Exploring Brain Somatic Mosaicism Through Robust Copy Number Variation Detection in Single Cell Whole Genome Sequencing Data

William Douglas Chronister

La Grange, Illinois

Bachelor of Science in Biological Sciences, University of Notre Dame
Notre Dame, Indiana, 2014

A Dissertation Presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

University of Virginia

December 2019

Dr. Michael McConnell

Dr. Stefan Bekiranov

Dr. Xiaowei Lu

Dr. Aakrosh Ratan

Abstract

The presence of intra-individual genomic diversity in neurons and other brain cells, known as brain somatic mosaicism, has gone largely unnoticed and unappreciated historically. But growing evidence accumulated in recent years shows that neurons are highly genomically heterogeneous as a consequence of mutations ranging in scale from single nucleotide variants to whole chromosome copy number variations (CNVs). Mosaicism can disrupt particular cell signaling pathways to cause disease but is also common in healthy individuals, wherein its role is currently unknown. Given the long lifespan of neurons and their critical role in neural circuits and networks, it is probable that mosaic mutations affect neurodevelopment and brain function in ways that are not yet understood. Instrumental in identifying these genomic variations have been single cell technology and next generation sequencing, which allow for high resolution, high throughput measurement of individual genomes.

To quantify the extent of mosaicism in healthy brains and explore its biological implications, we developed a robust analysis pipeline for CNV detection and analyzed single cell whole genome sequencing (scWGS) data from over 1200 human brain cells from 15 individuals. Notably, our results showed a significant anticorrelation between the percent of an individual's neurons containing CNVs and age. We also sequenced over 800 mouse neurons from 10 mice to examine mosaic CNVs in relationship to aging and perturbations of various genes of interest. In brief, our work demonstrates the value of large datasets in constructing a rigorous CNV detection pipeline, documents the prevalence and characteristics of mosaic CNVs in the brain, and contributes insights to our understanding of the causes and consequences of somatic mosaicism in the brain.

Dedication

"The God of the Bible is also the God of the genome. He can be worshipped in the cathedral or in the laboratory. His creation is majestic, awesome, intricate and beautiful – and it cannot be at war with itself."

Francis S. Collins

I dedicate this dissertation to my wife, Briana, who selflessly supported my decision to pursue a Ph.D. over the past 5 years and supplied unfailing optimism every step of the way; to my parents, Laura and Joe, who raised me in a loving home and always encouraged me to put my abilities to good use; to my brothers, Andrew and Stephen, whom I have always striven to emulate in academics and in life; to my *suegros*, Alfonso and Vianey, who never complained that I was the reason their only child was living 2,500 miles away from them; and to the rest of my friends and family, whose names are too many to mention, whose encouragement gave me the energy to reach the finish line.

Acknowledgments

Thank you to everyone who contributed to the research presented in this dissertation. Your efforts are what made this work possible.

Special thanks are due to my principal investigators and mentors, Mike McConnell and Stefan Bekiranov: Your instruction and guidance has made me a vastly better scientist than I was when I arrived at the University of Virginia in 2014. I will always be grateful that I was able to learn from two exceptional biologists who complemented each other's strengths so perfectly.

Thank you to Aakrosh Ratan and Xiaowei Lu, who served as committee members and provided valuable feedback and advice. Your input improved the quality of my research and made me consider my work in a broader context.

Thank you to Ian Burbulis, Matt Wolpert, Mark Haakenson, and Maggie Wierman for your tireless efforts in optimizing and carrying out protocols to produce high quality single cell sequencing data. Your work saved me from hundreds of hours in the lab and allowed me to focus entirely on computational analysis.

Thank you to undergraduate researchers Inusah Diallo, Aiden Smith, Ahnaf Khan, and Graham Smith; computational analysis has a steep learning curve, but you all rose to the occasion and made important contributions.

Thank you to Todd Stukenberg and all those involved with the Cell and Molecular Biology Training Grant. It was a privilege to learn from and interact with the outstanding faculty and student members.

Thank you to Joel Hockensmith and Debbie Sites. Your unfailing willingness to help with any problem big or small over the past 5 years provided tremendous stress relief.

Thank you to the BIMS program, the Department of Biochemistry and Molecular Genetics, and all associated faculty, administrators, and students. The University of Virginia is home to numerous outstanding research labs, but it also fosters a friendly, collegial environment that allows scientists to flourish. I hope that continues to be the case long into the future.

Thank you to all friends, labmates, and colleagues who I got to know inside and outside of lab. I hope our paths cross again soon.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AD	Alzheimer's disease
BIC	Bayesian information criterion
BSMN	Brain Somatic Mosaicism Network
CBS	circular binary segmentation
CGH	comparative genomic hybridization
cKO	conditional knockout
CNV	copy number variation
DOP-PCR	degenerate oligonucleotide-primed polymerase chain reaction
DSB	double-strand break
FACS	fluorescence-activated cell sorting
FANS	fluorescence-activated nucleus sorting
FDR	false discovery rate
FISH	fluorescence in situ hybridization
GO	gene ontology
HME	hemimegalencephaly
HMM	hidden Markov model
IGV	Integrative Genomics Viewer
kb	kilobase
KO	knockout
LINE1	long interspersed nuclear element 1
LTR	long terminal repeat
M-FISH	multi-fluor fluorescence in situ hybridization
MAD	median absolute deviation
MALBAC	multiple annealing and looping-based amplification cycles
MAPD	median absolute pairwise difference
Mb	megabase
MDA	multiple displacement amplification
NFT	neurofibrillary tangle
NGS	next-generation sequencing
NPC	neural progenitor cell
PCR	polymerase chain reaction
PMI	post-mortem interval
ROC	receiver operating characteristic
scWGS	single cell whole genome sequencing
SINE	short interspersed nuclear element
SKY	spectral karyotyping
SNV	single nucleotide variant
SV	structural variant
VS	variability score
WGA	whole genome amplification
WGS	whole genome sequencing
WT	wild-type

Chapter 1
Introduction

1.1 Mosaicism: Background and basics

Multicellular organisms are multi-genomic organisms. Each nucleated cell contains its own copy of the individual's DNA and protects it from damage using multiple mechanisms of DNA repair. Yet, inevitably, insults from internal and external sources mutate the genomes of single cells. The resulting coexistence of multiple distinct genotypes within a monozygotic individual is known as mosaicism.

The mutations that make up mosaicism can be genome alterations as minor as a single nucleotide substitution, as major as the loss or gain of multiple chromosomes, or any of a host of additional mutation classes such as small insertions and deletions, megabase-scale copy number variations (CNVs), inversions, and translocations. These variants can arise from a number of causes, such as nondisjunction during cell division, faulty excision and replacement of mismatched bases, misrepair of double-strand breaks, ionizing radiation, reactive oxygen species, or chemical exposure (Campbell et al., 2015; Carvalho and Lupski, 2016; McConnell et al., 2017). Variation can also arise due to retrotransposon activity by LINE-1 (Evrony et al., 2012). Some cell types undergo a controlled recombination process to produce cell-to-cell diversity, such as the V(D)J recombination employed by lymphocytes to generate the immune system's diverse array of antibodies and receptors. This class of genomic modification is also considered a form of mosaicism, though our research generally does not involve cell types that carry out these types of rearrangements.

Just as mutations can take a range of forms, these mutations can have a range of outcomes for the organism as a whole, from favorable to benign to pathological, depending on a number of contextual factors; these include the class of mutation, the

timing of the mutation in development, the size of the mutated region, the particular genomic location that is mutated (including any affected genes), the number of cells carrying the mutation, and the tissue affected by the mutation. Mosaicism cannot be categorized as an inherently beneficial or harmful phenomenon because its consequences vary from case to case.

Mosaicism can be divided into two subtypes: Germline mosaicism encompasses any mutation that occurs in an individual's gametes. Historically, diversity in sperm and egg cells has been recognized as providing the basis for natural selection and genome evolution (Zhang and Vijg, 2018). The rate of germline mosaicism can be estimated by sequencing the genomes of parents and children and detecting *de novo* variants in the child; recent studies estimate this rate to be $1.0\text{-}1.2 \times 10^{-8}$ per nucleotide per generation in humans (Conrad et al., 2011; Kong et al., 2012).

Meanwhile, somatic mosaicism, consisting of mutations in non-reproductive cells, has no direct evolutionary role, and has been little studied in comparison to germline mosaicism. For much of the history of biological research, it was generally assumed that the genome of every healthy non-gamete in a multicellular organism was essentially identical. At the same time, researchers were aware of various examples of apparent somatic mutations in disease, in particular skin abnormalities and cancer, and speculated that even healthy cells could contain somatic mutations (Frank, 2014; Hall, 1988). Until recent years, however, researchers lacked the technology and techniques to analyze single genomes deeply enough to confirm the presence of somatic mosaicism in healthy cells.

1.2 A brief history of genomics methods

Dating back to the 19th century, cytogeneticists characterized condensed chromosomes using microscopes, but the methods they employed often led to inaccurate reports; remarkably, it was not until 1956 that the correct number of human chromosomes was published as 46, thanks to an improved protocol for preparing and viewing metaphase spreads (Ferguson-Smith, 2015; Tjio and Levan, 1956). A few years later, in 1959, the correct count of human chromosomes enabled the discovery of trisomy 21 in Down syndrome, the first evidence of a constitutional aneuploidy in humans (Lejeune et al., 1959). The following year, it was found that lymphocyte cultures derived from peripheral blood samples could produce high resolution chromosomes, leading to widespread adoption of the practice in clinical laboratories (Ferguson-Smith, 2015; Moorhead et al., 1960). Further progress was made by the development of quinacrine and Giemsa staining, which could be utilized to distinguish specific chromosomes from each other based on patterns of banding (Caspersson et al., 1970; Speicher and Carter, 2005). These methods were crucial to the discovery of the reciprocal translocation between chromosomes 9 and 22 seen commonly in chronic myelogenous leukemia, known as the Philadelphia chromosome (Rowley, 1973).

Another major leap came with the arrival of fluorescence *in situ* hybridization in 1980, which utilized the binding specificity of a complementary RNA-DNA hybrid along with a coupled fluorochrome to illuminate a specific genomic locus (Bauman et al., 1980). A key advantage of this technique was the ability to detect specific loci in cells at any stage of the cell cycle, including post-mitotic cells, such as neurons. Subsequently, multi-fluor FISH (M-FISH) and spectral karyotyping (SKY) were introduced to enable

each human chromosome to be labeled by unique combinations of fluorescent probes simultaneously, allowing for greater efficiency in karyotyping metaphase spreads (Schrock et al., 1996; Speicher et al., 1996). Meanwhile, comparative genomic hybridization (CGH), improved upon by microarray-based versions (array CGH), was among the first methods capable of detecting sub-chromosomal copy number variations with high precision (Kallioniemi et al., 1992; Pinkel et al., 1998; Solinas-Toldo et al., 1997). Finally, in 2011, the first single cell whole genome sequencing (scWGS) study was published, demonstrating that individual nuclear genomes isolated from cancerous tumors could be amplified, sequenced using next generation sequencing (NGS), and analyzed for CNVs (Navin et al., 2011). This breakthrough resulted in a wave of publications using single cell techniques to explore individual genomes in oocytes, the skin, and the brain, among other tissues (Cai et al., 2014; Hou et al., 2013; Knouse et al., 2014; McConnell et al., 2013). To this day, scWGS remains the most comprehensive method for studying mosaicism, and new technologies are constantly being developed to improve and streamline the process of single cell genome analysis.

1.3 Mosaicism in the brain

Based on the evidence reported to date, it seems likely that somatic mosaicism exists at some level in every tissue in the body (Zhang and Vijg, 2018). Nonetheless, the brain is one of the most important repositories of mosaic variation for two chief reasons: first, neurons are among the longest-lived cells in the body, commonly arising during neurogenesis and surviving until death. This considerable lifespan means that abnormal neurons are not systematically replaced as they would be in other tissues; rather, the

effects of neuronal mutations are present throughout the life of the individual. Secondly, neurons function through complex neural networks, and the presence of even a small percentage of abnormal neurons can cause large-scale effects in brain function (McConnell et al., 2017).

1.3.1 Brain somatic mosaicism in disease

Similar to how visible skin abnormalities sparked interest in somatic mosaicism generally, reports of brain abnormalities have driven exploration of brain somatic mosaicism. Hemimegalencephaly (HME), a neurodevelopmental disorder resulting in the overgrowth of one hemisphere of the brain, is a classic example. An exome sequencing study of HME patients showed somatic mutations in the *PIK3CA*, *AKT3*, and *MTOR* genes in 6 of 20 individuals (30%) (Lee et al., 2012). Within the affected hemispheres of HME patients, the somatic mutations were found in between 8 and 40% of sequenced alleles. Further cases of disease-causing brain somatic mosaicism have been documented in focal cortical dysplasia, tuberous sclerosis, and a variety of other brain diseases (Jansen et al., 2015; Lim et al., 2015; McConnell et al., 2017; van Slegtenhorst et al., 1997). Collectively, these findings support the potential of somatic mosaicism to cause structural brain malformations.

The role of mosaicism in neuropsychiatric disease is also being explored, and has yielded some intriguing results. *De novo* mutations are a known contributor to risk of autism spectrum disorder, and *de novo* somatic mutations are enriched in autistic probands (Freed and Pevsner, 2016). In schizophrenia, increased levels of chromosome 1 loss and gain were reported in the brains of affected individuals compared to controls

(Yurov et al., 2008). Studies have also reported increased LINE-1 retrotransposition in neurons of individuals with schizophrenia and Rett syndrome (Bundo et al., 2014; Muotri et al., 2010). These studies have been motivated not only by the interest in finding causative genetic factors for each disease but also the observation that monozygotic twins show significant discordance in all psychiatric disorders (Insel, 2014). Somatic mutations, therefore, would serve as a plausible explanation for why a psychiatric disease arises in one twin but not the other, or, more generally, why it arises in some individuals with no known genetic risk loci, and not others.

1.3.2 Brain somatic mosaicism in healthy individuals

While disease relevance is an important motivation for studying brain somatic mosaicism, the discovery of mosaicism in normal (neurotypical) brains sparked tremendous curiosity about the role of somatic mutation in the brain and paved the way for many more studies seeking to characterize and quantify the genomic diversity of the brain. Rehen et al. (2001) used SKY to detect aneuploidy in neuroblasts of the developing mouse brain and found that ~33% of cells carried more or less than 46 chromosomes. A few years later, similar evidence of aneuploidy was found in the human brain using FISH (Pack et al., 2005; Rehen et al., 2005). These results were surprising given what is known about constitutive (whole-organism) aneuploidy, namely that the vast majority of whole-chromosome gains and losses are lethal. Monosomies are always lethal except in Turner's syndrome (X0), and trisomies of chromosomes 13, 18, and 21 are the only constitutive autosomal aneuploidies seen in humans surviving until birth. A few

additional aneuploidies can be tolerated in mosaic contexts, but in most cases, aneuploidy is incompatible with human life (D'Gama and Walsh, 2018).

Reports in the years that followed established the presence of other somatic mutations in the brain, including evidence that LINE-1 retrotransposition is active in neural progenitors and neurons (Evrony et al., 2012; Evrony et al., 2016; Muotri et al., 2005; Upton et al., 2015); single nucleotide variants (SNVs) are abundant in single neurons and neural progenitors (Bae et al., 2018; Lodato et al., 2018; Lodato et al., 2015); and megabase-scale CNVs are common in neurons (McConnell 2013, Cai 2014, Knouse 2016).

These reports were not always in agreement with prior results; for instance, the estimate of LINE-1 insertion rate in hippocampal neurons given by Upton et al. (2015) was subsequently re-evaluated and lowered by Evrony et al. (2016). Meanwhile, the high rates of aneuploidy observed in early SKY and FISH experiments (Pack et al., 2005; Rehen et al., 2001; Rehen et al., 2005) were not confirmed by scWGS studies (Cai et al., 2014; Knouse et al., 2014; McConnell et al., 2013; van den Bos et al., 2016), perhaps due to sub-chromosomal CNVs that were misinterpreted as aneuploidies in SKY and FISH assays. Even estimates of the percentages of neurons containing CNVs between comparable scWGS studies has varied from below 10% to over 40% (Knouse et al., 2016; McConnell et al., 2013). Nonetheless, the past two decades of research has made it clear that the brain harbors many flavors of genomic diversity in addition to the remarkable morphological and functional diversity of its neurons. But a key question remains unanswered: What is the role of mosaicism in the normal brain?

1.3.3 The role of mosaicism

Germline mosaicism, consisting of mutations that can be passed on to offspring, is understood to be a mechanism for generating genomic diversity in the population, providing the "substrate" that natural selection acts upon to drive evolution (Zhang and Vijg, 2018). Meanwhile, somatic mosaicism is not passed on to future generations and, in many cases, such as those discussed above, has deleterious effects. Given the costs of genome maintenance and risks of disorder, why have somatic mosaicism at all?

This remains a major unresolved question in the field of somatic mosaicism. The answer could be as simple (and bleak) as humans are not evolved to live forever, and somatic mutations accumulate in order to cause aging and, eventually, death (Vijg, 2014). Conversely, there are clues that point to somatic mutations serving a more complicated role, particularly in the brain.

DNA double strand breaks (DSBs), which can lead to CNVs, occur frequently in neural progenitor cells during neurogenesis (Wei et al., 2016). In mice, it has been shown that this process requires the presence of DNA repair proteins DNA Ligase IV (Lig4) and Xrcc4, otherwise resulting in embryonic lethality (Alt and Schwer, 2018; Barnes et al., 1998; Frank et al., 2000; Gao et al., 1998). These findings indicate that DNA repair in response to damage is an essential part of normal brain development and suggest that CNV formation may be a normal outcome of DSB repair. If so, it is plausible that mosaic CNVs could contain clues about their functional role that can be deduced from their location in the genome, affected genes, size, type (duplication or deletion), and other characteristics. To start the investigation, a robust CNV detection pipeline is required.

1.4 Designing a pipeline for single cell CNV detection

Understanding why mosaic CNVs exist in neurotypical individuals begins with designing a robust pipeline for accurate detection of CNVs. Single cell CNV pipelines do not adhere to a single, standard formula; rather, they allow room for customization and experimentation. These decisions include choice of whole genome amplification (WGA) method, bin size, CNV detection algorithm, and quality control metrics.

1.4.1 Choosing the right whole genome amplification (WGA) method

WGA is perhaps the most critical step in any scWGS pipeline. Because each cell is potentially (if not literally) genomically unique, there are no biological replicates that can serve to validate the measurements made in the original cell, which is destroyed in the process. For this reason, accurate WGA is crucial.

Current WGA methods fall into three main groups (Gawad et al., 2016). The first and oldest category encompasses purely polymerase chain reaction-based (PCR-based) amplification methods, including degenerate oligonucleotide primed PCR (DOP-PCR). These methods rely on degenerate or random priming and PCR amplification to amplify genomic material. The second group consists of isothermal methods, most commonly multiple displacement amplification (MDA), which involves random priming and extension by the Phi29 polymerase without the necessity of temperature cycling. Finally, hybrid methods feature steps of isothermal linear amplification and exponential amplification by PCR; these methods include commercial products such as SMARTer PicoPLEX (Takara; formerly “PicoPLEX” by Rubicon Genomics) and GenomePlex

(Sigma-Aldrich) as well as protocols like multiple annealing loop-based amplification cycles (MALBAC) (Zong et al., 2012).

In practice, isothermal and hybrid methods are used most often, but for different types of experiments. MDA generates long, branched amplicons thanks to Phi29's high processivity and strand displacement activity, and these factors, along with its low error rate, make it ideal for identifying SNVs. However, MDA tends to over-amplify the regions that are primed first, making hybrid methods more desirable for use in CNV detection, where evenness of amplification is a chief concern (Gawad 2016).

1.4.2 Choice of bin size

Many single cell CNV papers have used normalized read counts in genomic bins (sometimes referred to as "windows") as a strategy to detect large CNVs in low coverage scWGS data (Baslan et al., 2012; Cai et al., 2014; Knouse et al., 2016; McConnell et al., 2013; Navin et al., 2011). For detection of megabase-scale CNVs, resolution of ~500kb is adequate; furthermore, binning reads permits the sequencing of fewer reads per cell, which in turn enables more cells to be sequenced in a single run.

Genomic bins are usually designed to account for the nonuniform distribution of mappable bases in the genome; if they are not, the read count variability due to highly mappable and highly unmappable regions can lead to false positive duplications and deletions, respectively. Bin coordinates are typically determined in one of two ways: first, by defining an arbitrary number of uniquely mappable bases to be grouped into one bin, e.g. 500,000, and using a mappability track to determine the additional length necessary for bins to contain that number of bases. This method was used by McConnell et al.

(2013) and resulted in a mean bin length of 687kb, demonstrating the compensation required for repetitive or otherwise unmappable sequences in the genome. An alternative approach to defining bin boundaries is to set an arbitrary number of unique simulated mappable reads that will comprise a bin; Cai et al. (2014) used read counts of 6,000, 20,000, and 50,000, and the resulting bins were 458kb, 137kb, and 54kb in median length, respectively. Both methods are effective in minimizing the influence of nonuniform mappability.

The specific bin size one should select will depend on several factors, such as the uniformity of WGA and the tendency of the CNV detection algorithm to detect false positives under different bin sizes. Baslan et al. (2012) suggest a median read count of 35 per bin, but this number may not be appropriate in all pipelines.

1.4.3 Evaluating CNV detection algorithms

Many algorithms exist for copy number segmentation, but all are designed to perform the same task: to identify non-overlapping, contiguous regions of chromosomes that have the same copy number state. Two of the most commonly used algorithms in the single cell CNV field have been R packages DNAcopy and HMMcopy (Cai et al., 2014; Knouse et al., 2016; Knouse et al., 2014; McConnell et al., 2013). DNAcopy uses circular binary segmentation (CBS) to detect changepoints in copy number data (Olshen et al., 2004), while HMMcopy employs a hidden Markov model (HMM) to either extend segments or call changepoints. The sensitivity of both algorithms to detecting CNVs can be tuned via parameters alpha (DNAcopy) and E (HMMcopy).

A good way to test CNV detection algorithms is by analyzing data with known variants and confirming that true positives are found. Analysis of trisomic cells or other stable cell lines carrying known variants represent the most straightforward forms of algorithm validation (Cai et al., 2014; McConnell et al., 2013). Meanwhile, Rohrback et al. (2018a) used V(D)J recombination in immune cells as a way to validate and train their CNV caller, FUnC, on somatic CNVs. Conversely, it is also worthwhile to test whether a CNV detection algorithm is prone to calling false positive CNVs. This can be done by simulating copy number data with varying levels of autocorrelated noise, a known cause of false positives (Muggeo and Adelfio, 2011).

1.4.4 Defining quality metrics

scWGS papers reporting mosaic CNVs in the brain have used a variety of approaches to sort good cells from bad. McConnell et al. (2013) employed three requirements on their dataset: a cell needed at least 500,000 unique reads; a median absolute deviation (MAD) of bin copy number values no higher than 0.35; and a "confidence score" no lower than 0.85, indicating broad adherence of genome segments to integer copy number states. Cai et al. (2014) employed one quality filter, a maximum median absolute pairwise difference (MAPD) of 0.45; MAPD was calculated for each cell by finding the median of the log2 copy number ratios of neighboring bins. Knouse et al. (2014) used a maximum variability score (VS) of 0.34; this score was determined by calculating the standard deviation (SD) of $\log_2$-based copy number values using sliding windows of 30 bins within chromosomes, averaging the SDs by chromosome, and then averaging the three highest autosomal average SDs. While different metrics were

calculated by each research group, all three approaches focused on bin-to-bin variability as a measure of cell quality. This is a sensible approach given the risk of nonuniform WGA driving the detection of false positive CNVs.

1.5 Summary

Building a CNV detection pipeline requires careful consideration of many factors, but doing so is necessary in order to generate a list of accurate, precise CNV calls through which the role of mosaicism can be explored. The McConnell Lab and its collaborators designed their first iteration of a CNV pipeline several years ago, resulting in the first scWGS paper reporting the detection of mosaic CNVs in post-mortem human neurons (McConnell et al., 2013). In Chapter 2, I discuss the improved pipeline we developed to analyze over 1200 neural cells from 15 human individuals, produce a CNV atlas, and examine possible roles of mosaic CNVs in the brain. In Chapter 3, I describe our approach to detecting mosaic neuronal CNVs in 10 mice of various backgrounds and genotypes. Finally, in Chapter 4, I discuss the conclusions of the work detailed in this dissertation and future experiments that can build upon the research presented here.

Chapter 2
Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex

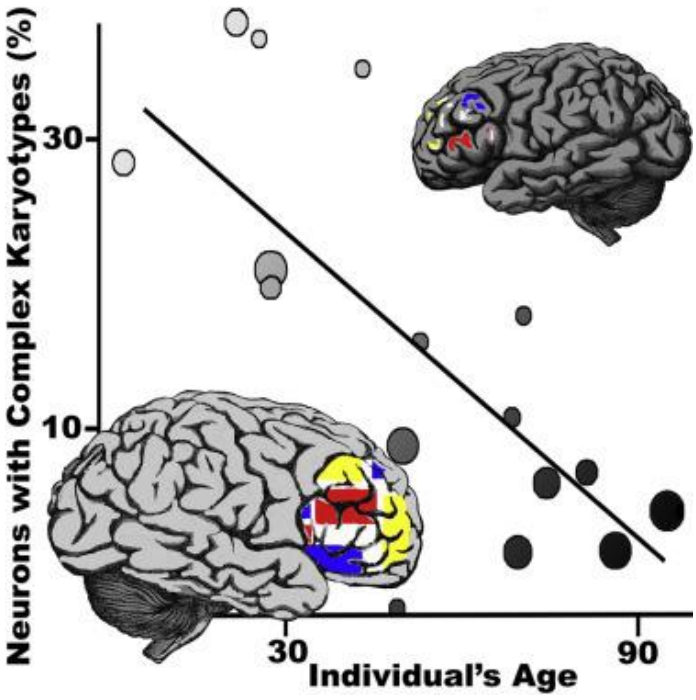
William D. Chronister, Ian E. Burbulis, Margaret B. Wierman, Matthew J. Wolpert, Mark F. Haakenson, Aiden C. B. Smith, Joel E. Kleinman, Thomas M. Hyde, Daniel R. Weinberger, Stefan Bekiranov, Michael J. McConnell

This chapter is adapted with minimal changes from: Chronister, W.D., Burbulis, I.E., Wierman, M.B., Wolpert, M.J., Haakenson, M.F., Smith, A.C.B., Kleinman, J.E., Hyde, T.M., Weinberger, D.R., Bekiranov, S., *et al.* (2019). Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell Reports *26*, 825-835 e827.

2.1 Summary

A subset of human neocortical neurons harbors complex karyotypes wherein megabase-scale copy-number variants (CNVs) alter allelic diversity. Divergent levels of neurons with complex karyotypes (CNV neurons) are reported in different individuals, yet genome-wide and familial studies implicitly assume a single brain genome when assessing the genetic risk architecture of neurological disease. We assembled a brain CNV atlas using a robust computational approach applied to a new dataset (>800 neurons from 5 neurotypical individuals) and to published data from 10 additional neurotypical individuals. The atlas reveals that the frequency of neocortical neurons with complex karyotypes varies widely among individuals, but this variability is not readily accounted for by tissue quality or CNV detection approach. Rather, the age of the individual is anti-correlated with CNV neuron frequency. Fewer CNV neurons are observed in aged individuals than in young individuals.

2.2 Graphical Abstract



2.3. Introduction

Neocortical neurons are among the most diverse and longest-lived mammalian cells. The mammalian cerebral cortex is often put forward as a pinnacle of evolutionary complexity, and human-specific brain phenotypes are attributed to neocortical expansion during evolution (Geschwind and Rakic, 2013; Lui et al., 2011). Aberrant development and maturation of neocortical circuits are likewise associated with neuropsychiatric and neurodegenerative diseases (Del Pino et al., 2018; Morrison and Baxter, 2012; Sudhof, 2017). Various approaches count ∼15–20 billion neurons and as many as 35 billion glia in the human cerebral cortex (Pakkenberg et al., 2003; von Bartheld et al., 2016). Single-cell transcriptomic approaches are beginning to comprehensively catalog human neuronal diversity (Lake et al., 2016; Nowakowski et al., 2017) and identify new subtypes of human neurons (Boldog et al., 2018). After decades of debate, it is now clear that human neocortical neurons are not normally regenerated during the human lifespan (Bhardwaj et al., 2006; Rakic, 2006). With some exceptions (Spitzer, 2017), neuronal cell types are also generally thought to be stable throughout life, but neuronal genomes are surprisingly labile.

Every human neocortical neuron may contain private somatic variants. Single nucleotide variants (SNVs) are especially common, with hundreds per neuron reported (Bae et al., 2018; Lodato et al., 2015) and with frequencies of >3,000 SNVs per neuron observed in aged individuals (Lodato et al., 2018). Endogenous mobile elements such as long interspersed nuclear element 1 (LINE1) retrotransposons are also active during brain development (Coufal et al., 2009; Muotri et al., 2005). Reported frequencies of *de novo* mobile element events range from <1 to >7 per neuron (Baillie et al., 2011; Evrony et al.,

2012; Upton et al., 2015). Mobile element activity has also been linked to the generation of copy-number variants (CNVs) (Erwin et al., 2016; Gilbert et al., 2002). Whole and subchromosomal CNVs bring about complex karyotypes through the duplication or deletion of several megabases (Mb) of genomic sequence in a subpopulation of neocortical neurons (Cai et al., 2014; Knouse et al., 2016; McConnell et al., 2013; Piotrowski et al., 2008). Gene density in the human genome averages >10 genes per Mb; thus, by contrast to other somatic variants, Mb-scale CNVs almost always affect multiple genes. A reanalysis of published data (Cai et al., 2014; Knouse et al., 2014; McConnell et al., 2013; van den Bos et al., 2016) herein found an average of 63 genes affected per neuronal CNV.

During the past decade, large CNVs have been recognized as major contributors to human genetic diversity (Conrad et al., 2010; Lupski, 2015; Redon et al., 2006). At the population level, SNVs are collectively more numerous than CNVs, but CNVs affect an order of magnitude more genome sequence (~10%), and some CNVs show evidence of positive selection during human evolution (Perry et al., 2007; Sudmant et al., 2015; Zarrei et al., 2015). In individuals, *de novo* CNVs represent rare variants with a strong contribution to the genetic risk of schizophrenia, autism, and other neurological disorders (Fromer et al., 2014; Iossifov et al., 2014; Marshall et al., 2017; Morrow, 2010; Sebat et al., 2007). Whereas the consequences of germline CNVs have been inferred from population-level studies, neuronal CNV studies to date have been underpowered to determine whether the genes affected by neuronal CNVs contribute to brain development, function, and disease.

We assembled a brain CNV atlas to evaluate how neuronal CNVs alter the genetic architecture of the neurotypical human cerebral cortex. A new dataset of 827 human cerebral cortical nuclei from 5 neurotypical individuals was combined with smaller published datasets (Cai et al., 2014; Knouse et al., 2014; McConnell et al., 2013; van den Bos et al., 2016) from 10 other neurotypical individuals. We developed an unbiased CNV detection approach based on population-level statistics and established a human neuronal CNV atlas with 507 CNVs. Initial analysis of the atlas identified substantial inter-individual variability in the frequency of neurons with complex karyotypes (CNV neurons), but also found support for the hypothesis (Weissman and Gage, 2016) that some long genomic loci shape the genetic architecture of neurotypical human brains.

2.4 Results

2.4.1 Generation of Single-cell Neuronal and Non-neuronal Genomic Data

Single neuronal and non-neuronal nuclei were isolated using fluorescence-activated nuclei sorting (FANS) of the prefrontal cortex from 5 non-diseased (neurotypical) male individuals aged 0.36, 26, 49, 86, and 95 years (Figure 2.1A; Table 1). Whole genome amplification (WGA) was performed using PicoPLEX (Rubicon Genomics), an approach that is similar to multiple annealing and loop-based amplification (MALBAC) (Zong et al., 2012), which produces Illumina-compatible libraries with 48 unique barcodes. Before library pooling and sequencing on Illumina platforms, we found that ~60% of WGA reactions produced a measurable product. Single-end or paired-end sequencing (50, 75, or 100 bp) of 48 pooled libraries on Illumina HiSeq Rapid platforms or of smaller pools (<17 libraries) on MiSeq platforms

routinely produced >1 million reads per library after duplicate removal. Neither paired-end nor longer read sequencing altered the data quality. Reads were aligned to hg19 and read depth was calculated across 4,505 genomic bins, each containing ~500 kb of uniquely mappable sequence (mean bin size = 687 ± 1,072 kb). This approach (Figure 2.1B) generated >134.2 Gb of genomic sequence from 827 male single-cell genome libraries (162.3 ± 115.1 Mb/cell).

