Transposable Elements in Health and Disease

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division of Medical Sciences

in the subject of Biomedical Informatics

have examined a dissertation entitled

Transposable Elements in Health and Disease

presented by Rebeca Borges Monroy

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: Po-Ru Loh
Po-Ru Loh (Jul 15, 2021 13:48 EDT)

Typed Name: Dr. Po-Ru Loh

Signature: Cynthia C. Morton, Ph.D.
Cynthia C. Morton, Ph.D. (Jul 15, 2021 12:42 EDT)

Typed Name: Dr. Cynthia Morton

Signature:

Typed Name: Dr. Tim Yu

Signature:

Typed Name: Dr. Guillaume Bourque

Date: June 30, 2021

Transposable Elements in Health and Disease

A dissertation presented

by

Rebeca Borges Monroy

to

The Division of Medical Sciences

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the subject of

Biomedical Informatics

Harvard University

Cambridge, Massachusetts

June 2021

Transposable Elements in Health and Disease

Abstract

Transposable elements are DNA sequences that can move within the genome. They play a pivotal role in genomic variability in humans and can cause diseases. Their repetitive nature requires specific computational algorithms to detect new retrotransposon insertions with high sensitivity. Therefore, they are often overlooked in clinical testing and basic research. This thesis intends to systematically determine the impact of transposable elements in certain neurological disorders and the healthy brain.

The first main aim of this work was to identify the role of retrotransposons in autism spectrum disorder (ASD). We analyzed whole-genome sequencing data from 2,288 families with an individual with ASD. This large cohort provided the opportunity to study the frequency of retrotransposons in healthy parental individuals as well. We detected 86,154 polymorphic retrotransposon insertions, most of which were novel, and 158 *de novo* insertions. We obtained precise estimates of 1 *de novo* insertion per 29, 104, and 192 births for Alu, L1, and SVA, respectively. As expected, rates of *de novo* retrotransposition were similar between individuals with ASD and their unaffected siblings. The main finding from this analysis was that ASD cases showed more *de novo* L1 insertions in ASD genes. Here, we identified a candidate causal *de novo* insertion and exonic insertions in loss-of-function genes.

Our second main aim was to study the rates of somatic retrotransposition in healthy aging brain cells and neurodegeneration. Rates of somatic retrotransposition in the brain have been controversial. The work presented here supports the argument that somatic retrotransposition is rare in the brain in healthy individuals and the neurological disorders studied here. We hoped that this research will clarify previous conflicting data, despite technical issues that still prevent us from detecting retrotransposons in a sensitive and precise manner in single cells.

This work offers important insights into germline and somatic retrotransposition. Although retrotransposition rates are modest, their impact can be substantial. This work highlights the importance of using high-throughput computational tools to study rare variants, including transposable element insertions.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am thankful for all the support I have received during my Ph.D. I would first like to express my deepest gratitude to my advisor Dr. Chris A. Walsh. Chris took the time to meet with me during my undergraduate studies, and his enthusiasm motivated me to join Harvard and his group. Throughout the years, he has maintained this enthusiasm for science and is a source of inspiration. He has guided and encouraged me while also providing intellectual freedom. His scientific advice has been instrumental to my experiments and analyses, publications, and this thesis.

I also express my warmest gratitude to my co-advisor Dr. Eunjung Alice Lee for taking me under her wings as her first graduate student. Alice provided thoughtful feedback throughout my Ph.D. and was frequently available to discuss my research. I would also like to thank her for providing the opportunity to attend various meetings and encouraging me to share my work with the transposon community.

I would like to recognize the valuable guidance from my dissertation advice committee: Dr. Michael Talkowski, Dr. Po-Ru Loh, and Dr. Evan Macosko, and I would like to thank in advance my thesis examination committee: Dr. Po-Ru Loh, Dr. Cynthia C. Morton, Dr. Guillaume Bourque, and Dr. Timothy Yu for taking the time to read this thesis and to serve on my committee.

I have been extremely lucky to work with talented people throughout my Ph.D. I would like to thank Dr. Matthew B. Johnson, Dr. Richard Smith, Dr. Chong Chu, and Eduardo Maury for giving me the opportunity to collaborate with them. I'm grateful for the support received from Dr. Robert Sean Hill and Jennifer Partlow for obtaining clinical samples and managing computational resources. I also appreciate the guidance I received during my rotations from Dr. Michael Greenberg, Dr. Peter J. Park, Dr. Clemens R. Scherzer, Dr. Sinisa Hrvatin, Dr. Rachel Rodin, and Dr. Xianjun Dong.

My mentors from the Harvard Graduate Women in Science and Engineering association: Dr. Loise Francisco-Anderson, Dr. Maria Gutierrez Arcelus, and Insa Mohr have been a source of inspiration and empowerment. I would like to sincerely thank them for their career and life advice.

One of the best things about graduate school has been the friendships I have made here. I would like to thank Christina Jayson, Atsushi Taguchi, Nishita Parnandi, and Sarah Erlandson in particular for the laughter, tears, picnics, and mountain views we have shared. My time in the lab has also been greatly enriched by my close friendships with my colleagues Ellen DeGennaro, Eduardo Maury, and Taehwan Shin. I appreciate our lunch and coffee breaks that have endured even throughout a global pandemic, you have been a great source of support and intellectual growth. Thank you, Ambar and Jorge, for making Boston feel like home.

I would like to thank my family, whom I love and appreciate very much. My parents, Zury and Gui, cultivated a passion for learning in me that led me here. I am thankful for the sacrifices they have made so that I could receive the best education possible and for their support during this challenging period of my life. Thank you, Isa, for always having my back and for being a great big sister. Finally, a special thanks go to my handsome cats Pixel and Dobby, you have been a source of joy and comfort.

Chapter 1.
Introduction

Overview

A surprising and important finding that came from the completion of the human genome is that almost half of it is composed of transposable elements (TEs) (Lander et al., 2001). The ability of these elements to mobilize and replicate has shaped our genome. TEs can disrupt genes and gene expression, producing genetic variability and on certain occasions causing disease (Kazazian & Moran, 2017). Despite their prevalence, previous human genetics studies have generally excluded TEs due to their repetitive nature and the bioinformatical challenges that this presents. This thesis aims to investigate the role of TEs in neurological and neurodevelopmental disorders. In Chapter 2 we identify the contribution of *de novo* transposable element insertions (TEIs) to autism spectrum disorders (ASD) and in Chapter 3 we study the frequency of somatic TEIs in single cells of the human brain in aging and disease. In this chapter we introduce this work by providing a general background to TEs and somatic mutations in the brain, followed by a description of our research aims and their significance.

Transposable Elements

Overview

TEs, also known as mobile elements or “jumping genes” were discovered in the 1940s by Barbara McClintock in maize (McClintock, 1950). In her revolutionary and Nobel-winning findings, she characterized the interaction of TEs with pigmentation genes in maize kernels, which gave rise to striking variegation patterns (Feschotte, Jiang, & Wessler, 2002; McClintock, 1950). Their ability to replicate and insert themselves in the genome has greatly impacted genomic size in eukaryotes (Feschotte et al., 2002). In humans, around 45% of the genome is composed of TEs (Figure 1.1A) (Lander et al., 2001). Previously considered “junk DNA”, researchers have now

established that TEs are an evolutionary force, impact gene expression and regulation, and can also cause diseases (Kazazian & Moran, 2017; Pennisi, 2007). Here, we revise the structure, classification, and mechanisms for retrotransposition of TEs in humans. We also discuss their impact on the genome in both health and disease.

Classification and structure

The first main class of TEs is DNA transposons. These are similar to bacterial TEs, and their sequences contain inverted repeats as well as a transposase (Figure 1.1B) (Kazazian & Moran, 2017; Lander et al., 2001). The transposase recognizes these inverted repeats and excises the DNA as a double strand that is then inserted into another region of the genome in a "cut-and-paste" manner (Feschotte & Pritham, 2007). However, DNA transposons are inactive in humans and the majority of mammals except for bats (Hancks & Kazazian, 2016; Platt, Vandewege, & Ray, 2018) and account for ~3% of the human genome (Lander et al., 2001). Therefore, we will not focus on DNA transposons in this study.

The second main class of TEs is retrotransposons. These elements mobilize using a "copy-and-paste" mechanism. Retrotransposons are subdivided into human endogenous retrovirus (HERV) or long terminal repeat (LTR) retrotransposons and poly(A) or non-LTR retrotransposons (Figure 1.1) (Kazazian & Moran, 2017). HERVs are considered to be autonomous TEs because they can replicate and retrotranspose autonomously yet are mostly unable to retrotranspose in the human genome, although there may be some exceptions (Kazazian & Moran, 2017; Wildschutte et al., 2016). HERV expression has been linked to disease (Li et al., 2015). These "fossil viruses" are composed of similar genes as exogenous retroviruses and contain the *gag* gene, which

encodes a capsid protein, the *env* gene, which encodes an envelope protein, as well as the *pol* gene, which specifies the enzymes necessary for reverse transcription, integration, and protein cleavage (Grandi & Tramontano, 2018; Nelson et al., 2003).

In this study, we will focus on the main active retrotransposons of the human genome: long interspersed nuclear elements (L1s or LINEs) and short interspersed nuclear elements (SINEs). L1s are autonomous non-LTR retrotransposons and account for ~21% of the genome (Figure 1.1A) (Kazazian & Moran, 2017). A full-length L1 element is around 6 kb long (A. F. Scott et al., 1987). L1s contain a 5' and a 3' untranslated region UTR as well as 3 open reading frames (ORFs) (Figure 1.1B) (Denli et al., 2015; A. F. Scott et al., 1987). The 5' UTR contains an antisense promoter (Speek, 2001) that was more recently discovered to drive ORF0 expression (Denli et al., 2015). ORF0 might play a role in enhancing L1 retrotransposition (Denli et al., 2015). The 5' UTR also contains a sense canonical promoter (Swergold, 1990) that drives the expression of ORF1 and ORF2. ORF1 encodes an L1 RNA binding protein (Kolosha & Martin, 1997) necessary for effective retrotransposition (Moran et al., 1996), while ORF2 encodes a protein with endonuclease (Feng, Moran, Kazazian, & Boeke, 1996) and reverse transcriptase activity (Mathias, Scott, Kazazian, Boeke, & Gabriel, 1991) that are also required for retrotransposition (Moran et al., 1996) as well as a conserved cysteine-rich motif (Fanning & Singer, 1987). Following the 3' UTR, L1 elements contain a poly(A) tract tail essential for retrotransposition (Doucet, Wilusz, Miyoshi, Liu, & Moran, 2015). However, ~95% of L1s in the human genome are truncated, and the majority are 5' truncated (Szak et al., 2002). Due to these truncations and other inactivating mutations, it is estimated that only 80-100 L1s in the human genome are intact and able to retrotranspose and only a small subset of these are active L1s that have contributed to retrotransposition in humans (Brouha et al., 2003).

Non-autonomous SINEs include Alu elements. These ~280 bp TEs comprise ~10% of the human genome with more than 1 million copies and are the most prevalent and successful TEs (Kazazian & Moran, 2017; A. F. Smit, 1996). Alu elements do not have ORFs and require the L1 ORF2 for their retrotransposition (Dewannieux, Esnault, & Heidmann, 2003). These TEs likewise have a poly(A) tail that is also essential for retrotransposition (Dewannieux et al., 2003). Alus are composed of two monomers derived from the 7SL RNA gene that are connected by an adenosine-rich linker (Figure 1.1B) (Deininger, 2011; Kazazian & Moran, 2017; Kojima, 2011; Ullu & Tschudi, 1984). Alu elements contain an RNA polymerase III promoter but do not contain a transcription terminator and instead may use terminator sequences at downstream regions to end transcription (Deininger, 2011). Similar to L1s, only a few Alu elements in the human genome remain active and most have gained inactivating mutations (Deininger, 2011).

SINE-VNTR-Alu (SVA) elements are also classified as non-autonomous SINEs. Named after its main components, SVAs contain a CCCTCT simple repeat region, an Alu-like domain from two antisense Alu fragments, a GC rich variable number of tandem repeat (VNTR) sequence, a SINE region (SINE-R) that contains a section derived from a HERV-K element and an *env* gene, as well as a poly(A) tail (Figure 1.1) (Hallmayer et al., 2011; Kwon et al., 2013; H. Wang et al., 2005). Their length is variable and can go up to 4,000 bp, but a canonical full-length insertion is ~2 kb (Hancks, Ewing, Chen, Tokunaga, & Kazazian, 2009; Hancks & Kazazian, 2010; H. Wang et al., 2005). SVAs are active in humans and also use the L1 machinery for their mobilization, requiring both the L1 ORF1 and ORF2 proteins (Hancks, Goodier, Mandal, Cheung, & Kazazian, 2011; Raiz et al., 2012).

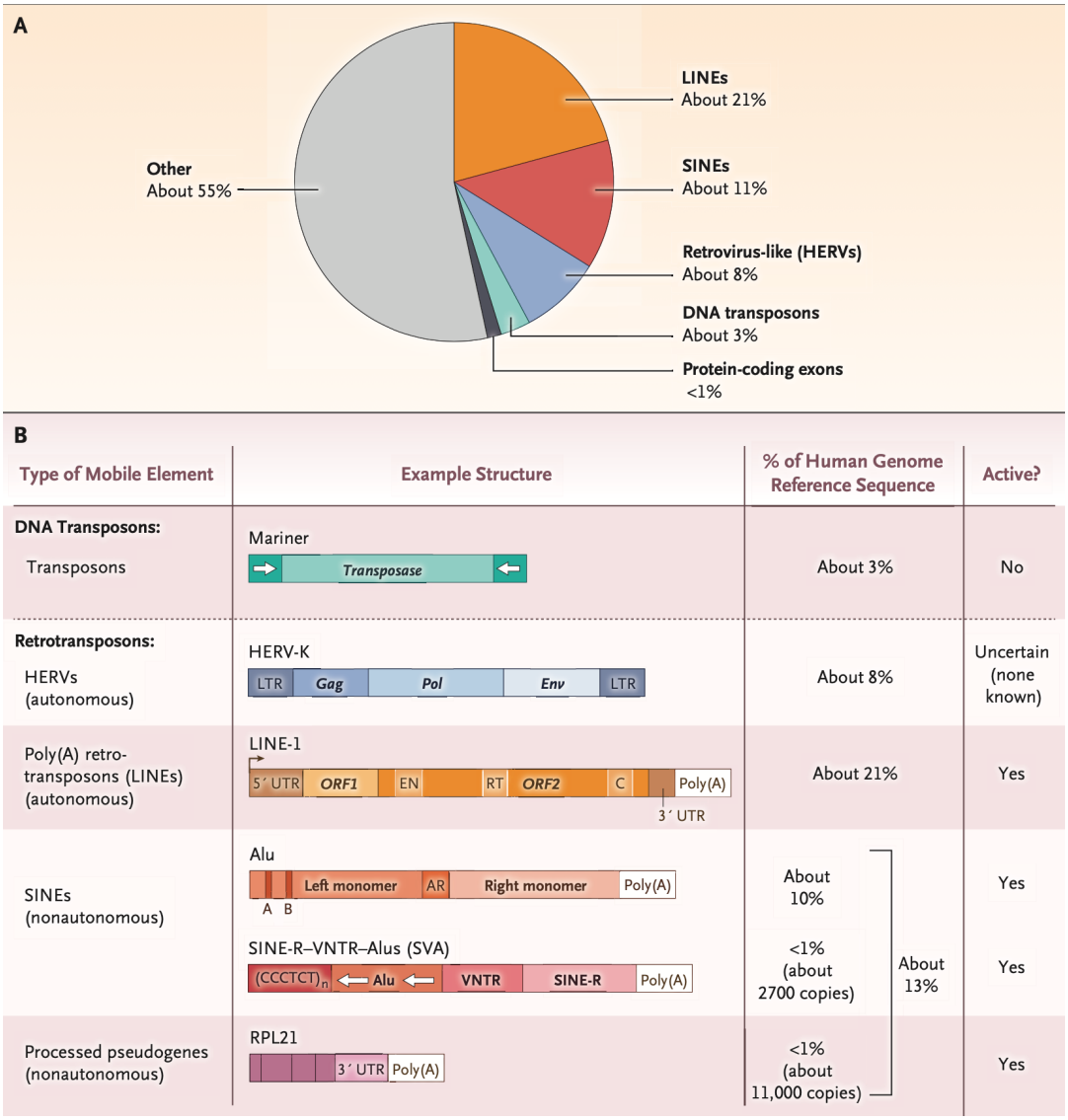

Figure 1.1 Transposable elements in humans.

Composition of the human genome. TEs account for almost 45% of the genome while protein-coding exons account for < 1%. B) Structure of TEs in the human genome (see main text for a detailed description). The two main types of mobile elements are DNA transposons and retrotransposons. Autonomous elements encode the necessary proteins for their retrotransposition, while non-autonomous TEs use the proteins encoded by autonomous elements such as L1s. Processed pseudogenes are mRNA sequences that were reverse transcribed by the L1 machinery and inserted in the genome without introns (Kazazian, 2014). LTR: long terminal repeat, UTR: untranslated region, ORF: open reading frame, EN: endonuclease, RT: reverse transcriptase, A and B represent Alu's promoter region, AR: adenosine-rich, VNTR: variable number of tandem repeat. This figure was adapted from (Kazazian & Moran, 2017).

Our genome contains different subfamilies of retrotransposons (A. F. Smit, Toth, Riggs, & Jurka, 1995). L1s are classified from PA1-PA17, and the oldest of these subfamilies were active in mammals millions of years ago before the origin of primates (A. F. Smit et al., 1995; Sultana et al., 2019). The currently active L1s in our genome belonging to the PA1 subfamily are a subset of the human-specific L1s (L1HS) (Richardson et al., 2015). We can distinguish this subfamily through specific mutations unique to L1HS elements (Richardson et al., 2015). Alu TEs are also subdivided into the following subfamilies: AluY, AluS, and AluJ (A. Smit, Hubley, R & Green, P. , 2013-2015). The youngest and most active elements belong to the AluY subfamily, and within this group, AluYa5 elements are the most active followed by AluYb8 (Konkel et al., 2015). SVAs are the youngest retrotransposons in humans and are divided into 6 subfamilies. The SVA_A to SVA_D subfamilies expanded before the divergence of humans and gorillas, chimpanzees, and orangutans, while SVA_E and SVA_F are the youngest and most active subfamilies in humans (Levy, Knisbacher, Levanon, & Havlin, 2017).

Mechanisms of retrotransposition

The poly(A) sequence of non-LTR retrotransposons is essential for their mobilization through a process called target-primed reverse transcription (TPRT) (Dewannieux et al., 2003; Hancks & Kazazian, 2010; Moran et al., 1996) (Figure 1.2). In L1s, L1 mRNA first binds L1 ORF1 and ORF2 proteins to form the ribonucleoprotein particle in the cytoplasm (Hohjoh & Singer, 1996; Kolosha & Martin, 1997; Kulpa & Moran, 2005). In the nucleus, the ORF2 endonuclease then cleaves DNA at a degenerate 5'-TTAAAA site (Feng et al., 1996; Jurka, 1997). The L1 mRNA then anneals

to the exposed poly(T) sequence and the ORF2 protein reverse transcribes the L1 mRNA, using it as a template (Kulpa & Moran, 2005). The second strand is cleaved near the initial cleavage, presumably by ORF2 (Faulkner & Garcia-Perez, 2017), although the mechanism for this is still being resolved (Khadgi, Govindaraju, & Christensen, 2019). The second strand of DNA is then synthesized (also through an unresolved ORF2 mediated mechanism (Kazazian & Moran, 2017)) and as the DNA is filled, a target site duplication (TSD) of the region surrounding the insertion is created (Pizarro & Cristofari, 2016; Szak et al., 2002). As mentioned previously, Alu and SVA insertions hijack the L1 ORF2 protein, and in the case of SVA elements the ORF1 protein as well, and also retrotranspose via TPRT (Dewannieux et al., 2003; Hancks et al., 2011; Kazazian & Moran, 2017; Raiz et al., 2012). This process creates a hallmark sequence consisting of two TSD surrounding the TE sequence along with a poly(A) tail that can then be used to identify these insertions computationally (Faulkner & Garcia-Perez, 2017).

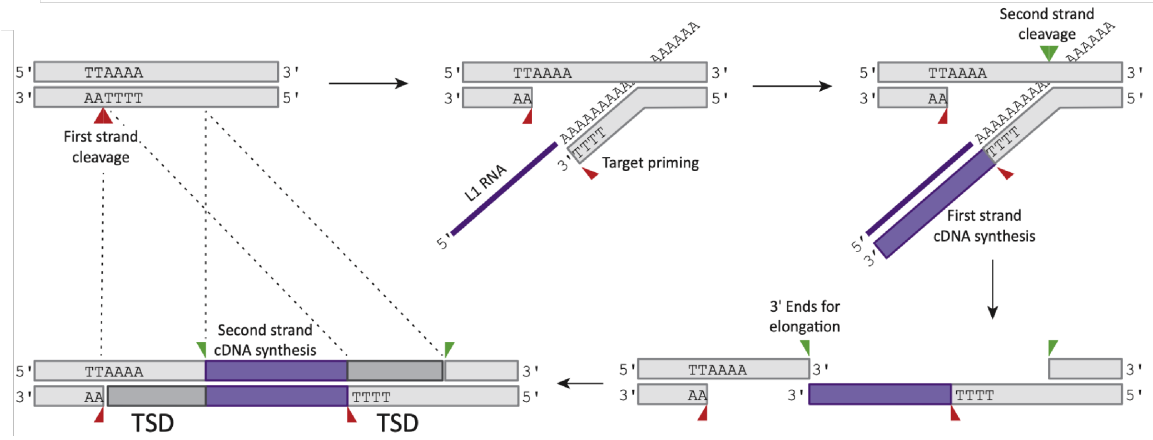

Figure 1.2 Target-primed reverse transcription.

L1, Alu, and SVA retrotransposition is mediated by a mechanism called target-primed reverse transcription. Here, the red arrow represents the first strand cleavage, and the green arrow the second strand cleavage. TSD: target site duplication. This figure was adapted from (Faulkner & Garcia-Perez, 2017).

Transposable elements in the human genome

TEs contribute to genomic diversity, can impact gene expression, promote structural variants or recombination, and have played an important role in mammalian evolution (Platt et al., 2018). TE sequences contain regulatory elements as well as transcription factor binding sites (Ali, Han, & Liang, 2021; Bourque et al., 2008; Platt et al., 2018; Polak & Domany, 2006). Transcription factor binding sites in different species including humans have been subject to evolutionary selection, can be lineage-specific, and likely regulate the expression of nearby genes (Bourque et al., 2008). In human embryonic stem cells, the binding of certain transcription factors in HERVH elements causes chimeric transcripts of HERVH and downstream long-noncoding RNAs (J. Wang et al., 2014). Additionally, via binding of the transcription factor OCT4, HERVH gains enhancer activity in human embryonic stem cells, and knockdown of HERVH in these cells drastically affects cell identity and morphology (Lu et al., 2014).

Because TEs are so prevalent in the human genome, their repetitive sequences can facilitate structural variants like deletions, duplications, inversions, and can promote non-allelic homologous recombination (Belancio, Deininger, & Roy-Engel, 2009; Deininger, 2011). Using comparative genomics between humans and chimpanzees, almost 500 Alu recombination-mediated deletions (Sen et al., 2006) and more than 70 L1 recombination-mediated deletions have been identified (Han et al., 2008), deleting ~400 kb and ~450 kb respectively. Inverted Alu and L1 elements in the genome have also induced at least 49 human and chimpanzee specific retrotransposon recombination-mediated inversions (Lee, Han, Meyer, Kim, & Batzer, 2008). These studies imply that not only do TEs contribute to genomic rearrangements, but they have also impacted the divergence between humans and other species.

In what has been described as an “arms race” between TEs and their host genome (Platt et al., 2018) (Goodier, 2016), a careful interplay between TE expression and repression takes place in our genomes. Humans and other species have evolved to restrict uncontrolled retrotransposition and expression as TEs have evolved to escape repression (Bourque et al., 2018). Several mechanisms have been described for TE repression in humans and include restriction factors in the cytoplasm, epigenetic silencing in the nucleus, DNA repair enzymes, and RNA silencing proteins and pathways (Goodier, 2016). The apolipoprotein B mRNA editing enzyme catalytic polypeptide 3 (APOBEC3) proteins are a family of proteins that inhibit retroviruses and retrotransposition in humans and other mammals (Cullen, 2006; Kinomoto et al., 2007; Koito & Ikeda, 2013; Schumann, 2007). There are at least seven different APOBEC3 proteins in humans, and they inhibit retrotransposition through different mechanisms such as via deamination of L1 cDNA during reverse transcription (Richardson, Narvaiza, Planegger, Weitzman, & Moran, 2014) or by sequestering Alu RNAs in the cytoplasm (Chiu et al., 2006). DNA methylation is another important mechanism for silencing TE expression in humans (Goodier, 2016; Thayer, Singer, & Fanning, 1993). CpG sites in L1 promoters are generally methylated, repressing L1 expression (Hata & Sakaki, 1997; Yu, Zingler, Schumann, & Stratling, 2001), a process that can be variable between individuals (Singer et al., 2012) and is disrupted in certain diseases like cancer (Phokaew, Kowudtitham, Subbalekha, Shuangshoti, & Mutirangura, 2008). Kruppel-associated box zinc finger proteins (KRAB-ZFPs) are a family of transcription factors that have evolved along with TEs and repress their expression by inducing DNA methylation and the formation of repressive heterochromatin (Thomas & Schneider, 2011; Yang, Wang, & Macfarlan, 2017). KRAB-ZFPs are expressed during embryogenesis, but also continue to regulate TEs in adult tissues and in the adult brain (Turelli et al., 2020). Piwi-interacting RNAs (piRNAs) are non-coding RNAs involved in TE silencing in animals

(Goriaux, Theron, Brasset, & Vaury, 2014). Although the piRNA pathway has mostly been studied in other species like *Drosophila*, piRNA expression has also been detected in human testis and ovaries (Ha et al., 2014; Williams et al., 2015).

Although TE expression is highly regulated, these elements including LINE, SINE, and LTR retrotransposons are expressed in humans in a tissue specific manner (Faulkner et al., 2009). The detection of TE expression requires specialized algorithms and pipelines, since these elements are very repetitive in the genome and pervasive transcription of mRNA from regions containing TEs can influence the quantification of these elements (Lanciano & Cristofari, 2020). In fact, more than 99% of L1 sequences detected from RNA are not from transcription initiated from L1 promoters (Deininger et al., 2017). Deregulation of TE expression has been described in human disease, in particular in cancer where increased L1 expression and protein levels have been detected in tumors (Asch et al., 1996; De Luca et al., 2016; Rodic et al., 2014). Increased ERV, L1, and SINE expression has also been observed in cells and tissues from individuals with Alzheimer's disease (AD) (Guo et al., 2018; He et al., 2021). TDP-43 pathology in amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration has also been associated with de-repression of TEs (Li, Jin, Prazak, Hammell, & Dubnau, 2012) and about 20% of ALS patients exhibit elevated levels of TE expression of ERV, L1, and SINEs (Tam et al., 2019).

Transposable elements as a mutagenic cause of disease

TE insertions may lead to genomic diversity, but deleterious insertions can disrupt gene function and cause disease (Prak & Kazazian, 2000). Since the first discovery of a disease causing L1 insertion, an insertion in the Factor VII gene in

hemophilia A (Kazazian et al., 1988), at least 124 other disease causing insertions have been reported and can affect coding sequences, splicing, and protein localization by mobilizing into exons and introns (Hancks & Kazazian, 2016). Examples include *de novo* Alu and LINE-1 retrotransposition events that disrupt splicing of *NF1*, cause autosomal dominant Neurofibromatosis Type I, and account for 0.4% of *NF1* mutations (Wimmer, Callens, Wernstedt, & Messiaen, 2011). An ancestral founder SVA insertion causes Fukuyama muscular dystrophy, a common recessive disorder in Japan by inserting into the 3'UTR of *FKTN*, which leads to protein mislocalization (Taniguchi-Ikeda et al., 2011). In 0.04% of patients with developmental disorders causal *de novo* exonic Alu or L1 insertions were identified (Gardner et al., 2019). Because TEs are relatively large, intronic insertions may lead to disease, as was observed with the insertion of an inherited autosomal dominant X-linked full-length LINE-1 element into the *RP2* gene in a patient with retinitis pigmentosa (Schwahn et al., 1998). A landmark study identified a deep intronic compound-heterozygous SVA insertion causing exon-trapping in a child with Batten disease, resulting in the development of a personalized antisense-oligonucleotide drug to correct the splicing defect (Kim et al., 2019). Thus, the identification of inherited and *de novo* TEs is important for increasing genetic diagnoses but also creates the promise of developing novel therapeutics for specific mutant alleles.

Somatic retrotransposition occurs at different rates in different types of cancer, and may contribute to hereditary cancer, its progression, and even initiation in certain cases (Hancks & Kazazian, 2016; Miki et al., 1992; Rodic et al., 2015; E. C. Scott et al., 2016; Shukla et al., 2013). Increased retrotransposition occurs particularly in cancers with DNA repair defects, including early gastrointestinal cancers (Ewing et al., 2015), Barrett's esophagus, and esophageal adenocarcimona (Casson et al., 2005; Doucet-O'Hare et al., 2015) but are rare in brain tumors (Achanta et al., 2016; E. Lee et al.,

2012). L1 insertions occur in genes that are frequently mutated in cancer (E. Lee et al., 2012). An example of a somatic TE directly causing cancer is a full-length L1 insertion detected in the *APC* gene, causing colorectal cancer (E. C. Scott et al., 2016). This suggests that somatic retrotransposition may also contribute to other diseases and disorders in humans.

Somatic mutations in the brain

Overview

Although cells in the human body all descend from a zygote with DNA inherited from the two parental germline cells, they accumulate new mutations throughout development, giving rise to a mosaic genetic landscape. These somatic mutations can occur early in development and be present in entire clones of cells derived from the initial mutant cell or appear after cell division and be present in a single cell (Rodin & Walsh, 2018). The genetic causes of neurodevelopmental and neurological disorders have been traditionally studied from the perspective of inherited or *de novo* mutations. However, there is now a growing body of literature that recognizes the role of somatic mutations in these disorders (D'Gama & Walsh, 2018; Freed, Stevens, & Pevsner, 2014). In this section, we will describe the classification of somatic mutations and existing methods for their detection. We will then discuss their involvement in aging and neurological and neurodevelopmental disorders.

Somatic mutations

First, it is important to define and distinguish somatic mutations from *de novo* and inherited mutations (Figure 1.3). Inherited mutations are mutations that are transmitted from the parent to their progeny via germ cells. This can occur in various ways: the mutation could be present in every cell in one or both parents and the child would inherit this mutation in a heterozygous or homozygous state respectively. The parent could also have a somatic mutation in a mosaic state in cells in the body, including germline cells, and the child would have this mutation in every cell as a heterozygous mutation. Additionally, a parent could have a somatic mutation only in germline cells that the child would then inherit also as a heterozygous mutation in all of their cells (Figure 1.3A) (Freed et al., 2014). *De novo* mutations are only detected in the child and usually arise during gametogenesis in the parent or in very early development (Figure 1.3B) (D'Gama & Walsh, 2018; Freed et al., 2014). These mutations are heterozygous and are present in every cell in the child. Somatic mutations are postzygotic mutations that occur during or after development in an individual's lifetime and are only present in a subset of cells (Figure 1.3C) (D'Gama & Walsh, 2018; Freed et al., 2014). These mutations are heterozygous in the cells carrying the mutation in the child, but the alternate allele frequency (AAF) detected in tissues is < 50% compared to inherited or *de novo* mutations that display an AAF = 50% (D'Gama & Walsh, 2018). Inherited and *de novo* mutations can be difficult to distinguish in clinical analyses because usually only blood samples from family members are sequenced. Because we have not sequenced parental germline cells in this study, we will consider any mutation detected only in the child as a *de novo* mutation, even though they might be parental mosaic mutations and inherited. These distinctions are important because *de novo* rates described in this study or previous literature may be inflated from inherited or somatic mutations.

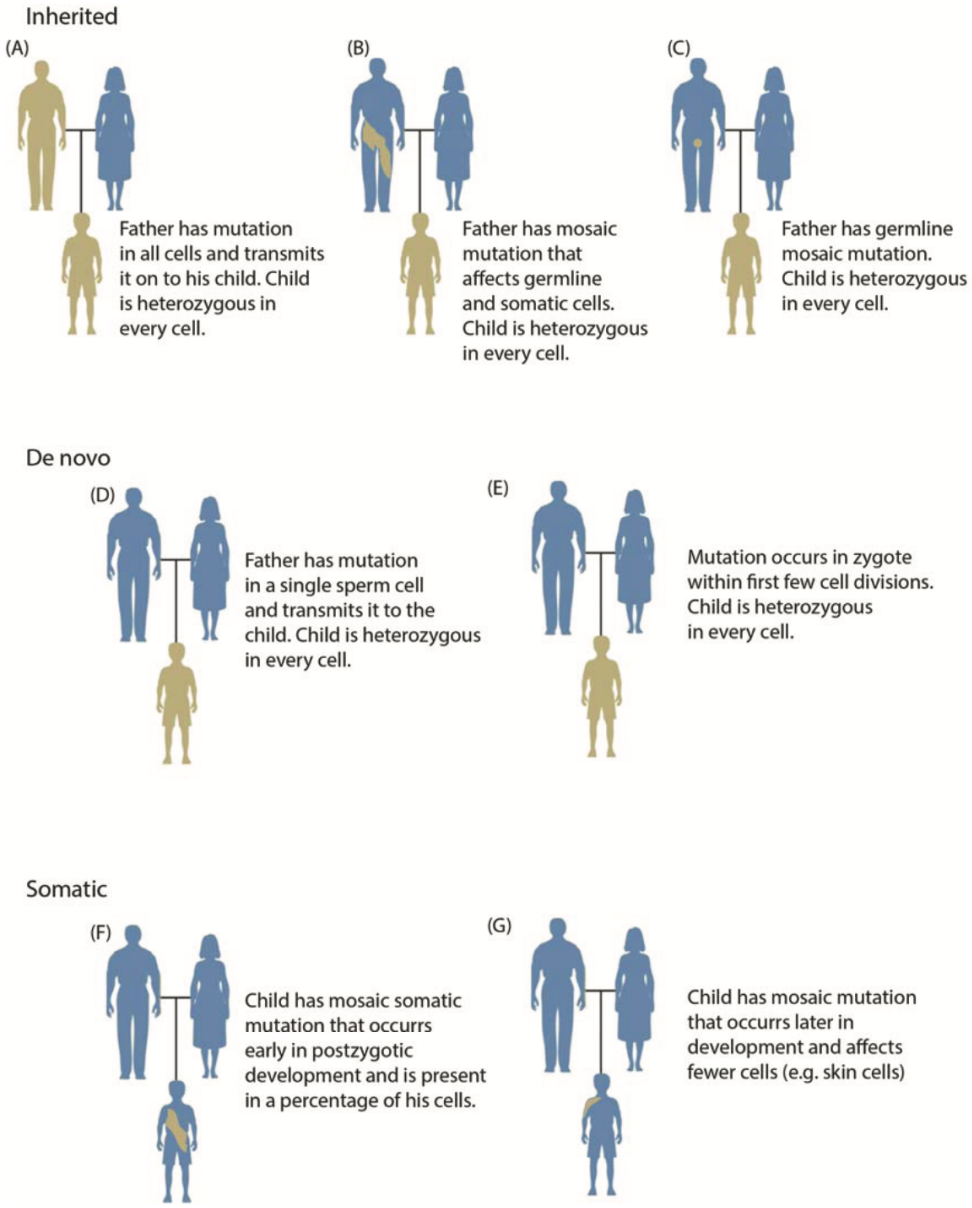

Figure 1.3 Classification of inherited, *de novo*, and somatic mutations.

A) Inherited mutations are transmitted from a parent to their child. B) *De novo* mutations are only detected in the child and can arise during gametogenesis in a parent or the first cell divisions of the zygote. C) Somatic mutations occur post-zygotically and are only present in a subset of the child's cells. This image was adapted from (Freed et al., 2014).

Studying somatic mutations is important because the types of mutations observed in this state may differ from *de novo* and inherited mutations present in every cell. The latter are subject to greater selective pressure because they will affect gene

function in every tissue in which this gene is expressed and will be segregated to future generations, whereas the impact of somatic mutations will depend on not only the type of mutation and functional consequence, but also the timing and location of expression. For example, although autosomal trisomy generally occurs in the maternal gamete during meiosis, *de novo* trisomy in chromosome 8 is usually developmentally lethal and is more frequently observed in a mosaic state (Karadima et al., 1998; Robinson et al., 1995). Additional examples of disorders that are only observed in a mosaic state include Proteus, Sturge–Weber, and McCune–Albright syndromes (Moog, Felbor, Has, & Zirn, 2020).

Somatic mutations may lead to milder phenotypes than germline inherited or *de novo* mutations in the same gene (D'Gama & Walsh, 2018). An example of this is mutations in *FLN1*, where somatic mutations cause a milder presentation of X-linked periventricular nodular heterotopia compared to inherited and *de novo* mutations (D'Gama & Walsh, 2018; Guerrini et al., 2004; Parrini, Mei, Wright, Dorn, & Guerrini, 2004). The timing of when a mutation arises in cell development is important since the proportion of cells carrying a somatic mutation may impact its phenotypic consequences. Mutations to genes in the mammalian target of rapamycin (mTOR) pathway can cause focal cortical dysplasia (FCS) or hemimegalencephaly (HME), with the former displaying small malformations in the brain's cortex and the latter large abnormal brain malformations affecting entire hemispheres in certain cases (Blumcke et al., 2011; D'Gama et al., 2017; Poduri et al., 2012). Somatic mutations in the mTOR pathway with a low AAF tend to cause FCD, whereas cases with a higher AAF generally present HME (Figure 1.4) (D'Gama et al., 2017). As shown in Figure 1.4, there is a continuum of the phenotype observed with the corresponding AAF of the somatic mutation, suggesting

that the phenotypic consequences of somatic mutations are not only dependent on AAF, but involve the location, cell types, affected genes, and type of mutation.

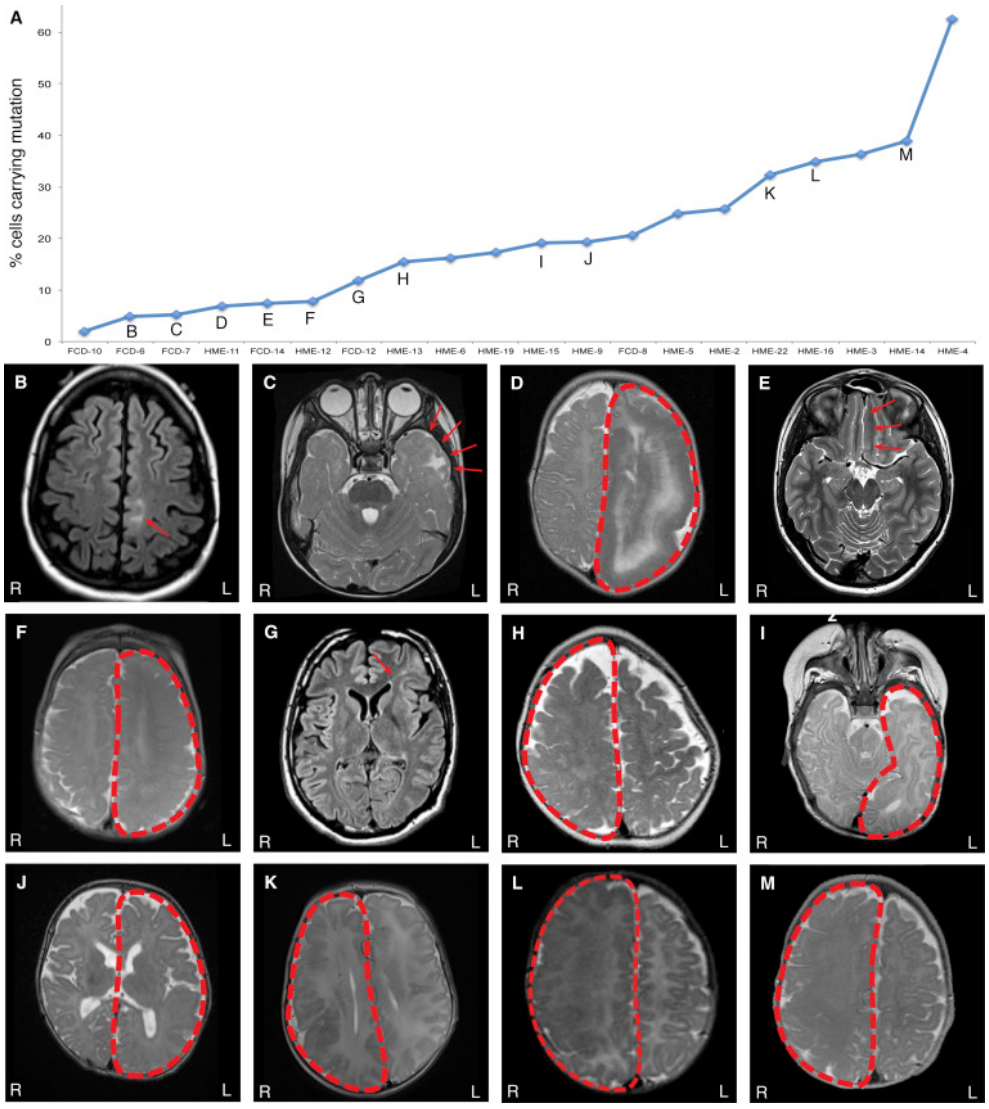

Figure 1.4 Somatic mutations in the mTOR pathway cause different forms of cortical malformations and dysplasia depending on their timing, location, and prevalence.

