaccarino, F. (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science, 356(6336), p.eaal1641.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), pp.1297-1303.

Meehl, P. (1962). Schizotaxia, schizotypy, schizophrenia. American Psychologist, 17(12), pp.827-838.

Metzker, M. (2009). Sequencing technologies — the next generation. Nature Reviews Genetics, 11(1), pp.31-46.

Mirzaa, G., Conti, V., Timms, A., Smyser, C., Ahmed, S., Carter, M., Barnett, S., Hufnagel, R., Goldstein, A., Narumi-Kishimoto, Y., Olds, C., Collins, S., Johnston, K., Deleuze, J., Nitschké, P., Friend, K., Harris, C., Goetsch, A., Martin, B., Boyle, E., Parrini, E., Mei, D., Tattini, L., Slavotinek, A., Blair, E., Barnett, C., Shendure, J., Chelly, J., Dobyns, W. and Guerrini, R. (2015). Characterisation of mutations of the phosphoinositide-3-kinase regulatory subunit, PIK3R2, in perisylvian polymicrogyria: a next-generation sequencing study. The Lancet Neurology, 14(12), pp.1182-1195.

Mirzaa, G., Campbell, C., Solovieff, N., Goold, C., Jansen, L., Menon, S., Timms, A., Conti, V., Biag, J., Olds, C., Boyle, E., Collins, S., Ishak, G., Poliachik, S., Girisha, K., Yeung, K., Chung, B., Rahikkala, E., Gunter, S., McDaniel, S., Macmurdo, C., Bernstein, J., Martin, B., Leary, R., Mahan, S., Liu, S., Weaver, M., Dorschner, M., Jhangiani, S., Muzny, D., Boerwinkle, E., Gibbs, R., Lupski, J., Shendure, J., Saneto, R., Novotny, E., Wilson, C., Sellers, W., Morrissey, M., Hevner, R., Ojemann, J., Guerrini, R., Murphy, L., Winckler, W. and Dobyns, W. (2016). Association ofMTORMutations With Developmental Brain Disorders, Including Megalencephaly, Focal Cortical Dysplasia, and Pigmentary Mosaicism. JAMA Neurology, 73(7), p.836.

Muotri, A. and Gage, F. (2006). Generation of neuronal variability and complexity. Nature, 441(7097), pp.1087-1093.

Nakashima, M., Saitsu, H., Takei, N., Tohyama, J., Kato, M., Kitaura, H., Shiina, M., Shirozu, H., Masuda, H., Watanabe, K., Ohba, C., Tsurusaki, Y., Miyake, N., Zheng, Y., Sato, T., Takebayashi, H., Ogata, K., Kameyama, S., Kakita, A. and Matsumoto, N. (2015). Somatic Mutations in theMTORgene cause focal cortical dysplasia type IIb. Annals of Neurology, 78(3), pp.375-386.

Ng, M., Levinson, D., Faraone, S., Suarez, B., DeLisi, L., Arinami, T., Riley, B., Paunio, T., Pulver, A., Irmansyah, Holmans, P., Escamilla, M., Wildenauer, D., Williams, N., Laurent, C., Mowry, B., Brzustowicz, L., Maziade, M., Sklar, P., Garver, D., Abecasis, G., Lerer, B., Fallin, M., Gurling, H., Gejman, P., Lindholm, E., Moises, H., Byerley, W., Wijsman, E., Forabosco, P., Tsuang, M., Hwu, H., Okazaki, Y., Kendler, K., Wormley, B., Fanous, A., Walsh, D., O'Neill, F., Peltonen, L., Nestadt, G., Lasseter, V., Liang, K., Papadimitriou, G., Dikeos, D., Schwab, S., Owen, M., O'Donovan, M., Norton, N., Hare, E., Raventos, H., Nicolini, H., Albus, M., Maier, W., Nimgaonkar, V., Terenius, L., Mallet, J., Jay, M., Godard, S., Nertney, D., Alexander, M., Crowe, R., Silverman, J., Bassett, A., Roy, M., Mérette, C., Pato, C., Pato, M., Roos, J., Kohn, Y., Amann-Zalcenstein, D., Kalsi, G., McQuillin, A., Curtis, D., Brynjolfson, J., Sigmundsson, T., Petursson, H., Sanders, A., Duan, J., Jazin, E., Myles-Worsley, M., Karayiorgou, M. and Lewis, C. (2008). Meta-analysis of 32 genome-wide linkage studies of schizophrenia. Molecular Psychiatry, 14(8), pp.774-785.

Purcell, S., Wray, N., Stone, J., Visscher, P., O'Donovan, M., Sullivan, P., Sklar, P., Purcell (Leader), S., Stone, J., Sullivan, P., Ruderfer, D., McQuillin, A., Morris, D., O'Dushlaine, C., Corvin, A., Holmans, P., O'Donovan, M., Sklar, P., Wray, N., Macgregor, S., Sklar, P., Sullivan, P., O'Donovan, M., Visscher, P., Gurling, H., Blackwood, D., Corvin, A., Craddock, N., Gill, M., Hultman, C., Kirov, G., Lichtenstein, P., McQuillin, A., Muir, W., O'Donovan, M., Owen, M., Pato, C., Purcell, S., Scolnick, E., St Clair, D., Stone, J., Sullivan, P., Sklar (Leader), P., O'Donovan, M., Kirov, G., Craddock, N., Holmans, P., Williams, N., Georgieva, L., Nikolov, I., Norton, N., Williams, H., Toncheva, D., Milanova, V., Owen, M., Hultman, C., Lichtenstein, P., Thelander, E., Sullivan, P., Morris, D., O'Dushlaine, C., Kenny, E., Quinn, E., Gill, M., Corvin, A., McQuillin, A., Choudhury, K., Datta, S., Pimm, J., Thirumalai, S., Puri, V., Krasucki, R., Lawrence, J., Quested, D., Bass, N., Gurling, H., Crombie, C., Fraser, G., Leh Kuan, S., Walker, N., St Clair, D., Blackwood, D., Muir, W., McGhee, K., Pickard, B., Malloy, P., Maclean, A., Van Beck, M., Wray, N., Macgregor, S., Visscher, P., Pato, M., Medeiros, H., Middleton, F., Carvalho, C., Morley, C., Fanous, A., Conti, D., Knowles, J., Paz Ferreira, C., Macedo, A., Helena Azevedo, M., Pato, C., Stone, J., Ruderfer, D., Kirby, A., Ferreira, M., Daly, M., Purcell, S., Sklar, P., Purcell, S., Stone, J., Chambert, K., Ruderfer, D., Kuruvilla, F., Gabriel, S., Ardlie, K., Moran, J., Daly, M., Scolnick, E. and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature.

Rivière, J., Mirzaa, G., O'Roak, B., Beddaoui, M., Alcantara, D., Conway, R., St-Onge, J., Schwartzentruber, J., Gripp, K., Nikkel, S., Worthylake, T., Sullivan, C., Ward, T., Butler, H., Kramer, N., Albrecht, B., Armour, C., Armstrong, L., Caluseriu, O., Cytrynbaum, C., Drolet, B., Innes, A., Lauzon, J., Lin, A., Mancini, G., Meschino, W., Reggin, J., Saggar, A., Lerman-Sagie, T., Uyanik, G., Weksberg, R., Zirn, B., Beaulieu, C., Majewski, J., Bulman, D., O'Driscoll, M., Shendure, J., Graham, J., Boycott, K. and Dobyns, W. (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nature Genetics, 44(8), pp.934-940.

Ronaghi, M. (1998). DNA SEQUENCING: A Sequencing Method Based on Real-Time Pyrophosphate. Science, 281(5375), pp.363-365.

Saunders, C., Wong, W., Swamy, S., Becq, J., Murray, L. and Cheetham, R. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics, 28(14), pp.1811-1817.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), pp.421-427.

Schulz, B. (1933). Zur Erbpathologie der Schizophrenie. Zeitschrift für die gesamte Neurologie und Psychiatrie, 143(1), pp.175-293.

Shirley, M., Tang, H., Gallione, C., Baugher, J., Frelin, L., Cohen, B., North, P., Marchuk, D., Comi, A. and Pevsner, J. (2013). Sturge–Weber Syndrome and Port-Wine Stains Caused by Somatic Mutation in GNAQ. New England Journal of Medicine, 368(21), pp.1971-1979.

Sicca, F., Kelemen, A., Genton, P., Das, S., Mei, D., Moro, F., Dobyns, W. and Guerrini, R. (2003). Mosaic mutations of the LIS1 gene cause subcortical band heterotopia. Neurology, 61(8), pp.1042-1046.

Tarasov, A., Vilella, A., Cuppen, E., Nijman, I. and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics, 31(12), pp.2032-2034.

Van Raamsdonk, C., Bezrookove, V., Green, G., Bauer, J., Gaugler, L., O'Brien, J., Simpson, E., Barsh, G. and Bastian, B. (2008). Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. Nature, 457(7229), pp.599-602.

WILSONIII, D. and BOHR, V. (2007). The mechanics of base excision repair, and its relationship to aging and disease. DNA Repair, 6(4), pp.544-559.

Zhang, J., Kobert, K., Flouri, T. and Stamatakis, A. (2013). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, 30(5), pp.614-620.

Chapter 3 Identification of RNA Level Single Molecule U6 Fusion Events

This chapter presents the computational contributions to the submitted manuscript entitled: “RNA ligation precedes the retrotransposition of U6/LINE-1 chimeric RNA” with Dr. John B. Moldovan, Dr. Stewart Shuman, Dr. Ryan E. Mills and Dr. John V. Moran. Dr. Stewart Shuman generously provided us the RtcB enzyme for biochemical experiments. A portion of the results section and the methods section from the original manuscript are presented here with minor modifications. Dr. John B. Moldovan performed all library preparation and biochemical experiments *in vivo and in vitro* and led the project with Dr. John V. Moran. I performed all the computational analysis of the RNA sequencing data as well as the junction search in HeLa genome and 1000 Genome 22 deep whole genome sequencing data, which I describe in detail in this chapter.

Introduction

Long INterspersed Element -1 (LINE-1, L1) and U6 snRNA in human genome

Long INterspersed Element-1 (LINE-1 or L1) and L1-derived sequences account for ~17% of human genomic DNA (Lander et al. 2001). L1 is the only known human autonomous non-Long Terminal Repeat (non-LTR) retrotransposon. L1

‘jumps’ in the genome through a ‘copy and paste’ mechanism termed retrotransposition. Most of the L1-derived sequences in the genome cannot move to new genomic locations because of 5’-trunction, insertions and/or deletions (indels) or single nucleotide variations in the sequences (Grimaldi et al, 1983; Scott et al. 1987; Lander et al., 2001). Among all L1-derived sequences in human genome, there are approximately 80-100 retrotransposition-competent L1’s present in human genome (Sassaman et al. 1997; Brouha et al. 2003; Beck et al. 2010). Details of L1 retrotransposition and structure are included in chapter 2.

U6 snRNA is a key component of spliceosome. The spliceosome, an intricate machine responsible for RNA splicing, is composed of five ribonucleoprotein (RNP) subunits (U1, U2, U4, U5, U6 and their associated proteins), along with a host of associated protein co-factors (Jurica et al. 2003; Wahl et al. 2009; Will et al. 2010; Matera et al. 2014). Multiple evidences show that U6 snRNA plays a role in the catalytic center of the spliceosome (Didychuk et al. 2018). Details of U6 structure and function are included in chapter 2.

There are over 900 copies of U6’s (Doucet et al. 2015) and thousands of L1’s composing 17% (Lander et al. 2001) of total sequence in human genome. As discussed in the last session, these repetitive sequences are excluded from the analysis for chimeric RNA discovery in all existing methods. However, 35% of full-length U6 sequences are in chimeric sequences with L1’s in human genome (Buzdin et al. 2003). The mechanism of formation for this chimeric pseudogene

remains unclear. To fill this gap of analysis of fusion events between repetitive RNAs, we developed and applied a computational approach to identify U6/L1 chimeric sequences in cells. In contrast to the previous hypothesis of the pseudogenization occurring via template switching during reverse transcription of L1 (Figure 3.1), we sought to explore whether the RNA ligase enzyme, RtcB (Englert et al. 2011; Tanaka et al. 2011; Popow et al. 2011) could join U6 RNAs ending in a 2',3'-cyclic phosphate to L1 or other mRNAs containing a 5'-OH group in HeLa cells. We identified U6/L1 RNA fusions in multiple cell lines, which could then be retrotransposed into genome by L1 machinery (Figure 3.2). We further investigated the overall fusion of U6 with other RNAs in different cell lines and characterized the genomic positions where the U6 sequence fused.

Method and Materials

RNA sequencing library preparation and sequencing

Adapted from Dr. John B. Moldovan's library preparation description

All cDNA library preparation and sequencing were conducted at the University of Michigan sequencing core facility (Ann Arbor, MI). Briefly, total RNA was collected from HeLa-JVM, HeLa-HA, and PA-1 cells using a RNeasy mini kit (Qiagen). Total RNA from hESC (Garcia-Perez et al. 2007; Macia et al. 2011), and hESC derived NPCs (Coufal et al. 2009) was a generous gift of Dr. Jose Garcia-Perez. To generate cDNA libraries, total RNA from each cell line was first depleted of ribosomal RNA using a Ribo-Zero rRNA removal kit (Illumina, San Diego, CA), and then cDNA libraries were generated from the rRNA-depleted

RNA using the TruSeq Stranded mRNA Library Prep Kit (Illumina) with random hexamers. Paired-end sequencing (100 bp reads) was performed on the Illumina HiSeq 2500. Sequencing data for PA-1, H9, and NPCs have been uploaded to the Sequence Read Archive (submission number: 3608651). HeLa sequencing data will be deposited to dbGaP.

RNA sequencing analysis pipeline

Trimmomatic (Bolger et al. 2014) was used to trim the sequencing adaptors from a total of ~1.1 X 109 RNA sequencing reads. We assessed the quality of our data using FastQC (Andrews S. et al. 2010). Samtools rmdup (Li et al. 2009) and Picard MarkDuplicates (http://broadinstitute.github.io/picard) were used to remove PCR duplicate reads. We aligned all reads that passed the quality check with BWA-MEM with default parameters (Li et al. 2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM to a custom built human reference genome from hg38 with all repeats masked using RepeatMasker and Repbase (Jurka et al. 2005), but including a single representative copy of a human specific L1 (L1.3; accession no. L19088) (Sassaman et al. 1997) and human U6 snRNA (accession no. X59362). FLASH (Magoc et al. 2011) then was used to reconstruct overlapping read pairs that aligned at one end to the 3' portion of U6 snRNA and the other end to L1. Merged U6/L1 sequences that contained U6 snRNA sequence at the 5' end conjoined to L1 sequence at the 3' end were then mapped back to the non-masked HGR (HGR/build GRCh38) using BWA-MEM in order to differentiate

events aligned to the genome from those which did not exhibit a clear mapping (Figure 3.3 a, b). We then applied blastn with the last 25 base pair of U6 snRNA sequence to ensure the existence of full length U6 in the reads. Our software for extracting these fusion reads from RNA-seq data can be found at https://github.com/mills-lab/U6L1. After alignments, non-aligned merged U6/L1 reads were hand curated to manually identify PCR duplicate reads. Merged U6/L1 reads that were either identical and/or differed by only a single nucleotide were marked as duplicates. All U6/L1 reads were manually aligned to the HGR using BLAT (Kent et al. 2002) to verify BWA-MEM alignments. The L1 portion of each U6/L1 read was manually aligned to the L1.3 sequence and consensus sequences from L1 subfamilies (L1PA1-L1PA13) (Khan at al. 2006) to determine the L1 subfamily and to derive L1 sequences for L1 junction analyses (Table 3.5).

In order to extend the capabilities of this process, we modified the above U6/L1 junction identification pipeline by including an option to consider all repetitive sequences in Repbase in addition to U6 and L1 sequences to identify putative RNA level U6 fusion events with other genes/repetitive elements in transcriptome. (Figure 3.6)

U6/L1 Junction Motif Search of HeLa cells and 1000 Genomes Project High Coverage Samples

Junction motifs across putative U6/L1 junctions were extracted from all merged reads as described above, including aligned, non-aligned and artifact junction sequences. Each 25 base pair junction motif contains U6 snRNA nucleotides 94-102 followed by 5-8 thymidines and ~8-11 nucleotides of L1 sequence (Table 3.6). All motifs and their reverse complements were used to interrogate HeLa cell genomic data from dbGaP (dbGaP accession number phs000640.v1.p1) (Mailman MD et al. 2007; Landry et al. 2013; Adey et al. 2013) and 23 high coverage PCR-free DNA sequencing samples from the 1000 Genomes Project (Table 3.7) (Genomes Project C, et al. 2015) to look for genomic evidence of each U6/L1 junction sequence. The script for 25 base pair motif searching is available at: https://github.com/mills-lab/U6L1. An exact match was required for labeling the existence of the junction from the HeLa genomic and 1000 Genomes DNA sequencing data. Two exceptions were noted in the 1000 genomes data, in which two genomes (NA20845 and HG03742) contained the same SNP within the U6 sequence for the U6/L1 chimera sequence with L1.3 junction 2052 and therefore did not initially exhibit an exact match to the genomic sequences of these samples (see Results and Table 3.6).

Random permutation of gene enrichment for U6 fusion events

In order to identify the RNAs that U6 fused with generally in the cell lines, we cross-referenced the fusion point identified from the pipeline of full Repbase sequences with GENCODE v24 annotation. We randomly selected genes from the annotation files for a background permutation test. For each round of

permutation test, we randomly selected 6321 genes (total number of fusion sites identified from five cell lines) from the annotation file. We repeated this process for 10,000 times. The distribution of genes selected is then served as the background model to be compared with RNAs fused with U6 in the five different cell lines.

Random permutation of motif analysis for U6 fusion junctions

For each U6 fusion point discovered from the five RNA libraries, we took the upstream and downstream 10 base pairs of the fusion points. After collecting all 20 base pair regions around the fusion points, we applied MEME Suite (Bailey et al. 2009) to discover the enriched motifs at the U6 fusion points with other RNAs.

To simulate the genomic context of motif at random positions in transcriptome, we randomly picked 20 base pair genomic fragments containing enriched 'AUA' motif for 6321 times each round of permutation. We performed the permutation for 10,000 times to assess the position of 'AUA' motif in random 20 base pair DNA fragments in transcriptome (Figure 3.8 b). We then compared the relative position of 'ATA' motif in random DNA fragments with the relative position of 'ATA' in U6 fusion points using a Kolmogorov-Smirnov test.

Random permutation of distance to loop in secondary structure for U6 fusion sites

We further investigated if any secondary structure feature could affect the fusion with U6 of certain RNAs. We utilized Vienna (Hofacker, 2003; Kerpedjiev et al. 2015) to predict the secondary structure of all snRNAs and snoRNAs. We only applied the prediction method to short RNAs since it is difficult to predict the secondary structure of longer RNAs. We started with calculating the relative distance of fusion sites to the closest loop in the secondary structure. The positions in loops are counted as 0's, and the positions right at the edge of the loops are counted as 1's (Figure 3.9 a).

Secondary structure background is then simulated for 6321 times by randomly selecting small RNAs as well as the position in the RNAs that were selected. We simulated for 10,000 rounds. The fusion sites distance to loops is then compared with the random permutation results (Figure 3.9 b).

Results

Endogenous U6/L1 RNA is part of the transcriptome in human cells (adapted from: RNA ligation precedes the retrotransposition of U6/LINE-1 chimeric RNA, Moldovan et at., 2019, PNAS, under review)

Data from our transfection experiments and *in vitro* ligation experiments performed by Dr. John B. Moldovan suggested that U6 snRNA could be ligated to L1 RNA *in vivo*, thus we sought to determine whether U6/L1 chimeric RNA could be part of the normal transcriptome in human cells. To accomplish this, we searched for U6/L1 junction reads in 100 base pair paired-end RNA sequencing

(RNA-seq) data generated from two independent HeLa cell lines (HeLa-JVM and HeLa-HA), a human embryonic carcinoma cell line (PA-1), a human embryonic stem cell line (H9-hESCs), and H9-derived neural progenitor cells (NPCs) (see Methods). Each of these cell lines can accommodate the retrotransposition of engineered human L1s *in vitro* (Coufal et al. 2009; Moran et al. 1996; Garcia-Perez et al. 2010; Garcia-Perez et al. 2007; Macia et al. 2017). We identified 398 U6/L1 chimeric RNA read-pairs out of ~1.1 X 10^9 RNA sequencing reads across the five cell lines. After removing PCR duplicate reads, we then merged overlapping reads to identify 64 intact U6/L1 junction sequences.

Both alignment back to non-masked genome and hand-annotation of the 64 U6/L1 junctions revealed that 53 (~83%) U6/L1 chimeric junction sequences consisted of the 3' end of U6 snRNA cDNA ending in ~4-8 thymidine nucleotides conjoined to a variably 5'-truncated L1 sequence (Figure 3.4; Table 3.4). Notably, 4 out of these 53 U6/L1 chimeras consisted of U6 ending in 5-7 thymidines conjoined to an L1 sequence in the antisense orientation, suggesting that U6 can become conjoined to both sense and anti-sense L1 RNAs. As above, there was not a specific sequence within L1 that appeared to facilitate U6/L1 chimera formation (Figure 3.4; Table 3.4). The remaining 11 out of 64 (~17%) U6/L1 sequences contained a 3'-truncated U6 snRNA conjoined to a 5'-truncated L1 and were excluded from further analysis, as they were structurally similar to template switching artifacts generated during cDNA synthesis described above

(Houseley et al. 2010) (Table 3.4). Thus, 53 bona fide unique U6/L1 chimeras were subjected to further analysis.

Most U6/L1 chimeric RNA sequences do not align to the genome (adapted from: RNA ligation precedes the retrotransposition of U6/LINE-1 chimeric RNA, Moldovan et at., 2019, PNAS, under review)

The low proportion of U6/L1 chimeras in our dataset suggested that U6/L1 chimeric RNA might represent a rare and unique subset of the total RNA in human cells. To determine whether the RNA-seq U6/L1 chimeras were derived from the transcription of an existing genomic U6/L1 or represented unique chimeric RNAs, the 53 unique U6/L1 sequences were used as probes in BLAT searches of the HGR (Figure 3.3; see Methods). Sixteen out of fifty-three (~30%) U6/L1 junctions were present in the HGR, suggesting that they could have resulted from the transcription of extant U6/L1 pseudogene insertion. Seven out of the sixteen putative transcribed U6/L1 chimeric RNAs were detected in multiple cell lines (Figure 3.4; Table 3.4) and seven were supported by multiple reads from the same cell line (Figure 3.4; Table 3.4). The 16 genomic U6/L1 chimeric pseudogenes that served as putative transcription templates that gave rise to chimeric U6/L1 RNAs exhibited L1 retrotransposition insertion structural hallmarks (Table 3.5). They consisted of a full-length U6 snRNA sequence ending in 5 to 7 thymidine nucleotides conjoined to a variably 5'-truncated L1, were flanked by 6-19 bp target site duplications, and inserted into a L1 EN consensus cleavage sequence. By comparison, 37 out of 53 (~70%) U6/L1

junction sequences did not align to the HGR and were unique to a single cell line (Figure 3.4). Thirty-one out of thirty-seven junctions were supported by a single merged read pair, 5/37 were supported by two merged reads that may represent PCR duplicates, and one junction was supported by three merged reads (Figure 3.4; see Methods).

Human-specific L1 insertions can be polymorphic with respect to presence/absence in the human population (Beck et al. 2011); thus, it is conceivable that some of the cell lines used to generate RNA-seq data could contain a genomic U6/L1 chimeric pseudogene that is absent from the HGR. To test this possibility, we used the 53 U6/L1 junctions as probes to query HeLa genome sequencing data available in the database of Genotypes and Phenotypes (dbGaP accession number phs000640.v1. p1) (Mailman MD et al. 2007; Landry et al. 2013; Adey et al. 2013). Controls revealed that the 16 U6/L1 junction sequences that aligned to the HGR were also present in the HeLa genome data (Figure 3.4, Table 3.4; see Methods). By comparison, the 37 non-aligned U6/L1 junction sequences were absent from HeLa genome data.

To further validate the uniqueness of the 37 U6/L1 junction sequences, we aligned the 53 U6/L1 junction sequences to 23 high-coverage individual genomes representing 23 distinct human geographic populations from the 1000 Genomes Project dataset (Figure 3.5; see Methods) (1000 Genomes Project et al. 2015). The 16 U6/L1 junction sequences that were present in the HGR and

HeLa cell genomic datasets also were present in each of the 23 of high coverage 1000 Genomes Project individual genomes; two genomes (NA20845 and HG03742) contained a SNP in the U6 portion of the junction sequences (Figure 3.5). In contrast, the 37 non-aligned U6/L1 junction sequences were absent from the high coverage 1000 Genomes Project individual genomes. Thus, the data suggest that 37 U6/L1 junctions detected in RNA-seq experiments do not to correspond to an existing genomic sequence in the HGR and that different cell types may contain a unique cohort of chimeric RNAs that are generated by post-transcriptional RNA ligation events.

Chimeric U6/L1 RNAs are present in human cells (adapted from: RNA ligation precedes the retrotransposition of U6/LINE-1 chimeric RNA, Moldovan et at., 2019, PNAS, under review)

RNA-seq experiments demonstrated that U6/L1 chimeric RNAs are a normal component of the transcriptome in human cancer cell lines, hESCs, and human NPCs (Figure 3.4). Approximately, 30% of the U6/L1 chimeric RNAs aligned to the HGR, HeLa, and 23 high-coverage genomes in the 1000 Genomes Project dataset (Figure 3.5), indicating that they are generated from existing U6/L1 chimeric pseudogenes. Vertebrate U6 snRNA is transcribed by RNA polymerase III and relies on upstream promoter elements to drive its transcription (Didychuk et al. 2018; Kunkel et al. 1986). Thus, unless U6/L1 chimeric pseudogenes fortuitously inserted downstream of a promoter that could augment RNA polymerase III transcription, it remains unlikely that U6/L1 chimeric RNAs are

transcribed as discrete transcription units. In contrast, the majority (~70%) of U6/L1 chimeric RNA-seq reads failed to align to the HGR, HeLa, or 23 high-coverage genomes in the 1000 Genomes Project dataset (Table 3.7), suggesting that they were generated de novo, by a post-transcriptional mechanism that joins U6 snRNA to L1 RNA. Our transfected cell RT-PCR experiments and *in vitro* ligation data suggest that U6 RNA is ligated to a variably 5'-truncated L1 RNA, and that there is not a specific sequence in L1 that serves as a ligation substrate. These data imply that U6/L1 ligation results in the formation of a unique U6/L1 chimeric RNA molecule. Consistent with this conclusion, the majority of the 37 "non-aligned" junctions was supported by a single RNA-seq read and was unique to a single cell line.

Other cellular genes are fused with U6

In addition to U6/L1 chimeric pseudogenes, the human genome also contains the fragments of full length U6 co-localized with other truncated RNAs (Garcia-Perez et al. 2007; Buzdin et al. 2003). We thus broadened our hypothesis to investigate potential fusion events between U6 and any other possible RNA fragments in cells. We identified 2,314 genes fused with U6 snRNA in total (Table 3.8 & Table 3.9). For all genes fused with U6, we observed that small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) are significantly enriched (Figure 3.7 b; Table 3.8). We also discovered that compared to snRNAs and snoRNAs, protein-coding genes contain the most unique genes fused with U6 snRNAs, but for each transcript, only a few molecules were fused with U6 (Table

3.9). As expected, L1 was also among the most frequent transcripts fused with U6 snRNAs. From the permutation of random selection from all annotated genes, protein coding genes, snRNAs and snoRNAs are all enriched with fusion of U6 (Figure 3.7 c).

An "AUA" sequence motif is enriched near the fusion sites with U6

Next, we looked for into more details of the characteristics of the junction sequences where U6 fuses into other RNAs. From the U6/L1 analysis, we were not able to identify any significant motif that were enriched at the fusion sites of L1s with U6 using MEME Suite (Bailey et al. 2009) because of the limited number of single molecule U6/L1 fusion events discovered from the five different cell lines. Motif finding algorithms are not able to build statistical model to find significant motifs from less than 50 junction sequences (Liu, Ma and Wang, 2008). With the larger number of candidate fusion events with U6 identified with other RNAs, we were able to discover an enriched motif at the fusion sites in the transcriptome with U6. From the motif enrichment analysis, we identified 'AUA as the motif enriched in all junction sequences by using MEME Suite (See Methods). In specific types of RNAs that U6 fused with, we were able to identify more specific motifs at junctions (i.e. 'AAAAAUA' for junctions of protein coding genes fused with U6, 'CTAUA' for junctions of snoRNAs fused with U6) (Figure 3.8 a).

To further investigate the randomness of 'AUA' motif enriched at fusion site with U6, we performed a random permutation of the position of 'AUA' in any random 20 base pairs in transcriptome (Figure 3.8 b). The 'AUA' motif is equally distributed in the 20 base pair fragments from random transcriptomic sequences, while in the 20 base pair fragments around junctions fused with U6, 'AUA' motif is significantly enriched around positive 5 base pairs from the junction point. The result of Kolmogorov-Smirnov test comparing the position of 'AUA' motif in U6 fusion points and random fragments returns a 2.2e-16 p-value, representing a statistically significant difference between the two sets of fragments.

Loops in RNA secondary structure are enriched near U6 fusion junctions

Another potential important aspect for RNA level fusion events is related to secondary structure. We analyzed the distance of fusion sites to constructed secondary structure loops of RNAs using Vienna (see Methods) (Hofacker, 2003; Kerpedjiev et al. 2015). We hypothesized that loops are more accessible to excision and fusion enzymes. Compared to random positions selected from all secondary structures constructed, U6 fusion sites are more enriched in positions inside as well as within 1 or 2 bases from the loops (Figure 3.9c). This provides additional support to our hypothesis that positions in loops or nearby loops in secondary structure are more accessible to U6 RNA level fusions.