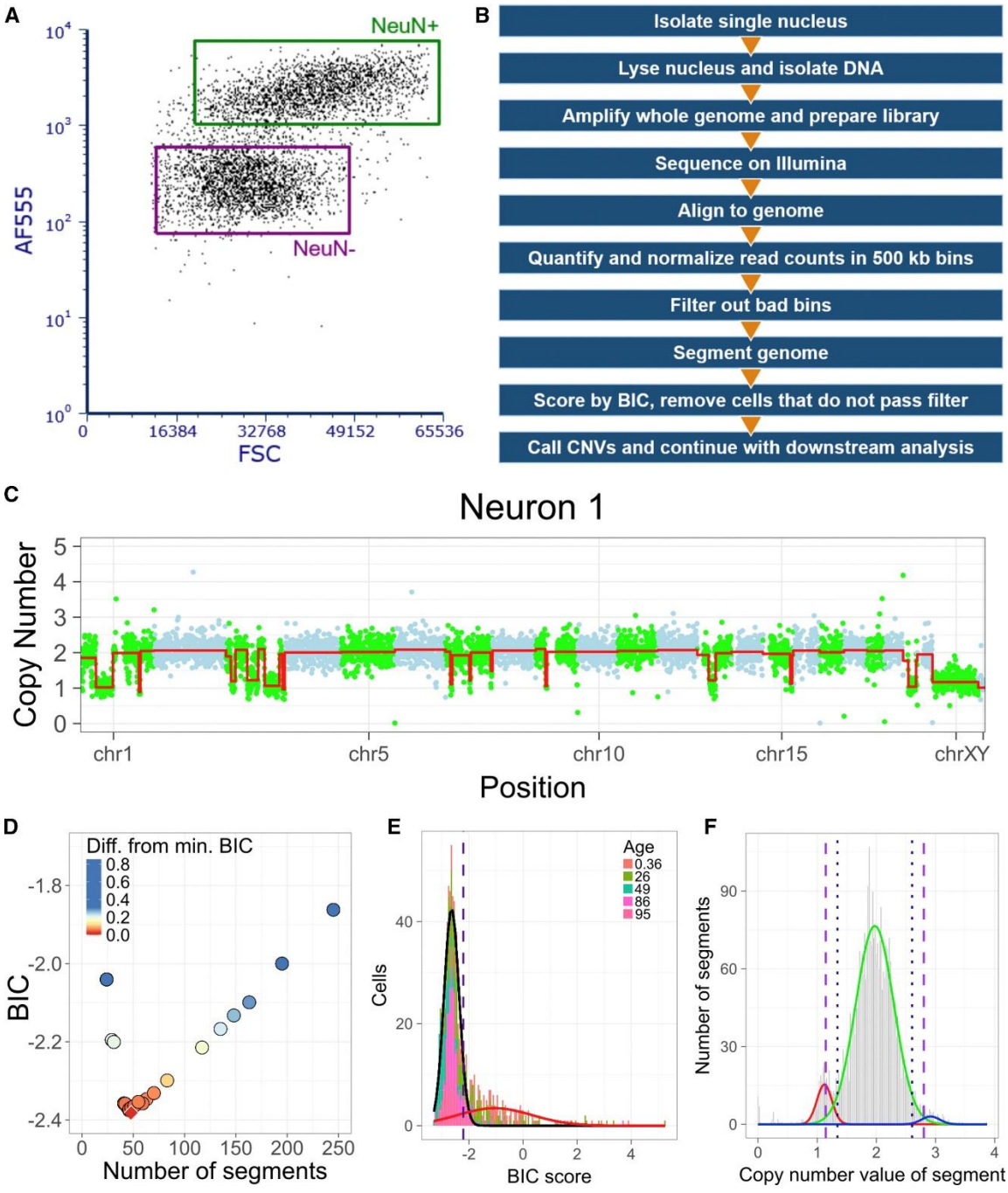


Figure 2.1: Optimization of Single-cell CNV Detection
(A and B) Representative NeuN$^{+/-}$ FANS **(A)** and summary of analysis pipeline **(B)**. **(C)** CNV profile of test Neuron 1. Read depth-derived CN values of genomic bins are colored alternately (green, blue) by chromosome. Red line indicates DNAcopy segmentation. **(D)** BIC scores for Neuron 1 across DNAcopy parameter space. Red diamond indicates lowest BIC score. **(E)** Histogram of BIC scores for new dataset. Gaussian distributions (black and red) were used to establish BIC cutoff (−2.21). **(F)** Segmentation output displays integer-like CN states (red, green, and blue) from which stringent (<1.14; >2.80; long dashes) and lenient (<1.34; >2.60; short dashes) CN state thresholds were established.

Table 2.1: Summary of Brain CNV Atlas

Age	Sex	Cell type	Data source	WGA	BIC pass/total cells (%)	Cells with CNV(s) (%), lenient	Cells with CNV(s) (%), stringent	Total CNVs (lenient)	Total CNVs (stringent)
0.36	Male	Neuron	This chapter	PicoPLEX	46/138 (33.3)	14 (30.4)	7 (15.2)	41	12
20	Female	Neuron	McConnell et al., 2013	GenomePlex	50/50 (100)	20 (40.0)	12 (24.0)	76	52
24	Female	Neuron	McConnell et al., 2013	GenomePlex	18/19 (94.7)	7 (38.9)	3 (16.7)	20	9
26*	Male	Neuron	This chapter	PicoPLEX	108/184 (58.7)	25 (23.1)	15 (13.9)	76	33
26*	Male	Non-neuron	This chapter	PicoPLEX	43/63 (68.3)	2 (4.7)	0 (0)	2	0
26*	Male	Neuron	McConnell et al., 2013	GenomePlex	41/41 (100)	9 (22.0)	6 (14.6)	18	13
42	Female	Neuron	Cai et al., 2014	GenomePlex	19/26 (73.1)	7 (36.8)	4 (21.1)	57	27
48	Female	Neuron	Knouse et al., 2014	GenomePlex	21/22 (95.5)	0 (0)	0 (0)	0	0
49	Male	Neuron	This chapter	PicoPLEX	99/101 (98.0)	11 (11.1)	8 (8.1)	113	75
49	Male	Non-neuron	This chapter	PicoPLEX	26/28 (92.9)	2 (7.7)	2 (7.7)	4	2
52	Male	Neuron	Knouse et al., 2014	GenomePlex	22/22 (100)	4 (18.2)	1 (4.5)	4	1
68	Male	Neuron	Knouse et al., 2014	GenomePlex	23/25 (92.0)	3 (13.0)	2 (8.7)	4	2
69	Male	Neuron	van den Bos et al., 2016	Strand-seq	78/81 (96.3)	3 (3.8)	2 (2.6)	3	2
70	Male	Neuron	Knouse et al., 2014	GenomePlex	20/20 (100)	4 (20.0)	4 (20.0)	9	4
74	Male	Neuron	van den Bos et al., 2016	Strand-seq	70/80 (87.5)	6 (8.6)	1 (1.4)	6	1
81	Female	Neuron	van den Bos et al., 2016	Strand-seq	43/72 (59.7)	4 (9.3)	2 (4.7)	15	6
86	Male	Neuron	This chapter	PicoPLEX	101/118 (85.6)	4 (4.0)	3 (3.0)	20	15
86	Male	Non-neuron	This chapter	PicoPLEX	46/55 (83.6)	4 (8.7)	2 (4.3)	9	3
95	Male	Neuron	This chapter	PicoPLEX	120/140 (85.7)	8 (6.7)	5 (4.2)	45	28

*Same individual

2.4.2 Optimization of Read Depth-based Single-cell Genomic Segmentation

Neuronal CNV detection is inherently challenging because one cannot know the state of the genome before WGA, and neuronal CNVs are rarely clonal. For this reason, CNV calling approaches have been used conservatively, with bias toward avoiding type I

errors at the risk of type II errors. We used a test dataset of 6 representative PicoPLEX WGA libraries (5 neurons and 1 trisomy 21 fibroblast) with varied subjective quality (Figures 2.1C and 2.S1A–2.S1E) to optimize our approach.

Parameter space in DNAcopy (Olshen et al., 2004) is defined by 3 user-tunable parameters: significance threshold (alpha), minimum number of genomic bins required to call a copy number state change (min.width), and the number of SDs between the levels of copy-number states to maintain the copy-number state change (undo.SD). We assessed DNAcopy parameters using Bayesian information criterion (BIC) (Schwarz, 1978), a log-likelihood estimate of the performance of the segmentation algorithm, for dozens of parameter combinations. The lowest, or near-lowest, BIC scores for each test cell library were observed at alpha = 0.001, min.width = 5, and undo.SD = 0 (Figures 2.1D and 2.S1G); these parameters identified monosomy X in all 6 male cells, and trisomy 21 in only the fibroblast. We also observed that minimum BIC scores were lower in WGA libraries with less overall bin-to-bin variation in read-depth (Figures 2.1C, 2.S1A–2.S1E, and 2.S1G), suggesting that low BIC scores represent an additional quality control filter.

We further assessed segmentation parameters with 2 in silico models (Figure 2.S1F) built from the read-depth statistics (DNAcopy segments, Gaussian noise, and autocorrelated noise [ϕ]) of each test cell. NULL model simulations contained no CNVs but did contain strong autocorrelated noise, a known source of DNAcopy false-positives (Muggeo and Adelfio, 2011). Alternative (ALT) model simulations harbored synthetic CNVs (i.e., DNAcopy calls) with residual autocorrelated noise. Segmentation outputs matched the ALT model, but not the NULL model, simulations well (Figures 2.S1H and

2.S1I). Concordance with ALT model cells (i.e., highest sensitivity and specificity) was also associated with the lowest BIC scores (Figure 2.S1J).

Analysis of 827 PicoPLEX datasets identified additional population-based filters. First, we identified 101 genomic bins that routinely deviated from median read depth and confound segmentation (Figure 2.S1K); these were excluded before the subsequent analysis. Second, we computed the 95th percentile of the low BIC score Gaussian distribution (−2.21) to establish an objective filter for the highest quality single-cell datasets (Figure 2.1E). Third, identified segments displayed 2 modes near integer-like copy number (CN) states of 2 (euploid, mean = 1.97) and 1 (deletion, mean = 1.12) and a heavy tail near the CN state of 3 (duplication, mean = 2.92) (Figure 2.1F). Lenient and stringent CN state thresholds were established, respectively, at a 2-tailed p value ≤ 0.05 (<1.34 for a deletion and >2.60 for a duplication) and ≤ 0.01 (<1.14 for a deletion and >2.80 for a duplication). The BIC threshold produced a final dataset of 589/827 (71.2%) neural nuclei, including 474/681 (69.6%) neuronal (NeuN+) nuclei and 115/146 (78.8%) non-neuronal (NeuN−) nuclei (Figures 2.S1L and S1M). The stringent CN state threshold identified monosomy X in 52.8% of male neural nuclei, while the lenient threshold identified 99.5% of these true-positives (Figure 2.S2A).

Additional NULL and ALT model simulations found that BIC cutoffs and CN state thresholds protect against false-positive CNVs brought about by simulated WGA-induced noise. We tested 6 neuronal WGA libraries in the BIC <−2.21 PicoPLEX dataset: 2 euploid, 2 with 1 CNV, 1 with 3 CNVs, and 1 with 6 CNVs. In the NULL model (Figure 2.S2B), all of the simulated cells from neuron 10 were excluded by our BIC threshold as a consequence of strong autocorrelated noise ($\phi = 0.663$; atlas mean $\phi =$

0.1394). In the remaining 1,000 simulated cells, only 41 and 1 small (3.9 ± 0.7 Mb) segments, respectively, passed the lenient and stringent CN state thresholds. Thus, upper-bound estimates of false-positive CNV neuron detection rates are ~3% (lenient threshold) and <0.1% (stringent threshold). All ALT model simulations (Figure 2.S2C) passed the BIC threshold, and only ~1% (14/1,200) of simulated cells contained DNAcopy segments that passed the lenient CN state thresholds without overlapping the synthetic CNV; these false-positive CNVs were also small (6.0 ± 2.7 Mb). As observed for monosomy X in the PicoPLEX dataset, improved detection of true-positives (synthetic CNVs) in ALT model cells was observed at lenient (95%) relative to stringent (81.8%) CN state thresholds.

2.4.3 Contribution of Mosaic CNVs to Neuronal Diversity

Large CNVs inevitably alter the CN state of a brain-expressed gene, as more of the genome is expressed in neurons than in other cell types (Uhlen et al., 2016). Our analysis (Table 1) identified 310 CNVs in 70 of 589 neural genomes (11.9%). CNVs ranged in size from 2.9 to 159.1 Mb (mean = 16.5 ± 20.1 Mb; Figure 2.2A). With stringent criteria that are prone to false-negatives, we still identify 168 CNVs (mean = 18.0 ± 22.3 Mb) in 42 neural nuclei (7.1%) (Figure 2.S3A). CNVs were detected in each individual examined and, given their size and frequency (Figures 2.2B and 2.S3D), represent a clear contribution to the genetic architecture of the brain (Figures 2.2C and 2.S3G). We note that much smaller percentages of phenotypically distinct cells can bring about focal epilepsies (Marin-Valencia et al., 2014) and are essential for normal brain

function (e.g., adult-born dentate granule neurons) (Anacker and Hen, 2017; Christian et al., 2014).

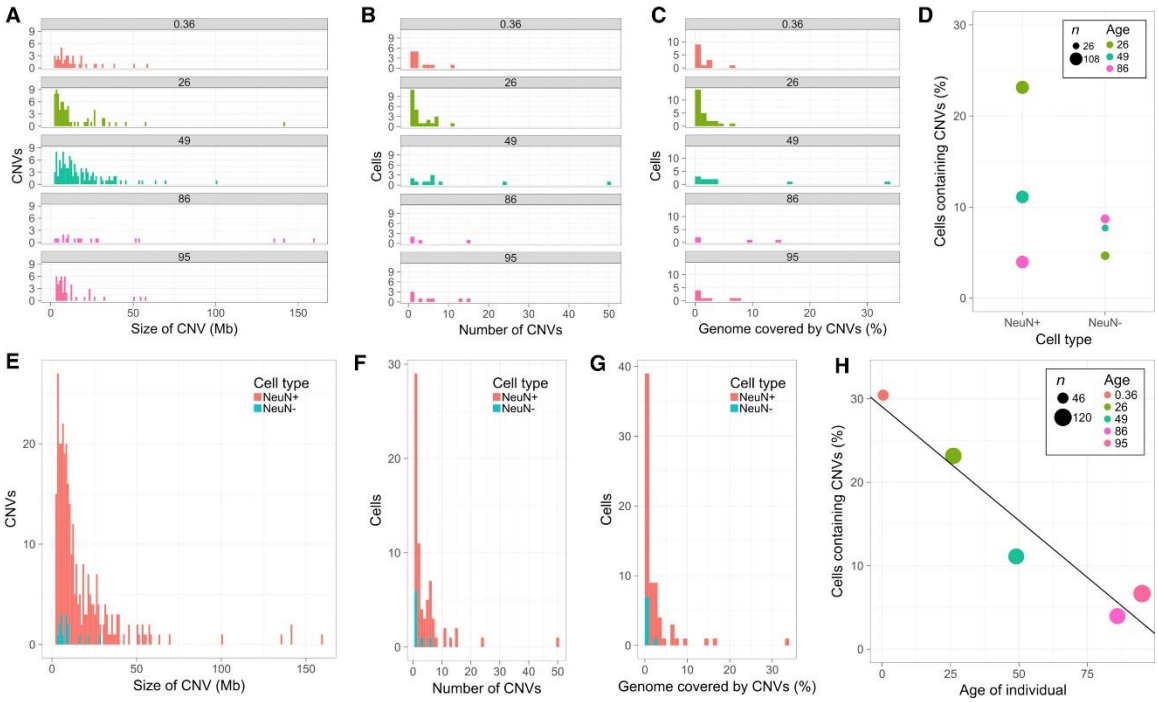


Figure 2.2 Mosaic CNVs Contribute to Neuronal Diversity
(A–C) Megabase-scale neuronal CNVs are observed across the human lifespan (ages 0.36–95 years) with varying size **(A)**, number per cell **(B)**, and percent genome coverage **(C)**. **(D)** Divergent CNV neuron (NeuN+), but similar non-neuronal (NeuN−), frequencies in 3 individuals. **(E–G)** CNVs have an increased impact on genetic architecture in neuronal genomes compared to non-neuronal genomes, as measured by size **(E)**, number per cell **(F)**, and percent genome coverage **(G)**. **(H)** Significant anti-correlation between age and CNV neuron frequency (linear fit, R2 = 0.9224, p = 0.0094). Lenient CN state thresholds throughout figure.

The 474 neuronal nuclei we analyzed represent the largest CNV dataset of neurotypical human brains generated by a single laboratory using a single WGA approach to date. We identified 62 CNV neurons (13.1%) with a mean CNV size of 16.8 Mb. We find that both neuronal and non-neuronal genomes harbor Mb-scale CNVs, which is consistent with previous reports (Cai et al., 2014; Knouse et al., 2016). The frequency (4%–23.1%) of complex karyotypes was more variable among neurons than among non-neuronal cells (4.7%–8.7%) from the same individuals (Figure 2.2D). CNV frequency in non-neuronal cells is similar to that reported in other somatic cells (e.g.,

keratinocytes). Relative to non-neuronal CNVs (Figures 2.S3B, 2.S3C, 2.S3E, 2.S3F, 2.S3H, and 2.S3I), neuronal CNVs are larger (Figure 2.2E), occur in greater numbers (Figure 2.2F), and affect more of the genome (1.3%–6% in neurons, 0.2%–0.8% in non-neurons; Figure 2.2G). Similar ratios of neuronal to non-neuronal CNVs were observed at stringent thresholds, but mean differences in non-euploid CN states were amplified (Figures 2.S3K– 2.S3M). A statistically significant linear decline in CNV neuron abundance was observed with age at both CN state thresholds (lenient: $R2 = 0.9224$, $p = 0.0094$, Figure 2.2H; stringent: $R2 = 0.9434$, $p = 0.0058$, Figure 2.S3O).

2.4.4 Analysis and Integration of Publicly Available Data

Four previous studies used different WGA and CNV detection approaches on 458 additional single neurons from 6 male and 5 female neurotypical individuals of different ages (Cai et al., 2014; Knouse et al., 2014; McConnell et al., 2013; van den Bos et al., 2016). WGA was performed using GenomePlex (Cai et al., 2014; Knouse et al., 2014; McConnell et al., 2013), a degenerate oligonucleotide-primed (DOP)-PCR-based approach, or Strand-seq (van den Bos et al., 2016), a "pre-amplification free" approach. We harmonized these with our PicoPLEX dataset (e.g., alignment to hg19, exclusion of outlier bins in Figures 2.S4A and 2.S4B, BIC thresholds in Figures 2.3A, 2.3B, 2.S4G, and 2.S4H). GenomePlex and Strand-seq datasets displayed distinct outlier bin profiles. Analysis of RepeatMasker (University of California, Santa Cruz [UCSC]) genomic features in outlier bins found a depletion (log2 fold change <-0.25) of DNA, LINE, and long terminal repeat (LTR) features, and enrichment (log2 fold change >0.47) for short interspersed nuclear element (SINE) and simple repeat features (Wilcoxon, $p < 2.2E{-}16$)

compared to remaining bins. BIC thresholds removed 11/225 GenomePlex cells (4.9%) and 41/233 Strand-seq cells (18.0%) from further analysis. The CN state thresholds reported above (Figure 2.1F) were inclusive of all of the WGA libraries passing BIC cutoffs. To further test the effectiveness of outlier bin removal, BIC cutoffs, and CNV thresholds, we re-analyzed split-amplification GenomePlex libraries from Knouse et al. (2016). One split-sample exceeded the BIC cutoff; the other with bad bins removed showed perfect concordance using our analysis pipeline (Figures 2.S4C–2.S4F).

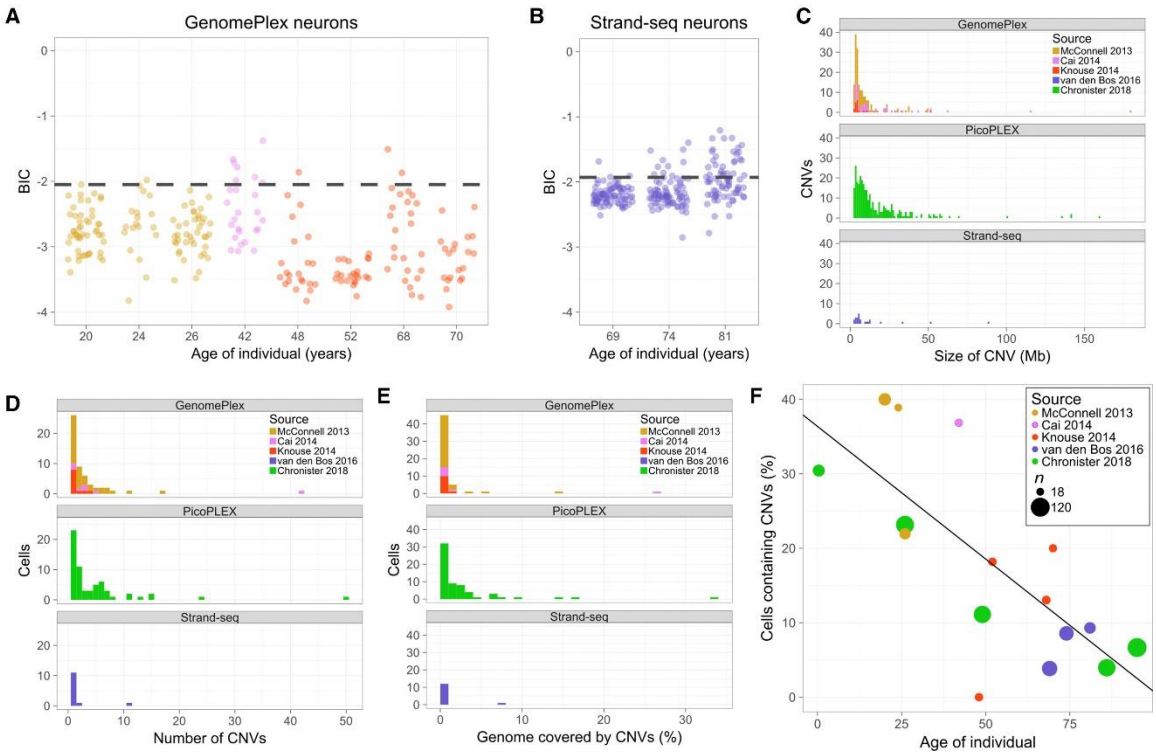


Figure 2.3 Brain CNV Atlas
(A and B) BIC scores established from population distribution of GenomePlex (**A**, <−2.05) and Strand-seq (**B**, <−1.93) WGA approaches. **(C–E)** CNV attributes including size **(C)**, number per cell **(D)**, and percent genome coverage **(E)** were similar regardless of WGA approach. **(F)** Significant atlas-wide anti-correlation between age and CNV neuron frequency (linear fit, R2 = 0.5521, p = 0.00097). Lenient CN state thresholds throughout figure.

The general characteristics of CNVs were similar regardless of WGA approach or sex (Figure 2.3C–2.3E). The mean length of single-cell CNVs was 14.8 Mb across all of the datasets, which is notably larger than most CNVs observed in bulk, germline human

genomes (1–10 kb) (MacDonald et al., 2014). As with PicoPLEX data, CNV size was variable among individuals (Figures 2.S4J, 2.S4K, 2.S4M, 2.S4N, 2.S4P, and 2.S4Q), but average CNV size was similar across WGA approaches (Figures 2.3C–2.3E). The average CNV size in GenomePlex libraries was 11.7 Mb, but it ranged from 4.1 to 15.9 Mb among individuals. The average CNV size in Strand-seq libraries was 13.2 Mb, but it ranged from 4.3 to 17.7 Mb among individuals. Similar but on average larger CNV sizes were observed at the stringent threshold (Figures 2.S4I, 2.S4L, and 2.S4O). In contrast to a previous report (Knouse et al., 2016), no enrichment for LINE1 sequence was apparent in CNVs or CNV borders compared to chance, and equal rates of telomeric CNVs (~14%) were observed in young and aged neurons (see Method Details).

Anti-correlation between CNV neuron prevalence and individual age is roughly an order of magnitude more significant in the complete atlas (lenient: $R2 = 0.5521$, $p = 0.00097$, Figure 2.3F; stringent: $R2 = 0.3941$, $p = 0.0092$, Figure 2.S4R). Despite narrower age ranges, CNV neuron prevalence was also anti-correlated with age in the published datasets (lenient: $R2 = 0.5315$, $p = 0.011$, Figure 2.S4S; stringent: $R2 = 0.3524$, $p = 0.054$, Figure 2.S4T). Potential confounding variables such as BIC scores or post-mortem interval showed no significant correlation with age (Figure 2.S5).

Our observation of fewer CNV neurons in aged individuals contrasts with the concept of genosenium (Lodato et al., 2018), which states that the accumulation of somatic mutations over one's lifetime is associated with aging-related cellular and molecular phenotypes. Thus, we also analyzed the number of CNVs per CNV neuron. Most CNV neurons (62.8%) contained only 1 or 2 CNVs, but the average number of CNVs per CNV neuron in the atlas was 3.9 (Figure 2.3D). Aged CNV neurons, although

rare, had more irregular karyotypes. For example, the 20-year-old individual had the highest percentage of CNV neurons (40% CNV neurons, 3.8 CNVs/CNV neuron), but the 49-year-old (11.1% CNV neurons) had the most CNVs/CNV neurons (10.3). Other individuals, such as the 24-year-old and the 86-year-old, also support this trend; 38.9% of neurons in the 24-year-old had CNVs with an average of 2.8 CNVs/CNV neuron, whereas only 4% of neurons in the 86-year-old had CNVs, but these CNV neurons averaged 5.0 CNVs. The individual with the highest mean CNV size (37.2 Mb) was 86 years old (Figure 2.S4J); this was due in large part to 1 neuron containing 2 trisomies and 1 monosomy (Figure 2.S3N). Genosenium, if true, may operate on different somatic mutations in distinct ways.

2.4.5 Long Gene Mosaicism and Neuronal Diversity

Brain CNVs, like all non-V(D)J CNVs, occur because DNA repair is not perfect. Transcription and replication lead to DNA double-strand breaks (DSBs), which in turn sometimes lead to CNVs. Gene length increases susceptibility to transcriptional and replicative genomic stress. Genes encoded by >100 kb of genomic sequence (i.e., long genes) tend to be neuronally expressed genes with roles in neuronal connectivity and synaptic plasticity (Zylka et al., 2015). Long genes also overlap with DNA fragile sites, and replicative stress can lead to large CNVs that encompass these loci (Wilson et al., 2015). Likewise, transcriptional stress leads to DNA DSBs in neurons (Madabhushi et al., 2015) and has a predominant effect on long gene transcript abundance (King et al., 2013). Recent studies link these observations to DNA DSBs during mouse neurodevelopment (Wei et al., 2016) and motivate the hypothesis that somatic mutations affecting long

genes mediate the functional consequences of brain somatic mosaicism (Weissman and Gage, 2016).

The assembled atlas identified 522 neural CNVs (Table 1). We performed overlap and random permutation analyses to determine whether subsets of 93 candidate long genes (Figure 2.4A) were associated with CNVs more frequently than expected by chance. Genomic regions (i.e., bins) that accumulated CNVs (i.e., duplications and deletions) at an increased population-wide frequency compared to the rest of the genome (i.e., hotspots; Figure 2.4B) were determined using different hotspot thresholds (see Method Details). Enrichment was assessed by calculating 184 raw p values (8 candidate gene lists, 23 hotspot thresholds) from 10,000 permutations per gene list per hotspot threshold. After correcting for multiple hypothesis testing (Benjamini-Hochberg false discovery rate [FDR], 5% FDR cutoff), dataset-wide significance was observed for the entire candidate gene list in some cases, and, notably, with putative hotspots that include 3 of the 4 common candidate genes: GPC6, NRXN3, and RBFOX1 (Figure 2.4C).

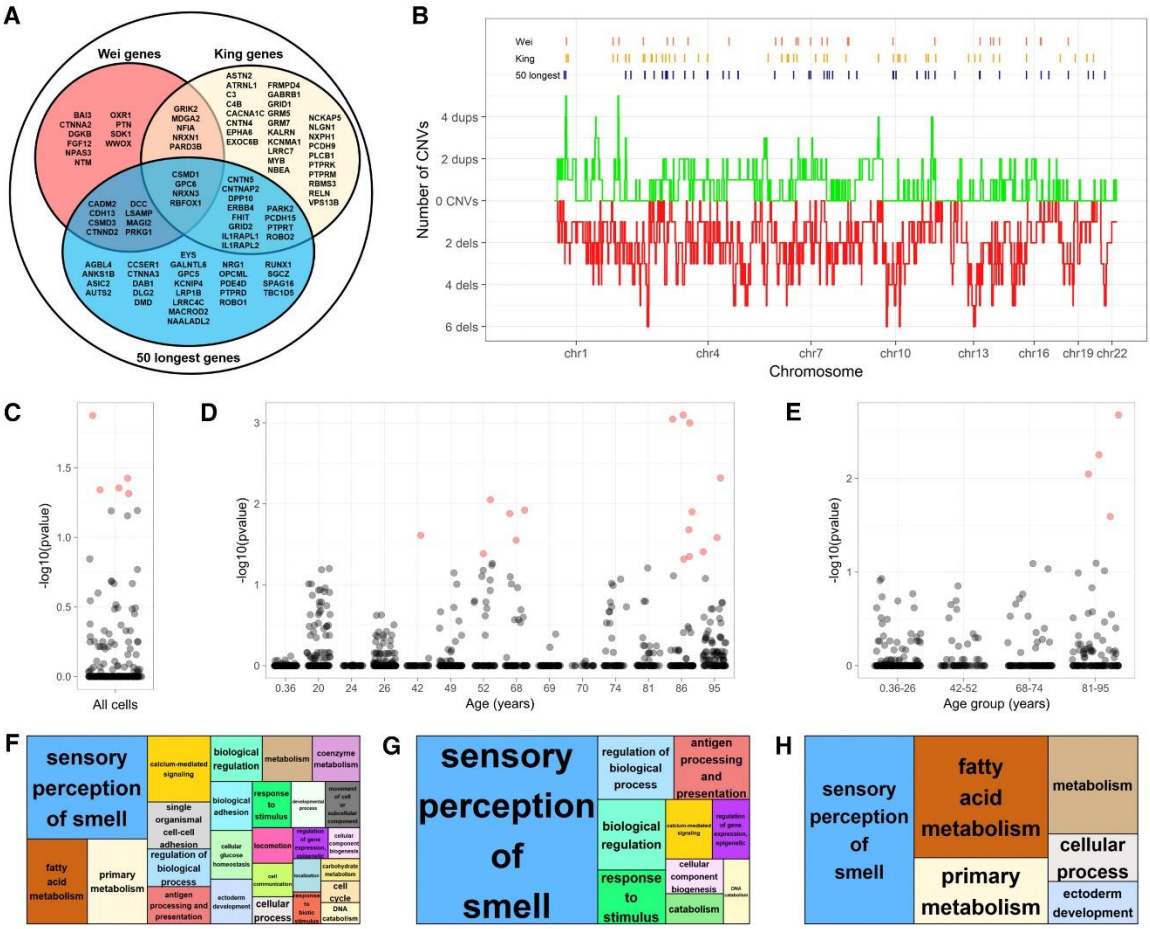


Figure 2.4: Long Gene Enrichment in Brain CNV Atlas
(A) Venn diagram to define putative hotspots. **(B)** Bin-level summary (lenient CN state threshold) of deletion (red) and duplication (green) occurrence in brain CNV atlas. **(C–E)** Enrichment results for hotspots from all cells **(C)**, individuals **(D)**, and age groups **(E)**. p values < 0.05 are red. **(F–H)** REViGO plots of enriched Gene Ontology (GO) terms for all of the data analyzed. The relative size of each category reflects significance; the largest groups have the lowest p values. GO enrichment determined using PANTHER analysis of CNV-affected genes in all neural data **(F)**, and neurons from age groups 68–74 **(G)** and 81–95 years old **(H)**.

CNV neurons were rare in aged individuals, so we assessed whether the genetic architecture of CNV neurons may also change during the lifespan. We further tested each individual in the atlas for candidate gene enrichment using relevant thresholds for recurrent hotspots (see Method Details) at stringent and lenient criteria. Significant corrected p values were not observed in the youngest individuals, but they were observed in aged individuals (Figure 2.4D). When individuals were pooled based on age groups, significant p values were also observed only in the most aged group (Figure 2.4E).

CNV-affected loci may not be restricted to long genes. We assembled a comprehensive list of all of the genes affected in the brain CNV atlas and used PANTHER (Mi et al., 2013) to calculate enrichment statistics and plotting scripts from REViGO (Supek et al., 2011) to visualize these results. Gene Ontology categories associated with “sensory perception of smell” and “calcium-mediated signaling” were notably enriched in the atlas-wide gene set (Figure 2.4F). When assessed by age group, as in Figure 2.4E, these enrichments were only detected in the aged groups (Figures 2.4G and 2.4H).

2.5 Discussion

Brain somatic mosaicism is a largely unexplored aspect of neuronal diversity (Harbom et al., 2018). In the human cerebral cortex, neuronal diversity is described in terms of electrophysiological properties (Contreras, 2004) and gene expression profiles (Lake et al., 2016) that are brought about by genetic programs (Lein et al., 2017). Current neurodevelopmental models implicitly assume that all somatic cells operate with identical genomes. In turn, population-based genetic studies of neurological disease typically sequence bulk blood DNA as a proxy for brain DNA. Somatic mutations affecting cell proliferation and survival pathways can alter neuronal diversity and lead to cortical overgrowth phenotypes ranging from hemimegalencephaly to focal dysplasia (Jamuar et al., 2014; Lee et al., 2012; Mirzaa et al., 2016; Poduri et al., 2012). Elevated levels of somatic mutations have also been associated with Rett syndrome (Muotri et al., 2010), neurodegenerative disease (Bushman et al., 2015; Iourov et al., 2009; Lee et al., 2018; Lodato et al., 2018; McConnell et al., 2004), schizophrenia (Bundo et al., 2014), and

altered behavior (Bedrosian et al., 2018). The Brain Somatic Mosaicism Network is an ongoing multi-site effort that aims to define how brain somatic mutations affect the genetic architecture of psychiatric disease (McConnell et al., 2017). However, the consequences of somatic mutations in neurotypical human brains remain a central unaddressed question. We report a neurotypical human brain somatic CNV atlas to begin to assess how mosaic somatic mutations affect neuronal diversity.

We assembled a brain CNV atlas from 1,285 single brain nucleus libraries, the accumulated work of our laboratory and 3 other laboratories. Development of a single, robust computational pipeline that both protected against false-positive CNV calls and minimized false-negative calls was essential. First, we identified and excluded a small group (101/4,505) of genomic bins that were consistently non-euploid (i.e., outliers) across our 5 analyzed individuals. Second, we applied an objective BIC cutoff to exclude WGA libraries that would be the most prone to aberrant segmentation calls. Third, we evaluated published datasets from 4 other sources that used different WGA methods. Population-level statistics were generally similar, but each WGA approach identified unique outlier bins and slightly different BIC cutoffs. Fourth, CN state distributions centered near CN = 1, 2, and 3 were apparent across all of the WGA approaches, so we used 2-tailed p values (lenient = 0.05, stringent = 0.01) from pan-method DNAcopy segmentation calls to define 2 sets of CN state thresholds. Fifth, we showed that these filters (outlier bin removal, BIC cutoffs, and CN state thresholds) effectively eliminate false-positive CNV calls (15/2,400; <1%). Lenient CN state thresholds are mildly (<3%) more prone to short (∼<6 Mb) false-positive calls, but far better at identifying true-positives (>95%). Our central findings are unchanged when only the largest (>6 Mb)

CNVs are considered. The brain CNV atlas comprises 879 neuronal and 115 non-neuronal genomic libraries; Mb-scale CNVs were identified in 129 neurons and 8 non-neurons.