MRI from individuals with somatic mutations in different genes of the mTOR pathway. FCD-6 (B), FCD-7 (C), HME-11 (D), FCD-14 (E), HME-12 (F), FCD-12 (G), HME-13 (H), HME-15 (I), HME-9 (J), HME-22 (K), HME-16 (L), and HME-14 (M). This figure was adapted from (D'Gama et al., 2017).

Methodologies for the detection of somatic mutations in the brain

Several methods are currently available for the detection and quantification of a diverse array of somatic mutations including SNVs, TEs, SVs, and copy number variants (CNVs). Early studies identified mosaic aneuploidies and copy number changes in neurodevelopmental and neurological disorders such as ASD, schizophrenia, and AD, using various cytogenetic methods including microscopy, fluorescent in situ hybridization, array comparative genomic hybridization, (Iourov, Vorsanova, & Yurov, 2006, 2008; Vorsanova, Yurov, Soloviev, & Iourov, 2010; Yurov et al., 2007), and SNP microarrays (Baugher, Baugher, Shirley, & Pevsner, 2013; Freed et al., 2014).

Next-generation sequencing has pushed forward the development of novel methods and approaches for detecting somatic variants with increased sensitivity, to the point where we are now able to detect single SNVs or TE insertions in individual cells. This can be applied to bulk tissues, where multiple cell types are represented, pools of sorted cells, clonally expanded single cells, or single cells (D'Gama & Walsh, 2018). Somatic mutation sequencing methodologies include Sanger sequencing, targeted panels and targeted sequencing including whole-exome sequencing (WES), and whole-genome sequencing (WGS) (D'Gama & Walsh, 2018; Dou, Gold, Luquette, & Park, 2018). Each of these methods and DNA sources has advantages and disadvantages. For example, targeted sequencing from bulk tissues can be applied to many samples in one study, can be simple to implement, and can provide AAF information. Somatic SNVs have been identified in thousands of blood WES samples from families with ASD-affected individuals (Dou et al., 2017; Freed & Pevsner, 2016; Krupp et al., 2017; Lim et al., 2017). However, the coverage of samples sequenced here (~60x) is not sufficiently high to detect and quantify very low AAF mutations, only mutations in the regions targeted (exons in this case) and present in the blood can be identified, and the analysis

required for this data is not trivial. Meanwhile, ultra-deep bulk WGS (~250x) from the prefrontal cortex of 74 individuals including 59 ASD cases identified somatic SNVs at lower AAFs in the brain but was limited by the smaller sample size and high cost for deep WGS (Rodin et al., 2021). These methodologies are under active development, and researchers in the field are working on implementing standardized practices for the detection of somatic mutations from brain tissues (Y. Wang et al., 2021). The advantages and disadvantages of using some of these approaches for the detection of somatic TEs will be discussed further in Chapter 3.

Single-cell WGS enables the detection of somatic mutations at the highest resolution and at resolutions that are currently technically challenging using other methods and DNA sources. Many types of mutations can be detected from this type of data and experimental validations are generally straightforward (D'Gama & Walsh, 2018). In this work, we focus on detecting somatic TEIs in postmitotic neurons that occur in adulthood. Single-cell sequencing is an advantageous method for this purpose since we expect these somatic mutations to be present at very low AAFs in tissue (Miller, Reed, & Walsh, 2021). However, single-cell sequencing faces two major challenges. The first one is isolating individual cells or neurons and the second is amplifying DNA at a sufficient quantity in a precise and uniform manner.

Current methods for isolating single cells include fluorescence-activated cell sorting (FACS), laser capture microdissection (LCM), and microfluidics (Hu, Zhang, Xin, & Deng, 2016). Brain cells are variable in morphology and size, so methods for isolating these cells must be able to distinguish and select the population of interest. LCM involves isolating cells directly from microscopic slides. This is a useful method for neuroscience research because specific populations of cells can be identified morphologically without a need for cell-type-specific antibodies (Morris & Mehta, 2018).

However, this method is technically challenging, can result in DNA damage, and is not scalable. FACS and microfluidics are more high-throughput methods. FACS is routinely used in single-cell DNA sequencing analysis since the use of fluorophore-conjugated monoclonal antibodies allows researchers to obtain pure populations of specific cell types, although a limitation of this method is that many cells are required and antibodies may not be available for all cell types of interest (Hu et al., 2016). It is common practice to use nuclear antibodies and to sequence single nuclei in neuroscience since neurons have a complex morphological and fragile structure. Microfluidic systems require lower starting volumes, are affordable, and highly scalable (Hu et al., 2016). These systems isolate cells based on properties like size and may require using them in combination with FACS to obtain specific brain cell type populations.

Single-cell DNA amplification methods may introduce artifacts including chimeras, which can occasionally result in false-positive somatic candidates, and therefore require meticulous bioinformatical analyses and experimental validations (Cai et al., 2014, 2015; Evrony, Lee, Park, & Walsh, 2016; Lasken & Stockwell, 2007). Additionally, since there are only two copies of DNA in individual cells, allelic dropout can lead to uneven amplification of the genome (Cai et al., 2014, 2015; Evrony et al., 2016). Popular methods for single-cell DNA sequencing includes PCR based methods such as degenerate oligonucleotide primed PCR (DOP-PCR) (Telenius et al., 1992), isothermal amplification methods like multiple displacement amplification (MDA) (Dean et al., 2002), or hybrid methods with both isothermal preamplification and additional PCR amplification such as multiple annealing and looping based amplification cycles (MALBAC) (Gawad, Koh, & Quake, 2016; Hu et al., 2016; Xing, Tan, Chang, Li, & Xie, 2021; Zong, Lu, Chapman, & Xie, 2012). MDA has a higher coverage and decreased false-positive rates, while MALBAC and DOP-PCR have more uniform coverage at

larger scales (Evrony et al., 2015; Gawad et al., 2016; Hou et al., 2015; Huang, Ma, Chapman, Lu, & Xie, 2015). New methods have recently been developed to improve amplification coverage such as multiplexed end-tagging amplification of complementary strands (META-CS) (Xing et al., 2021), NanoSeq (Abascal et al., 2021), and primary template-directed amplification (PTA) (Gonzalez et al., 2021). We will discuss the possibility of using these methods for the detection of somatic TEs in the Discussion section of this thesis (Chapter 4).

In this study and previous work from our group, we have selected MDA to amplify single neurons given its lower rates of false positives and artifacts, its overall even coverage at small scales, and large amplicons produced (Cai et al., 2014; Evrony et al., 2015). In this method, DNA is denatured, random primers bind to DNA, and amplification is then driven by a Phi29 DNA polymerase with high processivity in an exponential manner, yielding 20–30 μg of DNA (Dean et al., 2002). With MDA, a coverage of ~40x can be obtained and 98% ± 0.5% of the genome can be sequenced at a read depth ≥ 1x or 81% ± 2% can be sequenced at a depth ≥ 10x in neurons (Evrony et al., 2015). Limitations of this method include allelic dropout, GC-sequence bias, and the introduction of chimeric artifacts (Cai et al., 2014; Evrony et al., 2015; Lasken & Stockwell, 2007). Despite these limitations, our group and others have been able to successfully detect and quantify somatic SNVs (Lodato et al., 2018; Lodato et al., 2015), CNVs (Cai et al., 2014; Ning et al., 2015), and TEIs (Erwin et al., 2016; Evrony et al., 2015) in MDA amplified single neurons obtained from post-mortem tissue.

Somatic mutations in aging and disease in the brain

The potential deleterious consequences of somatic mutations in humans are clear in cancer, however, our understanding of the impact of these mutations in neurodevelopmental and neurological disorders is still quite incomplete. Although there are several reports of somatic mutations including SNVs and CNVs in brain malformations, epilepsy, ASD, intellectual disability (ID), and other neurological disorders (D'Gama & Walsh, 2018; Maury & Walsh, 2021; Rodin & Walsh, 2018), the rates of somatic mutations, and especially TEIs, in neurons and brain cells in neurological disorders is mostly unknown. The greatest advances in this field have stemmed from the research of brain malformations. As described previously, multiple studies have reported that somatic mutations in the mTOR pathway including SNVs (D'Gama et al., 2017; J. H. Lee et al., 2012; Poduri et al., 2012) and CNVs (Cai et al., 2014; Poduri et al., 2012) can cause HME or FCD (Blumcke et al., 2011; D'Gama et al., 2017; J. H. Lee et al., 2012; Poduri et al., 2012) (Figure 1.4). Somatic mutations contributing to disease in other pathways including somatic mutations in *GNAQ*, which were detected in 88% of individuals with Sturge-Weber syndrome, a neurocutaneous disorder that can cause seizures, ID, and brain vascular abnormalities (McConnell et al., 2017; Shirley et al., 2013).

Single-cell studies have been instrumental in determining somatic mutation rates in the brain. Fetal brains at around 15-21 weeks of gestation have already accumulated 200-400 somatic SNVs per cell (Bae et al., 2018), and this increase continues in human neurons with aging (Lodato et al., 2018; Xing et al., 2021). This occurs at a rate of ~16-23 somatic SNVs per year per cell for neurons in the pre-frontal cortex and ~40 somatic SNVs per year in neurons from the dentate gyrus in the hippocampus (Lodato et al., 2018; Xing et al., 2021). On the other hand, large CNVs $\geq$ 2 Mb not only do not increase

but appear to be anti-correlated with age in neurons (Chronister et al., 2019). Cockayne syndrome (CS) and Xeroderma pigmentosum (XP) are syndromes that display early-onset neurodegeneration and are caused by mutations in genes involved in the transcription-coupled repair and nucleotide excision repair pathways respectively (Bertola et al., 2006; Cleaver, 2005). In single neurons from affected individuals, somatic SNVs are increased by a ~2.3 fold in CS and a ~2.5 in XP (Lodato et al., 2018). This suggests that we might observe higher rates of somatic mutations in other neurodegenerative disorders (Miller et al., 2021). Researchers have hypothesized that a subset of AD cases may be due to somatic mutations in familial AD genes, given that many cases are sporadic and of an unknown genetic cause (Guerreiro, Gustafson, & Hardy, 2012; Miller et al., 2021). Yet, there is currently not enough data supporting this hypothesis, although previous analyses have been constrained by the AAF sensitivity limits from deep targeted sequencing (Sala Frigerio et al., 2015) or small sample sizes with single cells (Abascal et al., 2021; Miller et al., 2021). Most single-cell analyses of somatic mutations in neurological disorders have focused on SNVs or CNVs, but in Chapter 3 we will focus on what is currently known about somatic retrotransposition in the brain.

Conclusion and aims

Here, we have demonstrated that TEIs impact our genome in many ways. An important consequence of retrotransposition events is that these can alter gene function and cause disease (Hancks & Kazazian, 2016). In this work, we explore how TEIs impact the genome in neurological and neurodevelopmental disorders by harnessing state-of-the-art computational and experimental methodologies. Our first aim, which is

described in Chapter 2, was to assess the significance of *de novo* TEIs in ASD. We hypothesized that retrotransposition rates would be the same between ASD-affected and unaffected individuals, but that affected individuals would have more *de novo* TEIs in genes that are loss-of-function intolerant, are known to be ASD genes, or are involved in brain development and function. We tested this hypothesis by detecting *de novo* retrotransposition events from WGS data in ASD families. We then compared the rates of *de novo* TEIs in cases and controls and validated the TEIs detected. Our second aim, described in Chapter 3, was to study the rates of somatic TEIs in single neurons and glia throughout aging and neurodegeneration. We hypothesized that cells obtained from aging individuals or individuals with neurodegeneration would have increased somatic retrotransposition rates. We tested this by analyzing single-cell WGS data obtained from post-mortem human brain tissue.

TEs are not routinely screened for in the clinic in ASD cases and the impact of *de novo* coding and non-coding TEIs in ASD is poorly understood. Understanding the link between TEIs and ASD could help increase diagnostic rates. Additionally, studying retrotransposition rates in large WGS family cohorts will allow us to determine *de novo* rates with greater precision than has been possible using smaller cohorts. This project also provided an important opportunity to advance our understanding of polymorphic TEIs in the genome. Additionally, animal models have suggested that TE expression and insertions are increased in the aging brain (Li et al., 2013) and in neurodegeneration (Guo et al., 2018; Li et al., 2012; Sun, Samimi, Gamez, Zare, & Frost, 2018) and have even suggested that using antiretroviral therapy to decrease retrotransposition reduces neurodegeneration in these models (Sun et al., 2018). It is hoped that this research will contribute to a deeper understanding of the contribution of somatic TEIs to aging and disease in the human brain.

References

Abascal, F., Harvey, L. M. R., Mitchell, E., Lawson, A. R. J., Lensing, S. V., Ellis, P., . . . Martincorena, I. (2021). Somatic mutation landscapes at single-molecule resolution. *Nature, 593*(7859), 405-410. doi:10.1038/s41586-021-03477-4

Achanta, P., Steranka, J. P., Tang, Z., Rodic, N., Sharma, R., Yang, W. R., . . . Burns, K. H. (2016). Somatic retrotransposition is infrequent in glioblastomas. *Mob DNA, 7*, 22. doi:10.1186/s13100-016-0077-5

Ali, A., Han, K., & Liang, P. (2021). Role of Transposable Elements in Gene Regulation in the Human Genome. *Life (Basel), 11*(2). doi:10.3390/life11020118

Asch, H. L., Eliacin, E., Fanning, T. G., Connolly, J. L., Bratthauer, G., & Asch, B. B. (1996). Comparative expression of the LINE-1 p40 protein in human breast carcinomas and normal breast tissues. *Oncol Res, 8*(6), 239-247. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8895199

Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., . . . Vaccarino, F. M. (2018). Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science, 359*(6375), 550-555. doi:10.1126/science.aan8690

Baugher, J. D., Baugher, B. D., Shirley, M. D., & Pevsner, J. (2013). Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method. *BMC Genomics, 14*, 367. doi:10.1186/1471-2164-14-367

Belancio, V. P., Deininger, P. L., & Roy-Engel, A. M. (2009). LINE dancing in the human genome: transposable elements and disease. *Genome Med, 1*(10), 97. doi:10.1186/gm97

Bertola, D. R., Cao, H., Albano, L. M. J., Oliveira, D. P., Kok, F., Marques-Dias, M. J., . . . Hegele, R. A. (2006). Cockayne syndrome type A: novel mutations in eight typical patients. *J Hum Genet, 51*(8), 701-705. doi:10.1007/s10038-006-0011-7

Blumcke, I., Thom, M., Aronica, E., Armstrong, D. D., Vinters, H. V., Palmini, A., . . . Spreafico, R. (2011). The clinicopathologic spectrum of focal cortical dysplasias: a consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. *Epilepsia, 52*(1), 158-174. doi:10.1111/j.1528-1167.2010.02777.x

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., . . . Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biol, 19*(1), 199. doi:10.1186/s13059-018-1577-z

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., . . . Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res, 18*(11), 1752-1762. doi:10.1101/gr.080663.108

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., & Kazazian, H. H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A, 100*(9), 5280-5285. doi:10.1073/pnas.0831042100

Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A., & Walsh, C. A. (2014). Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep, 8*(5), 1280-1289. doi:10.1016/j.celrep.2014.07.043

Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A., & Walsh, C. A. (2015). Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep, 10*(4), 645. doi:10.1016/j.celrep.2015.01.028

Casson, A. G., Zheng, Z., Evans, S. C., Veugelers, P. J., Porter, G. A., & Guernsey, D. L. (2005). Polymorphisms in DNA repair genes in the molecular pathogenesis of esophageal (Barrett) adenocarcinoma. *Carcinogenesis, 26*(9), 1536-1541. doi:10.1093/carcin/bgi115

Chiu, Y. L., Witkowska, H. E., Hall, S. C., Santiago, M., Soros, V. B., Esnault, C., . . . Greene, W. C. (2006). High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition. *Proc Natl Acad Sci U S A, 103*(42), 15588-15593. doi:10.1073/pnas.0604524103

Chronister, W. D., Burbulis, I. E., Wierman, M. B., Wolpert, M. J., Haakenson, M. F., Smith, A. C. B., . . . McConnell, M. J. (2019). Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. *Cell Rep, 26*(4), 825-835 e827. doi:10.1016/j.celrep.2018.12.107

Cleaver, J. E. (2005). Cancer in xeroderma pigmentosum and related disorders of DNA repair. *Nat Rev Cancer, 5*(7), 564-573. doi:10.1038/nrc1652

Cullen, B. R. (2006). Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors. *J Virol, 80*(3), 1067-1076. doi:10.1128/JVI.80.3.1067-1076.2006

D'Gama, A. M., & Walsh, C. A. (2018). Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci, 21*(11), 1504-1514. doi:10.1038/s41593-018-0257-3

D'Gama, A. M., Woodworth, M. B., Hossain, A. A., Bizzotto, S., Hatem, N. E., LaCoursiere, C. M., . . . Walsh, C. A. (2017). Somatic Mutations Activating the mTOR Pathway in Dorsal Telencephalic Progenitors Cause a Continuum of Cortical Dysplasias. *Cell Rep, 21*(13), 3754-3766. doi:10.1016/j.celrep.2017.11.106

De Luca, C., Guadagni, F., Sinibaldi-Vallebona, P., Sentinelli, S., Gallucci, M., Hoffmann, A., . . . Sciamanna, I. (2016). Enhanced expression of LINE-1-encoded ORF2 protein in early stages of colon and prostate transformation. *Oncotarget, 7*(4), 4048-4061. doi:10.18632/oncotarget.6767

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., . . . Lasken, R. S. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A, 99*(8), 5261-5266. doi:10.1073/pnas.082089499

Deininger, P. (2011). Alu elements: know the SINEs. *Genome Biol, 12*(12), 236. doi:10.1186/gb-2011-12-12-236

Deininger, P., Morales, M. E., White, T. B., Baddoo, M., Hedges, D. J., Servant, G., . . . Belancio, V. P. (2017). A comprehensive approach to expression of L1 loci. *Nucleic Acids Res, 45*(5), e31. doi:10.1093/nar/gkw1067

Denli, A. M., Narvaiza, I., Kerman, B. E., Pena, M., Benner, C., Marchetto, M. C., . . . Gage, F. H. (2015). Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell, 163*(3), 583-593. doi:10.1016/j.cell.2015.09.025

Dewannieux, M., Esnault, C., & Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet, 35*(1), 41-48. doi:10.1038/ng1223

Dou, Y., Gold, H. D., Luquette, L. J., & Park, P. J. (2018). Detecting Somatic Mutations in Normal Cells. *Trends Genet, 34*(7), 545-557. doi:10.1016/j.tig.2018.04.003

Dou, Y., Yang, X., Li, Z., Wang, S., Zhang, Z., Ye, A. Y., . . . Wei, L. (2017). Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum

disorder and autistic traits and the origin of mutations. *Hum Mutat, 38*(8), 1002-1013. doi:10.1002/humu.23255
Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y., & Moran, J. V. (2015). A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol Cell, 60*(5), 728-741. doi:10.1016/j.molcel.2015.10.012
Doucet-O'Hare, T. T., Rodic, N., Sharma, R., Darbari, I., Abril, G., Choi, J. A., . . . Kazazian, H. H., Jr. (2015). LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc Natl Acad Sci U S A, 112*(35), E4894-4900. doi:10.1073/pnas.1502474112
Erwin, J. A., Paquola, A. C., Singer, T., Gallina, I., Novotny, M., Quayle, C., . . . Gage, F. H. (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci, 19*(12), 1583-1591. doi:10.1038/nn.4388
Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., . . . Walsh, C. A. (2015). Cell lineage analysis in human brain using endogenous retroelements. *Neuron, 85*(1), 49-59. doi:10.1016/j.neuron.2014.12.028
Evrony, G. D., Lee, E., Park, P. J., & Walsh, C. A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. *Elife, 5*. doi:10.7554/eLife.12966
Ewing, A. D., Gacita, A., Wood, L. D., Ma, F., Xing, D., Kim, M. S., . . . Solyom, S. (2015). Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res, 25*(10), 1536-1545. doi:10.1101/gr.196238.115
Fanning, T. G., & Singer, M. F. (1987). LINE-1: a mammalian transposable element. *Biochim Biophys Acta, 910*(3), 203-212. doi:10.1016/0167-4781(87)90112-6
Faulkner, G. J., & Garcia-Perez, J. L. (2017). L1 Mosaicism in Mammals: Extent, Effects, and Evolution. *Trends Genet, 33*(11), 802-816. doi:10.1016/j.tig.2017.07.004
Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., . . . Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet, 41*(5), 563-571. doi:10.1038/ng.368
Feng, Q., Moran, J. V., Kazazian, H. H., Jr., & Boeke, J. D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell, 87*(5), 905-916. doi:10.1016/s0092-8674(00)81997-2
Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet, 3*(5), 329-341. doi:10.1038/nrg793
Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet, 41*, 331-368. doi:10.1146/annurev.genet.40.110405.090448
Freed, D., & Pevsner, J. (2016). The Contribution of Mosaic Variants to Autism Spectrum Disorder. *PLoS Genet, 12*(9), e1006245. doi:10.1371/journal.pgen.1006245
Freed, D., Stevens, E. L., & Pevsner, J. (2014). Somatic mosaicism in the human genome. *Genes (Basel), 5*(4), 1064-1094. doi:10.3390/genes5041064
Gardner, E. J., Prigmore, E., Gallone, G., Danecek, P., Samocha, K. E., Handsaker, J., . . . Hurles, M. E. (2019). Contribution of retrotransposition to developmental disorders. *Nat Commun, 10*(1), 4630. doi:10.1038/s41467-019-12520-y
Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nat Rev Genet, 17*(3), 175-188. doi:10.1038/nrg.2015.16
Gonzalez, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., . . . Gawad, C. (2021). Accurate Genomic Variant Detection in Single Cells with Primary Template-Directed Amplification. *bioRxiv*, 2020.2011.2020.391961. doi:10.1101/2020.11.20.391961

Goodier, J. L. (2016). Restricting retrotransposons: a review. *Mob DNA, 7*, 16. doi:10.1186/s13100-016-0070-z
Goriaux, C., Theron, E., Brasset, E., & Vaury, C. (2014). History of the discovery of a master locus producing piRNAs: the flamenco/COM locus in Drosophila melanogaster. *Front Genet, 5*, 257. doi:10.3389/fgene.2014.00257
Grandi, N., & Tramontano, E. (2018). HERV Envelope Proteins: Physiological Role and Pathogenic Potential in Cancer and Autoimmunity. *Front Microbiol, 9*, 462. doi:10.3389/fmicb.2018.00462
Guerreiro, R. J., Gustafson, D. R., & Hardy, J. (2012). The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE. *Neurobiol Aging, 33*(3), 437-456. doi:10.1016/j.neurobiolaging.2010.03.025
Guerrini, R., Mei, D., Sisodiya, S., Sicca, F., Harding, B., Takahashi, Y., . . . Parrini, E. (2004). Germline and mosaic mutations of FLN1 in men with periventricular heterotopia. *Neurology, 63*(1), 51-56. doi:10.1212/01.wnl.0000132818.84827.4d
Guo, C., Jeong, H. H., Hsieh, Y. C., Klein, H. U., Bennett, D. A., De Jager, P. L., . . . Shulman, J. M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. *Cell Rep, 23*(10), 2874-2880. doi:10.1016/j.celrep.2018.05.004
Ha, H., Song, J., Wang, S., Kapusta, A., Feschotte, C., Chen, K. C., & Xing, J. (2014). A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. *BMC Genomics, 15*, 545. doi:10.1186/1471-2164-15-545
Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., . . . Risch, N. (2011). Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatry, 68*(11), 1095-1102. doi:10.1001/archgenpsychiatry.2011.76
Han, K., Lee, J., Meyer, T. J., Remedios, P., Goodwin, L., & Batzer, M. A. (2008). L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A, 105*(49), 19366-19371. doi:10.1073/pnas.0807866105
Hancks, D. C., Ewing, A. D., Chen, J. E., Tokunaga, K., & Kazazian, H. H., Jr. (2009). Exon-trapping mediated by the human retrotransposon SVA. *Genome Res, 19*(11), 1983-1991. doi:10.1101/gr.093153.109
Hancks, D. C., Goodier, J. L., Mandal, P. K., Cheung, L. E., & Kazazian, H. H., Jr. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum Mol Genet, 20*(17), 3386-3400. doi:10.1093/hmg/ddr245
Hancks, D. C., & Kazazian, H. H., Jr. (2010). SVA retrotransposons: Evolution and genetic instability. *Semin Cancer Biol, 20*(4), 234-245. doi:10.1016/j.semcancer.2010.04.001
Hancks, D. C., & Kazazian, H. H., Jr. (2016). Roles for retrotransposon insertions in human disease. *Mob DNA, 7*, 9. doi:10.1186/s13100-016-0065-9
Hata, K., & Sakaki, Y. (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene, 189*(2), 227-234. doi:10.1016/s0378-1119(96)00856-6
He, J., Babarinde, I. A., Sun, L., Xu, S., Chen, R., Shi, J., . . . Chen, J. (2021). Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. *Nat Commun, 12*(1), 1456. doi:10.1038/s41467-021-21808-x
Hohjoh, H., & Singer, M. F. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J, 15*(3), 630-639. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8599946

Hou, Y., Wu, K., Shi, X., Li, F., Song, L., Wu, H., . . . Wang, J. (2015). Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *Gigascience, 4*, 37. doi:10.1186/s13742-015-0068-3

Hu, P., Zhang, W., Xin, H., & Deng, G. (2016). Single Cell Isolation and Analysis. *Front Cell Dev Biol, 4*, 116. doi:10.3389/fcell.2016.00116

Huang, L., Ma, F., Chapman, A., Lu, S., & Xie, X. S. (2015). Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annu Rev Genomics Hum Genet, 16*, 79-102. doi:10.1146/annurev-genom-090413-025352

Iourov, I. Y., Vorsanova, S. G., & Yurov, Y. B. (2006). Chromosomal variation in mammalian neuronal cells: known facts and attractive hypotheses. *Int Rev Cytol, 249*, 143-191. doi:10.1016/S0074-7696(06)49003-3

Iourov, I. Y., Vorsanova, S. G., & Yurov, Y. B. (2008). Molecular cytogenetics and cytogenomics of brain diseases. *Curr Genomics, 9*(7), 452-465. doi:10.2174/138920208786241216

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences, 94*(5), 1872-1877. Retrieved from http://www.pnas.org/content/94/5/1872.abstract

Karadima, G., Bugge, M., Nicolaidis, P., Vassilopoulos, D., Avramopoulos, D., Grigoriadou, M., . . . et al. (1998). Origin of nondisjunction in trisomy 8 and trisomy 8 mosaicism. *Eur J Hum Genet, 6*(5), 432-438. doi:10.1038/sj.ejhg.5200212

Kazazian, H. H., Jr. (2014). Processed pseudogene insertions in somatic cells. *Mob DNA, 5*, 20. doi:10.1186/1759-8753-5-20

Kazazian, H. H., Jr., & Moran, J. V. (2017). Mobile DNA in Health and Disease. *N Engl J Med, 377*(4), 361-370. doi:10.1056/NEJMra1510092

Kazazian, H. H., Jr., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature, 332*(6160), 164-166. doi:10.1038/332164a0

Khadgi, B. B., Govindaraju, A., & Christensen, S. M. (2019). Completion of LINE integration involves an open '4-way' branched DNA intermediate. *Nucleic Acids Res, 47*(16), 8708-8719. doi:10.1093/nar/gkz673

Kim, J., Hu, C., Moufawad El Achkar, C., Black, L. E., Douville, J., Larson, A., . . . Yu, T. W. (2019). Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. *N Engl J Med, 381*(17), 1644-1652. doi:10.1056/NEJMoa1813279

Kinomoto, M., Kanno, T., Shimura, M., Ishizaka, Y., Kojima, A., Kurata, T., . . . Tokunaga, K. (2007). All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. *Nucleic Acids Res, 35*(9), 2955-2964. doi:10.1093/nar/gkm181

Koito, A., & Ikeda, T. (2013). Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. *Front Microbiol, 4*, 28. doi:10.3389/fmicb.2013.00028

Kojima, K. K. (2011). Alu monomer revisited: recent generation of Alu monomers. *Mol Biol Evol, 28*(1), 13-15. doi:10.1093/molbev/msq218

Kolosha, V. O., & Martin, S. L. (1997). In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc Natl Acad Sci U S A, 94*(19), 10155-10160. doi:10.1073/pnas.94.19.10155

Konkel, M. K., Walker, J. A., Hotard, A. B., Ranck, M. C., Fontenot, C. C., Storer, J., . . . Batzer, M. A. (2015). Sequence Analysis and Characterization of Active Human Alu Subfamilies Based on the 1000 Genomes Pilot Project. *Genome Biol Evol, 7*(9), 2608-2622. doi:10.1093/gbe/evv167

Krupp, D. R., Barnard, R. A., Duffourd, Y., Evans, S. A., Mulqueen, R. M., Bernier, R., . . . O'Roak, B. J. (2017). Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *Am J Hum Genet, 101*(3), 369-390. doi:10.1016/j.ajhg.2017.07.016

Kulpa, D. A., & Moran, J. V. (2005). Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet, 14*(21), 3237-3248. doi:10.1093/hmg/ddi354

Kwon, Y. J., Choi, Y., Eo, J., Noh, Y. N., Gim, J. A., Jung, Y. D., . . . Kim, H. S. (2013). Structure and Expression Analyses of SVA Elements in Relation to Functional Genes. *Genomics Inform, 11*(3), 142-148. doi:10.5808/GI.2013.11.3.142

Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nat Rev Genet, 21*(12), 721-736. doi:10.1038/s41576-020-0251-y

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Lasken, R. S., & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol, 7*, 19. doi:10.1186/1472-6750-7-19

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., 3rd, . . . Cancer Genome Atlas Research, N. (2012). Landscape of somatic retrotransposition in human cancers. *Science, 337*(6097), 967-971. doi:10.1126/science.1222077

Lee, J., Han, K., Meyer, T. J., Kim, H. S., & Batzer, M. A. (2008). Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS One, 3*(12), e4047. doi:10.1371/journal.pone.0004047

Lee, J. H., Huynh, M., Silhavy, J. L., Kim, S., Dixon-Salazar, T., Heiberg, A., . . . Gleeson, J. G. (2012). De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nat Genet, 44*(8), 941-945. doi:10.1038/ng.2329

Levy, O., Knisbacher, B. A., Levanon, E. Y., & Havlin, S. (2017). Integrating networks and comparative genomics reveals retroelement proliferation dynamics in hominid genomes. *Sci Adv, 3*(10), e1701256. doi:10.1126/sciadv.1701256

Li, W., Jin, Y., Prazak, L., Hammell, M., & Dubnau, J. (2012). Transposable elements in TDP-43-mediated neurodegenerative disorders. *PLoS One, 7*(9), e44099. doi:10.1371/journal.pone.0044099

Li, W., Lee, M. H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., . . . Nath, A. (2015). Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med, 7*(307), 307ra153. doi:10.1126/scitranslmed.aac8201

Li, W., Prazak, L., Chatterjee, N., Gruninger, S., Krug, L., Theodorou, D., & Dubnau, J. (2013). Activation of transposable elements during aging and neuronal decline in Drosophila. *Nat Neurosci, 16*(5), 529-531. doi:10.1038/nn.3368

Lim, E. T., Uddin, M., De Rubeis, S., Chan, Y., Kamumbu, A., Zhang, X., . . . Walsh, C. A. (2017). Rates, Distribution, and Implications of Post-zygotic Mutations in Autism Spectrum Disorder

. *Submitted*.

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., . . . Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science, 359*(6375), 555-559. doi:10.1126/science.aao4426

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., . . . Walsh, C. A. (2015). Somatic mutation in single human neurons tracks

developmental and transcriptional history. *Science, 350*(6256), 94-98. doi:10.1126/science.aab1785
Lu, X., Sachs, F., Ramsay, L., Jacques, P. E., Goke, J., Bourque, G., & Ng, H. H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol, 21*(4), 423-425. doi:10.1038/nsmb.2799
Mathias, S. L., Scott, A. F., Kazazian, H. H., Jr., Boeke, J. D., & Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science, 254*(5039), 1808-1810. doi:10.1126/science.1722352
Maury, E. A., & Walsh, C. A. (2021). Somatic copy number variants in neuropsychiatric disorders. *Curr Opin Genet Dev, 68*, 9-17. doi:10.1016/j.gde.2020.12.013
McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A, 36*(6), 344-355. doi:10.1073/pnas.36.6.344
McConnell, M. J., Moran, J. V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., . . . Brain Somatic Mosaicism, N. (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science, 356*(6336). doi:10.1126/science.aal1641
Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K. W., . . . Nakamura, Y. (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res, 52*(3), 643-645. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1310068
Miller, M. B., Reed, H. C., & Walsh, C. A. (2021). Brain Somatic Mutation in Aging and Alzheimer's Disease. *Annu Rev Genomics Hum Genet*. doi:10.1146/annurev-genom-121520-081242
Moog, U., Felbor, U., Has, C., & Zirn, B. (2020). Disorders Caused by Genetic Mosaicism. *Dtsch Arztebl Int, 116*(8), 119-125. doi:10.3238/arztebl.2020.0119
Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., & Kazazian, H. H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. *Cell, 87*(5), 917-927. doi:10.1016/s0092-8674(00)81998-4
Morris, R., & Mehta, P. (2018). The Isolation of Pure Populations of Neurons by Laser Capture Microdissection: Methods and Application in Neuroscience. *Methods Mol Biol, 1723*, 223-233. doi:10.1007/978-1-4939-7558-7_12
Nelson, P. N., Carnegie, P. R., Martin, J., Davari Ejtehadi, H., Hooley, P., Roden, D., . . . Murray, P. G. (2003). Demystified. Human endogenous retroviruses. *Mol Pathol, 56*(1), 11-18. doi:10.1136/mp.56.1.11
Ning, L., Li, Z., Wang, G., Hu, W., Hou, Q., Tong, Y., . . . He, J. (2015). Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. *Sci Rep, 5*, 11415. doi:10.1038/srep11415
Parrini, E., Mei, D., Wright, M., Dorn, T., & Guerrini, R. (2004). Mosaic mutations of the FLN1 gene cause a mild phenotype in patients with periventricular heterotopia. *Neurogenetics, 5*(3), 191-196. doi:10.1007/s10048-004-0187-y
Pennisi, E. (2007). Evolution. Jumping genes hop into the evolutionary limelight. *Science, 317*(5840), 894-895. doi:10.1126/science.317.5840.894
Phokaew, C., Kowudtitham, S., Subbalekha, K., Shuangshoti, S., & Mutirangura, A. (2008). LINE-1 methylation patterns of different loci in normal and cancerous cells. *Nucleic Acids Res, 36*(17), 5704-5712. doi:10.1093/nar/gkn571
Pizarro, J. G., & Cristofari, G. (2016). Post-Transcriptional Control of LINE-1 Retrotransposition by Cellular Host Factors in Somatic Cells. *Front Cell Dev Biol, 4*, 14. doi:10.3389/fcell.2016.00014

Platt, R. N., 2nd, Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res, 26*(1-2), 25-43. doi:10.1007/s10577-017-9570-z

Poduri, A., Evrony, G. D., Cai, X., Elhosary, P. C., Beroukhim, R., Lehtinen, M. K., . . . Walsh, C. A. (2012). Somatic activation of AKT3 causes hemispheric developmental brain malformations. *Neuron, 74*(1), 41-48. doi:10.1016/j.neuron.2012.03.010

Polak, P., & Domany, E. (2006). Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics, 7*, 133. doi:10.1186/1471-2164-7-133

Prak, E. T., & Kazazian, H. H., Jr. (2000). Mobile elements and the human genome. *Nat Rev Genet, 1*(2), 134-144. doi:10.1038/35038572

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., . . . Schumann, G. G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res, 40*(4), 1666-1683. doi:10.1093/nar/gkr863

Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., & Moran, J. V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr, 3*(2), MDNA3-0061-2014. doi:10.1128/microbiolspec.MDNA3-0061-2014

Richardson, S. R., Narvaiza, I., Planegger, R. A., Weitzman, M. D., & Moran, J. V. (2014). APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. *Elife, 3*, e02008. doi:10.7554/eLife.02008

Robinson, W. P., Binkert, F., Bernasconi, F., Lorda-Sanchez, I., Werder, E. A., & Schinzel, A. A. (1995). Molecular studies of chromosomal mosaicism: relative frequency of chromosome gain or loss and possible role of cell selection. *Am J Hum Genet, 56*(2), 444-451. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7847381

Rodic, N., Sharma, R., Sharma, R., Zampella, J., Dai, L., Taylor, M. S., . . . Burns, K. H. (2014). Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am J Pathol, 184*(5), 1280-1286. doi:10.1016/j.ajpath.2014.01.007

Rodic, N., Steranka, J. P., Makohon-Moore, A., Moyer, A., Shen, P., Sharma, R., . . . Burns, K. H. (2015). Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat Med, 21*(9), 1060-1064. doi:10.1038/nm.3919

Rodin, R. E., Dou, Y., Kwon, M., Sherman, M. A., D'Gama, A. M., Doan, R. N., . . . Walsh, C. A. (2021). The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. *Nat Neurosci, 24*(2), 176-185. doi:10.1038/s41593-020-00765-6

Rodin, R. E., & Walsh, C. A. (2018). Somatic Mutation in Pediatric Neurological Diseases. *Pediatr Neurol, 87*, 20-22. doi:10.1016/j.pediatrneurol.2018.08.008

Sala Frigerio, C., Lau, P., Troakes, C., Deramecourt, V., Gele, P., Van Loo, P., . . . De Strooper, B. (2015). On the identification of low allele frequency mosaic mutations in the brains of Alzheimer's disease patients. *Alzheimers Dement, 11*(11), 1265-1276. doi:10.1016/j.jalz.2015.02.007

Schumann, G. G. (2007). APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition. *Biochem Soc Trans, 35*(Pt 3), 637-642. doi:10.1042/BST0350637

Schwahn, U., Lenzner, S., Dong, J., Feil, S., Hinzmann, B., van Duijnhoven, G., . . . Berger, W. (1998). Positional cloning of the gene for X-linked retinitis pigmentosa 2. *19*, 327. doi:10.1038/1214

Scott, A. F., Schmeckpeper, B. J., Abdelrazik, M., Comey, C. T., O'Hara, B., Rossiter, J. P., . . . Margolet, L. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics, 1*(2), 113-125. doi:10.1016/0888-7543(87)90003-6

Scott, E. C., Gardner, E. J., Masood, A., Chuang, N. T., Vertino, P. M., & Devine, S. E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res, 26*(6), 745-755. doi:10.1101/gr.201814.115

Sen, S. K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P. A., . . . Batzer, M. A. (2006). Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet, 79*(1), 41-53. doi:10.1086/504600

Shirley, M. D., Tang, H., Gallione, C. J., Baugher, J. D., Frelin, L. P., Cohen, B., . . . Pevsner, J. (2013). Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. *N Engl J Med, 368*(21), 1971-1979. doi:10.1056/NEJMoa1213507

Shukla, R., Upton, K. R., Munoz-Lopez, M., Gerhardt, D. J., Fisher, M. E., Nguyen, T., . . . Faulkner, G. J. (2013). Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell, 153*(1), 101-111. doi:10.1016/j.cell.2013.02.032

Singer, H., Walier, M., Nusgen, N., Meesters, C., Schreiner, F., Woelfle, J., . . . El-Maarri, O. (2012). Methylation of L1Hs promoters is lower on the inactive X, has a tendency of being higher on autosomes in smaller genomes and shows inter-individual variability at some loci. *Hum Mol Genet, 21*(1), 219-235. doi:10.1093/hmg/ddr456

Smit, A., Hubley, R & Green, P. . (2013-2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev, 6*(6), 743-748. doi:10.1016/s0959-437x(96)80030-x

Smit, A. F., Toth, G., Riggs, A. D., & Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol, 246*(3), 401-417. doi:10.1006/jmbi.1994.0095

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol, 21*(6), 1973-1985. doi:10.1128/MCB.21.6.1973-1985.2001

Sultana, T., van Essen, D., Siol, O., Bailly-Bechet, M., Philippe, C., Zine El Aabidine, A., . . . Cristofari, G. (2019). The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Mol Cell, 74*(3), 555-570 e557. doi:10.1016/j.molcel.2019.02.036

Sun, W., Samimi, H., Gamez, M., Zare, H., & Frost, B. (2018). Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. *Nat Neurosci, 21*(8), 1038-1048. doi:10.1038/s41593-018-0194-1

Swergold, G. D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol, 10*(12), 6718-6729. doi:10.1128/mcb.10.12.6718-6729.1990

Szak, S. T., Pickeral, O. K., Makalowski, W., Boguski, M. S., Landsman, D., & Boeke, J. D. (2002). Molecular archeology of L1 insertions in the human genome. *Genome Biol, 3*(10), research0052. doi:10.1186/gb-2002-3-10-research0052

Tam, O. H., Rozhkov, N. V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., . . . Gale Hammell, M. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep, 29*(5), 1164-1177 e1165. doi:10.1016/j.celrep.2019.09.066

Taniguchi-Ikeda, M., Kobayashi, K., Kanagawa, M., Yu, C. C., Mori, K., Oda, T., . . . Toda, T. (2011). Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. *Nature, 478*(7367), 127-131. doi:10.1038/nature10456

Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjold, M., Ponder, B. A., & Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics, 13*(3), 718-725. doi:10.1016/0888-7543(92)90147-k

Thayer, R. E., Singer, M. F., & Fanning, T. G. (1993). Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. *Gene, 133*(2), 273-277. doi:10.1016/0378-1119(93)90651-i

Thomas, J. H., & Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. *Genome Res, 21*(11), 1800-1812. doi:10.1101/gr.121749.111

Turelli, P., Playfoot, C., Grun, D., Raclot, C., Pontis, J., Coudray, A., . . . Trono, D. (2020). Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. *Sci Adv, 6*(35), eaba3200. doi:10.1126/sciadv.aba3200

Ullu, E., & Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature, 312*(5990), 171-172. doi:10.1038/312171a0

Vorsanova, S. G., Yurov, Y. B., Soloviev, I. V., & Iourov, I. Y. (2010). Molecular cytogenetic diagnosis and somatic genome variations. *Curr Genomics, 11*(6), 440-446. doi:10.2174/138920210793176010

Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., Walker, J. A., & Batzer, M. A. (2005). SVA elements: a hominid-specific retroposon family. *J Mol Biol, 354*(4), 994-1007. doi:10.1016/j.jmb.2005.09.085

Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Rasko, T., Szvetnik, A., . . . Izsvak, Z. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature, 516*(7531), 405-409. doi:10.1038/nature13804

Wang, Y., Bae, T., Thorpe, J., Sherman, M. A., Jones, A. G., Cho, S., . . . Abyzov, A. (2021). Comprehensive identification of somatic nucleotide variants in human brain tissue. *Genome Biol, 22*(1), 92. doi:10.1186/s13059-021-02285-3

Wildschutte, J. H., Williams, Z. H., Montesion, M., Subramanian, R. P., Kidd, J. M., & Coffin, J. M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc Natl Acad Sci U S A, 113*(16), E2326-2334. doi:10.1073/pnas.1602336113

Williams, Z., Morozov, P., Mihailovic, A., Lin, C., Puvvula, P. K., Juranek, S., . . . Tuschl, T. (2015). Discovery and Characterization of piRNAs in the Human Fetal Ovary. *Cell Rep, 13*(4), 854-863. doi:10.1016/j.celrep.2015.09.030

Wimmer, K., Callens, T., Wernstedt, A., & Messiaen, L. (2011). The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet, 7*(11), e1002371. doi:10.1371/journal.pgen.1002371

Xing, D., Tan, L., Chang, C. H., Li, H., & Xie, X. S. (2021). Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc Natl Acad Sci U S A, 118*(8). doi:10.1073/pnas.2013106118

Yang, P., Wang, Y., & Macfarlan, T. S. (2017). The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends Genet, 33*(11), 871-881. doi:10.1016/j.tig.2017.08.006

Yu, F., Zingler, N., Schumann, G., & Stratling, W. H. (2001). Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res, 29*(21), 4493-4501. doi:10.1093/nar/29.21.4493

Yurov, Y. B., Vorsanova, S. G., Iourov, I. Y., Demidova, I. A., Beresheva, A. K., Kravetz, V. S., . . . Gorbachevskaya, N. L. (2007). Unexplained autism is frequently associated with low-level mosaic aneuploidy. *J Med Genet, 44*(8), 521-525. doi:10.1136/jmg.2007.049312

Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science, 338*(6114), 1622-1626. doi:10.1126/science.1229164

Chapter 2.
Whole-genome analysis of *de novo* insertions in autism spectrum disorders

Attribution

This chapter contains work from the manuscript “Whole-genome analysis of *de novo* and polymorphic retrotransposon insertions in Autism Spectrum Disorder”, which is currently available as a preprint in *bioRxiv*, 2021.01.29.428895; doi: https://doi.org/10.1101/2021.01.29.428895. Figures, tables, and text have been modified to accommodate the format of this thesis. Rebeca Borges Monroy was the first co-author along with Chong Chu. Rebeca Borges Monroy primarily contributed to the study design, analysis of data, experimental validation, and wrote this chapter. Chong Chu developed the *xTea* algorithm used in the analysis including the filtering and genotyping module and developed the docker for this algorithm. At the time of writing this, the manuscript for this algorithm is currently in press in a manuscript called “Comprehensive identification of transposable element insertions using multiple sequencing technologies” and will be available in *Nature Communications*, https://doi.org/10.1038/s41467-021-24041-8. Rebeca Borges-Monroy is listed as a second author in this work. Chong Chu obtained and provided the gold standard set of transposon insertions in HapMap sample HG002 and ran *xTea* on this sample. Eunjung Alice Lee and Kyu Park developed and modified the *memTea* algorithm used in the analysis. Kyu Park performed the simulation of the HuRef genome. Caroline Dias and Rebeca Borges Monroy reviewed the phenotypes of individuals with relevant mutations. Jaejoon Choi ran *xTea* on the samples on the cloud. Soohyun Lee assisted with running *xTea* on the cloud. Eunjung Alice Lee and Chris A. Walsh contributed to the design of the study and supervised the study.