Discussion

We identified both single molecule RNA level fusion events and expressed existing genomic U6/L1 fragments from five RNA sequencing libraries of different cell lines by applying our customized pipeline (Figure 3.4). Except for potential PCR duplicates, all single molecule RNA level fusion events are unique in every single cell line, while the expressed genomic events are mostly detected in multiple cell lines with multiple supportive reads. From the result of searching for 25 base pair junction sequence in the 22 high coverage 1000 Genomes samples, we were able to successfully identify all aligned U6/L1 junction sequences (3 events with SNPs in one of the 22 samples after manual checking) in all 22 samples. In contrast, the non-aligned U6/L1 junction sequences were not detected in any of the 22 samples, suggesting that the non-aligned U6/L1 events are unique, RNA level fusion events. Possible template switching artifacts that could have resulted during library preparation were not detected in our searches (Figure 3.5 & Table 3.6).

The biochemical analysis from our collaborators Dr. John Moldovan and Dr. John Moran showed that the U6/L1 fusion RNAs could be generated independently of L1 retrotransposition. This indicates that the U6/L1 events that we identified from the 5 different cell lines may not be the results of template switching during L1 retrotransposition as previously suggested (Figure 3.1) (Garcia-Perez et al. 2007). They also performed the experiment demonstrating that purified RtcB could ligate U6 RNA to L1 RNA *in vitro* as well as the necessity of a 2',3'-cyclic phosphate on U6 and a 5'-OH on L1 in the fusion RNA formation process.

Furthermore, two HeLa cell lines with reduced RtcB expression utilizing CRISPR/Cas9 gene-editing method showed a depletion of U6/L1 fusion events for ~ 4-5 folds compared to the negative control with RtcB regularly expressed. In addition to all the biochemical experiment results, our discovery of multiple U6/L1 fusion single molecule RNAs suggests that the U6/L1 RNA fusions found in the 5 cell lines were generated by RtcB-mediated RNA ligation.

As a critical part of spliceosome, U6 snRNA is enriched in the nucleus (Didychuk, Butcher and Brow, 2018; Zhang et al., 2014). RtcB is present in both cytoplasm and nucleus (Lu, Liang and Wang, 2014; Kosmaczewski et al., 2015). The co-localization of U6 snRNA, transcribed L1 RNA as well as RtcB in the nucleus, together with previously known mechanism for L1 retrotransposition, we propose a new mechanistic hypothesis for U6/L1 chimeric gene formation (Figure 3.2) in addition to the previously suggested mechanism of template switching during L1 retrotransposition (Figure 3.1) (Garcia-Perez et al. 2007). An endonuclease could cleave L1 RNAs and generate a 5'-OH end, together with the U6 snRNA 2',3'-cyclic phosphate and the co-localization of RtcB enzyme, the fusion between U6 snRNA and truncated L1 RNA could occur in nucleus. The 5'-OH end of L1, in turn, would assist the reverse transcription of the U6/L1 chimeric RNA into genome through L1 retrotransposition in *cis* (Buzdin, 2003; Garcia-Perez et al., 2007).

Previous genome studies showed that the majority pseudogenes identified from human genome contain a full-length U6 with a polyT tract conjoined with a collection of various 5'-truncated L1 sequences (Buzdin et al., 2002; Garcia-Perez et al., 2007; Doucet et al., 2015). The mechanism that we proposed provides a plausible explanation for the generation of U6/L1 pseudogenes in genome with full length U6. Furthermore, our model could also explain the formation of U6atac /L1 pseudogene (Garcia-Perez et al., 2007; Doucet et al., 2015) formation since U6atac snRNA also contains a 2',3'-cyclic phosphate (Shchepachev et al., 2015).

We have generated evidence from both biochemical experiments and computational analysis for the new mechanism for the formation of U6/L1 chimeric pseudogenes in the human genome. Other than the template switching mechanism suggested in previous studies (Garcia-Perez et al. 2007), our study found extra evidence supporting that the U6/L1 chimeric RNAs were ligated at the RNA level under RtcB catalysis first. The U6/L1 chimeric RNA is then inserted back to the human genome with the L1 retrotransposition machinery (Figure 3.2).

With the in-depth characterization of U6/L1 RNA fusion events, we then investigated what other RNAs were fused to U6 snRNA in our 5 cell lines. Several studies have shown that U6 is fused to other mRNAs in the human

genome (Garcia-Perez et al. 2007; Buzdin et al. 2003). Our collaborators showed that U6 snRNA and GFP were ligated in HeLa cell nuclear extracts.

We also observed that U6 snRNA is ligated to protein coding RNAs, other snRNAs as well as snoRNAs and that the enrichment of these RNA sequences fused to U6 is correlated to their relative abundance in the human transcriptome (Figure 3.7). Although the function of U6 fusion with L1 or other RNAs is not clear, the discovery of the enriched types of RNAs could help with future studies. In particular, the enriched fusion of U6 snRNA with other snRNAs and snoRNAs, which are also key components of spliceosome, suggests that the fusion phenomena between U6 snRNA and other RNAs is possibly related with RNA splicing or RNA degradation.

With more junctions identified after we broaden the analysis to U6 and all other possible RNAs, we were able to perform the motif enrichment analysis on the junction sequences where U6 snRNA were fused in. We discovered that compared to the random simulated background motif, ‘AUA’ motif is significantly enriched 7-8 base pairs downstream of the U6 snRNA fusion point towards the 3’-end of the RNAs. Further experiments and analysis are still needed to investigate the mechanism of this ligation process as well as the role of ‘AUA’ motif in this process.

Single stranded RNAs fold to secondary structures, for example, stem-loop structure (Svoboda et al. 2006) and pseudo-knot structure (Staple et al. 2005), in cells after being transcribed. The secondary structure of an RNA molecule has a critical impact on the function and stability of the RNA (Staple et al. 2005; Svoboda et al. 2006). Thus, we analyzed the impact of secondary structure of the RNAs ligated with U6 snRNAs. Because of the limitation of computational methods predicting secondary structures of longer RNAs, this analysis was limited to the snRNAs and snoRNAs ligated with U6 snRNAs. From this analysis, we discovered that U6 snRNAs ligation points are enriched in loop structures of other snRNAs, and in 1-2 bases away from loop structures in snoRNAs. Considering that bases in loop structures are more accessible to other molecules and enzymes, it is reasonable that the ligation point of other snRNAs and snoRNAs are enriched nearby loop structure. However, more precise secondary structure analysis is needed to make further conclusions about the hot spot of U6 snRNA ligation point in the secondary structure of RNAs.

Our study still has some limitations and caveats. For the biochemical analyses, since RtcB is a key ligase catalyzing tRNA splicing (Tanaka et al. 2011), we were not able to knock out RtcB entirely to show the necessity of RtcB to the ligation between U6 snRNA and L1. With this limitation, we were only able to show decreased U6/L1 ligation efficiency resulted from lower RtcB protein expression and that U6/L1 ligation efficiency increased in the same cell line with the transfected RtcB cDNA.

We identified the U6/L1 fusion reads from the RNA-seq data of 5 cell lines. However, due to the limitation of sequencing technique, we were not able to efficiently discard the PCR duplicates from the sequencing library itself. We utilized further analyses including searching for junction motifs in 22 non-related 1000 Genome high coverage samples to address this issue (Figure 3.5). With the improvement of bar-coded sequencing methods, we should be able to identify the single molecule level events better in the future.

The use of number of gene counts as the pool for random simulation could be another criticism of our study. Since what we were trying to characterize here is mostly related to repetitive sequences (i.e. U6 snRNAs, L1s, other snRNAs, snoRNAs etc.), the expression level of these elements are relatively difficult to evaluate from short read sequencing data. This could be improved by applying an estimated expression level for each of the repetitive elements and utilize the expression level as the pool for random simulation as negative control for gene type enrichment analysis.

We were only able to perform the secondary structure analysis in shorted RNAs such as snRNAs and snoRNAs due to the inaccuracy of computational method (Fallmann et al. 2017) to predict secondary structure and the high cost of experimental secondary structure prediction (Westhof et al. 2015). Despite this shortcoming, our result presented a reasonable explanation on how U6 snRNAs

could fuse with other RNAs at more flexible and accessible regions of the RNA secondary structure. Further studies on the 'AUA' motif enrichment correlation with the secondary structure positions are necessary for further conclusion of U6 snRNA fusion characteristics.

The detailed mechanism and function of U6 snRNA ligation process still remains unknown after this study. As part of the splicing machinery in cells, U6 snRNA is critical for cell development and survival. Further analysis for the function of RtcB catalyzed U6 snRNA fusion events with other RNAs could reveal possible mechanisms for other critical processes in cells, for example, RNA molecule degradation. As a highly active component of cell life cycle, we still need tremendous more effort both experimentally and computationally to further understand the function of U6 snRNA in cells.

Conclusion and Future Remarks

In sum, we have provided evidence both experimentally and computationally for a new mechanism for the formation of U6/L1 chimeric pseudogenes in the human genome. While the mechanism and function of U6 snRNA ligation with L1's and other RNAs remains unclear, our study presents strong evidence for the existence of this ligation process in cells.

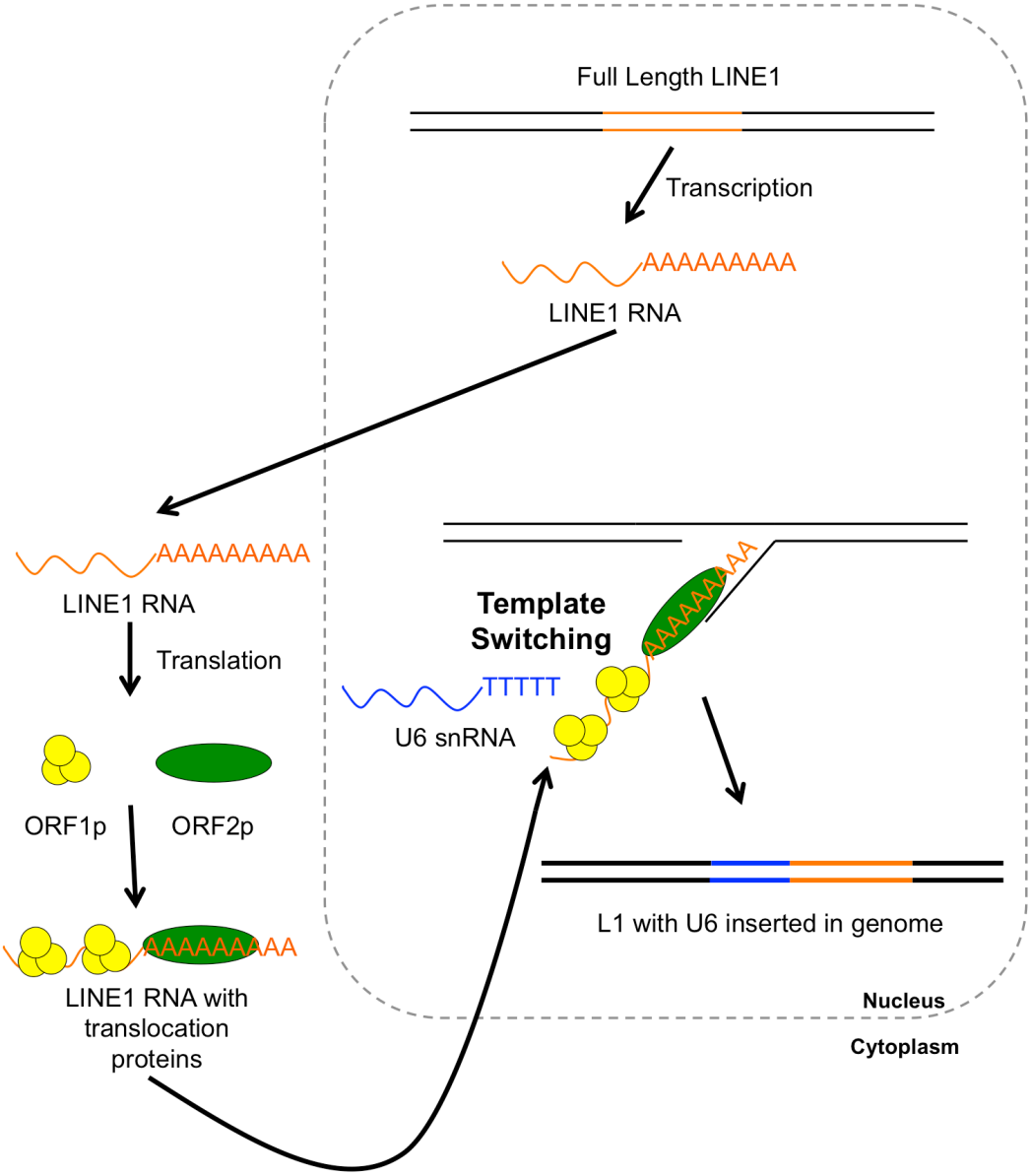


Figure 3.1: Previously suggested mechanism for U6/L1 fusion in human genome.
In previous research, U6/L1 fusion in genome happens when L1 is inserted to genome by template switching when reverse transcription takes place.

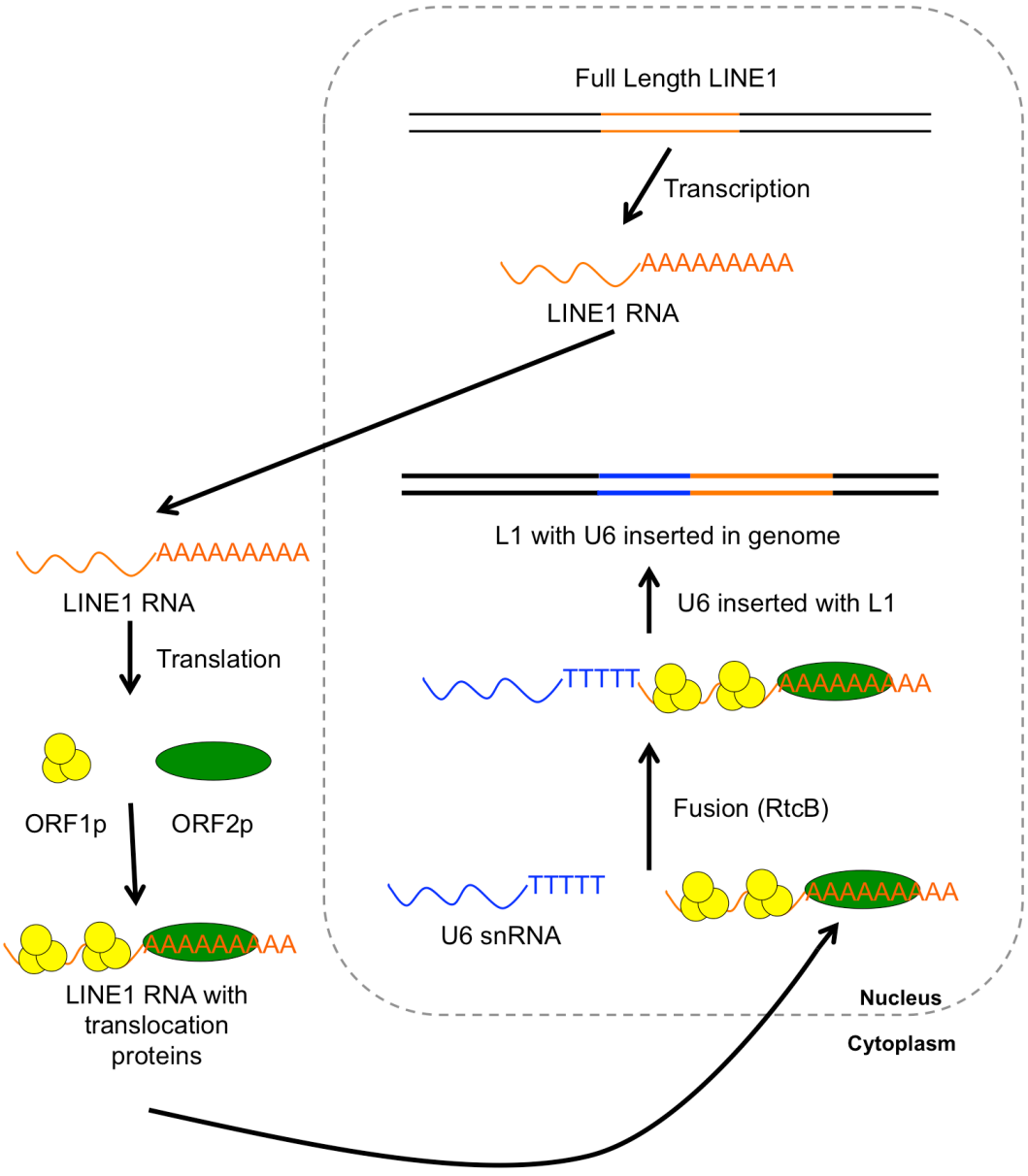


Figure 3.2: Formation of U6/L1 landmarks in genome by RNA level fusion between U6 and L1 catalyzed by RtcB.
In this study, we hypothesized that U6 and L1 RNA molecules form the fusion RNA first, then is inserted into human genome by L1 insertion mechanism catalyze by enzyme RtcB.

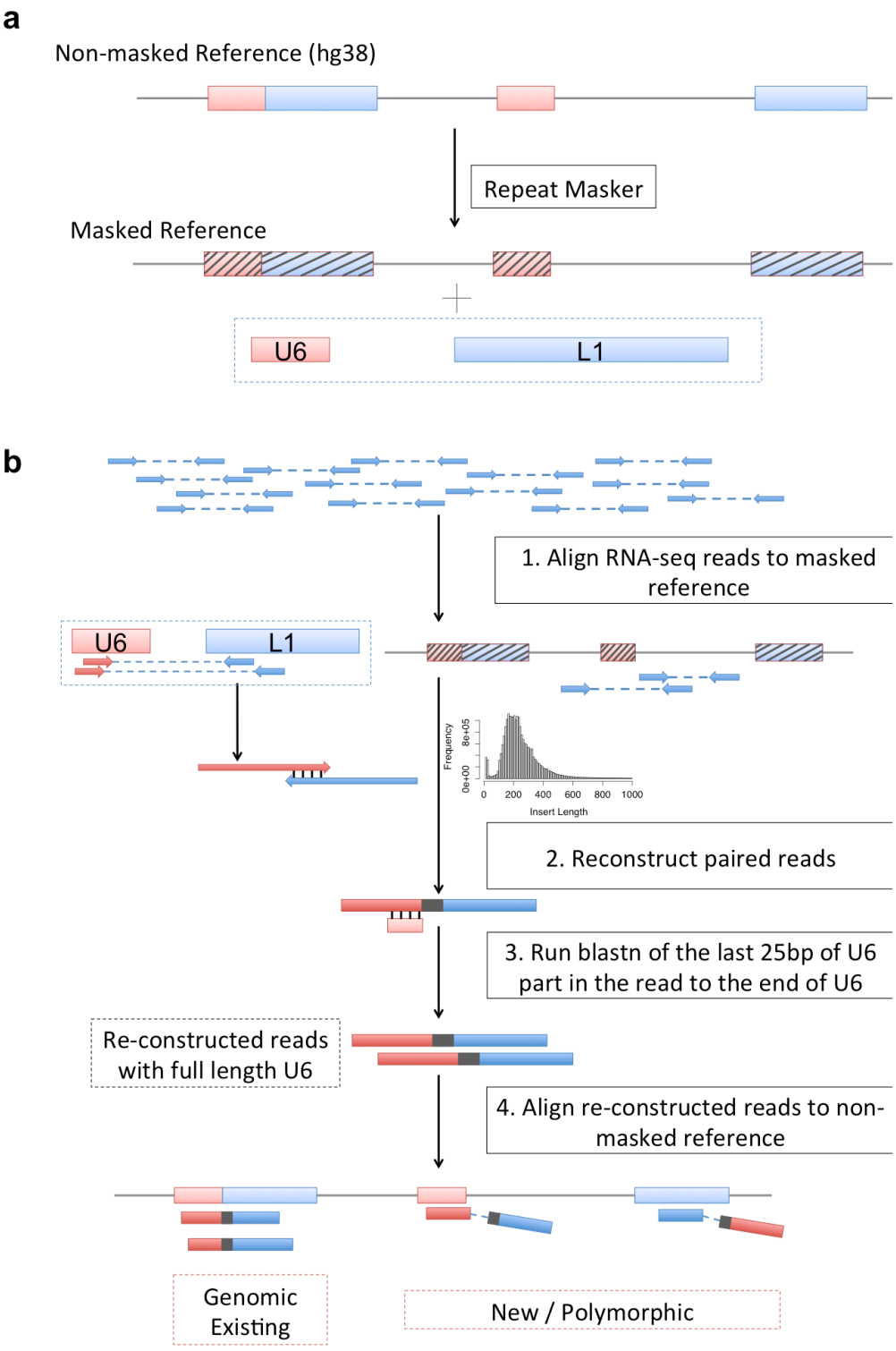

a
Non-masked Reference (hg38)
Repeat Masker
Masked Reference
U6
L1
b
1. Align RNA-seq reads to masked reference
U6
L1
Frequency
0e+00
8e+05
0
200
400
600
800
1000
Insert Length
2. Reconstruct paired reads
3. Run blastn of the last 25bp of U6 part in the read to the end of U6
Re-constructed reads with full length U6
4. Align re-constructed reads to non-masked reference
Genomic Existing
New / Polymorphic

Figure 3.3: Method for identification of U6/L1 fusion from RNA-seq data.
a) Customized human reference genome: We masked out all repetitive sequences from human genome GRCh38 using repeat masker, then added the only copy of U6 and L1 into the customized reference for future alignment of RNA-seq data.
b) Identification of RNA level U6/L1 fusion events: We applied bwa mem to align the reads to align the RNA-seq reads to the customized reference genome. We used FLASH to assemble the pair ended reads mapped to U6 and L1 to longer single ended reads. To ensure the reconstructed reads containing full length U6, we applied blastn of the assembled reads with the last 25 base pair of U6. We re-align the reads containing the last 25 base pair to the non-masked human genome GRCh38 to exclude the existing genomic U6/L1 events.

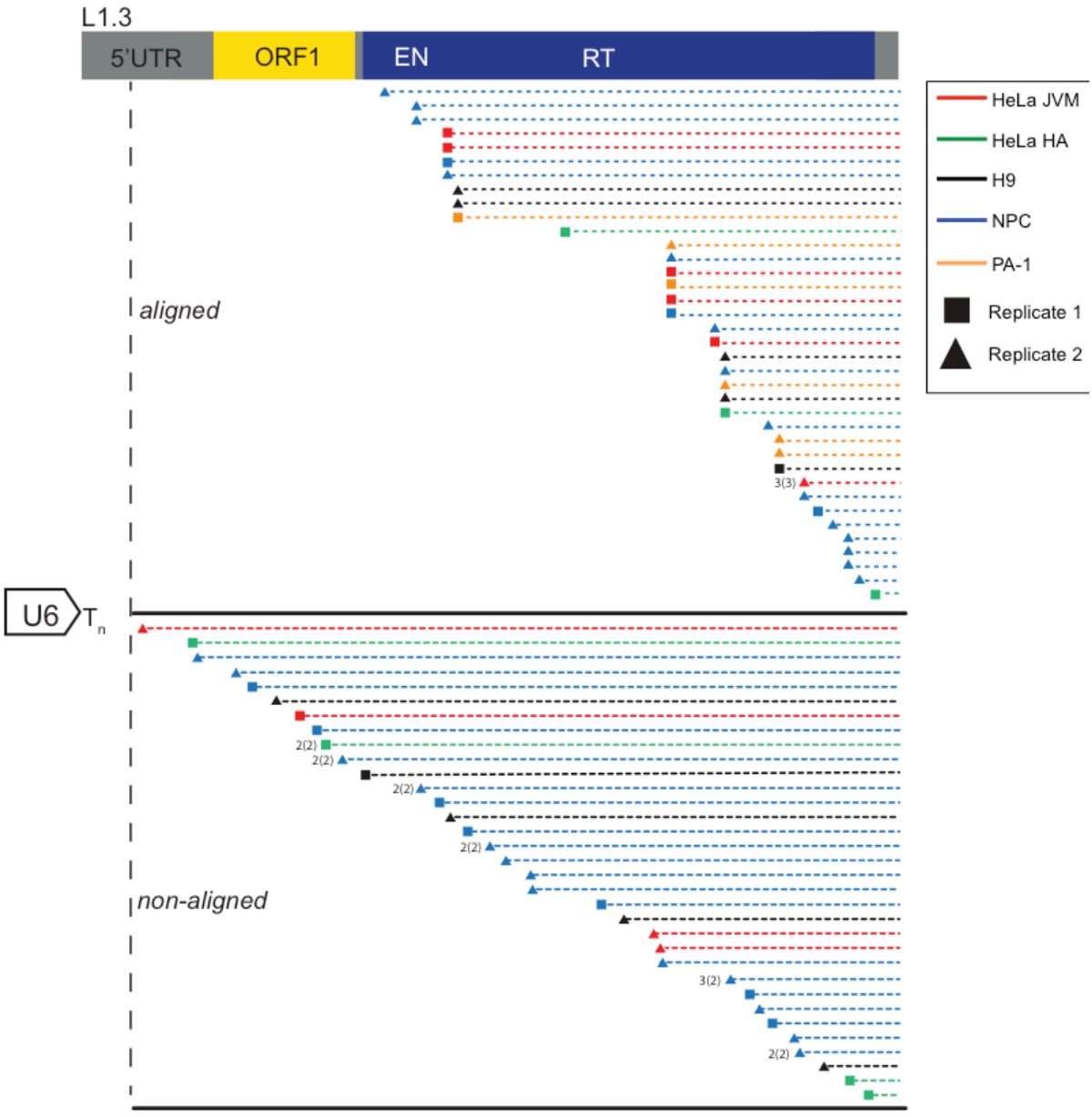

L1.3
5'UTR
ORF1
EN
RT
HeLa JVM
HeLa HA
H9
NPC
PA-1
Replicate 1
Replicate 2
aligned
U6
T_n
non-aligned

Figure 3.4: U6/L1 fusion events identified from 5 different samples RNA-seq data.

The top panel shows reads which could be aligned back to the genome, compared to the bottom panel where reads do not align back to the genome, most of the reads aligned have multiple supportive reads from different cell lines, representing the same genomic U6/L1 fragments got transcribed. The uniqueness of reads in the non-aligned panel after excluding the PCR duplicates shows the RNA level U6/L1 fusion events were happening at single molecule level. There is no specific hot-spot in L1 observed which preferentially fused with U6 snRNA.

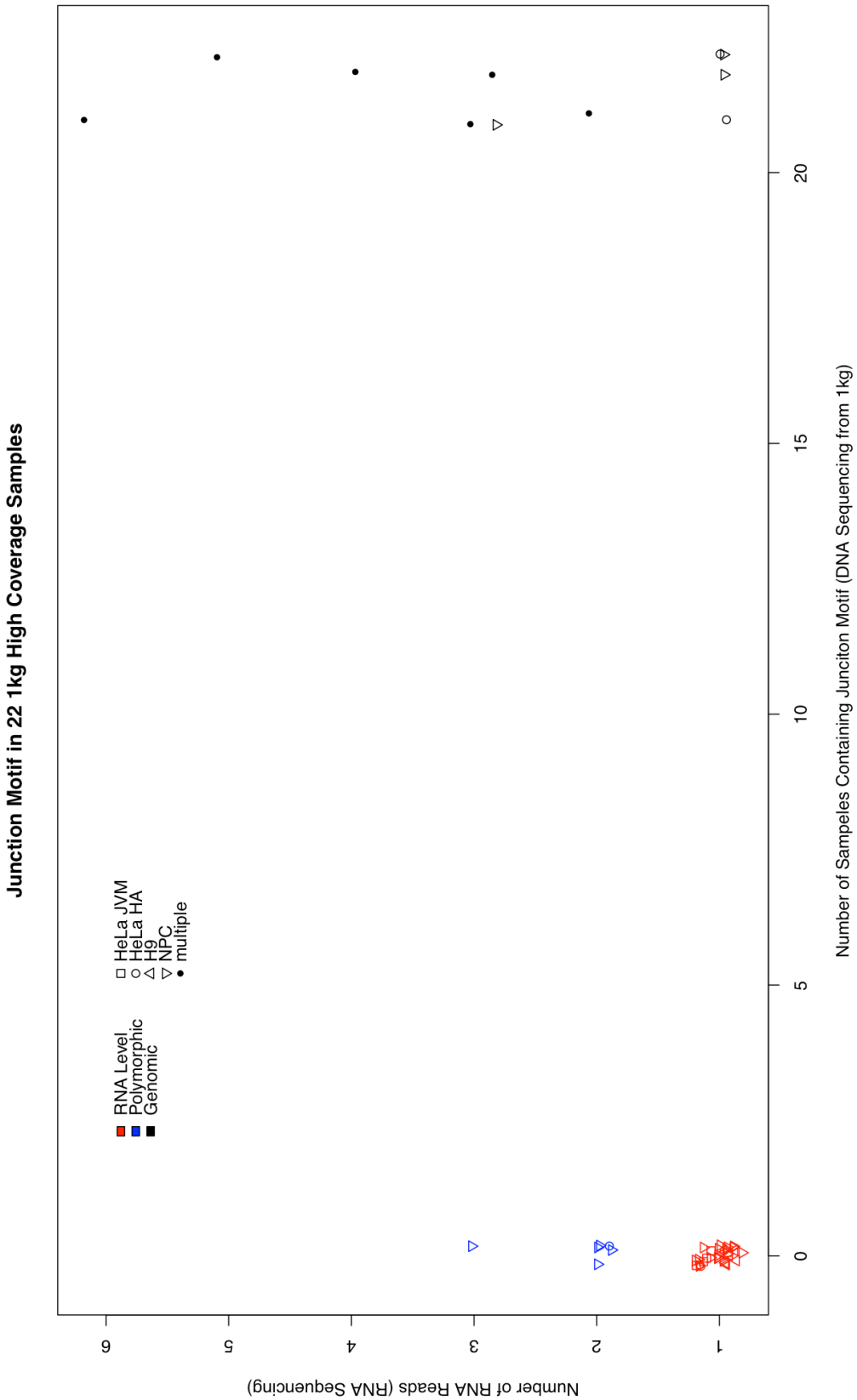

Junction Motif in 22 1kg High Coverage Samples
RNA Level
Polymorphic
Genomic
HeLa JVM
HeLa HA
H9
NPC
multiple
Number of RNA Reads (RNA Sequencing)
Number of Sampeles Containing Junciton Motif (DNA Sequencing from 1kg)
1
2
3
4
5
6
0
5
10
15
20

Figure 3.5: U6/L1 fusion events in high coverage 1000 Genome Project whole genome sequencing data.
For the 25 base pair *k*-mers around the junction of non-aligned reads (red (U6/L1 fusion with only 1 supportive read) and blue (U6/L1 fusion with more than 1 supportive reads, but are likely PCR duplicates)) identified from five cell lines, we were not able to find any matched *k*-mers in the 22 1000 Genome whole genome sequencing samples with high coverage. For the 25 base pair *k*-mers around the junction of aligned reads (black) identified from five cell lines, we were able to find all of them in at least 21 samples out of the 22. For the ones which were not found, we have identified SNP/sequencing errors which lead to the mismatches in *k*-mers searching.