The most salient feature of the brain CNV atlas is an anti-correlation between the age of an individual and the percentage of CNV neurons in the frontal cortex of that individual. This finding was significant in our new PicoPLEX dataset and in the published GenomePlex and Strand-seq datasets, and is highly significant using all of the available data ($p = 0.00097$). By contrast, the initial assessment of CNV location finds evidence for the enrichment of a subset of long genes and neurally associated Gene Ontology categories only in aged brains. Given the enrichment of these CNVs in aged, not young, neurons, CNVs affecting some genomic loci may be more compatible with neural survival than others. We found similar rates of CNV non-neurons at different ages; however, it will be interesting to determine whether other long-lived cells (e.g., cardiomyocytes) show a similar change in mosaic composition during aging.

We provide evidence that a functional consequence of CNV neurons may be selective vulnerability to aging-related cell death. Age-related cognitive decline is associated with notable decreases in cerebral cortical thickness, myelination, and synapse number accompanied by *ex vacuo* enlargement in ventricular volume (Morrison and Baxter, 2012). Although neuronal cell death is generally considered to be minimal in the healthy mature brain, rates of ~10% cerebral cortical neuron loss during adulthood are consistent with stereological counts in neurotypical individuals (Pakkenberg et al., 2003). The decline in CNV neuron prevalence that we observe between individuals <30 years old and individuals >70 years old is also strikingly consistent with selective CNV neuron

loss during a person's adult lifetime. We conclude that the most parsimonious interpretation of these data is that many, but not all, CNV neurons are selectively vulnerable to aging-associated atrophy. This cross-sectional finding highlights the unmet need for an increased longitudinal understanding of human neuronal genome dynamics during an individual's health span. Human pluripotent stem cell-based models (Brennand et al., 2015) represent a straightforward means to this end.

2.6 Methods

Table 2.2: Key Resources Table

Reagent or Resource	Source	Identifier
Antibodies		
mouse monoclonal anti-human NeuN IgG Alexa-Fluor 555 Conjugate clone A60	EMD via Millipore	Cat. # MAB377A5
Biological Samples		
Human pre-frontal cortex, Frozen	Lieber Institute for Brain Development	Anonymized Index #1845
Human pre-frontal cortex, Frozen	Lieber Institute for Brain Development	Anonymized Index #5154
Human pre-frontal cortex, Frozen	Lieber Institute for Brain Development	Anonymized Index #5401
Human pre-frontal cortex, Frozen	Lieber Institute for Brain Development	Anonymized Index #5570
Human pre-frontal cortex, Frozen	University of Maryland Brain and Tissue Bank	Anonymized Index #1583
Chemicals, Peptides, and Recombinant Proteins		
OptiPrep iodyxonol solution	Sigma-Aldrich	D1556-250ML
Acrylamide solution 37.5:1 (Acrylamide:Bis-acrylamide)	Bio-Rad	Cat. # 1610158
TEMED, 50 ml	Bio-Rad	Cat. # 1610801
Ammonium Persulfate (APS)	Bio-Rad	Cat. # 1610700
GelPilot DNA Loading Dye, 5x	QIAgen	Cat. # 239901
10 X Tris-Boric Acid-EDTA buffer	ThermoFisher	Cat. # B52
10 X Tris-Acetate-EDTA buffer	ThermoFisher	Cat. # B49
Magnesium Chloride anhydrous, ≥98%	Sigma-Aldrich	Cat. # M8266
Potassium chloride BioXtra, ≥99.0%	Sigma-Aldrich	Cat. # P9333

Reagent or Resource	Source	Identifier
Sodium chloride BioXtra, ≥99.5%	Sigma-Aldrich	Cat. # S7653
Sodium hydroxide	Sigma-Aldrich	Cat. # S8045
Polyoxyethylenesorbitan monolaurate CAS Number: 9005-64-5	Sigma-Aldrich	Cat. # P1379
Tris(hydroxymethyl)amino methane hydrochloride reagent grade, ≥99.0%	Sigma-Aldrich	Cat. # T3253
Sucrose for molecular biology, ≥99.5% (GC)	Sigma-Aldrich	Cat. # S0389
DL-Dithiothreitol BioUltra, for molecular biology, ≥99.5% (RT) (Sigma)	Sigma-Aldrich	Cat. # 43815
Trypan Blue Solution of 0.4%	ThermoFisher	Cat. # 15250061
Bovine Serum Albumin, lyophilized powder, suitable for (for molecular biology)	Sigma-Aldrich	Cat. # B6917
Hydrochloric acid ACS reagent, 37%	Sigma-Aldrich	Cat. # 320331
Glacial acetic acid	Sigma-Aldrich	Cat. # A6283
Ethylenediaminetetraacetic acid	Sigma-Aldrich	Cat. # EDS-100G
Ultra-Pure Low Melting Point agarose	ThermoFisher	Cat. # 16520050
EDTA-free Protease Inhibitor Cocktail; Tablets	Roche via Sigma-Aldrich	Cat. # 11873580001
SYTO 13 green fluorescent nucleic acid stain	ThermoFisher	Cat. # S7575
SYBR Gold nucleic acid gel stain	ThermoFisher	Cat. # S11494
Ethidium Bromide Solution (10 mg/mL)	ThermoFisher	Cat. # 17898
DAPI (4',6-Diamidino-2-Phenylindole, Dihydrochloride)	ThermoFisher	Cat. # D1306
UltraPure™ DNase/RNase-Free Distilled Water	ThermoFisher	Cat. # 10977015
Phosphate Buffered Saline, PBS (10X), pH 7.4	ThermoFisher	Cat. # 70011044
IsoFlow Sheath Fluid	Beckman Coulter	Cat. # 8546859
Critical Commercial Assays		
Rubicon PicoPLEX WGA Kit	via Agilent	Cat. # 5190-9533
QIAquick PCR purification columns Kit	QIAgen	Cat. # 28106
MiSeq Reagent Kit v2 (50 cycle)	Illumina	MS-102-2001

Reagent or Resource	Source	Identifier
HiSeq PE Rapid Cluster Kit v2	Illumina	PE-402-4002
Qubit™ dsDNA HS Assay Kit	ThermoFisher	Cat. # Q32851
Deposited Data		
Single cell sequencing data	This study (NDA Study ID 636)	https://dx.doi.org/10.15154/1503237
Software and Algorithms		
Bedtools version 2.17.0	Quinlan and Hall, 2010	https://bedtools.readthedocs.io/en/latest/index.html
biomaRt (R package) version 2.36.1	Smedley et al., 2015	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
BWA version 0.7.12	Li and Durbin, 2009	http://bio-bwa.sourceforge.net/
DNAcopy (R package) version 1.50.1	Seshan and Olshen, 2017	https://bioconductor.org/packages/release/bioc/html/DNAcopy.html
FASTX Toolkit version 0.0.13	http://hannonlab.cshl.edu/	http://hannonlab.cshl.edu/fastx_toolkit/download.html
mixtools (R package) version 1.1.0	Benaglia et al., 2009	https://cran.r-project.org/web/packages/mixtools/index.html
PANTHER	Mi et al., 2013	http://www.pantherdb.org/
Picard Tools version 1.105	https://broadinstitute.github.io/picard/	https://github.com/broadinstitute/picard/releases
Python version 2.6.6	https://www.python.org/	https://www.python.org/downloads/
R version 3.4.1	https://www.r-project.org/	https://cran.r-project.org/src/base/R-3/
REViGO	Supek et al., 2011	http://revigo.irb.hr/
Samtools version 1.1	Li et al., 2009	http://www.htslib.org/download/
Simulation script	This study	https://github.com/mcconnell-lab/scripts/blob/master/ALT_NULL_dataset_simulation.R
Other		
0.2 ml TempAssure PCR 8-tube strips	USA Scientific	Cat. # 1402-2700
Gloves large nitrile	Denville Scientific	Cat. # G4163
Dounce Tissue grinder Pestle A clearance 0.0030-0.0050 in. Pestle B clearance 0.0005-0.0025 in.	Kimble via Sigma-Aldrich	Cat. # D8938
SW 55 Ti Swinging-Bucket Rotor	Beckman-Coulter	Cat. # 342196
13 × 51 mm, 5 ml Thinwall polyallomer tubes	Beckman-Coulter	Cat. # 326819
Qubit Reader 3.0	ThermoFisher	Cat. # Q33216
Polytron PT 1300 D Manual Disperser	Kinematica Inc. via Fisher Scientific	Cat. # 08-451-71
100 bp DNA ladder	New England Biolabs	Cat. # N3231
Corning 0.22 µm bottle top filter	Sigma-Aldrich	Cat. # CLS430769
CellRaft System Kit for Inverted Microscopes	Cell Microsystems	P/N: CRK

Reagent or Resource	Source	Identifier
Syringe PP/PE without needle	Sigma-Aldrich	Cat. # Z230723

2.6.1 Human subjects

Brain tissue was taken from five human individuals. Descriptive details of these individuals, such as age, sex, and number and type of cells, are available in Table 1. Four samples (Lieber) were collected under the IRB approved Protocol Title: Collection of Postmortem Human Brain, Blood and Scalp Samples for Neuropsychiatric Research', registered to Thomas Hyde, MD, PhD, BA, Located on the Third Floor, 855 N. Wolfe Street, Baltimore, MD 21205, United States. Executed board action date: 06/19/2017. STUDY NUM: 1126332; ONLINE TRACKING: INVEST NUM: 165719.WO NUM: 1-1010280-1. STUDY APPROVAL EXPIRES: 07/18/2019 and is renewed annually. Informed consent was obtained from all subjects. The Western Institutional Review Board® (1019 39th Avenue SE. Suite 120 | Puyallup, WA 98374-2115; Office: (360) 252-2500 | Toll Free: (800) 562-4789; http://www.wirb.com/Pages/default.aspx) certifies the protocols for collecting and transferring clinical samples in the context of this study are in full compliance with good clinical practices as defined under the U.S. Food and Drug Administration regulations, and the International Conference on Harmonisation guidelines. These statements are true and correct as reflected in the records of the Western Institutional Review Board (WIRB), OHRP/FDA parent organization number IORG 0000432, IRB registration number IRB00000533. The fifth was obtained from the National Institute for Child Health and Human Development (NIH) Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD, contract HHSN2752009000011C, ref. no. N01-HD-9-011.

2.6.2 Single nuclei isolation

All procedures were done on ice. Solutions were made using nuclease-free water and supplemented with freshly prepared 50 X proteinase inhibitor just before use. Nuclei isolation media, optiprep nuclei diluent, nuclei storage buffer and blocking buffer (prior to adding tween) were 0.22 μm filtered.

Buffers used:

1. Nuclei Isolation Media (NIM): 25 mM KCl, 5 mM MgCl2, 10 mM Tris-Cl (pH 8.8), 250 mM sucrose, 1mM dithiothreitol (DTT).

2. OptiPrep Density Gradient Medium (ODGM): 60% Iodixanol solution.

3. OptiPrep Nuclei Diluent (OND): 150 mM KCl, 30 mM MgCl2, 60 mM Tris-Cl (pH 8.8), 250 mM sucrose.

4. Blocking Buffer (BB): 1x PBS, 1.0% Bovine Serum Albumin (BSA), and 0.1% Tween 20.

5. Nuclei Storage Buffer (NSB): 5 mM MgCl2, 50 mM Tris-Cl (pH 8.8), 166 mM sucrose, 1 mM dithiothreitol (DTT).

6. 50 X EDTA-free Protease Inhibitor Cocktail in nuclease-free water.

7. 10% Triton x-100 in nuclease-free water.

Post-mortem cortex was stored at −80°C until fragmented while frozen on dry ice in a pre-chilled mortar and pestle. Fragments (~100 mg) were completely solubilized in 1 mL of nuclei isolation media (NIM) by gently triturating using a 1000 μL pipette tip. We transferred slurry to a 5 mL round bottom, polypropylene tube and homogenized using a Polytron PT 1300 D tissue disruptor for 2 minutes. The sample appeared opaque and homogeneous. We added 1 μl of 10% Triton X-100 (final conc. 0.01%) and gently mixed

by rotation before transferring to a dounce with large and small pestle clearance (0.0030-0.0050 in. and 0.0005-0.0025 in., respectively). After 15 stokes with each clearance we microscopically verified cell disruption with 0.4% trypan blue solution. We centrifuged samples at 1,000 x g for 8 minutes at 4°C and re-suspended the pellet in 1 mL of 6:5:1 NIM:ODGM:OND (25% iodixanol). We used a 1 mL syringe (without needle) to apply the suspension onto 1 mL of 29:31 ODGM:OND solution (29% iodixanol) in a 5 mL thin-wall polyallomer ultracentrifuge tube (13 × 51 mm). Samples were centrifuged at 10,300 x g for 20 minutes at 4°C in a Beckman L8-M ultracentrifuge using a SW55 Ti rotor. The supernatant containing cell debris was removed, leaving ∼50 μL in the bottom of the tube. We confirmed the presence of nuclei by microscopy using 0.4% trypan blue and used immediately or stored up to 1 week in NSB at 4°C without reduction in whole genome amplification efficiency.

2.6.3 Single nuclei genome sequencing

We labeled nuclei derived from neurons by incubating with mouse monoclonal anti-human NeuN IgG clone A60 (Alexa Fluor 555 conjugate) diluted 1:250 in blocking buffer overnight at 4°C. We verified that NeuN+ nuclei also contained dsDNA by co-staining with either SYTO 13 green fluorescent nucleic acid stain at 500 nM or standard DAPI. We isolated individual NeuN+, DNA+ nuclei by flow sorting into 8-well thin-wall PCR tube strips. At certain points in the execution of these procedures we also isolated individual nuclei from the flow-sorted NeuN+, DNA+ nuclei pool using the CellRaft system (Cell Microsystems) to verify integrity of the nuclei and quality of the WGA on a small scale. There are videos and extensive literature on the use of the CellRaft at

https://www.cellmicrosystems.com and Wierman et al. (2017). These quality control steps ensured the nuclei were intact and contained genomic material suitable for WGA. The PicoPLEX Whole Genome Amplification Kit was applied to single nuclei according to manufacturer's instructions (Rubicon Genomics, Ann Arbor, MI). The PicoPLEX reaction enzymatically copies elements across the entire genome, fragments DNA products, and bar-codes the fragments with unique Illumina i5 and i7 index sequences. We confirmed reactions produced high molecular weight DNA by 1 X TBE, 1% agarose gel electrophoresis containing 1 μg/ml ethidium bromide. Productive reactions were purified using QIAquick PCR purification columns. We quantified yields using the high-sensitivity (HS) DNA Qubit 3.0 assay and combined equimolar portions of each purified PicoPLEX product into pools of 48 nuclei with compatible index combinations for multiplex sequencing. We electrophoresed pooled libraries into 0.75 mm thick 1 X TBE, 7.5% polyacrylamide (37.5:1 acrylamide:bisacrylamide) gels at 35 mA for 20 minutes and incubated with SYBR Gold DNA stain diluted 1:10,000 in 1 X TBE for 5 minutes before excising sections containing 450 bp to 800 bp DNA fragments on a UV transilluminator (long wave). We electro-eluted DNA from polyacrylamide sections into 1 X TAE, 1% low-melt agarose gel containing 1 μg/ml ethidium bromide at 100 V for 15 minutes and isolated chunks containing DNA on a UV transilluminator (long wave). DNA was purified from agarose using QG buffer according to the QIAquick PCR column protocol (QIAGEN). We quantified size-selected pooled library DNA by HS DNA Qubit 3.0 assay and diluted to 6 nM prior to sequencing on the Illumina (San Diego, CA) platform. We sequenced a total of 829 brain nuclei and used these for BIC and copy number state population statistics, including threshold determination. Two

single nuclei were flagged for ambiguous provenance. The CNV atlas includes 827 male brain nuclei that were amplified by the DNA WGA PicoPLEX kit. These single nucleus genome sequencing procedures were previously described in McConnell et al. (2013) and in greater detail in Wierman et al. (2017).

2.6.4 Analysis of single cell sequencing data

Sequence reads from Illumina were trimmed of PicoPLEX primers using the fastx_trimmer command (hannonlab.cshl.edu/fastx_toolkit/). Reads were then aligned to the human genome (version hg19) with BWA-aln V0.7.12 using default options (Li and Durbin, 2009) and converted to BAM format using Samtools V1.1 (Li et al., 2009). Duplicates were removed using MarkDuplicates (Picard tools V1.105, broadinstitute.github.io/picard). Using a 40-mer mappability track (UCSC, wgEncodeCrgMapabilityAlign40-mer.bigWig) to determine uniquely mappable bases, we divided the genome into 4,505 dynamically sized genomic bins, each containing 500kb of mappable sequence. The mean bin size was 687 kb. Read counts for each bin were determined by Bedtools V2.17.0 coverageBed (Quinlan and Hall, 2010). To avoid read count bias arising from GC content, bins were grouped into 16 roughly equal size groups according to GC percentage and each read count within a GC group was divided by the median read count of the group and multiplied by two.

Following analysis of several hundred single cell datasets, we observed that certain genomic bins were consistently above or below the euploid state, most likely due to biases arising from alignment or artifacts generated during WGA. To avoid biases in segmentation introduced by these outlier bins, namely false positive and false negative

CNVs, we used Tukey's Outlier Method on the median log-copy number values of all 4,505 bins in a sex and WGA-specific fashion, resulting in between 101 and 153 bins to be excluded from segmentation, depending on the sex of the individual and WGA used. In addition to the bins excluded by the outlier detection method, two Y chromosome bins were manually excluded from female Strand-seq dataset. Single cell datasets were segmented using DNAcopy (Seshan and Olshen, 2018), an R package (www.R-project.org) that implements circular binary segmentation (CBS) to detect copy number "changepoints" in genomic data. DNAcopy was run on the normalized bin data using parameters alpha = 0.001, undo.SD = 0, and min.width = 5.

2.6.5 BIC scoring and filtering

To determine which samples were of sufficient quality to merit further analysis, we implemented Bayesian Information Criterion (BIC) (Schwarz, 1978)

$$BIC = \ln(\sigma^2) + \frac{k_p}{n}\ln(n)$$

as a scoring metric where σ^2 is the variance of the data points about their respective segment means, k_p is equal to 1 + 2*(number of changepoints), and n is the number of bins (4,505) assuming a piecewise-constant, Gaussian error segmentation model (see Muggeo and Adelfio (2011) for details of segmentation modeling and parameterization). The equation above shows the usual BIC score divided by the number of bins, which is fixed across samples. BIC penalizes a segmentation that under-fits the data (i.e., allows high variance of data within segments) or over-fits the data (i.e., creates too many segments); thus, a cell with properly fitted data and relatively low bin value variance will receive a low BIC score.

To define threshold BIC scores for inclusion in further analysis, a histogram of BIC scores was generated for each WGA method. Using the R package mixtools (Benaglia et al., 2009), we fit two Gaussian distributions to the PicoPLEX and Strand-seq histograms, and one Gaussian distribution to the GenomePlex histogram due to its displaying a single mode. Using the Gaussian distribution with the lower mean (or, in the case of GenomePlex, the lone distribution), we set the threshold BIC score for inclusion to correspond to $p = 0.05$ on the upper tail. For PicoPLEX, cells scoring below −2.21 were selected for further analysis (Figure 2.1E); for GenomePlex, the threshold was −2.05 (Figure 2.S4G); for Strand-seq, the cutoff was −1.93 (Figure 2.S4H).

2.6.6 Defining CNVs

Because DNAcopy does not assign integer copy number values to the segments it outputs, it was necessary to define threshold copy number values for a segment to be considered a CNV. We set CNV thresholds by evaluating autosomal segments of sizes ranging from 5, our minimum number of bins required to call a CNV, and 45 bins, a length smaller than the shortest autosome, resulting in a distribution of copy numbers excluding the high number of whole chromosomes at or near copy number two. The segments were plotted by copy number value in a histogram, and we fit a three-Gaussian mixture model using mixtools (Benaglia et al., 2009) and plotted the resulting mixed Gaussian model of three distributions centered at the local peaks near copy number 1, 2, and 3. Using the central Gaussian, centered near 2, we calculated two sets of thresholds: the stringent thresholds, 2.80 and 1.14, the result of a cumulative two-tailed probability of 0.01; and the lenient thresholds, 2.60 and 1.34, determined by a two-tailed cumulative

probability of 0.05. Throughout our CNV analyses, we used both thresholds in order to give a wider range of possible results and avoid any influence of false positive or false negative CNVs. Additionally, CNVs were required to be at least 5 bins in length, and chromosomes X and Y were not examined for CNVs.

2.6.7 Test data simulation

DNAcopy is prone to calling CNV events using correlated noise as input (Muggeo and Adelfio, 2011), so we sought to determine the degree to which our single cell data contains correlated noise. We selected six single cells to simulate test data. In our "NULL model," we assumed that the data for our six cells were described by correlated noise and Gaussian noise about the euploid copy number state, and contained no real CNV events. In the "alternative model," we assumed that the six cells were described by real CNV events identified by DNAcopy and residual correlated and Gaussian noise. We simulated 200 test data cells for each of the cells under each of the models and found that the CNVs produced by the NULL model rarely matched those of the original cell in size or divergence from 2 (Figure 2.S1H), leading us to conclude that the ALT model was a more accurate representation of our data (Figure 2.S1I). To determine the best DNAcopy segmentation parameters, we used simulation data of the "alternative model" to explore values of alpha ranging from 0.05 to 10^{-5}, undo.SD ranging from 0 to 5, and min.width ranging from 2 to 5, and calculated a BIC score for each. We also tested the performance of each segmentation using a receiver operating characteristic (ROC) curve to determine the parameters at which sensitivity was maximized and false positive CNVs were minimized (Figure 2.S1J). We concluded that

the best parameters for segmentation were alpha = 0.001, undo.SD = 0, and min.width = 5.

We later simulated data based on 6 additional cells to assess the effect of our BIC and CNV thresholds on false positives and false negatives using ALT and NULL model simulations as described above (Figure 2.S2B-2.S2C).

2.6.8 RepeatMasker analysis of outlier bins

To see if outlier bins showed any differences in genomic features compared to normal bins, we selected 100 outlier bins and 100 normal bins of similar size and calculated base overlap with RepeatMasker (UCSC) feature types, such as LINE, SINE, simple repeats, long terminal repeats, etc. Following normalization of overlapping bases by bin size, we performed a Wilcoxon rank-sum test for each RepeatMasker feature type to test whether there was a difference in the fraction of overlapping bases between the bin groups.

2.6.9 RepeatMasker enrichment analysis of CNVs

To check for overlap of genomic features in the CNVs that we detected, we selected a subsample of CNVs corresponding to the 5th, 25th, 50th, 75th, and 95th percentiles of CNV size and used Bedtools shuffle (Quinlan and Hall, 2010) to generate 1000 randomized genomic locations for each CNV. We then computed the percent overlap of each simulated CNV with RepeatMasker (UCSC) feature types. From this, we computed the mean and standard deviation percent overlap of the randomized CNVs which we used to calculate a z-score associated with the actual overlap as a function of

CNV size. We also checked for border region enrichment by calculating the percent overlap of RepeatMasker feature types with the first and last megabase of each simulated CNV and deriving the associated z-score of the actual CNV border overlap as a function of CNV size.

2.6.10 Gene set enrichment

To explore the biological significance of large scale CNVs in the brain, we sought to follow up on the work of two papers that identified lists of long genes that may be predisposed to DNA breaks in mice (King et al., 2013; Wei et al., 2016). Building upon this idea, we also examined a third list containing the 50 longest human genes obtained using biomaRt (Smedley et al., 2015). These gene lists shared genes in common with one another. There were ultimately seven gene lists drawn from one, two, or three of the original lists, as well as an eighth gene list containing all 93 genes gathered from all three sources. To test for enrichment, we collected hotspot coordinates corresponding to a minimum number of CNV events, which varied depending on the subset of data being tested and the CNV threshold used. For each data subset, we set hotspot cutoffs for the required number of CNVs such that no cutoff should lead to greater than 20% coverage of the human genome. The resulting cutoffs for each subset and threshold are listed in Table 2.3.

Table 2.3: Hotspots and Coldspots Used for Gene Set Enrichment

Individual(s)	Threshold	CNVs	Dels.	Dups.
All	Lenient	0, 5-8, 6-8, 7-8, 8	0, 4-6, 5-6, 6	2-5, 3-5, 4-5, 5
0.36-26	Lenient	3-5, 4-5, 5	2-5, 3-5, 4-5, 5	2-4, 3-4, 4
42-52	Lenient	3-4, 4	3-4, 4	1
68-74	Lenient	1-2, 2	1-2, 2	1-2, 2
81-95	Lenient	2-4, 3-4, 4	2-3, 3	1-2, 2
0.36	Lenient	1-3, 2-3, 3	1-2, 2	1-2, 2

Individual(s)	Threshold	CNVs	Dels.	Dups.
20	Lenient	2-4, 3-4, 4	2	1-2, 2
24	Lenient	1	1	1
26 NeuN+	Lenient	2-5, 3-5, 4-5, 5	1-3, 2-3, 3	1-4, 2-4, 3-4, 4
26 NeuN-	Lenient	NA	NA	1
42	Lenient	2	2	1
48	Lenient	NA	NA	NA
49 NeuN+	Lenient	2-3, 3	3	1
49 NeuN-	Lenient	1	1	1
52	Lenient	1-2, 2	1-2, 2	1
68	Lenient	1	1	1
69	Lenient	1	1	1
70	Lenient		1	
74	Lenient	1-2, 2	1-2, 2	1
81	Lenient	1	1	1
86 NeuN+	Lenient	2	1	1
86 NeuN-	Lenient	1	1	1
95	Lenient	1-3, 2-3, 3	1-2, 2	1
All	Stringent	0, 4-6, 5-6, 6	3-5, 4-5, 5	2-4, 3-4, 4
0.36-26	Stringent	2-5, 3-5, 4-5, 5	2	1-3, 2-3, 3
42-52	Stringent	NA	2-3, 3	NA
68-74	Stringent	1-2, 2	1	1-2, 2
81-95	Stringent	2-4, 3-4, 4	2-3, 3	1
0.36	Stringent	1-3, 2-3, 3	1	1-2, 2
20	Stringent	2-4, 3-4, 4	1-2, 2	1-2, 2
24	Stringent	NA	1	NA
26 NeuN+	Stringent	1-3, 2-3, 3	1-2, 2	1-2, 2
26 NeuN-	Stringent	NA	NA	NA
42	Stringent	NA	1	NA
48	Stringent	NA	NA	NA
49 NeuN+	Stringent	NA	2	NA
49 NeuN-	Stringent	NA	1	NA
52	Stringent	NA	1	NA
68	Stringent	1	1	1
69	Stringent	1	1	1
70	Stringent	NA	1	NA
74	Stringent	NA	NA	1
81	Stringent	1	1	1
86 NeuN+	Stringent	NA	1	1
86 NeuN-	Stringent	NA	1	NA
95	Stringent	1-3, 2-3, 3	1-2, 2	1

For each set of hotspots (Table 2.3), we randomly shuffled the hotspot loci within the genome 10,000 times to generate a null model of CNV coverage. We then determined

the number of genes of interest found in the actual hotspots and the range of genes of interest found in the null model to generate enrichment p values, which were then FDR-corrected using the Benjamini-Hochberg method. Owing to the considerable inter-dependence among hotspot sets being tested, these p values were FDR-corrected within stratified sub-groups; for example, p values for hotspots derived from individuals were corrected separately from p values for hotspots derived from age groups. Likewise, p values for hotspots defined as regions of 3-5 CNVs were corrected separately from regions defined as regions of 2-5, 4-5, or 5 CNVs.

2.6.11 Gene Ontology (GO) term analysis

We compiled a list of genes with genomic coordinates overlapping each CNV and separated these lists by individual and, where applicable, cell type. These lists of CNV-affected genes were submitted to PANTHER (Mi et al., 2013) (pantherdb.org) to determine if any GO terms were enriched. The resulting GO terms and corresponding p values were then submitted to REViGO (Supek et al., 2011) (revigo.irb.hr) to aid visualization via downloadable plotting scripts.

2.6.12 Quantification and Statistical Analysis

Most statistical analyses were performed using R (version 3.4.1, www.R-project.org), utilizing base and downloaded packages. BIC score cutoffs and CNV thresholds were defined as explained in Method Details. Statistical tests and outlier detection are explained in figure legends and in Results text. Stratification of p value

correction in gene set enrichment is described above in Method Details. Statistical significance was defined as $p < 0.05$.

2.6.13 Data and Software Availability

Single cell sequencing data for PicoPLEX neurons and non-neurons is available through the NIH Data Archive (Study 636, https://dx.doi.org/10.15154/1503237). The R script used for simulating data using ALT and NULL models is available at https://github.com/mcconnell-lab/scripts/blob/master/ALT_NULL_dataset_simulation.R.

2.7 Acknowledgments

We thank J. Lannigan and M. Solga (University of Virginia [UVA] fluorescence-activated cell sorting [FACS] core), Y. Bao (UVA genome analysis and technology core), and A. Koeppel (UVA bioinformatics core) for their contributed expertise. We thank P. Lansdorp (University of British Columbia) for prompt sharing of unpublished metadata, and F.H. Gage (Salk Institute) and J.V. Moran (University of Michigan) for critical feedback. Human tissue was obtained from the National Institute for Child Health and Human Development (NIH) Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore (contract HHSN2752009000011C, ref. no. N01-HD-9-011). Data are documented through the Brain Somatic Mosaicism Network Knowledge Portal (synapse.org/BSMN) and made available at the National Institute of Mental Health (NIMH) Data Archive. NIH funding to M.J.M. (U01 MH106882), to D.R.W. (U01 MH106893), and to M.J.M. and S.B. (U01 MH106882-03S1) supported this work. I.E.B.

received support from the McDonnell Foundation, and W.D.C. received support from NIH T32 GM008136-30.

2.8 Author contributions

M.J.M. and D.R.W. designed the study. I.E.B., M.B.W., M.J.W., and M.F.H. performed WGA and sequencing. Samples were collected and selected by J.E.K., T.M.H., and D.R.W. W.D.C., S.B., and M.J.M. analyzed the data. A.C.B.S. contributed to the enrichment analysis. M.J.M., I.E.B., W.D.C., S.B., and D.R.W. wrote the manuscript.

2.9 Supplemental Figures

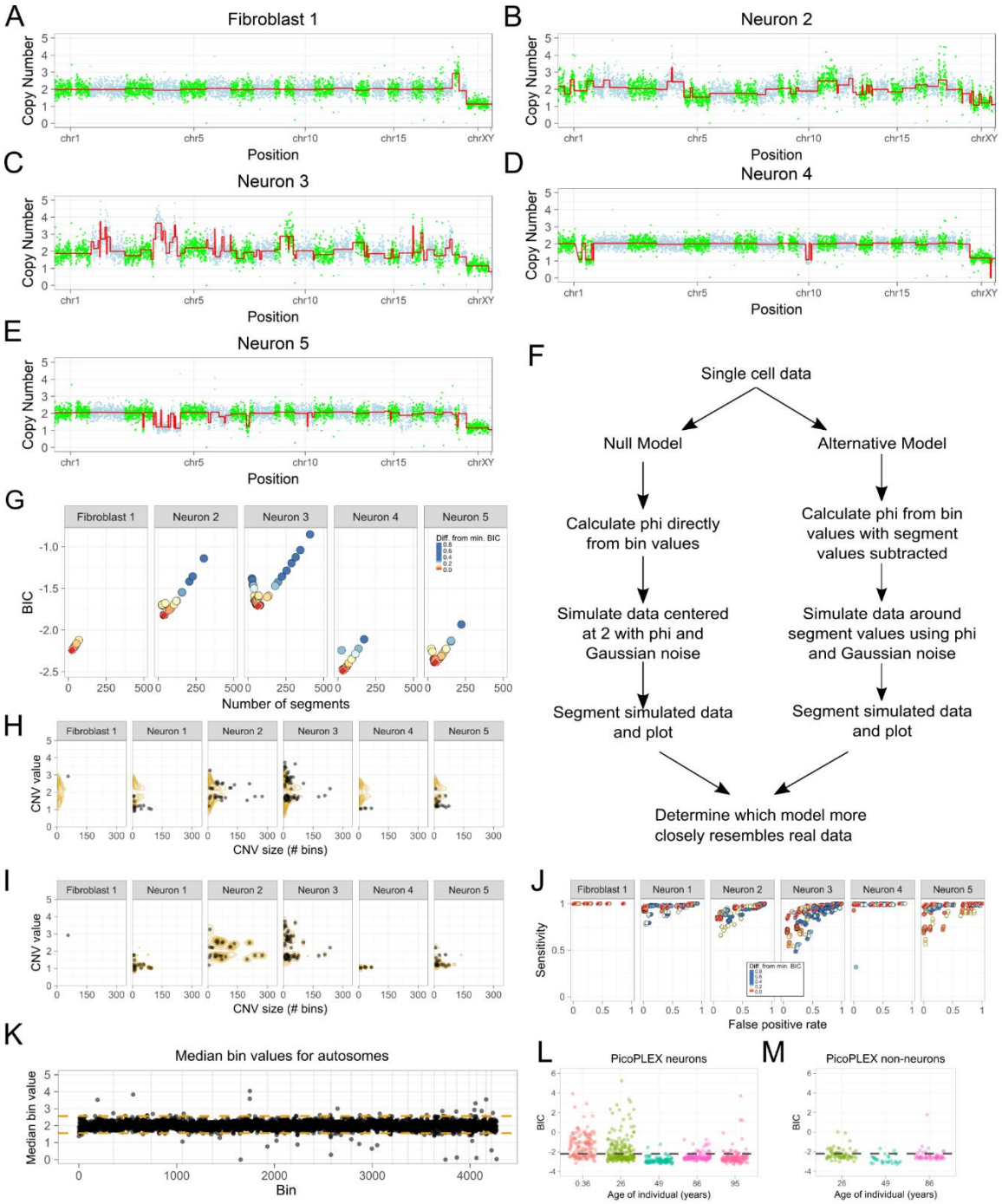


Figure 2.S1 (related to Figure 2.1). Optimization of sample filtration and segmentation
(A-E) CNV profile of test data Fibroblast **(A)**, Neuron 2,-5 **(B – E)**. Genome is arranged horizontally by chromosome. Read depth-derived CN values of genomic bins and are colored by chromosome, alternating between green and light blue. Red line represents segmentation output from DNAcopy. **(F)** Flowchart summarizing simulation of data based on authentic single cell data using the null and alternative models**.** **(G)** Comparison of BIC scores for segmentation of Fibroblast 1 and Neurons 2, 3, 4, and 5 using different settings of alpha, undo.SD, and min.width. A red diamond is used to highlight the lowest BIC score for each cell. **(H)** Comparison of CNVs detected in real data (black dots) to CNVs detected in simulated data

(yellow contours) under the null and **(I)** alternative models across six test cells. For the purposes of these tests, thresholds of CN < 1.825 for deletion and CN > 2.175 for duplication were used. Contours are based on a two-dimensional kernel density estimation of the simulated CNVs. **(J)** ROC curves demonstrating the sensitivity and false positive rates of CNV detection under different DNAcopy settings. Points are colored according to difference from minimum BIC score for each cell, as in panel G; similarly, red diamonds are used to highlight the best BIC scores for each cell. 10 CNV thresholds were used, ranging from < 1.825 for deletion and > 2.175 for duplication to < 1.99 for deletion and > 2.01 for duplication. For that reason, BIC scores from the same segmentation may appear up to 10 times in each plot depending on changes in sensitivity or false positive rate resulting from changes in the CNV thresholds used. **(K)** Median bin values for autosomal bins in PicoPLEX data. Orange dashed lines represent bin exclusion thresholds of > 2.57 and < 1.56, as determined by Tukey's outlier test**. (L)** BIC scores for PicoPLEX neurons and **(M)** non-neurons. Dashed line indicates BIC cutoff of < -2.21.