Summary

Retrotransposons can cause Mendelian disease (Hancks & Kazazian, 2016), but their role in autism spectrum disorder (ASD) has not been systematically defined. We analyzed whole-genome sequencing (WGS) data from the largest cohort of 2,288 ASD families from the Simons Simplex Collection (SSC) by establishing a scalable computational pipeline for retrotransposon insertion detection. We report 86,154 polymorphic retrotransposon insertions—including >60% not previously reported—and 158 *de novo* retrotransposition events. As expected, the overall burden of *de novo* events was similar between ASD individuals and unaffected siblings, with 1 *de novo* insertion per 29, 104, and 192 births for Alu, L1, and SVA respectively, and 1 *de novo* insertion per 20 births total. However, ASD cases showed more *de novo* L1 insertions than expected in the introns of ASD genes. Additionally, we observed exonic insertions in loss-of-function intolerant genes, including a likely pathogenic exonic insertion in *CSDE1*, only in ASD individuals. We detected a strong depletion of polymorphic germline retrotransposon insertions in regulatory regions of the genome during brain development, but this depletion was not observed with *de novo* insertions, and we identified a trend for more Alu insertions in active enhancers than expected. We achieved a high validation rate of 93% using full-length PCR. These findings suggest a modest, but important, impact of intronic and exonic retrotransposon insertions in ASD, show the importance of WGS for their analysis, and highlight the utility of specific bioinformatics tools for high-throughput detection of retrotransposon insertions.

Introduction

Methods for detection of polymorphic transposable elements

Transposable elements (TEs) comprise approximately half of the human genome (Lander et al., 2001). Due to their prevalence and repetitive nature, it is challenging to detect transposable element insertions (TEIs) with standard computational tools for structural variant detection, in particular in short-read data (Chu, Zhao, Park, & Lee, 2020; Goerner-Potvin & Bourque, 2018). Although many methods exist for identifying TEs in genome assemblies (Goerner-Potvin & Bourque, 2018), here we focus on identifying and genotyping polymorphic insertions in the human genome since this genome is already well annotated. Polymorphic TEIs are present in individuals in the population but are not present in the reference genome. Additionally, in this chapter, we refer to *de novo* TEIs as those insertions that are present in an individual as a result of a mutation that arose in one of their parent's germ cells or in the zygote during the first cell divisions. Since *de novo* TEIs are heterozygous in the individual carrying the mutation, they have similar characteristics to polymorphic insertions in WGS data and can be detected with the same algorithms.

There are many existing computational methods for the detection and annotation of TEIs in human WGS data (Bogaerts-Marquez et al., 2020; Gardner et al., 2017; Jiang, Chen, Huang, Liu, & Verdier, 2015; Keane, Wong, & Adams, 2013; Kroon et al., 2016; E. Lee et al., 2012; Rajaby & Sung, 2018; Thung et al., 2014; Tubio et al., 2014; Wu et al., 2014; W. Zhou et al., 2020; Zhuang, Wang, Theurkauf, & Weng, 2014), particularly for short-read sequencing data. These tools generally follow a similar approach: they first search for clusters of discordant paired reads where one read maps to the reference genome and the other read maps to a different location and/or a library of repetitive or TE sequences. They may also search for clusters of clipped reads. These are reads

spanning the insertion breakpoint which partially map to the reference genome and also to a TE library. The part of the read that does not map to the reference genome is "clipped" by alignment algorithms such as BWA (Heng Li, 2013). Finally, tools may detect or filter TEs based on genomic features specific to TE insertions (Chu et al., 2020; Ewing, 2015; Goerner-Potvin & Bourque, 2018). This includes the presence of a target site duplication (TSD) on both sides of the insertion, a poly(A) tail at the end of the insertion (Kazazian & Moran, 1998) as well as a 3' transduction (Goodier, Ostertag, & Kazazian, 2000). Some algorithms use combinations of these strategies while some methods use all three (Ewing, 2015; Goerner-Potvin & Bourque, 2018).

The Mobile Element Locator Tool (*MELT*) is a popular method for the detection of polymorphic germline and *de novo* TEIs (Gardner et al., 2017). *MELT* was developed for analyzing the 1000 Genomes data and has since been used for detecting *de novo* TEIs in other large cohorts (Feusier et al., 2019) including the Genome Aggregation Database (gnomAD) (Collins et al., 2020) and cohorts with developmental disorders (Gardner et al., 2019) and ASD (Belyeu et al., 2021). A comparison of *MELT* with other algorithms including *TEMP* (Zhuang et al., 2014), *RetroSeq* (Keane et al., 2013), *Tangram* (Wu et al., 2014), and *Mobster* (Thung et al., 2014), showed that *MELT* had a higher sensitivity, a faster runtime, and a higher specificity (Gardner et al., 2017). However, the intersection of several algorithms for TEI detection usually reveals low overlap (Ewing, 2015), and some researchers have proposed combining methods to increase their sensitivity (Feusier et al., 2019; Rishishwar, Marino-Ramirez, & Jordan, 2017). This is a reasonable approach for a small sample size but can be costly when running these tools on the cloud on thousands of files. Because of this, it is critical to select a tool with high sensitivity. Although a high specificity is also important, false positives may usually be filtered out post hoc to increase this after running a certain algorithm.

De novo retrotransposition rates in humans

The initial estimates of *de novo* retrotransposition rates that are still commonly cited in the literature are ~1 per birth 20 births for Alu elements (Cordaux, Hedges, Herke, & Batzer, 2006) (Xing et al., 2009), ~1 LINE-1 insertions per 212 births (A. D. Ewing & H. H. Kazazian, Jr., 2010; Xing et al., 2009), and 1 SVA insertions per 916 births (Xing et al., 2009). These estimates have been obtained indirectly and using only a few genomes. Cordaux *et al.* estimated Alu retrotransposition rates using two different methods. The first method is evolutionary-based and relies on a comparison of the chimpanzee and human genome to obtain the number of human-specific insertion events that are fixed in the human population. However, this method assumes an average retrotransposition rate of over 6 million years. The second method compares the number of disease-causing *de novo* retrotransposition events in the Human Gene Mutation Database to the number of disease-causing SNVs. A caveat of this method is that this database may be biased in the types of mutations reported (Cordaux et al., 2006). It also assumes that the impact of SNVs and TEIs on protein function are comparable. Xing *et al.* obtained these estimates by comparing the HuRef genome to the Human Genome Project reference genome (hg18), and calculating the time to the most recent common ancestor between these genomes based on assumptions of the SNV mutation rate (Xing et al., 2009). Ewing and Kazazian obtained this estimate by obtaining the per generation mutation rate μ from the Watterson estimator: $\theta = 4N_e\mu$, where the human effective population size N_e is 10,000 and θ was calculated from the number of segregating sites in 15 unrelated individuals, in which targeted LINE-1 sequencing had been performed (A. D. Ewing & H. H. Kazazian, Jr., 2010). To obtain a more accurate rate, larger sample sizes, deep sequencing, and accurate analyses of retrotransposition in these genomes are required (Campbell & Eichler, 2013).

More recent studies using large cohorts of WGS data have obtained different rates of *de novo* retrotransposition. SVA rates in particular appear to be higher than what was previously reported, with 1 SVA insertion detected per 63 births in an analysis performed on a three-generation human pedigree cohort (Feusier et al., 2019). Here, they also reported a lower rate of *de novo* Alu insertions, with 1 per 40 births, and 1 L1 TEI per 63 births. In another recent study analyzing 2,396 families of this same cohort along with ASD families, similar trends for higher SVA *de novo* TEI rates and lower *de novo* Alu TEI rates were detected, with *de novo* rates of 1 per 42 births for Alu, 1 per 309 births for SVA, and 1 per 231 births for L1. However, these rates are probably lower bound estimates, since researchers have not accounted for the loss of sensitivity that occurs when detecting TEIs in short-read data compared to long-read data (W. Zhou et al., 2020).

Genetics of autism spectrum disorder

ASD is a heterogeneous developmental disorder characterized by communication deficits, impaired social interactions, and repetitive behaviors (American Psychiatric Association, 2013). 1 in 54 children is diagnosed with ASD in the United States (Maenner et al., 2020). Although about 50% of the overall heritability of ASD reflects common variation at a population level (Alonso-Gonzalez, Rodriguez-Fontenla, & Carracedo, 2018; Gaugler et al., 2014), rare inherited and *de novo* copy-number variations and single nucleotide variations can confer an elevated risk to developing ASD (De Rubeis et al., 2014; Doan et al., 2019; Iossifov et al., 2014; Kosmicki et al., 2017; D. Levy et al., 2011; Sanders et al., 2015b; Sanders et al., 2012; Satterstrom et al., 2020; Sebat et al., 2007). Although recent ASD studies have included TEIs (Belyeu

et al., 2021; Brandler et al., 2018; Werling et al., 2018), the smaller sample size and the low rates of *de novo* TEIs limited their analyses, leaving the role of *de novo* TEIs in both exons and introns in ASD largely unknown.

The SSC has become a vital cohort in this field. The SSC collects and provides researchers DNA sequencing data obtained from blood, DNA material, and extensive phenotypic information from families with a single individual with moderate to severe ASD, two unaffected parents, and in many cases, an unaffected sibling (Fischbach & Lord, 2010). Since many ASD cases are sporadic, this familial structure provides the opportunity to study *de novo* variants in ASD with unaffected controls (Fischbach & Lord, 2010). The SSC samples have been thoroughly analyzed at an exome and whole-genome level, and these studies have confirmed that rare *de novo* single nucleotide variants (SNVs), indels, and copy number variants (CNVs) are enriched in these simplex families (Iossifov et al., 2014; D. Levy et al., 2011; O'Roak et al., 2012; Sanders et al., 2015a; Satterstrom et al., 2020), with *de novo* copy-number variations and single nucleotide variations contributing to 30% of cases in the SSC (Iossifov et al., 2014). However, these studies have focused on large CNVs and loss-of-function coding indels and SNVs. Initial studies analyzing non-coding indels, SNVs, and structural variants (SVs) including TEIs in 519 SSC families (Werling et al., 2018) and 829 SSC families (Brandler et al., 2018) did not find any association between these variants and ASD. A more recent analysis using 2,262 ASD SSC families did find a higher rate of *de novo* structural variants at a whole-genome resolution, with 0.206 SVs detected per individual in cases and 0.160 detected per individual in controls on average (Belyeu et al., 2021). This suggests that this association does exist but requires large sample sizes to be revealed, given the low rates of *de novo* SV and TEI.

In this study, we explored the role of TEs as a mutagenic cause of ASD using a sensitive and scalable pipeline for their detection that we developed for this work (Chu et al., 2021). Detecting TEIs in clinical samples is critical to help increase diagnostic rates (Torene et al., 2020). We analyzed 2,288 families from the SSC sequenced at a whole-genome resolution and assessed the phenotypes of individuals with *de novo* insertions to study the impact of TEIs in both coding and non-coding regions.

Results

Benchmarking

In this analysis, we used a tool for detecting TEIs called *xTea* (x-Transposable element analyzer) (Chu et al., 2021) (https://github.com/parklab/xTea), which was originally developed for bulk long-read whole-genome sequencing analysis and then adapted for short-read data and this study, in close collaboration with our group. We compared *xTea* with the *MELT* (Gardner et al., 2017). To determine the sensitivity of each pipeline, we ran them on the same HuRef sample (S. Levy et al., 2007). We used a set of previously experimentally validated germline non-reference TEIs detected in this individual (Xing et al., 2009) as a gold-standard. This gold-standard set includes 584 Alu, 52 L1, and 14 SVA TEIs. We tested the sensitivity of *xTea* and *MELT* first in Sanger sequencing data which was simulated as paired-end read data at a coverage of 40x, and then in 40x Illumina paired-end read WGS data once this became available (B. Zhou et al., 2018).

We detected more candidate TEIs with *xTea* than with *MELT* (Figure 2.1A). However, the overall number of filtered germline TEIs detected with *xTea* was similar to previous studies (Evrony et al., 2015; A. Ewing & H. Kazazian, 2010; E. Lee et al.,

2012). Additionally, we observed a higher sensitivity to detect gold-standard TEIs in both *xTea* raw and filtered candidates compared to *MELT* (Figure 2.1B). We observed a similar sensitivity in both the simulated and sequenced HuRef data, although Alu sensitivity was slightly lower in sequenced data for both methods, suggesting a decreased sensitivity in Illumina data that is not observed in Sanger sequencing data for these smaller variants. Since we detected more candidates with *xTea*, we overlapped *xTea* specific and *MELT* specific candidates to a set of known non-reference (KNR) germline insertions from other studies (A. D. Ewing & H. H. Kazazian, Jr., 2010; Gardner et al., 2017; Huang et al., 2010; Iskow et al., 2010; Stewart et al., 2011; Wang et al., 2006; Xing et al., 2009). More than 64% of *xTea* specific candidates overlapped with KNR TEIs, while less than 15% of *MELT* specific candidates did, suggesting that although we observe more insertions that were not included in the gold-standard set with *xTea* than *MELT*, the majority of these are likely true insertions and not false positives. This analysis confirmed that by using filtered *xTea* candidates, we would have a high sensitivity in ~40x WGS for the detection of polymorphic and *de novo* TEIs.

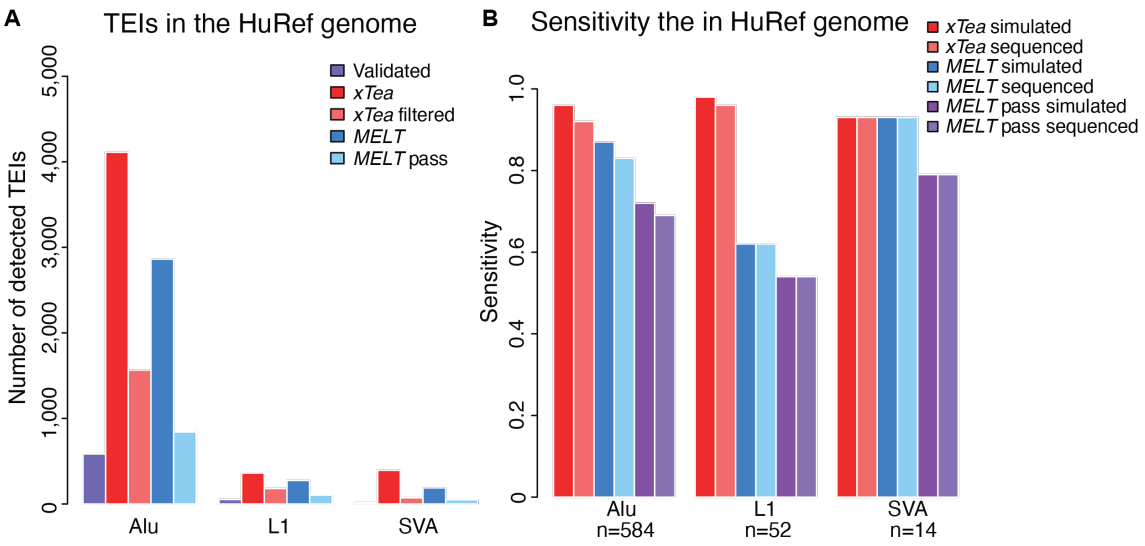

Figure 2.1 *xTea* and *MELT* benchmarking.

A) Number of TEIs in the gold standard set and TEIs detected with *xTea* and *MELT*. *xTea* filtered calls are those classified as “high confidence” by *xTea*, and *MELT* ‘pass' is a high confidence classification defined by the tool. B) Sensitivity is measured by the fraction of HuRef gold standard validated insertions detected by each method from 40x WGS simulated or sequenced WGS in the HuRef genome. The sample size “n” represents the total number of gold standard insertions for each family reported by Xing et al. (Xing et al., 2009).

Polymorphic insertions in the Simons Simplex Collection

In this study, we sought to define the role of TEIs in ASD by analyzing the largest cohort of 2,288 simplex families for *de novo* TEIs at whole-genome resolution (Figure 2.2A). We detected a total of 86,154 unique polymorphic TEIs (68,643 Alu, 12,076 L1, and 5,435 SVA) in the entire cohort (parents and children) (Figure 2.3A and Table 2.1). Each genome carried 1,618 polymorphic TEIs on average (1,385 Alu, 172 L1, and 61 SVA, comparable with previous analyses (Evrony et al., 2015; E. Lee et al., 2012), and the numbers were consistent across different family members (Figure 2.2B and Figure 2.3A). 74% of these TEIs (50,507 Alu, 9,247 L1, 4,273 SVA) were observed in either more than one individual in this cohort (71%; 48,189 Alu, 8,821 L1, 4,021 SVA) or previous studies (33%; 23,018 Alu, 3,663 L1, 1,982 SVA) (Figure 2.3B), suggesting that

the majority of these calls are bona fide. However, more than 60% of calls were novel and had not been detected before in gnomAD (Collins et al., 2020) or the 1000 genomes cohort (Gardner et al., 2017) (Figure 2.2C). In 4,577 unrelated parental samples in the SSC cohort, we detected 77,717 TEIs (dbVar “nstd203”), compared to the 79,632 insertions detected from 54,805 individuals in the gnomAD-SV cohort (Collins et al., 2020). Additionally, insertions in the SSC cohort had a higher overlap with previously published insertions from 2,534 individuals in the 1000 genomes cohort (Gardner et al., 2017) (Figure 2.2C). The majority of parental TEIs were rare, for example, >92% of TEIs having <1% population allele frequency (PAF) within the analyzed cohort (Figure 2.2D and Figure 2.4), which is similar to previous findings of structural variants (Collins et al., 2020).

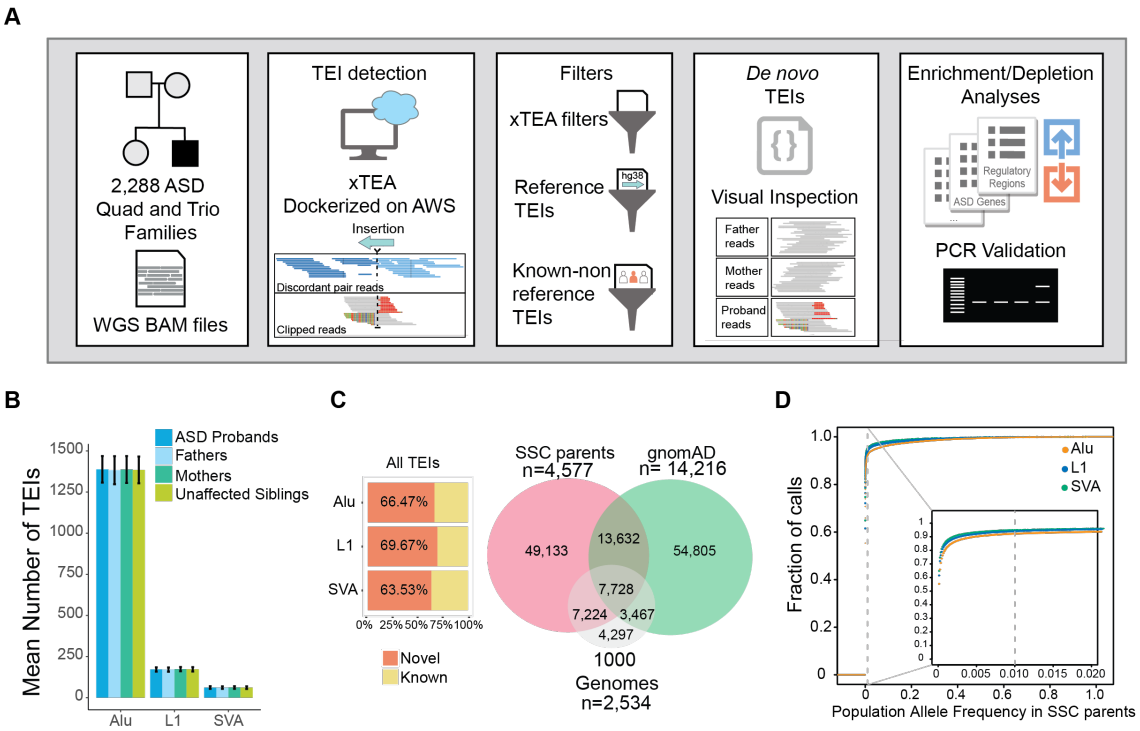

Figure 2.2 Detection of TEIs in the SSC cohort.

A) Pipeline and analysis overview. Quad and trio bam files were analyzed for TEIs using a dockerized version of *xTea* on the cloud in Amazon Web Services (AWS). Candidate TE insertions were filtered using xTea filters, and filters for regions of the genome with reference and known non-reference TEIs for a high confidence set. A custom pipeline for detection of *de novo* insertions was used, and candidates were manually inspected on the Integrative Genomics Viewer. Enrichment or depletion of TEIs in ASD genes, high pLI genes, genomic regions, and regulatory regions in fetal brain development was tested by simulation analyses. A subset of candidates was validated by full-length PCR. B) Mean number of TEIs detected in the SSC cohort with standard deviation. C) Percentage of insertions in the SSC cohort that were not found in previous studies (novel) or overlap with TEIs from previous analyses (known) for all TEIs including those in parents and children (left) and Venn diagram showing overlap with other large cohort studies for TEIs detected in unrelated parental samples in our cohort (right). D) Cumulative fraction of TEIs in unrelated parental samples which are found at a certain population allele frequency (PAF) within the SSC cohort. 94% L1, 92% Alu, and 95% SVA insertions show <1% PAF.

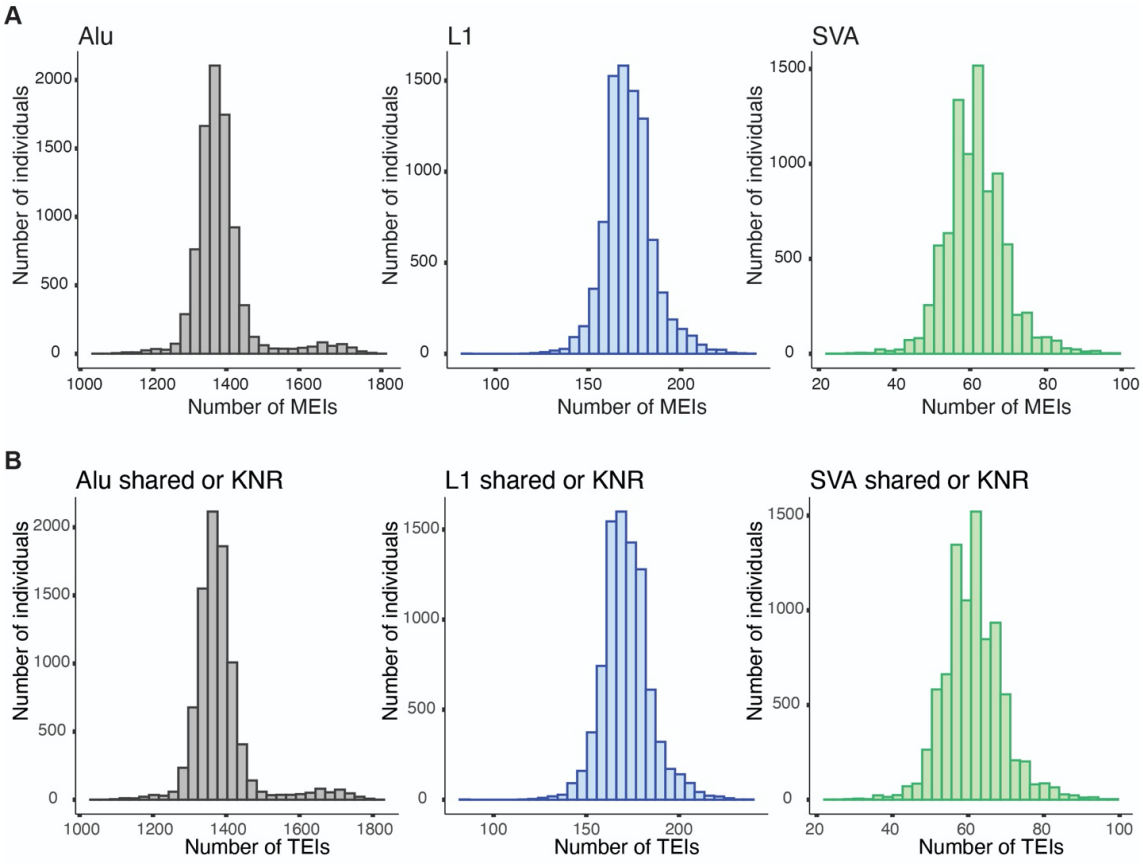

Figure 2.3 Polymorphic and *de novo* TEIs in the SSC cohort.

A) Number of TEIs detected per individual including parental, ASD, and unaffected siblings. Alu N=8,711, mean=1385.45, SD=82.12; L1 N=8,714, mean=171.94, SD=13; SVA N=8,720, mean=61.50, SD=7.75. B) Shared polymorphic and known non-reference TEIs in the SSC cohort. Number of TEIs detected per individual including parental, ASD, and unaffected siblings which are found in more than 2 individuals and/or in gnomAD (Collins et al., 2020) or 1000 genomes (Gardner et al., 2017). Alu N=8,711, mean=1383.26, SD= 81.23; L1 N=8,714, mean=171.62, SD= 12.89; SVA N=8,720, mean=61.36, SD=7.72.

Table 2.1 Polymorphic insertions sample sizes.

	L1	Alu	SVA
ASD Cases	2,286	2,285	2,287
Unaffected Siblings	1,856	1,855	1,858
Fathers	2,286	2,285	2,287
Mothers	2,286	2,286	2,288

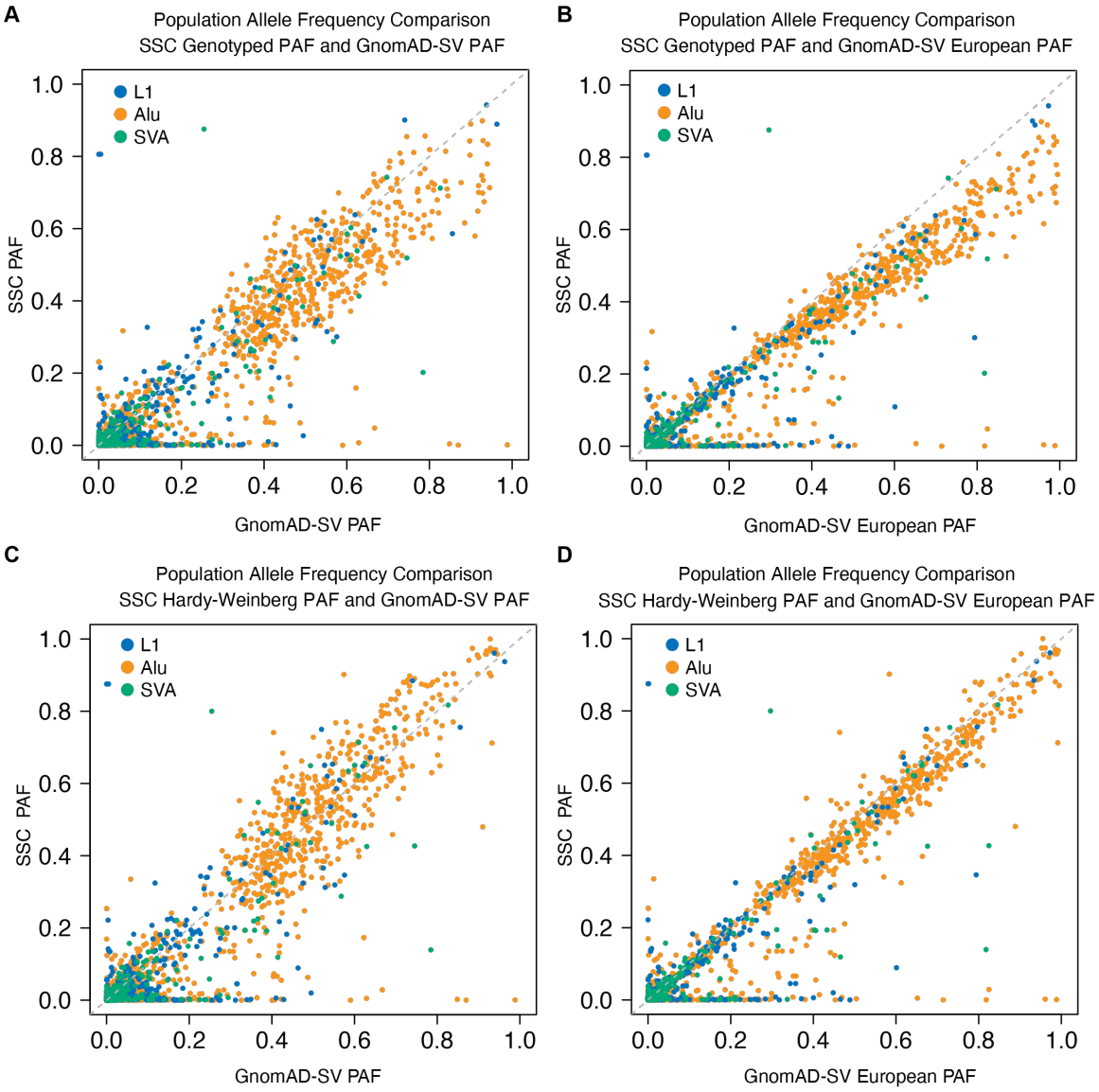

Figure 2.4 Comparison of population allele frequencies (PAFs) between unrelated parental individuals in the SSC cohort and gnomAD-SV TEIs.

A) Comparison using the SSC PAFs from genotyped TEIs. The PAF was defined as the number of alleles carrying the TEI in the population divided by the total number of chromosomes in the population. GnomAD-SV PAFs (Collins et al., 2020) are for the entire population or B) the European population. C) The same comparison as in Fig. S3A with SSC parental PAFs estimated using the Hardy-Weinberg principle compared to gnomAD-SV PAFs in the entire population and D) gnomAD-SV European PAFs.

De novo insertions in ASD

We identified 158 *de novo* TEIs from all children (Supplementary Table 2.1). Previous studies have generally reported *de novo* TEI rates based on the number of insertions found in their cohort without accounting for detection sensitivity (Belyeu et al., 2021; Feusier et al., 2019). Multiple factors, including filtered regions, low sensitivity of the algorithm being used, or false negatives due to the sequencing methodology, result in an underestimate of true *de novo* rates. For example, TEI detection in Illumina short-read sequencing data is less sensitive than in long-read data, particularly for L1 TEIs (W. Zhou et al., 2020). Therefore, we adjusted the observed *de novo* rates to account for sensitivity loss and to obtain precise estimates. We obtained adjusted *de novo* rates of 1 in 29 births for Alu (95% CI 24-34), 1 in 104 births for L1 (95% CI 77-146), and 1 in 192 births for SVA (95% CI 127-309) (Figure 2.5A and Table 2.2).

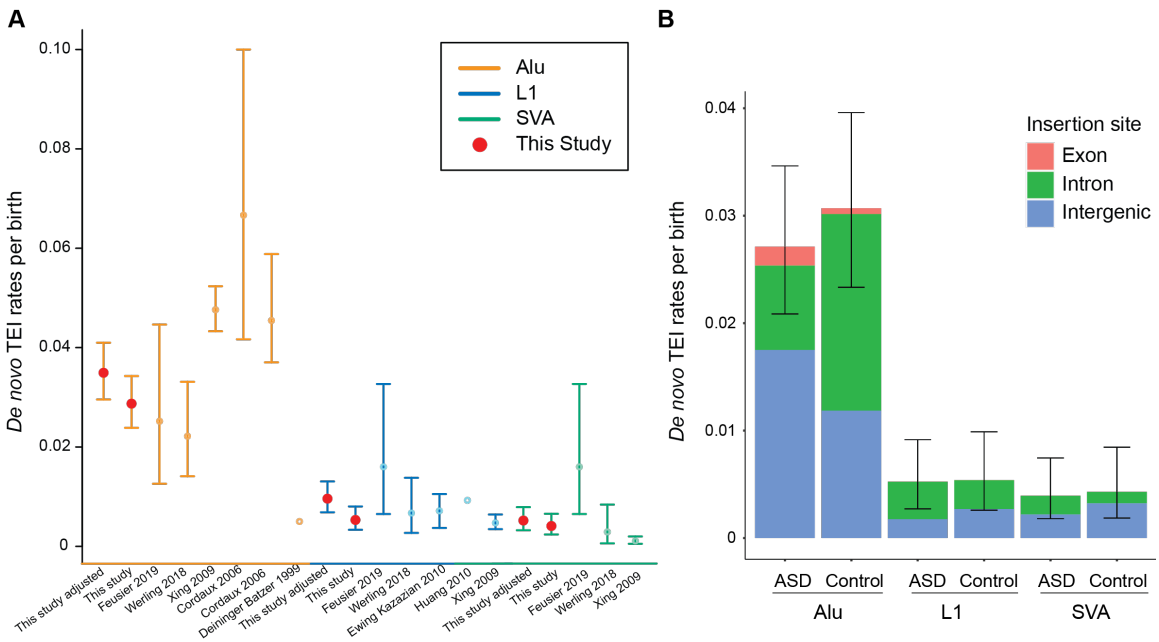

Figure 2.5 Rates of *de novo* TEIs.

A) Combined rates of *de novo* TEIs per birth for ASD and controls compared to previous studies. The adjusted rate in our study accounts for lower sensitivity for detecting TEIs in short-read Illumina data compared to long-read sequencing data. B) Rates of *de novo* TEIs per birth in probands with ASD and unaffected siblings (controls).

Table 2.2 ***De novo*** **insertion rates and sample sizes.**

	L1	Alu	SVA
ASD sample size	2,286	2,286	2,288
Controls sample size	1,856	1,857	1,860
ASD number of *de novo* insertions	12	62	9
Controls number of *de novo* insertions	10	57	8
ASD *de novo* rates	0.0052	0.0271	0.0039
Control *de novo* rates	0.0054	0.0307	0.0043
ASD + Control *de novo* insertions	22	119	17
ASD + Control sample size	4142	4143	4148
ASD + Control *de novo* rates	0.0053	0.0287	0.0041
1 in X births	188.27	34.82	244.00
Sensitivity 39.4x HG002	0.55	0.82	0.79
***De novo* rates adjusted**	0.0096	0.0349	0.0052
1 in X births adjusted	104.17	28.62	192.49
Confidence interval rates lower adjusted	0.0069	0.0296	0.0032
Confidence interval rates upper adjusted	0.0131	0.0410	0.0079
Confidence interval lower 1 in x adjusted	145.79	33.84	308.56
Confidence interval upper 1 in x adjusted	76.56	24.40	126.77

We detected 62 *de novo* Alu insertions in ASD (N=2,286) and 57 in controls (N=1,857), 12 *de novo* L1 insertions in ASD (N=2,286) and 10 in controls (N=1,856), and 9 *de novo* SVA insertions in ASD (N=2,288) and 8 in controls (N=1860) (Supplementary Table 2.1). We did not detect a difference in total *de novo* TEIs in ASD versus unaffected siblings (Figure 2.5B) but unexpectedly observed a higher rate of intronic Alu insertions in controls (p=0.003, two-sided Fisher's Exact Test) (Figure 2.5B). On the other hand, we observed a trend towards more exonic and intergenic Alu insertions in ASD than controls though not significant (p=0.388 for exonic insertions, p=0.157 for

intergenic insertions, two-sided Fisher's Exact Test) (Figure 2.5B) which leads to similar overall rates for *de novo* Alu insertions.

We observed *de novo* intronic L1 insertions in syndromic SFARI ASD genes (Abrahams et al., 2013) only in ASD and not in controls, and the rate in ASD was higher than expected (empirical two-sided p-value using 10,000 permutation runs, p=0.001, q-value=0.03) (A and Table 2.3). We also observed a trend for more *de novo* intronic L1 insertions in high pLI genes (Lek et al., 2016) in ASD than expected (empirical two-sided p-value, p=0.02, q-value > 0.05) (B). We observed *de novo* exonic insertions in genes with a high probability of LoF intolerance or haploinsufficiency (pLI $\geq$ 0.9) (Lek et al., 2016) only in affected individuals (Table 2.3 and Supplementary Table 2.1), including an exonic insertion in *CSDE1*, a gene recently implicated in patients with ASD and neurodevelopmental disabilities (Guo et al., 2019). There is a large overlap between SFARI genes and high pLI genes with *de novo* L1 insertions in cases; 80% (4/5) of SFARI genes with L1 insertions in ASD are also high pLI genes, suggesting that the *de novo* events can disrupt the haploinsufficient ASD genes and contribute to ASD risk (Table 2.3).

Table 2.3 Select *de novo* insertions in ASD and high pLI genes in affected individuals.

A subset of *de novo* TEIs observed in individuals with ASD in genes relevant to ASD or with a high probability of being loss-of-function intolerant (pLI>0.95) is shown. *SFARI annotations were obtained in 2019. LGD: likely gene disrupting; ID: intellectual disability, LoF: loss of function; ADHD: attention-deficit hyperactivity disorder, GI: gastrointestinal.