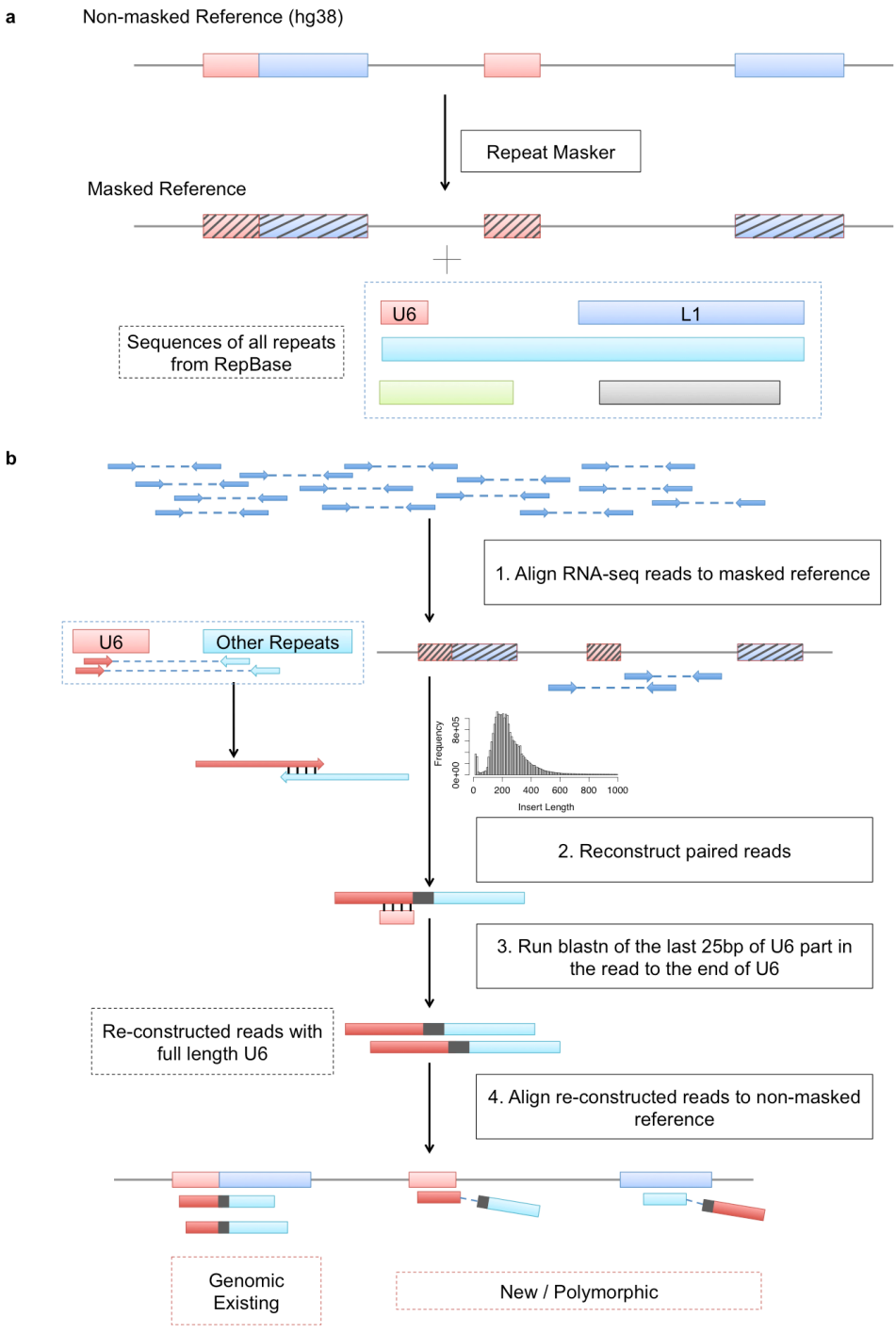
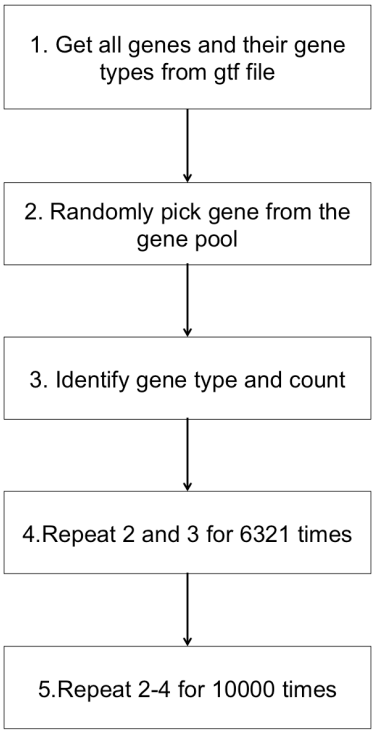

a
Non-masked Reference (hg38)
Repeat Masker
Masked Reference
Sequences of all repeats from RepBase
U6
L1
b
1. Align RNA-seq reads to masked reference
U6
Other Repeats
Frequency
0e+00
8e+05
0
200
400
600
800
1000
Insert Length
2. Reconstruct paired reads
3. Run blastn of the last 25bp of U6 part in the read to the end of U6
Re-constructed reads with full length U6
4. Align re-constructed reads to non-masked reference
Genomic Existing
New / Polymorphic

Figure 3.6: Method for identification of U6-other sequences fusion from RNA-seq data.
a) Customized human reference genome: We masked out all repetitive sequences from human genome GRCh38 using repeat masker, then added all sequences in RepBase into the customized reference for future alignment of RNA-seq data.
b) Identification of RNA level U6-other sequences fusion events: We applied bwa mem to align the reads to align the RNA-seq reads to the customized reference genome. We used FLASH to assemble the pair ended reads mapped to U6 and L1 to longer single ended reads. To ensure the reconstructed reads containing full length U6, we applied blastn of the assembled reads with the last 25 base pair of U6. We re-align the reads containing the last 25 base pair to the non-masked human genome GRCh38 to exclude the existing genomic U6-other sequences events.

a



b

Gene Name	Gene Type	Number of Events
SNORD3A	snoRNA	623
HERC3	protein coding	527
SNORD3C	snoRNA	425
U2	snRNA	202
PAK2	protein coding	181
RNU4-2	snRNA	155
RNU12	snRNA	89
MALAT1	lincRNA	88
RNU4-1	snRNA	54
L1	repetitive element	44
SNORD3B-2	snoRNA	43
RNU2-2P	snRNA	32
SNORD3B-1	snoRNA	31
SNORD15A	snoRNA	30
RNU2-1	snRNA	29
RNU5B-1	snRNA	26
SNORA8	snoRNA	26
SNORA70	snoRNA	24
RNU5A-1	snRNA	20

c

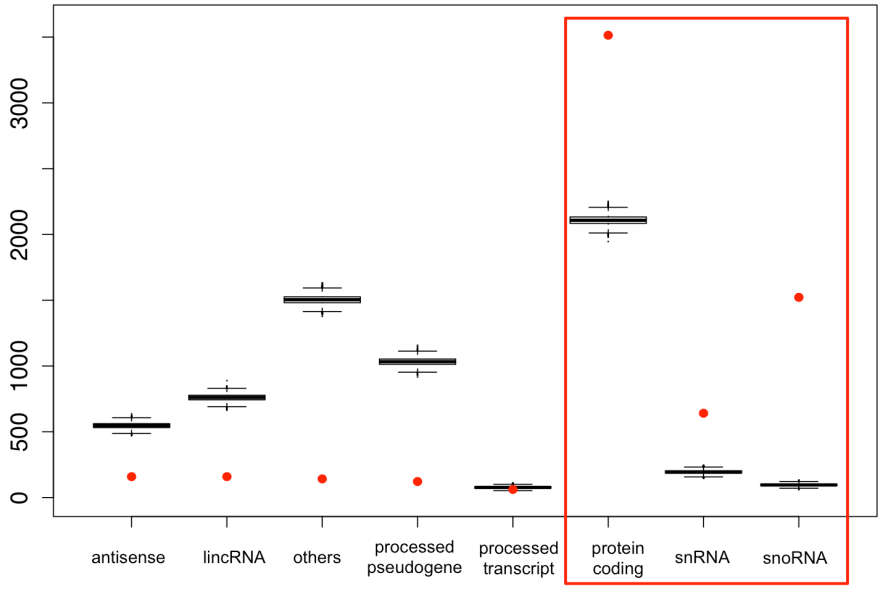
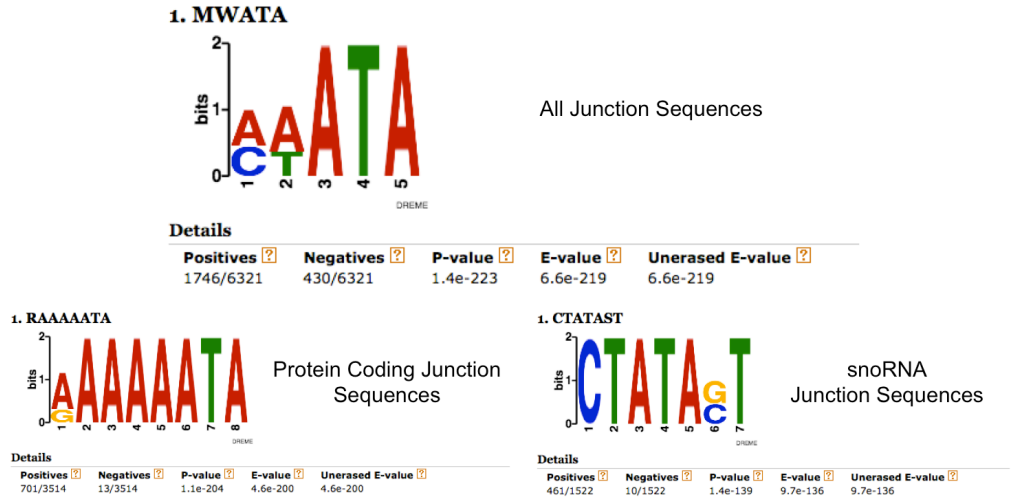
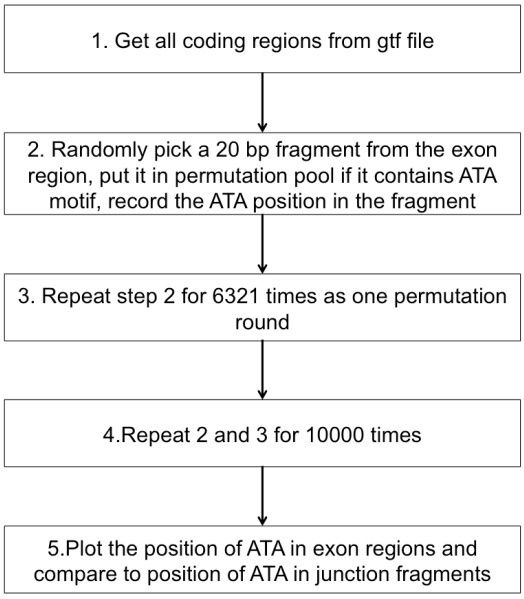


Figure 3.7: Analysis of gene type enrichment for U6 RNA level fusions.
a) Permutation method for enrichment analysis of genes fused with U6. In order to test if the number of genes fused with U6 in each gene categories, we permutated the number of genes by random selecting each different type of genes for 6321 times, which is the total number gene fusions identified. We repeated this process for 10,000 times.
b) Top genes fused with U6 at RNA level. Most of genes fused with U6 at RNA level are snRNAs or snoRNAs.
c) Genes enriched with U6 RNA fusions with permutation result. Compared with random permutation result, U6 fused with protein coding genes, snRNAs and snoRNAs are more enriched than the other kinds.

a


1. MWATA
bits
DREME
All Junction Sequences
Details
Positives 1746/6321
Negatives 430/6321
P-value 1.4e-223
E-value 6.6e-219
Unerased E-value 6.6e-219
1. RAAAAATA
Protein Coding Junction Sequences
Positives 701/3514
Negatives 13/3514
P-value 1.1e-204
E-value 4.6e-200
Unerased E-value 4.6e-200
1. CTATAST
snoRNA Junction Sequences
Positives 461/1522
Negatives 10/1522
P-value 1.4e-139
E-value 9.7e-136
Unerased E-value 9.7e-136

b


1. Get all coding regions from gtf file
2. Randomly pick a 20 bp fragment from the exon region, put it in permutation pool if it contains ATA motif, record the ATA position in the fragment
3. Repeat step 2 for 6321 times as one permutation round
4.Repeat 2 and 3 for 10000 times
5.Plot the position of ATA in exon regions and compare to position of ATA in junction fragments

c

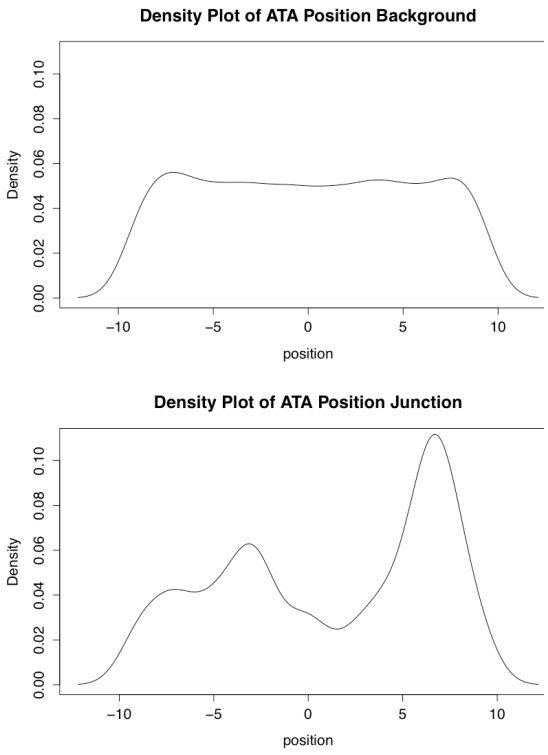
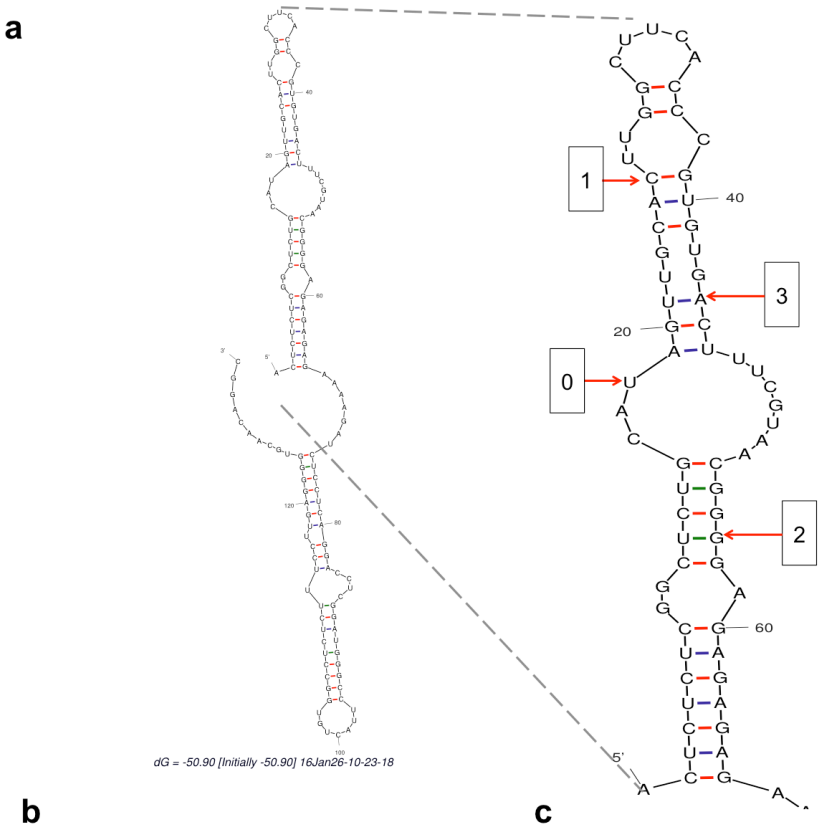

Density Plot of ATA Position Background
Density
position
Density Plot of ATA Position Junction
Density
position

Figure 3.8: Motif analysis at junctions of U6 fusions at RNA level.
a) AUA motif enriched in all U6 RNA level fusion junctions. Based on junction analysis, we identified 'AUA' motif enriched in all three kinds of most enriched gene types fused with U6 at the RNA level.
b) Permutation experiment for motif enrichment analysis. We selected all 20 base pair exon fragments containing 'AUA' motif from GRCh38 genome. From the pool of genetic fragments containing 'AUA' motif, we randomly selected the candidate fragments from the pool for 6321 times and recoded the position of 'AUA' motif in the selected gene fragments. We repeated this process for 10,000 times for background distribution of 'AUA' motif in the genomic context of the transcriptome.
c) 'AUA' motif position distributed in 20 base pair DNA fragments in random genome and in 20 base pair of junction motif with U6 insertion. We have observed a significant enrichment of 'AUA' motif at the positive position around 5 base pairs away from the insertion point compared to the negative control of 'AUA' motif in genomic context.



b

1. Get secondary structure for each RNA which has a fusion event by using Vienna

2. Randomly pick positions on each RNA for the same amount of times as the fusion events

3. Calculate the distance of the randomly generated positions to the closest loop, record the distance

4. Repeat step 2 and 3 for 10000 times

5.Plot the permutation result and compare to real distance result

c

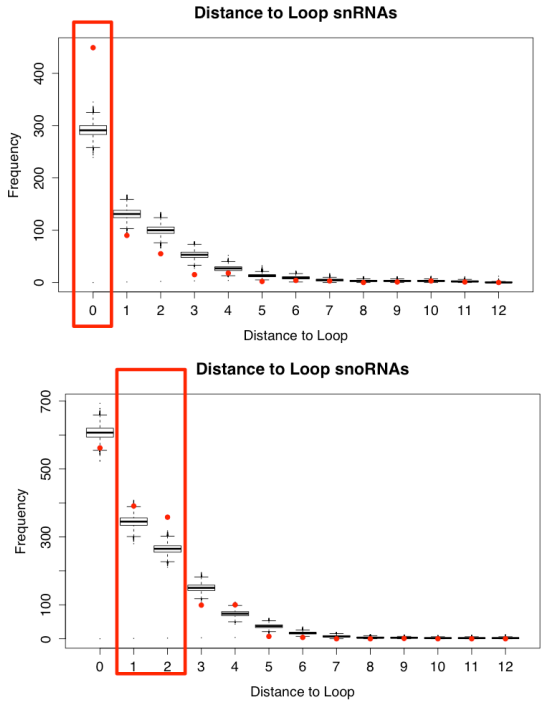


Figure 3.9: Method for identification of U6-other sequences fusion from RNA-seq data.
a) Illustration of secondary structure position counts using a secondary structure example for SNORA64. The relative distance of each base in the secondary structure to the closest loop is counted as the distance for U6 fusion point.
b) Permutation experiment for negative control of secondary structure analysis of U6 fusion genes. We applied Vienna to build secondary structure for all snRNAs and snoRNAs for permutation experiment. We randomly selected positions in these RNAs to calculate the distance to loops for 6321 times. We then repeated this process for 10,000 times as background distribution for genomic context of distance to loops.
c) U6 fusion point distance too loop compared to negative control from genomic context. For snRNAs, U6 fusion events are enriched in loops, while for snoRNAs, U6 fusion sites are enriched 2 bases away from the loop.

Table 3.1: Human U6 snRNA post-transcriptional modifications.

Modification	Position	Enzyme or guide RNA
Terminal Modifications		
5’ Capping	5’ γ-Monomethyl	Bin3/MePCE
3’ End trimming	3’ Terminal 2’,3’-cyclic phosphate	Usb1/Mpn1
3’ End oligouridylation	3’ End	U6 TUTase
3’ End oligoadenylation	3’ End	TRAMP?
Internal Modifications		
Pseudouridylation	U31	?
	U40	?
	U86	?
Ribose 2’-O-methylation	A47	mgU6-47
	A53	mgU6-53, mgU6-53B
	G54	?
	C60	MBII-166
	C62	?
	C63	?
	A70	?
	C77	mgU6-77
N-6-adenosine methylation	A43	METTL16
N-2-guanosine methylation	G72	?

Modified from Didychuk et al. 2018.

Table 3.2: Junction analysis of aligned RNA-seq U6/L1 sequences.

L1.3 Junction	L1 subfamily	U6/L1 junction -20	Junction Ts	U6/L1 junction + 20
2052	L1PA10	GTAAATGGGCTAAATGCCCC	5	AATTAAAAGACACAGAATGG
2234	L1PA3	AATCCTAGTCTCTGATAAAA	5	CAGACTTTAAACCAACAAAG
2568	L1PA4	AATCAACAGAATATACATTC	8	TTTTCAGCACCACACCACAC
2598	L1PA5	CCACATCACACTTATTCCAA	5	AATTGACCACATAGTTGGAA
3450	L1PA7	CTACCAGGAGTACAAAGAGG	5	AGCTGGTACCAATCCTTCTG
4268	L1PA5	ATGAGTGAACTCCCATTCAC	5	AATTGCTTCAAAGAGAATAA
4611	L1PA2	GGAGGCATCACACTACCTGA	5	CTTCAAACTATACTACAAGG
s4683	L1PA2	CAAAACAGAGATATAGATCA	5	ATGGAACAGAACAGAGCCCT
5030	L1PA7	AATTGACAAATGGGATCTAA	6	TTAAAATAAAGAGCTTCTGC
5095	L1PA7	TGAACAGACAACCTACAGAA	5	TGGAAGAAAATTTTTGCAAT
5281	L1PA2	ACATGAAAAAATGCTCATCA	5	TCACTGGCCATCAGAGAAAT
5358	L1PA5	GTTAGAATGGCGATCATTAA	5	AAAGTCAGGAAACAACAGGT
5558	L1PA7	TTATAAATCATTCTACTGTA	5	AAAACACATGCACACATGTT
5647	L1PA2	GTCCAACAATGATAGACTGG	5	ATTAAGAAAATGTGGCACAT
5720	L1PA5	TGATGAGTTCATGTCCTTTG	5	TAGGGACATGGATGAAGCTG
5906	L1PA3	GGGGGAGGGATAGCATTAGG	5	AGATATACCTAATGCTAAAT

Table 3.3: Junction analysis of non-aligned RNA-seq U6/L1 sequences.

L1.3 Junction	U6/L1 junction -20	Junction Ts	U6/L1 junction + 20
474	TGAGGCTTGAGTAGGTAAAC	5	AAAGTAGCCGGGAAGCTCGA
832	TTAGAAGGAAAACTAACAAC	5	CAGAAAGGACATCTACACCG
864	ATCTACACCGAAAACCCATC	6	TGTACATCACCATCATCAAA
1125	GTTAAAAACTTTGAAAAAAA	5	ATTAGACGAATGGCTAACTA
1231	GTGACGAATGCACAAGCTTC	5	AGTAGCCGATTCGATCAACT
1454	AAACACTCTGCAGGATATTA	6	TCCAGGAGAACTTCCCCAAT
1559	AAGAGCAACTCCAAGACACA	6	TAATTGTCAGATTCACCAAA
1748	AAAGAATTTTCAACCCAGAA	8	TTTCATATCCAGCCAAACTA
1752	AATTTTCAACCCAGAATTTC	5	ATATCCAGCCAAACTAAGCT
1824	GACAAGCAAATGTTGAGAGA	6	TTTTGTCACCACCAGGCCTG
2025	TCACACATAACAATATTAAC	6	TTTAAATATAAATGGACTAA
2395	TCAGTGACCTACAAAGAGAC	7	TTAGACTCCCACACATTAAT
2657	ATGTAAAAGAACAGAAATTA	6	TAACAAACTATCTCTCAGAC
2728	AGAATCTCACTCAAAGCCGC	6	TCAACTACATGGAAACTGAA
2849	AGACACCACATACCAGAATC	5	TCTGGGACGCATTCAAAGCA
2884	AAGCAGTGTGTAGAGGGAAA	5	TTTATAGCACTAAATGCCTA
3056	AATAGAGACACAAAAAACCC	6	TTCAAAAAATCAATGAATCC
3262	TCTACGCAAATAAACTAGAA	5	AATCTAGAAGAAATGGATAC
3307	ACACATACACTCTCCCAAGA	5	CTAAACCAGGAAGAAGTTGA
3311	CACATACACTCTCCCAAGAC	6	TAAACCAGGAAGAAGTTGAA
3872	TATTGATGGGACGTATTTCA	5	AAATAATAAGAGCTATCTAT
3945	CAAAAACTGGAAGCATTCCC	8	TTTGAAAACCGGCACAAGAC
4131	CTAGAAAACCCCATCGTCTC	6	AGCCCAAAATCTCCTTAAGC
4190	CAGGATACAAAATCAATGTA	5	CAAAAATCACAAGCATTCTT
4783	AATGGGGAAAGGATTCCCTA	8	TTTAATAAATGGTGCTGGGA
4854	CCCTTCCTTACACCTTATAC	6	AAAAATCAATTCAAGATGGA
4887	AATTCAAGATGGATTAAAGA	6	TTTAAACGTTAAACCTAAAA
5029	AAATTGACAAATGGGATCTA	5	ATTAAACTAAAGAGCTTCTG
5145	GACAAAGGGCTAATATCCAG	5	AATCTACAATGAACTCAAAC
5197	AAAAAACAAACAACCCCATC	7	AAAAAGTGGGCGAAGGACAT
5416	AAATAGGAACACTTTTACAC	5	TGTTGGTGGGACTGTAAACT
5593	GTATGTTTATTGCGGCACTA	6	TTCACAATAGCAAAGACTTG
5757	TTGGAAACCATCATTCTCAG	6	TAAACTATCGCAAGAACAAA

Table 3.4: Sequences features of the 25bp U6/L1 junction sequences motifs of the "aligned", "non-aligned", and putative "artifact" RNA-seq chimeras

Category	L1.3 Junction	25bp Junction Motif	# Supporting reads	Cell Line
aligned	2052	**5'-CATTCTGTA**TTTTTAATTAAAAGAC	1	NPC
aligned	2234	**5'-CGTTCTGTA**TTTTTCAGACTCTAAA	2	NPC
aligned	2568	**5'-CGTTCCATT**TCTTTTTTCAGCACCA	4	NPC, JVM
aligned	2598	**5'-CGTTCCATA**TTTTTAATTGACCACA	3	H9, PA-1
aligned	3450	**5'-CGTTCCATA**TTTTTACTGGTACCAT	1	HA
aligned	4268	**5'-CGTTCCATA**TTTTTAATTGCTTCAA	6	PA-1, NPC, JVM
aligned	4611	**5'-CGTTCCATA**TTTTTCTTCAAACTAT	2	NPC, JVM
aligned	4683	**5'-CGTTCCATA**TTTTTATGGAACAGAA	5	H9, NPC, PA-1, HA
aligned	5030	**5'-CGTTCCGTA**TTTTTTAAACTAAAGA	1	NPC
aligned	5095	**5'-CATTCCATA**TTTTTGGGAGAAAATT	3	PA-1, H9
aligned	5281	**5'-CGTTCCATA**TTTTTCACTGGCCATC	4	JVM, NPC
aligned	5358	**5'-CGTTCCATA**TTTTTAAAGTCAGGAA	1	NPC
aligned	5558	**5'-AGTTCCGTA**TTTTTAAAACACATGC	1	NPC
aligned	5647	**5'-CGTTCCATA**TTTTTATTAAGAAAAT	3	NPC
aligned	5720	**5'-CGTTCCATA**TTTTTAGGGACATGGA	1	NPC
aligned	5906	**5'-CGTTCCATA**TTTTTAGATATACCTA	1	HA
non-aligned	474	**5'-CGTTCCATA**TTTTTAAAGCGTCCTG	1	JVM
non-aligned	832	**5'-CGTTCCATA**TTTTTAACAGAAAGGA	1	HA
non-aligned	864	**5'-CGTTCCATA**TTTTTTGTACATCACC	1	NPC
non-aligned	1125	**5'-CGTTCCATA**TTTTTATTGACGAATG	1	NPC
non-aligned	1231	**5'-CGTTCCATA**TTTTTAGTAGCTGATT	1	NPC
non-aligned	1454	**5'-CGTTCCATA**TTTTTTCCAGGAGAAC	1	H9
non-aligned	1559	**5'-CGTTCCATA**TTTTTTAATTGTCAGA	1	JVM
non-aligned	1748	**5'-CGTTCCATA**TTTTTTTTCATATCCA	2(2)	HA
non-aligned	1752	**5'-CGTTCCATA**TTTTTATATCCAGCCA	1	NPC
non-aligned	1824	**5'-CGTTCCATA**TTTTTTTGTCACCACC	2(2)	NPC
non-aligned	2025	**5'-CGTTCCATA**TTTTTTAAATGTAAAT	1	H9
non-aligned	2395	**5'-CGTTCCATA**TTTTTTTAGACTCCCA	2(2)	NPC
non-aligned	2657	**5'-CGTTCCATA**TTTTTTAACAAACTGT	1	NPC
non-aligned	2728	**5'-CGTTCCATA**TTTTTTCAACTACATA	1	H9
non-aligned	2849	**5'-CGTTCCATA**TTTTTCTGGGACACAT	1	NPC
non-aligned	2884	**5'-CGTTCCATA**TTTTTATAGCACTAAA	2(2)	NPC
non-aligned	3056	**5'-CGTTCCATA**TTTTTTCAAAAAATCA	1	NPC
non-aligned	3262	**5'-CGTTCCATA**TTTTTAATCTAGAAGA	1	NPC
non-aligned	3307	**5'-CGTTCCATA**TTTTTAGGCTAAACCA	1	NPC

non-aligned	3311	**5'-CGTTCCATA**TTTTTTAAACCAGGCA	1	JVM
non-aligned	3872	**5'-CGTTCCATA**TTTTTAAATAATAAGA	1	NPC
non-aligned	3945	**5'-CGTTCCATA**TTTTTTTTGAAAACTG	1	H9
non-aligned	4131	**5'-CGTTCCATA**TTTTTTAGCCCAAAAT	1	JVM
non-aligned	4190	**5'-CGTTCCATA**TTTTTATGTTCAAAAA	1	NPC
non-aligned	4783	**5'-CGTTCCATA**TTTTTTTTAATAAATG	3(2)	NPC
non-aligned	4854	**5'-CGTTCCATA**TTTTTTATACAAAAAA	1	NPC
non-aligned	4887	**5'-CGTTCCATA**TTTTTTAAACGTTAGA	1	NPC
non-aligned	5029	**5'-CGTTCCATA**TTTTTATTAAACTAAA	1	NPC
non-aligned	5145	**5'-CGTTCCATA**TTTTTAATCTACAATG	1	NPC
non-aligned	5197	**5'-CGTTCCATA**TTTTTTTAAAAAGTGG	2(2)	NPC
non-aligned	5416	**5'-CGTTCCATA**TTTTTGTTGGTGGGAC	1	H9
non-aligned	5593	**5'-CGTTCCATA**TTTTTTCACAATAGCA	1	HA
non-aligned	5757	**5'-CGTTCCATA**TTTTTTAAACTATCGC	1	HA
non-aligned	934	**5'-CGTTCCATA**TTTTTTTCTGCTCTGT	1	NPC
non-aligned	4322	**5'-CGTTCCATA**TTTTTTTCACATCCCT	1	JVM
non-aligned	5259	**5'-CGTTCCATA**TTTTTTTGTTGGCCAC	1	NPC
non-aligned	5343	**5'-CGTTCCATA**TTTTTAATGATGACGT	1	NPC
artifact	-	**5'-TTCGTGAAGCGT**ATACACCAATAAC	1	NPC
artifact	-	**5'-GACACGCAAATTC**TATTGAGGGTTT	2(2)	H9
artifact	-	**5'-ACACGCAAATTC**ATCAGTGAATCCA	1	NPC
artifact	-	**5'-ACGCAAATTCG**ATAAAAATCCTAGA	2(2)	H9
artifact	-	**5'-GACACGCAAATTC**TTTTTATGGCTG	1	NPC
artifact	-	**5'-CACGCAAATTC**AAAATACTGGCAAA	1	HA
artifact	-	**5'-ACGCAAATTCGATGAA**ATAAAGCAT	1	NPC
artifact	-	**5'-GACACGCAAATTCTTG**GGTTGGTTC	2(2)	H9
artifact	-	5'-**CACGCAAATTCTTGAAG**ATGACATG	1	H9
artifact	-	**5'-CACGCAAATTCG**GTACCTGAAAGGA	1	NPC
artifact	-	**5'-ATGACACGCAAATTCG**ACAAAGGGC	1	NPC

Table 3.5: Characterization of 16 genomic U6/L1 chimeras.