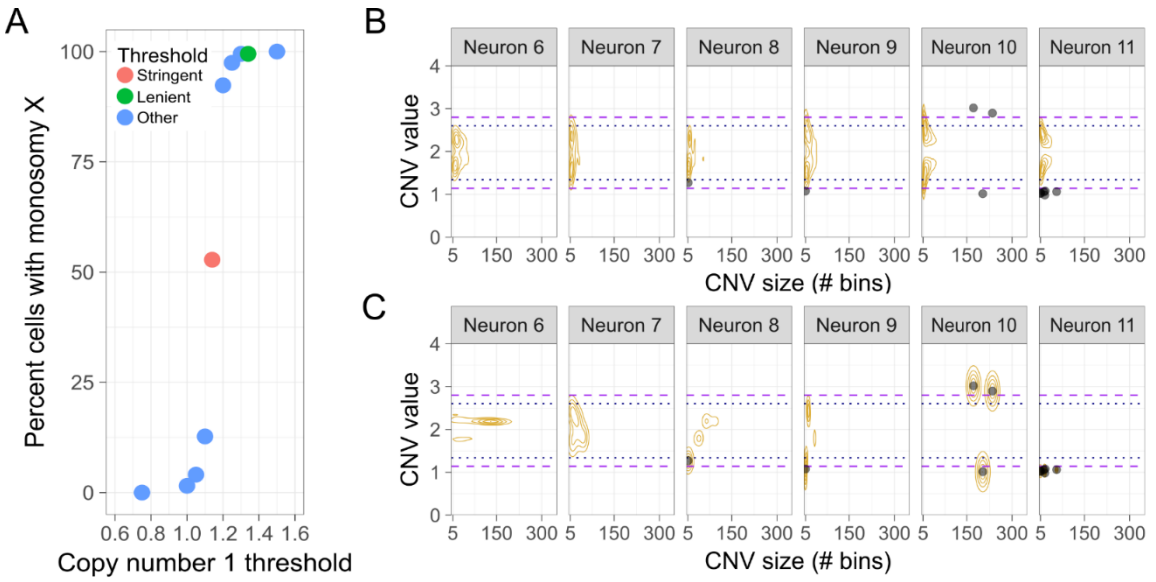


Figure 2.S2 (related to Figure 2.1). Lenient and stringent thresholds protect against false positive CNVs.
(A) Percent of cells with monosomy X called plotted against the threshold required for calling at a state of 1. In cases where chrX contained multiple segments, overall copy number was derived from an average of segment copy number values weighted by number of bins in each segment, then compared to each threshold. **(B)** Comparison of CNVs detected in real data (black dots) to CNVs detected in simulated data (yellow contours) under the null and **(C)** alternative models across six test cells. For the purposes of these tests, thresholds of CN < 1.825 for deletion and CN > 2.175 for duplication were used. Contours are based on a two-dimensional kernel density estimation of the simulated CNVs. Long dashed lines indicate stringent CNV cutoffs of < 1.14 and > 2.80, while short dashed lines indicate lenient CNV cutoffs of < 1.34 and > 2.60. Neuron 6 (BIC = -2.686, φ = 0.108) and Neuron 7 (BIC = -2.807, φ = 0.083) contained 0 CNV calls; Neuron 8 (BIC = -2.969, φ = 0.085) and Neuron 9 (BIC = -2.386, φ = 0.166) contained 1 CNV each; Neuron 10 (BIC = -2.558, φ = 0.663) contained 3 CNVs; Neuron 11 (BIC = -2.766, φ = 0.344) contained 6 CNVs.

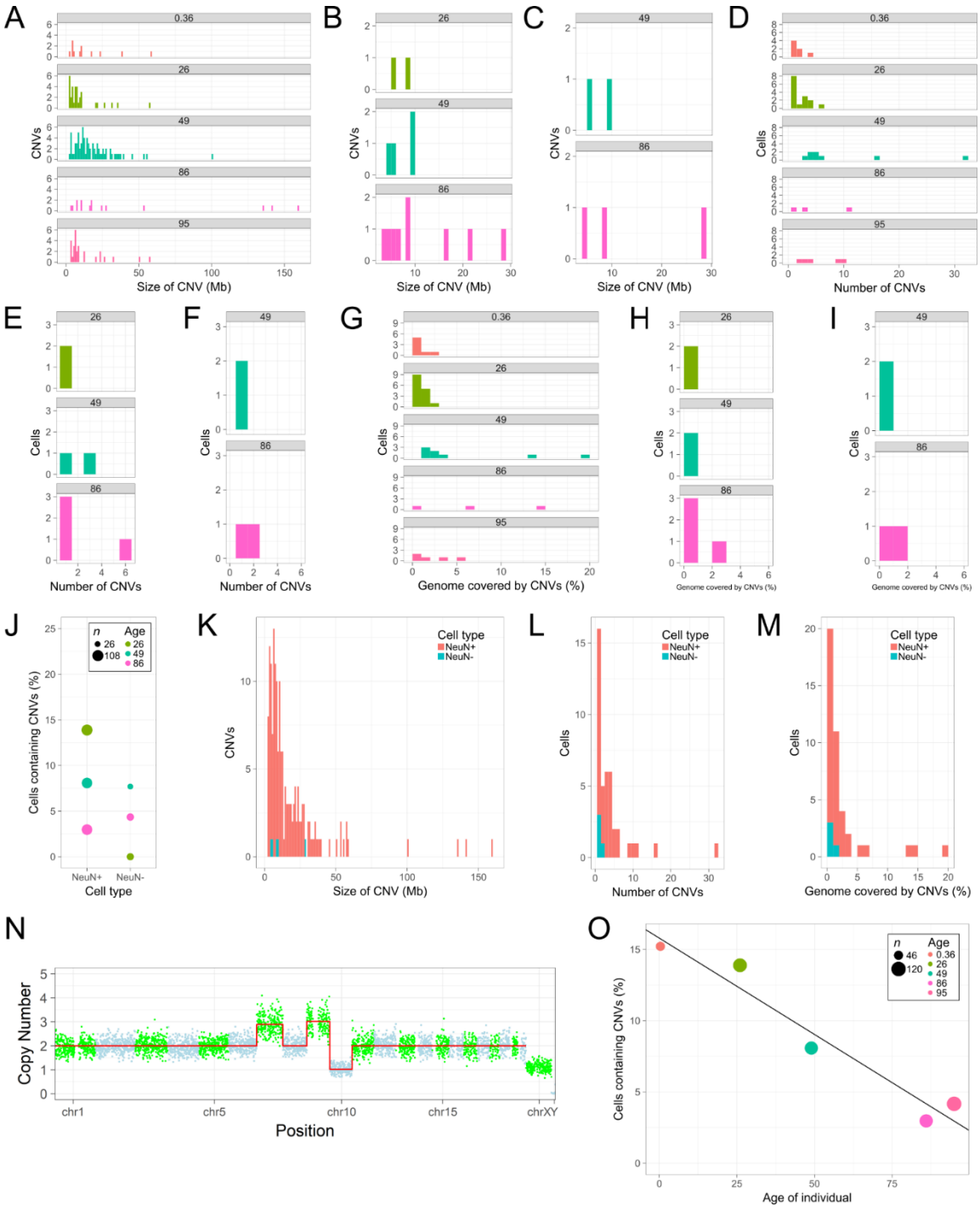


Figure 2.S3 (related to Figure 2.2). Contribution of mosaic CNVs to brain cell diversity across individuals and CNV thresholds

(A) Histogram showing CNV sizes among the neurons from five individuals aged 0.36 to 95. Stringent thresholds were used. **(B)** Histogram showing CNV sizes among the non-neurons from three individuals aged 49 to 86. CNVs were detected using lenient thresholds and **(C)** stringent thresholds. **(D)** Histogram showing number of CNVs in CNV neurons from five individuals. Neurons with 0 CNVs were excluded

from this plot. Stringent thresholds. **(E)** Histogram showing number of CNVs in CNV non-neurons from three individuals. Non-neurons with 0 CNVs were excluded from this plot. CNVs were detected using lenient thresholds and **(F)** stringent thresholds. **(G)** Histogram showing percent genome coverage by CNVs in CNV neurons from five individuals. Neurons with 0 CNVs were excluded from this plot. Stringent thresholds. **(H)** Histogram showing percent genome coverage by CNVs in CNV non-neurons from three individuals. Non-neurons with 0 CNVs were excluded from this plot. CNVs were detected using lenient thresholds and **(I)** stringent thresholds. **(J)** Comparison of percent cells containing CNVs between NeuN+ and NeuN- cells from three individuals. Stringent thresholds. **(K)** Comparison of CNV size between NeuN+ and NeuN- cells. Stringent thresholds. **(L)** Comparison of number of CNVs between NeuN+ and NeuN- cells. Cells with 0 CNVs were excluded from this plot. Stringent thresholds. **(M)** Comparison of percent genome coverage by CNVs between NeuN+ and NeuN- cells. Cells with 0 CNVs were excluded from this plot. Stringent thresholds. **(N)** CNV profile of neuron from 86 year old male displaying trisomy 7, trisomy 9, and monosomy 10. The cell also appears to show a Y chromosome loss. Red lines represent the CN states at each bin across the genome; CNV calls are represented at the precise CN determined by DNAcopy while euploid regions are depicted at exactly CN 2. **(O)** Percent of neurons containing CNVs plotted against the age of the individual from which they were collected (linear fit, $R^2 = 0.9434$, $p = 0.0058$). Stringent thresholds.

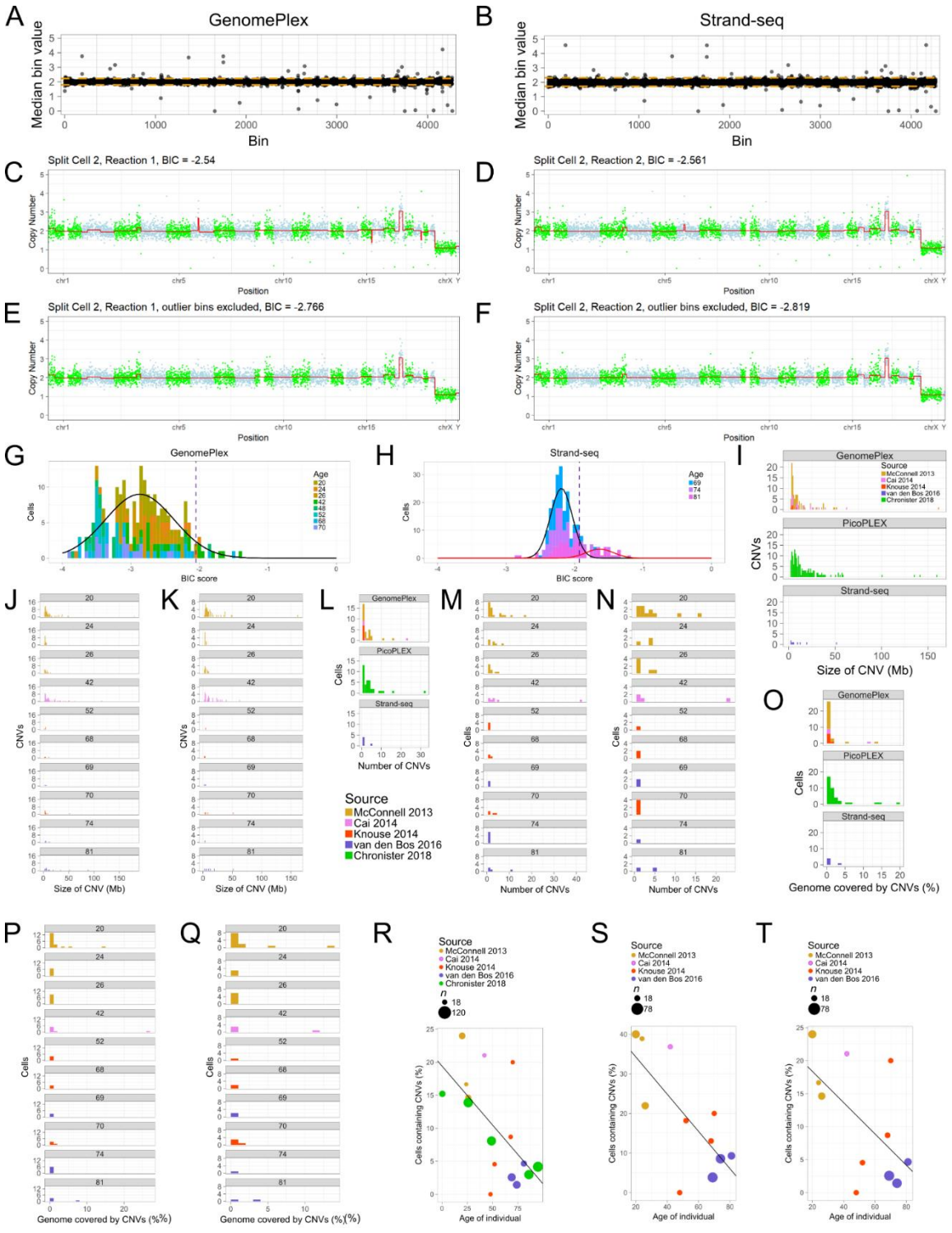


Figure 2.S4 (related to Figure 2.3). Publicly available data shows similarities to PicoPLEX data under both CNV thresholds

(A) Median bin values for autosomal bins in GenomePlex data. Orange dashed lines represent bin exclusion thresholds of > 2.23 and < 1.80, as determined by Tukey's outlier method. **(B)** Median bin values for autosomal bins in Strand-seq data. Orange dashed lines represent bin exclusion thresholds of > 2.32 and

< 1.73, as determined by Tukey's outlier method. **(C)** CNV profile of Reaction 1 of split-amplification cell from Knouse 2016 and **(D)** CNV profile of Reaction 2. GenomePlex outlier bins are included in both datasets. **(E)** CNV profile of Reaction 1 of split-amplification cell from Knouse 2016 and **(F)** CNV profile of Reaction 2. GenomePlex outlier bins were removed from both datasets. **(G)** Histogram of BIC scores for GenomePlex datasets. Gaussian distribution fit to data is depicted in black. Dashed line indicates BIC cutoff of < -2.05. **(H)** Histogram of BIC scores for Strand-seq datasets. Gaussian distributions fit to data are depicted in black and red. Dashed line indicates BIC cutoff of < -1.93. **(I)** Histogram showing CNV sizes among neurons across each WGA method and data source. Stringent thresholds were used. **(J)** Histogram showing CNV sizes in each individual from previously published data. CNVs were detected using lenient thresholds and **(K)** stringent thresholds. **(L)** Histogram showing number of CNVs in CNV neurons across each WGA method and data source. Neurons with 0 CNVs were excluded from this plot. Stringent thresholds. **(M)** Histogram showing number of CNVs in CNV neurons in each individual from previously published data. Neurons with 0 CNVs were excluded from this plot. CNVs were detected using lenient thresholds and **(N)** stringent thresholds. **(O)** Histogram showing percent genome coverage by CNVs in CNV neurons across each WGA method and data source. Neurons with 0 CNVs were excluded from this plot. Stringent thresholds. **(P)** Histogram showing percent genome coverage by CNVs in CNV neurons in each individual from previously published data. Neurons with 0 CNVs were excluded from this plot. CNVs were detected using lenient thresholds and **(Q)** stringent thresholds. **(R)** Percent of neurons containing CNVs plotted against the age of the individual from which they were collected (linear fit, $R^2 = 0.3941$, $p = 0.0092$). Stringent thresholds. **(S)** Percent of neurons containing CNVs plotted against the age of the individual using previously published data under lenient CNV thresholds (linear fit, $R^2 = 0.5315$, $p = 0.011$) and **(T)** stringent thresholds (linear fit, $R^2 = 0.3524$, $p = 0.054$).

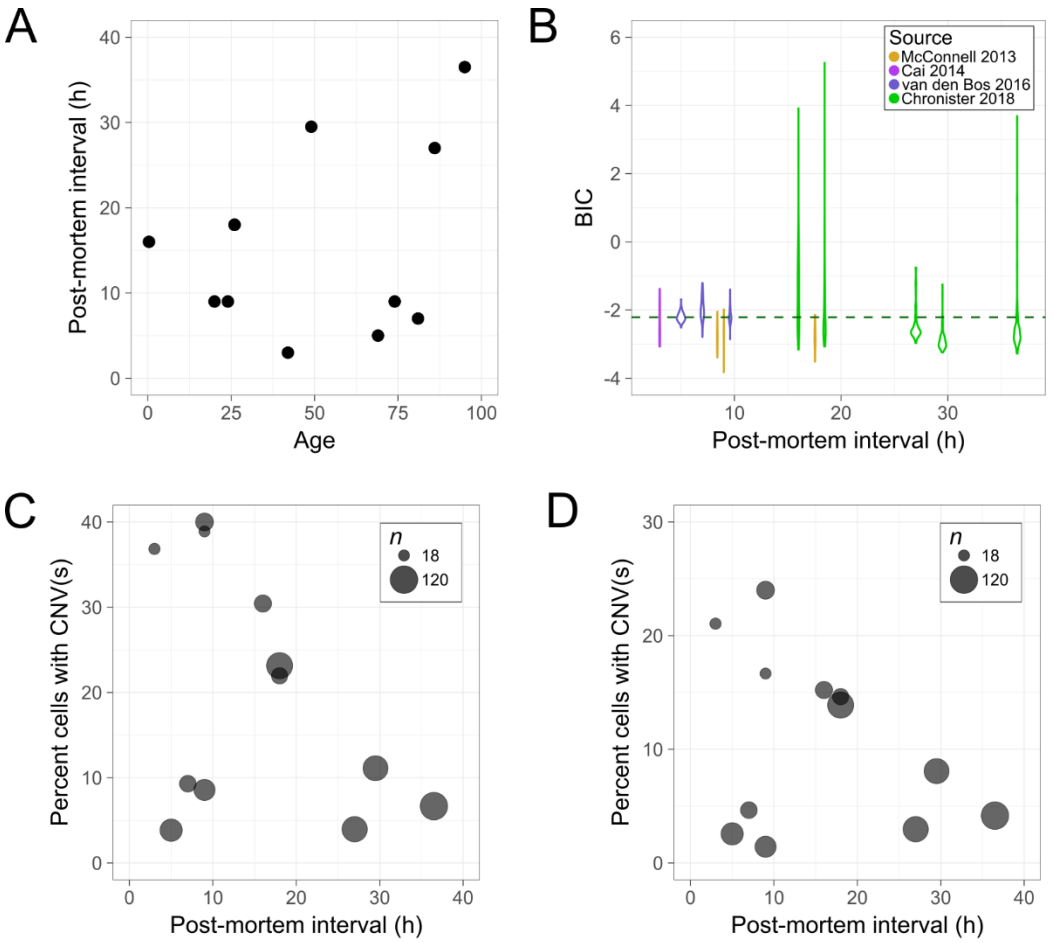


Figure 2.S5 (related to Figure 2.3). Post-mortem interval (PMI) does not impact sample quality or confound age-related trends.
(A) Scatter plot of PMI and age information for 11 individuals studied in this chapter. **(B)** Violin plots of BIC score and PMI for 11 individuals studied in this chapter. The 26 year old individual studied in both McConnell 2013 and this chapter has a PMI of 18 h and violins are shown for each group of cells analyzed. **(C)** Percent of neurons containing CNVs plotted against the PMI of the individual from which they were collected using lenient CNV thresholds and **(D)** stringent thresholds. The data points represent 11 of the individuals studied in this chapter, including the 26 year old for whom there are separate data points for the McConnell 2013 and Chronister 2018 datasets. The relationship was statistically insignificant in both cases.

Chapter 3
Exploring somatic mosaicism in the mouse brain

The Top1$^{-/-}$ and Top1$^{+/+}$ mouse analyses discussed in this chapter are part of a submitted manuscript currently under review at Nature Communications (Fragola et al., 2019). Figure 3.1 D-E were taken from a figure in that manuscript.

3.1 Abstract

Efforts to study brain somatic mosaicism in healthy human beings are hindered by the difficulty of obtaining disease-free brain tissue and the considerable genomic diversity of human individuals. Fortunately, brain somatic mosaicism is a conserved phenomenon in mice, which allow for easier experimentation. To probe the factors influencing CNVs in the mouse brain, we sequenced over 800 single neuronal nuclei from 10 mice, aged between 1 week and nearly 3 years of age. Several of these mice were bred to carry mutant genotypes hypothesized to influence CNV prevalence, including *Top1* knockout, *Trp53* knockout, and human tau (*htau*) transgene. We analyzed these single cell genomes through our custom CNV detection pipeline and found rates of CNV-containing neurons in the range of ~30-70%. We also created "bulk" data through the pooling of sequencing data from individual mice, which allowed us to identify clonal CNVs using structural variant callers.

3.2 Introduction

Mice have long been used as a model system to investigate research questions that cannot be as easily explored in human subjects. In recent years, this approach has been employed to study brain somatic mosaicism. Whereas human brain samples are scarce compared to other tissues and, as such, provide investigators with little flexibility to control for critical characteristics such as ancestry, environmental effects, disease state, and age, mouse brains are abundant and can be collected from individuals of identical genetic makeup, from the same controlled laboratory environment, with any of a variety of disease states, and at any desired time points in the development or aging process.

With regard to modeling brain somatic mosaicism, perhaps the most important aspect of mice has been that mouse neurons indeed show similar kinds of mosaicism to humans, thus making them a highly relevant model system.

Over the past two decades, various studies have identified the presence of somatic mosaicism in the mouse brain. Rehen et al. (2001) were the first to report evidence of widespread somatic mosaicism in the mouse nervous system, reporting that 33% of neuroblasts examined using spectral karyotyping (SKY) were aneuploid. They also performed DNA-fluorescence in situ hybridization (DNA-FISH) on the X and Y chromosomes in male cortical cells and reported aneusomy rates of 1.15% in the adult brain and 6.74% in the embryonic brain. Only two published studies to date have used single cell whole genome sequencing (scWGS) to explore mosaicism in the mouse brain: Knouse et al. (2014) identified evidence of aneuploidy in 1% of adult neurons and in 0% of embryonic neural progenitor cells (NPC). Meanwhile, Rohrback et al. (2018a) reported that CNVs as small as 250kb were detected in over 93% of embryonic neural progenitor cells and adult neurons, but detected whole-chromosome losses or gains in <1% of cells. Given the contradictory estimates of aneuploidy depending on the method employed, it has been proposed that the high rates of aneuploidy detected by SKY and FISH techniques may be accounted for by subchromosomal gains and losses, including chromothriptic events, which may be misidentified as aneuploidies by SKY and FISH (Knouse et al., 2014; Rohrback et al., 2018a; Rohrback et al., 2018b). In any case, the extent to which whole- and sub-chromosomal CNVs affect the mouse brain remains an open question and warrants further study.

We analyzed neurons from a variety of different mice to learn how brain somatic mosaicism manifests in mice of diverse genotypes. In our first experiment, we examined the effects of the presence or absence of topoisomerase 1 on large-scale CNV prevalence by studying 2 mice, 1 *Top1*$^{-/-}$ and 1 *Top1*$^{+/+}$. Top1 assists the expression of long genes by relieving torsional stress caused by DNA supercoiling, an important role in the brain, which utilizes a disproportionately high number of long genes compared to other tissues (King et al., 2013; Mabb et al., 2016; Zylka et al., 2015). Previous studies have shown that Top1 inhibition leads to downregulation of long neural genes and impairment of synaptic function in neuronal cultures (King et al., 2013; Mabb et al., 2014; Mabb et al., 2016). Given its role in mitigating transcriptional stress, a known cause of somatic mutations in neurons (Lodato et al., 2015), we hypothesized that the absence of Top1 would cause an increase in CNVs. To test this hypothesis, we analyzed 47 *Top1*$^{-/-}$ and 48 *Top1*$^{+/+}$ single neuronal nuclei for CNVs and found that loss of *Top1* resulted in an increase in both the number of neurons containing CNVs ("CNV neurons") and the size of CNVs.

We also analyzed a group of 8 mice of various genotypes selected to examine the impact of p53 absence and a human tau transgene (*htau*) on CNV occurrence. p53 (gene name *TP53* in humans, *Trp53* in mice) is a well-characterized tumor suppressor which is frequently mutated in cancer (Muller and Vousden, 2013). Given the functions of p53 in DNA damage response, which can entail the activation of DNA repair machinery or the initiation of apoptosis (Williams and Schumacher, 2016), we were interested to learn if deletion of one or both *Trp53* alleles would result in a higher or lower proportion of neurons containing CNVs compared to wild-type. We also wanted to study the impact of

htau on a mouse with its endogenous tau gene knocked out. *htau* causes mice to develop symptoms similar to those seen in various human tauopathies, including neurofibrillary tangles (NFTs) and neuronal death (Andorfer et al., 2005; Andorfer et al., 2003). Further, *htau* has been reported to induce abnormal cell cycle reentry and re-replication prior to cell death in mouse neurons (Andorfer et al., 2005). We were interested to see if *htau* mouse neurons carried CNVs that were in any way selected for by neurodegenerative disease; in addition, we were curious as to whether any neurons we analyzed would show evidence of cell cycle reentry in the form of duplication CNVs or hyperploidy. As a control, we also sequenced cells from a mouse without *htau* or the endogenous mouse *tau* gene. Our dataset thus included 2 $Trp53^{-/-}$ mice, 1 $Trp53^{+/-}$ mouse, 1 $htau^{+}$ $mtau^{-/-}$ mouse, 1 $htau^{-}$ $mtau^{-/-}$ mouse, and a total of 3 wild-type mice. Among these individuals, 2 wild-type mice and the $htau^{+}$ mouse were well over 2 years of age, enabling us to also examine CNV incidence in aged mice, though low sample sizes and lack of control mice in each genotype prevented direct analysis of aging effects. By carrying out a similar CNV analysis to the pipeline outlined in Chapter 2, we detected percentages of CNV neurons in the range of 30-70% across the 8 individuals.

Noticing that CNVs were more frequent than expected, we also employed a novel approach of pooling single cell data to create effective bulk data for each of the 8 mice in our second dataset. The creation of bulk data allowed us to run variant calling tools LUMPY, Delly, and SVXplorer to detect clonal CNVs in each individual. We detected highly variable numbers of CNVs across individuals and assessed the level of concordance of deletions called by these tools. We then attempted to verify the presence of deletions using in-depth analysis of individual mapped reads.

3.3 Methods

3.3.1 Analysis of Top1 cKO and WT mice

Top1 cKO and WT mice were generated as previously described (Mabb et al., 2016). Cortices were isolated from P7 brains of one *Top1*$^{+/+}$ mouse and one *Top1*$^{-/-}$ mouse and stored at -80^{o}C. Following mechanical dissociation of brain tissue and purification of nuclei, as described in Wierman et al. (2017), FACS was used to isolate NeuN^{+} nuclei as described in Chapter 2. Whole genome amplification was carried out using a customized MALBAC protocol (Burbulis et al., 2018), and libraries were barcoded and prepared for sequencing by NextFLEX (Bioo Scientific). Whole genome, single-end 100bp sequencing was carried out on the Illumina HiSeq Rapid platform. The resulting sequencing data from 95 cells (48 Top1+/+, 47 Top1-/-) was trimmed of MALBAC adapters using FASTX Toolkit (version 0.0.13) prior to being analyzed using a CNV detection pipeline similar to the one detailed in Chapter 2, with exceptions as noted here. Reads were counted in 5,126 non-overlapping bins by Bedtools (version 2.17.0). Bins were designed to contain 500kb of mappable sequence each (average bin size = 518 kb) using the 40mer mappability track for mm9 (wgEncodeCrgMapabilityAlign40mer.bigWig, available from UCSC). Binned reads were normalized to copy number estimates following the GC-correction method detailed in Chapter 2. Using Tukey's Outlier Method on the median values for each of the 5,126 genomic bins, we identified 182 outlier bins that consistently showed higher or lower than normal copy number estimates and excluded them from segmentation by DNAcopy (version 1.50.1). A BIC cutoff was set to exclude poor quality cells, excluding all cells scoring higher than -1.5. Copy number values of segments between 5 and 54 bins in size

were plotted as a histogram, and three Gaussian distributions were fit to the data using the R package mixtools (version 1.1.0) to identify peaks corresponding to 1-copy, 2-copy, and 3-copy regions of DNA. Using a two-tailed p-value of 0.001 on the distribution centered near 2, a deletion threshold of < 1.22 and a duplication threshold of > 2.82 were calculated.

3.3.2 Analysis of 8 mice

3.3.2.1 Traditional single cell CNV analysis

Our 8 mice (Table 3.1) were obtained from several sources. p53KOmouse1 and p53KOmouse2 were obtained from the Jackson Laboratory (strain B6.129S2). p53hetmouse and WTmouse1 were obtained from the laboratory of Noelle Dwyer (University of Virginia) and were bred to an approximate background of 70% C57BL/6 and 30% FVB. htau, htauKO, WTmouse2 and WTmouse3 were obtained from the laboratory of George Bloom (University of Virginia); htau and htauKO mice were bred as previously described (Dawson et al., 2001; Duff et al., 2000). Cortices were gathered and frozen in the same fashion as the *Top1* cKO and WT mice. FACS and MALBAC WGA were carried out as described in 3.3.1. Sequencing libraries were barcoded and prepared for Illumina sequencing by iGenomX. Paired-end 75bp sequencing was performed on the Illumina NextSeq platform to generate single cell data. 721 neurons contained sufficient read counts to be analyzed in our pipeline. We followed a similar single cell CNV detection pipeline to the one described for the Top1 data, with exceptions detailed here. We identified 166 outlier bins using Tukey's Outlier Method and excluded them from DNAcopy segmentation. BIC scores were calculated and plotted

as a histogram, to which we fit two Gaussian distributions. The cutoff was set at the 95th percentile score (-1.268), excluding all cells with a higher score. Lenient and stringent CNV cutoffs calculated in Chapter 2 with human data were utilized to detect CNVs.

3.3.2.2 Identifying clonal mutations using pooled single cell data

We then pooled the single cell data from each individual to generate 8 sets of "bulk" data. The pooling was accomplished by simple concatenation of all trimmed FASTQ files. The 8 bulk FASTQs were then run through the beginning steps of the CNV pipeline to generate BAM files. BAM files from each individual were analyzed by Lumpy (and genotyped by SVTyper), Delly, and SVXplorer according to the default protocols.

We created merged BAM files for each individual by using the Samtools (version 1.1) "merge" command on our single cell BAMs in order to preserve the read groups of each cell. The presence of these read group tags enabled us to identify which reads belonged to which cells when viewing in the Integrative Genomics Viewer (IGV, Broad Institute).

Clipped reads were isolated from regions 1-2bp upstream and downstream of putative deletions using Samtools. We submitted their sequences to Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) for clustering and viewed the resulting alignments in MView.

3.4 Results

3.4.1 Exploring effects of Top1 knockout on CNV incidence

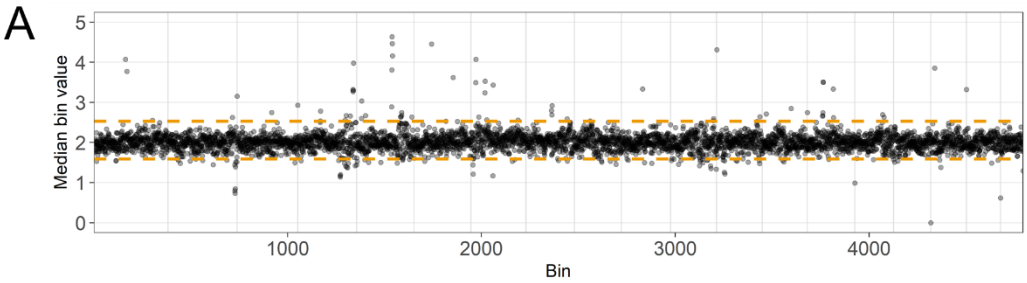
We analyzed sequencing data from 47 Top1 cKO neurons and 48 Top1 WT neurons using the same pipeline as in Chapter 2, with adjustments made for the mouse data. 183 outlier bins were identified using Tukey's Outlier Method on median bin values from the 95 cKO and WT neurons (Figure 3.1A). Following outlier bin removal, each cell was segmented using DNAcopy and scored using BIC. The average BIC scores were -2.14 (cKO) and -1.72 (WT). Median BIC scores were -2.28 (cKO) and -1.81 (WT). A cutoff was set at -1.5 for both groups, which resulted in 43/47 (91.5%) cKO neurons and 39/48 (81.3%) WT neurons being retained for downstream analysis (Figure 3.1B).

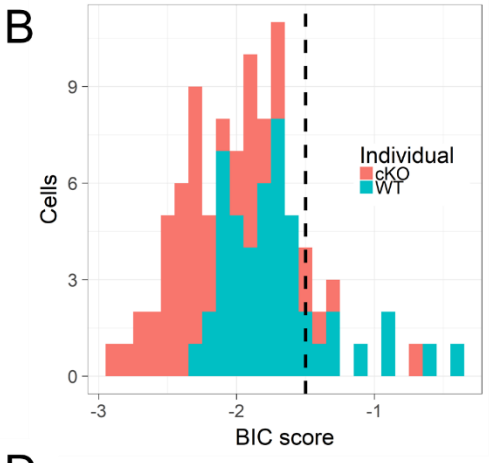
We plotted a histogram of DNAcopy segments from BIC-passing cells between 5 and 54 bins in size, which consisted of the shortest 20% of segments. Using the R package mixtools, we fit three Gaussian distributions which were centered near integer copy number states of 1 (1.24), 2 (2.02), and 3 (2.81) (Figure 3.1C). Using the distribution centered near 2, we calculated a two-tailed p-value of 0.01 to arrive at a deletion threshold of 1.22 and a duplication threshold of 2.82.