Proband ID	TE type	Gene	SFARI classification*	pLI	Genic region	Observed phenotype	Previous neurodevelopmental phenotype associated with gene	Reference
11859.p1	Alu	CSDE1	No	1	Exon	ASD, language delay, ID, macrocephaly, history of vision correction, normal EEG at 4 years	LGD variants associated with ASD, developmental delay, ID, seizures, macrocephaly, ADHD, anxiety, ocular abnormalities	(Guo et al., 2019)
14565.p1	Alu	KBTBD6	No	0.935	Exon	ASD, macrocephaly, uncoordinated, normal IQ, BMI Z-score -3.91		
12548.p1	Alu	APPBP2	No	0.999	Intron	ASD, normal IQ, macrocephaly		
12748.p1	Alu	SYT1	Syndromic	0.837	Intron	ASD, normal IQ, uncoordinated	Developmental delays, autistic features, hypotonia, ocular abnormalities, hyperkinetic movements associated with *de novo* missense variation	(Baker et al., 2018)
13931.p1	Alu	OTUD7A	Suggestive evidence	0.975	Intron	ASD, borderline IQ, normal EEG and brain imaging	Neurodevelopmental phenotype of ASD, developmental delay, ID, seizures associated with 15q13.3 microdeletion syndrome	(Yin et al., 2018), (Uddin et al., 2018)
13107.p1	Alu	TOX3	No	0.994	Intron	ASD, normal IQ		
14315.p1	Alu	JAZF1	No	0.958	Intron	ASD, borderline verbal IQ, normal nonverbal IQ, normal EEG		
11196.p1	L1	SRGAP3	Minimal Evidence	1	Intron	ASD, above average IQ, no history of seizures, heart problems reported	Case report of translocation breakpoint at loci posited to be LoF associated with hypotonia and severe ID	(Endris et al., 2002)
13684.p1	L1	HCN1	Syndromic	0.953	Intron	ASD, Tourette syndrome, above average IQ, GI problems, uncoordinated	Missense variation associated with syndrome of seizures, intellectual disability, and autistic features, gene also implicated in Tourette syndrome, role in striatal neuronal function and enteric nervous system	(Nava et al., 2014), (Tsetsos et al., 2020)
14080.p1	L1	DAB1	Hypothesized	0.981	Intron	ASD, uncoordinated, GI problems	ASD, GI problems, schizophrenia, spinocerebellar ataxia-37 associated with non-coding nucleotide repeats	(Corral-Juan et al., 2018), (Nawa et al., 2020)
14282.p1	L1	DPYD	Suggestive evidence	0	Intron	ASD, normal IQ	ASD, ID	(Carter et al., 2011), (Willemsen et al., 2011)
11234.p1	L1	NOTCH2	No	1	Intron	ASD, above average IQ		
13451.p1	L1	DPP10	Suggestive evidence	1	Intron	ASD, borderline IQ	ASD	(Marshall et al., 2008)
14404.p1	SVA	GRAMD1B	No	0.985	Intron	ASD, non-verbal, IQ in profound intellectual disability range, macrocephaly	Autosomal recessive intellectual disability	(Santos-Cortez et al., 2018)
14523.p1	SVA	ACACA	No	1	Intron	ASD, above average IQ, macrocephaly	Acetyl-CoA carboxylase deficiency	(Blom, de Muinck Keizer, & Scholte, 1981)

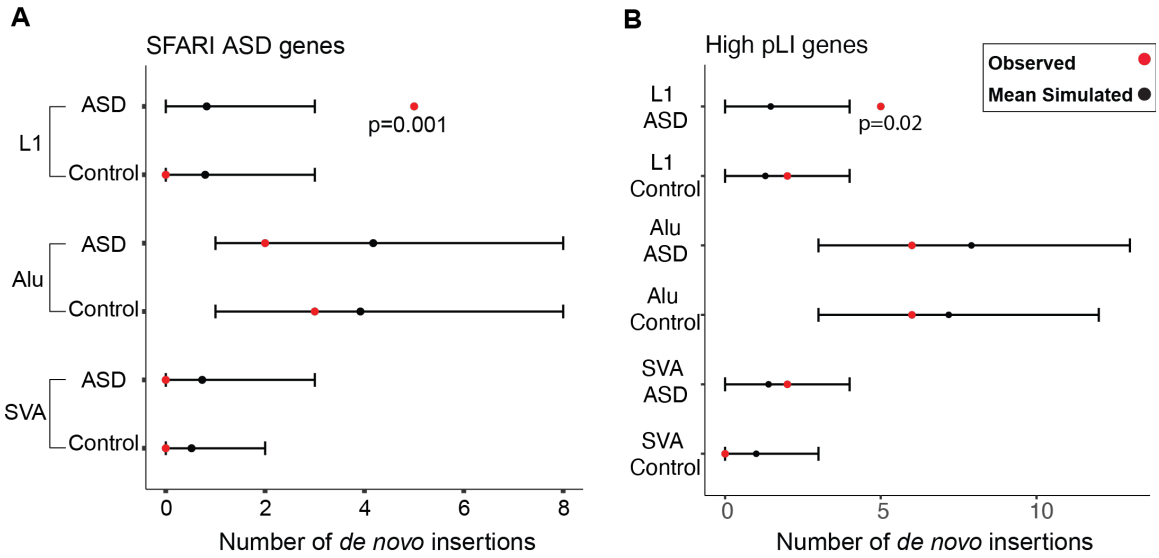

Figure 2.6 Enrichment of *de novo* TEIs in SFARI ASD and high probability of loss-of-function (pLI) intolerance genes.

A) Observed numbers of *de novo* TEIs in a list of complied ASD genes or B) high pLI genes (pLI >=0.90) (Lek et al., 2016) are marked by red dots. Black dots and lines represent mean numbers and 95% confidence intervals of expected TEIs based on 10,000 random simulations, respectively. More *de novo* L1 insertions in ASD genes than expected are observed in cases only. A trend for more L1 insertions in high pLI genes than expected is observed in cases (not significant after multiple testing correction with the Benjamini & Yekutieli method).

Genomic distribution of insertions

Since paternal and maternal age presents a risk to ASD (Croen, Najjar, Fireman, & Grether, 2007), we tested whether there was a difference in parental age at birth in children with and without *de novo* TEIs. We found a modest, but not significant, increase in paternal age for children with *de novo* TEIs compared to those without *de novo* TEIs (M= 33.94, SD= 5.63 *vs.* M=33.29, SD=4.71; t(163.42)=1.4452, p =0.1503) as well as increase in maternal age (M=31.62, SD=4.92 *vs*. M=31.12, SD=4.92; t(163.75)=1.29, p=0.198) (Figure 2.7). We also estimated the insertion size of polymorphic and *de novo* TEIs by mapping insertion-supporting reads from *xTea* output to TE consensus

sequences and obtaining the minimum and maximum mapping coordinates. The distribution of polymorphic L1 insertion size closely resembles previously published data (Gardner et al., 2017) (Figure 2.8A). Overall, *de novo* TEIs showed similar size distributions to polymorphic TEIs but had different patterns from somatic TEIs which showed more severe 5' truncation (E. Lee et al., 2012) (Figure 2.8B).

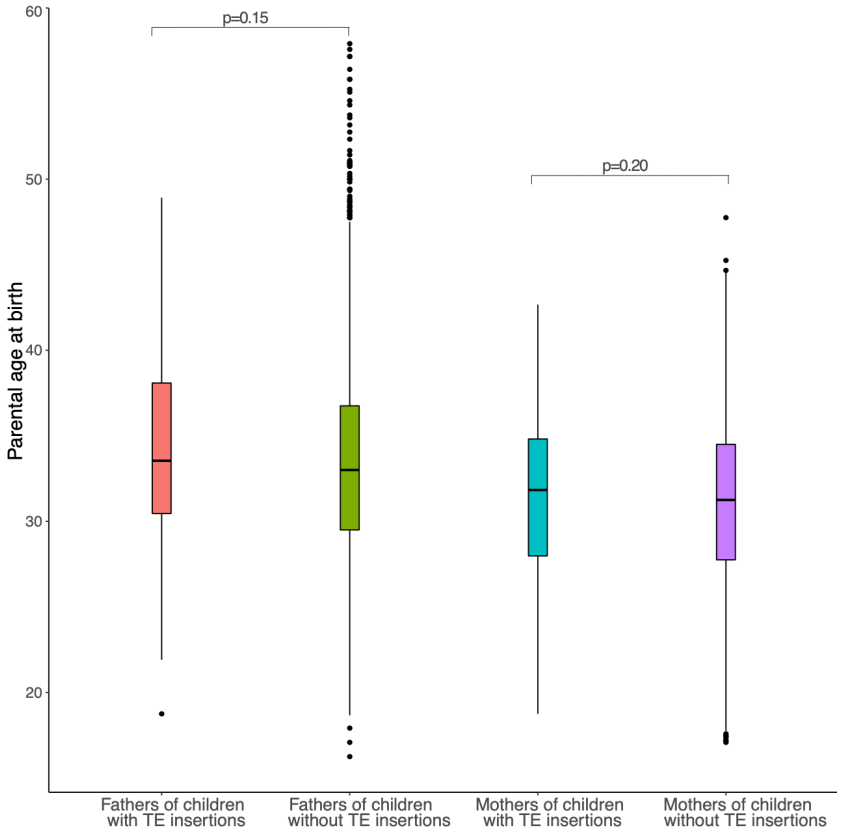

Figure 2.7 Parental age at birth of children with and without TEIs for cases and controls combined.

The median is represented with a line in the middle of the box plot and dots represent outlier samples.

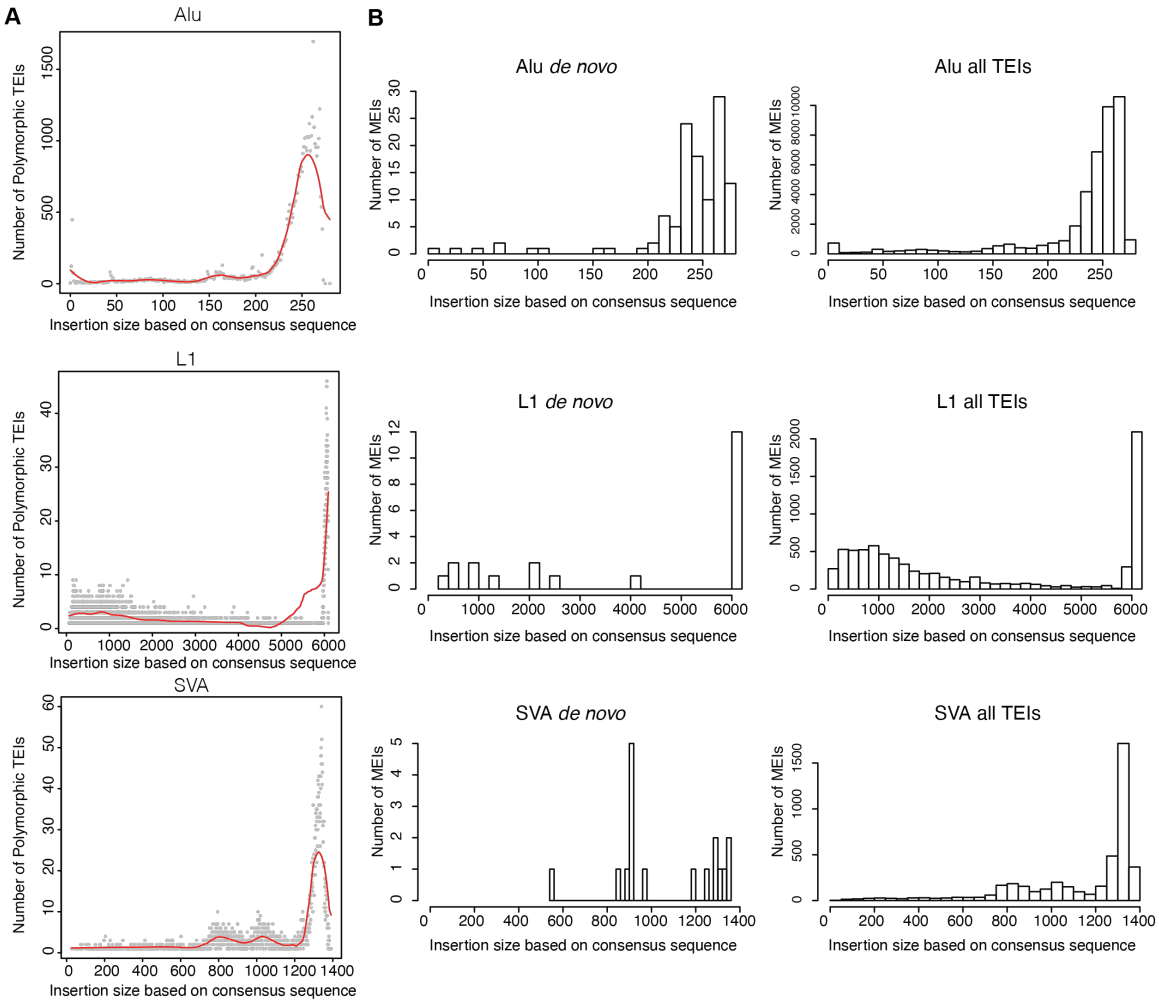

Figure 2.8 Estimated insertion size of TEIs.

A) Number of TEIs at a certain estimated insertion size for parental, ASD, and control individuals. The red line represents the Loess Regression with a 25% smoothing span. Alu N=42,045, L1 N=7,872, SVA N=4,375. B) Number of TEIs for each insertion size bin for *de novo* insertions or all insertions including both polymorphic and *de novo*. Since we used the position of clipped reads mapping to a consensus sequence to estimate size, this does not account for variable repeat expansion length or polyA tail variability. All: Alu N=42,045, L1 N=7,872, SVA N=4,375; *de novo:* Alu N=188, L1 N=22, SVA N=17.

Some genes with *de novo* TEIs in ASD are highly expressed in the brain at all stages of development (Table 2.4). We found an enrichment of *de novo* TEIs in ASD in genes upregulated in the prefrontal cortex, although this was not significant after multiple test correction (p-value=0.0017, Benjamini-Hochberg q-value=0.07), whereas no such enrichment was detected in controls. Additionally, we found that genes with *de novo*

TEIs were enriched for calcium-dependent phospholipid binding in ASD (adjusted p-value=0.034) but did not find enrichment for any Gene Ontology terms in controls. Several *de novo* TEIs were also observed in regions with enhancer and promoter chromatin marks in fetal brain development (Table 2.5). Thus, we evaluated the enrichment of polymorphic and *de novo* TEIs in different genomic and epigenomic regions using the Roadmap Epigenomics 25-state model (Roadmap Epigenomics et al., 2015). Polymorphic L1 and Alu insertions were depleted in exons, enhancers, and promoters (Figure 2.9A and Figure 2.9B; two-sided empirical $p<0.0005$, Benjamini–Yekutieli q-value < 0.0043 for each category) whereas SVAs did not show a significant depletion in those regions likely due to the limited number of insertions (Figure 2.9C and Table 2.6). *De novo* TEIs overall showed patterns within the expected ranges in most regions, however, we observed a trend for more *de novo* Alu insertions in active enhancer regions in the fetal brain in ASD than expected but not in controls (two-sided empirical p=0.018, Benjamini–Yekutieli q-value=0.3). This suggests the intriguing possibility that Alu insertions in neural enhancers might be a rare cause of ASD, though larger sample sizes are needed to test this.

Table 2.4 *De novo* insertions overlapping the top 10% expressed genes in the neocortex during development.

Overlap of genes with *de novo* insertions and the top 10% expressed genes in neocortex brain regions during development. Early prenatal: 8-19 postconceptional weeks (PCW), Late prenatal: 21-37 PCW, Childhood (4 months -11 years), Adolescence: 13-19 years, and Adulthood: 21-40 years. Genes with *de novo* SVA insertions did not overlap with any of the categories.

	Early Prenatal	Late Prenatal	Childhood	Adolescence	Adulthood
Genes with *de novo* Alu insertions in ASD	CSDE1 SYT1 KBTBD6 TCF25 EPS15	CSDE1 SYT1 TCF25 RPH3A EPS15	CSDE1 SYT1 TCF25 RPH3A EPS15	CSDE1 SYT1 TCF25 RPH3A EPS15	CSDE1 SYT1 TCF25 RPH3A EPS15
Genes with *de novo* Alu insertions in control	DCLK2 SF3A1	DCLK2	DCLK2	DCLK2	
Genes with *de novo* L1 insertions in ASD	DAB1				
Genes with *de novo* L1 insertions in control	EPHA7				

Table 2.5 Number of *de novo* insertions overlapping regions with epigenetic annotation in the fetal brain.

Sample and Genomic Region	Alu	L1	SVA
ASD introns	18	8	4
Control introns	34	5	2
ASD exons	4	0	0
Control Exons	1	0	0
ASD active enhancers	3	0	1
Control active enhancers	0	1	0
ASD other enhancers	4	0	2
Control other enhancers	1	1	0
ASD promoters	1	0	0
Control promoters	2	1	0

Table 2.6 Number of observed polymorphic insertions in parental SSC samples overlapping regions with epigenetic annotation in the fetal brain.

Genomic Region	Alu	L1	SVA
Introns	28,498	4,564	2,727
Exons	1,065	140	169
Active enhancers	254	44	34
Other enhancers	1,560	249	205
Promoters	449	42	81
Exon junctions	1,031	145	188

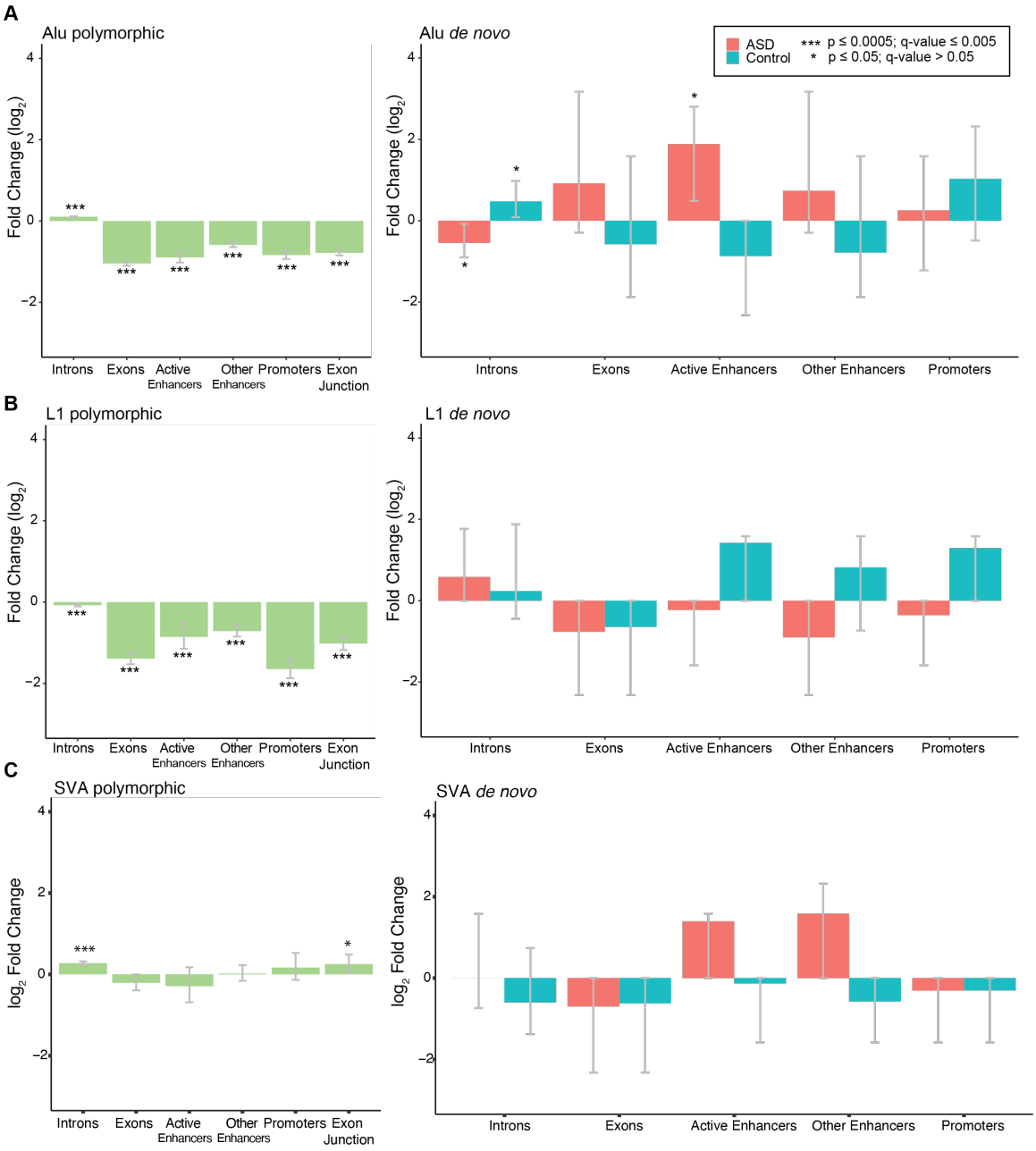

Figure 2.9 Genomic distribution of polymorphic and *de novo* TEIs.

10,000 random simulations were performed for both polymorphic and *de novo* TEIs based on the observed rates. The $\log_2$ fold change of observed compared to expected counts in different genomic regions is shown for coding and gene regulatory regions. 95% confidence intervals were estimated based on the empirical distribution of the random simulations. Polymorphic TEIs from parental individuals are depleted in exons and regulatory regions in the developing fetal brain. *De novo* Alu (A), L1 (B), and SVA TEIs (C) do not show this depletion compared to 10,000 random simulations. Two-sided empirical p-values and Benjamini–Yekutieli q-values based on multiple corrections of all enrichment and depletions performed are represented.

Experimental validation of ASD relevant TEIs

We selected *de novo* L1 and Alu insertions from both cases and controls in a subset of ASD and high pLI genes as well as in randomly selected genes for full-length PCR validation. We validated 22 of 23 (96%) Alu insertions and 6/7 (86%) L1 insertions, achieving a high validation rate of 93% (28/30). Validated insertions include a full-length *de novo* intronic L1 insertion in *DAB1,* a gene with a high probability of being loss-of-function intolerant (pLI=0.981) (Lek et al., 2016), and a hypothesized ASD gene (Abrahams et al., 2013; Nawa et al., 2020) implicated in regulating neuronal migration in development via the Reelin pathway in an isoform dependent manner (Gao & Godbout, 2013). We additionally validated an exonic Alu insertion in ASD gene *CSDE1* (Guo et al., 2019) in an ASD proband (Figure 2.10A). Our manual IGV inspection identified a single supporting clipped read at the breakpoint (Figure 2.10B) in the mother, suggesting that the exonic Alu insertion in *CSDE1* could be potentially mosaic at a low allelic fraction in the mother's blood, though low-level contamination from the proband's DNA cannot be completely ruled out. This insertion was fully validated in lymphoblastoid cell line (LCL) DNA in the individual with ASD and was absent in the mother, but LCLs might be expected to be limited in validating low-level mosaic variants (Figure 2.10A).

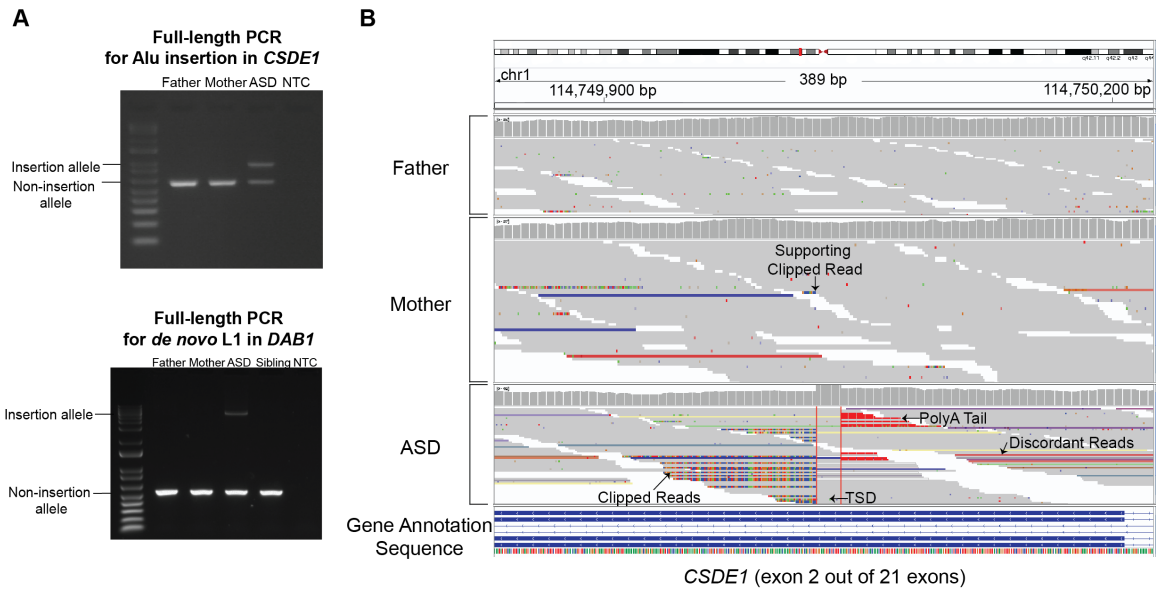

Figure 2.10 Full-length PCR validations and visual inspection.

A) Full-length PCR validation of the Alu insertion in *CSDE1* and the *de novo* L1 insertion in DAB1 in ASD cases. In lymphoblastoid cell line DNA, we validated the insertions in the ASD proband only. NTC: non-template control. B) Integrative Genomics Viewer image at the insertion site in gene *CSDE1* in an ASD case. For each individual, the sequencing coverage (top) and sequencing reads (bottom) are shown. The insertion shows the canonical signatures of target-primed reverse transcription (TPRT)-mediated retrotransposition: 15bp target site duplication (TSD) between the two insertion breakpoints, a poly-A tail, supporting clipped reads, and discordant reads with mates mapping to the consensus Alu sequence. The mother has one small clipped read sequence at the breakpoint that has the same sequence as in the proband, suggesting that the insertion could be mosaic at a low allele frequency in the mother's blood.

Discussion

In this chapter, we analyzed 2,288 a WGS dataset of 2,288 ASD families using *xTea*, a scalable and publicly available pipeline that we developed (Chu et al., 2021) to identify polymorphic and *de novo* TE insertions. Our systematic, large-scale genomic analysis has created the most rigorous and precise estimates of *de novo* TE insertions for each active TE family (L1, Alu, SVA) in the human population, adjusting for the limited detection sensitivity with short-read sequencing. We also created a catalog of 86,154 polymorphic TE insertions including >60% not previously reported and 158 *de*

novo insertions, and deposited it in dbVar*,* NCBI's database of human genomic structural variation, as a community data resource to enhance understanding of TE biology and population genetics and to facilitate identification of disease-associated insertions in many other genomic disease cohorts.

Overall, ASD cases and unaffected siblings had similar *de novo* TE insertion rates. However, only ASD cases were found to have more *de novo* L1 insertions than expected in known and hypothesized autism genes. Only in ASD did we observe *de novo* insertions in the exons of haploinsufficient genes, including a likely causative exonic insertion in *CSDE1,* resulting in an estimated rate of pathogenic exonic insertions of 1 in 2,288 ASD cases. Furthermore, the majority of *de novo* TE insertions detected were non-coding intronic insertions, which were enriched in active enhancers and found in highly expressed genes in the developing fetal brain and may have contributed to autism. Our comprehensive and focused analysis of a large WGS cohort not only produces a critical data resource of broad utility but also highlights an important impact of non-coding TE insertions in ASD implying diverse TE insertion-mediated pathogenic mechanisms beyond the insertional mutagenesis of protein-coding sequences.

The detection of TEIs in genome sequencing data requires specific pipelines, given their repetitive nature and short read length. These variants have previously been excluded from most routine genetic diagnoses and studies, including for ASD. Furthermore, accurate estimation of *de novo* TEIs in healthy individuals is important to understand the contribution of *de novo* TEIs in disease cohorts. Initial methods to determine de novo rates of TEIs relied on indirect methods that compared two reference genomes, making assumptions regarding the time to the most recent common ancestor between human reference genomes (Xing et al., 2009) and human-chimpanzee divergence time (Cordaux et al., 2006). To directly determine *de novo* retrotransposition

rates, large cohorts are necessary given the infrequency of these events. More recent studies using short-read sequencing technologies have included fewer than 1,000 families each, leading to uncertainties in estimates, especially for SVA insertions (Brandler et al., 2018; Feusier et al., 2019; Werling et al., 2018). They have also not accounted for the lower sensitivity of detection on TEIs using short-read sequencing (Belyeu et al., 2021). Compared to 1 in 20 (Cordaux et al., 2006) or 1 in 21 (Xing et al., 2009) Alu insertions per birth by earlier studies using evolutionary and mutational based methods, our estimate of 1 in 29 births is lower but within the range from more recent work using family genome sequencing data of 1 in 39.7 births (95% CI 22.4–79.4) (Feusier et al., 2019) (Figure 2.5A). L1 rates observed here of 1 in 104 births are also within the ranges observed previously of 1 in 63 births (95% CI 30.6–153.8) (Feusier et al., 2019) and 1 in 149.2 (95% CI 72.5-370.4) (Werling et al., 2018) but higher than the Xing et al. 2009 rate of 1 in 212 births (95% CI 156-289) (Xing et al., 2009) (Figure 2.5A). Our SVA *de novo* rates of 1 in 192 births are much higher than the Xing et al. 2009 rate of 1 in 916 births (95% CI 503-1,927) (Xing et al., 2009), but not as high as the Feusier et al. 2019 rate of 1 in 63 births (95% CI 30.6–153.8) (Figure 2.5A). The large sample size in our study produces more reliable estimates with smaller confidence intervals than previous analyses (Figure 2.5A), suggesting that our data provide the most accurate determination of TEI rates up to this time. Recently published work in this ASD cohort (Belyeu et al., 2021) detected fewer insertions and reported 31% (1 in 42 Alu), 55% (1 in 231 L1), and 38% (1 in 309 SVA) lower *de novo* insertion rates than ours, possibly due to their exclusion of mosaic insertions in their rate estimates, the use of a less sensitive pipeline (Gardner et al., 2017), and not adjusting for the lower sensitivity for detection of TEIs in short-read data.

Assigning causality to non-coding variants based on clinical phenotypes is challenging, given that most known ASD genes have been discovered in the context of coding LoF variants, yet the majority of individuals with ASD do not have LoF coding variants identified (Iossifov et al., 2014). To understand the clinical phenotypes of individuals with TEIs in high pLI (Lek et al., 2016) or known ASD genes (Abrahams et al., 2013), we reviewed the available clinical data and compared this to any known phenotypes associated with the gene, as well as the scientific literature more generally available (Table 2.3). Exonic insertions are likely to disrupt the coding sequence and are thus of particular interest. We observed one exonic Alu insertion in *CSDE1*, which has been recently associated with ASD (Guo et al., 2019). The affected proband shared clinical features, albeit non-specific, consistent with the previously described cohort, including ASD, intellectual disability, macrocephaly, and vision impairment. We additionally observed an exonic Alu insertion in *KBTBD6* (Table 2.3). Variation in this gene has not yet been associated with a reported neurodevelopmental phenotype that we are aware of. However, *KBTBD6* represents an intriguing candidate gene given its high pLI score (pLI=0.935) (Lek et al., 2016) as well as its molecular interactions with known ASD gene *CUL3*, to mediate the activity of another ASD gene, *RAC1* (Genau et al., 2015). Studying target genes of exonic *de novo* TEIs may shed novel biological insight not captured solely with more commonly studied forms of genetic variation in ASD.

We estimated a rate of underlying exonic TEIs of at least 1 in 2,288 in ASD, which is similar to the rate of 1 in 2,434 cases with developmental disorders reported in a recent exome sequencing study (Gardner et al., 2019). Although this is lower than other types of *de novo* genetic drivers of ASD, such as copy number variation, and the contribution of non-coding variants is thought to be smaller than coding LoF variants

(Werling et al., 2018), the strong depletion of polymorphic TEIs in regulatory non-coding regions and enrichment of large *de novo* L1 insertions (~6kb when full-length) in introns of ASD genes in cases but not in control suggest some of these non-coding events may contribute to ASD risk. Since intronic TEIs can affect gene function through various mechanisms, such as altering RNA expression and splicing (Hancks & Kazazian, 2016), TEIs contributing to ASD may present a phenotype different from known phenotypes caused by LoF coding variants or large CNVs in these genes. Including TEIs and structural variants in standard clinical genetic analyses for ASD will continue to expand our knowledge of non-coding variants and could increase the rates of genetic diagnoses. Our work also presents significant advances in scalable bioinformatic processing and identification of TEIs, which by their nature represent a challenging form of genomic variation to study. Future work, including both further development of computational methods, as well as experimental functional assessment of the effects and pathogenicity of non-coding TEIs, will be critical in understanding the role of these variants in autism.

Methods

Datasets and data processing with *xTea*

Whole-genome data from the SSC from phases: Pilot, Phase 1, Phase 2, Phase 3-1, Phase 3-1, and Phase 4 were analyzed. The analyzed data consists of ASD families with one affected individual, two unaffected parents, and for 1,860 of these families, one unaffected sibling was analyzed as the unaffected control. In order to process this massive amount of >9,000 individual whole genomes, we optimized for scalability a TEI detection computational tool, *xTea* (https://github.com/parklab/xTea), and implemented a dockerized version on Amazon Web Services where the SSC samples are hosted.

Tibanna (S. Lee et al., 2019) was used for managing jobs on AWS. For each job, a cloud instance with specific configurations was created, and each cram file was downloaded from the S3 bucket to the instance. *xTea* docker was pulled from dockerhub (https://hub.docker.com/repository/docker/warbler/xteab, v9). The reference genome and repeat libraries were also downloaded from the S3 bucket to each instance. *xTea* was run on each downloaded cram, and the results were compressed and saved to the Amazon S3 bucket. After removing outlier results and confirming that these were due to corrupted bam files with incomplete sequences or failed *xTea* runs, we analyzed WGS data from ~2,288 ASD-affected individuals and ~1,856 unaffected siblings with both parents sequenced (Table 2.1 and Table 2.2 for sample sizes per TE type). The approximate average sequencing depth, as determined by *xTea*, was 39.4x. Paired-end reads were 151 base pairs in length.

TEI identification with *xTea*

For each cram file, *xTea* ran three major steps to call TE insertions. First, raw candidate sites were collected based on whether there were enough qualified clipped reads at the breakpoints, where part of the read is aligned to the flanking region while the clipped part is well aligned to the consensus TE sequence. Second, for each passed candidate site we checked whether there was enough discordant reads support. Here, we consider a pair of reads with one read aligned to the flanking region and its mate aligned to the TE consensus sequence or other copies as discordant. Third, we ran TE family specific filters to reduce false positives in both polymorphic and *de novo* insertions).

TE family specific filters were implemented within *xTea* to remove false positives. In specific, while we used the default values for most of the parameters, there are three

major parameters (the number of clipped reads, the number of discordant pairs, and the number of clip and discordant reads) which can affect the sensitivity and specificity. We thoroughly evaluated these three parameters and required >= 3 clipped reads, >=5 discordant pairs, and >=1 clip and discordant pairs as the optimal ones to maintain high sensitivity and high specificity. As a consequence of target-primed reverse transcription, a poly(A) tail and TSD should be observed along with enough supporting clipped and discordant reads at both sides of the breakpoint. However, in many cases, not all of these features could be detected. *xTea* incorporates a confidence rating system that evaluates whether all these features are found and whether they are on one or both sides of the breakpoint. We selected only insertions classified as “high confidence”. Additional filters within *xTea* include examining the patterns of insertion-supporting clipped sequences and discordant reads mapped to the TE consensus sequences: the supporting reads should not be scattered across the consensus but instead form one cluster (c1) for 5’-clip reads and another cluster (c2) for 3’-clip reads; the mates of 3’ and 5' discordant reads should form two distinct clusters (d1 and d2). The distance between c1 and d2 and between c2 and d1 must be less than the average insert size ± 3*(standard deviation of the insert size). The supplementary *xTea* filter module with TE family specific filters implemented in our analysis is now part of the main *xTea* code in the latest version.

Annotation of non-redundant polymorphic TEIs

After obtaining the *xTea* high confidence insertions for each individual, we excluded calls where the clipped and discordant reads mapped above the consensus size *xTea* uses for mapping for AluY, L1HS, and SVA (282, 6,120, and 1,400 base pairs respectively). This removed some Alu insertions, which tended to be polyA expansion

artifacts. Since breakpoint positions can have slight differences between individuals, these insertions were given a 40 base pair margin from the midpoint of the breakpoints and were merged if they overlapped to obtain a unique set of non-redundant TEIs in the SSC cohort.

To determine whether insertions in the SSC cohort were known or novel, merged TEI calls were overlapped with the breakpoints from gnomAD (Collins et al., 2020), 1000 genomes (Gardner et al., 2017), or a compilation of other studies obtained from Evrony *et al.* (Evrony et al., 2015) to obtain insertions in our cohort that are not found in these studies (novel), known TEIs which overlap, as well as known TEIs which overlap to individual studies only. To obtain Venn diagrams for overlap with other cohorts, TEIs from unrelated parental individuals were given a 40 base pair margin from the midpoint of the breakpoints and were merged with bedtools (Quinlan & Hall, 2010) “merge” if they overlapped to obtain a unique set of non-redundant TEIs in unrelated individuals in the SSC cohort. Breakpoints from gnomAD (Collins et al., 2020) and 1000 genomes (Gardner et al., 2017) were also given a 40 base pair margin. The different datasets were overlapped using bedtools (Quinlan & Hall, 2010) “intersect” and counts were plotted in R with the VennDiagram library (H. Chen, 2018).

Calculation of TEI population allele frequency

The merged insertions were genotyped with the xTea genotyping module which uses a random forest model to genotype TEIs (https://github.com/parklab/xTea). The PAF was calculated using only parental genomes in the cohort, which were unaffected and unrelated. Specifically, PAF for each polymorphic TEI was defined as the number of alleles carrying the TEI in the parents divided by the total number of chromosomes in the population (i.e., 2 $\times$ the number of parents). As an additional approach, we estimated the

PAF within the SSC parental cohort using the Hardy-Weinberg principle. Here, p+q=1, where p is the frequency of the insertion allele in the population and q is the frequency of the non-insertion allele in the population. Assuming that q2 is the fraction of individuals without an insertion, we calculated the PAF as 1 - (sqrt((total parental individuals in the cohort - individuals with insertion allele) / total individuals in the cohort)). Merged breakpoints were overlapped with gnomAD TEIs (Collins et al., 2020) using a window of 40 base pairs to define overlap. We compared the PAF within the SSC cohort to both the overall PAF and the European PAF in gnomAD since 83% of fathers and 85% of mothers were classified as white (Figure 2.4).

Identification and rate estimation of *de novo* TEIs

To detect *de novo* insertions, we selected calls that did not have supporting reads in parental raw *xTea* output files and then confirmed this via manual inspection. We used the high confidence post-filtered insertions from *xTea* for this analysis for Alu, L1, and SVA (https://github.com/parklab/xTea). TEIs were given a 40 base pair margin from the midpoint of the breakpoints and were excluded if they overlapped with KNR insertions obtained from previous studies (Beck et al., 2010; A. D. Ewing & H. H. Kazazian, Jr., 2010; Ewing & Kazazian, 2011; Gardner et al., 2017; Hormozdiari et al., 2011; Huang et al., 2010; Iskow et al., 2010; Stewart et al., 2011; Wang et al., 2006) as well as reference SVA, reference young L1 (L1HS, L1PA2, L1PA3) or reference young Alu (AluY) (Smit, 2013-2015). To exclude inherited insertions that may have been missed in parents, we excluded insertions that had clipped or discordant reads in the raw parental files (clip_reads_tmp0 and discordant_reads_tmp0) in the *xTea* output.

We imaged *de novo* candidates on IGV 2.4.19 (Thorvaldsdottir, Robinson, & Mesirov, 2013) for manual inspection. We visually confirmed the absence of supporting

parental reads, as well as the presence of a TSD, a poly(A) tail, and clipped and discordant supporting reads that support a retrotransposition event (E. Lee et al., 2012). Insertions were scored as “high confidence *de novo*” if visual inspection of calls passed these criteria for *de novo* insertions, as “*de novo*” if there were some discordant reads in the parents but no clipped reads supporting the breakpoint in parents, as “somatic candidate” if there is strong read support with a poly(A) tail, TSD, clipped reads and discordant reads, but the supporting reads are a small fraction of the overall coverage at the breakpoint, as “parental mosaic candidate” if there were <=2 clipped reads in one of the parents and discordant reads at a low allele frequency, suggesting it might be mosaic in parent’s blood yet not called by *xTea* due to the low allele frequency, and “false negative parental” if it was not clear whether there was a false negative insertion or a mosaic blood insertion in the parental sample due to having few clipped reads but many discordant reads near the insertion site. Only insertions scored as “high confidence *de novo*”, “*de novo*”, “somatic candidate”, and “parental mosaic candidate” were included.

De novo retrotransposition rates were calculated as the number of *de novo* TEIs for both ASD affected and unaffected siblings divided by the total sample size. Samples that failed the *xTea* run were excluded from the analysis, resulting in a sample size of n=4,142 for L1, n=4,143 for Alu, and n=4,148 for SVA (Table 2.2). Rates and confidence intervals from previous studies were obtained from Feusier *et al*. (Feusier et al., 2019). The 95% confidence intervals for *de novo* rates in the SSC cohort were obtained in the same manner, with an exact binomial confidence interval estimate.