L1.3 Junction	L1 Subfamily	Genomic Position	Remarks
2052	L1PA10	chrX:102678813-102674130	ARMCX5-GPRASP2 Intron
2234	L1PA3	chr13:48987911-48988334	FNDC3A intron
2568	L1PA4	chr1:180758722-180762284	XPR1 intron
2598	L1PA5	chr3:98805084-98801701	DCBLD2 intron
3450	L1PA7	chr8:103384961-103387948	intergenic
4268	L1PA5	chr13:72706123-72704270	intergenic
4611	L1PA2	chr18:68858934-68860488	CCDC102B intron
4683	L1PA2	chr4:39296252-39297711	RFC1 intron
5030	L1PA7	chr1:42569034-42570125	CCDC30 intron
5095	L1PA7	chr14:37434573-37433236	MIPOL1 intron
5281	L1PA2	chr3:196784226-196785086	PAK2 intron
5358	L1PA5	chr4:109992325-109993102	EGF intron
5558	L1PA7	chr14:102865856-102866427	TRAF3 intron
5647	L1PA2	chr15:65553187-65552698	HACD3 intron
5720	L1PA5	chr4:76532327-76531908	SHROOM3 intron
5906	L1PA3	chr2:174558072-174557836	intergenic

Table 3.6: U6/L1 fusion junction search in 1000 Genomes Project samples.

25bp Junction Motif	Number of 1kg Sample (25bp)	Classification	Number of Reads Called	Sample(s)
CGTTCCATATTTTTGTTGGTGGGAC	0	non-aligned	1	H9
CGTTCCATATTTTTTCAACTACATA	0	non-aligned	1	H9
CGTTCCATATTTTTTAAACTATCGC	0	non-aligned	1	HA
CGTTCCATATTTTTAATGATGACGT	0	non-aligned	1	NPC
CGTTCCATATTTTTTTGTTGGCCAC	0	non-aligned	1	NPC
CGTTCCATATTTTTTAATTGTCAGA	0	non-aligned	1	JVM
CGTTCCATATTTTTTAAATGTAAAT	0	non-aligned	1	H9
CGTTCCATATTTTTTCCAGGAGAAC	0	non-aligned	1	H9
CGTTCCATATTTTTAAAGCGTCCTG	0	non-aligned	1	JVM
CGTTCCATATTTTTTAAACGTTAGA	0	non-aligned	1	NPC
CGTTCCATATTTTTATGTTCAAAAA	0	non-aligned	1	NPC
CGTTCCATATTTTTTTCTGCTCTGT	0	non-aligned	1	NPC
CGTTCCATATTTTTAGTAGCTGATT	0	non-aligned	1	NPC
CGTTCCATATTTTTATATCCAGCCA	0	non-aligned	1	NPC
CGTTCCATATTTTTTAGCCCAAAAT	0	non-aligned	1	JVM
CGTTCCATATTTTTTAAACCAGGCA	0	non-aligned	1	JVM
CGTTCCATATTTTTATTGACGAATG	0	non-aligned	1	NPC
CGTTCCATATTTTTAGGCTAAACCA	0	non-aligned	1	NPC
CGTTCCATATTTTTCTGGGACACAT	0	non-aligned	1	NPC
CGTTCCATATTTTTAACAGAAAGGA	0	non-aligned	1	HA
CGTTCCATATTTTTTGTACATCACC	0	non-aligned	1	NPC
CGTTCCATATTTTTTATACAAAAAA	0	non-aligned	1	NPC
CGTTCCATATTTTTTTCACATCCCT	0	non-aligned	1	JVM
CGTTCCATATTTTTAAATAATAAGA	0	non-aligned	1	NPC
CGTTCCATATTTTTTTTGAAAACTG	0	non-aligned	1	H9
CGTTCCATATTTTTTCACAATAGCA	0	non-aligned	1	HA
CGTTCCATATTTTTTAACAAACTGT	0	non-aligned	1	NPC
CGTTCCATATTTTTATTAAACTAAA	0	non-aligned	1	NPC
CGTTCCATATTTTTTCAAAAAATCA	0	non-aligned	1	NPC
CGTTCCATATTTTTAATCTACAATG	0	non-aligned	1	NPC

CGTTCCATATTTTTAATCTAGAAGA	0	non-aligned	1	NPC
CGTTCCATATTTTTTTTCATATCCA	0	non-aligned	2	HA
CGTTCCATATTTTTTTTAATAAATG	0	non-aligned	3	NPC
CGTTCCATATTTTTTTAGACTCCCA	0	non-aligned	2	NPC
CGTTCCATATTTTTTTGTCACCACC	0	non-aligned	2	NPC
CGTTCCATATTTTTATAGCACTAAA	0	non-aligned	2	NPC
CGTTCCATATTTTTTTAAAAAGTGG	0	non-aligned	2	NPC
CGTTCCATATTTTTAGATATACCTA	22	aligned	1	HA
CGTTCCATATTTTTCACTGGCCATC	22	aligned	4	multiple
CGTTCCATATTTTTATTAAGAAAAT	21	aligned	3	NPC
CGTTCCATATTTTTAATTGCTTCAA	21	aligned	6	multiple
CGTTCCATATTTTTAAAGTCAGGA	22	aligned	1	NPC
CGTTCCATATTTTTACTGGTACCA	21	aligned	1	HA
CGTTCCATATTTTTCTTCAAACTA	21	aligned	2	multiple
CGTTCCATATTTTTAGGGACATGGA	22	aligned	1	NPC
CGTTCCATATTTTTAATTGACCACA	22	aligned	3	multiple
CATTCCATATTTTTGGGAGAAAATT	21	aligned	3	multiple
CGTTCCATATTTTTATGGAACAGAA	22	aligned	5	multiple
TTCGTGAAGCGTATACACCAATAAC	0	artifact	1	NPC
GACACGCAAATTCTATTGAGGGTTT	0	artifact	1	H9
ACACGCAAATTCATCAGTGAATCCA	0	artifact	1	NPC
ACGCAAATTCGATAAAAATCCTAGA	0	artifact	1	H9
GACACGCAAATTCTTTTTATGGCTG	0	artifact	1	NPC
CACGCAAATTCAAAATACTGGCAAA	0	artifact	1	HA
ACGCAAATTCGATGAAATAAAGCAT	0	artifact	1	NPC
GACACGCAAATTCTTGGGTTGGTTC	0	artifact	1	H9
CACGCAAATTCTTGAAGATGACATG	0	artifact	1	H9
CACGCAAATTCGGTACCTGAAAGGA	0	artifact	1	NPC
ATGACACGCAAATTCGACAAAGGGC	0	artifact	1	NPC

Table 3.7: 1000 Genomes Project sample numbers with population codes.

Sample	Population
HG00096	GBR
HG00268	FIN
HG00419	CHS
HG00759	CDX
HG01051	PUR
HG01112	CLM
HG01500	IBS
HG01565	PEL
HG01583	PJL
HG01595	KHV
HG01879	ACB
HG02568	GWD
HG02922	YRI
HG03052	MSL
HG03642	STU
HG03742	ITU
NA18525	CHB
NA18939	JPT
NA19017	LWK
NA19625	ASW
NA19648	MXL
NA20502	TSI
NA20845	GIH

Table 3.8: Number of genes fused with U6 in different categories.

Number of Fusion Events		
Gene Type	Count	Percentage
Protein Coding	3514	56%
snoRNA	1522	24%
snRNA	641	10%
lincRNA	159	3%
Antisense	159	3%
Processed Pseudogene	122	2%
Processed Transcript	62	1%
Others	142	2%

Table 3.9: Number of unique genes fused with U6 in different categories.

Number of Unique Genes		
Gene Type	Count	Percentage
Protein Coding	1854	80%
snoRNA	94	4%
snRNA	20	1%
lincRNA	52	2%
Antisense	99	4%
Processed Pseudogene	101	4%
Processed Transcript	17	1%
Others	77	3%

Reference

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), pp.68-74

Adey, A., Burton, J., Kitzman, J., Hiatt, J., Lewis, A., Martin, B., Qiu, R., Lee, C. and Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature, 500(7461), pp.207-211.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bailey, T., Boden, M., Buske, F., Frith, M., Grant, C., Clementi, L., Ren, J., Li, W. and Noble, W. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, 37(Web Server), pp.W202-W208.

Basame, S., Wai-lun Li, P., Howard, G., Branciforte, D., Keller, D. and Martin, S. (2006). Spatial Assembly and RNA Binding Stoichiometry of a LINE-1 Protein Essential for Retrotransposition. Journal of Molecular Biology, 357(2), pp.351-357.

Beck, C., Collier, P., Macfarlane, C., Malig, M., Kidd, J., Eichler, E., Badge, R. and Moran, J. (2010). LINE-1 Retrotransposition Activity in Human Genomes. Cell, 141(7), pp.1159-1170.

Beck, C., Garcia-Perez, J., Badge, R. and Moran, J. (2011). LINE-1 Elements in Structural Variation and Disease. Annual Review of Genomics and Human Genetics, 12(1), pp.187-215.

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

Brouha, B., Schustak, J., Badge, R., Lutz-Prigge, S., Farley, A., Moran, J. and Kazazian, H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proceedings of the National Academy of Sciences, 100(9), pp.5280-5285.

Brow, D. and Guthrie, C. (1988). Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. Nature, 334(6179), pp.213-218.

Burns, K. (2017). Transposable elements in cancer. Nature Reviews Cancer, 17(7), pp.415-424.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y. and Sverdlov, E. (2002). A New Family of Chimeric Retrotranscripts Formed by a Full Copy of U6 Small Nuclear RNA Fused to the 3′ Terminus of L1. Genomics, 80(4), pp.402-406.

Buzdin, A. (2003). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Research, 31(15), pp.4385-4390.

Coufal, N., Garcia-Perez, J., Peng, G., Yeo, G., Mu, Y., Lovci, M., Morell, M., O'Shea, K., Moran, J. and Gage, F. (2009). L1 retrotransposition in human neural progenitor cells. Nature, 460(7259), pp.1127-1131.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nature Genetics, 35(1), pp.41-48.

Didychuk, A., Butcher, S. and Brow, D. (2018). The life of U6 small nuclear RNA, from cradle to grave. RNA, 24(4), pp.437-460.

Dombroski, B., Mathias, S., Nanthakumar, E., Scott, A. and Kazazian, H. (1991). Isolation of an active human transposable element. Science, 254(5039), pp.1805-1808.

Domitrovich, A. (2003). Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. Nucleic Acids Research, 31(9), pp.2344-2352.

Doucet, A., Hulme, A., Sahinovic, E., Kulpa, D., Moldovan, J., Kopera, H., Athanikar, J., Hasnaoui, M., Bucheton, A., Moran, J. and Gilbert, N. (2010). Characterization of LINE-1 Ribonucleoprotein Particles. PLoS Genetics, 6(10), p.e1001150.

Doucet, A., Droc, G., Siol, O., Audoux, J. and Gilbert, N. (2015). U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. Molecular Biology and Evolution, 32(7), pp.1815-1832.

Doucet, A., Wilusz, J., Miyoshi, T., Liu, Y. and Moran, J. (2015). A 3′ Poly(A) Tract Is Required for LINE-1 Retrotransposition. Molecular Cell, 60(5), pp.728-741.

Englert, M., Sheppard, K., Aslanian, A., Yates, J. and Soll, D. (2011). Archaeal 3'-phosphate RNA splicing ligase characterization identifies the missing

component in tRNA maturation. Proceedings of the National Academy of Sciences, 108(4), pp.1290-1295.

Ergün, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Strätling, W. and Schumann, G. (2004). Cell Type-specific Expression of LINE-1 Open Reading Frames 1 and 2 in Fetal and Adult Human Tissues. Journal of Biological Chemistry, 279(26), pp.27753-27763.

Esnault, C., Maestre, J. and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nature Genetics, 24(4), pp.363-367.

Ewing, A. and Kazazian, H. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Research, 20(9), pp.1262-1270.

Fallmann, J., Will, S., Engelhardt, J., Grüning, B., Backofen, R. and Stadler, P. (2017). Recent advances in RNA folding. Journal of Biotechnology, 261, pp.97-104.

Feng, Q., Moran, J., Kazazian, H. and Boeke, J. (1996). Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. Cell, 87(5), pp.905-916.

Fica, S., Tuttle, N., Novak, T., Li, N., Lu, J., Koodathingal, P., Dai, Q., Staley, J. and Piccirilli, J. (2013). RNA catalyses nuclear pre-mRNA splicing. Nature, 503(7475), pp.229-234.

Garcia-Perez, J., Doucet, A., Bucheton, A., Moran, J. and Gilbert, N. (2007). Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome Research, 17(5), pp.602-611.

Garcia-Perez, J., Marchetto, M., Muotri, A., Coufal, N., Gage, F., O'Shea, K. and Moran, J. (2007). LINE-1 retrotransposition in human embryonic stem cells. Human Molecular Genetics, 16(13), pp.1569-1577.

Garcia-Perez, J., Morell, M., Scheys, J., Kulpa, D., Morell, S., Carter, C., Hammer, G., Collins, K., O'Shea, K., Menendez, P. and Moran, J. (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature, 466(7307), pp.769-773.

Gilbert, N., Lutz, S., Morrish, T. and Moran, J. (2005). Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. Molecular and Cellular Biology, 25(17), pp.7780-7795.

Grimaldi, G. and Singer, M. (1983). Members of the KpnI family of long interspersed repeated sequences join and interrupt α-satellite in the monkey genome. Nucleic Acids Research, 11(2), pp.321-338.

Haas, B., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C., Doak, T., Pochet, N., Sun, J., Wu, C., Gingeras, T. and Regev, A. (2017). STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq.

Hancks, D., Goodier, J., Mandal, P., Cheung, L. and Kazazian, H. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. Human Molecular Genetics, 20(17), pp.3386-3400.

Hilcenko, C., Simpson, P., Finch, A., Bowler, F., Churcher, M., Jin, L., Packman, L., Shlien, A., Campbell, P., Kirwan, M., Dokal, I. and Warren, A. (2012). Aberrant 3' oligoadenylation of spliceosomal U6 small nuclear RNA in poikiloderma with neutropenia. Blood, 121(6), pp.1028-1038.

Hofacker, I. (2003). Vienna RNA secondary structure server. Nucleic Acids Research, 31(13), pp.3429-3431.

Hohjoh, H. and Singer, M. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. The EMBO Journal, 15(3), pp.630-639.

Houseley, J. and Tollervey, D. (2010). Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase In Vitro. PLoS ONE, 5(8), p.e12271.

Huang, C., Schneider, A., Lu, Y., Niranjan, T., Shen, P., Robinson, M., Steranka, J., Valle, D., Civin, C., Wang, T., Wheelan, S., Ji, H., Boeke, J. and Burns, K. (2010). Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome. Cell, 141(7), pp.1171-1182.

Iskow, R., McCabe, M., Mills, R., Torene, S., Pittard, W., Neuwald, A., Van Meir, E., Vertino, P. and Devine, S. (2010). Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. Cell, 141(7), pp.1253-1261.

Januszyk, K., Li, P., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J., Martin, S. and Clubb, R. (2007). Identification and Solution Structure of a Highly Conserved C-terminal Domain within ORF1p Required for Retrotransposition of Long Interspersed Nuclear Element-1. Journal of Biological Chemistry, 282(34), pp.24893-24904.

Jurica MS, Moore MJ. (2003). Pre-mRNA splicing: awash in a sea of proteins. Molecular cell, 12(1):5–14.

Jurka, J., Kapitonov, V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research, 110(1-4), pp.462-467.

Kent, W. (2002). BLAT---The BLAST-Like Alignment Tool. Genome Research, 12(4), pp.656-664.

Kerpedjiev, P., Hammer, S. and Hofacker, I. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics, 31(20), pp.3377-3379.

Khan, H. (2005). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Research, 16(1), pp.78-87.

Khazina, E. and Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proceedings of the National Academy of Sciences, 106(3), pp.731-736.

Kim, D. and Salzberg, S. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biology, 12(8), p.R72.

Kolosha, V. and Martin, S. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proceedings of the National Academy of Sciences, 94(19), pp.10155-10160.

Kunkel, G., Maser, R., Calvet, J. and Pederson, T. (1986). U6 small nuclear RNA is transcribed by RNA polymerase III. Proceedings of the National Academy of Sciences, 83(22), pp.8575-8579.

Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. Nature, 409(6822), pp.860-921.

Landry, J., Pyl, P., Rausch, T., Zichner, T., Tekkedil, M., Stütz, A., Jauch, A., Aiyar, R., Pau, G., Delhomme, N., Gagneur, J., Korbel, J., Huber, W. and Steinmetz, L. (2013). The Genomic and Transcriptomic Landscape of a HeLa Cell Line. G3: Genes|Genomes|Genetics, 3(8), pp.1213-1224.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), pp.2078-2079.

Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]

Liu, X., Ma, B. and Wang, L. (2008). Voting algorithms for the motif finding problem. In: Proc LSS Comput Syst Bioinform Conf. [online] Stanford: Proc LSS Comput Syst Bioinform Conf, pp.Vol. 7, p. 37-47. Available at: http://www.lifesciencessociety.org/CSB2008/toc/37.2008.html.

Lund, E. and Dahlberg, J. (1992). Cyclic 2',3'-phosphates and nontemplated nucleotides at the 3' end of spliceosomal U6 small nuclear RNA's. Science, 255(5042), pp.327-330.

Macia, A., Munoz-Lopez, M., Cortes, J., Hastings, R., Morell, S., Lucena-Aguilar, G., Marchal, J., Badge, R. and Garcia-Perez, J. (2010). Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells. Molecular and Cellular Biology, 31(2), pp.300-316.

Macia, A., Widmann, T., Heras, S., Ayllon, V., Sanchez, L., Benkaddour-Boumzaouad, M., Muñoz-Lopez, M., Rubio, A., Amador-Cubero, S., Blanco-Jimenez, E., Garcia-Castro, J., Menendez, P., Ng, P., Muotri, A., Goodier, J. and Garcia-Perez, J. (2016). Engineered LINE-1 retrotransposition in nondividing human neurons. Genome Research, 27(3), pp.335-348.

Magoc, T. and Salzberg, S. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 27(21), pp.2957-2963.

Mailman, M., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J. and Sherry, S. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nature Genetics, 39(10), pp.1181-1186.

Martin, S. and Bushman, F. (2001). Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. Molecular and Cellular Biology, 21(2), pp.467-475.

Matera, A. and Wang, Z. (2014). A day in the life of the spliceosome. Nature Reviews Molecular Cell Biology, 15(2), pp.108-121.

Mathias, S., Scott, A., Kazazian, H., Boeke, J. and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. Science, 254(5039), pp.1808-1810.

Moran, J., Holmes, S., Naas, T., DeBerardinis, R., Boeke, J. and Kazazian, H. (1996). High Frequency Retrotransposition in Cultured Mammalian Cells. Cell, 87(5), pp.917-927.

Mroczek, S., Krwawicz, J., Kutner, J., Lazniewski, M., Kucinski, I., Ginalski, K. and Dziembowski, A. (2012). C16orf57, a gene mutated in poikiloderma with neutropenia, encodes a putative phosphodiesterase responsible for the U6 snRNA 3' end modification. Genes & Development, 26(17), pp.1911-1925.

Mroczek, S. and Dziembowski, A. (2013). U6 RNA biogenesis and disease association. Wiley Interdisciplinary Reviews: RNA, 4(5), pp.581-592.

Popow, J., Englert, M., Weitzer, S., Schleiffer, A., Mierzwa, B., Mechtler, K., Trowitzsch, S., Will, C., Luhrmann, R., Soll, D. and Martinez, J. (2011). HSPC117 Is the Essential Subunit of a Human tRNA Splicing Ligase Complex. Science, 331(6018), pp.760-764.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Löwer, J., Strätling, W., Löwer, R. and Schumann, G. (2011). The non-autonomous retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery. Nucleic Acids Research, 40(4), pp.1666-1683.

Richardson, S., Morell, S. and Faulkner, G. (2014). L1 Retrotransposons and Somatic Mosaicism in the Brain. Annual Review of Genetics, 48(1), pp.1-27.

Sassaman, D., Dombroski, B., Moran, J., Kimberland, M., Naas, T., DeBerardinis, R., Gabriel, A., Swergold, G. and Kazazian, H. (1997). Many human L1 elements are capable of retrotransposition. Nature Genetics, 16(1), pp.37-43.

Sawa, H. and Abelson, J. (1992). Evidence for a base-pairing interaction between U6 small nuclear RNA and 5' splice site during the splicing reaction in yeast. Proceedings of the National Academy of Sciences, 89(23), pp.11269-11273.

Scott, A., Schmeckpeper, B., Abdelrazik, M., Comey, C., O'Hara, B., Rossiter, J., Cooley, T., Heath, P., Smith, K. and Margolet, L. (1987). Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence. Genomics, 1(2), pp.113-125.

Scott, E. and Devine, S. (2017). The Role of Somatic L1 Retrotransposition in Human Cancers. Viruses, 9(6), p.131.

Shchepachev, V., Wischnewski, H., Missiaglia, E., Soneson, C. and Azzalin, C. (2012). Mpn1, Mutated in Poikiloderma with Neutropenia Protein 1, Is a Conserved 3′-to-5′ RNA Exonuclease Processing U6 Small Nuclear RNA. Cell Reports, 2(4), pp.855-865.

Smit, A. (1996). The origin of interspersed repeats in the human genome. Current Opinion in Genetics & Development, 6(6), pp.743-748.

Sontheimer, E. and Steitz, J. (1993). The U5 and U6 small nuclear RNAs as active site components of the spliceosome. Science, 262(5142), pp.1989-1996.

Staple, D. and Butcher, S. (2005). Pseudoknots: RNA Structures with Diverse Functions. PLoS Biology, 3(6), p.e213.

Svoboda, P. and Cara, A. (2006). Hairpin RNA: a secondary structure of primary importance. Cellular and Molecular Life Sciences, 63(7-8), pp.901-908.

Tanaka, N., Meineke, B. and Shuman, S. (2011). RtcB, a Novel RNA Ligase, Can Catalyze tRNA Splicing andHAC1mRNA Splicingin Vivo. Journal of Biological Chemistry, 286(35), pp.30253-30257.

Wahl, M., Will, C. and Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell, 136(4), pp.701-718.

Wei, W., Gilbert, N., Ooi, S., Lawler, J., Ostertag, E., Kazazian, H., Boeke, J. and Moran, J. (2001). Human L1 Retrotransposition: cis Preference versus trans Complementation. Molecular and Cellular Biology, 21(4), pp.1429-1439.

Westhof, E. (2015). Twenty years of RNA crystallography. RNA, 21(4), pp.486-487.

Will, C. and Luhrmann, R. (2010). Spliceosome Structure and Function. Cold Spring Harbor Perspectives in Biology, 3(7), pp.a003707-a003707.

Yean, S., Wuenschell, G., Termini, J. and Lin, R. (2000). Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. Nature, 408(6814), pp.881-884.

Chapter 4 Seekmer: Expression Quantification for both Bulk and Single-cell RNA Sequencing

This chapter presents a working draft of the manuscript “Seekmer recovers single-cell transcriptomic expression profiles through weighted pooling” in preparation with Hongjiu Zhang, Dr. Ryan E. Mills and Dr. Yuanfang Guan. Hongjiu Zhang developed the original algorithm design. I performed all the benchmark analysis using different data sets and algorithm refinement based on the benchmark result.

Introduction

RNA sequencing (RNA-seq) technologies and RNA quantification methods

The DNA makeup of individual cells in a human body remains mostly the same, with the exception of somatic variation. However, RNA is highly dynamic and can vary in different cell types. The human transcriptome includes all transcripts in a cell as well as their quantity (Mortazavi et al., 2008; Wang, Gerstein and Snyder, 2009). The variability of a transcriptome comes from differences in expressed genes, differences in specific isoforms that are spliced from those genes, as well as differences in the quantity of transcripts expressed per gene. Thus, RNA sequencing based on Illumina sequencing of reverse transcribed cDNA library

from cells can represent the features of transcriptome for certain cells and tissues.

Bioinformatics tools for quantifying gene/transcript expression have been developed for decades. In general, there are two major kinds of RNA-seq quantifications tools. RNA-seq quantification was initially developed to quantify isoform expression based on read alignment, for example, TopHat (Trapnell, Pachter and Salzberg, 2009) + Cufflinks (Trapnell et al., 2012), STAR (Dobin et al, 2012) + RSEM (Li and Dewey, 2011). These alignment-based tools are relatively accurate to quantify isoforms in transcriptome (Teng et al., 2016), but take a long time and large memory footprint to perform due to the time and memory needed for read alignment. In recent years, multiple alignment-free RNA-seq quantification tools have been developed, including Kallisto (Bray et al. 2016), Sailfish (Patro, Mount and Kingsford, 2014) and Salmon (Patro et al., 2017). For alignment free methods, a pseudo-alignment is performed using the RNA-seq reads. The pseudo-counts for each isoform are then used as input for further quantification. Without alignment, these alignment-free methods have shortened the running time significantly compared to alignment-based methods (Patro, Mount and Kingsford, 2014; Bray et al. 2016; Patro et al., 2017). However, because of the relative lower accuracy of read assignment, alignment-free methods are slightly less accurate compared to alignment-based methods (Teng et al., 2016).

Single cell RNA-seq technologies and RNA quantification methods

Bulk RNA-seq is performed with RNA extracted from large population of cells as discussed in Chapter 2. However, the results from a bulk RNA-seq analysis will only represent the average isoform profile of thousands to millions of cells, and as such the quantification of the cell-to-cell difference is impossible to ascertain. In 2009, the first protocol of single cell RNA-seq (scRNA-seq) was published (Tang et al., 2009). Multiple technologies for scRNA-seq has been developed since then. There are a few major types of scRNA-seq technologies widely performed, including SMART-seq/SMART-seq2 (Ramsköld et al., 2012; Picelli et al., 2014), and UMI-tag based approaches (Islam et al., 2011; Hashimshony et al., 2012; Macosko et al., 2015). While UMI-barcoded can improve the accuracy by removing PCR bias using barcodes, these methods can only sequence the 5' or 3' end of the transcripts (Islam et al., 2011; Hashimshony et al, 2012). Thus, SMART-seq2 (Picelli et al., 2014) with full coverage on each transcript fits better for the purpose of isoform quantification and allele specific gene expression in single cell transcriptome.

Although experimental methods for scRNA-seq have been developed to capture the dynamics of each single cell transcriptome, the bioinformatics tools for analyzing scRNA-seq data remains limited and underdeveloped (Hwang, Lee and Bang, 2018). The random dropout of many genes (Wagner, Regev and Yosef, 2016) and biased amplification of certain genes (Bacher and Kendziorski, 2016) are two major challenges for most of scRNA-seq analysis tools.

Furthermore, with a relatively small amount of reads in scRNA-seq data compared to bulk RNA-seq data, the normalization models applied to bulk RNA-seq data do not perform well in scRNA-seq.