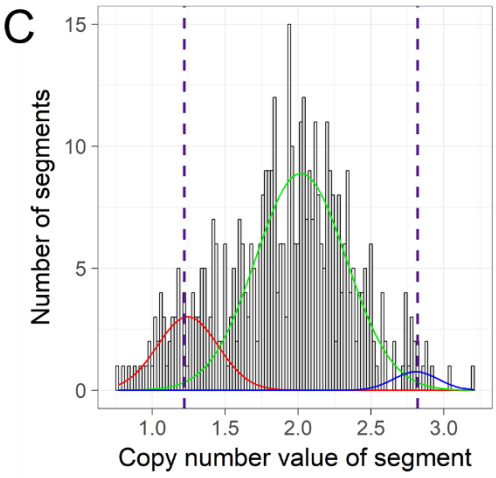
Using these CNV thresholds, we detected at least one deletion or duplication in 18 of 82 BIC-passing cells (22.0%). In total, 56 CNVs were identified (3.1 CNVs/neuron). 41 of the CNVs were detected in 14 Top1 cKO neurons (2.9 CNVs/neuron) and 15 were found in 4 Top1 WT neurons (3.8 CNVs/neuron), representing CNV neuron subpopulations of 14/43 (32.6%) in cKO cells and 4/39 (10.3%) in WT cells (Figure 3.1D-E). The mean size of CNVs in cKO neurons (15.6 Mb) was more than double that

of WT neurons (7.1 Mb), and this difference was determined to be statistically significant ($p = 0.01508$, two-sided student's t-test) (Figure 3.1F).

As previously observed in other datasets (Chronister et al., 2019; McConnell et al., 2013; Rohrback et al., 2018a), deletions were more common than duplications (80.4% to 19.6%). A similar ratio was found in cKO (78.0% to 22.0%) and WT (86.7% to 13.3%) neurons alike.

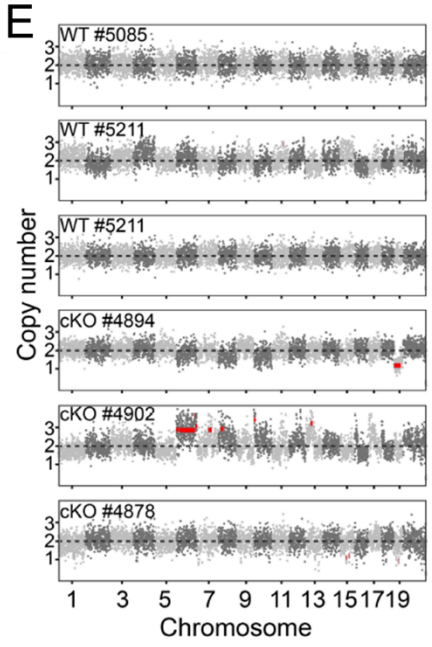






D

	WT	cKO
neurons with CNV	4/39	14/43
duplications	2/15	9/41
deletions	13/15	32/41
min-max (Mb)	2.5-14.8	2.5-111.6
mean (Mb)	7.1	15.6



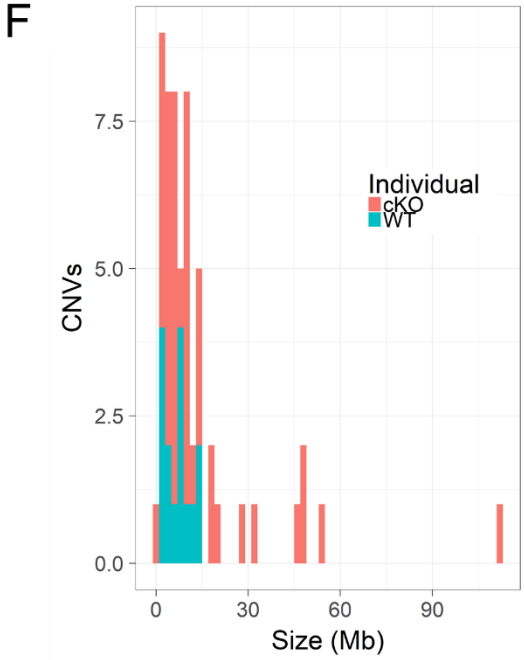


Figure 3.1: Knockout of *Top1* in neurons results in a higher proportion of CNV neurons and longer CNVs than wild-type neurons. (A) Plot showing median bin values for each bin across mouse genome. Orange dashed lines indicate upper and lower cutoffs of 2.53 and 1.59. **(B)** Histogram of BIC scores for *Top1* cKO and WT neurons. The dashed line indicates the maximum passing score of -1.50. **(C)** Histogram of segments between 5 and 54 bins in size. The red distribution is centered near the copy number 1 state, green is centered near 2, and blue is centered near 3. The dashed lines indicate the deletion threshold (1.22) and duplication threshold (2.82) used for Top1 cKO and WT neurons. **(D)** Summary table of CNV statistics for cKO and WT neurons. **(E)** Representative CNV profiles of 3 WT and 3 cKO neurons. CNVs are shown as red line segments. **(F)** Histogram showing the significant increase in CNV length in cKO cells.

3.4.2 Exploring somatic copy number variations across age, genotype, and background

We obtained brain samples from 8 mice with a variety of ages, genotypes, and genetic backgrounds (Table 3.1). Following isolation of nuclei, WGA, and library preparation, a total of 721 neurons were sequenced and analyzed for CNVs. Each single cell dataset was analyzed through the pipeline used for human data in Chapter 2 but modified for mouse sequencing data. 166 bins were identified as outliers using Tukey's Outlier Method and excluded from segmentation (Figure 3.2A). Following outlier bin removal and segmentation, BIC scores were calculated for each cell. These scores were plotted as a histogram, to which two Gaussian distributions were fitted, one centered on the bulk of the data and the other covering the long tail of poor quality cells. Using an upper-tailed p-value of 0.05 for the main Gaussian distribution, a quality threshold was set at -1.268 (Figure 3.2B). Cells scoring higher were excluded from downstream analysis, leaving 585/721 (81.1%) cells which were analyzed for CNVs. The BIC results for each individual are reported in Table 3.1.

Table 3.1 BIC statistics across individuals

Individual	Age (weeks)	Background	Cells passing BIC (%)	Mean BIC
htau	152	C57BL/6	82/98 (83.7%)	-1.800
htauKO	7	C57BL/6	19/23 (82.6%)	-1.378
p53hetmouse	8	FVB & C57BL/6	56/101 (55.4%)	-1.543
p53KOmouse1	8.5	B6.129S2	41/48 (85.4%)	-1.692
p53KOmouse2	12	B6.129S2	91/106 (85.8%)	-1.561
WTmouse1	11	FVB & C57BL/6	126/141 (89.4%)	-1.688
WTmouse2	123	C57BL/6	107/132 (81.1%)	-1.855
WTmouse3	123	C57BL/6	63/72 (87.5%)	-1.510

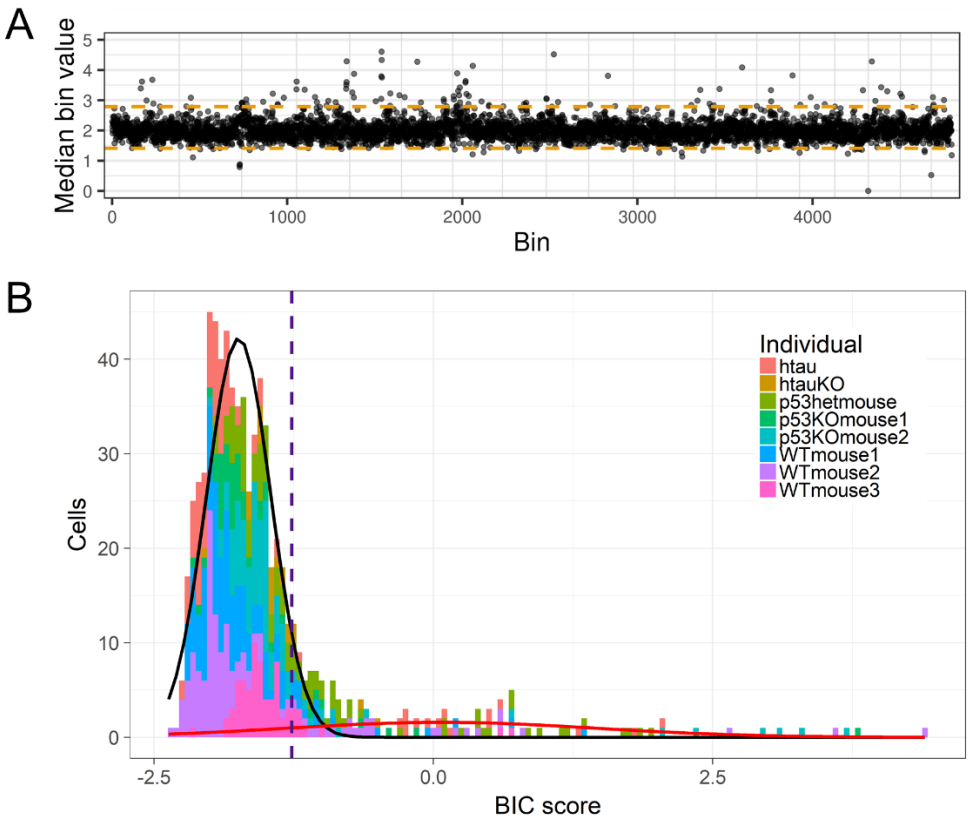


Figure 3.2: Outlier bins and a BIC cutoff were identified using data from all 8 mice. (A) Plot of median normalized bin values in autosomal bins. Orange dashed lines indicate upper and lower cutoffs of 2.79 and 1.41. **(B)** Histogram of BIC scores from 721 mouse cells. Gaussian distributions (black and red) were used to determine the cutoff of -1.268 (dashed line).

Lenient and stringent thresholds calculated in the human single cell analysis (see Chapter 2) were employed to identify deletion and duplication CNVs in our mouse cells (Table 3.2). Under lenient thresholds, the proportion of neurons containing at least one CNV varied from 30.4 to 69.4% by individual (dataset mean = 59.7%). Under stringent

thresholds, the proportion dropped to 13.0-56.9% (dataset mean 37.8%). The CNV neuron percentages did not show an age-related decline as our human analysis did; however, this lack of an aging effect may be attributable to confounding effects of mutations and transgenes.

Table 3.2 CNV results across individuals

Individual	% Cells with CNVs, lenient	Number of CNVs, lenient	Mean CNV size, lenient (Mb)	% Cells with CNVs, stringent	Number of CNVs, stringent	Mean CNV size, stringent (Mb)
htau	35.7	106	13.1	20.4	57	16.1
htauKO	30.4	23	17.0	13.0	9	12.8
p53hetmouse	34.7	74	9.5	22.8	41	6.5
p53KOmouse1	50.0	53	35.5	22.9	21	9.5
p53KOmouse2	67.9	182	7.6	53.8	99	5.9
WTmouse1	48.2	129	8.5	24.1	52	6.6
WTmouse2	43.9	120	6.9	24.2	50	4.4
WTmouse3	69.4	137	7.7	56.9	92	4.8

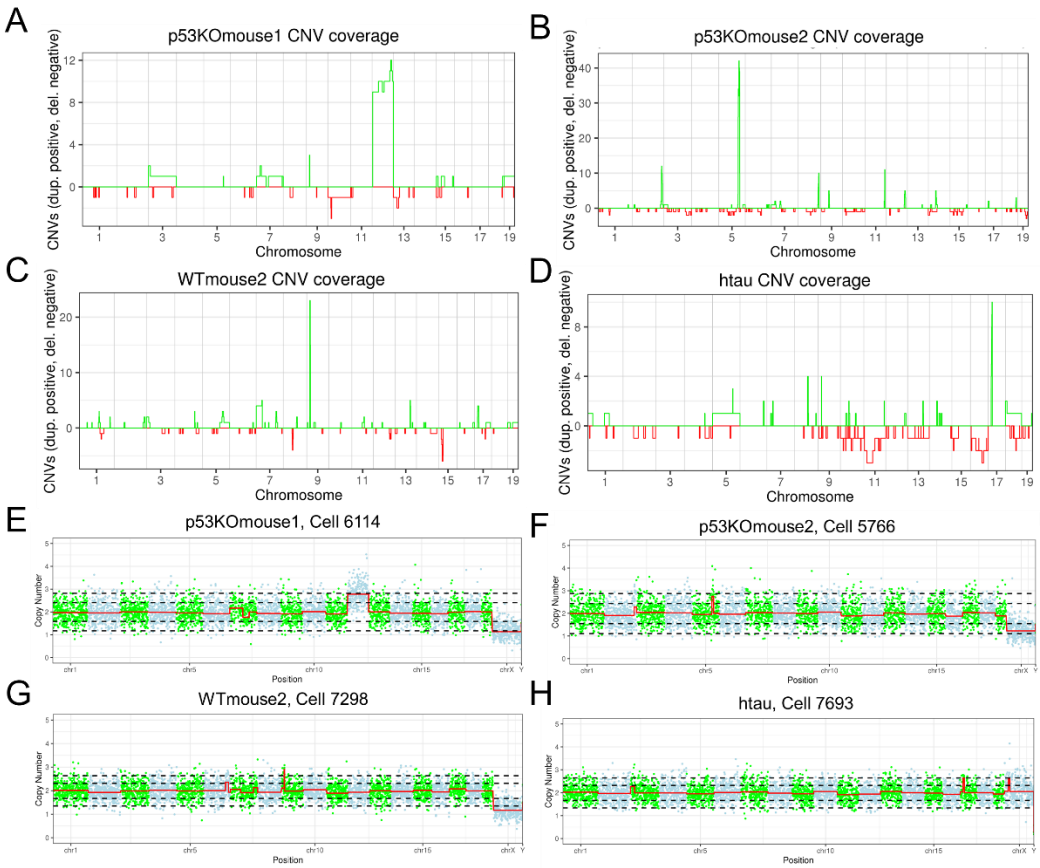


Figure 3.3: Individual mice show unique CNV hotspots. (A-D) Plots showing the cumulative CNVs detected in **(A)** p53KOmouse1, **(B)** p53KOmouse2, **(C)** WTmouse2, and **(D)** htau. Duplications are shown

in green and deletions in red. **(E-H)** Representative CNV profiles of single neurons containing examples of clonal duplications in **(E)** chromosome 12, **(F)** chromosome 5, **(G)** chromosome 9, and **(H)** chromosome 17. DNA copy output is shown in red while the black dashed lines indicate 1 and 2 median absolute deviations above and below copy number 2.

With regard to the specific genotypes of the individuals, the statistics of CNV neuron prevalence generally did not indicate any obvious biological causes; for example, although the two p53 KO mice were found to have 50-67.9% CNV neurons under lenient thresholds, the three wild-type mice were found to have very similar proportions of CNV neurons (43.9-69.4%). The htau mouse showed a modest 5-7% increase in CNV neurons compared to its control, htauKO, but the sample size for htauKO was also the lowest of the 8 mice with only 19 cells passing BIC.

Several individuals showed evidence of possible clonal duplications, which was a surprising finding due to the fact that reports of clonal CNVs have been scarce (Cai et al., 2014). In p53KOmouse1, trisomy of chromosome 12 was detected under lenient thresholds in 9 of 41 BIC-passing cells (22.0%; Figure 3.3A, E). Meanwhile, 42 of 91 p53KOmouse2 neurons (46.2%) showed a duplication located at chromosome 5:116,255,119-117,755,119 (Figure 3.3B, F). In WTmouse2, 23 of 107 cells (21.5%) contained a duplication at chromosome 9:17,555,807-19,555,807 (Figure 3.3C, G). Finally, 10 of 82 htau neurons (12.2%) showed a duplication in a region spanning chromosome 17:20,706,388-22,710,481 (Figure 3.3D, H).

3.4.3 Deeper investigation suggests results likely include false positive CNVs

During this extensive analysis, we noticed several indications that our analysis could be identifying artifactual CNVs due to subpar data quality. For instance, the dataset-wide percentages of neurons containing CNVs of 59.7% (lenient thresholds) and

37.8% (stringent thresholds) greatly surpassed our expectations based on published reports of mosaicism in the mouse brain. The most extensive published CNV analysis of mouse brain scWGS data, Rohrback et al. (2018a), found that 93% of NPCs and neurons contained at least one CNV, but only 12% of these were larger than 3.5 Mb. Given these statistics, we would expect that we would observe CNVs 3.5 Mb in size or larger in 11.2% of neurons (93% x 12%). When we excluded CNVs under 3.5 Mb, we still detected CNVs in 42.6% of cells using lenient thresholds and 21.0% of cells using stringent thresholds.

Another statistic that contrasted with published reports was the percentage of duplication CNVs; we observed 58.3% under lenient thresholds and 61.3% under stringent thresholds, both of which were considerably higher than the 28.5% observed in our human dataset under lenient thresholds and the 19.6% found in the *Top1* WT and cKO neuron dataset. Rohrback et al. (2018a) reported that duplications constituted ~20% of their detected CNVs.

We derived the BIC cutoff of -1.268 in the same fashion as our human BIC cutoffs, but it was possible to see by visual inspection of CNV profiles that the mouse neurons passing the BIC filter were not of the same quality as the human cells, for which cutoffs of -2.21, -2.05, and -1.93 had been employed, depending on WGA method used. We also noted that the BIC scores in the BIC-passing cells of the 8 mouse dataset (mean = -1.785) were worse than the *Top1* cKO and WT neuron average among BIC-passers (-2.06).

The mean phi across the 8 mouse dataset was 0.2262, which was 62% higher than the human dataset average in Chapter 2 (0.1394). Given the tendency of DNAcopy to

detect false positive CNVs in data containing high levels of autocorrelated noise (Muggeo and Adelfio, 2011), there was reason to suspect that the amount of CNV neurons we detected may be inflated by false positive CNVs driven by autocorrelated noise. To test this, we simulated 200 cells of six representative mouse neurons under the null model (see Chapter 2), which produced false positive CNVs in 49% of cells, where human simulations had only produced 3%. Similarly, alt simulations (see Chapter 2) produced false positive CNVs in 14.5% of cells, whereas human simulations only contained ~1%.

Taken together, these findings casted doubt on the authenticity of the CNVs detected by our pipeline. For that reason, we opted to use a different approach to utilize the data we had generated.

3.4.4 Performing bulk sequencing analysis from pooled single cell data

To avoid the problem of false positive CNVs in our single cell data, we sought an alternative analysis to utilize the large dataset we had generated. Given the availability of many open source structural variant (SV) callers, which run on bulk data, we decided to combine our scWGS data to reach sufficient coverage for SV detection. We concatenated the FASTQ files from each of the 8 individuals and aligned the pooled reads to the mm9 genome to generate a BAM file. Following this step, we ran our 8 pooled samples through 3 structural variant callers: LUMPY (and its genotyper, SVTyper), Delly, and SVXplorer (Chiang et al., 2015; Kathuria and Ratan, 2018; Layer et al., 2014; Rausch et al., 2012). With an eye towards the goal of validating deletion calls via PCR (Figure 3.4), we chose to focus on the deletion variants detected in our data. The number of deletions

called by each tool for each individual, as well as the number of consensus deletions, are listed in Table 3.3.

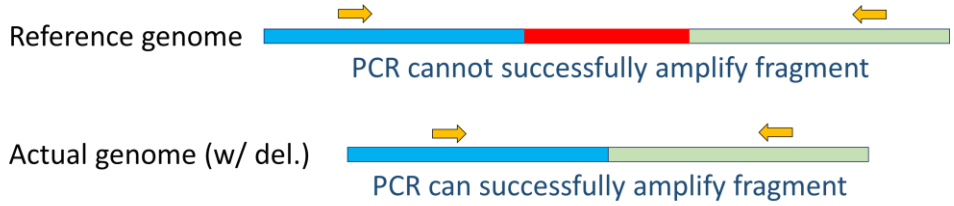


Figure 3.4: Schematic of PCR amplification of a deletion-spanning fragment to validate a putative deletion. The PCR primers (orange arrows) are located on either side of a deletion such that, in an unaltered genome (Reference genome), the fragment will not be amplified by PCR. If the putative deletion is, in fact, deleted (Actual genome), the primers will be located close enough to enable efficient amplification by PCR.

One benefit of using bulk data was the dramatic improvement in resolution, reflected in the median deletion size of around 8 kb. While bulk data prevented us from identifying CNVs in small subpopulations of cells, the massive increase in sequencing coverage enabled us to see copy number variations on a much smaller scale.

The number of mapped reads an individual had was a clear contributing factor in the number of deletions detected; in the case of htauKO, which had less than 20 million reads, there were a mere 57 deletions detected across the three callers, the fewest of any individual. Likewise, the individual with the most reads, WTmouse1, received by far the most deletion calls, with 5,040. However, there were some exceptions to this trend, such as WTmouse2, which, despite having only ~5 million fewer reads than WTmouse1, had a total of just 914 deletions called. We were unable to determine whether a biological or technical cause underlay this disparity.

Table 3.3 Results from three structural variant callers

Mouse	Cells analyzed	Total mapped reads	LUMPY-SVTyper deletions	DELLY deletions	SVXplorer deletions	Consensus deletions
htauKO	23	19,999,066	12	33	12	2
htau	98	176,641,994	176	387	218	68
p53hetmouse	101	110,418,410	466	722	753	322
p53KOmouse1	48	79,325,342	90	144	96	42
p53KOmouse2	106	158,738,676	187	446	156	47
WTmouse1	141	237,703,156	1341	1905	1794	998
WTmouse2	132	232,989,965	194	519	201	64
WTmouse3	72	146,070,639	84	148	97	25

We manually examined dozens of consensus deletions to determine if any would be good candidates for validation via targeted PCR amplification. One such example was found on chromosome 8 of p53KOmouse2. Through visualization of the merged p53KOmouse2 BAM in the Integrative Genomics Viewer (IGV, Broad Institute), we were able to see the presence of paired-end reads spanning the putative ~2kb deletion in some cells but not others (Figure 3.5A).

We also analyzed individual read sequences mapping on one side of a putative deletion, particularly any clipped reads (full-length sequencing reads which are shortened for purposes of stronger alignment to a locus in mm9). Using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/), we found that the clipped reads mapping on either side of a putative deletion at chromosome 6:115,972,422-115,974,818 in WTmouse1 aligned to one another (Figure 3.5B). This self-alignment indicated that while

these reads map to the reference genome in two groups located 2 kb apart from each other, the sequences are, in fact, “split reads;” that is, owing to the presence of a deletion, one portion of their sequence comes from the upstream side of a deletion junction and the other portion is from downstream of the deletion, and the location they map depends on whichever portion scores as a stronger match to the reference. While manual inspection was too labor intensive to carry out for all deletions found by our SV callers, it was helpful to partially validate the presence of a few selected deletions.

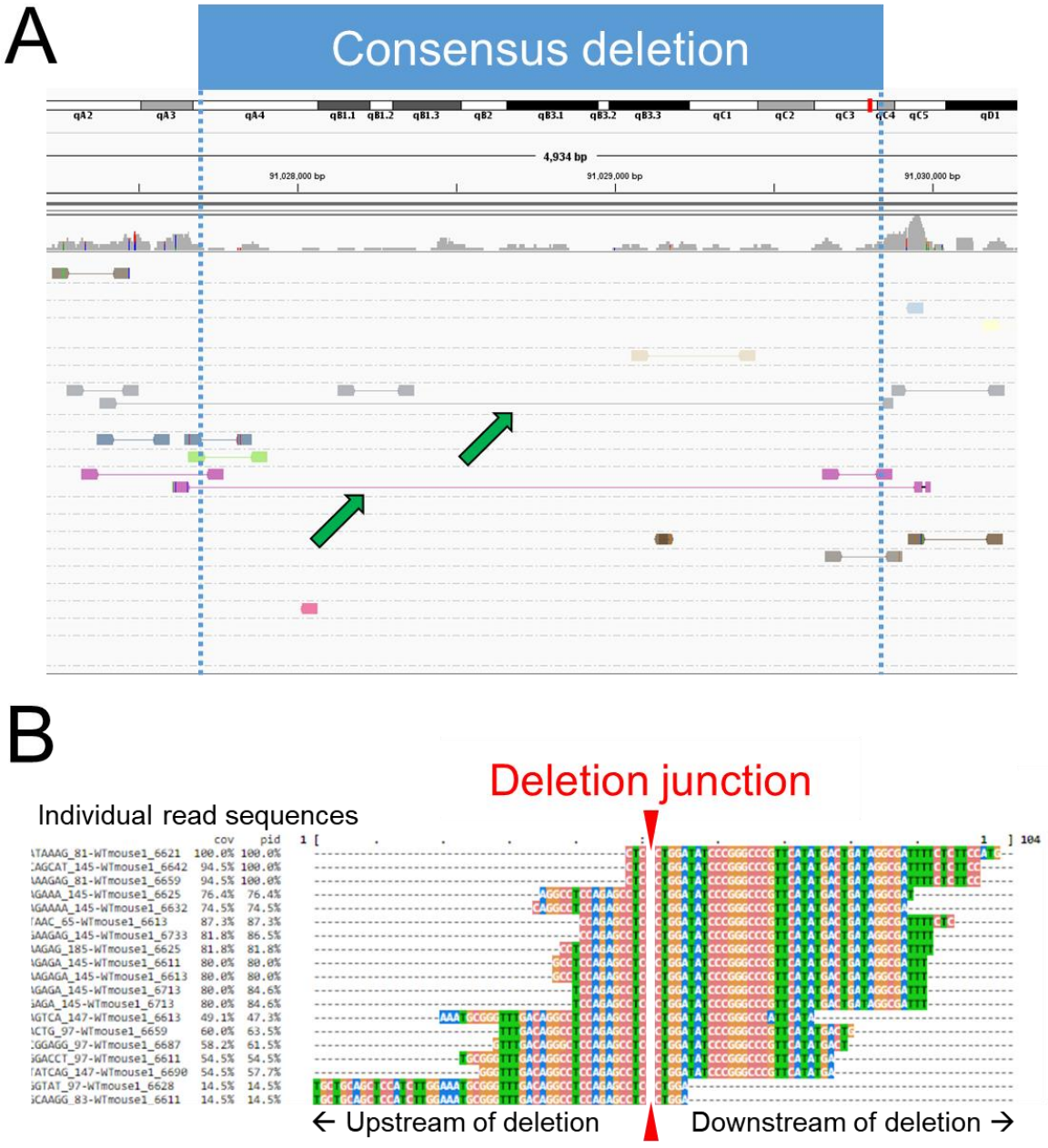


Figure 3.5: Inspection of putative deletions reveals strong candidate deletions. (A) IGV image showing reads mapped within and around a consensus deletion located on chromosome 8 in p53KOmouse2. Two sets of read pairs (indicated by green arrows) span the putative deletion (blue box and dashed lines). **(B)** Clipped read sequences that aligned on one side of the deletion junction at chromosome 6:115,972,422-115,974,818 in WTmouse1 show alignment to one another in Clustal Omega (exported to MView). This alignment indicates that the DNA template for these reads contained the deletion.

3.5 Discussion

Despite difficulties in the 8 mouse study that diminished our ability to draw conclusions from the results, our single cell analysis of 10 mice was successful in several aspects. Our study of Top1 cKO and WT neurons showed an increase in CNV incidence in cKO cells compared to wild-type. This result makes sense from a biological standpoint, given the known deleterious effects of *Top1* KO on neuronal health and function (Mabb et al., 2014). Our analysis also showed a significant increase in the size of KO CNVs over wild-type neurons. These results suggest that the absence of Top1 in neurons leads to an increase in the frequency of DNA damage instances and/or a worsening of the long-term damage to the genome resulting from DNA damage, manifested as large-scale DNA losses and gains. The mechanism of this increase in CNV neurons and CNV size remains an important question for improving our understanding of the role of Top1 in the brain.

Meanwhile, our efforts to apply our CNV detection pipeline to a panel of 8 mice of various ages and genotypes was hindered by subpar data quality. It is possible that more realistic estimates of CNV neuron prevalence in these mice could be reached by increasing the stringency of outlier bin analysis, BIC cutoff calculation, and/or CNV threshold determination. However, while the data quality issues cast doubt on findings such as the dataset-wide CNV neuron rate of ~60% and the duplication to deletion ratio of 3:2, they do not entirely account for the unique clonal CNVs we identified in four of the mice (Figure 3.3). Poor data quality alone does not explain why specific individuals so frequently carried duplication CNVs that were unique to them. It is possible that different individual mice have particular bins that amplify better than in other mice, and

these bins could account for some of the smaller clonal CNVs. To prevent this kind of false positive, one could employ an individual-specific outlier bin analysis; however, doing so would run the risk of obscuring real clonal CNVs. In any case, findings such as the 9 p53KOmouse1 neurons trisomic for chromosome 12 are difficult to dismiss as a product of technical error and are much more likely to be attributable to genomic instability caused by the homozygous deletion of *Trp53*. Further study of these mice will help clarify whether the other clonal duplications are legitimate.

In an attempt to salvage the single cell data, we found success in detecting clonal CNVs in "bulk" data derived from combining individual cells' sequence data, though we did not experimentally validate any of the CNVs. Due the disparate read counts that largely drove the number of deletions found by LUMPY-SVTyper, Delly, and SVXplorer, we were unable to interpret the number of deletions found by the 3 SV callers as a function of genotype or age; that being said, a solution we did not attempt would be to downsample reads from the most sequenced individuals in order to make individuals more comparable.

Another consequence of the reduced confidence in our CNV calls in the 8 mouse dataset was that we did not carry out gene set enrichment analysis or gene ontology (GO) analysis on the genes affected by CNVs. One option for future work would be to devise a methodology for selecting only the highest quality CNVs from the most reliable cells. This filter would limit the scope of our analysis but would enable the use of an approach similar to the one employed in Chapter 2, testing mouse CNV hotspots against a random model to determine if long genes are enriched in mouse CNVs. The results would be of particular interest because the gene lists identified by Wei et al. (2016) and King et al.

(2013) were identified in mouse cells. PANTHER analysis of affected genes in this set of high confidence CNVs would also be useful to see if any similar or different GO terms are enriched in mouse CNVs as compared to human. The benefit of our diverse mouse dataset is that the results of these analyses could be further broken down by individual to examine the potential impacts of genotype and age.

Similarly, gene set enrichment analysis could be carried out on CNVs detected in Top1 cKO and WT neurons. Identifying an enrichment of King genes in Top1 cKO neurons but not WT neurons would serve to validate the findings of King et al. (2013), who found their list of long neural genes by using topotecan to inhibit Top1. The CNVs detected in Top1 cKO and WT neurons could be also be analyzed in comparison to highly expressed genes in order to determine whether CNVs are indeed driven by expression. Furthermore, the CNV-affected genes from each genotype could be analyzed using PANTHER to identify any significant GO terms associated with Top1 absence.

Future work should also include selection of candidate deletions for validation by PCR. In particular, deletions that are present in only a subpopulation of neurons from a brain would be an outstanding proof of the concept that we began to explore here, namely that clonal mosaic mutations can be identified from pooled scWGS data. If mosaic CNVs can be successfully identified from pooled scWGS libraries, scWGS will become an even more useful technique beyond its chief purpose of studying individual genomes.

Chapter 4
Conclusions and Future Directions

Mosaicism has been observed using a variety of methods over the past several decades, but only in modern times, with the arrival of single cell analysis and next generation sequencing, has it come into clear focus. The work presented in this dissertation documents our efforts to uncover and understand the genomic diversity present in the human and mouse brains.

Our human brain CNV pipeline, presented in Chapter 2, identified CNVs in a dataset of over 1200 neural cells from 15 neurotypical individuals (Table 2.1). Over 800 of these cells were newly sequenced, representing the largest published dataset of neurons sequenced for CNV analysis. We leveraged the size of our dataset to inform identification of outlier bins, suitable BIC score cutoffs, and CNV thresholds, and tested the results against ALT and NULL simulation models to determine false positive rates (Figures 2.1, 2.S1, and 2.S2). Using this novel approach to single cell quality control, we discovered the surprising result that CNV neurons decline with age in 5 individuals, a trend strengthened by the incorporation of publicly available data from 10 additional individuals (Figures 2.2 and 2.3). Finally, we found evidence of CNV enrichment in long neural genes and neurally associated gene ontology terms, particularly in neurons from aged individuals (Figure 2.4).

We also carried out CNV analysis for over 800 mouse neurons from 10 individuals, as described in Chapter 3. CNVs were detected in neurons gathered from a genotypically diverse group of mice (Table 3.1). Analysis of $Top1^{-/-}$ mice showed an increased percentage of CNV neurons and an increase in CNV size compared to $Top1^{+/+}$, supporting the role of Top1 in maintaining genome integrity in neurons (Figure 3.1). The conclusions drawn from the analysis of our 8 mouse dataset were limited due to data

quality concerns, but the mice showed evidence of potential CNV hotspots not attributable to outlier bins (Figure 3.3). We attempted a novel approach of combining single cell datasets into “bulk” data and successfully identified thousands of CNVs using open source structural variant callers (Table 3.3); however, we did not experimentally validate any of the CNVs.

Between both species, we analyzed over 2,000 neurons across 25 individuals. Our work confirms that CNVs contribute to brain somatic mosaicism in great measure, though the role of mosaicism remains an open question. Future exploration of this field would benefit from further work focused on improving analysis and exploring biological questions with well-designed experiments.

4.1 Analytical improvements

The CNVs we detect in single cell data are only as precise as the read bins we use to detect them. Read binning is necessary in cases where read coverage is too sparse and/or uneven for direct detection of CNVs from sequencing read evidence. The read binning approach used to analyze our single cell data is by design, as it enables a larger sample size to be sequenced without losing the ability to detect megabase-scale CNVs. Yet read binning leads to inaccuracy at the edges of CNV calls; under our current pipeline, CNV start and end points are simply taken from the boundary of the bin in which they reside. In our human analysis, this means CNV coordinates may be inaccurate by hundreds of kilobases, given the average bin size of 687 kb. However, there are a variety of strategies that could be implemented to try to determine the boundaries of CNVs more accurately.