With long-read technologies, the sensitivity for detection of TEIs is higher (W. Zhou et al., 2020), suggesting that our raw rates are an underestimate. To account for genomic regions in which *xTea* is unable to detect TEIs given the lower sensitivity with

Illumina short-read data, as well as for the reference filters we used for *de novo* insertions, we calculated our sensitivity for detecting germline TEIs in the Genome in a Bottle sample NA24385/HG002 (Zook et al., 2016) which has been sequenced with both long and short-read technologies. A curated set of 9,970 (>50bp) insertions was obtained from Genome in a Bottle V0.6 (Zook et al., 2020) and integrated with 15,268 (>50bp) insertions from a haplotype assembly of the samples (H. Li, Feng, & Chu, 2020) (https://github.com/parklab/xTea). RepeatMasker (Smit, 2013-2015) was used to annotate Alu, L1, and SVA sequences, and then insertions were confirmed by manual inspection of poly-A tails and TSD or deletions on IGV (Robinson et al., 2011). This resulted in 1,642 (1,355 Alu, 197 L1, and 90 SVA) high-confidence TEIs detected in the NA24385/HG002 genome but not in the reference genome. We downsampled the HG002 Illumina bam file to the average coverage of SSC samples (39.4x) and detected TEIs with *xTea*. We excluded calls that overlapped reference SVA, reference young L1 (L1HS, L1PA2, L1PA3), or reference young Alu (AluY) (Smit, 2013-2015), as performed for the SSC analysis, and calculated the sensitivity of our pipeline to detect the set of curated TEIs. We adjusted the number of total ASD and control *de novo* insertions by dividing the observed rate by the sensitivity and obtained an exact binomial 95% confidence interval.

Annotation of TEIs

UCSC Table Browser (Karolchik et al., 2004) was used to obtain RefSeq gene annotations and coordinates for exons, and introns. ASD gene annotations were obtained from SFARI (Abrahams et al., 2013) in March 2019. Categories S (Syndromic),1 (High Confidence), 2 (Strong candidate), 3 (Suggestive evidence), 4 (Minimal evidence), and 5 (Hypothesized but untested) were included. *De novo* insertion

candidates were also annotated with the probability of being loss-of-function intolerant pLI (Lek et al., 2016).

Chromatin states from fetal brain tissue (E081 and E082) were downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/ (Roadmap Epigenomics et al., 2015). States were classified as: 13_EnhA1 and 12_EnhA2 = active enhancers; 15_EnhAF, 16_EnhW1, 17_EnhW2, 18_EnhAc = other enhancers (weak, flank, acetylation only); 2_PromU, 3_PromD1, 4_PromD2 = promoters. TEIs were overlapped with the two fetal brain regions using bedtools (Quinlan & Hall, 2010) and the number of unique calls in each category was obtained.

Enrichment analysis using simulated TEIs

We performed simulations to calculate the probability of the number of observed insertions in ASD genes, high pLI (pLI $\geq$ 0.9) genes, or in the annotated genomic regions with different chromatin states. Using the observed *de novo* TEIs candidates in ASD and unaffected siblings and the number of unique polymorphic insertions in parents, we simulated the same number of insertions of the same size in random regions of the genome, while also excluding the same young reference TE regions and KNR regions we excluded when detecting *de novo* TEIs for *de novo* simulations and excluding young reference TE regions for polymorphic insertions. We performed 10,000 simulations and determined the number of random insertions which overlapped a SFARI gene, high pLI gene, or region of interest per simulation. We determined if the observed value fell on the upper or lower end of the observed distribution, with a pseudo count of 1, to obtain a p-value. For example, the upper p-value is defined as (r+1)/(n+1), where r is the number of simulations greater or equal to the observed value and n is the number of simulations.

This value was multiplied by 2 for an empirical two-sided p-value. 95% CIs were calculated by obtaining the 0.025 and 0.975 percentiles of the null distribution. The log2 FC was calculated as the log2(observed value/mean of the null distribution) and the 95% CIs are plotted as their log2 values. These p-values were corrected for multiple testing with the Benjamini & Yekutieli method, to account for dependency between tests.

Mobile Element Insertion Size

To determine the size of polymorphic and *de novo* insertions, we only included TEIs which had supporting clipped reads on both breakpoints and which did not overlap with reference. Since *xTea* maps reads to several subfamilies of retrotransposons, we excluded calls where the clipped and discordant reads mapped above the consensus size *xTea* uses for mapping for AluY, L1HS, and SVA (282, 6,120, and 1,400 base pairs respectively) since, particularly for Alu calls, these tended to be poly(A) expansion artifacts. The consensus sequences within *xTea* were obtained from RepBase23.02 (Bao, Kojima, & Kohany, 2015) and the L1HS consensus sequence was manually constructed by multiple sequence alignment of full-length reference sequences. The position of clipped and discordant reads mapping to reference retrotransposon sequences was obtained for each insertion. The minimum position was subtracted from the maximum position to obtain the predicted size. If the maximum position was larger than the consensus length, this was set to the consensus length. The resulting estimated insertion size is an approximation since we are unable to account for repeat expansions and different poly(A) tail lengths (Grandi, Rosser, & An, 2013; Hancks & Kazazian, 2010). For all insertions including polymorphic and *de novo* TEIs, calls were given a 40 base pair margin from the midpoint of the breakpoints and were merged based on the overlap of these coordinates. The median size for all samples with each insertion is

reported. Loess regressions were performed in R (R Core Team, 2019) with a 25% smoothing span.

De novo insertions in brain expressed genes

For overlap of insertions with brain expressed genes, we selected the neocortex regions from Brainspan (Hawrylycz et al., 2012; Miller et al., 2014) (ventrolateral prefrontal cortex (VFC), dorsolateral prefrontal cortex (DFC), medial prefrontal cortex (MFC), primary visual cortex (V1C), primary motor cortex (M1C), orbitofrontal cortex (OFC), primary association cortex (A1C), inferior parietal cortex (IPC), primary somatosensory cortex (S1C), superior temporal cortex (STC), inferior temporal cortex (ITC)) (Parikshak et al., 2013), and obtained the mean expression of each gene per sample in these tissues and then obtained the mean expression per age group for the following categories: Early prenatal: 8-19 PCW, Late prenatal: 21-37 PCW, Childhood (4 months -11 years), Adolescence: 13-19 years, and Adulthood: 21-40 years (Table 2.4). We then overlapped genes with insertions in ASD and controls with the top 10% of gene expression observed.

Gene list enrichment in genes with TEIs

We tested whether genes with TEIs in ASD or controls were enriched for genes overexpressed in tissues in the Human Gene Atlas list using Enrichr (E. Y. Chen et al., 2013). We also tested for enrichment of gene ontology terms in the subset of genes with TEIs using g:Profiler (Raudvere et al., 2019) in only annotated genes, with a user threshold of 0.05 and a significant threshold for multiple testing correction with the g:SCS threshold method.

PCR validations

24 cases were chosen for full-length PCR validation based on their clinical relevance by selecting variants that occurred in SFARI, high pLI, or brain expressed genes and 13 cases were randomly selected for a total of 12 L1 insertions and 25 Alu insertions. We developed a pipeline for designing specific primers and tested and optimized the PCR protocols for each primer pair in control DNA before validating them in the SSC samples. We excluded events overlapping duplicated regions or reference insertions of the same class to reduce amplification artifacts. A custom pipeline, based on a previously developed pipeline (Evrony et al., 2012), was used to obtain primer sequences for full-length validation. Sequences from -800 to -100 and +100 to +800 base pairs from the insertion breakpoint were used to select primers with Primer3 (Untergasser et al., 2012). InSilico PCR from UCSC (https://genome.ucsc.edu/cgi-bin/hgPcr) was then implemented to assess whether these primers would amplify a unique region in the genome. Blat (Kent, 2002) (-stepSize=5 -minScore=20 -minIdentity=80) was then used to confirm unique mapping to the genome. If these steps failed, a masked genome (https://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigZips/hg38.fa.masked.gz) was used for the first step.

Validations were performed with 20ng of DNA from each available family member from lymphoblastoid cell lines which were provided by the Rutgers University Cell and DNA Repository. This was done by confirming the presence of both an insertion and a non-insertion allele band near or at the expected insertion size in the samples with

predicted insertions, and only a non-insertion allele band in the other family members. Water was used instead of DNA as a non-template control for each primer pair.

PCRs were performed using Phusion Hot Start II High-Fidelity DNA Polymerase (F549L, Thermo Fisher Scientific) (Table 2.7 and Table 2.8). Primers were tested and optimized using the Genome in a Bottle sample NA24385/HG002 (Zook et al., 2016), where we had sequencing data and high confidence insertions from the gold standard available. DNA was quantified using a Quant-iT™ dsDNA Assay Kit (Q33120, Thermo Fisher Scientific) before running at least 70 ng of PCR product, when possible, on a 2% agarose gel (for Alu) or a 1% agarose gel and with a Genomic DNA ScreenTape Analysis (5067-5366, Agilent) on an Agilent TapeStation (for L1) for a higher resolution at determining the insertion amplicon size. A 1kb Plus DNA ladder (10787-026, Invitrogen) was used.

Table 2.7 Protocols for PCR.

Reagent	20µl reaction (µl)
H2O	11.4
5X Phusion HS Buffer	4
10mM dNTPs	0.4
Primer 1 10µM (to 0.5µM)	1
Primer 2 10µM (to 0.5µM)	1
DNA (10ng/µl)	2
Phusion Hot Start II DNA Polymerase (2U/µl)	0.2
Total	20

Table 2.8 PCR cycling instructions.

	Temp	Time Alu	Time L1	Cycles
Denaturation	98°C	30 s	30 s	1
Denaturation	98°C	5 s	10 s	
Annealing	68°C	20 s	20 s	28
Extension	72°C	27s	3 mins 45 s	
Final extension	72°C	5 min	8 min	1
Hold	4°C	inf	inf	

Some primer pairs produced additional amplification bands or artifact bands and were further optimized by increasing the annealing temperature and/or decreasing the number of amplification cycles, and some primer pairs produced lower concentrations of DNA and were optimized by increasing the number of amplification cycles and/or decreasing the annealing temperature. If primer pairs did not amplify a unique non-insertion allele and had artifact bands, we did not proceed with validation of those insertions using those primers. We were unable to confirm the presence of positive control germline SVA TEIs in the control DNA NA24385/HG002 sample with our PCR conditions and therefore did not test any *de novo* SVA insertions. Out of 12 L1 primer pairs designed for validations of *de novo* insertions, we were able to optimize 9 primer pairs, and we optimized 23 Alu primer pairs out of 25. 2 of the L1 primers were previously classified as mosaic candidates in 1 case and 1 control and were considered separately for validation rates. These 2 cases did not validate in lymphoblastoid cell line DNA.

References

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., . . . Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism, 4*(1), 36. doi:10.1186/2040-2392-4-36

Alonso-Gonzalez, A., Rodriguez-Fontenla, C., & Carracedo, A. (2018). De novo Mutations (DNMs) in Autism Spectrum Disorder (ASD): Pathway and Network Analysis. *Front Genet, 9*, 406. doi:10.3389/fgene.2018.00406

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (Fifth edition. ed.). Washington, DC: American Psychiatric Publishing.

Baker, K., Gordon, S. L., Melland, H., Bumbak, F., Scott, D. J., Jiang, T. J., . . . Raymond, F. L. (2018). SYT1-associated neurodevelopmental disorder: a case series. *Brain, 141*(9), 2576-2591. doi:10.1093/brain/awy209

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA, 6*, 11. doi:10.1186/s13100-015-0041-9

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., . . . Moran, J. V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell, 141*(7), 1159-1170. doi:10.1016/j.cell.2010.05.021

Belyeu, J. R., Brand, H., Wang, H., Zhao, X., Pedersen, B. S., Feusier, J., . . . Quinlan, A. R. (2021). De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am J Hum Genet*. doi:10.1016/j.ajhg.2021.02.012

Blom, W., de Muinck Keizer, S. M., & Scholte, H. R. (1981). Acetyl-CoA carboxylase deficiency: an inborn error of de novo fatty acid synthesis. *N Engl J Med, 305*(8), 465-466. doi:10.1056/NEJM198108203050820

Bogaerts-Marquez, M., Barron, M. G., Fiston-Lavier, A. S., Vendrell-Mir, P., Castanera, R., Casacuberta, J. M., & Gonzalez, J. (2020). T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. *Bioinformatics, 36*(4), 1191-1197. doi:10.1093/bioinformatics/btz727

Brandler, W. M., Antaki, D., Gujral, M., Kleiber, M. L., Whitney, J., Maile, M. S., . . . Sebat, J. (2018). Paternally inherited cis-regulatory structural variants are associated with autism. *Science, 360*(6386), 327-331. doi:10.1126/science.aan2261

Campbell, C. D., & Eichler, E. E. (2013). Properties and rates of germline mutations in humans. *Trends Genet, 29*(10), 575-584. doi:10.1016/j.tig.2013.04.005

Carter, M. T., Nikkel, S. M., Fernandez, B. A., Marshall, C. R., Noor, A., Lionel, A. C., . . . Scherer, S. W. (2011). Hemizygous deletions on chromosome 1p21.3 involving the DPYD gene in individuals with autism spectrum disorder. *Clin Genet, 80*(5), 435-443. doi:10.1111/j.1399-0004.2010.01578.x

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., . . . Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics, 14*, 128. doi:10.1186/1471-2105-14-128

Chen, H. (2018). VennDiagram: Generate High-Resolution Venn and Euler Plots. R package

version 1.6.19. https://CRAN.R-project.org/package=VennDiagram.
Chu, C., Borges-Monroy, R., Viswanadham, V. V., Lee, S., Li, H., Lee, E. A., & Park, P. J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*.
Chu, C., Zhao, B., Park, P. J., & Lee, E. A. (2020). Identification and Genotyping of Transposable Element Insertions From Genome Sequencing Data. *Curr Protoc Hum Genet, 107*(1), e102. doi:10.1002/cphg.102
Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alfoldi, J., Francioli, L. C., . . . Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature, 581*(7809), 444-451. doi:10.1038/s41586-020-2287-8
Cordaux, R., Hedges, D. J., Herke, S. W., & Batzer, M. A. (2006). Estimating the retrotransposition rate of human Alu elements. *Gene, 373*(Supplement C), 134-137. doi:https://doi.org/10.1016/j.gene.2006.01.019
Corral-Juan, M., Serrano-Munuera, C., Rabano, A., Cota-Gonzalez, D., Segarra-Roca, A., Ispierto, L., . . . Matilla-Duenas, A. (2018). Clinical, genetic and neuropathological characterization of spinocerebellar ataxia type 37. *Brain, 141*(7), 1981-1997. doi:10.1093/brain/awy137
Croen, L. A., Najjar, D. V., Fireman, B., & Grether, J. K. (2007). Maternal and paternal age and risk of autism spectrum disorders. *Arch Pediatr Adolesc Med, 161*(4), 334-340. doi:10.1001/archpedi.161.4.334
De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., . . . Buxbaum, J. D. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature, 515*(7526), 209-215. doi:10.1038/nature13772
Doan, R. N., Lim, E. T., De Rubeis, S., Betancur, C., Cutler, D. J., Chiocchetti, A. G., . . . Yu, T. W. (2019). Recessive gene disruptions in autism spectrum disorder. *Nat Genet, 51*(7), 1092-1098. doi:10.1038/s41588-019-0433-8
Endris, V., Wogatzky, B., Leimer, U., Bartsch, D., Zatyka, M., Latif, F., . . . Rappold, G. A. (2002). The novel Rho-GTPase activating gene MEGAP/ srGAP3 has a putative role in severe mental retardation. *Proc Natl Acad Sci U S A, 99*(18), 11754-11759. doi:10.1073/pnas.162241099
Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., . . . Walsh, C. A. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell, 151*(3), 483-496. doi:10.1016/j.cell.2012.09.035
Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., . . . Walsh, C. A. (2015). Cell lineage analysis in human brain using endogenous retroelements. *Neuron, 85*(1), 49-59. doi:10.1016/j.neuron.2014.12.028
Ewing, A., & Kazazian, H. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome research, 20*(9), 1262-1270. doi:citeulike-article-id:7209974
doi: 10.1101/gr.106419.110
Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mob DNA, 6*, 24. doi:10.1186/s13100-015-0055-3
Ewing, A. D., & Kazazian, H. H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res, 20*(9), 1262-1270. doi:10.1101/gr.106419.110
Ewing, A. D., & Kazazian, H. H., Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res, 21*(6), 985-990. doi:10.1101/gr.114777.110

Feusier, J., Watkins, W. S., Thomas, J., Farrell, A., Witherspoon, D. J., Baird, L., . . . Jorde, L. B. (2019). Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res, 29*(10), 1567-1577. doi:10.1101/gr.247965.118

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron, 68*(2), 192-195. doi:10.1016/j.neuron.2010.10.006

Gao, Z., & Godbout, R. (2013). Reelin-Disabled-1 signaling in neuronal migration: splicing takes the stage. *Cell Mol Life Sci, 70*(13), 2319-2329. doi:10.1007/s00018-012-1171-6

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., . . . Devine, S. E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res, 27*(11), 1916-1929. doi:10.1101/gr.218032.116

Gardner, E. J., Prigmore, E., Gallone, G., Danecek, P., Samocha, K. E., Handsaker, J., . . . Hurles, M. E. (2019). Contribution of retrotransposition to developmental disorders. *Nat Commun, 10*(1), 4630. doi:10.1038/s41467-019-12520-y

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., . . . Buxbaum, J. D. (2014). Most genetic risk for autism resides with common variation. *Nat Genet, 46*(8), 881-885. doi:10.1038/ng.3039

Genau, H. M., Huber, J., Baschieri, F., Akutsu, M., Dotsch, V., Farhan, H., . . . Behrends, C. (2015). CUL3-KBTBD6/KBTBD7 ubiquitin ligase cooperates with GABARAP proteins to spatially restrict TIAM1-RAC1 signaling. *Mol Cell, 57*(6), 995-1010. doi:10.1016/j.molcel.2014.12.040

Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nat Rev Genet, 19*(11), 688-704. doi:10.1038/s41576-018-0050-x

Goodier, J. L., Ostertag, E. M., & Kazazian, H. H., Jr. (2000). Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet, 9*(4), 653-657. doi:10.1093/hmg/9.4.653

Grandi, F. C., Rosser, J. M., & An, W. (2013). LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. *Mol Biol Evol, 30*(3), 503-512. doi:10.1093/molbev/mss251

Guo, H., Li, Y., Shen, L., Wang, T., Jia, X., Liu, L., . . . Xia, K. (2019). Disruptive variants of CSDE1 associate with autism and interfere with neuronal development and synaptic transmission. *Sci Adv, 5*(9), eaax2166. doi:10.1126/sciadv.aax2166

Hancks, D. C., & Kazazian, H. H., Jr. (2010). SVA retrotransposons: Evolution and genetic instability. *Semin Cancer Biol, 20*(4), 234-245. doi:10.1016/j.semcancer.2010.04.001

Hancks, D. C., & Kazazian, H. H., Jr. (2016). Roles for retrotransposon insertions in human disease. *Mob DNA, 7*, 9. doi:10.1186/s13100-016-0065-9

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., . . . Jones, A. R. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature, 489*(7416), 391-399. doi:10.1038/nature11405

Hormozdiari, F., Alkan, C., Ventura, M., Hajirasouliha, I., Malig, M., Hach, F., . . . Eichler, E. E. (2011). Alu repeat discovery and characterization within human genomes. *Genome Res, 21*(6), 840-849. doi:10.1101/gr.115956.110

Huang, C. R., Schneider, A. M., Lu, Y., Niranjan, T., Shen, P., Robinson, M. A., . . . Burns, K. H. (2010). Mobile interspersed repeats are major structural variants in the human genome. *Cell, 141*(7), 1171-1182. doi:10.1016/j.cell.2010.05.026

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., . . . Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature, 515*(7526), 216-221. doi:10.1038/nature13908

Iskow, R. C., McCabe, M. T., Mills, R. E., Torene, S., Pittard, W. S., Neuwald, A. F., . . . Devine, S. E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell, 141*(7), 1253-1261. doi:10.1016/j.cell.2010.05.020

Jiang, C., Chen, C., Huang, Z., Liu, R., & Verdier, J. (2015). ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics, 16*, 72. doi:10.1186/s12859-015-0507-2

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res, 32*(Database issue), D493-496. doi:10.1093/nar/gkh103

Kazazian, H. H., Jr., & Moran, J. V. (1998). The impact of L1 retrotransposons on the human genome. *Nat Genet, 19*(1), 19-24. doi:10.1038/ng0598-19

Keane, T. M., Wong, K., & Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics, 29*(3), 389-390. doi:10.1093/bioinformatics/bts697

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664. doi:10.1101/gr.229202

Kosmicki, J. A., Samocha, K. E., Howrigan, D. P., Sanders, S. J., Slowikowski, K., Lek, M., . . . Daly, M. J. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet, 49*(4), 504-510. doi:10.1038/ng.3789

Kroon, M., Lameijer, E. W., Lakenberg, N., Hehir-Kwa, J. Y., Thung, D. T., Slagboom, P. E., . . . Ye, K. (2016). Detecting dispersed duplications in high-throughput sequencing data using a database-free approach. *Bioinformatics, 32*(4), 505-510. doi:10.1093/bioinformatics/btv621

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., 3rd, . . . Cancer Genome Atlas Research, N. (2012). Landscape of somatic retrotransposition in human cancers. *Science, 337*(6097), 967-971. doi:10.1126/science.1222077

Lee, S., Johnson, J., Vitzthum, C., Kirli, K., Alver, B. H., & Park, P. J. (2019). Tibanna: software for scalable execution of portable pipelines on the cloud. *Bioinformatics, 35*(21), 4424-4426. doi:10.1093/bioinformatics/btz379

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. doi:10.1038/nature19057

Levy, D., Ronemus, M., Yamrom, B., Lee, Y. H., Leotta, A., Kendall, J., . . . Wigler, M. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron, 70*(5), 886-897. doi:10.1016/j.neuron.2011.05.015

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., . . . Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biol, 5*(10), e254. doi:10.1371/journal.pbio.0050254

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biol, 21*(1), 265. doi:10.1186/s13059-020-02168-z

Maenner, M. J., Shaw, K. A., Baio, J., Washington, A., Patrick, M., DiRienzo, M., . . . Dietz, P. M. (2020). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. *MMWR Surveill Summ, 69*(4), 1-12. doi:10.15585/mmwr.ss6904a1

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., . . . Scherer, S. W. (2008). Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet, 82*(2), 477-488. doi:10.1016/j.ajhg.2007.12.009

Miller, J. A., Ding, S. L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., . . . Lein, E. S. (2014). Transcriptional landscape of the prenatal human brain. *Nature, 508*(7495), 199-206. doi:10.1038/nature13185

Nava, C., Dalle, C., Rastetter, A., Striano, P., de Kovel, C. G., Nabbout, R., . . . Depienne, C. (2014). De novo mutations in HCN1 cause early infantile epileptic encephalopathy. *Nat Genet, 46*(6), 640-645. doi:10.1038/ng.2952

Nawa, Y., Kimura, H., Mori, D., Kato, H., Toyama, M., Furuta, S., . . . Ozaki, N. (2020). Rare single-nucleotide DAB1 variants and their contribution to Schizophrenia and autism spectrum disorder susceptibility. *Hum Genome Var, 7*(1), 37. doi:10.1038/s41439-020-00125-7

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., . . . Eichler, E. E. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature, 485*(7397), 246-250. doi:10.1038/nature10989

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., . . . Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell, 155*(5), 1008-1021. doi:10.1016/j.cell.2013.10.031

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rajaby, R., & Sung, W. K. (2018). TranSurVeyor: an improved database-free algorithm for finding non-reference transpositions in high-throughput sequencing data. *Nucleic Acids Res, 46*(20), e122. doi:10.1093/nar/gky685

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res, 47*(W1), W191-W198. doi:10.1093/nar/gkz369

Rishishwar, L., Marino-Ramirez, L., & Jordan, I. K. (2017). Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinform, 18*(6), 908-918. doi:10.1093/bib/bbw072

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi:10.1038/nature14248

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., . . . State, M. W. (2015a). Insights into Autism Spectrum Disorder Genomic

Architecture and Biology from 71 Risk Loci. *Neuron, 87*(6), 1215-1233. doi:10.1016/j.neuron.2015.09.016

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., . . . State, M. W. (2015b). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron, 87*(6), 1215-1233. doi:10.1016/j.neuron.2015.09.016

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., . . . State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature, 485*(7397), 237-241. doi:10.1038/nature10945

Santos-Cortez, R. L. P., Khan, V., Khan, F. S., Mughal, Z. U., Chakchouk, I., Lee, K., . . . Leal, S. M. (2018). Novel candidate genes and variants underlying autosomal recessive neurodevelopmental disorders with intellectual disability. *Hum Genet, 137*(9), 735-752. doi:10.1007/s00439-018-1928-6

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J. Y., . . . Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*. doi:10.1016/j.cell.2019.12.036

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., . . . Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science, 316*(5823), 445-449. doi:10.1126/science.1138659

Smit, A., Hubley, R & Green, P. . (2013-2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Stewart, C., Kural, D., Stromberg, M. P., Walker, J. A., Konkel, M. K., Stutz, A. M., . . . Genomes, P. (2011). A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet, 7*(8), e1002236. doi:10.1371/journal.pgen.1002236

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017

Thung, D. T., de Ligt, J., Vissers, L. E., Steehouwer, M., Kroon, M., de Vries, P., . . . Hehir-Kwa, J. Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol, 15*(10), 488. doi:10.1186/s13059-014-0488-x

Torene, R. I., Galens, K., Liu, S., Arvai, K., Borroto, C., Scuffins, J., . . . Retterer, K. (2020). Mobile element insertion detection in 89,874 clinical exomes. *Genet Med, 22*(5), 974-978. doi:10.1038/s41436-020-0749-x

Tsetsos, F., Yu, D., Sul, J. H., Huang, A. Y., Illmann, C., Osiecki, L., . . . Paschou, P. (2020). Synaptic processes and immune-related pathways implicated in Tourette Syndrome. *medRxiv*, 2020.2004.2024.20047845. doi:10.1101/2020.04.24.20047845

Tubio, J. M. C., Li, Y., Ju, Y. S., Martincorena, I., Cooke, S. L., Tojo, M., . . . Group, I. P. C. (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science, 345*(6196), 1251343. doi:10.1126/science.1251343

Uddin, M., Unda, B. K., Kwan, V., Holzapfel, N. T., White, S. H., Chalil, L., . . . Singh, K. K. (2018). OTUD7A Regulates Neurodevelopmental Phenotypes in the 15q13.3 Microdeletion Syndrome. *Am J Hum Genet, 102*(2), 278-295. doi:10.1016/j.ajhg.2018.01.006

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Res, 40*(15), e115. doi:10.1093/nar/gks596

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M. A., & Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat, 27*(4), 323-329. doi:10.1002/humu.20307

Werling, D. M., Brand, H., An, J. Y., Stone, M. R., Zhu, L., Glessner, J. T., . . . Sanders, S. J. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet, 50*(5), 727-736. doi:10.1038/s41588-018-0107-y

Willemsen, M. H., Valles, A., Kirkels, L. A., Mastebroek, M., Olde Loohuis, N., Kos, A., . . . Kleefstra, T. (2011). Chromosome 1p21.3 microdeletions comprising DPYD and MIR137 are associated with intellectual disability. *J Med Genet, 48*(12), 810-818. doi:10.1136/jmedgenet-2011-100294

Wu, J., Lee, W. P., Ward, A., Walker, J. A., Konkel, M. K., Batzer, M. A., & Marth, G. T. (2014). Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics, 15*, 795. doi:10.1186/1471-2164-15-795

Xing, J., Zhang, Y., Han, K., Salem, A. H., Sen, S. K., Huff, C. D., . . . Jorde, L. B. (2009). Mobile elements create structural variation: analysis of a complete human genome. *Genome Res, 19*(9), 1516-1526. doi:10.1101/gr.091827.109

Yin, J., Chen, W., Chao, E. S., Soriano, S., Wang, L., Wang, W., . . . Schaaf, C. P. (2018). Otud7a Knockout Mice Recapitulate Many Neurological Features of 15q13.3 Microdeletion Syndrome. *Am J Hum Genet, 102*(2), 296-308. doi:10.1016/j.ajhg.2018.01.005

Zhou, B., Arthur, J. G., Ho, S. S., Pattni, R., Huang, Y., Wong, W. H., & Urban, A. E. (2018). Extensive and deep sequencing of the Venter/HuRef genome for developing and benchmarking genome analysis tools. *Sci Data, 5*, 180261. doi:10.1038/sdata.2018.261

Zhou, W., Emery, S. B., Flasch, D. A., Wang, Y., Kwan, K. Y., Kidd, J. M., . . . Mills, R. E. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res, 48*(3), 1146-1163. doi:10.1093/nar/gkz1173

Zhuang, J., Wang, J., Theurkauf, W., & Weng, Z. (2014). TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res, 42*(11), 6826-6838. doi:10.1093/nar/gku323

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., . . . Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data, 3*, 160025. doi:10.1038/sdata.2016.25

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., . . . Salit, M. (2020). A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*. doi:10.1038/s41587-020-0538-8

Chapter 3.
Somatic transposable element insertions in the brain at a single-cell resolution

Attribution

Rebeca Borges Monroy primarily contributed to the study design, analysis of data and wrote this chapter. Eunjung Alice Lee and Kyu Park developed and modified the memTea algorithm used in the analysis. The code used to detect somatic mutations in data analyzed by memTea was adapted from code written by Eunjung Alice Lee. Junho Kim developed a pipeline for annotating poly(A) tract artifacts in single-cell data. Chong Chu developed the xTea algorithm used in the analysis. At the time of writing this, the manuscript for this algorithm is currently in press in a manuscript called “Comprehensive identification of transposable element insertions using multiple sequencing technologies”, and will be available in Nature Communications, https://doi.org/10.1038/s41467-021-24041-8. Rebeca Borges-Monroy is listed as a second author in this work. Phased deletions in single cells were detected by Junho Kim using an algorithm that he developed. The single-cell data used in this analysis was experimentally obtained by Rachel E. Rodin, Michael A. Lodato, Michael E. Coulter, Michael B. Miller, Sangita Choudhury, Javier Ganz, Sarah Bizzotto, and Zinan Zhou. Eunjung Alice Lee and Chris A. Walsh contributed to the design of the study and supervised the study.

Summary

Retrotransposons can cause mutations in the human genome. Work in cancer has highlighted the extent of somatic retrotransposition in different tissues as a consequence, and potentially, a source, of disease (Evrony et al., 2012; Rodic et al., 2015; Scott et al., 2016). This has been possible with the development of both experimental methods to isolate individual cells and amplify their genome, as well as

computational methods to detect regions where these insertions have occurred. The issue of whether somatic retrotransposition is a rare or common occurrence in non-dividing brain cells remains controversial, with varying rates obtained from different methods and computational pipelines (Dubnau, 2018; Terry & Devine, 2019). This chapter analyses somatic retrotransposon insertions in hundreds of single cells isolated from human post-mortem brain tissue at a whole-genome resolution. We utilized more than 260 single cells obtained from healthy individuals from a wide age range as well as from individuals with neurological disorders. Neurons and glia from these samples were sorted and whole-genome amplified. We used *xTea* and *memTea,* two algorithms for the detection of transposable element insertions (TEIs), and adapted them for the detection of insertions in single cells. Despite some caveats in the sensitivity for detection of TEIs in single cells, our results confirm that somatic retrotransposition rates in the brain are low, with a low-end estimate of ~0.03 somatic TEIs detected per neuron and ~0.13 somatic TEIs in glia before PCR validation. Our data is the first to systematically analyze somatic retrotransposition at a whole-genome resolution in aging human brain single cells, as well as in Cockayne syndrome (CS), Xeroderma pigmentosum (XP), Ataxia telangiectasia (AT), Alzheimer's disease (AD), and amyotrophic lateral sclerosis (ALS).

Introduction

Methods for detection of transposable elements in single cells

The goal of the work in this chapter is to investigate the rates of somatic TEIs in the human brain in aging and disease. Previous work has demonstrated that somatic retrotransposition can occur during neurogenesis and in neural progenitor cells (Coufal et al., 2009; Evrony et al., 2012; Evrony et al., 2015; Muotri et al., 2005). Several

different methods for detecting and quantifying somatic insertions have been developed. Early studies investigating somatic L1 insertions in the brain used retrotransposition reporter constructs in rat neural progenitor cells and rat neural stem cells (Muotri et al., 2005) and later in human neural progenitor cells (Coufal et al., 2009). In these assays, retrotransposition of an L1 element results in the expression of a reported gene such as GFP, allowing sensitive identification of L1 transposition events. Importantly, hippocampus-derived rat primary neurons and astrocytes were tested but, unlike dividing progenitor cells, did not show L1 retrotransposition in the same assay (Muotri et al., 2005). These L1 reporter constructs have the caveat of expressing single L1s removed from their natural genomic context or even their native species and may lack the same regulatory mechanisms as native L1s. This could cause atypical elevated L1 expression patterns and false-positive insertions. Additionally, this method does not quantify somatic insertions of host native L1s, resulting in type II errors.

Other early work on retrotransposition in brains quantified L1 somatic mutations in bulk brain tissue by measuring L1 ORF2 copy number using a quantitative PCR assay. Results suggested rates of up to 80 somatic L1 insertions per genome (Coufal et al., 2009). However, this method requires a plasmid L1 spike-in for calibration (Coufal et al., 2009; Reilly, Faulkner, Dubnau, Ponomarev, & Gage, 2013), might not discriminate between reverse-transcribed L1 elements that are not inserted in the genome and retrotransposed inserted elements (Reilly et al., 2013), and may also not be precise due to a lack of resolution in discriminating between older inactive and active L1 elements (Evrony, Lee, Park, & Walsh, 2016). Because of this, this method and its estimates are currently considered unreliable and should be interpreted with caution (Terry & Devine, 2019).

Alternative methods for the detection of somatic retrotransposition events include targeted insertion profiling methods. Many varieties of this approach exist, including targeted hybridization capture arrays followed by PCR amplification in bulk tissue (Baillie et al., 2011) or amplified single cells (Upton et al., 2015) and PCR-based targeted L1 sequencing in single cells (Erwin et al., 2016; Evrony et al., 2012) or in bulk tissue (B. Zhao et al., 2019). These methods target conserved regions on the 3' and 5' end or only the 3' end of retrotransposons to amplify them specifically. A caveat of these methods is that they also amplify vast numbers of TEs present in the genome which are fixed and are known to no longer retrotranspose, requiring complex computational and statistical methods to separate these fixed, known, reference TEs, and well as other potential noise, from bona fide somatic TEIs (Erwin et al., 2016; Evrony et al., 2012). An additional caveat of these methods is the presence of PCR and single-cell amplification chimeras which contribute additional noise and are challenging to remove computationally (Evrony, Lee, Park, & Walsh, 2016).

TEIs may also be resolved through both low and high coverage whole-genome sequencing analysis (Ewing & Kazazian, 2011; Stewart et al., 2011). In cancer, somatic TEIs have been resolved by sequencing tumor and normal bulk tissue pairs (Lee et al., 2012; Rodriguez-Martin et al., 2017). In the brain, somatic TEIs have been detected with single-cell multiple displacement amplification (MDA) (Dean et al., 2002) followed by whole-genome sequencing (Evrony et al., 2015) or by analyzing deep (200x) whole-genome sequencing data (Zhu et al., 2021). Although these methods are currently more expensive than other approaches like targeted sequencing, there are several advantages. The data produced can be used for other studies and different types of variants can be analyzed beyond TEs. Additionally, both breakpoints of the insertion can be precisely detected, which is not always the case with targeted sequencing

approaches. Another advantage of using MDA amplified single genomes for TEI detection, is that amplicons are large (~23 kb) (Hou et al., 2012) and can therefore provide a large enough template for full-length PCR validation of L1 insertions which can be up to 6 kb in length. This is a crucial advantage to other approaches since this allows us to discern between true insertions and chimeric artifacts when an insertion is detected in a single cell and not present in bulk tissue.

Rates of mosaic retrotransposition in the brain

The rates of somatic retrotransposition in the brain remain a controversial topic in the field (Dubnau, 2018). Previous work from our group performed L1 profiling in single cells of the cortex and caudate nucleus and detected somatic TEIs, albeit at low rates of ~0.04-0.6 insertions per neuron (Evrony et al., 2012). Whole-genome sequencing (WGS) of MDA-amplified DNA from 12 single neurons from a single human brain confirmed that most neurons do not contain somatic TEIs, but that some clonal somatic insertions were detected and validated (Evrony et al., 2015). This work has highlighted a need for thorough validation of TEIs by full-length PCR validation as well as a careful and stringent selection of parameters when detecting variants computationally, to filter out false positives that arise from single-cell amplification chimeras and artifacts (Evrony et al., 2016; Lee et al., 2012). Other studies estimating more than 13 somatic TEIs per neuron (Upton et al., 2015) have been criticized for this lack of stringency (Evrony et al., 2016), and additional subsequent independent studies using targeted L1 sequencing have reported similarly low rates of 0.63–1.66 L1 somatic insertions per neuron in healthy individuals (B. Zhao et al., 2019) and ~0.58–1 L1 TEIs and retrotransposition independent L1 deletions per cell (Erwin et al., 2016). Despite the considerable evidence

that many studies claiming high rates of somatic insertions suffer from artifacts and methodological shortcomings (Evrony et al., 2016), studies continue to claim high rates of somatic mosaicism in healthy and pathological brains, usually without using stringent evidence or validation (Jacob-Hirsch et al., 2018). Recent review articles tend to report both higher and lower rates, often attributing differences in rates merely to differences in methodologies or samples used, rather than demanding stringent evidence (Faulkner & Billon, 2018; Terry & Devine, 2019).

Somatic retrotransposition in the aging brain and brain diseases

Previous studies have implicated dysregulation of TE expression in an array of brain diseases and disorders in both animal models and humans. Higher rates of retrotransposition events were reported when using an enhanced green fluorescent protein retrotransposition reporter construct followed by qPCR in a mouse knockout model of MeCP2, which models human Rett syndrome, a syndromic form of autism spectrum disorder, and in neuronal progenitor cells derived from Rett syndrome patients (Muotri et al., 2010). Increased retrotransposition was also reported in an ataxia telangiectasia mutated (ATM) mouse model, which models a neurodegenerative autosomal recessive DNA repair disorder, and in *ATM* knockdown human neural progenitor cells, using a retrotransposition reporter construct (Coufal et al., 2011). Increased expression of TEs has been detected in ALS and frontotemporal lobar degeneration animal models of TDP-43 misexpression (Krug et al., 2017; W. Li, Jin, Prazak, Hammell, & Dubnau, 2012). In human cells, TDP-43 binds TEs which were also found to be overexpressed in a subset of ~20% of ALS cortical samples (Tam et al., 2019). In another example, an increased L1 copy number, determined by RT-PCR, was

reported in human schizophrenia brains. However, WGS bulk analysis of those very same samples did not confirm this overall increase TE insertion in schizophrenia patients compared to unaffected controls, finding instead equivalent overall TE insertion rates in schizophrenia and control, with a nominal enrichment in insertions in genes encoding synaptic proteins in schizophrenia compared to control (Bundo et al., 2014), a finding that has not yet been reproduced (Zhu et al., 2021).

Elevated rates of TE insertion have also been reported with aging. In senescent human diploid fibroblasts, the chromatin of Alu, SVA, and L1 TEs was reported to be more open, and these elements were increased in expression and copy number as measured by DNA and RNA qPCR (De Cecco et al., 2013). In *Drosophila,* GFP reporter constructs have also suggested increased retrotransposition in the brain with aging (W. Li et al., 2013). In AD, Tau burden was associated with overexpression of certain TEs in human brains and *Tau* mutations in a *Drosophila* model led to de-repression of certain TEs which were further overexpressed with aging (Guo et al., 2018). Although these studies suggest that certain diseases and aging cells may display increased expression of TEs and potentially somatic retrotransposition in the brain, these experiments have mainly focused on cell lines and animal models or are based on reporter construct assays or PCR based assays for TE copy number, which as discussed previously, have major flaws.

The most direct and reliable evidence for rates of retrotransposition in the human brain in aging and disease come from genomic sequencing of brain tissues, either single cells or bulk DNA at high coverage. In a study using deep whole-genome sequencing of bulk brain-derived DNA (Zhu et al., 2021), two somatic intronic L1 insertions in genes that are in loci associated with schizophrenia and neurodevelopmental disorders were detected and validated in one individual with schizophrenia, suggesting rates far lower

than that proposed from *in vitro* experiments by Coufal *et al*. (2009) (Coufal et al., 2009). An analysis of targeted human-specific L1 (L1HS) sequencing in tissues from 5 individuals with Rett syndrome and healthy controls did not find an increase in TEIs in Rett syndrome but found a depletion of somatic L1HS insertions in exons in Rett patients compared to controls, suggesting that these exonic insertions may affect the viability of neurons in these patients (B. Zhao et al., 2019). These types of exonic insertions are presumably under negative selection in both case and controls, but these researchers hypothesized that since cases already carry an *MECP2* mutation, their neurons are less tolerant to additional damaging mutations. To date, there is no experimentally validated direct evidence from genome sequencing showing an increase of somatic retrotransposition in the brain in aging or disease. Here, we used single-neuron, whole-genome sequencing from MDA-amplified DNA from hundreds of single brain cells from healthy, aged, and affected individuals to explore whether and of these conditions detectably affect somatic TE insertion rates.

Results

Pipeline Sensitivity

We used two methods for the detection of TEIs in MDA whole-genome amplified single cells. The first method, *memTea*, is an adaptation of a previously published algorithm, *scTea* (Evrony et al., 2015; Lee et al., 2012), modified for the analysis of BWA-Mem mapped data (Heng Li, 2013). The second method, *xTea* (Chu et al., 2021), which at the time of this writing remains under active development (https://github.com/parklab/xTea), was developed for bulk long and short-read sequencing data and adapted for this single-cell analysis. This second method performs

additional filtering steps beyond those used in *memTea*, such as mapping supporting insertion reads to reference TE sequences.