Here we developed Seekmer, an RNA-seq quantification tool that combines the advantages of both alignment-based and alignment-free methods to quantify isoform expression level in both bulk and single cell RNA-seq data. We further present an imputation method designed for scRNA-seq to better quantify isoform expression in a smaller total number reads scRNA-seq library by cell pooling based on the initial isoform expression quantification.

Method and Materials

Seekmer Algorithm

Seekmer consists of two parts, an alignment-free mapper and an abundance estimator (Figure 4.1 a). The alignment-free mapper is similar to existing alignment-free transcript quantification tools such as Kallisto (Bray et al. 2016). Seekmer mapper generates an index of reference transcriptomic sequences. To build an index, Seekmer first collects all possible *k*-mers from reference transcriptomic sequences. Each *k*-mer is associated with a set of transcripts to which the *k*-mer can be mapped. Due to the high similarities between sequences of splicing isoforms from same genes, many *k*-mers can be mapped to more than one transcript. Seekmer groups together *k*-mers that are contiguous on the transcript sequences and share same sets of mappable transcripts. These

grouped *k*-mers form contigs, as they occupy continuous segment of transcript sequences. Each contig has two terminal *k*-mers (and their reverse complements). These grouped *k*-mers, together with their sets of mappable transcripts, are the index for later mapping process.

When mapping a read, the mapper looks for a *k*-mer in the read that is also present in the Seekmer index. If a read has no such *k*-mer, the read is discarded. Starting from the matched *k*-mer, the mapper keeps extending the match by iteratively jumping over the terminal *k*-mers of the contigs and mapping *k*-mers. By jumping over the contig, the mapper skips over *k*-mers that have the same mappable target transcripts, both speeding up the process and avoiding potential sequencing errors or mutations in the middle of the contig. The output of the mapper for each cell is a long list of transcript sets and how many reads can be mapped to these sets. These data are the input for the abundance estimator.

The abundance estimator of Seekmer takes the read mapping data and performs an initial estimation the abundance of genes. The initial estimation optimizes a uniform mixture model. The log-likelihood function is expressed as:

$$\log L = \sum_{i=1}^{N_{\text{read}}} \log \sum_{j=1}^{N_{\text{transcript}}} \frac{\alpha_j}{l_j},$$

where i, j is the effective length of the j-th transcript and is the abundance of the j-th transcript. The estimator estimates the gene abundance of each cell by adding the abundance of all transcripts from same genes.

Based on the initial estimation, the abundance estimator then calculates the Pearson correlation coefficients of gene expression for all pairs of cells as

$$\bar{g}_i = \frac{1}{N_{\text{gene}}} \sum_{k=1}^{N_{\text{gene}}} g_{ik}$$

$$r_{ij} = \frac{\sum_{k=1}^{N_{\text{gene}}} \left(g_{ik} - \hat{g}_i \right) \left(g_{jk} - \hat{g}_j \right)}{\sqrt{\sum_{k=1}^{N_{\text{gene}}} \left(g_{ik} - \hat{g}_i \right)^2} \sqrt{\sum_{k=1}^{N_{\text{gene}}} \left(g_{jk} - \hat{g}_j \right)^2}}$$

where g_{ik} is the expression level of k-th gene in i-th cell, and r_{ij} is the Pearson correlation coefficient between the gene expression profile of i-th cell and j-th cell. The estimator then applies K-mean clustering (K = 2) on the coefficients in the matrix and zeros out (Figure 4.1 b). The cluster of the lower coefficients in the correlation matrix is zero out. Seekmer then raise all elements in the processed correlation matrix to higher power to get the weight matrix W for all cells. Given a target cell to impute, Seekmer build a model similarly to the uniform expression cells to the same number of the target cell. Then the read counts are multiplied by the pre-calculated weights. By optimizing the likelihood function of the mixture model, Seekmer estimates the transcript-level abundance of the cells.

Real-world RNA sequencing data

Public available real RNA sequencing data in this work includes: 1) bulk RNA sequencing data of two uniformly used samples, Universal Human Reference RNA (UHRR) and Human Brain Reference RNA (HBRR) (SEQC/MAQC-III

Consortium et al. 2014.); 2) SMRT-seq sequencing of mouse embryonic cells with ERCC and SIRV spike-ins (ID: E-MTAB-5485) (Svensson et al. 2017). The quantitative reverse transcription polymerase chain reaction (qRT-PCR) assay results for UHRR and HBRR were generated from SEQC project as a 'gold standard' for RNA-seq quantification (GEO Accession: GSM1361812, GSM1361813) (SEQC/MAQC-III Consortium et al. 2014).

Simulated sequencing data

All simulated sequencing data in this work were generated using RSEM v1.2.28 (coupled with STAR v2.7.0e aligner). STAR and RSEM aligned and estimated the transcript abundance of a UHRR sample (SRR950078) and a HBRR sample (SRR950079) against ENSEMBL 90 human reference cDNA sequences. RSEM simulator then took the quantification results as the expression level for each transcript and simulated 48 RNA sequencing libraries based on the UHRR sample and 48 based on the HBRR sample. The total number of reads for the 96 RNA-seq libraries is randomly selected from a normal distribution.

Benchmarks

The UHRR, HBRR, simulated, and Fluidigm Polaris samples were mapped against ENSEMBL 90 human reference cDNA sequences. The mouse embryonic stem cells with spike-ins were mapped against ENSEMBL 90 mouse reference cDNA sequences.

The quantification and imputation results of simulated 96-cell dataset, the mouse embryonic stem cell dataset, and the Fluidigm Polaris dataset were also analyzed using Seurat.

Results

Seekmer fills the gap of alignment-free and alignment-based RNA quantification methods in real bulk RNA sequencing benchmark by more accurate pseudo-alignment

We performed Kallisto (Bray et al. 2016), Sailfish (Patro, Mount and Kingsford, 2014), Salmon (Patro et al., 2017), STAR+RSEM (Li and Dewey, 2011; Dobin et al., 2012) and Seekmer on the two uniformly used samples, Universal Human Reference RNA (UHRR) and Human Brain Reference RNA (HBRR) (SEQC/MAQC-III Consortium et al. 2014). Seekmer performs slightly better than all other alignment-free methods and have a similar performance with alignment-based method (STAR+RSEM) in both UHRR and HBRR libraries (Figure 4.2; Figure 4.3; Table 4.1; Table 4.2). With a similar performance of Seekmer with STAR+RSEM, Seekmer is on average 20 times faster than STAR+RSEM (Table 4.3) and is slightly slower than the other three alignment-free methods.

The local alignment applied in non-matching *k*-mers increased the accuracy of read alignment for Seekmer. The total number of mapped reads from Seekmer was observed to be consistent with those mapped using STAR (Figure 4.4 a). In order to test whether the aligned reads in Seekmer were accurately mapping to

the correct positions in the genome, we further generated a set of negative control “pseudo” reads with 0-50 base pair mapped to a position in the reference and the remaining portion as non-reference randomly generated sequence. We performed all five methods on the negative control data. Only Seekmer and STAR+RSEM did not align any of these pseudo-reads to the reference, while all three alignment-free methods aligned 48% reads from the negative control set (Figure 4.4 b).

Seekmer single cell imputation improves single cell RNA sequencing quantification in simulated single cell RNA sequencing data

With fewer reads in the single cell RNA sequencing libraries, we tested the performance of Seekmer imputation on single cell data using a set of simulated libraries, each with a different total number of reads. We observed that the performance of Seekmer drops when the total number of reads falls under 1,000,000 (Figure 4.5 a, b, c).

We also tested the log-Pearson and Spearman correlation of Seekmer imputation quantification in simulated single cell RNA data with known gene and transcript level expression levels as well as qRT-PCR expression of 20801 genes. With imputation, Seekmer performs much better on both higher mean log-Pearson and Spearman correlation as well as exhibiting smaller standard error among different libraries.

We then also tested different power (exponent) levels of the weight matrix used for single cell pooling in the simulated single cell data. From both log-Pearson and Spearman correlation, power=16 performed the best with a higher mean compared to power 1-12 and with a smaller standard error compared to power 32. Thus, we chose power=16 for future Seekmer imputation analysis (Figure 4.7; Table 4.4). Furthermore, the performance of Seekmer imputation is not affected by different ratio of cells in different clusters (Figure 4.9; Figure 4.10).

Both PCA and tSNE plots show a distinct clustering of the two types of simulated cells (Figure 4.8 a, b). The cells further away from the center of clusters show a clear fewer number of reads (Figure 4.8 c).

Seekmer single cell imputation improves single cell RNA sequencing quantification in SIRV spike-in RNA sequencing data

SIRV is a spike-in isoform test set that is used for quantification assessment. We were able to thus assess the accuracy of Seekmer imputation against real RNA sequencing data with SIRV. We observed a clear clustering of cells after imputation compared to raw quantification results (Figure 4.11; Figure 4.12). Both log-Pearson and Spearman correlation between Seekmer quantification result of SIRV transcripts and original spike-in amount increased after imputation compared to the result before imputation (Figure 4.13), representing a huge amount of performance improvement after imputation performed.

Seekmer single cell imputation correctly identified gene markers for different cell types in Fluidigm Polaris sequencing on K562 and SUM149

As a case study, we applied Seekmer to a Fluidigm Polaris sequencing data on two cell lines. One is the leukemia cell line K562, and the other is the triple-negative breast cancer cell line SUM149. Both cell lines are well studied and have many splicing isoforms have been characterized.

K562 cells clustered distinctly from SUM149 cells, with some outliers of cells containing fewer total number of reads (Figure 4.14; Figure 4.15). We observed both transcripts of CD44 markers and both transcripts of EGFR markers better enriched in SUM149 after imputation, as well as a better enrichment of GYPA and GYPB in K562 after imputation. (Figure 4.16; Figure 4.17; Figure 4.18)

Discussion

We have developed Seekmer to perform RNA quantification for both bulk RNA sequencing and single cell RNA sequencing with imputation. We have shown that Seekmer has similar performance with alignment-based method, STAR+RSEM, with a significantly shorter run time. We also showed that with imputation, Seekmer could significantly improve the performance of RNA quantification from simulated single cell RNA sequencing data and SIRV spike-in data. We also showed a case study of RNA sequencing on K562 and SUM149 with Seekmer imputation and presented the correct enrichment of certain marker genes previously known in both of the cell lines.

While Seekmer has been shown to perform well in these scenarios, there are a few limitations. One limitation is that Seekmer calculates the similarity between cells based on read counts of all genes. Considering most platforms can detect only a few thousand genes in each cell, more than half of the elements in gene expression vectors of cells are zeros. Due to the majority of zeros in the expression vectors in every cell, the expression profile of cells tends to be similar, and it is difficult for Seekmer to cluster the similar cells. Also, for studies involving cells from the same tissues or organs, there may be many genes with similar expression levels. This may lead to high correlations between cells and cause Seekmer to be insensitive to minor differences between them. However, we did not observe this in our analysis. A potential reason for the absence may be our 2-way clustering and our use of higher power of correlation coefficients. This issue may be alleviated by using the most variable genes instead of all genes to calculate the correlation matrix.

Another limitation is that the current implementation of Seekmer is not scalable to handle cases where are many thousands of cells. Inference for more than a few thousand cells may take many days to complete. This is currently acceptable as major full-length single-cell sequencing platforms accept at most a few hundred cells per batch but may be limiting in the future as technology continues to advance. At the same time, this can be easily improved in algorithm implementation and future hardware upgrades.

Conclusion

This work presents Seekmer, an imputation method that estimates single-cell transcript abundance through read pooling. The weighted pooling approach enables estimating the abundance of the transcripts per cell much more accurately. The algorithm is able to differentiate splicing isoforms from same genes in the imputation process and performs well even for limited numbers of cells. In sum, Seekmer provides more accurate transcript profiling analysis and empowers researchers in single-cell splicing studies.

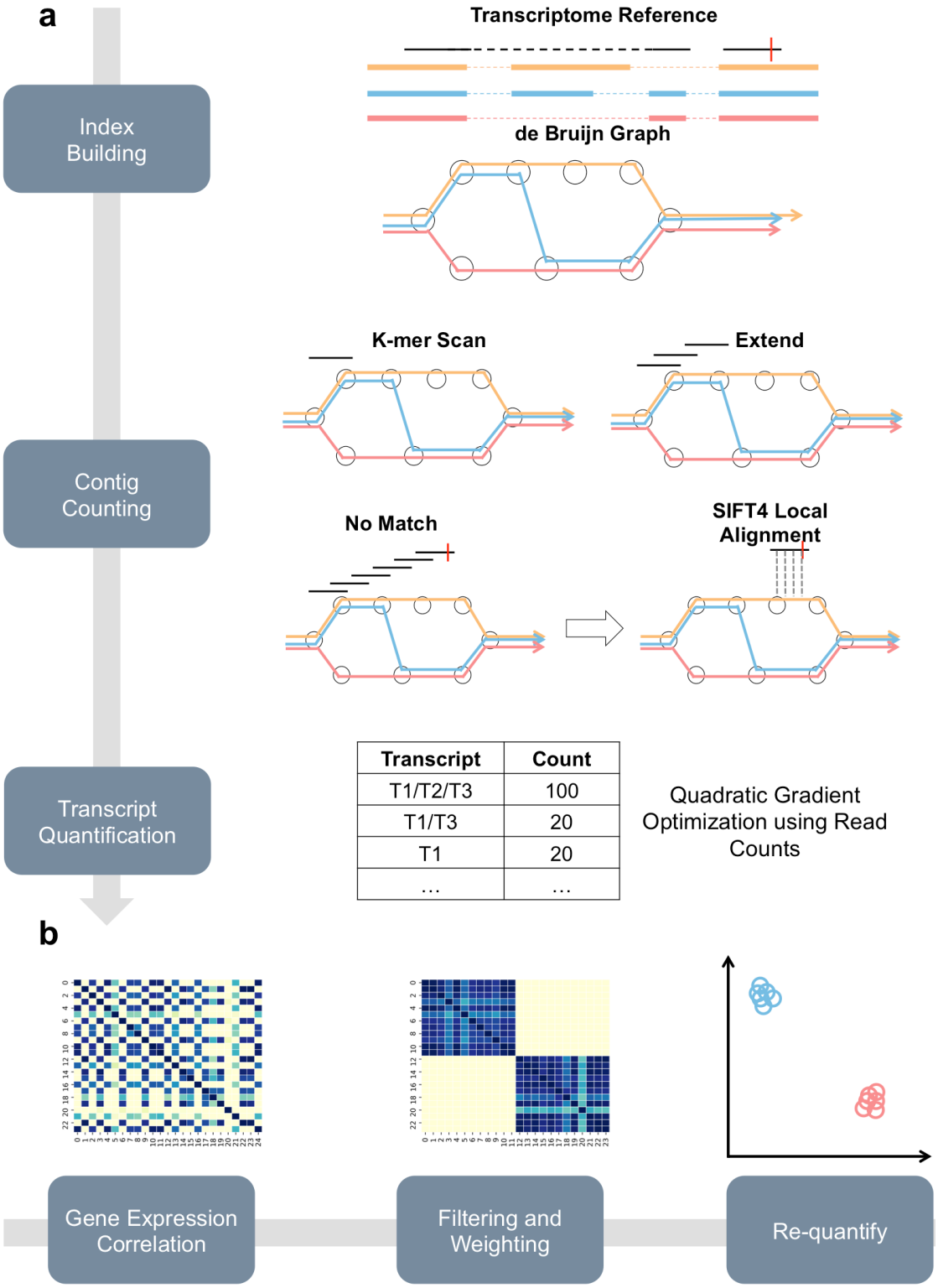

a
Index Building
Contig Counting
Transcript Quantification
Transcriptome Reference
de Bruijn Graph
K-mer Scan
Extend
No Match
SIFT4 Local Alignment
Transcript
Count
T1/T2/T3
100
T1/T3
20
T1
20
…
…
Quadratic Gradient Optimization using Read Counts
b
Gene Expression Correlation
Filtering and Weighting
Re-quantify

Figure 4.1: Seekmer RNA quantification method and single cell pooling strategy.
a) Seekmer RNA sequencing pseudo-alignment and quantification. Seekmer RNA quantification tool looks for *k*-mers in reads in the pre-built de Bruijn graph of index for transcriptome. Once one *k*-mer is matched with reference, the *k*-mer search would be extended along the reference. If there is no exact match when extending *k*-mer search, a local alignment (SIFT alignment) would be performed to ensure if the reason for non-match is sequencing error or SNP in certain samples. After counts for each transcript, a quadratic gradient optimization is performed to get the final expression level of each transcript.
b) Seekmer single cell pooling and imputation. The RNA quantification result for each single cell is used to calculate the correlation between the single cells. The cells are then pooled based on the correlation. Seekmer imputation then re-quantifies the expression level of each single cell based on the most correlated single cells.

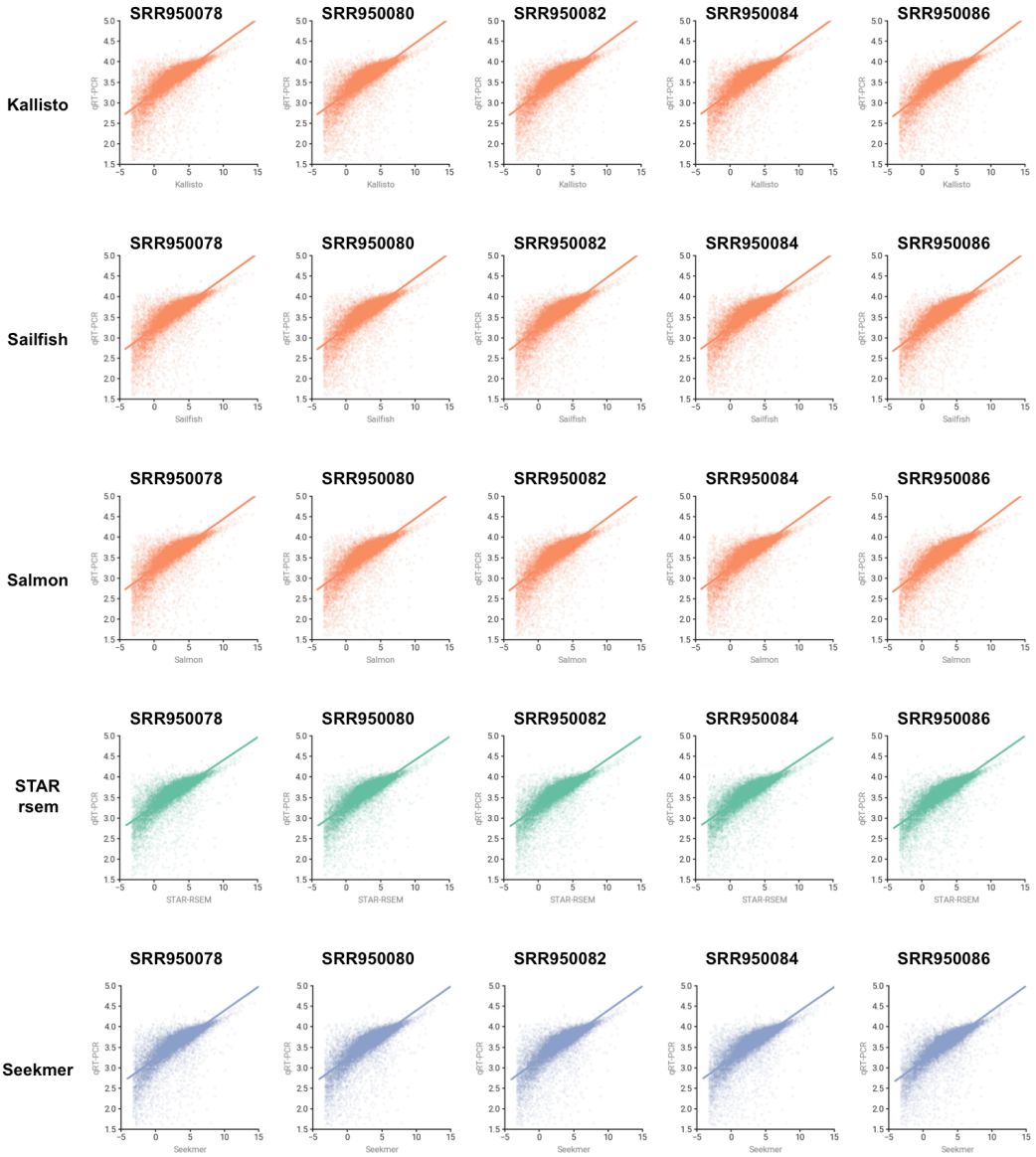


Figure 4.2: Seekmer RNA quantification performance on bulk RNA sequencing compared to other methods (UHRR).
Correlation of the TPM calculated by different algorithms (Kallisto, Sailfish, Salmon, STAR-RSEM and Seekmer) with the qRT-PCR quantification of 20801genes in five UHRR RNA sequencing libraries.

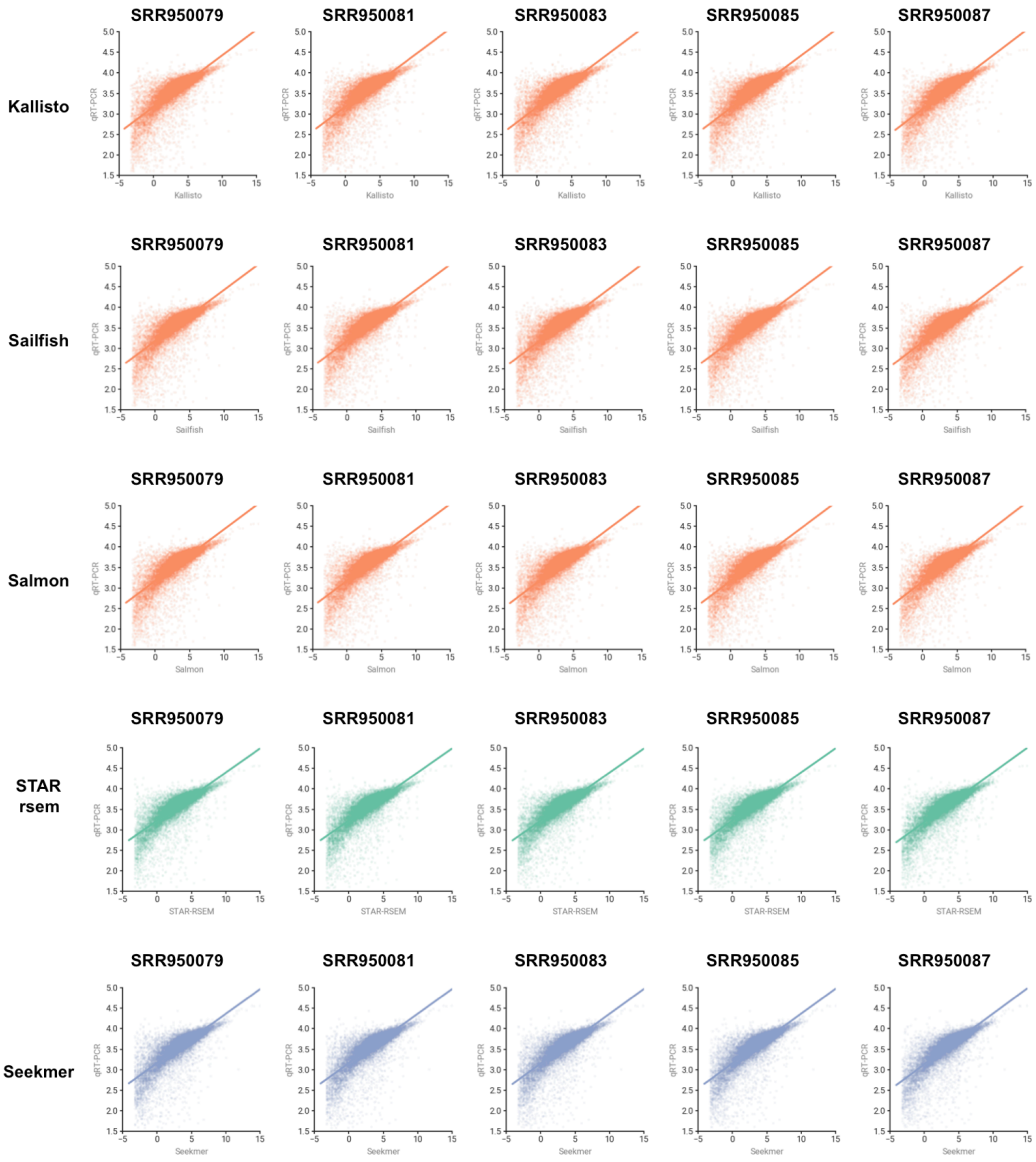


Figure 4.3: Seekmer RNA quantification performance on bulk RNA sequencing compared to other methods (HBRR)
Correlation of the TPM calculated by different algorithms (Kallisto, Sailfish, Salmon, STAR+RSEM and Seekmer) with the qRT-PCR quantification of 20801genes in five HBRR RNA sequencing libraries.

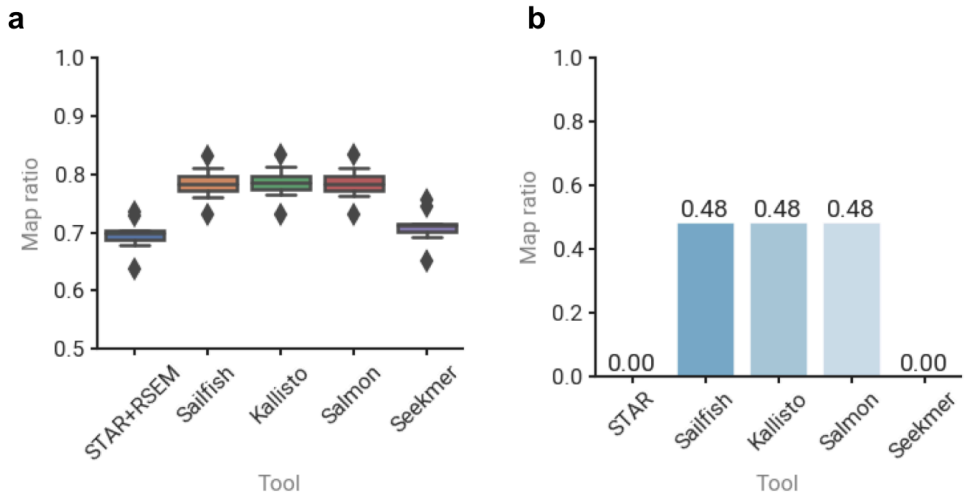


Figure 4.4: Seekmer RNA quantification performance on bulk RNA sequencing data (UHRR and HBRR) map ratio compared to other alignment free methods and alignment based methods.
a) Percentage of reads mapped in bulk RNA sequencing data. The percentage of reads mapped to reference from 10 different bulk RNA sequencing libraries in different methods including both alignment-free and alignment-based methods. Seekmer has a closer mapped rate of reads compared to STAR-RSEM which is the alignment-based methods with more alignment accuracy.
b) *Percentage of reads mapped in simulated negative control*. With negative control where only less than half of the reads contain the *k*-mer in genome, all alignment-free methods would still align 48% reads, while Seekmer performs as STAR, which did not align any simulated negative control reads.

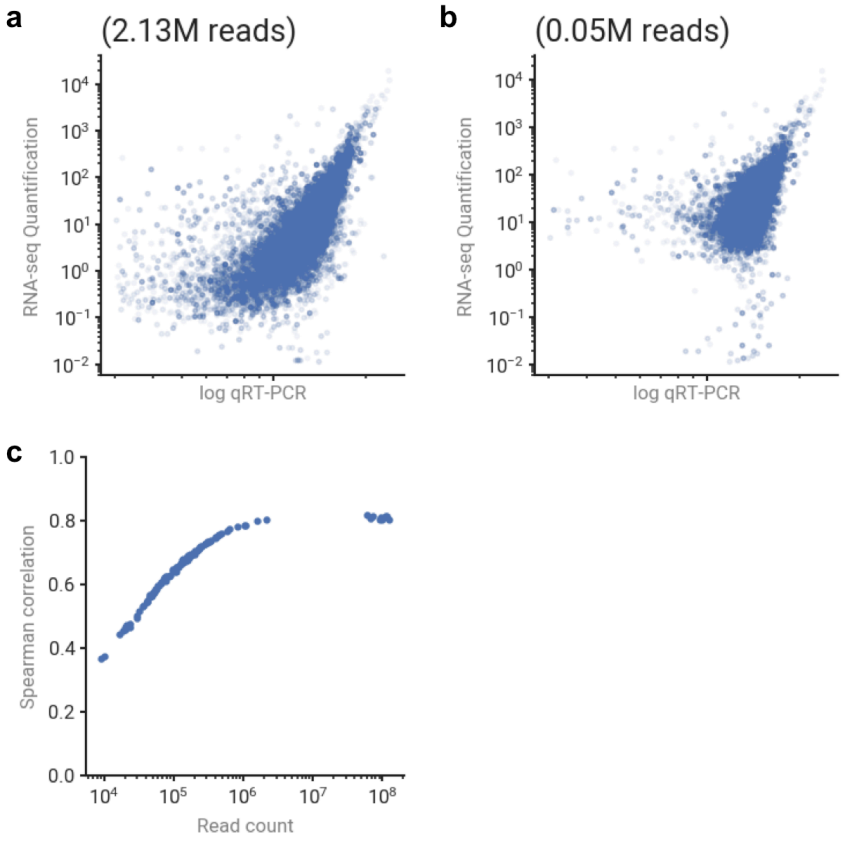


Figure 4.5: Total number of reads and Seekmer RNA quantification performance on simulated single cell RNA sequencing data.
a) Seekmer RNA quantification result compared with qRT-PCR result with 2.13M reads in the simulated library. The quantification of single cell RNA sequencing library with relatively high number of reads (2.13M) has a high correlation with the qRT-PCR result.
b) Seekmer RNA quantification result compared with qRT-PCR result with 0.05M reads in the simulated library. The quantification of single cell RNA sequencing library with relatively low number of reads (0.05M) has a lower correlation with the qRT-PCR result.
c) Spearman correlation of Seekmer quantification result with qRT-PCR result from simulation. Spearman correlation increases when the number of reads in single cell RNA-seq library increases. It reaches a flat curve after ~10^6 total reads in the library.