One such idea is to design smaller bins than our existing scripts use (Figure 4.1). Binning reads over regions smaller than 500kb of mappable sequence is certainly not a new concept (Cai et al., 2014; Navin et al., 2011; Rohrback et al., 2018a), but could be implemented to great effect in re-analysis of our human and mouse data. Smaller bins carry the risk of more variable copy number estimates, potentially causing false positive and false negative calls to be made. However, this risk is a hypothesis that can be tested, and the ALT simulation strategy detailed in Chapter 2 is well-suited for doing so. The experiment would be carried out as follows: Repeat the same method of counting mapped reads using Bedtools in bins designed to contain 400kb, 300kb, 200kb, and 100kb of mappable sequence; normalize read counts to copy number values, accounting for GC content; segment using DNAcopy and identify CNVs; and, finally, create simulated datasets based on the characteristics of the bin value data at each bin size to monitor changes in false positive and false negative rates. If FPRs and FNRs become prohibitively high at a particular bin size, select the smallest bin size that does not result in high numbers of FP or FN CNVs.

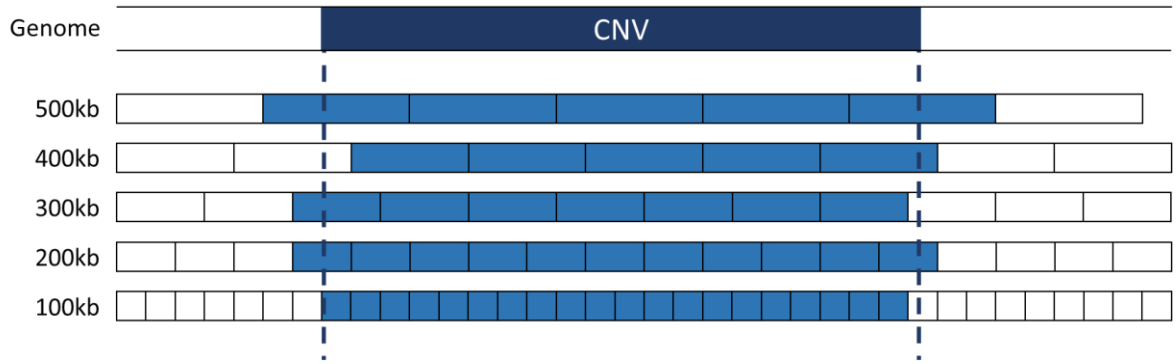


Figure 4.1: Bin sizes determine precision of CNV detection. This schematic demonstrates the hypothesis that as bin sizes decrease from 500kb to 100kb, the bins included in the CNV call by the segmentation algorithm (light blue) will be more closely aligned to the true coordinates of the CNV (navy blue).

One benefit of this technique is that it can be tailored to individual cells depending on their quality, as measured by FPR and FNR. For example, if Cell A is analyzed with 500kb and 100kb bins and shows the same CNVs each time, and if ALT simulations show that Cell A is not prone to FPs or FNs at 100kb bin resolution, the 100kb bins can be used for CNV detection in Cell A. Meanwhile, if simulations show that Cell B's data becomes unreliably noisy at 300kb resolution, the 400kb bins can be used for analyzing Cell B.

As the edges of CNVs become more sharply defined, even if only in a subset of cells, the lists of genes affected by CNVs become more accurate. In Chapter 2, we reported mild enrichment for long neural genes and some neurally relevant GO terms in CNVs. But with more precise CNV boundaries, it is possible that more significant enrichments will emerge that will provide greater insight into the role and function of mosaicism in the brain.

Another approach which has not been tried in our data is the use of uniquely mappable reads as a basis for bin coordinates. This approach, which has been used in prior CNV studies (Baslan et al., 2012; Cai et al., 2014; Navin et al., 2011), would involve setting a number of uniquely mappable reads that will constitute a bin and then determining sequential, non-overlapping regions of the genome that contain the specified number of unique read mappings. This method is a sound strategy for defining equally mappable regions of the genome wherein reads can be counted, and as such, could assist in further resolving CNV coordinates determined via the mappable sequence method. It is also plausible that this approach could lead to less inherent bin-to-bin variation compared

to mappable sequence-derived bins and potentially obviate the need for outlier bin analysis and removal.

A final analytical approach worthy of exploration is alternative forms of segmentation algorithms to DNAcopy. DNAcopy is widely used in the single cell CNV field (Knouse et al., 2016; McConnell et al., 2013) but has been critiqued for yielding inaccurate results in data with high autocorrelated noise (Muggeo and Adelfio, 2011). While we have ascertained that the impact of autocorrelated noise on our results is likely very low, it would nonetheless be useful to obtain CNV calls from other commonly used segmentation algorithms, such as HMMcopy and FUnC (Knouse et al., 2016; Knouse et al., 2014; Rohrback et al., 2018a). The use of multiple single cell segmentation algorithms would allow us to evaluate the merits of different computational approaches and generate a consensus set of CNVs, similar to our strategy incorporating LUMPY, DELLY, and SVXplorer in Chapter 3.

4.2 Experimental exploration of the biological implications of mosaicism

Given the myriad unresolved biological questions surrounding brain somatic mosaicism, there are a variety of future experiments that would build upon our work thus far.

Using the mouse brain as a model for studying mosaicism is an excellent approach for the reasons discussed in Chapter 3. However, there could be some improvements to the mouse experiments we carried out. The most pressing, but also most fixable, concern was the questionable quality of data in the 8 mouse study; this problem could be avoided in the future by a switch to PicoPLEX amplification, which was used in

the human data, or optimizing the MALBAC protocol. Nevertheless, we attempted to make use of the data by combining it into “bulk” data and detecting CNVs using SV callers. A key follow-up experiment that would confirm the presence of these CNVs is PCR validation. In the case of a deletion, PCR validation is carried out using primers oriented towards each other that amplify a region starting upstream and ending downstream of the deleted region. In duplications, PCR validation can be slightly more difficult depending on the duplication type; however, if the region is duplicated in tandem (i.e., the new DNA is directly adjacent to the original sequence), primers are designed in opposite orientation at either end of the duplicated region and a fragment spanning the end of one copy and the beginning of the next is amplified. These experiments would give legitimacy to our method of detecting CNVs in pooled single cell sequencing data.

The experimental design of the 8 mouse study could also be reconfigured to focus on a more specific question. The mice selected for the analysis were from multiple gene knockout statuses, ages, and genetic backgrounds (Table 3.1). Thus, the surveying of CNV prevalence across a diverse group of mice, some of which lacked biological replicates, was perhaps too broadly focused to produce conclusive results about the effects of any particular gene knockout, age, or genetic background. The issue was compounded by the varying number of cells sequenced from each mouse, resulting in sample sizes that hindered direct comparison of CNV results among the 8 mice. A simpler, age-matched comparison of just p53KO and p53WT individuals, or just htau and htauKO individuals, including multiple mice and equal numbers of cells sequenced for each condition, might allow for more definitive conclusions to be drawn from the results.

Our research would also benefit from an improved enrichment analysis using a broader group of relevant gene sets. Using a similar random model to the one implemented in Chapter 2, and perhaps adding improvements such as adjustments for heterochromatic and euchromatic regions as well as accounting for gene length variability, we could explore additional gene lists for their overlap with CNV hotspots. One group worthy of exploration is haploinsufficient genes, which are defined as those genes for which a single intact copy is incapable of maintaining the gene's normal function in the organism (Huang et al., 2010). These genes are of particular interest in the context of deletion CNVs, wherein entire blocks of genes are reduced to single copies, but the cell nonetheless survives and, presumably, maintains some level of functionality. In such cases, we can test whether known sets of haploinsufficient genes (Petrovski et al., 2013), annotated by sources such as the Online Mendelian Inheritance in Man (OMIM) database, are observed to be deleted less often in neurons as a result of the genes' importance. Alternatively, it may be the case that neurons are resilient to the loss of these typically critical genes, perhaps as a result of having their own separate haploinsufficiency gene set. Moreover, each of these possible outcomes could vary by the genotype of the individual; for instance, it is conceivable that individuals with p53 KO may demonstrate greater tolerance for loss of haploinsufficient genes than wild-type individuals, owing to the absence of apoptotic signaling by p53.

These gene set analyses could be further expanded to investigate questions relating CNV-affected genes to the ages of the individuals. It may be the case that genomic diversity in neurons is a dynamic phenomenon during development and aging, with distinct manifestations in young and old neurons. For example, it would be

worthwhile to examine whether aged neurons are more likely to contain deletions of neurodevelopmental genes (including any that are haploinsufficient) than young neurons. If so, this finding would suggest that aged neurons experience and tolerate the accumulation of deletions in locations that would result in cell death or organismal lethality if present in young neurons. Conversely, if deletions tend to affect the same regions in both young and old individuals, it would imply a general robustness of single neurons to cope with DNA damage, even in important genes at important stages of development, challenging the very notion of haploinsufficiency in neurodevelopment at the single cell level. In the course of performing these analyses, it would be beneficial to compare and contrast the findings in humans and mice in order to elucidate similarities and differences between the species, in regards to both overall trends of haploinsufficiency in neurodevelopment as well as specific conserved genes of interest.

Neurodevelopmental disease represents another branch of brain somatic mosaicism deserving further exploration. To build upon the findings linking *de novo* mutations to diseases such as autism spectrum disorder and schizophrenia (Freed and Pevsner, 2016; Yurov et al., 2008), an analysis of neurons from the brains of affected individuals would allow us to search for known disease-causing CNVs and assess their prevalence. While these *de novo* events are often detected from bulk sequencing of peripheral blood samples, disease-causing mutations also occur in mosaic fashion and can be detected via single cell sequencing. In cases where blood sequencing is unable to identify mutations causing neurodevelopmental disease, single cell sequencing of the affected tissue may be better-suited to provide answers.

One aspect of brain somatic mosaicism that we have not explored is the prevalence of CNV neurons in regions beyond the frontal cortex. Like most single neuron CNV studies to date, our work has focused on frontal cortical neurons, but these cells might not be representative of the prevalence of CNV neurons in other important regions of the brain. We could obtain additional samples from our post-mortem brains and determine whether CNV neuron rates differ among brain regions. Moreover, this approach would allow us to identify any clonal CNVs detected across multiple regions, which would help elucidate the timing of somatic mutations in neurodevelopment.

Exploring different brain regions would also be of interest from a neurodegenerative disease perspective. Our aging trend (Figure 2.3) demonstrates the apparent die-off of CNV neurons during normal aging, resulting in a higher proportion of euploid neurons. The entorhinal cortex is known to exhibit degeneration beginning in early stage Alzheimer's disease (AD) (Van Hoesen et al., 1991), but few studies to date have examined this particular tissue for evidence of any forms of mosaicism (Arendt et al., 2015; Sala Frigerio et al., 2015). A study comparing CNVs detected in the frontal and entorhinal cortices between age-matched healthy individuals and AD patients could reveal a relationship between CNVs, or some subset of CNVs, and AD. Are CNV neurons the first to die in the entorhinal cortex of AD patients, as they appear to be in the frontal cortex of non-diseased individuals? If so, are some CNV neurons more likely to survive than others? Exploring these questions would establish the prevalence of CNV neurons across brain regions in AD and non-AD individuals as well as inform future hypotheses about the interaction of mosaicism and brain disease.

4.3 Summary

Thomas R. Insel, the Director of the National Institute of Mental Health from 2002 to 2015, once mused that brain somatic mosaicism was the “dark matter of psychiatric genetics” (Insel, 2014). Fortunately, unlike dark matter, we have the technological means to make direct observations about brain somatic mosaicism. The experiments detailed in this dissertation and those proposed in this chapter are aimed towards accumulating observations about mosaicism in the brain such that, in time, we will understand its causes and consequences.

Appendix A
Genomic Coordinates of CNVs Detected in Chapter 2

Table A.1: CNVs detected under lenient thresholds in human brain scWGS data analyzed in Chapter 2

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
0.36	Neuron	UVA_neuron_124	del	1	6	125622700	143454539
0.36	Neuron	UVA_neuron_124	del	1	6	151290699	159861731
0.36	Neuron	UVA_neuron_124	del	1	11	55912100	62524869
0.36	Neuron	UVA_neuron_124	del	1	12	78512659	97060333
0.36	Neuron	UVA_neuron_124	del	1	15	30151267	42421648
0.36	Neuron	UVA_neuron_124	del	1	16	30720835	58149348
0.36	Neuron	UVA_neuron_138	dup	3	1	13787191	52515582
0.36	Neuron	UVA_neuron_138	dup	3	1	156424280	167003471
0.36	Neuron	UVA_neuron_138	del	1	5	2620428	21076454
0.36	Neuron	UVA_neuron_138	dup	3	6	20567405	25955160
0.36	Neuron	UVA_neuron_138	dup	3	9	91039916	141213431
0.36	Neuron	UVA_neuron_138	del	1	10	60270052	71230921
0.36	Neuron	UVA_neuron_138	dup	3	10	127503786	135534747
0.36	Neuron	UVA_neuron_138	del	1	11	26047735	37299097
0.36	Neuron	UVA_neuron_138	dup	3	13	96410970	110306856
0.36	Neuron	UVA_neuron_138	dup	3	16	33939045	60655383
0.36	Neuron	UVA_neuron_138	del	1	18	24597918	43571620
0.36	Neuron	UVA_neuron_139	del	1	1	162849228	172498733
0.36	Neuron	UVA_neuron_139	del	1	2	171681989	203532394
0.36	Neuron	UVA_neuron_139	del	1	2	220317777	233581946
0.36	Neuron	UVA_neuron_139	dup	3	5	134261522	137814368
0.36	Neuron	UVA_neuron_139	del	1	13	78700363	82857520
0.36	Neuron	UVA_neuron_150	dup	3	7	84453688	93543079
0.36	Neuron	UVA_neuron_150	dup	3	8	64008552	67021221
0.36	Neuron	UVA_neuron_161	dup	3	1	163417199	166368036
0.36	Neuron	UVA_neuron_177	del	1	2	219052240	232928932
0.36	Neuron	UVA_neuron_177	del	1	6	76026361	86110507
0.36	Neuron	UVA_neuron_177	dup	3	7	83293756	86219163
0.36	Neuron	UVA_neuron_177	dup	3	11	20663030	25429692
0.36	Neuron	UVA_neuron_200	dup	3	16	4776984	10318764
0.36	Neuron	UVA_neuron_202	dup	3	10	76874399	83096856
0.36	Neuron	UVA_neuron_211	dup	3	1	186531620	245047728
0.36	Neuron	UVA_neuron_211	dup	3	2	56157960	64100151
0.36	Neuron	UVA_neuron_212	dup	3	8	92131456	96970694
0.36	Neuron	UVA_neuron_213	dup	3	3	25296049	47155485
0.36	Neuron	UVA_neuron_213	dup	3	6	137478177	144028598
0.36	Neuron	UVA_neuron_214	dup	3	9	116087618	122906838
0.36	Neuron	UVA_neuron_214	dup	3	22	34341949	40645768
0.36	Neuron	UVA_neuron_226	dup	3	6	90697068	104915253
0.36	Neuron	UVA_neuron_226	dup	3	13	42467337	45384290

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
0.36	Neuron	UVA_neuron_228	del	1	14	40336394	50128849
20	Neuron	SRR1006161	dup	3	2	79201843	85053312
20	Neuron	SRR1006161	del	1	5	33413701	39255824
20	Neuron	SRR1006164	del	1	3	52001019	56678590
20	Neuron	SRR1006164	del	1	3	72096579	81325034
20	Neuron	SRR1006164	del	1	10	18626666	26068464
20	Neuron	SRR1006182	dup	3	2	114470301	117440682
20	Neuron	SRR1006182	dup	3	20	45065873	48857276
20	Neuron	SRR1006183	del	1	2	0	37597556
20	Neuron	SRR1006183	del	1	2	54339151	234240781
20	Neuron	SRR1006183	del	1	4	16126824	21983997
20	Neuron	SRR1006183	del	1	7	76860004	84453687
20	Neuron	SRR1006183	del	1	8	12771698	20358881
20	Neuron	SRR1006183	del	1	8	56808295	94495143
20	Neuron	SRR1006183	del	1	8	136030810	146364022
20	Neuron	SRR1006183	del	1	9	19390496	33256284
20	Neuron	SRR1006183	del	1	9	92230894	103279747
20	Neuron	SRR1006183	del	1	9	107511035	112314022
20	Neuron	SRR1006183	del	1	13	30526230	44217199
20	Neuron	SRR1006183	del	0	13	44217200	50380053
20	Neuron	SRR1006183	del	1	13	50380054	99402742
20	Neuron	SRR1006183	del	1	14	24151499	35249857
20	Neuron	SRR1006183	del	0	14	35249858	61869005
20	Neuron	SRR1006183	del	0	14	79690457	95097802
20	Neuron	SRR1006183	del	1	20	32848484	36251126
20	Neuron	SRR1006191	del	1	12	27691237	32230557
20	Neuron	SRR1006193	dup	3	3	53195998	57302460
20	Neuron	SRR1006193	dup	3	3	170606996	194021919
20	Neuron	SRR1006211	dup	3	1	192946632	200746732
20	Neuron	SRR1006211	dup	3	1	206863148	211687065
20	Neuron	SRR1006211	dup	3	1	226358227	229496990
20	Neuron	SRR1006211	dup	3	2	166862141	169817535
20	Neuron	SRR1006211	dup	3	7	130715094	134795986
20	Neuron	SRR1006211	dup	3	11	125603664	130318113
20	Neuron	SRR1006211	dup	3	15	23757739	29546675
20	Neuron	SRR1006211	dup	3	15	48444762	52150669
20	Neuron	SRR1006229	dup	3	3	97673841	104431111
20	Neuron	SRR1006229	dup	3	7	48809135	53621695
20	Neuron	SRR1006363	del	1	1	201321553	204560509
20	Neuron	SRR1006363	del	1	2	210869348	233581946
20	Neuron	SRR1006363	del	1	4	109800045	140454702

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
20	Neuron	SRR1006363	del	1	4	148044344	185573695
20	Neuron	SRR1006363	del	1	8	29818780	38353038
20	Neuron	SRR1006363	del	1	13	84578903	91639787
20	Neuron	SRR1006380	dup	3	2	56157960	86326604
20	Neuron	SRR1006380	dup	3	2	184210526	187222439
20	Neuron	SRR1006380	dup	3	5	63641913	68455175
20	Neuron	SRR1006380	dup	3	5	107391861	112244909
20	Neuron	SRR1006380	dup	3	8	47308005	54276092
20	Neuron	SRR1006380	dup	3	8	57457789	62800234
20	Neuron	SRR1006380	dup	3	8	78558090	82781055
20	Neuron	SRR1006394	dup	3	15	86512096	89424074
20	Neuron	SRR1006396	dup	3	12	113174179	116091173
20	Neuron	SRR1006399	dup	3	3	13262437	16306586
20	Neuron	SRR1006403	dup	3	18	0	6758006
20	Neuron	SRR1006404	dup	3	7	155552938	159138663
20	Neuron	SRR1006409	del	1	6	64411760	68034963
20	Neuron	SRR1006409	del	1	9	102080528	105096382
20	Neuron	SRR1006411	dup	3	2	113543645	118701713
20	Neuron	SRR1006412	dup	3	1	210416295	224373408
20	Neuron	SRR1006412	dup	3	4	8901947	15526654
20	Neuron	SRR1006412	dup	3	4	21983998	70843066
20	Neuron	SRR1006412	dup	3	4	89600533	96193924
20	Neuron	SRR1006412	dup	3	4	107990746	125787612
20	Neuron	SRR1006412	dup	3	4	141626199	151620553
20	Neuron	SRR1006412	dup	3	4	166012100	177420690
20	Neuron	SRR1006412	dup	3	8	109678767	120710738
20	Neuron	SRR1006412	dup	3	9	35501954	71370308
20	Neuron	SRR1006412	dup	3	14	59409245	65558275
20	Neuron	SRR1006412	dup	3	14	80253742	88978761
20	Neuron	SRR1006415	del	1	1	17334882	26166885
20	Neuron	SRR1006415	del	1	2	112792600	122785261
20	Neuron	SRR1006415	del	1	3	135469191	139367866
20	Neuron	SRR1006415	dup	3	5	167797876	171938590
20	Neuron	SRR1006415	del	1	8	29818780	42765932
20	Neuron	SRR1006415	del	1	17	0	6324645
20	Neuron	SRR1006416	dup	3	17	30458696	33384294
24	Neuron	SRR1006018	del	1	3	68597333	72096578
24	Neuron	SRR1006018	del	1	7	128673484	131910587
24	Neuron	SRR1006018	del	1	11	75904906	79008352
24	Neuron	SRR1006018	del	1	13	102403082	107439681
24	Neuron	SRR1006019	del	1	1	224373409	228266772

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
24	Neuron	SRR1006019	del	1	9	106891097	109948494
24	Neuron	SRR1006019	del	1	12	16268912	19227374
24	Neuron	SRR1006019	del	1	12	110402896	113771179
24	Neuron	SRR1006019	del	1	17	62937467	69819402
24	Neuron	SRR1006019	del	1	18	25735093	30994269
24	Neuron	SRR1006019	del	1	20	20741862	23764315
24	Neuron	SRR1006021	del	1	17	0	5725558
24	Neuron	SRR1006024	del	1	10	16952438	22968237
24	Neuron	SRR1006026	del	1	4	10234780	14347120
24	Neuron	SRR1006026	del	1	9	97625074	101516985
24	Neuron	SRR1006026	dup	3	18	41263826	45898880
24	Neuron	SRR1006036	del	1	2	224992645	228498133
24	Neuron	SRR1006036	del	1	4	170923452	173893331
24	Neuron	SRR1006036	del	1	10	0	4704894
24	Neuron	SRR1006036	del	1	12	6066151	10033192
26	Non-neuron	UVA_nonneuron_079	dup	3	12	125847967	133851895
26	Non-neuron	UVA_nonneuron_090	dup	3	11	40924064	46208223
26	Neuron	SRR1006042	del	1	10	86032194	89483221
26	Neuron	SRR1006043	del	1	1	15573437	25427883
26	Neuron	SRR1006043	del	1	21	40764653	48129895
26	Neuron	SRR1006055	del	1	5	14714111	18488898
26	Neuron	SRR1006055	del	1	5	167797876	172550230
26	Neuron	SRR1006055	del	1	10	86032194	93667392
26	Neuron	SRR1006055	del	1	20	0	3130305
26	Neuron	SRR1006081	del	1	13	32330089	35221749
26	Neuron	SRR1006088	dup	3	16	85061182	90354753
26	Neuron	SRR1006099	del	1	2	47782132	55492676
26	Neuron	SRR1006099	del	1	2	61457641	70046549
26	Neuron	SRR1006099	del	1	2	204230789	207159955
26	Neuron	SRR1006099	del	1	10	78634443	81764654
26	Neuron	SRR1006099	del	1	13	32893809	36999872
26	Neuron	SRR1006101	del	1	6	147026482	149991518
26	Neuron	SRR1006101	del	1	20	39681629	42618074
26	Neuron	SRR1006137	del	1	12	78512659	93236899
26	Neuron	SRR1006143	del	1	1	240347054	244387281
26	Neuron	UVA_neuron_251	dup	3	5	0	141048954
26	Neuron	UVA_neuron_255	dup	3	6	37402144	44514766
26	Neuron	UVA_neuron_255	dup	3	6	147026482	151290698
26	Neuron	UVA_neuron_255	dup	3	12	125221928	133851895
26	Neuron	UVA_neuron_260	dup	3	4	77277690	104265509

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
26	Neuron	UVA_neuron_260	dup	3	4	170923452	173893331
26	Neuron	UVA_neuron_260	dup	3	6	0	24165067
26	Neuron	UVA_neuron_260	dup	3	10	108109809	114006282
26	Neuron	UVA_neuron_260	dup	3	13	57146240	60642041
26	Neuron	UVA_neuron_260	dup	3	16	19944950	46833323
26	Neuron	UVA_neuron_260	dup	3	17	31112224	34082077
26	Neuron	UVA_neuron_270	dup	3	8	32850243	37657413
26	Neuron	UVA_neuron_274	dup	3	2	75036542	79201842
26	Neuron	UVA_neuron_276	dup	3	19	28416236	33117740
26	Neuron	UVA_neuron_282	dup	3	11	63983040	73405105
26	Neuron	UVA_neuron_282	dup	3	11	112056503	135006516
26	Neuron	UVA_neuron_283	dup	3	11	112056503	121970557
26	Neuron	UVA_neuron_289	dup	4	1	14407453	21563445
26	Neuron	UVA_neuron_289	dup	3	1	28995833	36027949
26	Neuron	UVA_neuron_289	dup	3	7	14766268	24279826
26	Neuron	UVA_neuron_289	dup	3	11	66542079	70033066
26	Neuron	UVA_neuron_289	dup	3	11	113779841	120234173
26	Neuron	UVA_neuron_289	dup	3	13	99992280	102951309
26	Neuron	UVA_neuron_289	dup	3	18	21649083	24597917
26	Neuron	UVA_neuron_290	dup	3	6	11502203	16826185
26	Neuron	UVA_neuron_299	dup	3	2	110317438	120427479
26	Neuron	UVA_neuron_299	dup	3	3	140492003	172377702
26	Neuron	UVA_neuron_299	dup	3	6	149991519	158006608
26	Neuron	UVA_neuron_299	dup	3	7	13601725	17061332
26	Neuron	UVA_neuron_305	del	1	13	92199121	95759123
26	Neuron	UVA_neuron_320	del	1	1	75753124	79414169
26	Neuron	UVA_neuron_320	dup	3	3	116335167	119237862
26	Neuron	UVA_neuron_320	del	1	5	4856870	62425621
26	Neuron	UVA_neuron_320	del	1	5	102798905	110434237
26	Neuron	UVA_neuron_320	del	1	8	99422239	103755081
26	Neuron	UVA_neuron_320	del	1	8	113174309	122528361
26	Neuron	UVA_neuron_324	dup	3	7	36089030	43210146
26	Neuron	UVA_neuron_324	dup	3	7	101739495	112052331
26	Neuron	UVA_neuron_332	del	1	1	168802725	200746732
26	Neuron	UVA_neuron_332	del	1	1	222913998	228266772
26	Neuron	UVA_neuron_332	del	1	5	35584463	46321785
26	Neuron	UVA_neuron_332	del	1	5	148177528	168391175
26	Neuron	UVA_neuron_332	del	1	12	104332379	111126214
26	Neuron	UVA_neuron_334	dup	3	11	113779841	120234173
26	Neuron	UVA_neuron_339	dup	3	1	41088519	67427161
26	Neuron	UVA_neuron_339	dup	3	1	92910060	109071838

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
26	Neuron	UVA_neuron_339	dup	3	3	63866113	96506298
26	Neuron	UVA_neuron_339	dup	3	4	6011375	10234779
26	Neuron	UVA_neuron_339	dup	3	6	3072097	29455580
26	Neuron	UVA_neuron_341	dup	3	11	96145811	100504487
26	Neuron	UVA_neuron_346	dup	3	15	0	39850794
26	Neuron	UVA_neuron_346	dup	3	15	46094911	52150669
26	Neuron	UVA_neuron_359	del	1	6	79835303	90697067
26	Neuron	UVA_neuron_359	del	1	6	110505516	117202244
26	Neuron	UVA_neuron_379	del	1	5	42954016	55143420
26	Neuron	UVA_neuron_379	del	1	5	148741542	151740334
26	Neuron	UVA_neuron_391	dup	3	9	22454535	25441940
26	Neuron	UVA_neuron_403	del	1	18	54218500	57198996
26	Neuron	UVA_neuron_411	del	1	3	124789364	133726592
26	Neuron	UVA_neuron_411	del	1	4	0	8901946
26	Neuron	UVA_neuron_411	del	1	4	20265895	53126901
26	Neuron	UVA_neuron_411	del	1	4	57587710	103549342
26	Neuron	UVA_neuron_411	del	1	4	106745991	120960022
26	Neuron	UVA_neuron_411	del	1	4	144045630	151620553
26	Neuron	UVA_neuron_411	del	1	4	155241292	191154276
26	Neuron	UVA_neuron_411	del	1	6	13230225	24165067
26	Neuron	UVA_neuron_411	del	1	7	106680490	129336084
26	Neuron	UVA_neuron_411	del	1	10	86032194	91213881
26	Neuron	UVA_neuron_411	del	1	21	30644951	41939010
26	Neuron	UVA_neuron_421	del	1	1	49933878	53202683
26	Neuron	UVA_neuron_421	del	1	1	236828852	240958964
26	Neuron	UVA_neuron_421	del	1	2	175956929	179477039
26	Neuron	UVA_neuron_421	del	1	2	228498134	232230040
26	Neuron	UVA_neuron_421	del	1	6	6680836	28141567
26	Neuron	UVA_neuron_421	del	1	6	121458003	124461076
26	Neuron	UVA_neuron_421	del	1	6	160476597	167006545
42	Neuron	SRR1708616	del	1	4	6011375	16126823
42	Neuron	SRR1708616	del	1	4	76639006	84643973
42	Neuron	SRR1708616	del	1	4	95607321	99213573
42	Neuron	SRR1708616	del	1	4	109192152	120237850
42	Neuron	SRR1708616	del	1	15	72659223	77329142
42	Neuron	SRR1708620	del	1	2	82672633	85702751
42	Neuron	SRR1708620	del	1	3	9550350	13262436
42	Neuron	SRR1708620	del	1	4	156394126	160563379
42	Neuron	SRR1708621	dup	3	2	238939047	243199373
42	Neuron	SRR1708621	del	1	3	24745615	27751638
42	Neuron	SRR1708621	del	1	6	124461077	127402842

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
42	Neuron	SRR1708622	del	1	15	47305386	52784858
42	Neuron	SRR1708631	del	1	6	21793482	28763731
42	Neuron	SRR1708631	del	1	6	154359324	159861731
42	Neuron	SRR1708633	del	1	1	46057641	55654886
42	Neuron	SRR1708633	del	0	1	55654887	60255842
42	Neuron	SRR1708633	del	1	1	60255843	89069042
42	Neuron	SRR1708633	del	1	1	113311583	156424279
42	Neuron	SRR1708633	del	0	1	156424280	167003471
42	Neuron	SRR1708633	del	1	1	183553816	191781617
42	Neuron	SRR1708633	del	1	1	205122961	228266772
42	Neuron	SRR1708633	del	1	1	238007532	242151660
42	Neuron	SRR1708633	del	1	2	20650180	27895222
42	Neuron	SRR1708633	del	1	2	150863336	162288618
42	Neuron	SRR1708633	dup	3	4	21983998	26138396
42	Neuron	SRR1708633	del	1	4	38401590	42279611
42	Neuron	SRR1708633	del	1	5	72831185	121909397
42	Neuron	SRR1708633	del	1	7	7089517	40356336
42	Neuron	SRR1708633	del	1	7	106680490	128673483
42	Neuron	SRR1708633	del	1	7	149087096	159138663
42	Neuron	SRR1708633	del	1	9	27854572	79213739
42	Neuron	SRR1708633	del	1	11	14020783	17610261
42	Neuron	SRR1708633	del	1	11	36132876	43338768
42	Neuron	SRR1708633	del	1	11	111426297	135006516
42	Neuron	SRR1708633	del	1	13	0	115169878
42	Neuron	SRR1708633	del	0	14	0	39659648
42	Neuron	SRR1708633	del	1	14	39659649	45198271
42	Neuron	SRR1708633	del	0	14	45198272	55099765
42	Neuron	SRR1708633	del	1	14	55099766	58796222
42	Neuron	SRR1708633	del	0	14	58796223	68685230
42	Neuron	SRR1708633	del	1	14	68685231	71727396
42	Neuron	SRR1708633	del	0	14	71727397	82041045
42	Neuron	SRR1708633	del	1	14	82041046	103188689
42	Neuron	SRR1708633	del	0	14	103188690	107349540
42	Neuron	SRR1708633	del	0	15	23757739	39850794
42	Neuron	SRR1708633	del	1	15	39850795	47305385
42	Neuron	SRR1708633	del	0	15	47305386	55863189
42	Neuron	SRR1708633	del	1	15	55863190	72659222
42	Neuron	SRR1708633	del	0	15	72659223	102531392
42	Neuron	SRR1708633	del	1	16	2774872	7268606
42	Neuron	SRR1708633	del	0	16	12958031	75785091
42	Neuron	SRR1708633	del	0	16	80914072	90354753

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
42	Neuron	SRR1708633	del	1	17	35657382	49213544
42	Neuron	SRR1708633	del	1	19	18101063	40213068
42	Neuron	SRR1708633	del	0	21	15737919	48129895
42	Neuron	SRR1708633	del	0	22	0	23892724
42	Neuron	SRR1708635	del	1	9	104493886	107511034
49	Non-neuron	UVA_nonneuron_002	del	1	14	73534859	78597161
49	Non-neuron	UVA_nonneuron_015	dup	3	2	75036542	79201842
49	Non-neuron	UVA_nonneuron_015	del	1	7	6098775	15320919
49	Non-neuron	UVA_nonneuron_015	del	1	7	45180623	54831586
49	Neuron	UVA_neuron_018	del	1	5	33413701	44343860
49	Neuron	UVA_neuron_018	del	1	5	64855087	76532539
49	Neuron	UVA_neuron_018	del	1	5	79664376	107391860
49	Neuron	UVA_neuron_018	del	1	5	110434238	130481983
49	Neuron	UVA_neuron_022	dup	3	16	5531492	11027809
49	Neuron	UVA_neuron_023	del	1	2	187222440	190803979
49	Neuron	UVA_neuron_027	del	1	6	44514767	47465986
49	Neuron	UVA_neuron_027	del	1	11	63326410	67209449
49	Neuron	UVA_neuron_031	del	1	1	145035037	167586869
49	Neuron	UVA_neuron_031	del	1	1	193539023	233052991
49	Neuron	UVA_neuron_031	del	1	1	242902919	249250621
49	Neuron	UVA_neuron_031	del	1	10	52397003	74184936
49	Neuron	UVA_neuron_031	del	1	10	86032194	93667392
49	Neuron	UVA_neuron_035	del	1	1	109071839	116456523
49	Neuron	UVA_neuron_035	del	1	6	0	42642124
49	Neuron	UVA_neuron_035	del	1	6	72836840	111141025
49	Neuron	UVA_neuron_035	del	1	6	125622700	171115067
49	Neuron	UVA_neuron_035	del	1	7	6098775	75361496
49	Neuron	UVA_neuron_035	del	1	7	88032352	99331103
49	Neuron	UVA_neuron_035	del	1	7	124340583	127335585
49	Neuron	UVA_neuron_035	del	1	8	0	37657413
49	Neuron	UVA_neuron_035	del	0	8	37657414	53068846
49	Neuron	UVA_neuron_035	del	1	8	53068847	69476869
49	Neuron	UVA_neuron_035	del	0	8	69476870	90950790
49	Neuron	UVA_neuron_035	del	1	8	90950791	109678766
49	Neuron	UVA_neuron_035	del	0	8	109678767	124319612
49	Neuron	UVA_neuron_035	del	1	8	124319613	146364022
49	Neuron	UVA_neuron_035	del	0	9	0	34194622
49	Neuron	UVA_neuron_035	del	1	9	81064076	95539901
49	Neuron	UVA_neuron_035	del	1	9	98794162	108127109