We analyzed 159 single neurons from the prefrontal cortex and dentate gyrus from both healthy individuals and individuals with CS and XP (Lodato et al., 2018) using *memTea*. Most of these datasets had previously been analyzed for SNVs (Lodato et al, 2018), but only a small number of them had been analyzed for TE insertions (Evrony et al, 2015). Single neuronal nuclei were sorted using fluorescence-activated nuclei sorting (FANS), gating for the largest-sized, NeuN positive nuclei, which likely selects strongly for nuclei of large, pyramidal neurons. Nuclei were sorted into individual wells of microtiter plates and lysed on ice with NaOH as previously described (Evrony et al, 2012; Lodato et al, 2018). Single-cell DNA samples were then MDA amplified (Dean et al., 2002), and PCR-based and low coverage WGS-based quality control analysis was used to select a minority of the most evenly-amplified genomes from each individual to be subjected to WGS, as part of a previous analysis for single-nucleotide variants in these genomes (Lodato et al., 2018).

We selected high confidence TEIs based on the parameters used previously in *scTea* (Evrony et al., 2015) and adjusted these for more stringent parameters (see Methods) since in this data mapped with BWA-Mem the previous parameters resulted in more false-positive artifact calls. Importantly, although the selected parameters were more stringent, we still detected all previously validated somatic TEIs from previous analyses (Evrony et al., 2012; Evrony et al., 2015), suggesting no apparent sacrifice in sensitivity.

We analyzed the number of detected high confidence insertions that overlapped with previously published germline known non-reference (KNR) Alu, L1, and SVA insertions (Gardner et al., 2017; Stewart et al., 2011; Wang et al., 2006; Xing et al.,

2009) in single neurons and ~30-60x bulk sequencing data obtained from the same individual to determine the sensitivity of the algorithm and parameters for detecting TEIs in single neurons. KNR insertions were chosen for the sensitivity analysis since they are polymorphic in the population and generally heterozygous, suggesting that they would have signal characteristics similar to somatic TEIs, also expected to be heterozygous. As expected, bulk samples showed a similar and higher number of KNR insertions than single cells. If cells had even and full coverage, we would expect to be able to detect all of the KNR TEIs detected in bulk in single cells as well. However, MDA amplification of single cells is uneven and can cause locus dropout, leading to a lower sensitivity for detection of TEIs in single cells (Evrony et al., 2015). In cells isolated from postmortem brains of children and adolescents, the observed detection sensitivity for KNR insertions was similar to the previously reported sensitivity from a single adolescent individual (Evrony et al., 2015). In contrast, aged, CS, and XP neurons had a significantly lower number of KNR TEIs detected (CS vs. adolescents, p=0.0002; XP vs. adult, p= 0.0002; unpaired two-sample Wilcoxon test with Benjamini-Hochberg FDR correction) (Figure 3.1A).

One likely explanation for the lower sensitivity of KNR detection in aged and diseased neurons is that they had uneven or lower genome coverage. We used the median absolute pairwise deviation (MAPD) (Cai et al., 2014) as a metric to estimate genome coverage variability. The MAPD score measures the median of the difference of the $\log_2$ copy number of adjacent bins in the genome. A higher MAPD score reflects more noise and more uneven amplification in single cells. We observed that, despite WGS only being performed on the most evenly amplified cells from each individual, aged and diseased neurons still showed higher MAPD scores (Figure 3.1B). The sensitivity for detection of KNR TEIs in single neurons is strongly and inversely correlated (Pearson

correlation coefficient, r = -0.94, p-value < 2.2e-16 in the prefrontal cortex and r = -0.96, p-value 9.5e-11 in the dentate gyrus) with the MAPD score (Figure 3.1C and Figure 3.1D). Cells with noisier cell coverage amplification show a lower sensitivity. Importantly, the sample “1465” which was obtained from an adolescent individual (17.5 years old) and was analyzed in-depth in previous work from our group (Evrony et al., 2015) shows the highest sensitivity and is similar to what was observed in this previous work. The basis for the persistent difference in amplification coverage between young neurons and aged and diseased neurons is not known, but plausible explanations would include various forms of genomic damage, such as bulky adducts of covalent DNA modifications being more common in age and disease states, which would block DNA synthesis by the MDA Phi29 polymerase (Baumer, Fisch, Wedler, Reinecke, & Korfhage, 2018). Nonetheless, the consistent differences in amplification require sensitivity corrections for the TEI detection rate.

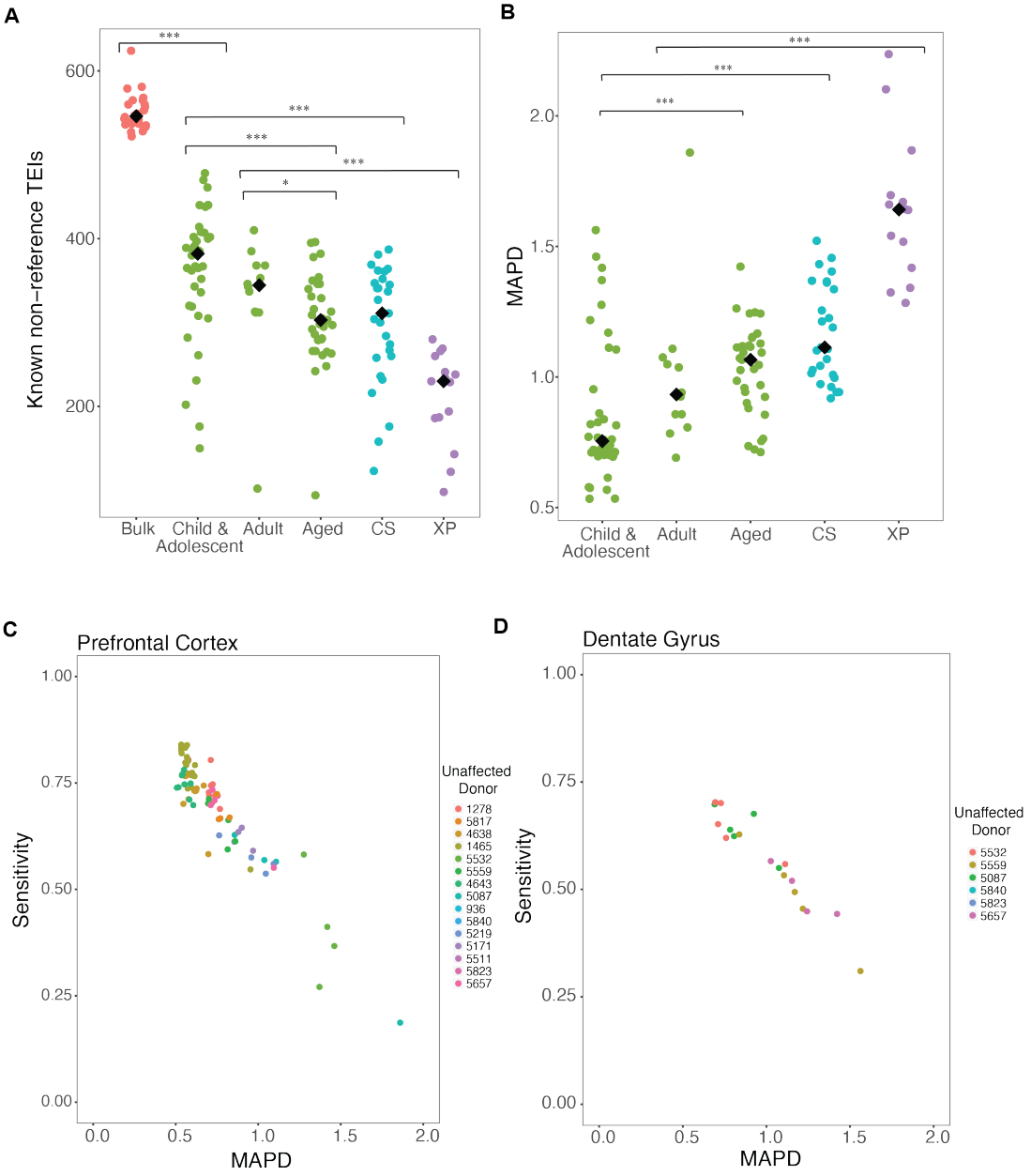

Figure 3.1 Known non-reference sensitivity in single neurons with memTea.

A) Known non-reference (KNR) TEIs for bulk and single cells of individuals of multiple ages and diagnosis. Sensitivity differs between samples and must be accounted for. Sensitivity of TE detection in aged and diseased neurons is lower than for younger neurons or neurons of the same age range in healthy individuals respectively. CS (Cockayne syndrome); XP (Xeroderma pigmentosum). Each point is a single cell or bulk sample. B) Median of the Absolute values of all Pairwise differences score for each of the single cells (Lodato et al., 2018). A higher MAPD score is generally associated with poor cell quality and lower TEI detection sensitivity. Aged and diseased neurons tend have higher MAPD scores in this dataset. Diamonds represent the group median. **, and *** denote P≤ 0.01, 0.001 respectively using an unpaired two samples Wilcoxon test. C) Sensitivity for detection of KNR insertions in single cells by MAPD score in single neurons obtained from the prefrontal cortex and from the D) dentate gyrus.

Next, we ran *xTea* to detect KNR and somatic L1 insertions with a subset of these same aging, CS and XP neurons along with 24 unpublished neuronal genomes from 4 individuals with ALS, 25 neurons from 4 individuals with AD, and 10 neurons from 2 individuals with AT for a total of 207 single cells (Supplementary Table 3.1). With this method, we also observed that neurons from aged and diseased individuals had a lower sensitivity than neurons from healthy individuals, with neurons from individuals with AD having the lowest sensitivity (Figure 3.2). Here, we also detected all somatic insertions which were detected previously in individual “1465” (Evrony et al., 2015). However, this method has a very high false-positive rate for single neurons, with many calls in single neurons that are not found in bulk tissue and which were determined to be likely artifacts based upon manual inspection in the Integrative Genomics Viewer (IGV) (Thorvaldsdottir, Robinson, & Mesirov, 2013). Because of this, we used stringent parameters requiring ≥ 6 supporting clipped reads and ≥ 6 supporting discordant reads for the detection of somatic TEIs at the expense of having a reduced sensitivity (see Methods) (Table 3.1).

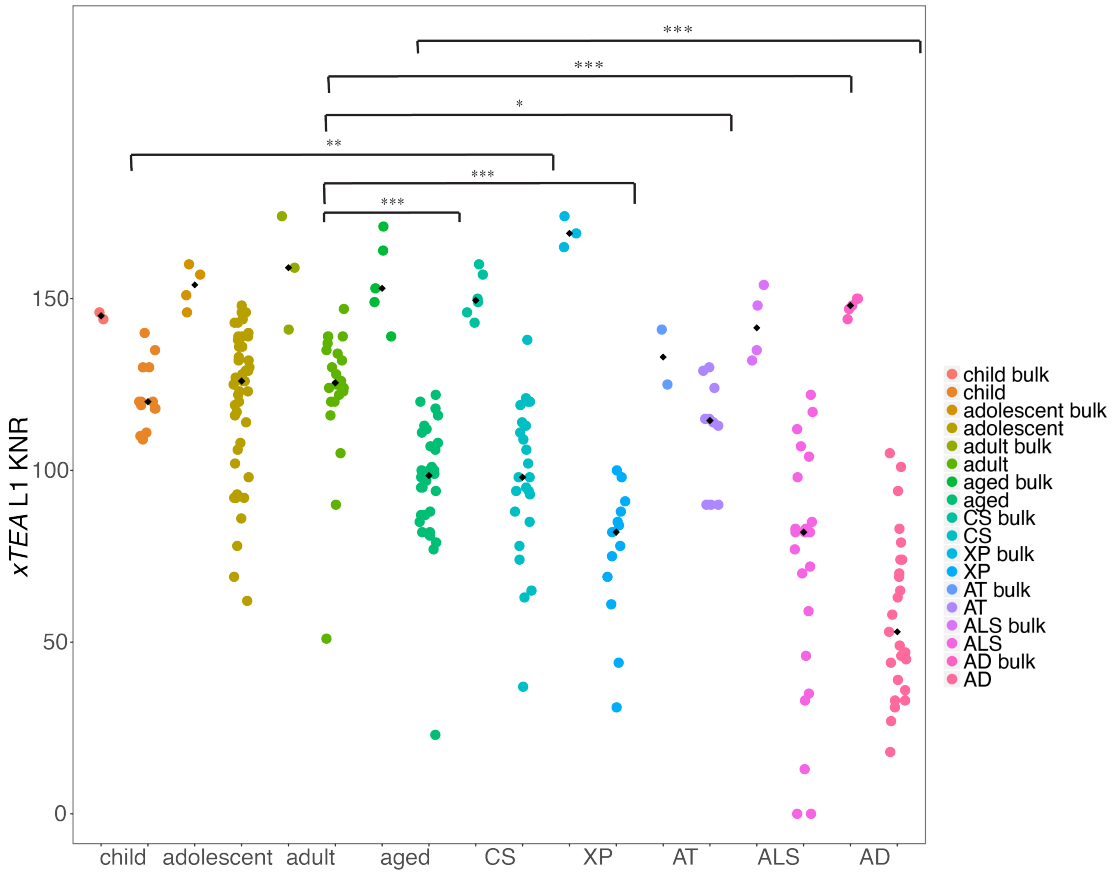

Figure 3.2 Known non-reference L1 insertions detected with xTea in single neurons and bulk.

Known non-reference (KNR) L1 TEIs for bulk and single cells of individuals of multiple ages and diagnosis. The sensitivity of TE detection in aged and diseased neurons is lower than for younger neurons or neurons of the same age range in healthy individuals respectively. CS (Cockayne syndrome); XP (xeroderma pigmentosum); AT (ataxia telangiectasia); ALS (amyotrophic lateral sclerosis); AD (Alzheimer's disease). Each point is a single cell or bulk sample. Asterisks *, **, and *** denote P≤ 0.05, 0.005, 0.0005 respectively using an unpaired two samples Wilcoxon test and were significant after multiple testing correction with the Benjamini-Hochberg method.

Table 3.1 Number of TEI calls in single neurons and bulk tissue with xTea.

The number of all calls identified. Calls were categorized by *xTea* as "high confidence", or filtered "high confidence" calls in the adolescent individual "1465". These calls were also overlapped with known non-reference (KNR) insertions to determine their overlap. Single neurons have many more false-positive calls than bulk tissue, but reduced sensitivity for KNR insertions than bulk. Neurons from individual 1465 have a low MAPD score and low amplification noise. These selected neurons all had validated previously published somatic TEIs (Evrony et al., 2012; Evrony et al., 2015) that were also detected with *xTea.*

Tissue	All L1 candidates	L1 candidates KNR overlap	L1 HC candidates	L1 HC KNR overlap	L1 HC candidates ≥ 6 supporting reads	L1 HC candidates ≥ 6 supporting reads KNR overlap
1465 cortex 18	2,017	186	697	143	316	131
1465 cortex 2	2,671	157	1,040	126	294	115
1465 cortex 51	2,808	148	1,158	119	287	104
1465 cortex 6	2,343	156	957	126	270	103
1465 heart bulk DNA	712	191	238	151	167	135

Somatic retrotransposition in the aging and diseased brain

We developed a pipeline to detect somatic TEIs from *mem*Tea calls. We observed that BWA-Mem aligned single neuron data contained more false positives than the BWA-Sampe aligned data analyzed previously (Evrony et al., 2015), due to the extensive clipped read mapping performed by this algorithm. This resulted in 13.9

somatic candidates per neuron mapped with BWA-Mem and analyzed with *memTea*, vs. 1.69 calls per neuron mapped with BWA-Sampe and analyzed with *scTea*. We increased the stringency of the parameters required to detect high confidence somatic TEIs. This included requiring a poly(A) tail, a target site duplication (TSD) of 5-30 base pairs in length, and < 2 clipped or discordant reads within 5 base pairs of the predicted insertion breakpoint in bulk tissue. Since chimeras and poly(A) tracts in the region can cause false positives and artifacts (Evrony et al., 2016), we manually inspected the insertions on the IGV browser (Thorvaldsdottir et al., 2013) where we confirmed the presence of supporting clipped and discordant reads in the expected orientation, the presence of a TSD and a poly(A) tail, and the absence of supporting reads for a heterozygous insertion in the bulk tissue.

We detected four novel, somatic TEI candidates in the 159 single neurons from healthy individuals and individuals with Cockayne syndrome and xeroderma pigmentosum (Supplementary Table 3.1). We additionally detected the four previously experimentally validated somatic candidates in individual “1465”, but excluded cells that were included in this previous analysis from the aging analysis and rate estimates since they were obtained and sequenced in a biased manner to confirm previously detected insertions from a targeted sequencing analysis (Evrony et al., 2012; Evrony et al., 2015), but we included neurons from this individual which were selected randomly in a subsequent study (Lodato et al., 2018). Somatic TEI counts were adjusted by dividing the count by the observed KNR sensitivity (Figure 3.3) to account for the lower sensitivity of TEI detection in neurons from aged and diseased individuals. With this low number of somatic candidates, we were unable to observe a difference in rates of TEIs with aging.

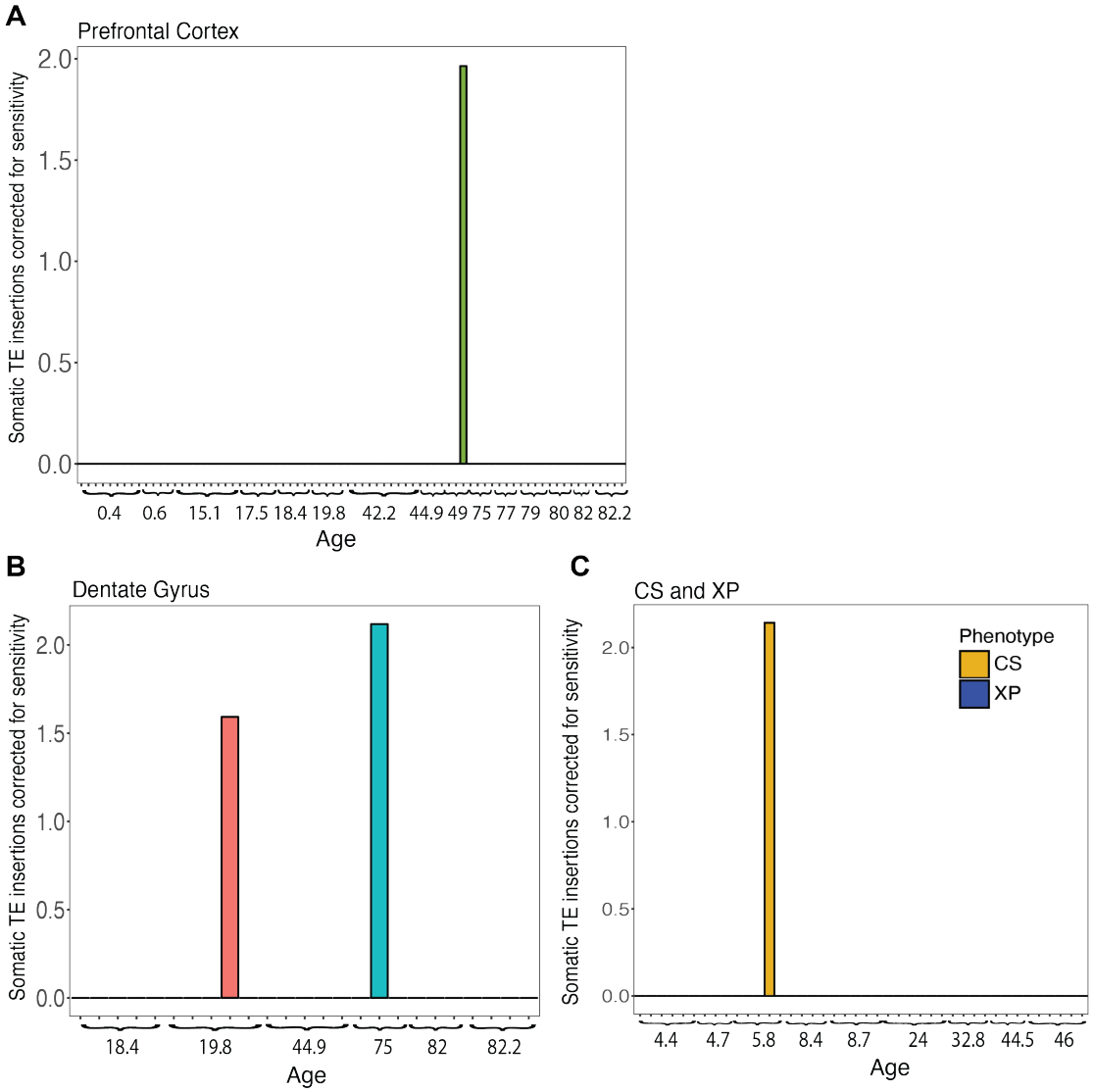

Figure 3.3 Pre-validation somatic TEIs in aging, Cockayne syndrome, and xeroderma pigmentosum detected with *mem*Tea.

A) Pre-validation somatic TE insertion counts in single cells of different ages in the prefrontal cortex after visual inspection filtration. Each bar represents the sensitivity corrected count of candidate insertions in a single cell. Non-integral values reflect this correction which was obtained by dividing the count by the sensitivity (percentage of germline KNR in bulk also detected in single cells). Many cells have no detectable somatic insertions as shown by the absence of a bar. B) Pre-validation somatic TE insertions counts in single cells of different ages in the dentate gyrus after visual inspection filtration. D) Pre-validation somatic TE insertions counts in single cells of individuals diagnosed with Cockayne syndrome (CS) and xeroderma pigmentosum (XP).

Next, we used *xTEA* and a custom somatic mutation calling pipeline to detect somatic L1 TEIs in these neurons along with 24 neurons from four individuals with ALS, 25 premotor cortex or hippocampus CA1 neurons from four individuals with AD, and ten neurons from two individuals with AT (Figure 3.4). Because these cells had more candidate TEI calls, which visual inspection suggested were likely artifacts and false positives, candidate TEIs were scored to prioritize those with more discordant reads in the expected orientation for visual inspection with IGV (see Methods). Despite the high number of candidates, none of the high-scoring ALS or AD candidates looked like real TEIs with two breakpoints, a poly-A tail, and a TSD upon inspection on IGV. These data confirm previous analyses (Evrony et al, 2012, 2015) that TEIs are quite uncommon in single neurons of normal human postmortem brains but extends these findings to show that TEI rates are similarly low in neurons from aged brains, and in brains from individuals who died with ALS, AD, CS, and XP.

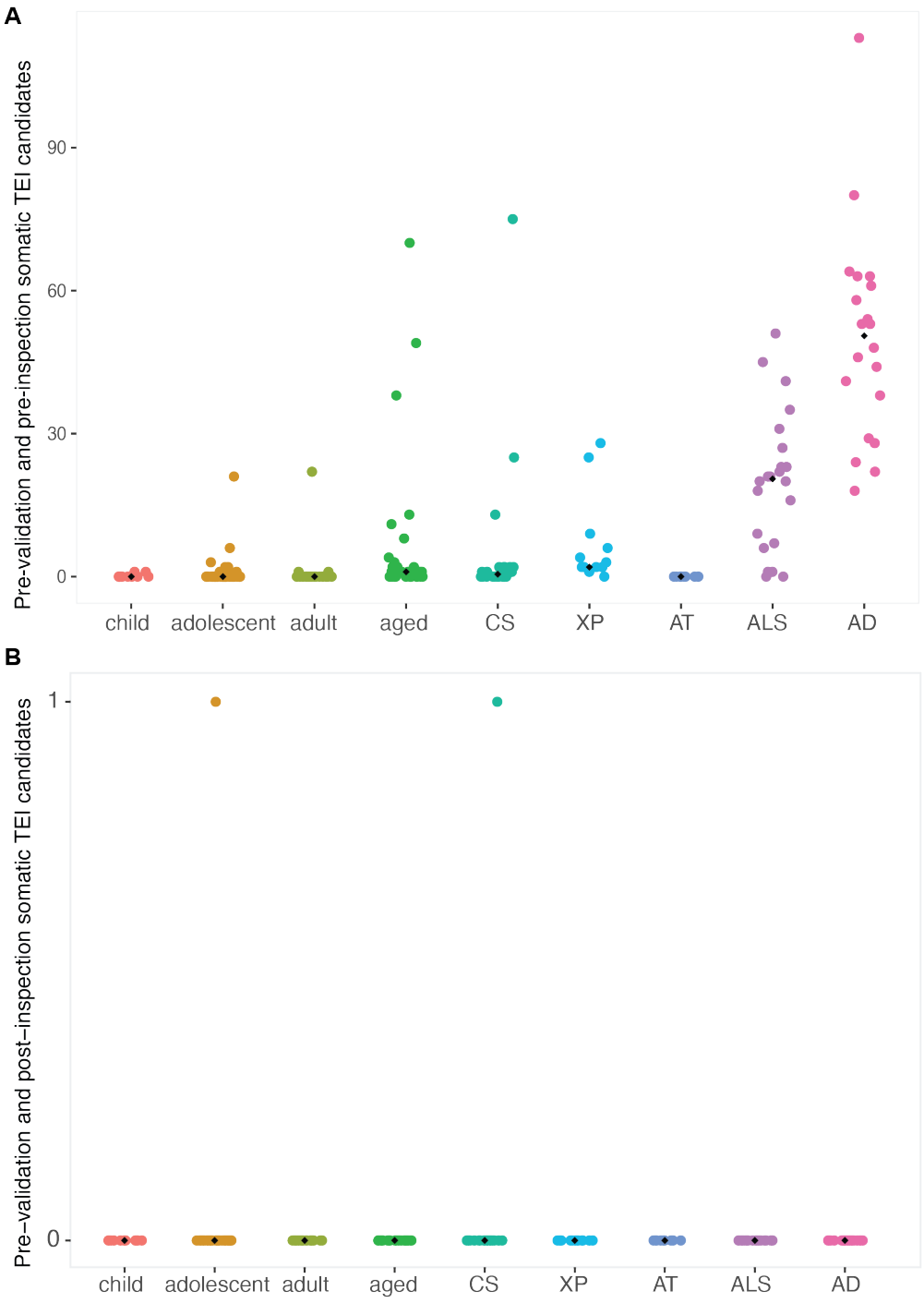

Figure 3.4 Number of somatic candidates before validation detected in single neurons with xTEA in aging and disease.

A) Only candidates classified with the most stringent criteria requiring 6 clipped reads and 6 discordant reads, a poly(A) tail, and a target site duplication are included. B) Most of these calls were false positives after inspection on IGV. *xTea* identified 1 somatic candidate that had not been detected by *memTea* in CS but filtered out 3 *memTea* candidates. Each dot represents a single neuron. Cockayne syndrome (CS), Xeroderma pigmentosum (XP), Ataxia telangiectasia (AT), Amyotrophic lateral sclerosis (ALS), Alzheimer's disease (AD).

Somatic retrotransposition in non-neuronal cells

We analyzed 30 single nuclei from glia obtained from the postmortem prefrontal cortex in 3 individuals, and 20 diploid and tetraploid heart cells from the left ventricle from 3 individuals including young, adult, and aged individuals, and compared somatic TEI rates in these cells with neurons (Supplementary Table 3.1). Glial cells were selected using FANS with anti-GFAP and anti-SOX10 antibodies in the NeuN negative nuclei population. The GFAP positive NeuN negative nuclei are represented by various glial cell types and are referred to here as “glia”, whereas SOX10 positive NeuN negative nuclei select a pure population of oligodendrocyte cells. Human heart cells from the left ventricle were sorted with FANS using the cardiomyocyte-specific antibodies cardiac troponin T and PCM1). Cardiomyocytes can be diploid or polyploid (Derks & Bergmann, 2020), therefore both diploid (2n) and polyploid (4n) nuclei were selected during FANS. This analysis detected one glial-specific clonal insertion in two nuclei that was not detected in the analyzed neurons or oligodendrocytes of the same individual. Although heart cells showed a large number of candidate L1 somatic insertions before manual inspection, the inspection showed that they did not have the expected patterns of TEIs and were either artifacts or other types of structural variants.

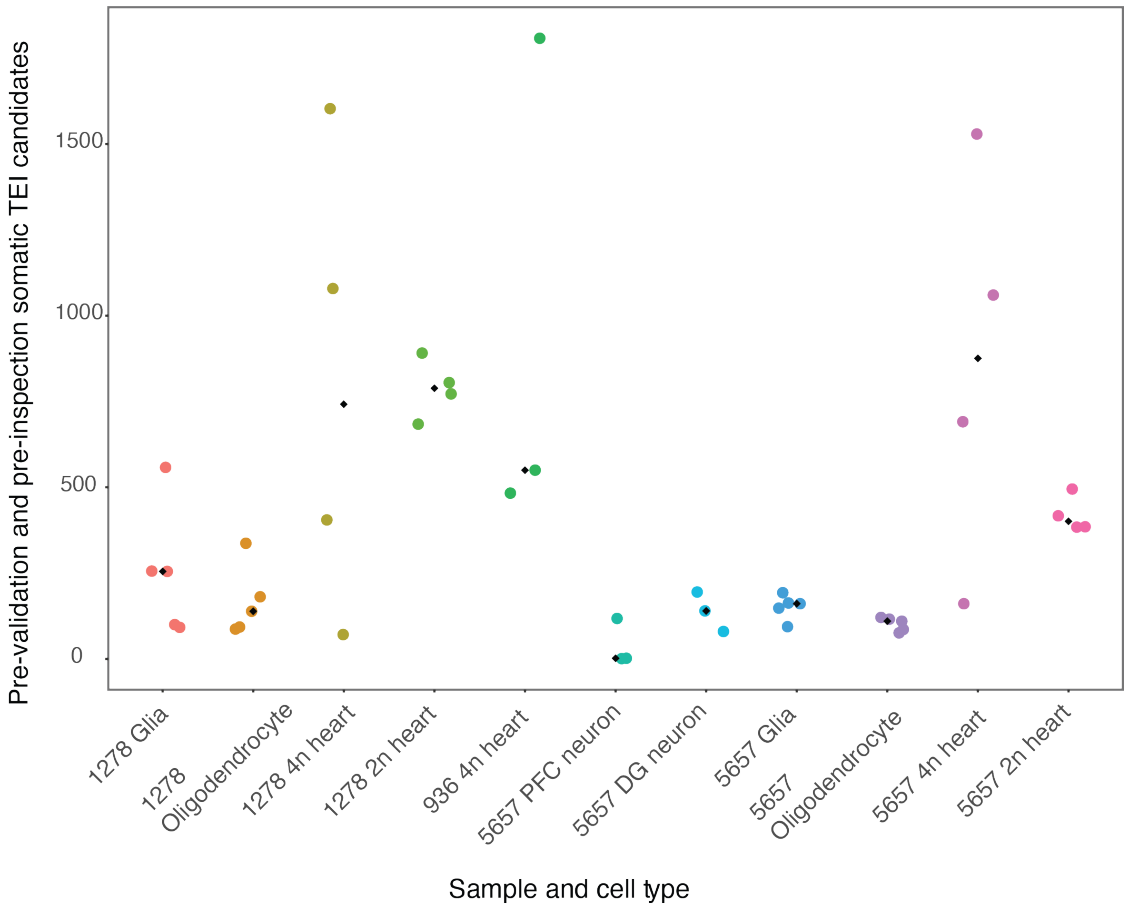

Figure 3.5 Number of somatic candidates before manual inspection and validation detected in glia and heart cells.

Somatic candidates from called classified as "filtered high confidence" by *xTea* are shown. As shown here, less stringent calling criteria, results in a high number of somatic candidates. Most of these calls are false positives after inspection on IGV. Each dot represents a single cell. Individual 1278 is 0.4 years old, 936 is 49.2 years old, and 5657 is 82.2 years old. PFC (prefrontal cortex); DG (dentate gyrus). 4n cells were sorted as tetraploid cells and 2n as diploid cells.

Overall, we detected six new candidate somatic L1 insertions using both *memTea* and *xTea* combined. We detected three *memTea* and one *xTea* specific TEIs, suggesting that *xTea*'s additional filters may be too stringent for TE detection in single cells. However, all four previously validated insertions were called with both algorithms. Keeping in mind that strict parameters were used which excludes clonal somatic insertions present in bulk tissue and reduces the sensitivity of detection of TEIs, glial cells had nominally the highest rate of somatic retrotransposition with 0.13 insertions per

single cell compared to 0.013 per cell in neurons from the prefrontal cortex (p=0.056, one-sided two-sample chi-squared test for equality of proportions with continuity correction, X-squared = 2.53, df = 1) (Table 3.2).

Table 3.2 Pre-validation rates of somatic retrotransposition after visual inspection of calls on IGV.

Results shown are for somatic TEIs detected with both *memTea* and *xTea*. PFC (prefrontal cortex); DG (dentate gyrus); Cockayne syndrome (CS); xeroderma pigmentosum (XP). The rates shown are a lower bound since strict calling parameters were used. Cells that were sequenced in a biased manner due to having previous knowledge of clonal TEIs in these cells were excluded from these estimates. The p-values were obtained using a one-sided two-sample chi-squared test for equality of proportions with continuity correction by comparing to somatic TEI rates in PFC neurons.

Cell	Cells with insertions / total cells analyzed	Rate
Pfc neurons	1/76	0.013
DG neurons	2/25 (p=0.16)	0.08
CS neurons	2/27 (p=0.17)	0.07
XP Neurons	0/15 (p=0.5)	0
PFC glia	2/15 (p=0.056)	0.13
PFC oligodendrocytes	0/15 (p=0.5)	0

Retrotransposon mediated deletions in single cells

Previous work has identified retrotransposition-independent somatic L1 associated deletions in the brain (Erwin et al., 2016) and cancer (Rodriguez-Martin et al., 2017). We hypothesized that some of the detected candidates which did not show canonical TEI patterns upon visual inspection could be L1 mediated structural variants such as these deletions. We overlapped the set of *xTea* somatic calls in the aging single neurons, CS, and XP single cells with deletions detected in these cells. These deletions were obtained using *PhaseDel* (J. Kim *et al.* 2021 under preparation), a method

developed in our group that requires haplotype phasing of the supporting reads for quality control to minimize chimeric artifact false positives. Deletions detected with DELLY2 (Rausch et al., 2012) in this method ranged from 13 bp to 158,812,543 bp, and the mean deletion size was 2,812 bp long. All candidates that overlapped both phased deletion breakpoints, as well as *xTea* breakpoints, were manually inspected with IGV. We detected a total of 5 candidates; 4 were present in the same aged healthy individual in 2 different neurons and 1 was present in a single neuron from a healthy adolescent individual. A detailed sequence and visual inspection of these events suggested that these candidates were deletions mediated by nearby reference L1 retrotransposon sequences. This suggests that although a small number of *xTea* calls may be attributed to deletions, most candidates observed may either be other types of structural variants that were not detected here or chimeras and amplification and sequencing artifacts.

Discussion

We tested whether somatic retrotransposition increases in the single cells of the brain and heart with aging and disease by analyzing 159 single nuclei for somatic TEIs using two novel methods for BWA-MEM data. We detected somatic L1 TEIs in six single nuclei, including three neurons from healthy individuals, one neuron from an individual with CS, and one clonal mutation in two glial cells. Additionally, we detected 5 candidate somatic L1-associated deletions in 5 single neurons. With the constraints of our current sample size and our sensitivity for detection of TEIs in MDA single-cell data, we did not find evidence for an increase in somatic retrotransposition in aging neurons, glia, or heart cells, nor in the single neurons analyzed from individuals with CS, AP, AT, ALS, or AD. Instead, single cells from aged or diseased individuals showed very low rates of

somatic TEI indistinguishable from cells from younger individuals. However, aged and diseased cells generally showed more artifacts and uneven genome amplification, undercutting the sensitivity for detection of TEIs in these cells. These results confirm previous findings suggesting low somatic retrotransposition in the human brain (Erwin et al., 2016; Evrony et al., 2012; Evrony et al., 2015), and support the hypothesis that most neuronal retrotransposition occurs during development.

Compared to previous studies which have reported rates of <0.6 somatic TEIs per neuron and a post-validation rate of <0.04 somatic TEIs per neuron (Evrony et al., 2012), the pre-validation retrotransposition rates observed here for single neurons of ~0.03 somatic TEIs per neuron are within that estimate. Including retrotransposition independent somatic L1-associated variants, we observed a pre-validation rate of 0.08 variants per single neuron. A previous study in glia and neurons estimated ~0.58-1 event per cell (Erwin et al., 2016). However, the upper limits of these estimates are based on their 3' validations, which are insufficient to exclude MDA artifacts (Evrony et al., 2016). Without considering an adjustment for their validation rates, here they observed only two L1 TEIs which were validated by 3' PCR and flanking PCR with Sanger sequencing (2/89)--which is a rate of 0.022 and is similar to the number of L1 somatic TEIs detected in this study--and 2 loci with loss of heterozygosity and a confirmed deletion using a flanking PCR assay and Sanger sequencing (2/89) with a rate of 0.022 retrotransposition independent L1-mediated deletions as well.

To obtain a precise TEI rate, we would need to consider the lower sensitivity for TEI detection due to sequencing bulk tissue and single cells with short reads (Chu et al., 2021; X. Zhao et al., 2021; Zhou et al., 2020), the lower sensitivity of our somatic pipeline which uses stringent parameters to reduce the number of false negatives and increase the precision, as well as the lower sensitivity of detection in single cells, which

for MDA single-cell aging and diseased cells is even lower than in cells from healthy and young individuals. In the previous chapter, we estimated that the sensitivity for detection of L1 TEIs in Illumina short-read sequencing data with *xTea* in ~40x coverage is 55%. With *memTea*, we detected an average of 551 L1, Alu, and SVA KNR insertions in a bulk tissue using stringent calling parameters, suggesting that our sensitivity with the stringent parameters selected for *memTea* ~49% based on previous analyses (Evrony et al., 2015). On average in cells from adolescent and adult individuals using *xTea*, we detected 121.6 L1 KNR or 80% the number of KNR L1 TEIs detected in bulk (151.2 L1 high confidence KNR TEIs on average in bulk), whereas in aged neurons, we detected only 63.7% of these. In neurons from individuals with Alzheimer's disease, which have the lowest sensitivity observed, we only detected an average of 57.4 KNR TEIs, or 38% the number of KNR TEIs observed in bulk tissue. By requiring a high number of supporting reads for detecting and filtering somatic TEI candidates, we reduced the sensitivity to 75% in adolescent neurons (Table 3.1). Therefore, while the somatic TEI rate observed in healthy young neurons is similar to the rates detected in previous studies, they are still a lower bound estimate. However, even when adjusting for the stringent parameters used and the lower sensitivity in short-read data and MDA data, we would still expect <1 somatic TEI in healthy neurons based on the observed rates. The rates observed in neurons from aged individuals and individuals with neurodegenerative and DNA damage diseases are even more likely to be an underestimate of the true underlying rate, so although we detected few insertions in these conditions, if there is a small underlying increase in somatic TEI rates in these conditions it would be difficult to detect with these current methods and sample sizes. These results highlight the importance of either sequencing many cells at once through methods like targeted sequencing, improving single-cell amplification methods for achieving greater coverage evenness for better detection of structural variants including TEIs, using long-read

sequencing, or ideally a combination of these three methods. Some new methods that might help resolve these issues are considered in the Overall Discussion (Chapter 4).

Additionally, a limitation to this analysis was the higher false-positive rates observed with both *memTea* and *xTea* in BWA-MEM aligned single cell data compared to in bulk tissue. We observed more extensive read clipping in BWA-MEM aligned data than in data aligned with its predecessor BWA-SAMPE. A large number of chimeras present in MDA single-cell data (Evrony et al., 2015; Evrony et al., 2016; Jiao et al., 2011; Lasken & Stockwell, 2007) likely led to a higher number of artifact candidate TEIs detected by these tools due to their clipped read patterns. We increased the stringency for TEI detection to reduce the number of artifact calls, limiting our sensitivity as well as our ability to detect clonal TEIs which have supporting clipped and discordant reads in the bulk tissue at a low allele frequency. Since the main goal of this analysis was to determine whether there is an increase in somatic retrotransposition after brain development in aging and disease, we prioritized obtaining high confidence TEI candidates. Accurate assessment of clonal TEI rates in these cells would require further modification of these computational tools to further filter false positives while maintaining high sensitivity. Machine learning techniques that use relevant TEI features such as the number of supporting reads, read alignment scores and the distance between paired reads as has been implemented before for targeted sequencing methods (Erwin et al., 2016) could improve these methods.

Methods

Datasets and samples

Post-mortem human tissues from neurotypical individuals of different ages, individuals with CS, and XP (Evrony et al., 2015; Lodato et al., 2018) and AT were obtained from the NIH NeuroBioBank. ALS tissues were obtained from the Massachusetts Alzheimer's Disease Research Center. Post-mortem frozen human tissues were obtained from the Massachusetts Alzheimer's Disease Research Center, with AD cases selected based on a clinical history of dementia consistent with AD, AD neuropathologic change (Braak stage V-VI), and no significant other neurodegenerative pathology.