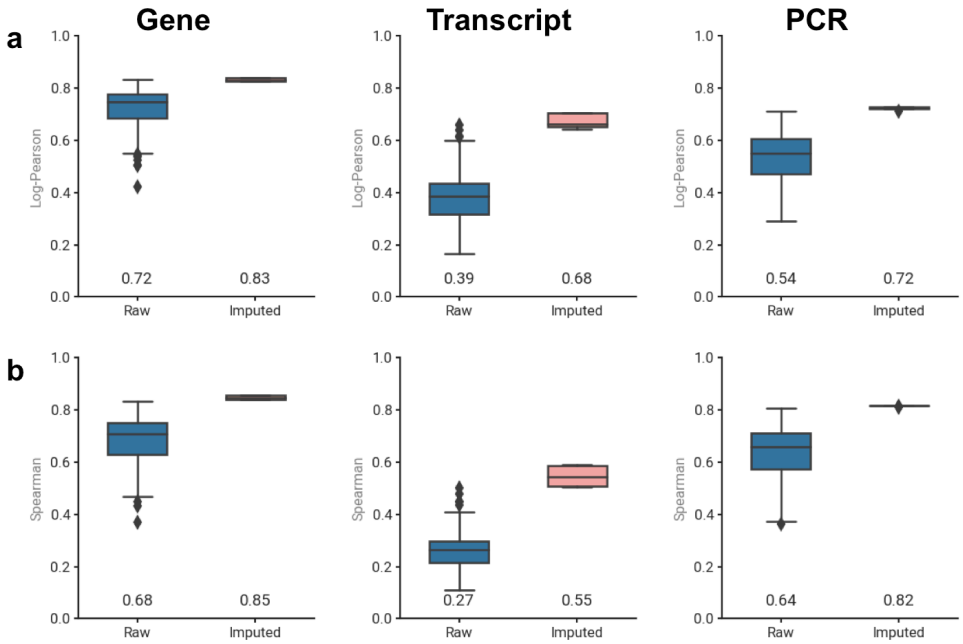


Figure 4.6: Seekmer RNA quantification performance on simulated single cell RNA sequencing data.
a) Log-Pearson correlation of Seekmer quantification result with different true answers before and after imputation. The imputed log-Pearson correlation of Seekmer quantification result after imputation with gene level true answer, with transcirpt level true answer and qRT-PCR quantification result is all better (higher mean and smaller standrad deviation) than the log-Pearson correlation of Seekmer quantification result before imputation.
b) Spearman correlation of Seekmer quantification result with different true answers before and after imputation. The imputed *Spearman* correlation of Seekmer quantification result after imputation with gene level true answer, with transcirpt level true answer and qRT-PCR quantification result is all better (higher mean and smaller standrad deviation) than the log-Pearson correlation of Seekmer quantification result before imputation.

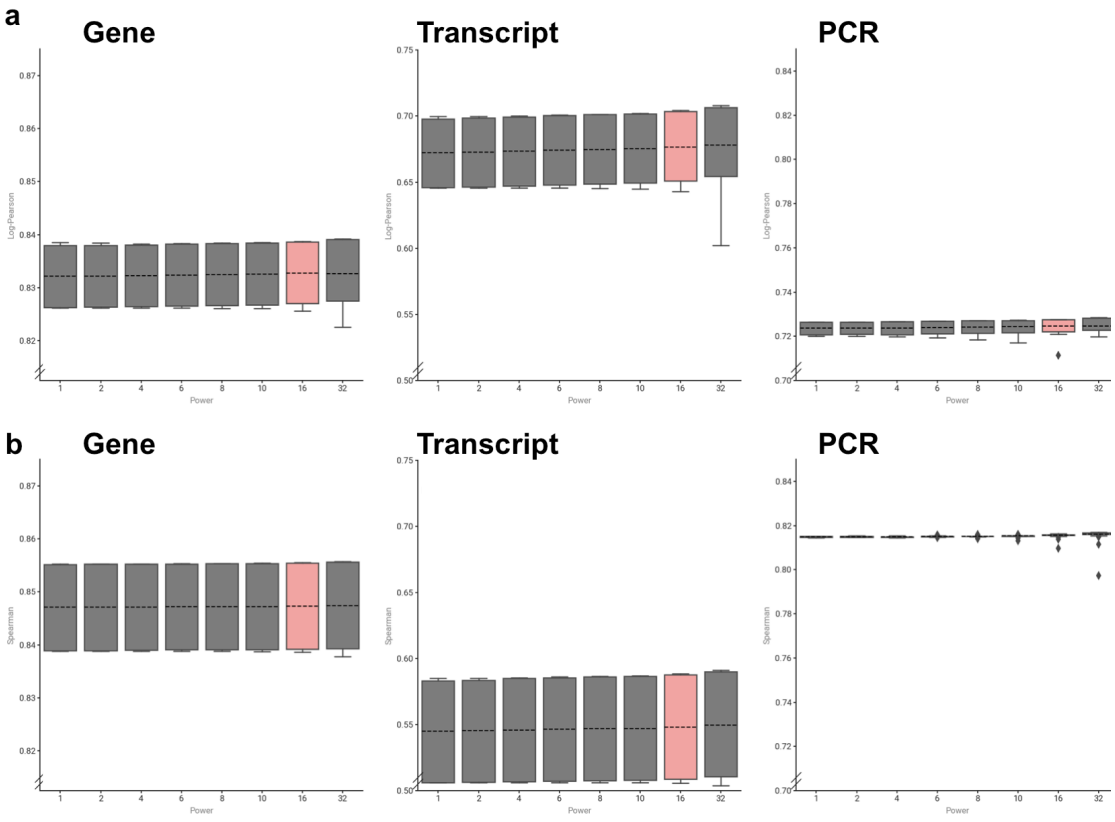


Figure 4.7: Seekmer RNA quantification performance on simulated single cell RNA sequencing data with different power.
a) Log-Pearson correlation of Seekmer quantification result with different true answers with different power of imputation. The log-Pearson correlation of Seekmer quantification result with gene level true answer, with transcirpt level true answer and qRT-PCR quantification result is best with power=16, with the higher mean as well as a lower standard deviation.
b) Spearman correlation of Seekmer quantification result with different true answers with different power of imputation. The Spearman correlation of Seekmer quantification result with gene level true answer, with transcirpt level true answer and qRT-PCR quantification result is best with power=16, with the higher mean as well as a lower standard deviation.

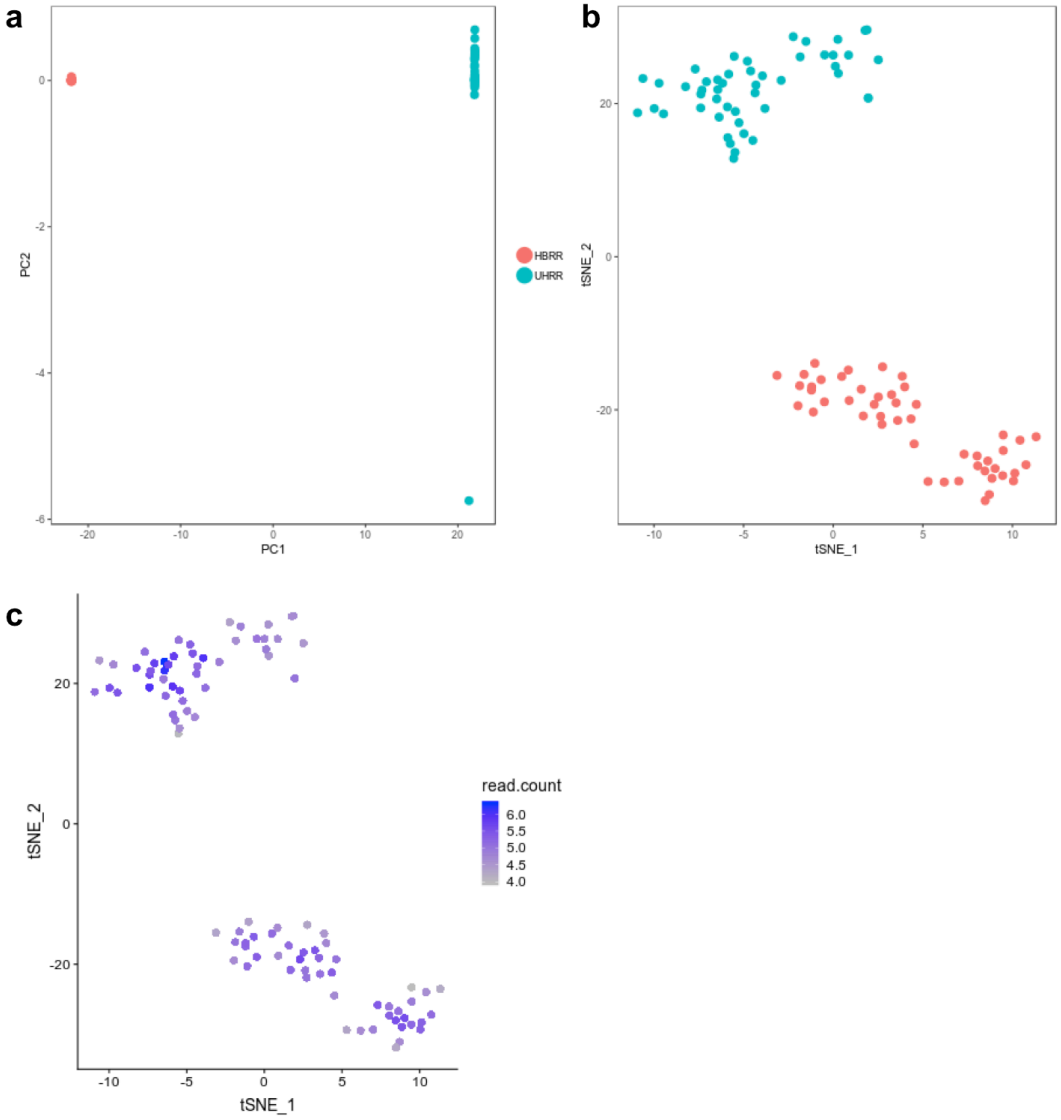
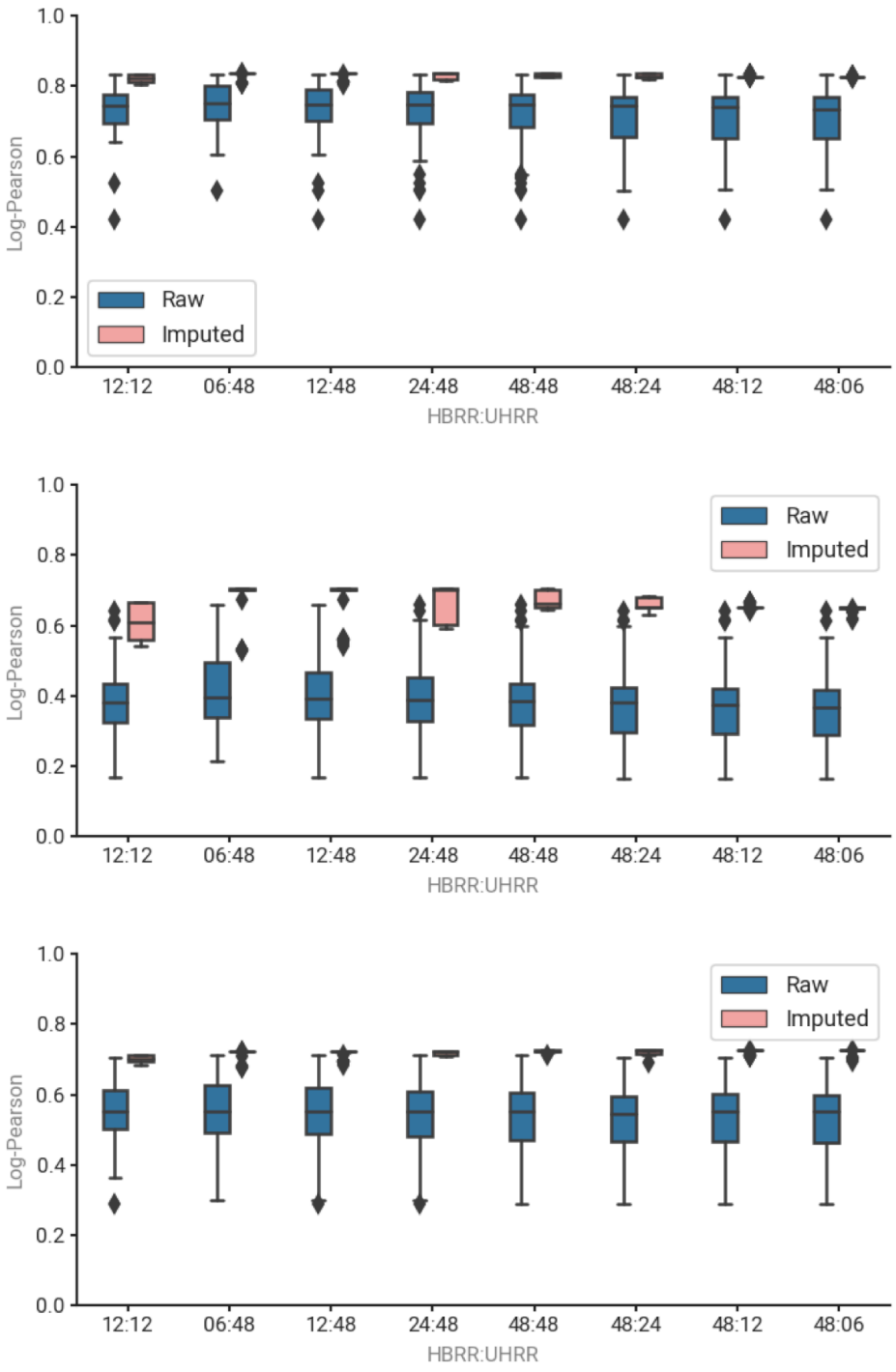

a
PC2
PC1
HBRR
UHRR
b
tSNE_2
tSNE_1
c
tSNE_2
tSNE_1
read.count
6.0
5.5
5.0
4.5
4.0

Figure 4.8: Clustering after imputation on simulated single cell RNA sequencing data.
a) PCA plot of simulated single cells after imputation. PCA plot of simulated single cell RNA sequencing data from UHRR and HBRR of different gene expression profiles. Except for a few samples, all UHRR and HBRR samples are cluestered together after imputation.
b) tSNE plot of simulated single cells after imputation. tSNE plot of simulated single cell RNA sequencing data from UHRR and HBRR of different gene expression profiles. UHRR and HBRR samples clustered into two groups, but there are a few samples with a few of samples furhter away from the center of each cluster.
c) tSNE plot of simulated single cells after imputation colored with read count. tSNE plot of simulated single cell RNA sequencing data from UHRR and HBRR of different gene expression profiles colored with read count of each single cell RNA sequencing library. The two further away small clusters from the center of the two major clusters both have much fewer reads compared to the other samples closer to the center of the two clusters.

a


1.0
0.8
0.6
0.4
0.2
0.0
Log-Pearson
Raw
Imputed
12:12
06:48
12:48
24:48
48:48
48:24
48:12
48:06
HBRR:UHRR

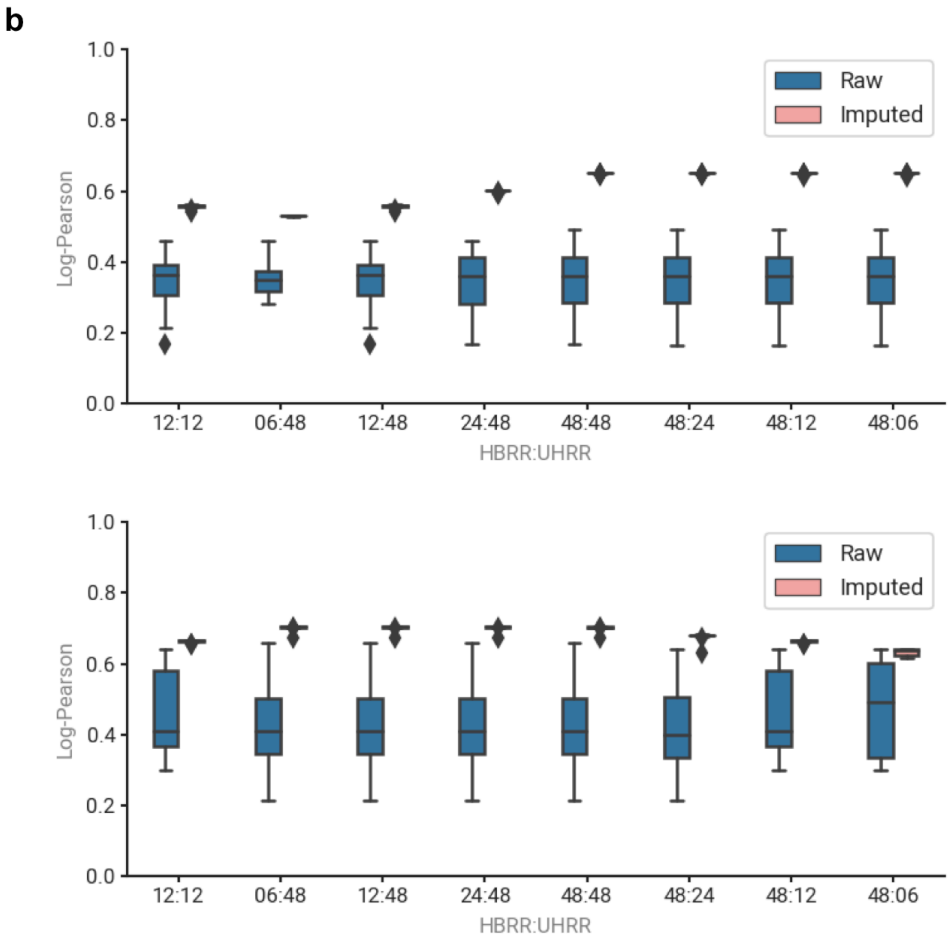
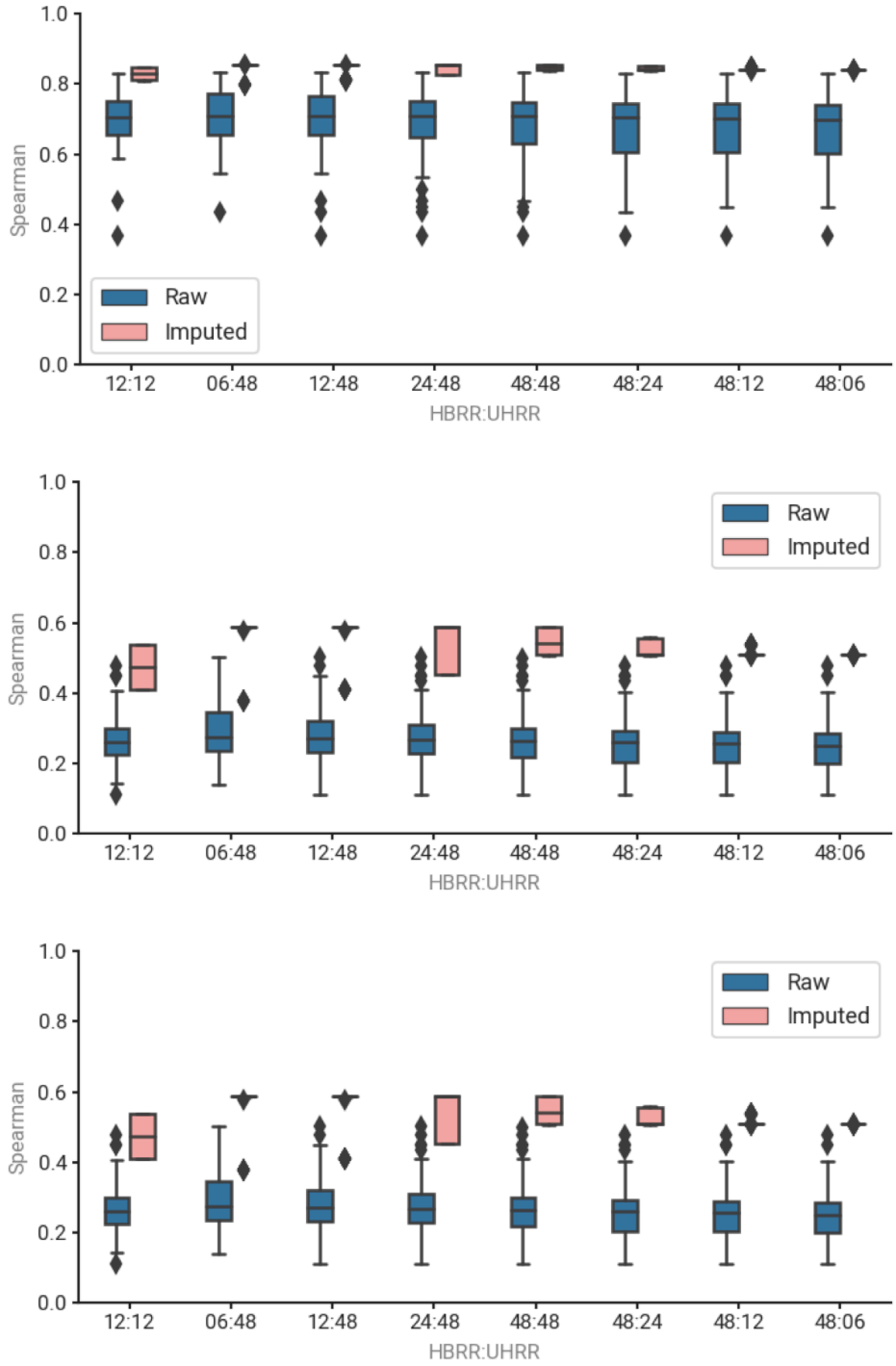


Figure 4.9: Performance (Log-Pearson) of Seekmer on different ratio of cells with simulated single cell data.
a) Log-Pearson correlation of Seekmer quantification result for all cells with gene level true answer (upper), with transcript level true answer (middle) and qRT-PCR quantification result (lower) before and after imputation with different ratio of numbers of single cells from two classes. Different ratio of cells from different clusters did not significantly affect the performance of Seekmer from log-Pearson correlation.
b) Log-Pearson correlation of Seekmer quantification result for HBRR cells (upper) and UHRR cells (lower) with transcript level true answer before and after imputation with different ratio of numbers of single cells from two classes. Different ratio of cells from different clusters did not significantly affect the performance of Seekmer from log-Pearson correlation in the two clusters of cells.

a


1.0
0.8
0.6
0.4
0.2
0.0
Spearman
Raw
Imputed
12:12
06:48
12:48
24:48
48:48
48:24
48:12
48:06
HBRR:UHRR

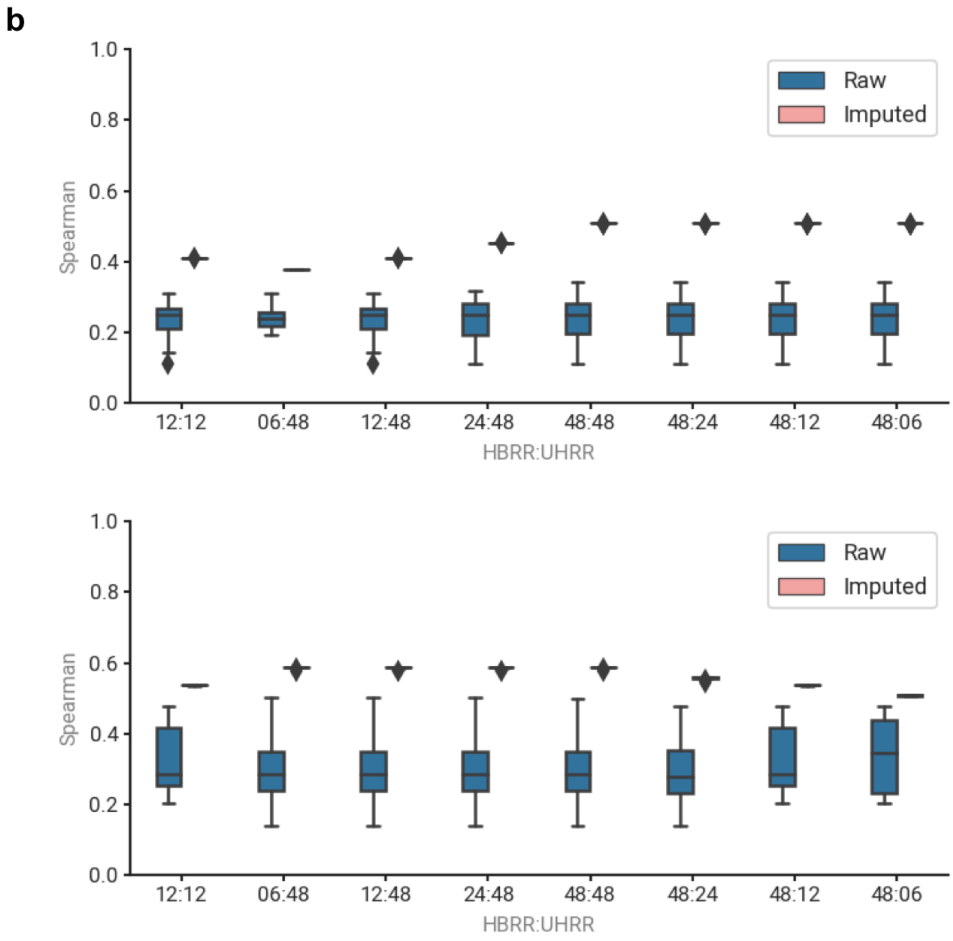


Figure 4.10: Performance (Spearman) of Seekmer on different ratio of cells with simulated single cell data.
a) Spearman correlation of Seekmer quantification result for all cells with gene level true answer (upper), with transcript level true answer (middle) and qRT-PCR quantification result (lower) before and after imputation with different ratio of numbers of single cells from two classes. Different ratio of cells from different clusters did not significantly affect the performance of Seekmer from Spearman correlation.
b) Spearman correlation of Seekmer quantification result for HBRR cells (upper) and UHRR cells (lower) with transcript level true answer before and after imputation with different ratio of numbers of single cells from two classes. Different ratio of cells from different clusters did not significantly affect the performance of Seekmer from Spearman correlation in the two clusters of cells.

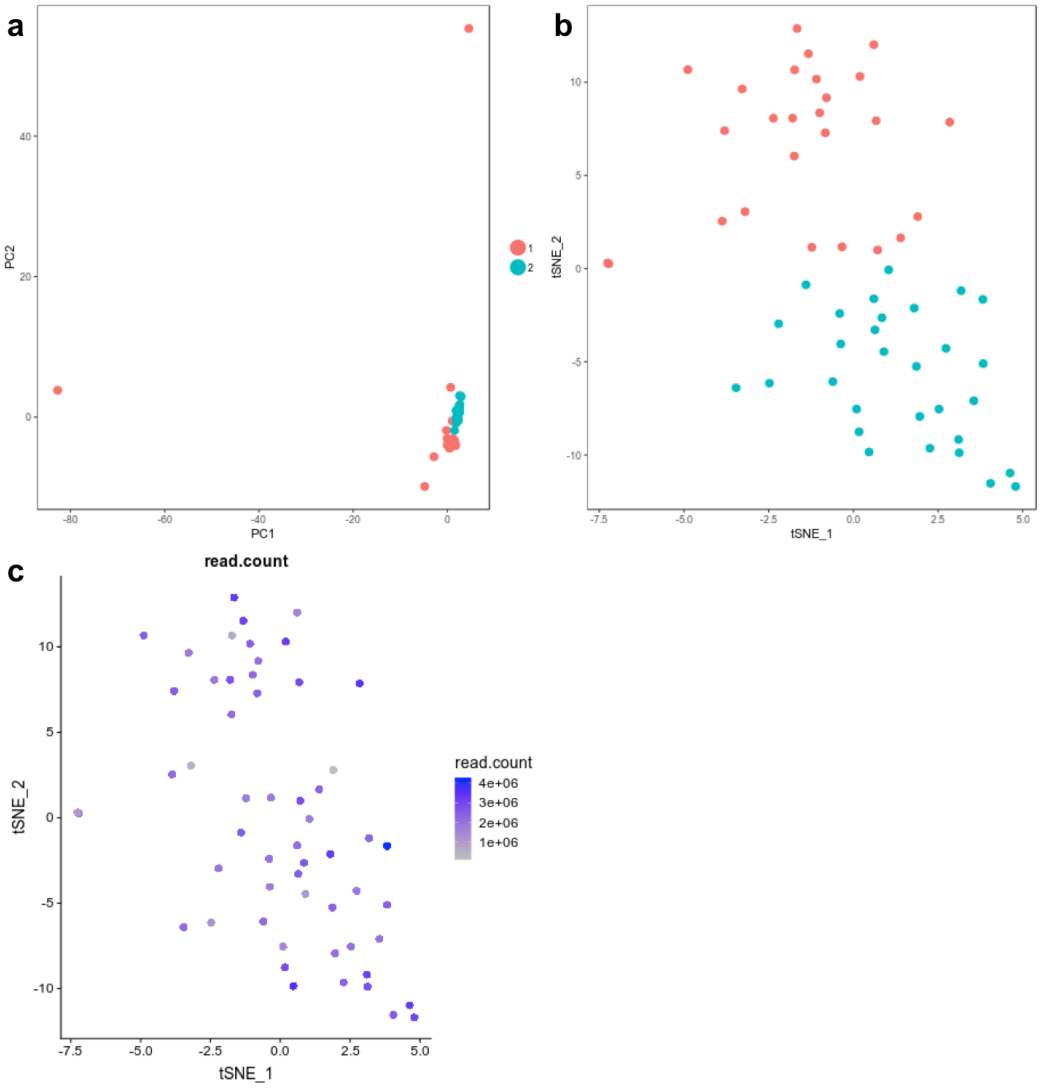

a
PC2
PC1
1
2
b
tSNE_2
tSNE_1
c
read.count
tSNE_2
tSNE_1
read.count
4e+06
3e+06
2e+06
1e+06

Figure 4.11: Clustering of single cells before imputation in SIRV spike-in single cell data.
a) PCA plot of SIRV spike-in single cell data before imputation. In PCA plot of SIRV single cell RNA sequencing data, all single cells are clustered togrther before imputation with the exception of a few single cells.
b) tSNE plot of SIRV spike-in single cell data before imputation. In tSNE plot of of SIRV single cell RNA sequencing data, samples do not cluster well before imputation.
c) tSNE plot of SIRV spike-in single cell data before imputation colored with read count. In tSNE plot of SIRV single cell RNA sequencing data colored with read count of each single cell RNA sequencing library, there is not significant cluster signature observed.