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
49	Neuron	UVA_neuron_035	del	0	9	108127110	114148695
49	Neuron	UVA_neuron_035	del	1	9	114148696	119539668
49	Neuron	UVA_neuron_035	del	0	9	119539669	128404855
49	Neuron	UVA_neuron_035	del	1	9	128404856	136083447
49	Neuron	UVA_neuron_035	del	1	10	17566785	27413532
49	Neuron	UVA_neuron_035	del	1	10	114006283	135534747
49	Neuron	UVA_neuron_035	del	1	11	0	63983039
49	Neuron	UVA_neuron_035	del	1	11	69398811	108422355
49	Neuron	UVA_neuron_035	del	1	11	116619094	135006516
49	Neuron	UVA_neuron_035	del	1	12	79632098	93236899
49	Neuron	UVA_neuron_035	del	0	13	0	25710053
49	Neuron	UVA_neuron_035	del	1	13	25710054	34065686
49	Neuron	UVA_neuron_035	del	0	13	34065687	46039591
49	Neuron	UVA_neuron_035	del	1	13	46039592	49757523
49	Neuron	UVA_neuron_035	del	0	13	49757524	79841898
49	Neuron	UVA_neuron_035	del	1	13	79841899	115169878
49	Neuron	UVA_neuron_035	del	0	14	0	30871657
49	Neuron	UVA_neuron_035	del	1	14	30871658	39659648
49	Neuron	UVA_neuron_035	del	0	14	39659649	55099765
49	Neuron	UVA_neuron_035	del	1	14	55099766	58796222
49	Neuron	UVA_neuron_035	del	0	14	58796223	68685230
49	Neuron	UVA_neuron_035	del	1	14	68685231	107349540
49	Neuron	UVA_neuron_035	del	0	15	0	33214513
49	Neuron	UVA_neuron_035	del	1	15	33214514	44563098
49	Neuron	UVA_neuron_035	del	0	15	44563099	66186243
49	Neuron	UVA_neuron_035	del	1	15	66186244	79300783
49	Neuron	UVA_neuron_035	del	0	15	79300784	102531392
49	Neuron	UVA_neuron_035	del	1	17	8294318	27390811
49	Neuron	UVA_neuron_035	del	0	19	19495532	33856129
49	Neuron	UVA_neuron_035	del	1	19	33856130	59128983
49	Neuron	UVA_neuron_035	del	1	20	24987091	32848483
49	Neuron	UVA_neuron_035	del	0	21	15737919	43155186
49	Neuron	UVA_neuron_035	del	1	21	43155187	48129895
49	Neuron	UVA_neuron_052	del	1	2	80355685	85702751
49	Neuron	UVA_neuron_052	del	1	2	120427480	133330191
49	Neuron	UVA_neuron_052	del	1	2	136314476	145525085
49	Neuron	UVA_neuron_052	del	1	10	31961281	48580042
49	Neuron	UVA_neuron_052	del	1	18	13621501	22263013
49	Neuron	UVA_neuron_052	del	1	18	69934661	78077248
49	Neuron	UVA_neuron_059	del	1	1	32239335	37834006
49	Neuron	UVA_neuron_059	del	1	2	24059871	43042109

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
49	Neuron	UVA_neuron_059	del	1	2	101601235	107606695
49	Neuron	UVA_neuron_059	del	1	2	159891567	170436761
49	Neuron	UVA_neuron_059	del	1	4	130089480	142844627
49	Neuron	UVA_neuron_059	del	1	14	87218045	94514383
49	Neuron	UVA_neuron_061	del	1	2	97380358	107606695
49	Neuron	UVA_neuron_061	del	1	2	122785262	159271170
49	Neuron	UVA_neuron_061	del	1	2	188432074	200950169
49	Neuron	UVA_neuron_061	del	1	2	207159956	219724241
49	Neuron	UVA_neuron_061	del	1	2	232928933	236071756
49	Neuron	UVA_neuron_061	del	1	12	85615589	88600828
49	Neuron	UVA_neuron_082	del	1	3	42908150	45923636
49	Neuron	UVA_neuron_082	del	1	3	107361480	142424833
49	Neuron	UVA_neuron_082	del	1	5	94651057	102230903
49	Neuron	UVA_neuron_082	del	1	5	119507447	122511514
49	Neuron	UVA_neuron_082	del	1	6	0	11502202
49	Neuron	UVA_neuron_082	del	1	6	49238865	64411759
49	Neuron	UVA_neuron_082	del	1	6	116594010	138659323
49	Neuron	UVA_neuron_082	del	1	6	145843887	151950282
49	Neuron	UVA_neuron_099	del	1	4	94379967	106745990
49	Neuron	UVA_neuron_099	del	1	5	28505588	55143420
49	Neuron	UVA_neuron_099	del	1	5	68455176	88102197
49	Neuron	UVA_neuron_099	del	1	5	94651057	119507446
49	Neuron	UVA_neuron_099	del	1	5	130481984	144594516
49	Neuron	UVA_neuron_099	del	1	7	16471244	22319107
49	Neuron	UVA_neuron_099	del	1	7	32538381	43812011
49	Neuron	UVA_neuron_099	del	1	7	75361497	100174991
49	Neuron	UVA_neuron_099	del	1	7	123129112	127335585
49	Neuron	UVA_neuron_099	del	1	8	0	53637582
49	Neuron	UVA_neuron_099	del	1	8	85764299	104984539
49	Neuron	UVA_neuron_099	del	1	8	127938903	140076363
49	Neuron	UVA_neuron_099	del	1	10	0	100673442
49	Neuron	UVA_neuron_099	del	1	10	103869749	135534747
49	Neuron	UVA_neuron_099	del	1	14	40336394	45830753
49	Neuron	UVA_neuron_099	del	1	14	61244093	75544359
49	Neuron	UVA_neuron_099	del	1	15	38095761	45451863
49	Neuron	UVA_neuron_099	del	1	15	48993249	60151637
49	Neuron	UVA_neuron_099	del	1	15	66787090	74551715
49	Neuron	UVA_neuron_099	del	0	18	0	3835264
49	Neuron	UVA_neuron_099	del	1	18	3835265	59556928
49	Neuron	UVA_neuron_099	del	0	18	59556929	78077248
49	Neuron	UVA_neuron_099	del	1	20	0	25638025

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
49	Neuron	UVA_neuron_099	del	1	20	47558424	58364310
52	Neuron	SRR1548998	del	1	8	0	4410697
52	Neuron	SRR1549003	dup	3	10	77480813	80329798
52	Neuron	SRR1549005	del	1	3	0	4233531
52	Neuron	SRR1549010	del	1	8	0	4961867
68	Neuron	SRR1549033	dup	3	1	156424280	160119602
68	Neuron	SRR1549038	dup	3	5	0	4283072
68	Neuron	SRR1549040	dup	3	5	146406506	157255568
68	Neuron	SRR1549040	del	1	13	92828702	95759123
69	Neuron	ERR1391234	dup	3	1	55037497	60821075
69	Neuron	ERR1391319	del	1	18	73891587	78077248
69	Neuron	ERR1391336	dup	3	17	68669368	71609825
70	Neuron	SRR1548977	del	1	22	0	51304566
70	Neuron	SRR1548980	del	0	1	245671145	249250621
70	Neuron	SRR1548980	del	1	4	76639006	79730141
70	Neuron	SRR1548980	del	1	10	11832186	16952437
70	Neuron	SRR1548982	del	1	6	0	4275439
70	Neuron	SRR1548983	del	1	3	87150266	90176747
70	Neuron	SRR1548983	del	1	4	99840114	104265509
70	Neuron	SRR1548983	del	1	11	102999982	106489970
70	Neuron	SRR1548983	del	1	13	66003034	74587812
74	Neuron	ERR1391282	del	1	16	0	12344471
74	Neuron	ERR1391284	dup	3	17	70389170	73499935
74	Neuron	ERR1391291	del	1	3	44715393	49993328
74	Neuron	ERR1391293	del	1	16	0	10318764
74	Neuron	ERR1391357	dup	3	1	57393964	61970029
74	Neuron	ERR1391380	dup	3	22	47080595	51304566
81	Neuron	ERR1391176	dup	3	9	118393554	121768891
81	Neuron	ERR1391176	dup	3	17	53411024	56376647
81	Neuron	ERR1391183	dup	3	16	48745856	55674867
81	Neuron	ERR1391220	del	1	2	224992645	234240781
81	Neuron	ERR1391220	del	1	3	58633446	64412452
81	Neuron	ERR1391220	del	1	3	71522747	83068200
81	Neuron	ERR1391220	del	1	3	131283290	183233764
81	Neuron	ERR1391220	del	1	3	187050157	192262855
81	Neuron	ERR1391220	del	1	4	17284317	37194969
81	Neuron	ERR1391220	del	1	6	12660844	24798645
81	Neuron	ERR1391220	del	1	6	47465987	135683261
81	Neuron	ERR1391220	del	1	6	154359324	158006608
81	Neuron	ERR1391220	del	1	10	18626666	52397002
81	Neuron	ERR1391220	del	1	12	28302778	34370916

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
81	Neuron	ERR1391229	dup	3	1	55037497	60255842
86	Non-neuron	UVA_nonneuron_103	dup	3	2	75633395	79201842
86	Non-neuron	UVA_nonneuron_108	del	1	3	0	8971642
86	Non-neuron	UVA_nonneuron_108	del	1	3	15695348	37372096
86	Non-neuron	UVA_nonneuron_108	del	1	3	69210684	97673840
86	Non-neuron	UVA_nonneuron_108	del	1	3	162224600	178911918
86	Non-neuron	UVA_nonneuron_108	del	1	9	6276769	11594685
86	Non-neuron	UVA_nonneuron_108	del	1	9	16360992	24841287
86	Non-neuron	UVA_nonneuron_117	del	1	6	141078599	147593774
86	Non-neuron	UVA_nonneuron_145	del	1	19	0	4431118
86	Neuron	UVA_neuron_450	del	1	6	7880049	10896382
86	Neuron	UVA_neuron_485	dup	3	7	0	159138663
86	Neuron	UVA_neuron_485	dup	3	9	0	141213431
86	Neuron	UVA_neuron_485	del	1	10	0	135534747
86	Neuron	UVA_neuron_505	dup	3	2	75036542	78027311
86	Neuron	UVA_neuron_533	del	1	1	167586870	176885373
86	Neuron	UVA_neuron_533	del	1	1	183553816	200746732
86	Neuron	UVA_neuron_533	del	1	1	236828852	244387281
86	Neuron	UVA_neuron_533	del	1	2	61457641	88809038
86	Neuron	UVA_neuron_533	del	1	2	147197476	161693615
86	Neuron	UVA_neuron_533	del	1	2	183583489	190803979
86	Neuron	UVA_neuron_533	del	1	5	58749287	77144122
86	Neuron	UVA_neuron_533	del	1	5	82729349	134261521
86	Neuron	UVA_neuron_533	del	1	8	48124278	58665873
86	Neuron	UVA_neuron_533	del	1	8	119501301	124319612
86	Neuron	UVA_neuron_533	del	1	13	52941095	77518723
86	Neuron	UVA_neuron_533	del	1	13	99992280	110875528
86	Neuron	UVA_neuron_533	del	1	15	34392199	62406148
86	Neuron	UVA_neuron_533	del	1	15	84203290	100751400
86	Neuron	UVA_neuron_533	del	1	18	21049924	74475930
95	Neuron	UVA_neuron_584	dup	3	8	22668482	26845698
95	Neuron	UVA_neuron_586	dup	3	1	54415206	109071838
95	Neuron	UVA_neuron_586	dup	3	1	205736766	226358226
95	Neuron	UVA_neuron_586	dup	3	3	0	5450845
95	Neuron	UVA_neuron_586	dup	3	3	58633446	66824981
95	Neuron	UVA_neuron_586	dup	3	3	70968285	79005714

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
95	Neuron	UVA_neuron_586	dup	3	3	114067040	118003951
95	Neuron	UVA_neuron_586	dup	3	3	145905641	169326061
95	Neuron	UVA_neuron_586	dup	3	3	174097863	198022430
95	Neuron	UVA_neuron_586	dup	3	5	149920972	154153043
95	Neuron	UVA_neuron_586	dup	3	14	33417964	45830753
95	Neuron	UVA_neuron_586	dup	3	14	50806069	58796222
95	Neuron	UVA_neuron_586	dup	3	14	77944607	90194806
95	Neuron	UVA_neuron_586	dup	3	14	94514384	101310646
95	Neuron	UVA_neuron_586	dup	3	21	15737919	19392504
95	Neuron	UVA_neuron_586	dup	3	21	34866097	38458544
95	Neuron	UVA_neuron_593	dup	3	11	130899822	135006516
95	Neuron	UVA_neuron_617	del	1	13	61238839	69996157
95	Neuron	UVA_neuron_617	del	1	19	20336620	30237337
95	Neuron	UVA_neuron_617	del	1	21	34252154	37853829
95	Neuron	UVA_neuron_630	del	1	1	36688308	41704920
95	Neuron	UVA_neuron_630	del	1	1	52515583	109720251
95	Neuron	UVA_neuron_630	del	1	1	197191179	202697510
95	Neuron	UVA_neuron_630	del	1	3	16306587	31786628
95	Neuron	UVA_neuron_630	del	1	3	71522747	84223958
95	Neuron	UVA_neuron_630	del	1	3	130066701	180154507
95	Neuron	UVA_neuron_630	del	1	3	190485047	198022430
95	Neuron	UVA_neuron_630	del	1	7	16471244	22961576
95	Neuron	UVA_neuron_630	del	1	7	81002469	85598303
95	Neuron	UVA_neuron_630	del	1	9	32615299	37996021
95	Neuron	UVA_neuron_630	del	1	13	38771764	62432544
95	Neuron	UVA_neuron_630	del	1	15	92004944	98897575
95	Neuron	UVA_neuron_630	del	1	21	18062271	44311358
95	Neuron	UVA_neuron_636	del	1	3	163439789	172377702
95	Neuron	UVA_neuron_636	del	1	4	21983998	25519245
95	Neuron	UVA_neuron_636	del	1	7	86782458	95312683
95	Neuron	UVA_neuron_636	del	1	10	57220098	66757276
95	Neuron	UVA_neuron_636	del	1	10	106385076	114580194
95	Neuron	UVA_neuron_636	del	1	11	25429693	28436562
95	Neuron	UVA_neuron_647	del	1	4	4184323	10813482
95	Neuron	UVA_neuron_647	del	1	4	36015173	42882320
95	Neuron	UVA_neuron_647	del	1	4	184962175	191154276
95	Neuron	UVA_neuron_647	del	1	18	0	12366979
95	Neuron	UVA_neuron_647	del	1	18	45344626	78077248
95	Neuron	UVA_neuron_648	del	1	1	212928493	218133161

Table A.2: CNVs detected under stringent thresholds in human brain scWGS data analyzed in Chapter 2

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
0.36	Neuron	UVA_neuron_124	del	1	6	125622700	143454539
0.36	Neuron	UVA_neuron_138	dup	3	1	13787191	52515582
0.36	Neuron	UVA_neuron_138	dup	3	1	156424280	167003471
0.36	Neuron	UVA_neuron_138	dup	3	6	20567405	25955160
0.36	Neuron	UVA_neuron_138	dup	3	9	91039916	114148695
0.36	Neuron	UVA_neuron_139	del	1	1	162849228	172498733
0.36	Neuron	UVA_neuron_139	del	1	13	78700363	82857520
0.36	Neuron	UVA_neuron_161	dup	3	1	163417199	166368036
0.36	Neuron	UVA_neuron_177	del	1	6	76026361	86110507
0.36	Neuron	UVA_neuron_177	dup	3	11	20663030	25429692
0.36	Neuron	UVA_neuron_211	dup	3	1	186531620	245047728
0.36	Neuron	UVA_neuron_212	dup	3	8	92131456	96970694
20	Neuron	SRR1006161	dup	3	2	79201843	85053312
20	Neuron	SRR1006164	del	1	3	52001019	56678590
20	Neuron	SRR1006164	del	1	3	72096579	81325034
20	Neuron	SRR1006164	del	1	10	18626666	26068464
20	Neuron	SRR1006182	dup	3	2	114470301	117440682
20	Neuron	SRR1006182	dup	3	20	45065873	48857276
20	Neuron	SRR1006183	del	1	2	0	37597556
20	Neuron	SRR1006183	del	1	2	54339151	69415571
20	Neuron	SRR1006183	del	1	2	73134358	234240781
20	Neuron	SRR1006183	del	1	4	16126824	21983997
20	Neuron	SRR1006183	del	1	7	76860004	84453687
20	Neuron	SRR1006183	del	1	8	12771698	20358881
20	Neuron	SRR1006183	del	1	8	56808295	94495143
20	Neuron	SRR1006183	del	1	9	19390496	33256284
20	Neuron	SRR1006183	del	1	9	92230894	103279747
20	Neuron	SRR1006183	del	1	9	107511035	112314022
20	Neuron	SRR1006183	del	1	13	30526230	44217199
20	Neuron	SRR1006183	del	0	13	44217200	50380053
20	Neuron	SRR1006183	del	1	13	50380054	99402742
20	Neuron	SRR1006183	del	1	14	24151499	35249857
20	Neuron	SRR1006183	del	0	14	35249858	61869005
20	Neuron	SRR1006183	del	0	14	79690457	95097802
20	Neuron	SRR1006193	dup	3	3	53195998	57302460
20	Neuron	SRR1006193	dup	3	3	170606996	194021919
20	Neuron	SRR1006211	dup	3	1	206863148	211687065
20	Neuron	SRR1006211	dup	3	1	226358227	229496990
20	Neuron	SRR1006211	dup	3	7	130715094	134795986
20	Neuron	SRR1006211	dup	3	11	125603664	130318113

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
20	Neuron	SRR1006211	dup	3	15	23757739	29546675
20	Neuron	SRR1006229	dup	3	7	48809135	53621695
20	Neuron	SRR1006363	del	1	1	201321553	204560509
20	Neuron	SRR1006363	del	1	4	148044344	185573695
20	Neuron	SRR1006363	del	1	8	29818780	38353038
20	Neuron	SRR1006363	del	1	13	84578903	91639787
20	Neuron	SRR1006380	dup	3	2	56157960	86326604
20	Neuron	SRR1006380	dup	3	2	184210526	187222439
20	Neuron	SRR1006380	dup	3	8	47308005	54276092
20	Neuron	SRR1006380	dup	3	8	57457789	62800234
20	Neuron	SRR1006411	dup	3	2	113543645	118701713
20	Neuron	SRR1006412	dup	3	1	210416295	224373408
20	Neuron	SRR1006412	dup	3	4	8901947	15526654
20	Neuron	SRR1006412	dup	3	4	21983998	70843066
20	Neuron	SRR1006412	dup	3	4	89600533	96193924
20	Neuron	SRR1006412	dup	3	4	107990746	125787612
20	Neuron	SRR1006412	dup	3	4	141626199	151620553
20	Neuron	SRR1006412	dup	3	4	166012100	177420690
20	Neuron	SRR1006412	dup	3	8	109678767	120710738
20	Neuron	SRR1006412	dup	3	9	35501954	71370308
20	Neuron	SRR1006412	dup	3	14	59409245	65558275
20	Neuron	SRR1006412	dup	3	14	80253742	88978761
20	Neuron	SRR1006415	del	1	1	17334882	26166885
20	Neuron	SRR1006415	del	1	2	112792600	122785261
24	Neuron	SRR1006018	del	1	3	68597333	72096578
24	Neuron	SRR1006018	del	1	7	128673484	131910587
24	Neuron	SRR1006018	del	1	11	75904906	79008352
24	Neuron	SRR1006018	del	1	13	102403082	107439681
24	Neuron	SRR1006019	del	1	9	106891097	109948494
24	Neuron	SRR1006019	del	1	12	110402896	113771179
24	Neuron	SRR1006019	del	1	18	25735093	30994269
24	Neuron	SRR1006019	del	1	20	20741862	23764315
24	Neuron	SRR1006026	del	1	9	97625074	101516985
26	Neuron	SRR1006042	del	1	10	86032194	89483221
26	Neuron	SRR1006043	del	1	1	15573437	25427883
26	Neuron	SRR1006055	del	1	5	14714111	18488898
26	Neuron	SRR1006055	del	1	5	167797876	172550230
26	Neuron	SRR1006055	del	1	10	86032194	93667392
26	Neuron	SRR1006055	del	1	20	0	3130305
26	Neuron	SRR1006099	del	1	2	47782132	55492676
26	Neuron	SRR1006099	del	1	2	61457641	70046549

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
26	Neuron	SRR1006099	del	1	2	204230789	207159955
26	Neuron	SRR1006099	del	1	10	78634443	81764654
26	Neuron	SRR1006099	del	1	13	32893809	36999872
26	Neuron	SRR1006101	del	1	20	39681629	42618074
26	Neuron	SRR1006143	del	1	1	240347054	244387281
26	Neuron	UVA_neuron_255	dup	3	6	37402144	44514766
26	Neuron	UVA_neuron_255	dup	3	6	147026482	151290698
26	Neuron	UVA_neuron_255	dup	3	12	125221928	133851895
26	Neuron	UVA_neuron_260	dup	3	17	31112224	34082077
26	Neuron	UVA_neuron_274	dup	3	2	75036542	79201842
26	Neuron	UVA_neuron_276	dup	3	19	28416236	33117740
26	Neuron	UVA_neuron_289	dup	4	1	14407453	21563445
26	Neuron	UVA_neuron_289	dup	3	1	28995833	36027949
26	Neuron	UVA_neuron_289	dup	3	7	14766268	24279826
26	Neuron	UVA_neuron_289	dup	3	11	113779841	120234173
26	Neuron	UVA_neuron_289	dup	3	13	99992280	102951309
26	Neuron	UVA_neuron_289	dup	3	18	21649083	24597917
26	Neuron	UVA_neuron_290	dup	3	6	11502203	16826185
26	Neuron	UVA_neuron_299	dup	3	2	110317438	120427479
26	Neuron	UVA_neuron_299	dup	3	3	140492003	172377702
26	Neuron	UVA_neuron_299	dup	3	7	13601725	17061332
26	Neuron	UVA_neuron_320	dup	3	3	116335167	119237862
26	Neuron	UVA_neuron_320	del	1	5	4856870	62425621
26	Neuron	UVA_neuron_320	del	1	5	102798905	110434237
26	Neuron	UVA_neuron_320	del	1	8	113174309	122528361
26	Neuron	UVA_neuron_332	del	1	5	35584463	46321785
26	Neuron	UVA_neuron_332	del	1	5	148177528	168391175
26	Neuron	UVA_neuron_332	del	1	12	104332379	111126214
26	Neuron	UVA_neuron_339	dup	3	6	3072097	29455580
26	Neuron	UVA_neuron_359	del	1	6	79835303	90697067
26	Neuron	UVA_neuron_359	del	1	6	110505516	117202244
26	Neuron	UVA_neuron_379	del	1	5	148741542	151740334
26	Neuron	UVA_neuron_403	del	1	18	54218500	57198996
26	Neuron	UVA_neuron_411	del	1	4	155241292	191154276
26	Neuron	UVA_neuron_421	del	1	1	236828852	240958964
26	Neuron	UVA_neuron_421	del	1	2	175956929	179477039
26	Neuron	UVA_neuron_421	del	1	6	6680836	28141567
26	Neuron	UVA_neuron_421	del	1	6	160476597	167006545
42	Neuron	SRR1708616	del	1	4	76639006	84643973
42	Neuron	SRR1708621	del	1	3	24745615	27751638
42	Neuron	SRR1708621	del	1	6	124461077	127402842

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
42	Neuron	SRR1708633	del	0	1	55654887	60255842
42	Neuron	SRR1708633	del	0	1	156424280	167003471
42	Neuron	SRR1708633	del	1	1	205122961	228266772
42	Neuron	SRR1708633	del	1	1	238007532	242151660
42	Neuron	SRR1708633	del	1	4	38401590	42279611
42	Neuron	SRR1708633	del	1	11	36132876	43338768
42	Neuron	SRR1708633	del	1	13	89333186	115169878
42	Neuron	SRR1708633	del	0	14	0	39659648
42	Neuron	SRR1708633	del	0	14	45198272	55099765
42	Neuron	SRR1708633	del	1	14	55099766	58796222
42	Neuron	SRR1708633	del	0	14	58796223	68685230
42	Neuron	SRR1708633	del	1	14	68685231	71727396
42	Neuron	SRR1708633	del	0	14	71727397	82041045
42	Neuron	SRR1708633	del	0	14	103188690	107349540
42	Neuron	SRR1708633	del	0	15	23757739	39850794
42	Neuron	SRR1708633	del	0	15	47305386	55863189
42	Neuron	SRR1708633	del	1	15	55863190	72659222
42	Neuron	SRR1708633	del	0	15	72659223	102531392
42	Neuron	SRR1708633	del	1	16	2774872	7268606
42	Neuron	SRR1708633	del	0	16	12958031	75785091
42	Neuron	SRR1708633	del	0	16	80914072	90354753
42	Neuron	SRR1708633	del	0	21	15737919	48129895
42	Neuron	SRR1708633	del	0	22	0	23892724
42	Neuron	SRR1708635	del	1	9	104493886	107511034
49	Non-neuron	UVA_nonneuron_002	del	1	14	73534859	78597161
49	Non-neuron	UVA_nonneuron_015	del	1	7	6098775	15320919
49	Neuron	UVA_neuron_018	del	1	5	33413701	44343860
49	Neuron	UVA_neuron_018	del	1	5	64855087	76532539
49	Neuron	UVA_neuron_018	del	1	5	79664376	107391860
49	Neuron	UVA_neuron_018	del	1	5	110434238	130481983
49	Neuron	UVA_neuron_031	del	1	1	145035037	167586869
49	Neuron	UVA_neuron_031	del	1	1	193539023	233052991
49	Neuron	UVA_neuron_031	del	1	1	242902919	249250621
49	Neuron	UVA_neuron_031	del	1	10	52397003	74184936
49	Neuron	UVA_neuron_031	del	1	10	86032194	93667392
49	Neuron	UVA_neuron_035	del	1	6	125622700	171115067
49	Neuron	UVA_neuron_035	del	1	7	88032352	99331103
49	Neuron	UVA_neuron_035	del	1	8	0	37657413
49	Neuron	UVA_neuron_035	del	0	8	37657414	53068846
49	Neuron	UVA_neuron_035	del	1	8	53068847	69476869

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
49	Neuron	UVA_neuron_035	del	0	8	69476870	90950790
49	Neuron	UVA_neuron_035	del	1	8	90950791	109678766
49	Neuron	UVA_neuron_035	del	0	8	109678767	124319612
49	Neuron	UVA_neuron_035	del	0	9	0	34194622
49	Neuron	UVA_neuron_035	del	1	9	81064076	95539901
49	Neuron	UVA_neuron_035	del	1	9	98794162	108127109
49	Neuron	UVA_neuron_035	del	0	9	108127110	114148695
49	Neuron	UVA_neuron_035	del	0	9	119539669	128404855
49	Neuron	UVA_neuron_035	del	0	13	0	25710053
49	Neuron	UVA_neuron_035	del	1	13	25710054	34065686
49	Neuron	UVA_neuron_035	del	0	13	34065687	46039591
49	Neuron	UVA_neuron_035	del	1	13	46039592	49757523
49	Neuron	UVA_neuron_035	del	0	13	49757524	79841898
49	Neuron	UVA_neuron_035	del	1	13	79841899	115169878
49	Neuron	UVA_neuron_035	del	0	14	0	30871657
49	Neuron	UVA_neuron_035	del	1	14	30871658	39659648
49	Neuron	UVA_neuron_035	del	0	14	39659649	55099765
49	Neuron	UVA_neuron_035	del	1	14	55099766	58796222
49	Neuron	UVA_neuron_035	del	0	14	58796223	68685230
49	Neuron	UVA_neuron_035	del	0	15	0	33214513
49	Neuron	UVA_neuron_035	del	1	15	33214514	44563098
49	Neuron	UVA_neuron_035	del	0	15	44563099	66186243
49	Neuron	UVA_neuron_035	del	1	15	66186244	79300783
49	Neuron	UVA_neuron_035	del	0	15	79300784	102531392
49	Neuron	UVA_neuron_035	del	0	19	19495532	33856129
49	Neuron	UVA_neuron_035	del	1	20	24987091	32848483
49	Neuron	UVA_neuron_035	del	0	21	15737919	43155186
49	Neuron	UVA_neuron_052	del	1	2	80355685	85702751
49	Neuron	UVA_neuron_052	del	1	2	120427480	133330191
49	Neuron	UVA_neuron_052	del	1	10	31961281	48580042
49	Neuron	UVA_neuron_052	del	1	18	13621501	22263013
49	Neuron	UVA_neuron_052	del	1	18	69934661	78077248
49	Neuron	UVA_neuron_059	del	1	2	24059871	43042109
49	Neuron	UVA_neuron_059	del	1	2	101601235	107606695
49	Neuron	UVA_neuron_059	del	1	14	87218045	94514383
49	Neuron	UVA_neuron_061	del	1	2	97380358	107606695
49	Neuron	UVA_neuron_061	del	1	2	122785262	159271170
49	Neuron	UVA_neuron_061	del	1	2	188432074	200950169
49	Neuron	UVA_neuron_061	del	1	2	207159956	219724241
49	Neuron	UVA_neuron_061	del	1	2	232928933	236071756
49	Neuron	UVA_neuron_061	del	1	12	85615589	88600828

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
49	Neuron	UVA_neuron_082	del	1	3	42908150	45923636
49	Neuron	UVA_neuron_082	del	1	6	0	11502202
49	Neuron	UVA_neuron_082	del	1	6	49238865	64411759
49	Neuron	UVA_neuron_082	del	1	6	116594010	138659323
49	Neuron	UVA_neuron_099	del	1	4	94379967	106745990
49	Neuron	UVA_neuron_099	del	1	5	28505588	55143420
49	Neuron	UVA_neuron_099	del	1	5	68455176	88102197
49	Neuron	UVA_neuron_099	del	1	7	75361497	100174991
49	Neuron	UVA_neuron_099	del	1	7	123129112	127335585
49	Neuron	UVA_neuron_099	del	1	8	0	53637582
49	Neuron	UVA_neuron_099	del	1	8	85764299	104984539
49	Neuron	UVA_neuron_099	del	1	10	0	100673442
49	Neuron	UVA_neuron_099	del	1	10	103869749	135534747
49	Neuron	UVA_neuron_099	del	1	14	61244093	75544359
49	Neuron	UVA_neuron_099	del	1	15	48993249	60151637
49	Neuron	UVA_neuron_099	del	0	18	0	3835264
49	Neuron	UVA_neuron_099	del	1	18	3835265	59556928
49	Neuron	UVA_neuron_099	del	0	18	59556929	78077248
49	Neuron	UVA_neuron_099	del	1	20	0	25638025
49	Neuron	UVA_neuron_099	del	1	20	47558424	58364310
52	Neuron	SRR1548998	del	1	8	0	4410697
68	Neuron	SRR1549033	dup	3	1	156424280	160119602
68	Neuron	SRR1549040	del	1	13	92828702	95759123
69	Neuron	ERR1391319	del	1	18	73891587	78077248
69	Neuron	ERR1391336	dup	3	17	68669368	71609825
70	Neuron	SRR1548977	del	1	22	0	51304566
70	Neuron	SRR1548980	del	0	1	245671145	249250621
70	Neuron	SRR1548982	del	1	6	0	4275439
70	Neuron	SRR1548983	del	1	13	66003034	74587812
74	Neuron	ERR1391284	dup	3	17	70389170	73499935
81	Neuron	ERR1391176	dup	3	17	53411024	56376647
81	Neuron	ERR1391220	del	1	2	224992645	234240781
81	Neuron	ERR1391220	del	1	3	58633446	64412452
81	Neuron	ERR1391220	del	1	3	131283290	183233764
81	Neuron	ERR1391220	del	1	4	17284317	37194969
81	Neuron	ERR1391220	del	1	6	12660844	24798645
86	Non-neuron	UVA_nonneuron_108	del	1	3	0	8971642
86	Non-neuron	UVA_nonneuron_108	del	1	3	69210684	97673840
86	Non-neuron	UVA_nonneuron_145	del	1	19	0	4431118
86	Neuron	UVA_neuron_450	del	1	6	7880049	10896382

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
86	Neuron	UVA_neuron_485	dup	3	7	0	159138663
86	Neuron	UVA_neuron_485	dup	3	9	0	141213431
86	Neuron	UVA_neuron_485	del	1	10	0	135534747
86	Neuron	UVA_neuron_533	del	1	1	183553816	200746732
86	Neuron	UVA_neuron_533	del	1	1	236828852	244387281
86	Neuron	UVA_neuron_533	del	1	2	61457641	88809038
86	Neuron	UVA_neuron_533	del	1	2	183583489	190803979
86	Neuron	UVA_neuron_533	del	1	8	48124278	58665873
86	Neuron	UVA_neuron_533	del	1	8	119501301	124319612
86	Neuron	UVA_neuron_533	del	1	13	52941095	77518723
86	Neuron	UVA_neuron_533	del	1	13	99992280	110875528
86	Neuron	UVA_neuron_533	del	1	15	44563099	62406148
86	Neuron	UVA_neuron_533	del	1	15	84203290	100751400
86	Neuron	UVA_neuron_533	del	1	18	21049924	74475930
95	Neuron	UVA_neuron_586	dup	3	1	205736766	226358226
95	Neuron	UVA_neuron_586	dup	3	3	0	5450845
95	Neuron	UVA_neuron_586	dup	3	3	114067040	118003951
95	Neuron	UVA_neuron_586	dup	3	3	145905641	169326061
95	Neuron	UVA_neuron_586	dup	3	3	174097863	198022430
95	Neuron	UVA_neuron_586	dup	3	5	149920972	154153043
95	Neuron	UVA_neuron_586	dup	3	14	50806069	58796222
95	Neuron	UVA_neuron_586	dup	3	14	77944607	90194806
95	Neuron	UVA_neuron_586	dup	3	14	94514384	101310646
95	Neuron	UVA_neuron_586	dup	3	21	34866097	38458544
95	Neuron	UVA_neuron_617	del	1	13	61238839	69996157
95	Neuron	UVA_neuron_617	del	1	21	34252154	37853829
95	Neuron	UVA_neuron_630	del	1	1	52515583	109720251
95	Neuron	UVA_neuron_630	del	1	1	197191179	202697510
95	Neuron	UVA_neuron_630	del	1	3	71522747	84223958
95	Neuron	UVA_neuron_630	del	1	3	130066701	180154507
95	Neuron	UVA_neuron_630	del	1	3	190485047	198022430
95	Neuron	UVA_neuron_630	del	1	7	16471244	22961576
95	Neuron	UVA_neuron_630	del	1	9	32615299	37996021
95	Neuron	UVA_neuron_630	del	1	15	92004944	98897575
95	Neuron	UVA_neuron_630	del	1	21	18062271	44311358
95	Neuron	UVA_neuron_636	del	1	3	163439789	172377702
95	Neuron	UVA_neuron_636	del	1	7	86782458	95312683
95	Neuron	UVA_neuron_636	del	1	11	25429693	28436562
95	Neuron	UVA_neuron_647	del	1	4	4184323	10813482
95	Neuron	UVA_neuron_647	del	1	4	36015173	42882320
95	Neuron	UVA_neuron_647	del	1	4	184962175	191154276

Age (yrs)	Cell type	Cell	Type	CN state	Chrom	Start base	End base
95	Neuron	UVA_neuron_647	del	1	18	45344626	78077248

Appendix B
File Locations of Sequencing Data from Cells Sequenced in Chapters 2 and 3

Table B.1: Locations of human and mouse brain scWGS data generated for analysis in Chapters 2 and 3. FASTQs are organized in directories by sequencing run, WGA method, species, individual, and cell type. All directories are located on Rivanna in /nv/vol97/mcconnell_lab/cnvpipe/wgs_data/.