Single-cell isolation, amplification, and whole-genome sequencing

Single nuclei were isolated, and their DNA was amplified as described previously (Evrony et al., 2012; Evrony et al., 2015; Lodato et al., 2018). Samples were stored at -80°C, sections from the region of interest were obtained with a scalpel in a cryostat and then homogenized in a lysis buffer in a chilled Dounce homogenizer (0.32M Sucrose, 5mM CaCl2, 3mM Mg(Acetate)2, 0.1mM EDTA, 10mM Tris-HCl pH 8, 1mM DTT, 0.1% Triton X-100). The lysate was then placed on a sucrose cushion buffer (1.8M Sucrose, 3mM Mg(Acetate)2, 10mM Tris-HCl pH 8, 1mM DTT) and for 1-2 hours at 30,000G at 4°C. The supernatant was removed, and nuclear pellets were resuspended in a 3mM MgCl2 solution and filtered with a 40μM cell strainer. Neuronal nuclei were stained with anti-NeuN antibodies, oligodendrocytes were stained with anti-SOX10 antibodies, glia with anti-GFAP antibodies, and heart cells with anti-Cardiac Troponin T and anti-PCM1 antibodies. Glial cells were selected from a NeuN negative population and selected for

SOX10 or GFAP. Heart cells were selected based on their ploidy status for 2n or 4n cells. Nuclei were sorted with flow cytometry into 96 or 384 well plates into lysis buffer (2.8 µl 200 mM KOH, 5 mM EDTA, 40 mM DTT). Nuclei were lysed for 15-30 minutes and then 1.4 µl of neutralization buffer was added (400 mM HCl, 600 mM Tris-HCl, pH 7.5) MDA was performed for previously published normal aging, CS, XP, and AT samples in a 20 µl reaction with 0.4 µl repliPHI polymerase (40U) (Epicentre) for 16 hours at 30°C followed by inactivation at 65°C for 3 minutes. More recent AD, heart, glia, and ALS samples performed MDA for 2 hours at 30°C followed by inactivation at 65°C for 3 minutes in 20 µl reactions with 0.105 µl 1M DTT, 2.675 µl nuclease-free water, 12.18 µl Repli-G Reaction Buffer (Qiagen) and 0.84 ul Repli-G DNA polymerase (Qiagen).

Single-cell DNA was quantified using a Quant-iT™ dsDNA Assay Kit and a 4 loci multiplex PCR was performed for quality control. Previously published cells from individual 1465 were prepared using a NEXTflex DNA sequencing kit and paired-end sequenced with Illumina HiSeq 2000 (100 or 101bp x 2) (Evrony et al., 2012; Evrony et al., 2015). Previously published nuclei from individuals 4638 and 4643 were prepared with the Illumina TruSeq Nano LT sample preparation kit and paired-end sequenced on a HiSeq X10 instrument (150bp x 2) (Lodato et al., 2015). More recently published and non-published neuronal and glial nuclei were prepared with an Illumina Tru-Seq Kit (150bp x 2) and sequenced on an Illumina HiSeq X10 instrument (Lodato et al., 2018). Heart cells library preparations were performed with the Illumina Tru-Seq Nano LT sample preparation kit and paired-end sequenced (150bp×2) on a HiSeq X10 instrument.

Read mapping

Reads were aligned to the GRCh37 human reference genome with decoy using BWA-MEM (H. Li & Durbin, 2009). Picard tools MarkDuplicates was used to mark duplicates and indel realignment and base quality score recalibration was performed using the Genome Analysis Toolkit (McKenna et al., 2010).

TEI identification with *memTea*

Single-cell WGS data were systematically analyzed for TEIs using *memTea.* This algorithm is a revised version of *scTea* (Evrony et al., 2015; Lee et al., 2012) for data aligned with BWA-MEM. The *memTea* algorithm detects TEIs using two main types of supporting sequencing reads: 1) discordant reads or repeat-anchored mate reads which map uniquely to the reference genome but their paired mate read maps to a curated library of consensus TE sequences, and 2) clipped reads which are reads that partially map uniquely to the reference genome, but part of the read is masked or clipped by BWA-MEM (the part mapping to TE sequences). These two types of reads are expected at a TEI breakpoint site. Clipped reads are piled up at the breakpoint site, so the insertion breakpoints can be determined with a single nucleotide resolution. Additional features which support a target-primed reverse transcription mediated TEI include a poly(A) tail, which is supported by clipped reads with a poly(A) tract, and a TSD.

Since the BWA-MEM aligner performs extensive read clipping, which resulted in a higher number of false-positive calls than previous BWA versions, strict calling parameters were used, and additional filtering methods were implemented. Since *memTea* false positives occurred in genomic regions with poly(A) and poly(T) tracts, we used a custom algorithm to annotate and exclude candidates in these regions. A

sufficient number of supporting reads is required to distinguish chimera artifacts from true TEIs (Evrony et al., 2016). Using similar parameters to previously published work (Evrony et al., 2015), we required ≥ 2 discordant reads on the positive strand and ≥ 2 discordant reads on the negative strand, ≥ 4 aligned clipped reads, and a signal to noise score ≥ 9. The purpose of this score is to separate true insertions from noise driven by MDA chimeras (Evrony et al., 2015). Score = $2\sqrt{d_1 d_2} - (w_1 + w_2)$, where (d_1) are the number of discordant reads on the positive sense strand on the left side of the predicted insertion breakpoint and (d_2) on the negative strand on the right side of the insertion that support the predicted insertion, and to this, we subtract reads which do not support the predicted insertion, discordant reads on the negative strand and left side of the insertion (w_1) and discordant reads on the plus strand and right side of the insertion (w_2). We initially tested including insertions with and without a poly(A) tail and a TSD of size -15 to 30 but changed these parameters to requiring a poly(A) tail and a TSD of 5-30 base pairs to reduce the number of false positives.

TEI identification with *xTea*

We detected TEIs with *xTea* (Chu et al., 2021) (https://github.com/parklab/xTea). Candidate insertion breakpoint sites are first determined by collecting regions with enough clipped reads. The clipped region of the reads is aligned to a library of consensus TE sequences for additional support. Next, enough discordant reads support is required at the site as well, and *xTea* confirms whether its mate is aligned to the TE consensus sequence library. Finally, TE family-specific filters are implemented to reduce false positives. *xTea* candidates were classified as “high” or “low” confidence insertions depending on whether enough insertion supporting clipped and discordant were

distributed on both sides of the breakpoint or only one side, and on whether a poly(A) tail and TSD are detected.

We selected only insertions classified as “high confidence”. Since *xTea* was initially designed for bulk WGS sequencing data, we set MAX_COV_TIMES = 100 instead of 4. Regions where the coverage at the breakpoint $\geq$ average coverage * MAX_COV_TIMES are flagged by *xTea*. The purpose of increasing this is to account for uneven and high coverage which occurs in MDA sequencing data.

Annotation of known non-reference TEIs and sensitivity

To test the sensitivity for detection of TEIs in single cells obtained with *memTea*, we first determined whether the detected insertions overlapped with the breakpoints of TEIs previously detected in several published studies (Ewing & Kazazian, 2010; Gardner et al., 2017; Huang et al., 2010; Iskow et al., 2010; Stewart et al., 2011; Wang et al., 2006) obtained from Evrony *et al*. (Evrony et al., 2015). Insertion candidates obtained with *xTea* were also overlapped with the breakpoints observed in the 1000 genomes data (Gardner et al., 2017). Since TEIs were detected using different algorithms in these studies and insertion breakpoints may be imprecise, a 50 bp margin was added to the observed candidates in single nuclei before performing this comparison.

Sensitivity was defined as the number of KNR insertions detected in single cells divided by the number of germline KNR insertions detected for an individual sample in bulk. This is a lower bound of the true sensitivity, since this does not account for false negatives from the filtering parameters used to detect high confidence TEIs, nor insertions that were not detected in the WGS bulk data at lower confidence levels. To address this, we also compared the sensitivity for detection of KNR L1 TEIs in all calls

compared to calls classified as “high confidence” by *xTea.* In single nuclei from the adolescent individual 1465, 89% of KNR TEIs detected in a single neuron were classified as “high confidence”. The sensitivity of detection of KNR TEIs detected in heart bulk tissue in these single neurons was 73% with high confidence filtering and 81% sensitivity without high confidence filtering. However, single cells had almost twice as many candidate insertions without “high confidence” with 1,963 candidates on average per single neuron compared to 1,019 per neuron high confidence candidates and 578 candidates and 287 “high confidence” candidates in bulk tissue. Given the increased precision, we decided to only include candidates classified as “high confidence”.

Identification of somatic TEIs

A custom R script (R Core Team, 2015) was used to detect somatic TEIs in single nuclei called with *memTea*. We tested multiple thresholds for the number of clipped and discordant reads and required < 5 reads each at the most stringent level. We required < 2 clipped reads and < 2 discordant reads within 5 bp of the insertion breakpoints in bulk tissue to exclude germline TEIs. A custom bash and R pipeline was used to detect somatic TEIs in single nuclei called with *xTea.* L1 candidates classified as “high confidence” were given a 50 base pair margin from the midpoint of the breakpoints and were excluded if they overlapped with KNR insertions obtained from previous studies (Beck et al., 2010; Ewing & Kazazian, 2010, 2011; Gardner et al., 2017; Hormozdiari et al., 2011; Huang et al., 2010; Iskow et al., 2010; Stewart et al., 2011; Wang et al., 2006) as well as reference SVA, reference young L1 (L1HS, L1PA2, L1PA3) or reference young Alu (AluY) (Smit, 2013-2015). Clipped or discordant reads in the bulk tissue (clip_reads_tmp0 and discordant_reads_tmp0) raw files in the *xTea* output were also given a 50 bp margin and candidates were excluded if there were any

supporting reads in the bulk tissue in these regions overlapping with the candidate insertion breakpoints. Somatic candidates were manually inspected using IGV 2.4.19 (Thorvaldsdottir et al., 2013) to visually confirmed the absence of supporting reads in the bulk tissue and to confirm the presence of reads that support a retrotransposition event (Lee et al., 2012). All of the candidates from healthy aging neurons were manually inspected. We scored candidates from disease tissues where there were too many candidates to visually inspect using the signal-to-noise scoring metric described previously for *memTea* Score = $2\sqrt{d_1 d_2} - (w_1 + w_2)$. We then visually inspected the highest-scoring candidates.

Deletions and overlap with TEI candidates

We overlapped *xTEA* candidates that did not pass the visual inspection for retrotransposon insertion with deletion candidates from the same single cell. Since single-cell WGS MDA data has uneven amplification and chimeric artifacts which may resemble structural variants, a novel computational called *PhaseDel* (J. Kim, *et al.* article under preparation) was developed to detect high confidence focal somatic deletions. *PhaseDel* first uses DELLY2 (Rausch et al., 2012) to obtain candidate deletions using linkage information between nearby single nucleotide polymorphisms (SNPs) and the deletion breakpoints, in a similar manner to the method described previously by our group to detect high confidence phased somatic single-nucleotide variants in single cells (Bohrson et al., 2019; Lodato et al., 2018).

TEI breakpoints were given a 40 bp margin and overlapped with deletion candidate breakpoints. Retrotransposon-mediated deletion candidates were visually inspected using IGV 2.4.19 (Thorvaldsdottir et al., 2013) to confirm deletions. We obtained the sequences of clipped reads at both deletion breakpoints and mapped these

to the human reference genome with BLAT (Kent, 2002). A candidate was considered to be a retrotransposon mediated deletion when the clipped reads at the initial *xTea* called breakpoint overlapping the first DELLY2 breakpoint mapped to the second deletion breakpoint. At this second breakpoint, we confirmed that there was a reference germline retrotransposon sequence present, and clipped reads in this second breakpoint mapped to the reference genome at the site of the original *xTea* breakpoint and first deletion breakpoint.

References

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., . . . Faulkner, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature, 479*(7374), 534-537. doi:10.1038/nature10531

Baumer, C., Fisch, E., Wedler, H., Reinecke, F., & Korfhage, C. (2018). Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification. *Sci Rep, 8*(1), 7476. doi:10.1038/s41598-018-25895-7

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., . . . Moran, J. V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell, 141*(7), 1159-1170. doi:10.1016/j.cell.2010.05.021

Bohrson, C. L., Barton, A. R., Lodato, M. A., Rodin, R. E., Luquette, L. J., Viswanadham, V. V., . . . Park, P. J. (2019). Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat Genet, 51*(4), 749-754. doi:10.1038/s41588-019-0366-2

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., . . . Iwamoto, K. (2014). Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron, 81*(2), 306-313. doi:10.1016/j.neuron.2013.10.053

Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A., & Walsh, C. A. (2014). Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep, 8*(5), 1280-1289. doi:10.1016/j.celrep.2014.07.043

Chu, C., Borges-Monroy, R., Viswanadham, V. V., Lee, S., Li, H., Lee, E. A., & Park, P. J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*.

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Marchetto, M. C., Muotri, A. R., Mu, Y., . . . Gage, F. H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proc Natl Acad Sci U S A, 108*(51), 20382-20387. doi:10.1073/pnas.1100273108

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Yeo, G. W., Mu, Y., Lovci, M. T., . . . Gage, F. H. (2009). L1 retrotransposition in human neural progenitor cells. *Nature, 460*(7259), 1127-1131. doi:10.1038/nature08248

De Cecco, M., Criscione, S. W., Peckham, E. J., Hillenmeyer, S., Hamm, E. A., Manivannan, J., . . . Sedivy, J. M. (2013). Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. *Aging Cell, 12*(2), 247-256. doi:10.1111/acel.12047

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., . . . Lasken, R. S. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A, 99*(8), 5261-5266. doi:10.1073/pnas.082089499

Derks, W., & Bergmann, O. (2020). Polyploidy in Cardiomyocytes: Roadblock to Heart Regeneration? *Circ Res, 126*(4), 552-565. doi:10.1161/CIRCRESAHA.119.315408

Dubnau, J. (2018). The Retrotransposon storm and the dangers of a Collyer's genome. *Curr Opin Genet Dev, 49*, 95-105. doi:10.1016/j.gde.2018.04.004

Erwin, J. A., Paquola, A. C., Singer, T., Gallina, I., Novotny, M., Quayle, C., . . . Gage, F. H. (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci, 19*(12), 1583-1591. doi:10.1038/nn.4388

Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., . . . Walsh, C. A. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell, 151*(3), 483-496. doi:10.1016/j.cell.2012.09.035

Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., . . . Walsh, C. A. (2015). Cell lineage analysis in human brain using endogenous retroelements. *Neuron, 85*(1), 49-59. doi:10.1016/j.neuron.2014.12.028

Evrony, G. D., Lee, E., Park, P. J., & Walsh, C. A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. *Elife, 5*. doi:10.7554/eLife.12966

Ewing, A. D., & Kazazian, H. H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res, 20*(9), 1262-1270. doi:10.1101/gr.106419.110

Ewing, A. D., & Kazazian, H. H., Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res, 21*(6), 985-990. doi:10.1101/gr.114777.110

Faulkner, G. J., & Billon, V. (2018). L1 retrotransposition in the soma: a field jumping ahead. *Mob DNA, 9*, 22. doi:10.1186/s13100-018-0128-1

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., . . . Devine, S. E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res, 27*(11), 1916-1929. doi:10.1101/gr.218032.116

Guo, C., Jeong, H. H., Hsieh, Y. C., Klein, H. U., Bennett, D. A., De Jager, P. L., . . . Shulman, J. M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. *Cell Rep, 23*(10), 2874-2880. doi:10.1016/j.celrep.2018.05.004

Hormozdiari, F., Alkan, C., Ventura, M., Hajirasouliha, I., Malig, M., Hach, F., . . . Eichler, E. E. (2011). Alu repeat discovery and characterization within human genomes. *Genome Res, 21*(6), 840-849. doi:10.1101/gr.115956.110

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., . . . Wang, J. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell, 148*(5), 873-885. doi:10.1016/j.cell.2012.02.028

Huang, C. R., Schneider, A. M., Lu, Y., Niranjan, T., Shen, P., Robinson, M. A., . . . Burns, K. H. (2010). Mobile interspersed repeats are major structural variants in the human genome. *Cell, 141*(7), 1171-1182. doi:10.1016/j.cell.2010.05.026

Iskow, R. C., McCabe, M. T., Mills, R. E., Torene, S., Pittard, W. S., Neuwald, A. F., . . . Devine, S. E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell, 141*(7), 1253-1261. doi:10.1016/j.cell.2010.05.020

Jacob-Hirsch, J., Eyal, E., Knisbacher, B. A., Roth, J., Cesarkas, K., Dor, C., . . . Rechavi, G. (2018). Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders. *Cell Res, 28*(2), 187-203. doi:10.1038/cr.2018.8

Jiao, X., Rosenlund, M., Hooper, S. D., Tellgren-Roth, C., He, L., Fu, Y., . . . Sjoblom, T. (2011). Structural alterations from multiple displacement amplification of a human genome revealed by mate-pair sequencing. *PLoS One, 6*(7), e22250. doi:10.1371/journal.pone.0022250

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664. doi:10.1101/gr.229202

Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W. W., Morrill, K., . . . Dubnau, J. (2017). Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. *PLoS Genet, 13*(3), e1006635. doi:10.1371/journal.pgen.1006635

Lasken, R. S., & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol, 7*, 19. doi:10.1186/1472-6750-7-19

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., 3rd, . . . Cancer Genome Atlas Research, N. (2012). Landscape of somatic retrotransposition in human cancers. *Science, 337*(6097), 967-971. doi:10.1126/science.1222077

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, W., Jin, Y., Prazak, L., Hammell, M., & Dubnau, J. (2012). Transposable elements in TDP-43-mediated neurodegenerative disorders. *PLoS One, 7*(9), e44099. doi:10.1371/journal.pone.0044099

Li, W., Prazak, L., Chatterjee, N., Gruninger, S., Krug, L., Theodorou, D., & Dubnau, J. (2013). Activation of transposable elements during aging and neuronal decline in Drosophila. *Nat Neurosci, 16*(5), 529-531. doi:10.1038/nn.3368

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., . . . Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science, 359*(6375), 555-559. doi:10.1126/science.aao4426

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., . . . Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science, 350*(6256), 94-98. doi:10.1126/science.aab1785

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Muotri, A. R., Chu, V. T., Marchetto, M. C., Deng, W., Moran, J. V., & Gage, F. H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature, 435*(7044), 903-910. doi:10.1038/nature03663

Muotri, A. R., Marchetto, M. C., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K., & Gage, F. H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature, 468*(7322), 443-446. doi:10.1038/nature09544

R Core Team. (2015). R: A language and environment for statistical computing. . Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics, 28*(18), i333-i339. doi:10.1093/bioinformatics/bts378

Rodic, N., Steranka, J. P., Makohon-Moore, A., Moyer, A., Shen, P., Sharma, R., . . . Burns, K. H. (2015). Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat Med, 21*(9), 1060-1064. doi:10.1038/nm.3919

Rodriguez-Martin, B., Alvarez, E. G., Baez-Ortega, A., Demeulemeester, J., Ju, Y. S., Zamora, J., . . . Tubio, J. M. C. (2017). Pan-cancer analysis of whole genomes

reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/08/24/179705.abstract

Scott, E. C., Gardner, E. J., Masood, A., Chuang, N. T., Vertino, P. M., & Devine, S. E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res, 26*(6), 745-755. doi:10.1101/gr.201814.115

Smit, A., Hubley, R & Green, P. . (2013-2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Stewart, C., Kural, D., Stromberg, M. P., Walker, J. A., Konkel, M. K., Stutz, A. M., . . . Genomes, P. (2011). A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet, 7*(8), e1002236. doi:10.1371/journal.pgen.1002236

Tam, O. H., Rozhkov, N. V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., . . . Gale Hammell, M. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep, 29*(5), 1164-1177 e1165. doi:10.1016/j.celrep.2019.09.066

Terry, D. M., & Devine, S. E. (2019). Aberrantly High Levels of Somatic LINE-1 Expression and Retrotransposition in Human Neurological Disorders. *Front Genet, 10*, 1244. doi:10.3389/fgene.2019.01244

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017

Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sanchez-Luque, F. J., Bodea, G. O., . . . Faulkner, G. J. (2015). Ubiquitous L1 mosaicism in hippocampal neurons. *Cell, 161*(2), 228-239. doi:10.1016/j.cell.2015.03.026

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M. A., & Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat, 27*(4), 323-329. doi:10.1002/humu.20307

Xing, J., Zhang, Y., Han, K., Salem, A. H., Sen, S. K., Huff, C. D., . . . Jorde, L. B. (2009). Mobile elements create structural variation: analysis of a complete human genome. *Genome Res, 19*(9), 1516-1526. doi:10.1101/gr.091827.109

Zhao, B., Wu, Q., Ye, A. Y., Guo, J., Zheng, X., Yang, X., . . . Huang, A. Y. (2019). Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genet, 15*(4), e1008043. doi:10.1371/journal.pgen.1008043

Zhao, X., Collins, R. L., Lee, W. P., Weber, A. M., Jun, Y., Zhu, Q., . . . Talkowski, M. E. (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am J Hum Genet*. doi:10.1016/j.ajhg.2021.03.014

Zhou, W., Emery, S. B., Flasch, D. A., Wang, Y., Kwan, K. Y., Kidd, J. M., . . . Mills, R. E. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res, 48*(3), 1146-1163. doi:10.1093/nar/gkz1173

Zhu, X., Zhou, B., Pattni, R., Gleason, K., Tan, C., Kalinowski, A., . . . Urban, A. E. (2021). Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia. *Nat Neurosci, 24*(2), 186-196. doi:10.1038/s41593-020-00767-4

Chapter 4.
Discussion and Future Direction

Summary

In this dissertation, we explored the prevalence of transposable element insertions (TEIs) as a mutagenic cause and consequence of neurological disorders, neurodegeneration, and aging. In Chapter 2, we examined a large cohort of autism spectrum disorder (ASD) simplex families and analyzed cases and controls for *de novo* TEIs at a genome-wide resolution. We discovered that although *de novo* rates between cases and controls are similar, cases had more *de novo* L1 TEIs in ASD genes than expected. The majority of these TEIs occurred in non-coding genic regions, suggesting that non-coding insertions could have a phenotypic impact. We also detected exonic TEIs in loss-of-function (LoF) genes in cases, including a causal exonic Alu insertion in *CSDE1*, a known ASD gene (Guo et al., 2019). In Chapter 3 we focused on somatic TEIs in the human brain in aging glia and neurons as well as Cockayne syndrome (CS), Xeroderma pigmentosum (XP), Ataxia telangiectasia (AT), Alzheimer's disease (AD), and Amyotrophic lateral sclerosis (ALS) neurons. We confirmed previous findings of low rates of somatic retrotransposition in the brain (Erwin et al., 2016; Evrony et al., 2012; Evrony et al., 2015; B. Zhao et al., 2019) are low and observed <1 event per single cell. We additionally reported retrotransposition-independent L1 mediated deletions as observed previously (Erwin et al., 2016). We did not observe an increase in TEI rates with aging or neurological disorders with our current sample size and pipeline sensitivity. In this section, we will address these results within the broader context of current research and suggest future research approaches and directions for this field.

Discussion

Assessing the effects of TEIs

In Chapter 2 we discovered TEIs in individuals with ASD in genes associated with ASD, genes expressed in the fetal brain as well as genes annotated as LoF intolerant. The majority of these insertions occurred in non-coding regions of the genome, including introns. TEIs may cause disease, not only by disrupting exonic coding sequences, but intronic and non-coding insertions can affect gene splicing (Hancks & Kazazian, 2016) and can cause exon skipping, new exonization, or even disrupt RNA polymerase kinetics (Kaer & Speek, 2012). To determine whether some of the observed insertions were causal, we compared the phenotypes of ASD individuals carrying these insertions to other individuals with other types of mutations in the same genes from the existing literature (Table 2.3). Although we were able to match the phenotype of an individual carrying an exonic Alu insertion in *CSDE1* to the observed phenotype in individuals with LoF mutations in this gene (Guo et al., 2019), we were unable to accurately match the phenotypes of the other cases carrying intronic *de novo* TEIs to phenotypes in the existing literature. In some cases, such as in an affected individual carrying a *de novo* L1 intronic insertion in *HCN1*, certain characteristics did match previously reported phenotypes. This individual presented ASD, Tourette's syndrome, and gastrointestinal problems (Table 2.3) (Marini et al., 2018; Tsetsos et al., 2020), yet did not present signs of intellectual disability or epilepsy. Individuals carrying causal intronic TEIs may present a milder phenotype if this insertion causes aberrant splicing instead of loss of gene function. Because of this, we hypothesize that by studying non-coding structural variants (SVs) including TEIs, novel ASD genes that are lethal as LoF mutations may be discovered. To do so, we recommend focusing on intronic variants in the high probability of LoF intolerance (pLI) genes with high brain expression levels

during development. From the insertions detected in the SSC cohort, *DAB1* is an interesting gene candidate. We detected and validated a full-length *de novo* intronic L1 insertion in this high pLI gene (pLI=0.981) (Lek et al., 2016) with high gene expression in the developing cortex (Miller et al., 2014). *DAB1* plays a role in neuronal migration through the Reelin pathway in an isoform-dependent manner (Gao & Godbout, 2013), suggesting that splicing mutations could cause phenotypic consequences. The Reelin pathway is strongly associated with neurodevelopmental disorders including ASD and schizophrenia (Chen et al., 2017; Folsom & Fatemi, 2013; Lammert, Middleton, Pan, Olson, & Howell, 2017; Li, Guo, & Xiao, 2015; Nawa et al., 2020; Z. Wang et al., 2014). There is currently not enough evidence to classify *DAB1* as a high-confidence ASD gene since only rare missense mutations have been detected in ASD cases, and previous large-scale analyses studying LoF exonic mutations have not detected an association of *DAB1* with ASD (Abrahams et al., 2013; Nawa et al., 2020)

The consequences of non-coding TEIs on gene expression are currently difficult to predict without extensive experiments. For example, Ganguly *et al.* reported an intronic X-linked Alu insertion at the 3' end of intron 18 in the Factor VIII gene, which they hypothesized was the cause of hemophilia A in a child (Ganguly, Dunbar, Chen, Godmilow, & Ganguly, 2003). However, their bioinformatical analyses using both GENESCAN (http://argonaute.mit.edu/GENSCAN.html) and GeneSplicer (Pertea, Lin, & Salzberg, 2001) did not predict any consequences to gene splicing. Using reverse transcription PCR in lymphoblast cell lines (LCL) from the affected individual and in his unaffected mother, they were able to show that the Alu insertion caused skipping of exon 19.

Using cell lines and cellular modeling is a direct, albeit resource and time-intensive approach to experimentally validating the consequences of intronic TEIs for

neurological disorders and diseases. In X-linked Dystonia-Parkinsonism, a neurodegenerative disorder with a founder haplotype, researchers used patient and controls' derived skin fibroblasts, induced pluripotent stem cells (iPSCs), iPSC-derived neural stem cells, and IPSCs derived neurons to study the effects of an SVA insertion in intron 32 of the *TAF1* gene (Aneichyk et al., 2018). Using strand-specific RNA-sequencing and long-read target mRNA capture sequencing, they were able to show that this insertion resulted in aberrant alternative splicing and intron retention. Importantly, they demonstrated that normal gene expression could be rescued by using CRIPSR/Cas9 to remove the SVA insertion. These experiments are elegant ways of demonstrating the effects of intronic TEIs but are difficult to scale and to use for the analysis of large cohorts.

In a higher-throughput analysis, Payer *et al.* developed an ectopic minigene splicing assay to assess whether inherited polymorphic intronic Alu insertions alter gene splicing (Payer et al., 2019). They focused on Alu insertions within 100 bp of exons, where they confirmed previous findings of depletion of polymorphic Alus (Lev-Maor et al., 2008; Payer et al., 2019; Zhang, Romanish, & Mager, 2011). With this method, they were able to clone an exon and surrounding intronic region from individuals with and without the Alu insertions into a vector containing two rat insulin exons. They tested 23 different loci and observed that intronic Alus promoted exon skipping in 4 loci as well as increased exon inclusion in one locus. This included an Alu intronic insertion which has been previously associated with multiple sclerosis in *CD58* (Payer et al., 2017). Although this method is promising, cloning TEIs further away from exons or which are larger (~6kb for a full-length L1 insertion compared to ~300bp for Alus) has not been tested and may not be feasible. The development of more high throughput methods for testing gene expression of non-coding TEs will be very valuable.

Predicting the impact of non-coding TEIs on gene expression using computational methods is currently an important challenge. A basic approach which we applied in Chapter 2 is to overlap the TEI coordinates with genomic features including regulatory regions and transcription binding factor sites, epigenetic landscapes, or to identify regions under evolutionary selection (Goerner-Potvin & Bourque, 2018). Another approach has been through the identification of TE-associated expression quantitative trait loci (eQTLs) and splicing quantitative trait loci (sQTLs). Here, genomic regions, in this case polymorphic TEIs, are associated with a phenotype such as gene expression or alternative splicing isoforms using statistical and computational methods (Nica & Dermitzakis, 2013). TE associated eQTLs have been predicted for lymphoblastoid cell lines (Goubert, Zevallos, & Feschotte, 2020a; Spirito, Mangoni, Sanges, & Gustincich, 2019; Sudmant et al., 2015; L. Wang, Norris, & Jordan, 2017) as well as iPSCs (Goubert et al., 2020a). Using GTEx gene expression data from multiple tissues in 639 individuals, Cao *et al*. also identified 2,422 polymorphic TEIs with a correlation with gene expression in 6,342 genes and 1,427 polymorphic TEIs correlated with splicing (Cao et al., 2020). Importantly, they also showed that these TE-associated eQTLs and sQTLs are tissue-specific and that TEs in genic regions or 10 kb upstream and downstream to genes are more likely to impact expression levels and splicing. Although these findings are useful for studying the consequences of polymorphic TEIs, this approach does not address rare *de novo* TEs.

Important advances are being made with machine learning, where existing gene expression and genotype data has been used to train and predict the effects of mutations on alternative splicing (Goubert, Zevallos, & Feschotte, 2020b) or to predict whether non-coding variants are pathogenic (Wells et al., 2019). Researchers determined that *de novo* SNVs and indels predicted to alter splicing were enriched in

individuals with ASD using a method called SpliceAI (Goubert et al., 2020a). These predictions had high validation rates and Goubert *et al.* estimated that 9%-11% of cases with neurodevelopmental disorders could be attributed to splicing altering mutations. In a different approach, researchers used pathogenic non-coding variants from different human databases and trained a supervised machine learning model incorporating genomic features such as chromatin structure, existing deleteriousness, and essentiality metrics, and gene expression information to predict the impact of non-coding mutations (Wells et al., 2019). Here they tested non-coding deletions previously associated with ASD (Brandler et al., 2018) and reported more non-coding deletions in non-coding regions essential to gene regulation than expected by chance in cases only. These methods are promising but have not been designed for or rigorously tested with TEIs. However, they highlight the impact of non-coding mutations in neurodevelopmental disorders and ASD. A similar approach using polymorphic TE-associated eQTLs and sQTLs as well as regulatory genomic features could be implemented to predict the impact of TEs in non-coding regions.

Mosaic TEIs in ASD

In Chapter 2, we focused on *de novo* TEIs by excluding regions of the genome where we detected parental insertion read support at the predicted breakpoints. However, during visual inspections using the Integrative Genomics Browser (Busan & Weeks, 2017), we detected 4 parental mosaic candidates with ≤ 2 clipped reads, below the threshold for detection by *xTEA,* and/or discordant reads at a low alternate allele frequency (AAF) compared to the observed read depth in that region in the parents of individuals with candidate TEIs. This included a candidate parental mosaic heterozygous exonic Alu TEI which we predicted to be causal in the proband, where the insertion was

detected as a heterozygous mutation in the ASD gene *CSDE1* (Guo et al., 2019). We performed full-length PCR validation of the insertion in both parental samples and the affected individual in LCL DNA (Figure 2.10), and only validated the insertion in the proband. Although we cannot exclude sample contamination at the time of sample and library preparation or sequencing, another possibility is that this TEI was mosaic in the parent and was inherited by the proband. This suggests that mosaic parental TEIs may cause ASD, although further sequencing of the parental blood is necessary to confirm this. In an analysis of exome sequencing data in ASD families from the Simons Simplex Collection (SSC), it was estimated that 7%-10% of parental mosaic SNVs are transmitted to their children and account for 6.8% of SNVs which were previously presumed to be *de novo* (Krupp et al., 2017). If this rate is similar for TEIs, this suggests that we filtered most inherited parental mosaic TEIs.

During visual inspection of candidate *de novo* TEIs in Chapter 2, we also classified 5 candidate TEIs as “somatic candidates”. These insertions had a low AAF with clear read support, target site duplications (TSD), and poly(A) tails, but only a few supporting reads compared to the overall read depth at the insertion site. We performed full-length PCR validation in LCL DNA for 2 of these candidates. One of these candidate somatic TEIs was in *DPYD*, a gene involved in the catabolism of pyrimidine bases, and a SFARI ASD gene with a “suggestive evidence” classification (Abrahams et al., 2013). Microdeletions in this gene have been associated with ASD previously (Carter et al., 2011). However, we did not observe a supporting PCR band. This suggests that these insertions are either false positives or low AAF mosaic blood insertions which we were unable to validate in LCL DNA. Further testing in the proband’s blood is necessary to confirm this. Detecting mosaic TEIs based on their AAF in ~40x short-read whole-genome sequencing (WGS) data is challenging. Unlike SNVs, we observed that

heterozygous polymorphic TEIs are not always observed at an AAF near 0.5. Reads mapping to retrotransposons are frequently multi-mappers (Goerner-Potvin & Bourque, 2018) due to their repetitive sequences, Poly(A) stretches, and the prevalence of TEs in the genome. Read depth is also important when detecting mosaic variants, especially at low AAFs. MosaicHunter, a tool for detecting mosaic SNVs in WGS and whole exome sequencing (WES) data, has a sensitivity lower than 40% for AAF variants < 0.1 in 40x data but this jumps to more than 70% at a 60x read depth coverage (Huang et al., 2017). This suggests that we have a low sensitivity for accurately detecting mosaic TEIs in this data based on AAF read support.

There has been no detailed investigation of somatic TE mosaicism in ASD, and to our knowledge, the candidates detected in Chapter 2 are the first mosaic TEIs impacting ASD genes detected in cases. A previous analysis detected higher levels of L1 ORF2 sequences in post-mortem brain tissue from individuals with Rett syndrome (Muotri et al., 2010), a neurodevelopmental disorder with autistic features. However, they used qPCR for this analysis, which as described in Chapter 3 is not a reliable technique for quantifying TEIs. The methods for the detection of somatic TEIs described in Chapter 3 could be applied to blood or brain tissues from individuals with ASD to explore this further. This includes bulk targeted sequencing (B. Zhao et al., 2019) or deep WGS TEI analysis (Zhu et al., 2021) of bulk blood or brain tissue. Somatic SNVs in ASD have been studied by several groups in WES data from the SSC and the Autism Sequencing Consortium (Dou et al., 2017; Freed & Pevsner, 2016; Krupp et al., 2017; Lim et al., 2017). Results between these studies were varied, but they estimated a contribution of mosaic SNVs to ~3-5% of ASD cases (D'Gama & Walsh, 2018). The average number of mosaic SNVs per individual is low, with < 0.02 loss-of-function mosaic variants per individual (Freed & Pevsner, 2016; Lim et al., 2017). Given the lower

rates of *de novo* and somatic retrotransposition compared to SNVs (Acuna-Hidalgo, Veltman, & Hoischen, 2016; Rodin et al., 2021), these studies would probably require larger sample sizes than is currently feasible with these methods or higher throughput methods for analyzing somatic retrotransposition for sufficient power to observe a significant difference between cases and controls for damaging mosaic TEIs.

Detecting TEIs with long-read sequencing

Long-read sequencing provides many advantages for the detection of TEIs in sequencing data. The two most popular and validated long-read sequencing technologies currently are single-molecule real-time (SMRT) sequencing from Pacific BioSciences (PacBio) (Rhoads & Au, 2015) and Oxford Nanopore Technologies (ONT) sequencing (Jain, Olsen, Paten, & Akeson, 2016). These methods produce reads 1-1.5 x 10^4 bp and 2-5 x 10^3 bp long on average respectively compared to Illumina HiSeq 2500 short-read sequencing which produces reads with a maximum length of 250 bp (Rhoads & Au, 2015). The main advantage of using these technologies for the detection of SVs and TEIs is that long reads may contain the entire insertion sequence in a single read, providing an opportunity to clearly detect and describe complex insertion events (Goerner-Potvin & Bourque, 2018; X. Zhao et al., 2021). Additionally, SVs may be identified more easily with long-read sequencing technologies in regions that have been challenging to analyze previously with short-reads due to their poor mapability, including segmental duplications, simple repeat regions (X. Zhao et al., 2021), centromeres, telomeres, and regions with polymorphic and reference TEs (Chu, Zhao, Park, & Lee, 2020).

Several tools developed for detecting SVs in long-read sequencing data identify TEs as well (Chu et al., 2020; Goerner-Potvin & Bourque, 2018) including *SNIFFLES*

(Sedlazeck et al., 2018), *pbsv* (https://github.com/PacificBiosciences/pbsv), *SVIM* (Heller & Vingron, 2019), and *SMRT-SV* (Huddleston et al., 2017). Tools designed specifically for detecting TEs in long-read sequencing data may be preferable for a more detailed TE annotation (Chu et al., 2020). *PALMER* detects L1 insertions in PacBio data (Zhou et al., 2020). *xTea* was developed in collaboration with our group to detect TE insertions in PacBio, ONT, and 10x genomics sequencing libraries as well as in short-read data (Chu et al., 2021). In *Drosophila*, LoRTE was developed to identify TEIs from PacBio data (Disdero & Filee, 2017).

We observed that the sensitivity for detection of TEIs with *xTea* was higher for PacBio High Fidelity reads, with a 91%, 93%, and 90% sensitivity for Alu, L1, and SVA compared to the specificity in Illumina data of 88%, 93%, and 86% respectively (Chu et al., 2021). In a previous study of two haploid human genomes sequenced with SMRT-seq, 89% of the SVs identified had not been detected in the 1000 Genomes Project (Huddleston et al., 2017). In a recent comparison of 3 families using short-read and long-read sequencing technologies, ~11,000 SVs including TEIs were detected per individual with short-reads compared to ~25,000 SVs with long reads (X. Zhao et al., 2021). Additionally, 25.1%, 57.4%, and 49.3% of short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and long-terminal repeat (LTR) transposons were only detected with long-read WGS (X. Zhao et al., 2021). Importantly, 95.8% of these long-read specific TEIs had a significant number of supporting reads at the insertion site but were missed by *MELT* in short-read data, suggesting that improved algorithms close the sensitivity gap between short and long-read data.

Although this technology is promising, long-read WGS is at least ~6 fold more expensive than short-read WGS (X. Zhao et al., 2021), limiting the sample size for these studies. Currently, most available sequencing data from large cohorts have been

sequenced with short-read WGS or WES. Future studies will probably continue to analyze these datasets with a lower sensitivity until sequencing many individuals with long-reads is cost-effective. Additionally, long-read sequencing error rates are higher, with an average of ~13% error rate for ONT and PacBio and a 3.7% and 8% insertion error rate respectively (Dohm, Peters, Stralis-Pavese, & Himmelbauer, 2020), compared to an error rate of ~0.1% with Illumina HiSeq (Stoler & Nekrutenko, 2021). Using short-read Illumina data to correct long-read sequencing errors can be a useful approach. This decreases the error rate to 0.85% for ONT and 1.15% for PacBio (Dohm et al., 2020).

Novel methods for detecting somatic mutations and TEIs

We analyzed WGS data from 159 single cells amplified these with multiple displacement amplification (MDA) (Dean et al., 2002). As discussed in Chapter 3, we encountered uneven amplification, reducing our sensitivity for detecting TEIs. This reduced sensitivity has been observed previously in MDA amplified single cells and is primarily driven by amplification dropout in satellite DNA regions and high-GC content regions (Evrony et al., 2015). Interestingly, we observed an even lower sensitivity and uneven amplification in aged and diseased cells (Figure 3.1 and Figure 3.2). The MDA DNA's polymerase Phi29 has a high processivity, low error rates, proofreading activity, and obtains amplicons >10kb (Dean et al., 2002). However, Phi29 may be sensitive to template fragmentation as well as stop sites and alterations in the DNA templates such as abasic sites, double-strand DNA breaks, DNA crosslinking, and thymine dimers (Baumer, Fisch, Wedler, Reinecke, & Korfhage, 2018). Undamaged DNA may outcompete DNA templates with damage and stop sites, decreasing the amplification in these regions (Baumer et al., 2018). We hypothesize that the aged and diseased cells

analyzed may have more damage, resulting in lower sensitivity for the detection of TEIs and uneven amplification by MDA.

Recently, alternative single-cell amplification methods, as well as methods for detecting somatic mutations with lower error rates, have been developed. One of these methods, multiplexed end-tagging amplification of complementary strands (META-CS), fragments and tags both complementary DNA strands with various transposon sequences before amplification to identify and sequence both strands in single cells (Xing, Tan, Chang, Li, & Xie, 2021). Somatic SNVs must be detected in both strands, reducing the number of false positives present in their data. Using this method, researchers were able to replicate previous findings of an increasing number of somatic SNVs in single neurons with age (Lodato et al., 2018), although their proposed somatic SNV rate was lower, which could indicate the presence of false positives in the previously analyzed MDA data (Xing et al., 2021).

Duplex sequencing in bulk data uses a similar approach and has a very low theoretical error rate of $< 10^{-9}$ (Salk, Schmitt, & Loeb, 2018). By sequencing barcoded molecules of both strands of DNA multiple times and achieving sequencing error rates lower than human mutation rates using methods like NanoSeq (Abascal et al., 2021) or BotSeqs (Hoang et al., 2016), researchers identified somatic SNVs in tissues without single-cell amplification. Although these duplex bulk WGS methods have not been applied to the detection of TEIs, they are interesting approaches to test in the future.

