Development of Computational Methods for Identifying Mosaic Copy Number Variation with Single-cell Sequencing

by

Chen Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Ryan E. Mills, Chair
Professor Margit Burmeister
Professor Jeffrey Kidd
Assistant Professor Joshua Welch
Professor Xiaoquan Wen

Chen Sun

cnsun@umich.edu

ORCID iD: 0000-0003-0657-207X

Dedication

To everyone who has been part of this wonderful journey!

Acknowledgements

I would like to express my sincere gratitude to my PhD advisor, Dr. Ryan Mills, for his unwavering support and guidance throughout my doctoral journey. His deep knowledge and expertise in the field of bioinformatics has been invaluable in shaping my research questions and methodology. Ryan's constructive feedback and insightful comments have pushed me to think critically and to continuously improve my work. I am grateful for his encouragement and mentorship, which have played a crucial role in my academic and personal growth. I am honored to have worked with such an exceptional advisor and mentor, and I will continue to carry his lessons with me throughout my academic and professional career.

I would like to extend my appreciation to the collaborators who have contributed to the success of my research. Dr. Michael McConnel and Dr. Kunal Kathuria from the Lieber Institute, and Dr. Jeffrey Kidd and Dr. John Moran from the Department of Human Genetics, have been invaluable partners in my research journey. Their expertise, insights, and willingness to engage with my ideas have enriched my work and led to exciting discoveries. I am fortunate to have had the opportunity to collaborate with such talented individuals.

I would like to thank all the past and current members of Mills lab who have contributed to my growth as a researcher and to the success of my project. I have learned so much from their diverse backgrounds and experiences, and I am grateful for their support and encouragement throughout my graduate studies. I am thankful to having worked with Dr. Weichen Zhou, Dr. Yifan Wang, Dr. Marcus Sherman, Dr. Alex Weber, Dr. Tony Chun, Steve Ho and Wenjin Gu. I

am proud to have been a part of such an exceptional group, and I wish them all the best in their future.

I also want to thank my friends in Ann Arbor, in Michigan, in the US and in China. I feel incredibly fortunate to have had their company, wisdom, and motivation during the best and worst times of my PhD journey. The memories we have shared together, from exploring Ann Arbor's hidden gems to traveling to new places, have been among the most meaningful and enjoyable experiences of my life. I am grateful for their friendship, and I look forward to continuing to share life's joys and challenges with them.

I am deeply grateful for the constant support and encouragement of my parents throughout my graduate studies, despite being thousands of miles away. Their consistent presence and reassurance have helped me navigate the challenges of graduate school and given me the strength to persevere. Their support has been particularly invaluable during the pandemic when boundaries seem to be everywhere. I am deeply appreciative for their love, and I hope to make them proud in all my future endeavors.

Finally, I want to thank my wife for her endless support, encouragement, and love during my graduate studies. She has been my constant source of inspiration and strength, and her unwavering belief in me has kept me going through the toughest times. Her presence in my life has been invaluable, providing me with stability and comfort when I needed it the most. I am deeply grateful for the sacrifices she has made and the endless hours she has spent supporting me. I look forward to a lifetime of shared adventures and memories with her.

Table of Contents

List of Tables

List of Figures

Abstract

Somatic mutations are genetic variations that occur in a subset of cells during an individual's lifespan and have implications for various biological processes and diseases. While high-throughput sequencing technology has made it possible to profile these mutations genome-wide, detecting somatic mutations from bulk tissue samples poses significant challenges. Single-cell DNA sequencing is a powerful technique that can reveal the presence and extent of somatic genetic variations, such as copy number variations (CNVs), at the resolution of individual cells; however, this method also presents technical challenges due to low coverage, high noise, amplification bias, and artifacts.

To address these challenges, this dissertation presents three main contributions. First, I developed SCOVAL, a method that verifies the single-cell CNV calls from a read coverage-based approach with additional phased loss-of-heterozygosity information. This method was applied to 2,125 frontal cortical neurons from a neurotypical human brain and discovered 226 CNV neurons, including a novel class of neurons with complex karyotypes characterized by whole or substantial losses of multiple chromosomes. Second, it is difficult to directly validate CNVs present in only a single cell and thus ground truth sets are problematic to obtain. I therefore developed a single-cell CNV simulator to generate realistic single-cell DNA sequencing data with predefined somatic CNVs, which was used to benchmark existing tools for somatic CNV detection and to provide a ground truth dataset for the development and evaluation of new methods. Third, I developed a deep learning-based method named ScovalNN that leverages Long Short-Term Memory (LSTM) neural networks to detect somatic CNVs from

single-cell DNA sequencing data, and showed that it performs well on both simulated and real data and outperforms existing methods in terms of multiple metrics.

Overall, this dissertation provides significant advancements in the field of somatic CNV detection using single-cell DNA sequencing data and highlights the challenges that still need to be addressed. These contributions have important implications for the understanding of somatic mutations in non-cancer diseases and the development of personalized medicine.

Chapter 1 Introduction

1.1 Somatic mosaicism

Somatic mosaicism refers to the occurrence of genetically diverse cells within an individual, derived from a postzygotic mutation (Freed et al., 2014). Somatic mosaic mutations are present in only a subset of cells of a single individual. These mutations can arise spontaneously or as a result of exposure to environmental factors (Oota, 2020). Compared to inherited mutations, which are present in every cell of an individual and are passed down to subsequent generations, somatic mosaic mutations may affect only a limited portion of the body and are not transmitted to progeny ("Definition of somatic mutation - NCI Dictionary of Cancer Terms - NCI," 2011).