Directory	Dataset
cphg-Pool58_PicoPLEX_human_0.36yro-NeuNpos	Human
cphg-Pool58_PicoPLEX_human_26yro-NeuNneg	Human
cphg-Pool58_PicoPLEX_human_26yro-NeuNpos	Human
cphg-Pool58_PicoPLEX_human_95yro-NeuNpos	Human
lieber-Pool77_PicoPLEX_human_26yro-NeuNneg	Human
lieber-Pool77_PicoPLEX_human_26yro-NeuNpos	Human
lieber-Pool77_PicoPLEX_human_86yro-NeuNneg	Human
lieber-Pool77_PicoPLEX_human_86yro-NeuNpos	Human
lieber-Pool77_PicoPLEX_human_95yro-NeuNpos	Human
lieber-Pool78_PicoPLEX_human_26yro-NeuNneg	Human
lieber-Pool78_PicoPLEX_human_26yro-NeuNpos	Human
lieber-Pool78_PicoPLEX_human_86yro-NeuNneg	Human
lieber-Pool78_PicoPLEX_human_86yro-NeuNpos	Human
lieber-Pool78_PicoPLEX_human_95yro-NeuNpos	Human
lieber-Pool79_PicoPLEX_human_26yro-NeuNneg	Human
lieber-Pool79_PicoPLEX_human_26yro-NeuNpos	Human
lieber-Pool79_PicoPLEX_human_86yro-NeuNneg	Human
lieber-Pool79_PicoPLEX_human_86yro-NeuNpos	Human
lieber-Pool79_PicoPLEX_human_95yro-NeuNpos	Human
lieber-Pool80_PicoPLEX_human_26yro-NeuNneg	Human
lieber-Pool80_PicoPLEX_human_26yro-NeuNpos	Human
lieber-Pool80_PicoPLEX_human_86yro-NeuNneg	Human
lieber-Pool80_PicoPLEX_human_86yro-NeuNpos	Human
lieber-Pool80_PicoPLEX_human_95yro-NeuNpos	Human
lieber-Pool81_PicoPLEX_human_26yro-NeuNneg	Human
lieber-Pool81_PicoPLEX_human_26yro-NeuNpos	Human
lieber-Pool81_PicoPLEX_human_86yro-NeuNneg	Human
lieber-Pool81_PicoPLEX_human_86yro-NeuNpos	Human
lieber-Pool81_PicoPLEX_human_95yro-NeuNpos	Human
lieber-Pool82_PicoPLEX_human_26yro-NeuNneg	Human
lieber-Pool82_PicoPLEX_human_26yro-NeuNpos	Human
lieber-Pool82_PicoPLEX_human_86yro-NeuNneg	Human
lieber-Pool82_PicoPLEX_human_86yro-NeuNpos	Human
lieber-Pool82_PicoPLEX_human_95yro-NeuNpos	Human
novogene-Pool58_PicoPLEX_human_0.36yro-NeuNpos	Human
novogene-Pool58_PicoPLEX_human_26yro-NeuNneg	Human
novogene-Pool58_PicoPLEX_human_26yro-NeuNpos	Human
novogene-Pool58_PicoPLEX_human_95yro-NeuNpos	Human

Directory	Dataset
novogene-Pool71_PicoPLEX_human_0.36yro-NeuNpos	Human
novogene-Pool71_PicoPLEX_human_26yro-NeuNpos	Human
novogene-Pool72_PicoPLEX_human_0.36yro-NeuNpos	Human
novogene-Pool72_PicoPLEX_human_26yro-NeuNneg	Human
novogene-Pool72_PicoPLEX_human_26yro-NeuNpos	Human
novogene-Pool72_PicoPLEX_human_49yro-NeuNneg	Human
nygc-Pool60_PicoPLEX_human_0.36yro-NeuNpos	Human
nygc-Pool60_PicoPLEX_human_26yro-NeuNneg	Human
nygc-Pool60_PicoPLEX_human_26yro-NeuNpos	Human
nygc-Pool60_PicoPLEX_human_95yro-NeuNpos	Human
nygc-Pool61_PicoPLEX_human_0.36yro-NeuNpos	Human
nygc-Pool61_PicoPLEX_human_26yro-NeuNneg	Human
nygc-Pool61_PicoPLEX_human_26yro-NeuNpos	Human
nygc-Pool61_PicoPLEX_human_95yro-NeuNpos	Human
salk-Pool62_PicoPLEX_human_49yro-NeuNneg	Human
salk-Pool62_PicoPLEX_human_49yro-NeuNpos	Human
salk-Pool63_PicoPLEX_human_26yro-NeuNneg	Human
salk-Pool63_PicoPLEX_human_26yro-NeuNpos	Human
salk-Pool63_PicoPLEX_human_49yro-NeuNneg	Human
salk-Pool63_PicoPLEX_human_49yro-NeuNpos	Human
salk-Pool64_PicoPLEX_human_26yro-NeuNneg	Human
salk-Pool64_PicoPLEX_human_26yro-NeuNpos	Human
salk-Pool64_PicoPLEX_human_49yro-NeuNneg	Human
salk-Pool64_PicoPLEX_human_49yro-NeuNpos	Human
salk-Pool64_PicoPLEX_human_95yro-NeuNpos	Human
uva036-Pool00_PicoPLEX_human_0.36yro-NeuNpos	Human
uva042-Pool1_PicoPLEX_human_0.36yro-NeuNpos	Human
uva044-Pool3_PicoPLEX_human_0.36yro-NeuNpos	Human
uva045-Pool4_PicoPLEX_human_0.36yro-NeuNpos	Human
uva054-Pool8_PicoPLEX_human_0.36yro-NeuNpos	Human
uva055-Pool9_PicoPLEX_human_0.36yro-NeuNpos	Human
uva057-Pool10_PicoPLEX_human_0.36yro-NeuNpos	Human
uva058-Pool11_PicoPLEX_human_0.36yro-NeuNpos	Human
uva074-Pool12_PicoPLEX_human_0.36yro-NeuNpos	Human
uva075-Pool13_PicoPLEX_human_0.36yro-NeuNpos	Human
uva076-Pool14_PicoPLEX_human_0.36yro-NeuNpos	Human
uva077-Pool15_PicoPLEX_human_0.36yro-NeuNpos	Human
uva095-Pool32_PicoPLEX_human_0.36yro-NeuNpos	Human
uva098-Pool34_PicoPLEX_human_26yro-NeuNpos	Human
uva099-Pool35_PicoPLEX_human_26yro-NeuNpos	Human
uva100-Pool36_PicoPLEX_human_26yro-NeuNpos	Human
uva101-Pool37_PicoPLEX_human_26yro-NeuNpos	Human

Directory	Dataset
uva102-Pool38_PicoPLEX_human_26yro-NeuNpos	Human
uva109-Pool39_PicoPLEX_human_26yro-NeuNpos	Human
uva110-Pool40_PicoPLEX_human_26yro-NeuNpos	Human
uva111-Pool41_PicoPLEX_human_26yro-NeuNpos	Human
uva112-Pool42_PicoPLEX_human_26yro-NeuNpos	Human
uva113-Pool43_PicoPLEX_human_0.36yro-NeuNpos	Human
uva113-Pool43_PicoPLEX_human_26yro-NeuNpos	Human
uva114-Pool44_PicoPLEX_human_26yro-NeuNpos	Human
uva121-Pool27_PicoPLEX_human_0.36yro-NeuNpos	Human
uva123-Pool29_PicoPLEX_human_0.36yro-NeuNpos	Human
uva128-Pool48_PicoPLEX_human_95yro-NeuNpos	Human
uva129-Pool47_PicoPLEX_human_95yro-NeuNpos	Human
uva130-Pool46_PicoPLEX_human_95yro-NeuNpos	Human
uva136-Pool49_PicoPLEX_human_26yro-NeuNpos	Human
uva136-Pool49_PicoPLEX_human_95yro-NeuNpos	Human
uva137-Pool50_PicoPLEX_human_95yro-NeuNpos	Human
uva138-Pool51_PicoPLEX_human_95yro-NeuNpos	Human
uva150-Pool67_PicoPLEX_human_0.36yro-NeuNpos	Human
uva150-Pool67_PicoPLEX_human_26yro-NeuNpos	Human
uva151-Pool68_PicoPLEX_human_26yro-NeuNneg	Human
uva151-Pool68_PicoPLEX_human_26yro-NeuNpos	Human
uva152-Pool69_PicoPLEX_human_26yro-NeuNneg	Human
uva152-Pool69_PicoPLEX_human_26yro-NeuNpos	Human
uva153-Pool70_PicoPLEX_human_26yro-NeuNpos	Human
uva155-Pool73_PicoPLEX_human_26yro-NeuNpos	Human
uva156-Pool74_PicoPLEX_human_26yro-NeuNneg	Human
uva156-Pool74_PicoPLEX_human_26yro-NeuNpos	Human
uva156-Pool74_PicoPLEX_human_49yro-NeuNneg	Human
unc-Pool89_MALBAC-LAMP_mouse_Top1KOmouse	Mouse
unc-Pool89_MALBAC-LAMP_mouse_Top1WTmouse	Mouse
uva387-Pool93_MALBAC-LAMP_mouse_p53KOmouse1	Mouse
uva387-Pool93_MALBAC-LAMP_mouse_WTmouse1	Mouse
uvaUnknownRun-July2018Pool-Riptide768_MALBAC-LAMP_mouse_htau	Mouse
uvaUnknownRun-July2018Pool-Riptide768_MALBAC-LAMP_mouse_htauKO	Mouse
uvaUnknownRun-July2018Pool-Riptide768_MALBAC-LAMP_mouse_p53hetmouse	Mouse
uvaUnknownRun-July2018Pool-Riptide768_MALBAC-LAMP_mouse_p53KOmouse2	Mouse
uvaUnknownRun-July2018Pool-Riptide768_MALBAC-LAMP_mouse_WTmouse1	Mouse
uvaUnknownRun-July2018Pool-Riptide768_MALBAC-LAMP_mouse_WTmouse2	Mouse
uvaUnknownRun-July2018Pool-Riptide768_MALBAC-LAMP_mouse_WTmouse3	Mouse

References

Alt, F.W., and Schwer, B. (2018). DNA double-strand breaks as drivers of neural genomic change, function, and disease. DNA Repair (Amst) *71*, 158-163.

Anacker, C., and Hen, R. (2017). Adult hippocampal neurogenesis and cognitive flexibility - linking memory and mood. Nature reviews Neuroscience *18*, 335-346.

Andorfer, C., Acker, C.M., Kress, Y., Hof, P.R., Duff, K., and Davies, P. (2005). Cell-cycle reentry and cell death in transgenic mice expressing nonmutant human tau isoforms. The Journal of neuroscience : the official journal of the Society for Neuroscience *25*, 5446-5454.

Andorfer, C., Kress, Y., Espinoza, M., de Silva, R., Tucker, K.L., Barde, Y.A., Duff, K., and Davies, P. (2003). Hyperphosphorylation and aggregation of tau in mice expressing normal human tau isoforms. J Neurochem *86*, 582-590.

Arendt, T., Bruckner, M.K., and Losche, A. (2015). Regional mosaic genomic heterogeneity in the elderly and in Alzheimer's disease as a correlate of neuronal vulnerability. Acta Neuropathol *130*, 501-510.

Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., Pletikos, M., Pattni, R., Chen, B.J., Venturini, E., *et al.* (2018). Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science *359*, 550-555.

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., *et al.* (2011). Somatic

retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Barnes, D.E., Stamp, G., Rosewell, I., Denzel, A., and Lindahl, T. (1998). Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. Current Biology *8*, 1395-1398.

Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B., *et al.* (2012). Genome-wide copy number analysis of single cells. Nature protocols *7*, 1024-1041.

Bauman, J.G., Wiegant, J., Borst, P., and van Duijn, P. (1980). A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. Exp Cell Res *128*, 485-490.

Bedrosian, T.A., Quayle, C., Novaresi, N., and Gage, F.H. (2018). Early life experience drives structural variation of neural genomes in mice. Science *359*, 1395-1399.

Benaglia, T., Chauveau, D., Hunter, D.R., and Young, D.S. (2009). mixtools: An R Package for Analyzing Mixture Models. 2009 *32*, 29.

Bhardwaj, R.D., Curtis, M.A., Spalding, K.L., Buchholz, B.A., Fink, D., Bjork-Eriksson, T., Nordborg, C., Gage, F.H., Druid, H., Eriksson, P.S., *et al.* (2006). Neocortical neurogenesis in humans is restricted to development. Proceedings of the National Academy of Sciences of the United States of America *103*, 12564-12568.

Boldog, E., Bakken, T.E., Hodge, R.D., Novotny, M., Aevermann, B.D., Baka, J., Borde, S., Close, J.L., Diez-Fuertes, F., Ding, S.L., *et al.* (2018). Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type. Nature neuroscience *21*, 1185-1195.

Brennand, K.J., Marchetto, M.C., Benvenisty, N., Brustle, O., Ebert, A., Izpisua Belmonte, J.C., Kaykas, A., Lancaster, M.A., Livesey, F.J., McConnell, M.J., *et al.* (2015). Creating Patient-Specific Neural Cells for the In Vitro Study of Brain Disorders. Stem cell reports *5*, 933-945.

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., *et al.* (2014). Increased l1 retrotransposition in the neuronal genome in schizophrenia. Neuron *81*, 306-313.

Burbulis, I.E., Wierman, M.B., Wolpert, M., Haakenson, M., Lopes, M.B., Schiff, D., Hicks, J., Loe, J., Ratan, A., and McConnell, M.J. (2018). Improved molecular karyotyping in glioblastoma. Mutation research *811*, 16-26.

Bushman, D.M., Kaeser, G.E., Siddoway, B., Westra, J.W., Rivera, R.R., Rehen, S.K., Yung, Y.C., and Chun, J. (2015). Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. eLife *4*.

Cai, X., Evrony, G.D., Lehmann, H.S., Elhosary, P.C., Mehta, B.K., Poduri, A., and Walsh, C.A. (2014). Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell reports *8*, 1280-1289.

Campbell, I.M., Shaw, C.A., Stankiewicz, P., and Lupski, J.R. (2015). Somatic mosaicism: implications for disease and transmission genetics. Trends Genet *31*, 382-392.

Carvalho, C.M., and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nature reviews Genetics *17*, 224-238.

Caspersson, T., Zech, L., Johansson, C., and Modest, E.J. (1970). Identification of human chromosomes by DNA-binding fluorescent agents. Chromosoma *30*, 215-227.

Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature methods *12*, 966-968.

Christian, K.M., Song, H., and Ming, G.L. (2014). Functions and dysfunctions of adult hippocampal neurogenesis. Annual review of neuroscience *37*, 243-262.

Chronister, W.D., Burbulis, I.E., Wierman, M.B., Wolpert, M.J., Haakenson, M.F., Smith, A.C.B., Kleinman, J.E., Hyde, T.M., Weinberger, D.R., Bekiranov, S., *et al.* (2019). Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell reports *26*, 825-835 e827.

Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., *et al.* (2011). Variation in genome-wide mutation rates within and between human families. Nature genetics *43*, 712-714.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., *et al.* (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704-712.

Contreras, D. (2004). Electrophysiological classes of neocortical neurons. Neural networks : the official journal of the International Neural Network Society *17*, 633-646.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127-1131.

D'Gama, A.M., and Walsh, C.A. (2018). Somatic mosaicism and neurodevelopmental disease. Nature neuroscience *21*, 1504-1514.

Dawson, H.N., Ferreira, A., Eyster, M.V., Ghoshal, N., Binder, L.I., and Vitek, M.P. (2001). Inhibition of neuronal maturation in primary hippocampal neurons from tau deficient mice. J Cell Sci *114*, 1179-1187.

Del Pino, I., Rico, B., and Marin, O. (2018). Neural circuit dysfunction in mouse models of neurodevelopmental disorders. Current opinion in neurobiology *48*, 174-182.

Duff, K., Knight, H., Refolo, L.M., Sanders, S., Yu, X., Picciano, M., Malester, B., Hutton, M., Adamson, J., Goedert, M., *et al.* (2000). Characterization of pathology in transgenic mice over-expressing human genomic and cDNA tau transgenes. Neurobiol Dis *7*, 87-98.

Erwin, J.A., Paquola, A.C., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I., Butcher, C.R., Herdy, J.R., *et al.* (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nature neuroscience.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A., *et al.* (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483-496.

Evrony, G.D., Lee, E., Park, P.J., and Walsh, C.A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. eLife *5*.

Ferguson-Smith, M.A. (2015). History and evolution of cytogenetics. Mol Cytogenet *8*, 19.

Frank, K.M., Sharpless, N.E., Gao, Y., Sekiguchi, J.M., Ferguson, D.O., Zhu, C., Manis, J.P., Horner, J., DePinho, R.A., and Alt, F.W. (2000). DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. Molecular cell *5*, 993-1002.

Frank, S.A. (2014). Somatic mosaicism and disease. Current biology : CB *24*, R577-R581.

Freed, D., and Pevsner, J. (2016). The Contribution of Mosaic Variants to Autism Spectrum Disorder. PLoS genetics *12*, e1006245.

Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M., *et al.* (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature *506*, 179-184.

Gao, Y., Sun, Y., Frank, K.M., Dikkes, P., Fujiwara, Y., Seidl, K.J., Sekiguchi, J.M., Rathbun, G.A., Swat, W., Wang, J., *et al.* (1998). A Critical Role for DNA End-Joining Proteins in Both Lymphogenesis and Neurogenesis. Cell *95*, 891-902.

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

Geschwind, D.H., and Rakic, P. (2013). Cortical evolution: judge the brain by its cover. Neuron *80*, 633-647.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell *110*, 315-325.

Hall, J.G. (1988). Review and hypotheses: somatic mosaicism: observations related to clinical genetics. Am J Hum Genet *43*, 355-363.

Harbom, L.J., Michel, N., and McConnell, M.J. (2018). Single-cell analysis of diversity in human stem cell-derived neurons. Cell and tissue research *371*, 171-179.

Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., Xie, X.S., *et al.* (2013). Genome analyses of single human oocytes. Cell *155*, 1492-1506.

Huang, N., Lee, I., Marcotte, E.M., and Hurles, M.E. (2010). Characterising and Predicting Haploinsufficiency in the Human Genome. PLoS genetics *6*.

Insel, T.R. (2014). Brain somatic mutations: the dark matter of psychiatric genetics? Molecular psychiatry *19*, 156-158.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., *et al.* (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216-221.

Iourov, I.Y., Vorsanova, S.G., Liehr, T., Kolotii, A.D., and Yurov, Y.B. (2009). Increased chromosome instability dramatically disrupts neural genome integrity and mediates cerebellar degeneration in the ataxia-telangiectasia brain. Human molecular genetics *18*, 2656-2669.

Jamuar, S.S., Lam, A.T., Kircher, M., D'Gama, A.M., Wang, J., Barry, B.J., Zhang, X., Hill, R.S., Partlow, J.N., Rozzo, A., *et al.* (2014). Somatic mutations in cerebral cortical malformations. The New England journal of medicine *371*, 733-743.

Jansen, L.A., Mirzaa, G.M., Ishak, G.E., O'Roak, B.J., Hiatt, J.B., Roden, W.H., Gunter, S.A., Christian, S.L., Collins, S., Adams, C., *et al.* (2015). PI3K/AKT pathway mutations cause a spectrum of brain malformations from megalencephaly to focal cortical dysplasia. Brain *138*, 1613-1628.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science *258*, 818-821.

Kathuria, K., and Ratan, A. (2018).

King, I.F., Yandava, C.N., Mabb, A.M., Hsiao, J.S., Huang, H.S., Pearson, B.L., Calabrese, J.M., Starmer, J., Parker, J.S., Magnuson, T., *et al.* (2013). Topoisomerases facilitate transcription of long genes linked to autism. Nature *501*, 58-62.

Knouse, K.A., Wu, J., and Amon, A. (2016). Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome research *26*, 376-384.

Knouse, K.A., Wu, J., Whittaker, C.A., and Amon, A. (2014). Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. Proceedings of the National Academy of Sciences of the United States of America *111*, 13409-13414.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., *et al.* (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471-475.

Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.L., Chen, S., *et al.* (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science *352*, 1586-1590.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome biology *15*, R84.

Lee, J.H., Huynh, M., Silhavy, J.L., Kim, S., Dixon-Salazar, T., Heiberg, A., Scott, E., Bafna, V., Hill, K.J., Collazo, A., *et al.* (2012). De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nature genetics *44*, 941-945.

Lee, M.H., Siddoway, B., Kaeser, G.E., Segota, I., Rivera, R., Romanow, W.J., Liu, C.S., Park, C., Kennedy, G., Long, T., *et al.* (2018). Somatic APP gene recombination in Alzheimer's disease and normal neurons. Nature *563*, 639-645.

Lein, E.S., Belgard, T.G., Hawrylycz, M., and Molnar, Z. (2017). Transcriptomic Perspectives on Neocortical Structure, Development, Evolution, and Disease. Annual review of neuroscience *40*, 629-652.

Lejeune, J., Gautier, M., and Turpin, R. (1959). [Study of somatic chromosomes from 9 mongoloid children]. C R Hebd Seances Acad Sci *248*, 1721-1722.

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

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

Lim, J.S., Kim, W.I., Kang, H.C., Kim, S.H., Park, A.H., Park, E.K., Cho, Y.W., Kim, S., Kim, H.M., Kim, J.A., *et al.* (2015). Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. Nature medicine *21*, 395-400.

Lodato, M.A., Rodin, R.E., Bohrson, C.L., Coulter, M.E., Barton, A.R., Kwon, M., Sherman, M.A., Vitzthum, C.M., Luquette, L.J., Yandava, C.N., *et al.* (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. Science *359*, 555-559.

Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S., Chittenden, T.W., D'Gama, A.M., Cai, X., *et al.* (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science *350*, 94-98.

Lui, J.H., Hansen, D.V., and Kriegstein, A.R. (2011). Development and evolution of the human neocortex. Cell *146*, 18-36.

Lupski, J.R. (2015). Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environmental and molecular mutagenesis *56*, 419-436.

Mabb, A.M., Kullmann, P.H., Twomey, M.A., Miriyala, J., Philpot, B.D., and Zylka, M.J. (2014). Topoisomerase 1 inhibition reversibly impairs synaptic function. Proceedings of the National Academy of Sciences of the United States of America *111*, 17290-17295.

Mabb, A.M., Simon, J.M., King, I.F., Lee, H.M., An, L.K., Philpot, B.D., and Zylka, M.J. (2016). Topoisomerase 1 Regulates Gene Expression in Neurons through

Cleavage Complex-Dependent and -Independent Mechanisms. PloS one *11*, e0156439.

MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic acids research *42*, D986-992.

Madabhushi, R., Gao, F., Pfenning, A.R., Pan, L., Yamakawa, S., Seo, J., Rueda, R., Phan, T.X., Yamakawa, H., Pao, P.C., *et al.* (2015). Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. Cell *161*, 1592-1605.

Marin-Valencia, I., Guerrini, R., and Gleeson, J.G. (2014). Pathogenetic mechanisms of focal cortical dysplasia. Epilepsia *55*, 970-978.

Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., *et al.* (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nature genetics *49*, 27-35.

McConnell, M.J., Kaushal, D., Yang, A.H., Kingsbury, M.A., Rehen, S.K., Treuner, K., Helton, R., Annas, E.G., Chun, J., and Barlow, C. (2004). Failed clearance of aneuploid embryonic neural progenitor cells leads to excess aneuploidy in the Atm-deficient but not the Trp53-deficient adult cerebral cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience *24*, 8090-8096.

McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R.S., Vermeesch, J.R., Hall, I.M., *et al.* (2013). Mosaic copy number variation in human neurons. Science *342*, 632-637.

McConnell, M.J., Moran, J.V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J.A., Fasching, L., Flasch, D.A., Freed, D., *et al.* (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science *356*.

Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nature protocols *8*, 1551-1566.

Mirzaa, G.M., Campbell, C.D., Solovieff, N., Goold, C.P., Jansen, L.A., Menon, S., Timms, A.E., Conti, V., Biag, J.D., Olds, C., *et al.* (2016). Association of MTOR Mutations With Developmental Brain Disorders, Including Megalencephaly, Focal Cortical Dysplasia, and Pigmentary Mosaicism. JAMA neurology *73*, 836-845.

Moorhead, P.S., Nowell, P.C., Mellman, W.J., Battips, D.M., and Hungerford, D.A. (1960). Chromosome preparations of leukocytes cultured from human peripheral blood. Exp Cell Res *20*, 613-616.

Morrison, J.H., and Baxter, M.G. (2012). The ageing cortical synapse: hallmarks and implications for cognitive decline. Nature reviews Neuroscience *13*, 240-250.

Morrow, E.M. (2010). Genomic copy number variation in disorders of cognitive development. Journal of the American Academy of Child and Adolescent Psychiatry *49*, 1091-1104.

Muggeo, V.M., and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. Bioinformatics *27*, 161-166.

Muller, P.A., and Vousden, K.H. (2013). p53 mutations in cancer. Nat Cell Biol *15*, 2-8.

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

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

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., *et al.* (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90-94.

Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., *et al.* (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318-1323.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557-572.

Pack, S.D., Weil, R.J., Vortmeyer, A.O., Zeng, W., Li, J., Okamoto, H., Furuta, M., Pak, E., Lubensky, I.A., Oldfield, E.H., *et al.* (2005). Individual adult human neurons display aneuploidy: detection by fluorescence in situ hybridization and single neuron PCR. Cell Cycle *4*, 1758-1760.

Pakkenberg, B., Pelvig, D., Marner, L., Bundgaard, M.J., Gundersen, H.J., Nyengaard, J.R., and Regeur, L. (2003). Aging and the human neocortex. Experimental gerontology *38*, 95-99.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., *et al.* (2007). Diet and the evolution of human amylase gene copy number variation. Nature genetics *39*, 1256-1260.

Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS genetics *9*.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., *et al.* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature genetics *20*, 207-211.

Piotrowski, A., Bruder, C.E., Andersson, R., Diaz de Stahl, T., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., *et al.* (2008). Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat *29*, 1118-1124.

Poduri, A., Evrony, G.D., Cai, X., Elhosary, P.C., Beroukhim, R., Lehtinen, M.K., Hills, L.B., Heinzen, E.L., Hill, A., Hill, R.S., *et al.* (2012). Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron *74*, 41-48.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rakic, P. (2006). Neuroscience. No more cortical neurons for you. Science *313*, 928-929.

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

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., *et al.* (2006). Global variation in copy number in the human genome. Nature *444*, 444-454.

Rehen, S.K., McConnell, M.J., Kaushal, D., Kingsbury, M.A., Yang, A.H., and Chun, J. (2001). Chromosomal variation in neurons of the developing and adult mammalian nervous system. Proceedings of the National Academy of Sciences of the United States of America *98*, 13361-13366.

Rehen, S.K., Yung, Y.C., McCreight, M.P., Kaushal, D., Yang, A.H., Almeida, B.S., Kingsbury, M.A., Cabral, K.M., McConnell, M.J., Anliker, B., *et al.* (2005). Constitutional aneuploidy in the normal human brain. The Journal of neuroscience : the official journal of the Society for Neuroscience *25*, 2176-2180.

Rohrback, S., April, C., Kaper, F., Rivera, R.R., Liu, C.S., Siddoway, B., and Chun, J. (2018a). Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. Proceedings of the National Academy of Sciences of the United States of America *115*, 10804-10809.

Rohrback, S., Siddoway, B., Liu, C.S., and Chun, J. (2018b). Genomic mosaicism in the developing and adult brain. Dev Neurobiol *78*, 1026-1048.

Rowley, J.D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature *243*, 290-293.

Sala Frigerio, C., Lau, P., Troakes, C., Deramecourt, V., Gele, P., Van Loo, P., Voet, T., and De Strooper, B. (2015). On the identification of low allele frequency mosaic

mutations in the brains of Alzheimer's disease patients. Alzheimers Dement *11*, 1265-1276.

Schrock, E., du Manoir, S., Veldman, T., Schoell, B., Wienberg, J., Ferguson-Smith, M.A., Ning, Y., Ledbetter, D.H., Bar-Am, I., Soenksen, D., *et al.* (1996). Multicolor spectral karyotyping of human chromosomes. Science *273*, 494-497.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics *6*, 461-464.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., *et al.* (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445-449.

Seshan, V.E., and Olshen, A. (2018). DNAcopy: DNA copy number data analysis.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., *et al.* (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic acids research *43*, W589-598.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer *20*, 399-407.

Speicher, M.R., and Carter, N.P. (2005). The new cytogenetics: blurring the boundaries with molecular biology. Nature reviews Genetics *6*, 782-792.

Speicher, M.R., Gwyn Ballard, S., and Ward, D.C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nature genetics *12*, 368-375.

Spitzer, N.C. (2017). Neurotransmitter Switching in the Developing and Adult Brain. Annual review of neuroscience *40*, 1-19.

Sudhof, T.C. (2017). Synaptic Neurexin Complexes: A Molecular Code for the Logic of Neural Circuits. Cell *171*, 745-769.

Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., *et al.* (2015). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one *6*, e21800.

Tjio, J.H., and Levan, A. (1956). The Chromosome Number of Man. Hereditas *42*, 1-6.

Uhlen, M., Hallstrom, B.M., Lindskog, C., Mardinoglu, A., Ponten, F., and Nielsen, J. (2016). Transcriptomics resources of human tissues and organs. Molecular systems biology *12*, 862.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sanchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M., *et al.* (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228-239.

van den Bos, H., Spierings, D.C., Taudt, A.S., Bakker, B., Porubsky, D., Falconer, E., Novoa, C., Halsema, N., Kazemier, H.G., Hoekstra-Wakker, K., *et al.* (2016). Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons. Genome biology *17*, 116.

Van Hoesen, G.W., Hyman, B.T., and Damasio, A.R. (1991). Entorhinal cortex pathology in Alzheimer's disease. Hippocampus *1*, 1-8.

van Slegtenhorst, M., de Hoogt, R., Hermans, C., Nellist, M., Janssen, B., Verhoef, S., Lindhout, D., van den Ouweland, A., Halley, D., Young, J., *et al.* (1997). Identification of the tuberous sclerosis gene TSC1 on chromosome 9q34. Science *277*, 805-808.

Vijg, J. (2014). Somatic mutations, genome mosaicism, cancer and aging. Current opinion in genetics & development *26*, 141-149.

von Bartheld, C.S., Bahney, J., and Herculano-Houzel, S. (2016). The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. The Journal of comparative neurology *524*, 3865-3895.

Wei, P.C., Chang, A.N., Kao, J., Du, Z., Meyers, R.M., Alt, F.W., and Schwer, B. (2016). Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. Cell *164*, 644-655.

Weissman, I.L., and Gage, F.H. (2016). A Mechanism for Somatic Brain Mosaicism. Cell *164*, 593-595.

Wierman, M.B., Burbulis, I.E., Chronister, W.D., Bekiranov, S., and McConnell, M.J. (2017). Single-Cell CNV Detection in Human Neuronal Nuclei. In Genomic Mosaicism in Neurons and Other Cell Types, J.M. Frade, and F.H. Gage, eds. (New York, NY: Springer New York), pp. 109-131.

Williams, A.B., and Schumacher, B. (2016). p53 in the DNA-Damage-Repair Process. Cold Spring Harb Perspect Med *6*.

Wilson, T.E., Arlt, M.F., Park, S.H., Rajendran, S., Paulsen, M., Ljungman, M., and Glover, T.W. (2015). Large transcription units unify copy number variants and common fragile sites arising under replication stress. Genome research *25*, 189-200.

Yurov, Y.B., Iourov, I.Y., Vorsanova, S.G., Demidova, I.A., Kravetz, V.S., Beresheva, A.K., Kolotii, A.D., Monakchov, V.V., Uranova, N.A., Vostrikov, V.M., *et al.* (2008). The schizophrenia brain exhibits low-level aneuploidy involving chromosome 1. Schizophr Res *98*, 139-147.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nature reviews Genetics *16*, 172-183.

Zhang, L., and Vijg, J. (2018). Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. Annual review of genetics *52*, 397-419.

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

Zylka, M.J., Simon, J.M., and Philpot, B.D. (2015). Gene length matters in neurons. Neuron *86*, 353-355.