Another promising single-cell amplification method, primary template-directed amplification (PTA), also uses a Phi29 DNA polymerase but adds exonuclease-resistant terminators to the amplification reaction (Gonzalez et al., 2021). This creates a quasi-linear amplification instead of an exponential amplification and produces more even amplification coverage compared to MDA, while also reducing the amplification of errors

(Gonzalez et al., 2021). Nevertheless, MDA still has certain advantages to some of these methods. The most significant one is that MDA amplicons are the largest, and at >10kb (Dean et al., 2002) they can contain full-length L1 TEIs. This allows researchers to perform full-length PCR validation of these insertions in the original single-cell DNA product, which is crucial given the high rates of chimera-induced false positives in single-cell amplification methods. Additionally, in our preliminary analyses of META-CS data, we were unable to accurately detect polymorphic germline TEIs with *xTea*, given the uneven amplification and amplification blocks with definite borders obtained with this method. Our preliminary analyses of PTA single-cell data showed more undefined TEI breakpoint borders compared to MDA data, and more clipped reads. Since clipped reads are detected in the first step of *xTea*, the current algorithm was unable to handle the analysis of this data and will require further modifications. Finally, these single-cell amplification methods still present the same challenges as MDA in terms of a high cost for sequencing each cell and the resulting small sample size obtained. A combination of targeted sequencing with improved single-cell amplification techniques, or the refinement of bulk methodologies for the detection of somatic TEIs will greatly advance the field and could improve sensitivity in aged and diseased cells.

Somatic retrotransposition in post-mitotic neurons

A systematic understanding of whether somatic retrotransposition is supported by mature non-dividing neurons in the human brain is still lacking. There is a growing body of literature that recognizes that somatic retrotransposition occurs during neuronal cell division and brain development. *In vitro* support from neural stem cells (Muotri et al., 2005) and neural progenitor cells (Coufal et al., 2009) using reporter constructs first suggested this, but the most direct evidence has been through the discovery and

extensive validation of clonal TEIs from post-mortem human tissue (Evrony et al., 2015; Zhu et al., 2021). In these two studies, droplet digital PCR was used to quantify levels of mosaicism of validated TEIs detected from WGS and determined that these insertions were present in varying levels of mosaicism throughout the brain, suggesting that the initial insertion occurred during brain development and was later expanded during cell division. Determining whether somatic TEIs are exclusive to individual neurons in the human brain and occur in non-dividing cells is more challenging since the brain contains ~86 billion neurons (Azevedo et al., 2009) and this would require sequencing every neuron, or at least every neuron within its clade. Currently, the only alternative is to detect a significant increase of retrotransposition after brain development in certain conditions, such as with aging or neurodegeneration in multiple individuals. In our analysis of aging and diseased neurons, we did not detect an increase in somatic TEIs.

Several lines of evidence suggest that cell divisions are necessary to promote retrotransposition (Kubo et al., 2006; Shi, Seluanov, & Gorbunova, 2007; Xie et al., 2013). However, data from *Drosophila* have suggested that there is an increase in somatic retrotransposition in the aging brain (Li et al., 2013). Here, flies were injected with a construct containing a green fluorescent protein (GFP) reporter that became expressed when *gypsy*, an LTR element, retrotransposed into the *ovo* locus. This element preferentially inserts in the *ovo* gene in flies (Dej, Gerasimova, Corces, & Boeke, 1998). Researchers observed a dramatic increase in brain GFP expression in aged flies (Li et al., 2013), suggesting elevated somatic *gypsy* insertions. Using WGS from pooled neurons from αβ neurons from the *Drosophila* mushroom body, other researchers identified more than 200 somatic TEIs and estimated 129 somatic insertions per single neuron (Perrat et al., 2013). However, in a subsequent analysis from the same group, the authors concluded that most observed insertions were chimeric and

experimental artifacts, and they did not observe an increase in somatic TEI rates in aging flies (Treiber & Waddell, 2017). These results parallel findings in humans, where *in vitro* and artificial retrotransposition assays suggest high levels of somatic insertion rates, but careful studies from sequencing data predict low levels.

One of the main current studies supporting somatic retrotransposition in mature non-dividing human neurons (Faulkner & Garcia-Perez, 2017; Terry & Devine, 2019) is an *in vitro study* which used an L1 enhanced GFP (EGFP) reporter in mature neurons derived from H9 human embryonic stem cells derived neural progenitor cells (Macia et al., 2017). This assay requires the introduction via viral infection of a cassette containing a retrotransposition-competent L1 which was previously developed to express EGFP presumably only upon retrotransposition of the L1 and EGFP in the cassette. Here, the EGFP sequence is interrupted by an intronic sequence which is then spliced out (Coufal et al., 2009). The copy number of these presumed TEIs were then quantified via qPCR and these researchers observed higher L1-EGFP copy numbers in these transformed mature neurons compared to neural progenitor cells (Macia et al., 2017). As discussed in Chapter 3, qPCR may be inaccurate (Evrony, Lee, Park, & Walsh, 2016; Reilly, Faulkner, Dubnau, Ponomarev, & Gage, 2013; Terry & Devine, 2019), and this method does not quantify endogenous L1 retrotransposition. Additionally, TE regulation and repression could potentially be affected through cell reprogramming, *in vitro* maintenance, and cell infection, and direct and carefully validated evidence of endogenous TE retrotransposition in human brains will be the highest level of support of somatic retrotransposition which can be achieved. These examples highlight the importance of investigating somatic retrotransposition rates using alternative methods and of confirming *in vitro* findings *in vivo* or at least *ex vivo*.

Conclusion

The present research aimed to examine the prevalence of retrotransposition events in humans both as *de novo* variants in a large ASD cohort as well as somatic insertions in the human brain. This research has shown that *de novo* TEIs play a small role in ASD, but insertions may be severe enough to cause this disorder. The second major finding was that non-coding L1 insertions are present in cases at higher rates than expected by chance. Whilst the investigation of somatic mutations in the human brain did not confirm an increase in aging and disease, it did partially substantiate the overall low rates of somatic L1 insertions in neurons. This is also the largest study so far analyzing retrotransposition at a whole-genome resolution in single cells. The issue of somatic retrotransposition in non-dividing neurons is an intriguing question that could be explored in further research using novel technologies for sequencing and analysis.

References

Abascal, F., Harvey, L. M. R., Mitchell, E., Lawson, A. R. J., Lensing, S. V., Ellis, P., . . . Martincorena, I. (2021). Somatic mutation landscapes at single-molecule resolution. *Nature, 593*(7859), 405-410. doi:10.1038/s41586-021-03477-4

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., . . . Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism, 4*(1), 36. doi:10.1186/2040-2392-4-36

Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology, 17*(1), 241. doi:10.1186/s13059-016-1110-1

Aneichyk, T., Hendriks, W. T., Yadav, R., Shin, D., Gao, D., Vaine, C. A., . . . Talkowski, M. E. (2018). Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell, 172*(5), 897-909 e821. doi:10.1016/j.cell.2018.02.011

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., . . . Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol, 513*(5), 532-541. doi:10.1002/cne.21974

Baumer, C., Fisch, E., Wedler, H., Reinecke, F., & Korfhage, C. (2018). Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification. *Sci Rep, 8*(1), 7476. doi:10.1038/s41598-018-25895-7

Brandler, W. M., Antaki, D., Gujral, M., Kleiber, M. L., Whitney, J., Maile, M. S., . . . Sebat, J. (2018). Paternally inherited cis-regulatory structural variants are associated with autism. *Science, 360*(6386), 327-331. doi:10.1126/science.aan2261

Busan, S., & Weeks, K. M. (2017). Visualization of RNA structure models within the Integrative Genomics Viewer. *RNA, 23*(7), 1012-1018. doi:10.1261/rna.060194.116

Cao, X., Zhang, Y., Payer, L. M., Lords, H., Steranka, J. P., Burns, K. H., & Xing, J. (2020). Polymorphic mobile element insertions contribute to gene expression and alternative splicing in human tissues. *Genome Biol, 21*(1), 185. doi:10.1186/s13059-020-02101-4

Carter, M. T., Nikkel, S. M., Fernandez, B. A., Marshall, C. R., Noor, A., Lionel, A. C., . . . Scherer, S. W. (2011). Hemizygous deletions on chromosome 1p21.3 involving the DPYD gene in individuals with autism spectrum disorder. *Clin Genet, 80*(5), 435-443. doi:10.1111/j.1399-0004.2010.01578.x

Chen, N., Bao, Y., Xue, Y., Sun, Y., Hu, D., Meng, S., . . . Shi, J. (2017). Meta-analyses of RELN variants in neuropsychiatric disorders. *Behav Brain Res, 332*, 110-119. doi:10.1016/j.bbr.2017.05.028

Chu, C., Borges-Monroy, R., Viswanadham, V. V., Lee, S., Li, H., Lee, E. A., & Park, P. J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun*.

Chu, C., Zhao, B., Park, P. J., & Lee, E. A. (2020). Identification and Genotyping of Transposable Element Insertions From Genome Sequencing Data. *Curr Protoc Hum Genet, 107*(1), e102. doi:10.1002/cphg.102

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Yeo, G. W., Mu, Y., Lovci, M. T., . . . Gage, F. H. (2009). L1 retrotransposition in human neural progenitor cells. *Nature, 460*(7259), 1127-1131. doi:10.1038/nature08248

D'Gama, A. M., & Walsh, C. A. (2018). Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci, 21*(11), 1504-1514. doi:10.1038/s41593-018-0257-3

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., . . . Lasken, R. S. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A, 99*(8), 5261-5266. doi:10.1073/pnas.082089499

Dej, K. J., Gerasimova, T., Corces, V. G., & Boeke, J. D. (1998). A hotspot for the Drosophila gypsy retroelement in the ovo locus. *Nucleic Acids Res, 26*(17), 4019-4025. doi:10.1093/nar/26.17.4019

Disdero, E., & Filee, J. (2017). LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mob DNA, 8*, 5. doi:10.1186/s13100-017-0088-x

Dohm, J. C., Peters, P., Stralis-Pavese, N., & Himmelbauer, H. (2020). Benchmarking of long-read correction methods. *NAR Genom Bioinform, 2*(2), lqaa037. doi:10.1093/nargab/lqaa037

Dou, Y., Yang, X., Li, Z., Wang, S., Zhang, Z., Ye, A. Y., . . . Wei, L. (2017). Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Hum Mutat, 38*(8), 1002-1013. doi:10.1002/humu.23255

Erwin, J. A., Paquola, A. C., Singer, T., Gallina, I., Novotny, M., Quayle, C., . . . Gage, F. H. (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci, 19*(12), 1583-1591. doi:10.1038/nn.4388

Evrony, G. D., Cai, X., Lee, E., Hills, L. B., Elhosary, P. C., Lehmann, H. S., . . . Walsh, C. A. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell, 151*(3), 483-496. doi:10.1016/j.cell.2012.09.035

Evrony, G. D., Lee, E., Mehta, B. K., Benjamini, Y., Johnson, R. M., Cai, X., . . . Walsh, C. A. (2015). Cell lineage analysis in human brain using endogenous retroelements. *Neuron, 85*(1), 49-59. doi:10.1016/j.neuron.2014.12.028

Evrony, G. D., Lee, E., Park, P. J., & Walsh, C. A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. *Elife, 5*. doi:10.7554/eLife.12966

Faulkner, G. J., & Garcia-Perez, J. L. (2017). L1 Mosaicism in Mammals: Extent, Effects, and Evolution. *Trends Genet, 33*(11), 802-816. doi:10.1016/j.tig.2017.07.004

Folsom, T. D., & Fatemi, S. H. (2013). The involvement of Reelin in neurodevelopmental disorders. *Neuropharmacology, 68*, 122-135. doi:10.1016/j.neuropharm.2012.08.015

Freed, D., & Pevsner, J. (2016). The Contribution of Mosaic Variants to Autism Spectrum Disorder. *PLoS Genet, 12*(9), e1006245. doi:10.1371/journal.pgen.1006245

Ganguly, A., Dunbar, T., Chen, P., Godmilow, L., & Ganguly, T. (2003). Exon skipping caused by an intronic insertion of a young Alu Yb9 element leads to severe hemophilia A. *Hum Genet, 113*(4), 348-352. doi:10.1007/s00439-003-0986-5

Gao, Z., & Godbout, R. (2013). Reelin-Disabled-1 signaling in neuronal migration: splicing takes the stage. *Cell Mol Life Sci, 70*(13), 2319-2329. doi:10.1007/s00018-012-1171-6

Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nat Rev Genet, 19*(11), 688-704. doi:10.1038/s41576-018-0050-x

Gonzalez, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., . . . Gawad, C. (2021). Accurate Genomic Variant Detection in Single Cells with Primary Template-Directed Amplification. *bioRxiv*, 2020.2011.2020.391961. doi:10.1101/2020.11.20.391961

Goubert, C., Zevallos, N. A., & Feschotte, C. (2020a). Contribution of unfixed transposable element insertions to human regulatory variation. *Philos Trans R Soc Lond B Biol Sci, 375*(1795), 20190331. doi:10.1098/rstb.2019.0331

Goubert, C., Zevallos, N. A., & Feschotte, C. (2020b). Correction to 'Contribution of unfixed transposable element insertions to human regulatory variation'. *Philos Trans R Soc Lond B Biol Sci, 375*(1811), 20200084. doi:10.1098/rstb.2020.0084

Guo, H., Li, Y., Shen, L., Wang, T., Jia, X., Liu, L., . . . Xia, K. (2019). Disruptive variants of CSDE1 associate with autism and interfere with neuronal development and synaptic transmission. *Sci Adv, 5*(9), eaax2166. doi:10.1126/sciadv.aax2166

Hancks, D. C., & Kazazian, H. H., Jr. (2016). Roles for retrotransposon insertions in human disease. *Mob DNA, 7*, 9. doi:10.1186/s13100-016-0065-9

Heller, D., & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics, 35*(17), 2907-2915. doi:10.1093/bioinformatics/btz041

Hoang, M. L., Kinde, I., Tomasetti, C., McMahon, K. W., Rosenquist, T. A., Grollman, A. P., . . . Papadopoulos, N. (2016). Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A, 113*(35), 9846-9851. doi:10.1073/pnas.1607794113

Huang, A. Y., Zhang, Z., Ye, A. Y., Dou, Y., Yan, L., Yang, X., . . . Wei, L. (2017). MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Res, 45*(10), e76. doi:10.1093/nar/gkx024

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., . . . Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res, 27*(5), 677-685. doi:10.1101/gr.214007.116

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol, 17*(1), 239. doi:10.1186/s13059-016-1103-0

Kaer, K., & Speek, M. (2012). Intronic retroelements: Not just "speed bumps" for RNA polymerase II. *Mob Genet Elements, 2*(3), 154-157. doi:10.4161/mge.20774

Krupp, D. R., Barnard, R. A., Duffourd, Y., Evans, S. A., Mulqueen, R. M., Bernier, R., . . . O'Roak, B. J. (2017). Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *Am J Hum Genet, 101*(3), 369-390. doi:10.1016/j.ajhg.2017.07.016

Kubo, S., Seleme, M. C., Soifer, H. S., Perez, J. L., Moran, J. V., Kazazian, H. H., Jr., & Kasahara, N. (2006). L1 retrotransposition in nondividing and primary human somatic cells. *Proc Natl Acad Sci U S A, 103*(21), 8036-8041. doi:10.1073/pnas.0601954103

Lammert, D. B., Middleton, F. A., Pan, J., Olson, E. C., & Howell, B. W. (2017). The de novo autism spectrum disorder RELN R2290C mutation reduces Reelin secretion and increases protein disulfide isomerase expression. *J Neurochem, 142*(1), 89-102. doi:10.1111/jnc.14045

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. doi:10.1038/nature19057

Lev-Maor, G., Ram, O., Kim, E., Sela, N., Goren, A., Levanon, E. Y., & Ast, G. (2008). Intronic Alus influence alternative splicing. *PLoS Genet, 4*(9), e1000204. doi:10.1371/journal.pgen.1000204

Li, W., Guo, X., & Xiao, S. (2015). Evaluating the relationship between reelin gene variants (rs7341475 and rs262355) and schizophrenia: A meta-analysis. *Neurosci Lett, 609*, 42-47. doi:10.1016/j.neulet.2015.10.014

Li, W., Prazak, L., Chatterjee, N., Gruninger, S., Krug, L., Theodorou, D., & Dubnau, J. (2013). Activation of transposable elements during aging and neuronal decline in Drosophila. *Nat Neurosci, 16*(5), 529-531. doi:10.1038/nn.3368

Lim, E. T., Uddin, M., De Rubeis, S., Chan, Y., Kamumbu, A., Zhang, X., . . . Walsh, C. A. (2017). Rates, Distribution, and Implications of Post-zygotic Mutations in Autism Spectrum Disorder

. *Submitted*.

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., . . . Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science, 359*(6375), 555-559. doi:10.1126/science.aao4426

Macia, A., Widmann, T. J., Heras, S. R., Ayllon, V., Sanchez, L., Benkaddour-Boumzaouad, M., . . . Garcia-Perez, J. L. (2017). Engineered LINE-1 retrotransposition in nondividing human neurons. *Genome Res, 27*(3), 335-348. doi:10.1101/gr.206805.116

Marini, C., Porro, A., Rastetter, A., Dalle, C., Rivolta, I., Bauer, D., . . . Depienne, C. (2018). HCN1 mutation spectrum: from neonatal epileptic encephalopathy to benign generalized epilepsy and beyond. *Brain, 141*(11), 3160-3178. doi:10.1093/brain/awy263

Miller, J. A., Ding, S. L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., . . . Lein, E. S. (2014). Transcriptional landscape of the prenatal human brain. *Nature, 508*(7495), 199-206. doi:10.1038/nature13185

Muotri, A. R., Chu, V. T., Marchetto, M. C., Deng, W., Moran, J. V., & Gage, F. H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature, 435*(7044), 903-910. doi:10.1038/nature03663

Muotri, A. R., Marchetto, M. C., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K., & Gage, F. H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature, 468*(7322), 443-446. doi:10.1038/nature09544

Nawa, Y., Kimura, H., Mori, D., Kato, H., Toyama, M., Furuta, S., . . . Ozaki, N. (2020). Rare single-nucleotide DAB1 variants and their contribution to Schizophrenia and autism spectrum disorder susceptibility. *Hum Genome Var, 7*(1), 37. doi:10.1038/s41439-020-00125-7

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci, 368*(1620), 20120362. doi:10.1098/rstb.2012.0362

Payer, L. M., Steranka, J. P., Ardeljan, D., Walker, J., Fitzgerald, K. C., Calabresi, P. A., . . . Burns, K. H. (2019). Alu insertion variants alter mRNA splicing. *Nucleic Acids Res, 47*(1), 421-431. doi:10.1093/nar/gky1086

Payer, L. M., Steranka, J. P., Yang, W. R., Kryatova, M., Medabalimi, S., Ardeljan, D., . . . Burns, K. H. (2017). Structural variants caused by Alu insertions are associated with risks for many human diseases. *Proc Natl Acad Sci U S A, 114*(20), E3984-E3992. doi:10.1073/pnas.1704117114

Perrat, P. N., DasGupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., & Waddell, S. (2013). Transposition-driven genomic heterogeneity in the Drosophila brain. *Science, 340*(6128), 91-95. doi:10.1126/science.1231965

Pertea, M., Lin, X., & Salzberg, S. L. (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res, 29*(5), 1185-1190. doi:10.1093/nar/29.5.1185

Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I., & Gage, F. H. (2013). The role of transposable elements in health and diseases of the central nervous system. *J Neurosci, 33*(45), 17577-17586. doi:10.1523/JNEUROSCI.3369-13.2013

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics, 13*(5), 278-289. doi:10.1016/j.gpb.2015.08.002

Rodin, R. E., Dou, Y., Kwon, M., Sherman, M. A., D'Gama, A. M., Doan, R. N., . . . Walsh, C. A. (2021). The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. *Nat Neurosci, 24*(2), 176-185. doi:10.1038/s41593-020-00765-6

Salk, J. J., Schmitt, M. W., & Loeb, L. A. (2018). Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet, 19*(5), 269-285. doi:10.1038/nrg.2017.117

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods, 15*(6), 461-468. doi:10.1038/s41592-018-0001-7

Shi, X., Seluanov, A., & Gorbunova, V. (2007). Cell divisions are required for L1 retrotransposition. *Mol Cell Biol, 27*(4), 1264-1270. doi:10.1128/MCB.01888-06

Spirito, G., Mangoni, D., Sanges, R., & Gustincich, S. (2019). Impact of polymorphic transposable elements on transcription in lymphoblastoid cell lines from public data. *BMC Bioinformatics, 20*(Suppl 9), 495. doi:10.1186/s12859-019-3113-x

Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform, 3*(1), lqab019. doi:10.1093/nargab/lqab019

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81. doi:10.1038/nature15394

Terry, D. M., & Devine, S. E. (2019). Aberrantly High Levels of Somatic LINE-1 Expression and Retrotransposition in Human Neurological Disorders. *Front Genet, 10*, 1244. doi:10.3389/fgene.2019.01244

Treiber, C. D., & Waddell, S. (2017). Resolving the prevalence of somatic transposition in Drosophila. *Elife, 6*. doi:10.7554/eLife.28297

Tsetsos, F., Yu, D., Sul, J. H., Huang, A. Y., Illmann, C., Osiecki, L., . . . Paschou, P. (2020). Synaptic processes and immune-related pathways implicated in Tourette Syndrome. *medRxiv*, 2020.2004.2024.20047845. doi:10.1101/2020.04.24.20047845

Wang, L., Norris, E. T., & Jordan, I. K. (2017). Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. *Front Microbiol, 8*, 1418. doi:10.3389/fmicb.2017.01418

Wang, Z., Hong, Y., Zou, L., Zhong, R., Zhu, B., Shen, N., . . . Miao, X. (2014). Reelin gene variants and risk of autism spectrum disorders: an integrated meta-analysis. *Am J Med Genet B Neuropsychiatr Genet, 165B*(2), 192-200. doi:10.1002/ajmg.b.32222

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., . . . di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat Commun, 10*(1), 5241. doi:10.1038/s41467-019-13212-3

Xie, Y., Mates, L., Ivics, Z., Izsvak, Z., Martin, S. L., & An, W. (2013). Cell division promotes efficient retrotransposition in a stable L1 reporter cell line. *Mob DNA, 4*(1), 10. doi:10.1186/1759-8753-4-10

Xing, D., Tan, L., Chang, C. H., Li, H., & Xie, X. S. (2021). Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc Natl Acad Sci U S A, 118*(8). doi:10.1073/pnas.2013106118

Zhang, Y., Romanish, M. T., & Mager, D. L. (2011). Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput Biol, 7*(5), e1002046. doi:10.1371/journal.pcbi.1002046

Zhao, B., Wu, Q., Ye, A. Y., Guo, J., Zheng, X., Yang, X., . . . Huang, A. Y. (2019). Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genet, 15*(4), e1008043. doi:10.1371/journal.pgen.1008043

Zhao, X., Collins, R. L., Lee, W. P., Weber, A. M., Jun, Y., Zhu, Q., . . . Talkowski, M. E. (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am J Hum Genet*. doi:10.1016/j.ajhg.2021.03.014

Zhou, W., Emery, S. B., Flasch, D. A., Wang, Y., Kwan, K. Y., Kidd, J. M., . . . Mills, R. E. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res, 48*(3), 1146-1163. doi:10.1093/nar/gkz1173

Zhu, X., Zhou, B., Pattni, R., Gleason, K., Tan, C., Kalinowski, A., . . . Urban, A. E. (2021). Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia. *Nat Neurosci, 24*(2), 186-196. doi:10.1038/s41593-020-00767-4

Appendix

Supplementary material for Chapter 2

Supplementary Table 2.1: *De novo* insertions in ASD and controls.

Family	Phenotype	TE Type	Chr	Start	End	Gene	pLI	SFARI 2019 score	Genic Region	Strand
11859	ASD	Alu	chr1	114750025	114750040	CSDE1	0.99995156	NA	exon	-
12548	ASD	Alu	chr17	60473237	60473252	APPBP2	0.99907128	NA	intron	-
12748	ASD	Alu	chr12	79291451	79291465	SYT1	0.83718036	Syndromic	intron	-
13931	ASD	Alu	chr15	31732090	31732104	OTUD7A	0.97459544	3	intron	+
13107	ASD	Alu	chr16	52468265	52468277	TOX3	0.99433911	NA	intron	-
13195	ASD	Alu	chr9	76506036	76506050	GCNT1	0.06657664	NA	exon	-
14315	ASD	Alu	chr7	28013356	28013370	JAZF1	0.95830238	NA	intron	+
14565	ASD	Alu	chr13	41129294	41129307	KBTBD6	0.93481936	NA	exon	+
11305	ASD	Alu	chr16	89907224	89907238	TCF25	0.05177828	NA	intron	-
11190	ASD	Alu	chr4	4481646	4481656	STX18-IT1	1.17E-06	NA	exon	-
11128	ASD	Alu	chr2	41757966	41757980	NA	NA	NA	NA	+
11236	ASD	Alu	chr15	62869456	62869473	NA	NA	NA	NA	+
11432	ASD	Alu	chr20	56789176	56789188	NA	NA	NA	NA	+
11484	ASD	Alu	chr4	160407103	160407118	NA	NA	NA	NA	-
11484	ASD	Alu	chr3	104331631	104331640	NA	NA	NA	NA	+
11551	ASD	Alu	chr4	168130363	168130379	ANXA10	1.28E-08	NA	intron	-
11563	ASD	Alu	chr2	126224744	126224757	NA	NA	NA	NA	+
11565	ASD	Alu	chr18	61615604	61615618	NA	NA	NA	NA	-
11917	ASD	Alu	chr6	76111910	76111924	NA	NA	NA	NA	-
11946	ASD	Alu	chr18	27242467	27242482	NA	NA	NA	NA	-
12158	ASD	Alu	chr8	114736589	114736606	NA	NA	NA	NA	-
12357	ASD	Alu	chr6	69858208	69858216	NA	NA	NA	NA	-
12403	ASD	Alu	chr8	62690037	62690057	NKAIN3	0.01017944	NA	intron	-
12434	ASD	Alu	chr20	58475382	58475395	APCDD1L	0.01755922	NA	intron	+
12817	ASD	Alu	chr5	73156900	73156912	NA	NA	NA	NA	+
13102	ASD	Alu	chr5	51652114	51652128	NA	NA	NA	NA	-
13233	ASD	Alu	chr12	749189	749200	NA	NA	NA	NA	+
13287	ASD	Alu	chr3	96330136	96330152	NA	NA	NA	NA	+
13310	ASD	Alu	chr5	12875343	12875360	NA	NA	NA	NA	+
13370	ASD	Alu	chr10	56385473	56385479	NA	NA	NA	NA	-
13718	ASD	Alu	chr21	20850708	20850717	NA	NA	NA	NA	+
13775	ASD	Alu	chr6	45015716	45015732	SUPT3H	0.00010991	NA	intron	-
13783	ASD	Alu	chr12	112755126	112755139	RPH3A	0.73860348	NA	intron	-

Supplementary Table 2.1 (Continued)

Family	Phenotype	TE Type	Chr	Start	End	Gene	pLI	SFARI 2019 score	Genic Region	Strand
13920	ASD	Alu	chr17	40109779	40109794	NA	NA	NA	NA	-
13968	ASD	Alu	chr1	73079560	73079574	NA	NA	NA	NA	+
13984	ASD	Alu	chr14	24561452	24561466	NA	NA	NA	NA	+
14006	ASD	Alu	chr7	91631371	91631387	NA	NA	NA	NA	-
14075	ASD	Alu	chr15	62571150	62571164	NA	NA	NA	NA	+
14159	ASD	Alu	chr4	146284607	146284621	SLC10A7	0.02880743	NA	intron	-
14161	ASD	Alu	chr13	89961196	89961212	NA	NA	NA	NA	-
14174	ASD	Alu	chr5	5380825	5380844	NA	NA	NA	NA	+
14244	ASD	Alu	chr1	51491363	51491378	EPS15	0.24170707	NA	intron	+
14397	ASD	Alu	chr11	26046470	26046479	NA	NA	NA	NA	+
14417	ASD	Alu	chr7	109826976	109826990	NA	NA	NA	NA	+
14532	ASD	Alu	chr20	55919775	55919790	NA	NA	NA	NA	-
11432	ASD	Alu	chr5	85282625	85282637	NA	NA	NA	NA	+
11709	ASD	Alu	chr1	154026720	154026735	NUP210L	0.14464062	NA	intron	+
11944	ASD	Alu	chr19	3734541	3734555	TJP3	2.20E-21	NA	intron	-
12058	ASD	Alu	chr12	83316693	83316705	NA	NA	NA	NA	+
12297	ASD	Alu	chr3	137876063	137876078	NA	NA	NA	NA	+
13522	ASD	Alu	chr8	93053572	93053589	NA	NA	NA	NA	+
13555	ASD	Alu	chrX	144903787	144903801	NA	NA	NA	NA	+
13673	ASD	Alu	chr10	36524093	36524105	NA	NA	NA	NA	+
13425	ASD	Alu	chr6	7372046	7372060	CAGE1	2.62E-06	NA	intron	-
13190	ASD	Alu	chr9	33406406	33406421	NA	NA	NA	NA	+
13390	ASD	Alu	chr12	119359462	119359478	CCDC60	8.62E-06	NA	intron	+
11227	ASD	Alu	chr4	35603676	35603688	NA	NA	NA	NA	+
12764	ASD	Alu	chr13	32559157	32559169	NA	NA	NA	NA	+
11634	ASD	Alu	chr20	40059902	40059920	NA	NA	NA	NA	-
14441	ASD	Alu	chr14	93148136	93148147	NA	NA	NA	NA	+
12750	ASD	Alu	chr17	75087065	75087079	NA	NA	NA	NA	+
14545	Control	Alu	chr19	19329357	19329372	MAU2	0.99971776	NA	intron	+
11521	Control	Alu	chr4	150155536	150155550	DCLK2	0.9663101	NA	intron	+
11121	Control	Alu	chr9	97192772	97192779	ANKRD18CP	NA	NA	intron	-
11218	Control	Alu	chr13	61335389	61335403	NA	NA	NA	NA	+
11218	Control	Alu	chr12	107035018	107035038	CRY1	4.24E-07	NA	intron	+
11245	Control	Alu	chr18	76688657	76688666	NA	NA	NA	NA	-
11265	Control	Alu	chr17	73322491	73322505	NA	NA	NA	NA	-
11300	Control	Alu	chr6	65570284	65570290	EYS	8.21E-09	NA	intron	+

Supplementary Table 2.1 (Continued)

Family	Phenotype	TE Type	Chr	Start	End	Gene	pLI	SFARI 2019 score	Genic Region	Strand
11414	Control	Alu	chr12	126100649	126100663	LINC02359	NA	NA	intron	-
11571	Control	Alu	chr2	43584185	43584194	THADA	1.24E-28	NA	intron	-
11634	Control	Alu	chr1	54385944	54385960	SSBP3	0.9997562	NA	intron	-
12086	Control	Alu	chr5	74486390	74486404	LINC01331	NA	NA	intron	+
12540	Control	Alu	chr19	30950093	30950109	NA	NA	NA	NA	-
12605	Control	Alu	chr8	9555306	9555321	NA	NA	NA	NA	+
12686	Control	Alu	chr8	142074057	142074073	NA	NA	NA	NA	+
12748	Control	Alu	chr1	104639289	104639309	NA	NA	NA	NA	-
12864	Control	Alu	chr22	30344096	30344107	SF3A1	0.9940754	NA	intron	-
12891	Control	Alu	chr11	18117066	18117076	NA	NA	NA	NA	+
13000	Control	Alu	chr2	196155020	196155036	STK17B	0.08982575	NA	intron	-
13077	Control	Alu	chr4	81049933	81049945	BMP3	0.11024112	NA	intron	-
13080	Control	Alu	chr3	188327270	188327280	LPP	0.58239991	NA	intron	+
13083	Control	Alu	chr15	54868108	54868123	NA	NA	NA	NA	-
13180	Control	Alu	chr9	36113758	36113772	RECK	0.0001968	NA	intron	+
13266	Control	Alu	chr9	120958711	120958728	C5	5.61E-11	NA	intron	-
13298	Control	Alu	chr13	110085913	110085924	NA	NA	NA	NA	+
13403	Control	Alu	chr3	37558360	37558376	ITGA9	0.09619175	NA	intron	+
13512	Control	Alu	chr5	145756636	145756653	PRELID2	1.73E-05	NA	exon	+
13630	Control	Alu	chr11	59862602	59862616	TCN1	3.86E-16	NA	intron	-
13814	Control	Alu	chr9	38009756	-1	SHB	0.74418363	NA	intron	+
13931	Control	Alu	chr11	87390981	87390998	NA	NA	NA	NA	+
13972	Control	Alu	chr6	16694693	16694705	ATXN1	0.39866115	NA	intron	-
14075	Control	Alu	chr10	94516242	94516251	TBC1D12	0.00264365	NA	intron	-
14234	Control	Alu	chr13	19510615	19510628	TPTE2	3.69E-08	NA	intron	-
14254	Control	Alu	chr12	41700293	41700307	NA	NA	NA	NA	-
14279	Control	Alu	chr2	15560601	15560617	NBAS	1.89E-25	NA	intron	-
14350	Control	Alu	chr18	51610860	51610874	NA	NA	NA	NA	-
14384	Control	Alu	chr13	30127030	30127045	NA	NA	NA	NA	+
14522	Control	Alu	chr1	190720800	190720815	LINC01720	NA	NA	intron	-
14560	Control	Alu	chr16	13153070	13153086	SHISA9	NA	NA	intron	+
14579	Control	Alu	chr1	69244068	69244085	NA	NA	NA	NA	+
14600	Control	Alu	chr7	79763348	79763360	NA	NA	NA	NA	-
14652	Control	Alu	chr1	219079728	219079743	MIR548F3	NA	NA	intron	-
14660	Control	Alu	chr5	62637898	62637912	NA	NA	NA	NA	-
14697	Control	Alu	chr20	45045736	45045751	STK4	0.11831192	NA	intron	-
11397	Control	Alu	chr3	193261418	193261430	PLAAT1	0.00071783	NA	intron	-

Supplementary Table 2.1 (Continued)

Family	Phenotype	TE Type	Chr	Start	End	Gene	pLI	SFARI 2019 score	Genic Region	Strand
11499	Control	Alu	chr3	142532612	142532628	ATR	0.70129783	NA	intron	+
12630	Control	Alu	chr17	62247762	62247791	NA	NA	NA	NA	-
14075	Control	Alu	chr7	70854677	70854702	NA	NA	NA	NA	+
14428	Control	Alu	chr4	120548410	120548418	NA	NA	NA	NA	-
14533	Control	Alu	chr16	47757456	47757471	NA	NA	NA	NA	+
13732	Control	Alu	chr14	36160087	36160100	PTCSC3	NA	NA	intron	-
13095	Control	Alu	chr21	25284467	25284475	NA	NA	NA	NA	+
14682	Control	Alu	chr12	129724850	129724860	TMEM132D	0.99901039	NA	intron	-
11190	Control	Alu	chr3	139467257	139467271	RBP2	0.43355451	NA	intron	+
14091	ASD	Alu	chr12	14555244	14555254	PLBD1	9.67E-13	NA	intron	+
11196	ASD	L1	chr3	9093236	9093252	SRGAP3	0.99999775	4	intron	-
13684	ASD	L1	chr5	45576963	45576976	HCN1	0.9531385	Syndromic	intron	-
14080	ASD	L1	chr1	58103235	58103249	DAB1	0.98140857	5	intron	+
14282	ASD	L1	chr1	97416060	97416075	DPYD	4.19E-09	4	intron	+
13617	ASD	L1	chr21	25428297	25428314	LINC00158	NA	NA	intron	-
12925	ASD	L1	chr1	166451358	166451362	NA	NA	NA	NA	-
13043	ASD	L1	chr8	17458428	17458446	NA	NA	NA	NA	+
11234	ASD	L1	chr1	120063769	120063777	NOTCH2	0.99999999	NA	intron	-
12778	ASD	L1	chr6	63962148	63962163	EYS	8.21E-09	NA	intron	+
11420	ASD	L1	chr1	4227772	4227780	NA	NA	NA	NA	-
11445	ASD	L1	chr8	6366924	6366940	NA	NA	NA	NA	-
13451	ASD	L1	chr2	115698283	115698289	DPP10	0.99998929	3	intron	+
11671	Control	L1	chr9	74570049	74570058	RORB	0.9985517	NA	intron	-
12879	Control	L1	chr6	93400531	93400545	EPHA7	0.98775937	NA	intron	-
12162	Control	L1	chr3	69370702	69370717	FRMD4B	0.80701371	NA	intron	-
11329	Control	L1	chrX	67156505	67156522	NA	NA	NA	NA	+
11378	Control	L1	chr20	1086525	1086539	NA	NA	NA	NA	+
13869	Control	L1	chr11	4736054	4736072	NA	NA	NA	NA	+
12902	Control	L1	chr14	35508860	35508871	NA	NA	NA	NA	-
14459	Control	L1	chr2	194830322	194830336	NA	NA	NA	NA	+
14679	Control	L1	chr6	8063530	8063545	BLOC1S5	1.60E-05	NA	intron	-
12483	Control	L1	chr2	39295438	39295454	MAP4K3	0.80470105	NA	intron	-
14404	ASD	SVA	chr11	123585563	123585574	GRAMD1B	0.98485077	NA	intron	-
14523	ASD	SVA	chr17	37343893	-1	ACACA	1	NA	intron	+
13316	ASD	SVA	chr16	-1	14473268	PARN	0.39	NA	intron	-
12891	ASD	SVA	chr7	112711571	112711590	NA	NA	NA	NA	-

Supplementary Table 2.1 (Continued)

Family	Phenotype	TE Type	Chr	Start	End	Gene	pLI	SFARI 2019 score	Genic Region	Strand
13633	ASD	SVA	chr2	44102836	-1	NA	NA	NA	NA	+
13924	ASD	SVA	chrX	44748345	44748352	NA	NA	NA	NA	-
14320	ASD	SVA	chr5	176538776	176538792	NA	NA	NA	NA	-
11350	ASD	SVA	chr3	127970669	127970674	KBTBD12	9.96E-13	NA	intron	+
13153	ASD	SVA	chr17	7270387	-1	NA	NA	NA	NA	+
11073	Control	SVA	chr9	134164721	-1	RNU6ATAC	NA	NA	up	+
14560	Control	SVA	chr6	33635903	33635921	ITPR3	4.98E-12	NA	intron	-
12351	Control	SVA	chr12	25063225	25063234	LRMP	0.02091271	NA	intron	-
11010	Control	SVA	chrX	115483428	-1	NA	NA	NA	NA	+
11057	Control	SVA	chr6	7096099	7096114	NA	NA	NA	NA	-
11501	Control	SVA	chr13	36921669	36921686	NA	NA	NA	NA	+
12630	Control	SVA	chr19	14367280	14367297	NA	NA	NA	NA	-
13349	Control	SVA	chr12	55249219	-1	NA	NA	NA	NA	+

Supplementary material for Chapter 3

Supplementary Table 3.1: Case information and single cells analyzed.

Case ID	Sex	Age (years)	Diagnosis	PFC neurons		DG neurons		PFC NeuN - (glia)	PFC oligodendrocytes	Heart 2n	Heart 4n	CA1 neurons
				xTea	*memTea*	*xTea*	*memTea*	*xTea*	*xTea*	*xTea*	*xTea*	*xTea*
Infant												
1278	M	0.4	Normal	9	9	-	-	5	5	4	4	-
5817	M	0.6	Normal	4	4	-	-	-	-	-	-	-
Adolescent												
4638	F	15.1	Normal	10	9	-	-	-	-	-	-	-
1465	M	17.5	Normal	22	22	-	-	-	-	-	-	-
5532	M	18.4	Normal	4	4	5	5	-	-	-	-	-
5559	F	19.8	Normal	4	5	3	5	-	-	-	-	-
Adult												
4643	F	42.2	Normal	10	10	-	-	-	-	-	-	-
5087	M	44.9	Normal	4	4	5	5	-	-	-	-	-
936	F	49.2	Normal	3	3	-	-	-	-	-	4	-
Aged												
5840	M	75.3	Normal	3	3	3	3	-	-	-	-	-
5219	F	77	Normal	4	4	-	-	-	-	-	-	-
5171	M	79.2	Normal	9	4	-	-	5	5			-
5511	F	80.2	Normal	3	3	-	-	-	-	-	-	-
5657	M	82	Normal	5	5	5	4	5	5	4	4	-
5823	F	82.7	Normal	2	3	3	3	-	-	-	-	-
Cockayne syndrome												
1762	F	4.4	CS (*CSB*)	6	6	-	-	-	-	-	-	-
1124	F	4.7	CS (*CSB*)	3	3	-	-	-	-	-	-	-
1286	M	5.8	CS (*CSB*)	3	4	-	-	-	-	-	-	-
580	F	8.4	CS (*CSB*)	2	4	-	-	-	-	-	-	-
5105	M	8.7	CS (*CSB*)	6	6	-	-	-	-	-	-	-
682	M	32.8	CS (*CSB*)	4	4	-	-	-	-	-	-	-
Xeroderma pigmentosum												
5379	F	24	XP (*XPA*)	4	6	-	-	-	-	-	-	-
5316	F	44.5	XP (*XPA*)	3	3	-	-	-	-	-	-	-
5416	F	46	XP (*XPD*)	6	6	-	-	-	-	-	-	-
Alzheimer's disease												
MA1647	F	80	AD (Braak VI)	5	-	-	-	-	-	-	-	-
MA1828	F	91	AD (Braak VI)	5	-	-	-	-	-	-	-	5
UMB5222	F	80	AD (Braak VI)	-	-	-	-	-	-	-	-	5
Ataxia telangiectasia												
1459	F	19.9	AT	3	-	-	-	-	-	-	-	-
1485	F	24.9	AT	7	-	-	-	-	-	-	-	-
Amyotrophic lateral sclerosis												
1653	M	91	ALS	5	-	-	-	-	-	-	-	-
1798	M	66	ALS	5	-	-	-	-	-	-	-	-
1844	NA	51	ALS	5	-	-	-	-	-	-	-	-
2091	NA	53	ALS	5	-	-	-	-	-	-	-	-
Total				173	134	24	25	15	15	8	12	10

Total *xTea*	Total *memTea*	Total cells
257	159	266