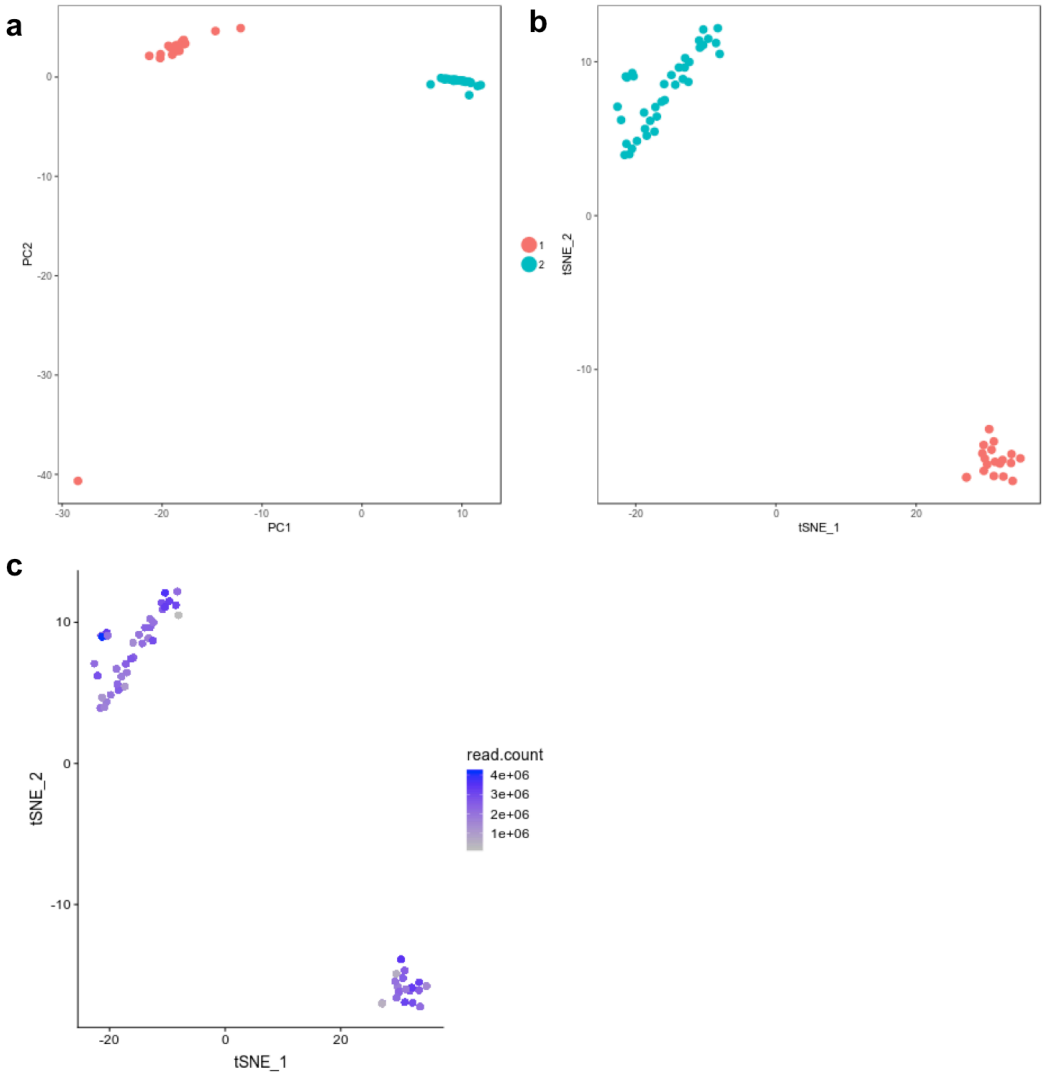

a
PC2
PC1
b
tSNE_2
tSNE_1
1
2
c
tSNE_2
tSNE_1
read.count
4e+06
3e+06
2e+06
1e+06

Figure 4.12: Clustering of single cells after imputation in SIRV spike-in single cell data.
a) PCA plot of SIRV single cell data after imputation. In PCA plot of SIRV single cell RNA sequencing data, all single cells significantly split into two clusters after imputation in PCA plot.
b) tSNE plot of SIRV single cell data after imputation. In tSNE plot of of SIRV single cell RNA sequencing data, all single cells significantly split into two clusters after imputation in tSNE plot similar to PCA plot.
c) tSNE plot of SIRV single cell data after imputation colored with read count. In tSNE plot of SIRV single cell RNA sequencing data colored with read count of each single cell RNA sequencing library, total number of reads do not differ too much in the SIRV spike in singe cell data and did not affect clustering.

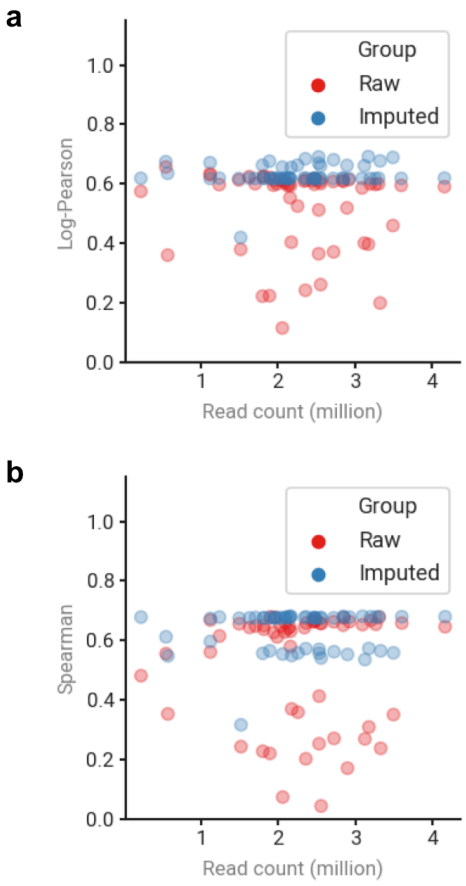


Figure 4.13: Performance of Seekmer in SIRV spike-in single cell data. *a) Log-Pearson correlation of SIRV transcripts quantification with true answer before and after imputation with different read counts in RNA sequencing libraries.* Log-Pearson performance after imputation is much higher than the performance before imputation. The performance after imputation does not change much compared to the performance before imputation with read counts. *b) Spearman correlation of SIRV transcripts quantification with true answer before and after imputation with different read counts in RNA sequencing libraries.* Spearman performance after imputation is much higher than the performance before imputation. The performance after imputation does not change much compared to the performance before imputation with read counts.

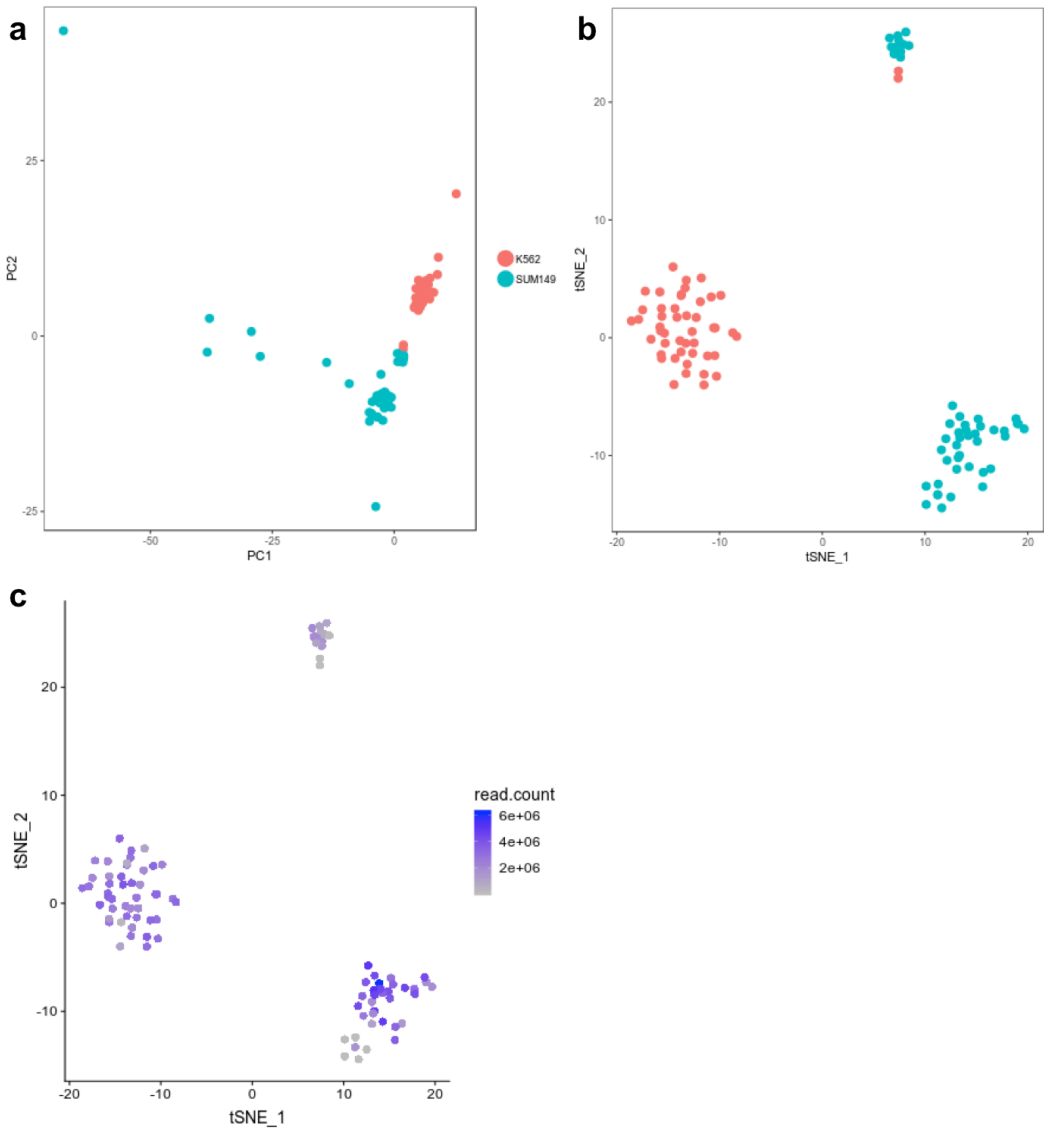

a
PC2
PC1
K562
SUM149
b
tSNE_2
tSNE_1
c
tSNE_2
tSNE_1
read.count
6e+06
4e+06
2e+06

Figure 4.14: Clustering of Seekmer in real single cell data before imputation.
a) PCA plot of real single cell data before imputation. In PCA plot of real single cell RNA sequencing data, all single cells are clustered togrther before imputation with the exception of a few single cells.
b) tSNE plot of real single cell data before imputation. In tSNE plot of of SIRV single cell RNA sequencing data, most of K562 and SUM149 cells culstered separately with a few exceptions.
c) tSNE plot of real single cell data before imputation colored with read count. In tSNE plot of SIRV single cell RNA sequencing data colored with read count of each single cell RNA sequencing library, the smallest cluster with both K562 and SUM149 contain all the cells with the fewst reads.

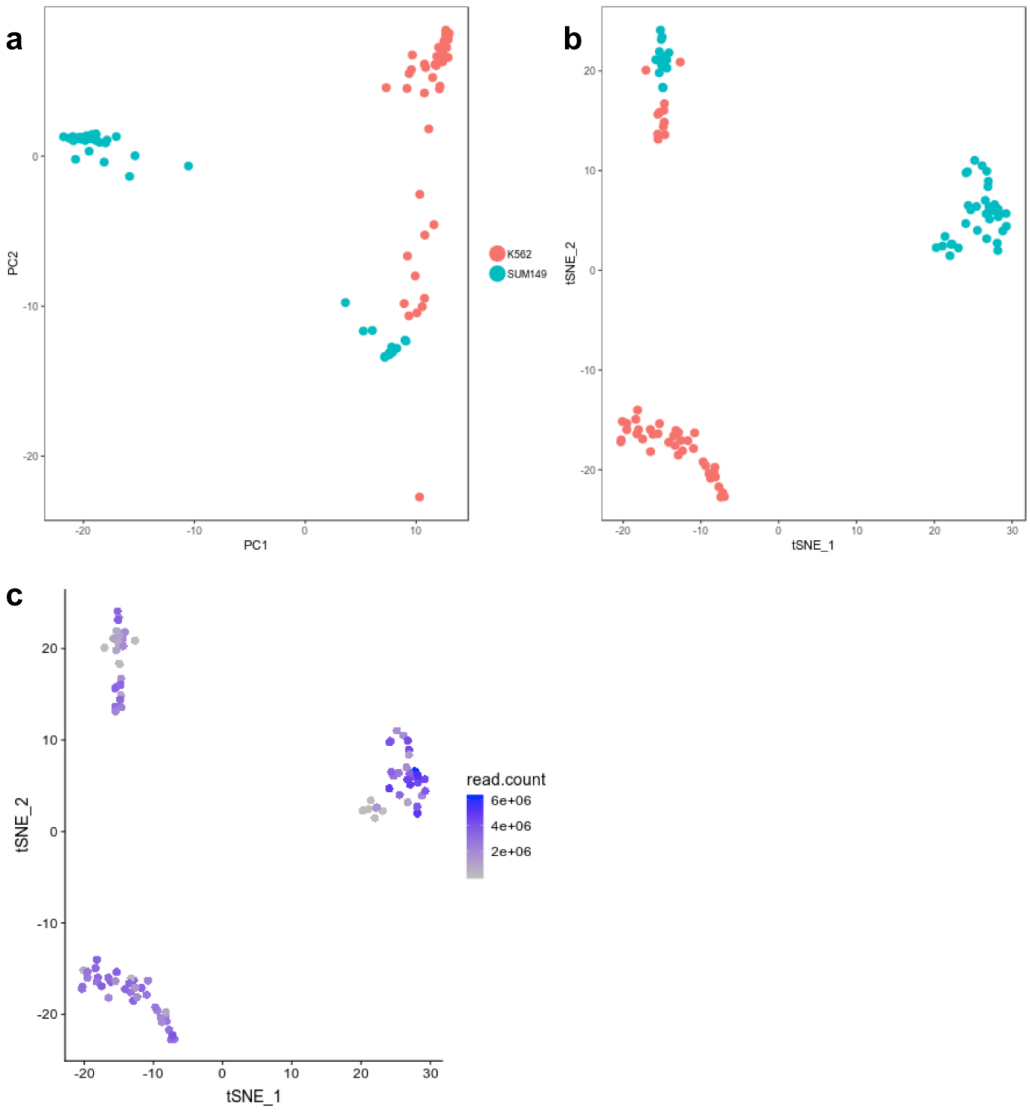

a
PC2
PC1
K562
SUM149
b
tSNE_2
tSNE_1
c
tSNE_2
tSNE_1
read.count
6e+06
4e+06
2e+06

Figure 4.15: Clustering of Seekmer in real single cell data after imputation.
a) PCA plot of real single cell data after imputation. In PCA plot of real single cell RNA sequencing data, three distinct clusters could be observed.
b) tSNE plot of real single cell data after imputation. In tSNE plot of of SIRV single cell RNA sequencing data, three distinct clusters could be observed.Mmost of K562 and SUM149 cells culstered separately in two major clusters, while there is one small cluster with both of the cell types.
c) tSNE plot of real single cell data after imputation colored with read count. In tSNE plot of SIRV single cell RNA sequencing data colored with read count of each single cell RNA sequencing library, the small cluster with both K562 and SUM149 contain all the cells with the fewst reads.

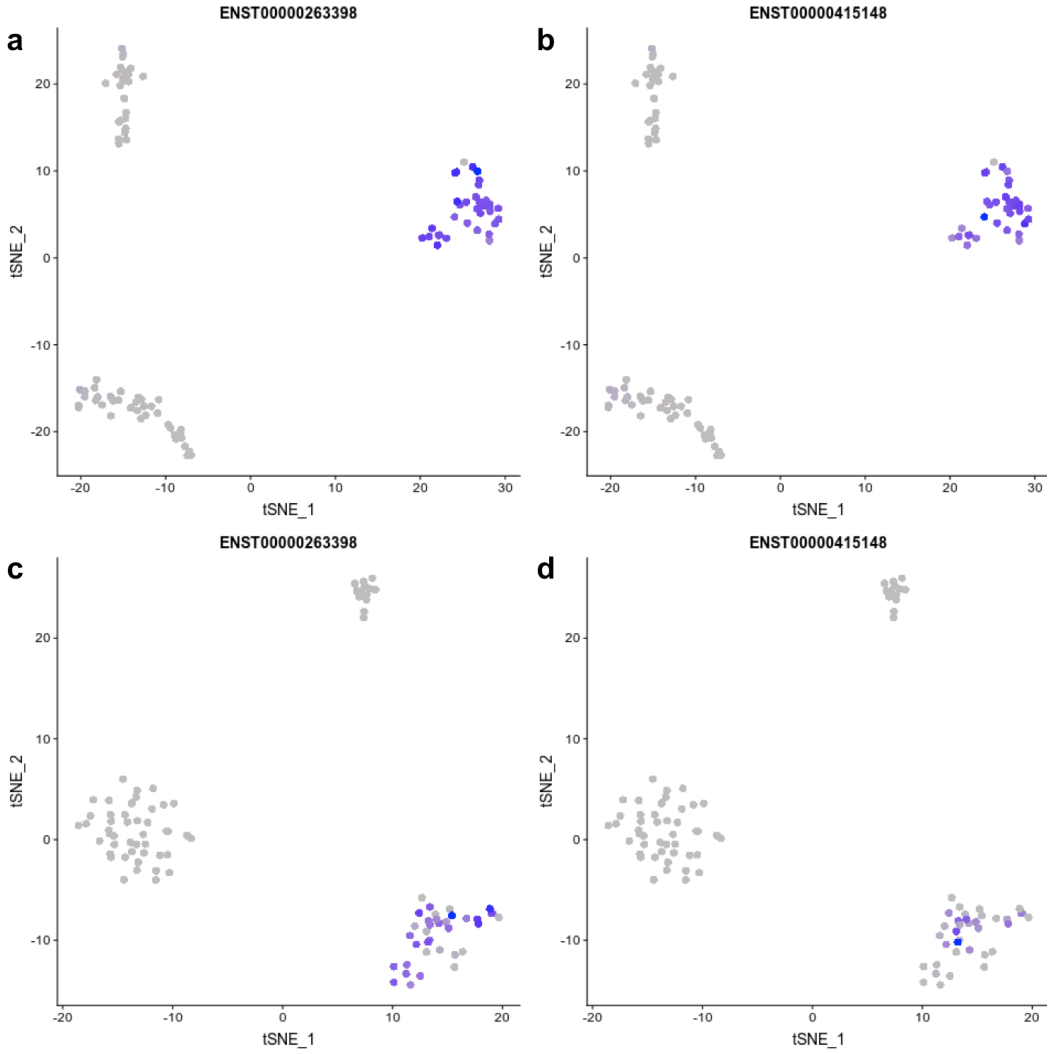

a
ENST00000263398
tSNE_2
tSNE_1
b
ENST00000415148
tSNE_2
tSNE_1
c
ENST00000263398
tSNE_2
tSNE_1
d
ENST00000415148
tSNE_2
tSNE_1

Figure 4.16: Identification of different transcripts of CD44 in real single cell RNA-sequencing data.
tSNE plot of real single cell RNA sequencing data with K562 and SUM149. The color represents the expression of certain transcripts in each single cell. Seekmer imputation significantly improves the quantification of gene marker CD44 in SUM149.
a) CD44-201 expression level quantified by Seekmer after imputation.
b) CD44-206 expression level quantified by Seekmer after imputation.
c) CD44-201 expression level quantified by Seekmer before imputation.
d) CD44-206 expression level quantified by Seekmer before imputation.

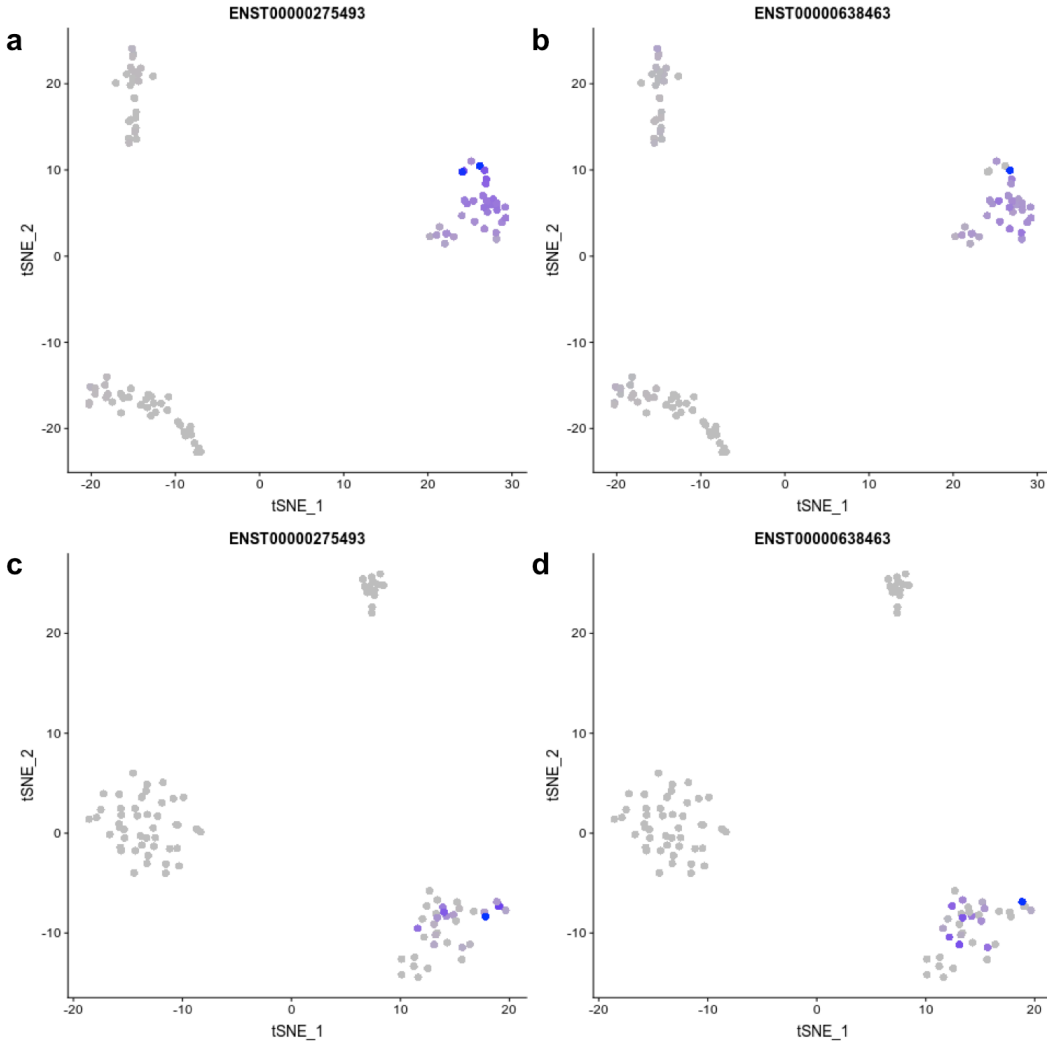

a
ENST00000275493
b
ENST00000638463
c
ENST00000275493
d
ENST00000638463
tSNE_1
tSNE_2

Figure 4.17: Identification of different transcripts of EGFR in real single cell RNA-seq data.
tSNE plot of real single cell RNA sequencing data with K562 and SUM149. The color represents the expression of certain transcripts in each single cell. Seekmer imputation significantly improves the quantification of gene marker EGFR in SUM149.
a) EGFR-201 expression level quantified by Seekmer after imputation.
b) EGFR-211 expression level quantified by Seekmer after imputation.
c) EGFR-201 expression level quantified by Seekmer before imputation.
d) EGFR-211 expression level quantified by Seekmer before imputation.

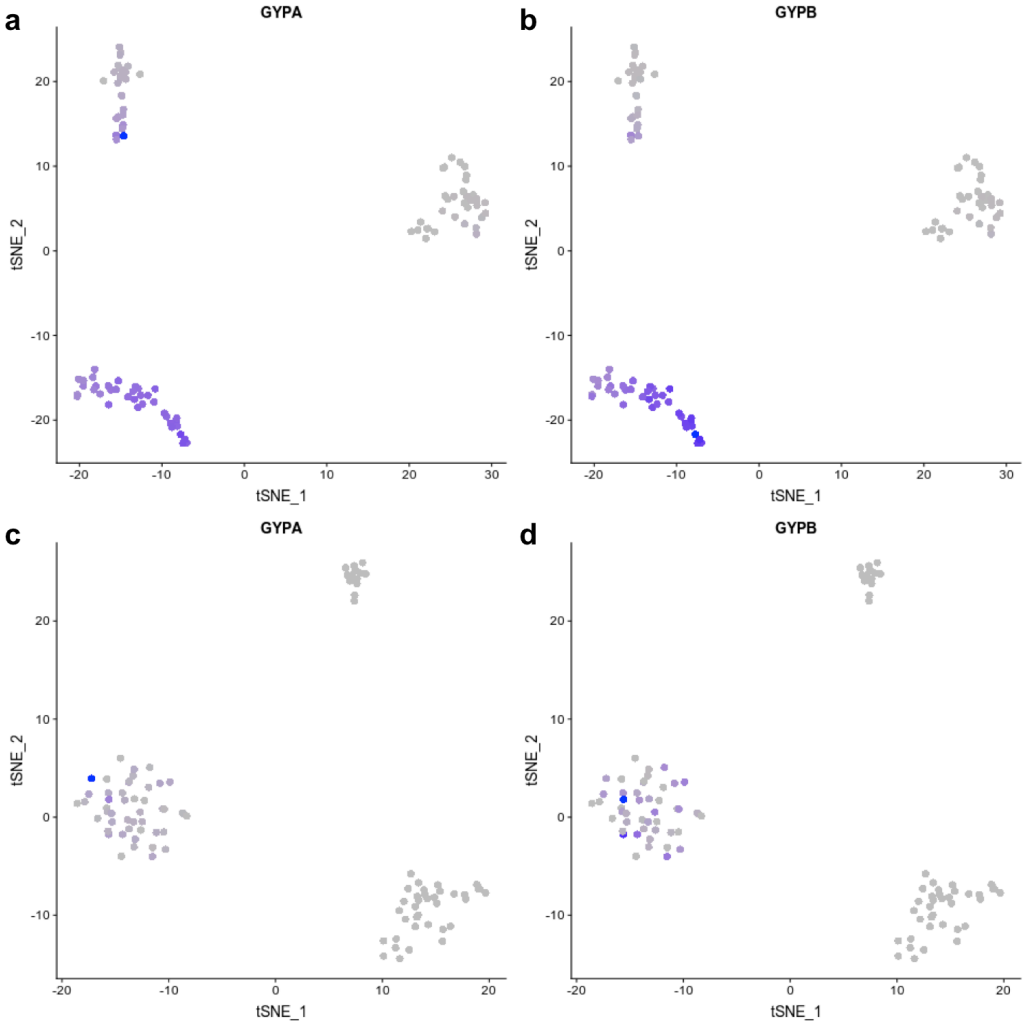

a
GYPA
b
GYPB
c
GYPA
d
GYPB
tSNE_1
tSNE_2

Figure 4.18: Identification of GYPA and GYPB in real single cell RNA-seq data.
tSNE plot of real single cell RNA sequencing data with K562 and SUM149. The color represents the expression of certain transcripts in each single cell. Seekmer imputation significantly improves the quantification of gene marker GYPA and GYPAB in K562.
a) GYPA expression level quantified by Seekmer after imputation.
b) GYPB expression level quantified by Seekmer after imputation.
c) GYPA expression level quantified by Seekmer before imputation.
d) GYPB expression level quantified by Seekmer before imputation.

Table 4.1: Performance of Seekmer compared to other methods (HBRR).

		SRR950079	SRR950081	SRR950083	SRR950085	SRR950087
Log-Pearson	Kallisto	0.7601	0.7565	0.7583	0.7566	0.7641
	Sailfish	0.7595	0.7558	0.7576	0.7561	0.7638
	Salmon	0.7599	0.7562	0.7584	0.7567	0.7641
	STAR-RSEM	0.7760	0.7727	0.7753	0.7731	0.7779
	Seekmer	0.7747	0.7719	0.7747	0.7729	0.7760
Spearman	Kallisto	0.8165	0.8133	0.8152	0.8126	0.8197
	Sailfish	0.8162	0.8129	0.8147	0.8123	0.8195
	Salmon	0.8163	0.8131	0.8151	0.8126	0.8195
	STAR-RSEM	0.8237	0.8211	0.8232	0.8207	0.8266
	Seekmer	0.8252	0.8224	0.8253	0.8228	0.8273

Table 4.2: Performance of Seekmer compared to other methods (UHRR).

		SRR950078	SRR950080	SRR950082	SRR950084	SRR950086
Log-Pearson	Kallisto	0.7325	0.7350	0.7352	0.7301	0.7464
	Sailfish	0.7321	0.7344	0.7348	0.7295	0.7461
	Salmon	0.7327	0.7350	0.7353	0.7301	0.7466
	STAR-RSEM	0.7456	0.7483	0.7477	0.7428	0.7575
	Seekmer	0.7481	0.7500	0.7511	0.7461	0.7578
Spearman	Kallisto	0.8043	0.8061	0.8077	0.8032	0.8161
	Sailfish	0.8041	0.8056	0.8073	0.8028	0.8159
	Salmon	0.8045	0.8061	0.8077	0.8033	0.8162
	STAR-RSEM	0.8105	0.8117	0.8132	0.8088	0.8213
	Seekmer	0.8141	0.8147	0.8169	0.8131	0.8223

Table 4.3: Run time (mins) of Seekmer compared to other methods.

		SRR950079	SRR950081	SRR950083	SRR950085	SRR950087
HBRR	Kallisto	29	25	21	18	11
	Sailfish	84	75	72	64	40
	Salmon	82	63	142	61	43
	STAR-RSEM	1157	1151	900	953	552
	Seekmer	51	55	54	52	32
		SRR950078	SRR950080	SRR950082	SRR950084	SRR950086
UHRR	Kallisto	26	24	20	30	15
	Sailfish	66	61	53	81	42
	Salmon	91	62	47	81	43
	STAR-RSEM	1060	920	651	1203	741
	Seekmer	52	45	36	73	40

Table 4.4: Performance of Seekmer single cell imputation with different power.

	Power	1	2	4	6	8	10	16	32
Log-Pearson	Gene	0.8321	0.8322	0.8323	0.8324	0.8325	0.8326	0.8327	0.8327
	Transcript	0.6723	0.6727	0.6734	0.6741	0.6747	0.6752	0.6766	0.6781
	PCR	0.7236	0.7237	0.7238	0.7240	0.7242	0.7243	0.7246	0.7247
Spearman	Gene	0.8471	0.8471	0.8471	0.8472	0.8472	0.8472	0.8473	0.8473
	Transcript	0.5449	0.5452	0.5459	0.5464	0.5467	0.5470	0.5480	0.5495
	PCR	0.8149	0.8149	0.8149	0.8150	0.8152	0.8153	0.8156	0.8161

Reference

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology, 11(10), p.R106.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biology, 17(1).

Bray NL, Pimentel H, Melsted P, Pachter L. Erratum: Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;34(8):888-888. doi:10.1038/nbt0816-888d.

Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), pp.15-21.

Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports, 2(3), pp.666-673.

Hwang, B., Lee, J. and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine, 50(8).

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J., Lonnerberg, P. and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research, 21(7), pp.1160-1167.

Li, B. and Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12(1).

Li, J., Witten, D., Johnstone, I. and Tibshirani, R. (2011). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics, 13(3), pp.523-538.

Macosko, E., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A. and McCarroll, S. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell, 161(5), pp.1202-1214.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 5(7), pp.621-628.

Patro, R., Mount, S. and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology, 32(5), pp.462-464.

Patro, R., Duggal, G., Love, M., Irizarry, R. and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14(4), pp.417-419.

Picelli, S., Faridani, O., Björklund, Å., Winberg, G., Sagasser, S. and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols, 9(1), pp.171-181.

Ramsköld, D., Luo, S., Wang, Y., Li, R., Deng, Q., Faridani, O., Daniels, G., Khrebtukova, I., Loring, J., Laurent, L., Schroth, G. and Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology, 30(8), pp.777-782.

Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3), p.R25.

SEQC/MAQC-III Consortium, A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology. 2014;32(9):903-914. doi:10.1038/nbt.2957.

Svensson V, Natarajan KN, Ly L-H, et al. Power analysis of single-cell RNA-sequencing experiments. Nature Methods. 2017;14(4):381-387. doi:10.1038/nmeth.4220

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B., Siddiqui, A., Lao, K. and Surani, M. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods, 6(5), pp.377-382.

Teng, M., Love, M., Davis, C., Djebali, S., Dobin, A., Graveley, B., Li, S., Mason, C., Olson, S., Pervouchine, D., Sloan, C., Wei, X., Zhan, L. and Irizarry, R. (2016). A benchmark for RNA-seq quantification pipelines. Genome Biology, 17(1).

Trapnell, C., Pachter, L. and Salzberg, S. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), pp.1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D., Pimentel, H., Salzberg, S., Rinn, J. and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), pp.562-578.

Wagner, A., Regev, A. and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), pp.1145-1160.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), pp.57-63.

Chapter 5 Conclusion

Overview

In this dissertation, I have outlined three diverse projects that are unified by the theme of detecting low frequency events in complex sequencing data. I have demonstrated a series of best practices to identify somatic single nucleotide variants in non-tumor tissues, in particular somatic SNVs with low allele fractions in brain tissues, using primarily whole genome and whole exome sequencing data with the additional haplotype information from 10X Genomics data.

Furthermore, I have presented a pipeline I developed to detect reads supporting single molecule RNA level fusion events between two repetitive sequences, U6 and L1, to investigate the mechanism of the formation of U6/L1 pseudogenes in the human genome and have extended this analysis to included other RNAs as well. The fusion RNA detection methods between repetitive sequences can help with future studies of possible function of fusion RNAs in cells.

Finally, I described an approach, Seekmer, that fills the gap between alignment-free methods and alignment-based methods in the performance of isoform quantification in both simulated and real RNA sequencing data. Seekmer

provides a better quantification of isoform expression particularly in single cell RNA sequencing data by collecting information from cells with similar expression profiles. The more accurate isoform quantification of single cells can facilitate the research of single cell level dynamics of transcriptome in different individuals, tissues and diseases.

Somatic single nucleotide variations (SNVs) in genome of non-tumor cells

In chapter 2, I have proposed a series of best practice to discover somatic SNVs in non-tumor tissue, in particular, for somatic SNVs with low allele fraction in tissues. A more accurate and sensitive pipeline for somatic SNV identification can facilitate future studies of somatic SNVs in different tissues and diseases particularly in non-tumor tissues without clonal expansion that have been less studied previously.

What is the overall level of mosaicisms and the function of somatic SNVs in brain and other tissues?

With the more sensitive somatic SNV detection methods, we can now begin to discover the potential functional role that somatic SNVs in brain tissues may play and how they are related to neurological diseases. Researchers began to characterize loss of function (LoF) genomic variants in protein coding regions with the large scale of sequencing projects in both healthy individuals and patients with different diseases (MacArthur and Tyler-Smith, 2010; MacArthur et

al., 2012). Previous studies showed the relationship between genomic mutations and phenotypes, diseases as well as the role of somatic mutations in evolution. With our best practice, we can now start to study the somatic LoF SNVs in protein coding region in different tissues.

Starting from the brain, with the whole exome sequencing data available we can analyze the somatic LoF SNVs in normal brains and discover the relationship between somatic LoF SNVs and age. Although existing study showed an increase of somatic SNVs in brain with age (Lodato et al., 2017), whether number of low allele fraction somatic LoF SNVs will increase or decrease with age still remains unknown. Somatic LoF SNVs can accumulate over time when brain grows old, however, if certain somatic LoF SNVs are lethal enough the kill the neurons with certain somatic SNVs, the number of somatic LoF SNVs can also decrease as the brain ages. We can also compare the age matched brains of neurotypical individuals and schizophrenic individuals to discover the somatic SNVs associated with schizophrenia. We can also investigate if somatic LoF SNVs are selected against to the same extent as observed with the germline LoF SNVs in previous studies. Furthermore, similar studies can be performed using other different tissues. Since neurons are unique due to their long life span and no mitosis during adult life compared to most of the other tissues which still undergo mitosis, the somatic LoF SNV pattern may be different in tissues other than brain tissue.

With the discovered somatic LoF SNVs in protein coding regions, we then can investigate into RNA sequencing data from the same tissue to identify whether there is an impact of the somatic LoF SNVs on the type of isoform expressed as well as the expression level of the isoforms. This is a challenge to which our Seekmer approach is well suited. The presence of the impact of isoform expression of somatic LoF SNVs would provide strong evidence that certain somatic SNVs act an important role in the tissue that we study.

How does the somatic mutation accumulate in brain tissues?

Previous studies show that a single neuron in a healthy human brain can harbor as many as 1458 to 1580 SNVs (Lodato et al., 2015). In addition, at least one megabase CNV is present in 13-41% human frontal cortex neurons (McConnell et al., 2013). Explanations of relatively high somatic mutation rate in neurons include long life span of neurons in human brain, which provides great opportunities for accumulation of different variations in cells (McConnell et al. 2017).

Somatic mutations could accumulate in long life span neurons due to both endogenous and exogenous factors. Exogenous factors include various kinds of environmental mutagenesis including radiations, mutagenic chemicals and so forth (Perera and Herbstman, 2011). Endogenous factors include failure of DNA repair for DNA damages from transcription, mobile element insertion and so on. The active transcription of genes in neuron cells can bring a huge burden to DNA

repair. Existing studies presented that the single strand DNA formed during transcription is exposed to multiple possible DNA damage factors that will induce the formation of SNVs, CNVs and large structural variations by recombination (Aguilera, 2002).

In addition to somatic mutations occur post-mitotically, somatic mutations in neurons can also be accumulated in mitosis during brain development at early stage. Compared to the mutations occurred post-mitotically, the mutations accumulated during brain development are more likely to be present in more neurons and have a larger impact. There are in total ~ 80 billion neurons in human brain (McConnell et al. 2017). The rudimentary brain structure and central nervous system are formed during the embryonic period. Although proliferation, migration, and differentiation continue postnatally, the ~80 billion neurons in an adult human brain have been generated during embryonic development (Stiles and Jernigan, 2010). The mechanism of how this tremendous feat of cell divisions completing in such a short time window with such a low error rate for the DNA replication remain unclear to the field. A reasonable hypothesis of the somatic mutations accumulated in neurons can be that the somatic mutations start to accumulate during the rapid cell division of brain cell generation in embryonic brain development already. Future studies on how and when the somatic mutation occurs during development can start from the somatic mutations in mouse brains at different developmental stages. With the existing technologies, researchers have begun to discover the somatic mutation rate in

neurons, however, the studies of the mechanisms of how the somatic mutations accumulate in neurons still remain unclear. Better discovery methods of somatic mutation in non-tumor tissues, like the best practice summarized in Chapter 2 can facilitate more studies of somatic mutations in brain in the future.

What is the possible function or effect of accumulated high mosaicism in brain tissues?

Somatic mutations in neurons have been associated with multiple neurological diseases in previous studies (Poduri et al., 2013). However, somatic mutations are also identified in normal human postmortem neurons from multiple studies (McConnell et al., 2013; Cai et al., 2014; Bae et al., 2017; Lodato et al., 2017). Researchers showed the increasing number of somatic SNVs in elder human brains compared to younger ones (Lodato et al., 2017). While CNVs and complex karyotypes are discovered to be rarer in elder brains (Chronister et al., 2019). Although somatic mutations have been associated with aging of brains in previous studies (Lodato et al., 2017), it still cannot explain the decrease of CNVs and complex karyotypes in elder brains. The accumulation of SNVs and decrease CNVs overtime in human brains can be associated with cell apoptosis and maintaining the large size and active transcription in neurons. The difference between somatic SNVs and somatic CNVs can also be due to the different lethality of different kind of mutations. In this situation, somatic CNVs may be more lethal to neurons leading to the apoptosis of neurons with accumulated CNVs. While somatic SNVs are more neutral for most of the cases, the

accumulation of somatic SNVs did not kill the neurons. This could also explain why somatic SNVs accumulate in neurons while number of somatic CNVs decreases during aging. However, additional research is needed, for example, the total number of neurons in brains of different ages and so on, before any conclusion can be drawn.

What if we apply best practice of somatic SNV discovery in normal human cells to tumor cells?

We have demonstrated a series of best practices for discovery of somatic SNVs in non-tumor tissues in chapter 2. With these approaches, we were able to filter the somatic SNV candidates from the output of existing methods for somatic SNVs with allele fraction as low as 1%. Given that the sensitivity of cancer somatic SNV detection methods dramatically decreases when allele fraction is less than 5% (Cibulskis et al., 2013), if we could apply our best practice somatic SNV identification pipeline to tumor samples, we may be able to find more somatic SNVs with lower allele fraction in tumors. The ability to discover lower allele fraction somatic SNVs in tumor could help to better characterize the evolution of tumors. As discussed in chapter 2, the detailed evolution mechanism for cancer cell progression and evolution still remain unclear (Stratton, Campbell and Futreal, 2009). Thus, with a better profile of somatic SNVs in tumor tissues can possibly facilitate building the evolutionary process of tumor cells.

Limitations of somatic SNV identification from current sequencing methods

Homopolymer regions are highly prone to sequencing errors, however, they are also the regions where mutations could occur because of the DNA polymerase slippage during DNA duplication (Streisinger et al., 1966; Kunkel, 2004). Due to the limitations of current next generation sequencing methods, we have not been able to ascertain the validity of candidates in homopolymer regions. As such, we have excluded the somatic SNV candidates in homopolymer regions to retain a high specificity in our accuracy. Most of the methods for sequencing or base identification requires PCR amplification, for example, ddPCR and next generation sequencing. PCR amplification of DNA libraries before sequencing or other methods is applied to amplify the signals for downstream detection. We thus have difficulty distinguishing amplification error from true somatic SNVs if there is error-prone amplification step involved in the DNA preparation steps. With current technologies, we could make use of restriction enzymes if a restriction site is created or removed by the mutation of certain bases inside a homopolymer region. However, this has a both high tissue and labor requirements if validating all candidate somatic SNVs identified in homopolymer regions. Furthermore, this method would only be limited to the sites where restriction sites could be created. In the future, methods without an amplification step could solve this issue to distinguish between sequencing artifacts and true somatic SNVs in homopolymer regions. However, even third generation sequencing technologies like Oxford Nanopore Technologies (ONT) still have relatively high sequencing errors in homopolymer regions due to the same signal

released for a string of homopolymers longer than the *k*-mer length of reading (Rang, Kloosterman and de Ridder, 2018).

Another possible limitation is that methods for discovery of structural variations (SV) or copy number variants (CNV) from next generation sequencing data are not perfect. False positive somatic SNV candidates in structural variant or copy number variant regions could still not be excluded from the raw sites due to the inaccuracy of the SV/CNV calls themselves. Third generation sequencing methods with extremely long read length can better identify structural variation and CNVs, however, current methods are still expensive to apply in large numbers of samples for further disease related somatic SNV studies.

U6 snRNA chimeric events in human transcriptome

Chimeric U6/L1 RNAs are present in human cells

In chapter 3, I developed a method to identify individual RNA-Seq reads exhibiting evidence of a U6/L1 chimeric RNA. We were able to distinguish reads with U6/L1 chimeric RNA supportive evidence from U6/L1 chimeric reads generated from template switching and U6/L1 reads transcribed from existing genomic U6/L1 pseudogenes (Figure 3.4). We have also searched the 25 base unique pair junction sequences formed at the fusion position in 22 high coverage 1000 Genomes samples. We were able to successfully identify all aligned U6/L1 junction sequences (3 events with SNPs in one of the 22 samples after manual checking) in all 22 samples. This indicates that the U6/L1 chimeric reads aligned

to genome were generated from existing U6/L1 pseudogenes. In contrast, the non-aligned U6/L1 junction sequences were not detected in any of the 22 samples, suggesting that the non-aligned U6/L1 events are unique, RNA level fusion events. Possible template switching artifacts that could have resulted during library preparation were not detected in our searches.

The uniqueness of the junction sequences and the fact that these sequences all containing the full length U6 sequence strongly suggest that the U6/L1 chimeric sequences that we identified are formed from U6/L1 RNA level ligation rather than RNA fusion artifacts formed by template switching during RNA sequencing library preparation.

RNA ligation generated chimeric U6/L1 RNA with RtcB ligase

The biochemical analysis from our collaborators Dr. John Moldovan and Dr. John Moran showed that U6 snRNAs and L1 RNAs could be ligated to form U6/L1 chimeric RNA with RtcB ligase *in vitro*. Furthermore, two HeLa cell lines with reduced RtcB expression utilizing CRISPR/Cas9 gene-editing method showed a depletion of U6/L1 fusion events for ~ 4-5 folds compared to the negative control with RtcB regularly expressed. These biochemical and genetic experiments all showed that U6/L1 chimeric RNAs were ligated under the catalysis of RtcB enzyme.

U6 chimeric RNAs presents in the transcriptome of human cells

In chapter 3, I have then further studied the RNAs fused with U6 snRNAs in human cancer cell lines, hESCs, and human NPCs. With the existence of U6/other RNA fused pseudogenes in human genome (Buzdin et al. 2003) (Figure 5.1), we hypothesized that RtcB could also fuse U6 with other RNAs in the transcriptome. We found multiple reads supporting U6 snRNA fusion with other RNAs in human cell transcriptome (Figure 3.7). With all the junction sequences identified from U6 snRNA chimeric events, we were able to briefly characterize the motif and secondary structure of enriched U6 fusions (Figure 3.8; Figure 3.9). To further understand the mechanism of the formation of U6 snRNA fusion at RNA level, more studies are necessary including both biochemical and computational experiments.

A model of U6/other pseudogene formation

With all previous evidence, we suggest that the mechanism of U6/other pseudogene formation takes the same fusion step as the formation of U6/L1 chimeric RNAs. Fewer chimeric RNAs could then be retrotransposed to genome *in trans* with the L1 retrotransposition machinery. Since the retrotransposition of these chimeric RNAs requires the L1 machinery *in trans*, it does not occur as frequent as U6/L1 chimeric RNA retrotransposition *in cis* (Wei et al., 2001). This explains the fewer U6/mRNA and U6/Alu pseudogenes in genome compared to the U6/L1 pseudogenes (Figure 5.1) (Buzdin et al. 2003). Future study is required for understanding the formation of 5'-OH group by endonucleases in the other RNAs.

What is possible function of U6 chimeric RNAs?

From our gene fusion discovery with U6 snRNA from five human cell lines, we were able to identify the enrichment of snRNAs and snoRNAs fused with U6. The co-localization of U6 and other snRNAs in the spliceosome (Papasaikas and Valcárcel, 2016) makes the fusion events occur easier. In addition, both the formation of mature mRNA and snoRNA involves RNA splicing catalyzed by spliceosome. The fusion of U6 snRNA with mRNAs (Shi, 2017) as well as snoRNAs (Dupuis-Sandoval, Poirier and Scott, 2015) could form when splicing takes place in cells. Future studies of possible motif or secondary structures recognized by certain endonuclease to create the 5'-OH group, which is necessary for U6 fusion to happen, is also required for more detailed studies of U6 fusion phenomena in general.

The detailed mechanism of formation of U6/other RNA formation remains unclear. Thus, the function of U6 fusion with other genes is still not clear. The high involvement of U6 fusion genes with RNA splicing may provide a hypothesis that the formation of U6 and other RNAs is related with RNA degradation. The hypothesis would be the RNAs needed to be degraded fuse with U6 as a signal for further degradation steps. However, there have been no studies successfully demonstrating the detailed mechanism of how RNA degradation happens efficiently in human cells (Arraiano et al., 2010). Further studies are necessary to demonstrate the function of U6 snRNA fusion phenomena in human cells.

Limitations of identification of single molecule events from next generation sequencing

As we described in the method of Chapter 3, we were cautious of the PCR duplicates generated from RNA sequencing library preparation reverse transcription. However, there is no perfect solution for this issue except for checking the read similarity due to the PCR amplification step of the next generation sequencing method. In order to identify single molecule level events in transcriptome, we need to apply barcoded method to distinguish the PCR duplicates from reads generated from two different molecules. Current technologies with barcoded RNA sequencing could better solve this PCR duplicates problem in the future.

RNA sequencing quantification

In chapter 4, I have demonstrated that Seekmer can accurately quantify isoform expression in particular in single cell RNA sequencing by collecting information from single cells with similar isoform expression patterns. With a better isoform quantification in single cells, we expect to perform more analysis and investigate the dynamics of transcriptome at single cell level.

What can we improve with single cell RNA sequencing data using Seekmer?

Existing methods for RNA sequencing expression analysis either do not perform imputation or only perform imputation at gene level. The development of

Seekmer can facilitate with future single cell RNA sequencing analysis by providing better isoform quantification for each single cell. Previous studies using single cell RNA sequencing data mostly focuses on identifying signature genes expressed in different cell types by clustering single cells with similar expression level together (Gupta et al., 2018). Due to the lack of better isoform quantification method, limited studies have been focusing on investigating the difference among single cells in the same tissue. While the most important feature of single cell technologies is that we can now study each single cell and study the dynamics of different cells in the same tissue. With Seekmer, although we utilized the information provided by other single cells with similar isoform expression profile, we should still to some extent be able to study the difference among different cells at the transcriptome level.

Concluding Remarks

My thesis has developed methods to identify rare events including somatic single nucleotide variants, single molecule RNA level repetitive element fusion as well as single cell RNA sequencing quantification.

I have generated the best practice to discover somatic single nucleotide variants from non-tumor tissue with the information collected from multiple different sequencing libraries. I was able to identify the supportive evidence for single molecule RNA level U6/L1 chimeric events and helped with forming the new mechanism of U6/L1 pseudogene formation. I also showed the possibility that U6

also fuses with other RNAs in cells and possibly to have a cellular function related with RNA degradation from the genes that fused with U6. I also developed a single cell RNA sequencing quantification tool with a better performance than existing methods by collecting information from cells with similar expression profiles.

With the current sequencing technologies, there are artifacts that we are not able to exclude using computational methods. However, we have demonstrated that with cautious filtering and collecting extra information from other methods or other cells, we could to the largest extent utilize the current methods to study the characters and possible functions of the rare events in human genome and transcriptome.

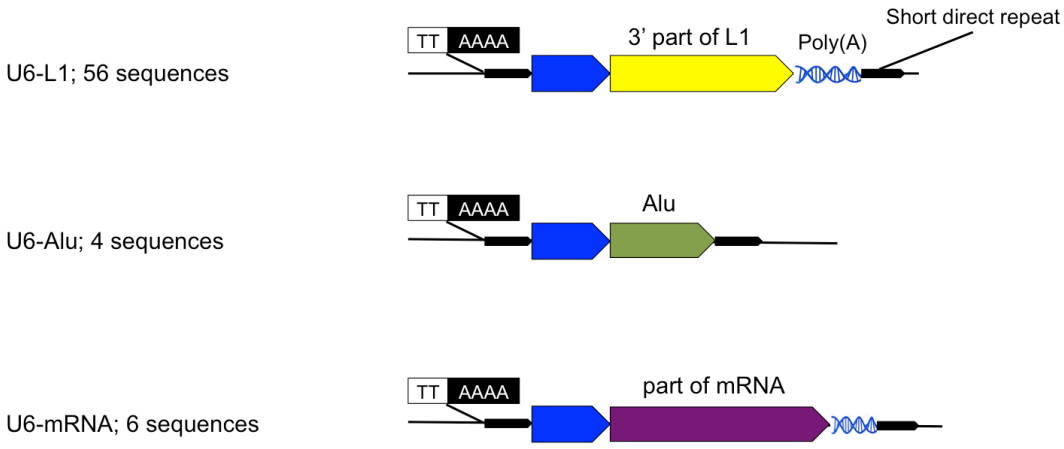


Figure 5.1 Schematic representation of the U6 chimeric pseudogenes in public databases.
*Adapted from Buzdin et al. 2003.

Reference

Aguilera, A. (2002). The connection between transcription and genomic instability. The EMBO Journal, 21(3), pp.195-201.

Arraiano, C., Andrade, J., Domingues, S., Guinote, I., Malecki, M., Matos, R., Moreira, R., Pobre, V., Reis, F., Saramago, M., Silva, I. and Viegas, S. (2010). The critical role of RNA processing and degradation in the control of gene expression. FEMS Microbiology Reviews, 34(5), pp.883-923.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biology, 17(1).

Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., Pletikos, M., Pattni, R., Chen, B., Venturini, E., Riley-Gillis, B., Sestan, N., Urban, A., Abyzov, A. and Vaccarino, F. (2017). Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science, 359(6375), pp.550-555.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y. and Sverdlov, E. (2002). A New Family of Chimeric Retrotranscripts Formed by a Full Copy of U6 Small Nuclear RNA Fused to the 3′ Terminus of L1. Genomics, 80(4), pp.402-406.

Buzdin, A. (2003). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Research, 31(15), pp.4385-4390.

Cai, X., Evrony, G., Lehmann, H., Elhosary, P., Mehta, B., Poduri, A. and Walsh, C. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Reports, 8(5), pp.1280-1289.

Chronister, W., Burbulis, I., Wierman, M., Wolpert, M., Haakenson, M., Smith, A., Kleinman, J., Hyde, T., Weinberger, D., Bekiranov, S. and McConnell, M. (2019). Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell Reports, 26(4), pp.825-835.e7.

Cibulskis, K., Lawrence, M., Carter, S., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology, 31(3), pp.213-219.

Didychuk, A., Butcher, S. and Brow, D. (2018). The life of U6 small nuclear RNA, from cradle to grave. RNA, 24(4), pp.437-460.

Doucet, A., Droc, G., Siol, O., Audoux, J. and Gilbert, N. (2015). U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. Molecular Biology and Evolution, 32(7), pp.1815-1832.

Dupuis-Sandoval, F., Poirier, M. and Scott, M. (2015). The emerging landscape of small nucleolar RNAs in cell biology. Wiley Interdisciplinary Reviews: RNA, 6(4), pp.381-397.

Garcia-Perez, J., Doucet, A., Bucheton, A., Moran, J. and Gilbert, N. (2007). Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome Research, 17(5), pp.602-611.

Gupta, I., Collier, P., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A., Sloan, S., Luo, W., Fedrigo, O., Ross, M. and Tilgner, H. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nature Biotechnology, 36(12), pp.1197-1202.

Kosmaczewski, S., Han, S., Han, B., Irving Meyer, B., Baig, H., Athar, W., Lin-Moore, A., Koelle, M. and Hammarlund, M. (2015). RNA ligation in neurons by RtcB inhibits axon regeneration. Proceedings of the National Academy of Sciences, 112(27), pp.8451-8456.

Kunkel, T. (2004). DNA Replication Fidelity. Journal of Biological Chemistry, 279(17), pp.16895-16898.

Lodato, M., Woodworth, M., Lee, S., Evrony, G., Mehta, B., Karger, A., Lee, S., Chittenden, T., D'Gama, A., Cai, X., Luquette, L., Lee, E., Park, P. and Walsh, C. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science, 350(6256), pp.94-98.

Lodato, M., Rodin, R., Bohrson, C., Coulter, M., Barton, A., Kwon, M., Sherman, M., Vitzthum, C., Luquette, L., Yandava, C., Yang, P., Chittenden, T., Hatem, N., Ryu, S., Woodworth, M., Park, P. and Walsh, C. (2017). Aging and neurodegeneration are associated with increased mutations in single human neurons. Science, 359(6375), pp.555-559.

Lu, Y., Liang, F. and Wang, X. (2014). A Synthetic Biology Approach Identifies the Mammalian UPR RNA Ligase RtcB. Molecular Cell, 55(5), pp.758-770.

MacArthur, D. and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. Human Molecular Genetics, 19(R2), pp.R125-R130.

MacArthur, D., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J., Montgomery, S., Albers, C., Zhang, Z., Conrad, D., Lunter, G., Zheng, H., Ayub, Q., DePristo, M., Banks, E., Hu, M., Handsaker, R., Rosenfeld, J., Fromer, M., Jin, M., Mu, X., Khurana, E., Ye, K., Kay, M., Saunders, G., Suner, M., Hunt, T., Barnes, I., Amid, C., Carvalho-Silva, D., Bignell, A., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D., Xue, Y., Romero, I., Wang, J., Li, Y., Gibbs, R., McCarroll, S., Dermitzakis, E., Pritchard, J., Barrett, J., Harrow, J., Hurles, M., Gerstein, M. and Tyler-Smith, C. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science, 335(6070), pp.823-828.

McConnell, M., Lindberg, M., Brennand, K., Piper, J., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R., Vermeesch, J., Hall, I. and Gage, F. (2013). Mosaic Copy Number Variation in Human Neurons. Science, 342(6158), pp.632-637.

McConnell, M., Moran, J., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J., Fasching, L., Flasch, D., Freed, D., Ganz, J., Jaffe, A., Kwan, K., Kwon, M., Lodato, M., Mills, R., Paquola, A., Rodin, R., Rosenbluh, C., Sestan, N., Sherman, M., Shin, J., Song, S., Straub, R., Thorpe, J., Weinberger, D., Urban, A., Zhou, B., Gage, F., Lehner, T., Senthil, G., Walsh, C., Chess, A., Courchesne, E., Gleeson, J., Kidd, J., Park, P., Pevsner, J. and Vaccarino, F. (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science, 356(6336), p.eaal1641.

Papasaikas, P. and Valcárcel, J. (2016). The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends in Biochemical Sciences, 41(1), pp.33-45.

Perera, F. and Herbstman, J. (2011). Prenatal environmental exposures, epigenetics, and disease. Reproductive Toxicology, 31(3), pp.363-373.

Poduri, A., Evrony, G., Cai, X. and Walsh, C. (2013). Somatic Mutation, Genomic Variation, and Neurological Disease. Science, 341(6141), p.1237758.

Rang, F., Kloosterman, W. and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biology, 19(1).

SEQC/MAQC-III Consortium, A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology. 2014;32(9):903-914. doi:10.1038/nbt.2957.

Shchepachev, V., Wischnewski, H., Soneson, C., Arnold, A. and Azzalin, C. (2015). Human Mpn1 promotes post-transcriptional processing and stability of U6atac. FEBS Letters, 589(18), pp.2417-2423.

Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nature Reviews Molecular Cell Biology, 18(11), pp.655-670.

Stiles, J. and Jernigan, T. (2010). The Basics of Brain Development. Neuropsychology Review, 20(4), pp.327-348.

Stratton, M., Campbell, P. and Futreal, P. (2009). The cancer genome. Nature, 458(7239), pp.719-724.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. and Inouye, M. (1966). Frameshift Mutations and the Genetic Code. Cold Spring Harbor Symposia on Quantitative Biology, 31(0), pp.77-84.

Villela, D., Suemoto, C., Leite, R., Pasqualucci, C., Grinberg, L., Pearson, P. and Rosenberg, C. (2018). Increased DNA Copy Number Variation Mosaicism in Elderly Human Brain. Neural Plasticity, 2018, pp.1-9.

Wagner, A., Regev, A. and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology, 34(11), pp.1145-1160.

Wei, W., Gilbert, N., Ooi, S., Lawler, J., Ostertag, E., Kazazian, H., Boeke, J. and Moran, J. (2001). Human L1 Retrotransposition: cis Preference versus trans Complementation. Molecular and Cellular Biology, 21(4), pp.1429-1439.

Zhang, B., Gunawardane, L., Niazi, F., Jahanbani, F., Chen, X. and Valadkhan, S. (2014). A Novel RNA Motif Mediates the Strict Nuclear Localization of a Long Noncoding RNA. Molecular and Cellular Biology, 34(12), pp.2318-2329.