The spectrum of somatic mutations encompasses a wide range of variants, ranging from small-scale single nucleotide variants (SNVs) (Wang et al., 2021; Zaidi et al., 2020) to large-scale structural variations, such as the loss or gain of entire chromosomes (aneuploidy) (Valind et al., 2013). These mutations have the potential to alter cellular function through changes in gene expression and regulation (Ding et al., 2015; Jia and Zhao, 2017; Mathelier et al., 2015). Different types of somatic mutations include single nucleotide variations (SNVs), insertions and deletions (indels), copy number variations (CNVs), and structural variations (SVs). A noteworthy type of structural variation is caused by long interspersed nucleotide element 1 (LINE-1 or L1), which occurs due to the insertion of a copy of the L1 retrotransposable element into a new location in the genome, leading to activation or inactivation of genes and subsequent cellular dysfunction and disease (Brouha et al., 2003; Kazazian and Moran, 2017, 1998; Larson

et al., 2018; Moran et al., 1996; Ostertag et al., 2000; Ostertag and Kazazian, 2001; Singer et al., 2010; Smit, 1999; Zhao et al., 2019; Zhou et al., 2020). To advance our understanding of the underlying mechanisms of disease and develop effective treatments, it is essential to thoroughly investigate the full spectrum of somatic mosaic mutations, from small-scale SNVs to large-scale structural variations.

Somatic mutations also play a crucial role in the development and progression of various diseases, both cancerous and non-cancerous. In particular, cancer is heavily influenced by somatic mutations (Martincorena and Campbell, 2015). These mutations accumulate over time, driving the evolution of malignant cells and contributing to the development of tumors (Fernández et al., 2016). In tumor tissues, somatic mutations can alter the expression and regulation of key genes, leading to cellular transformation and the hallmark features of cancer, such as uncontrolled cell growth and the ability to evade normal cellular controls (Nenclares and Harrington, 2020). Cancer evolves from somatic mutations through a process of clonal expansion, genetic diversification, and clonal selection. It is derived from a single ancestor cell, whose progenies that are positively selected by acquisition of driver mutations (Greaves and Maley, 2012; Greenman et al., 2007). Clonal expansion driven by the acquisition of different mutations leads to the development of a population of cancer cells (Merlo et al., 2006). Such expansions are accompanied by passenger alterations that may transform into driver aberrations if the selective pressures change. During tumor progression, the mutational rate changes, leading to genetic heterogeneity within the tumor (Greaves and Maley, 2012; Hanahan and Weinberg, 2011). Clonal evolution has been demonstrated in multiple cancers. Using next-generation sequencing, the tumor clonal structures and evolutionary history of several cancers have been identified (Gerlinger et al., 2012; Gundem et al., 2015; Harbst et al., 2016; Jamal-Hanjani et al.,

2017; Roerink et al., 2018; Yates et al., 2015), directly confirming that the cancer tissues comprise highly heterogeneous cell populations in terms of somatic mutation.

On the other hand, somatic mutations also play a role in a diverse range of non-cancerous diseases. Autoimmune disorders, such as systemic lupus erythematosus (SLE), can be influenced by somatic mutations, as these mutations can trigger an immune response against the individual's own cells (Law et al., 2022). Neurodegenerative diseases, such as Huntington's disease (Roy et al., 2021), Alzheimer's disease (Miller et al., 2022) and Parkinson's disease (Veeriah et al., 2010), are also thought to be influenced by somatic mutations, which can contribute to the progressive degeneration of neural cells (Proukakis, 2020). In addition, somatic mutations have been linked to a variety of neurodevelopmental disorders (D'Gama and Walsh, 2018), including epileptic encephalopathies (Stosser et al., 2018), intellectual disability (Gilissen et al., 2014), Schizophrenia (Bundo et al., 2014) and autism spectrum disorder (ASD) (Dou et al., 2017). A previous study suggested that somatic mutations can contribute to 3-5% of simplex ASD diagnoses (D'Gama, 2021).

In addition, recent research has demonstrated that somatic mutations play a role in the aging process and can contribute to the causes of age-related diseases (Vijg, 2014). However, the exact relationship between aging and somatic mutation rate remains to be fully elucidated. Some studies found that the somatic mutation rate appears to remain constant during aging, leading to a gradual linear mutation accumulation over time (Manders et al., 2021). In contrast, other studies suggest that somatic mutation rates may be inversely correlated with lifespan across species and tissues, indicating a more complex relationship between aging and somatic mutations (Cagan et al., 2022; Chronister et al., 2019).

Overall, the relationship between somatic mutations and disease is complex and multi-faceted, encompassing both cancerous and non-cancerous conditions. A comprehensive understanding of this relationship is essential for the development of effective diagnostic and therapeutic strategies aimed at improving patient outcomes and quality of life.

1.2 Somatic mosaicism in human neurons

Neurons are produced by neural stem cells (NSCs), including neuroepithelial cells (NECs), radial glial cells (RGCs), basal progenitors (BPs), intermediate neuronal precursors (INPs), subventricular zone astrocytes, and subgranular zone radial astrocytes, each of which has a distinct role in the process of neurogenesis (Kandel et al., 2021). During neurogenesis, NSCs divide and differentiate to form mature neurons. Mature neurons are not capable of dividing after birth, and do not renew themselves. However, NSC populations that divide rapidly can also acquire somatic mutations due to errors in DNA replication. Subsequently, the clonal expansion of variant genomes can contribute to the somatic mosaicism (Malinverno et al., 2019).

Somatic single nucleotide variants (SNVs) can have significant effects on gene expression, cellular function and disease susceptibility (Huang and Lee, 2022). Somatic SNVs are particularly relevant for studying brain disorders, as the brain is composed of diverse cell types and regions that may harbor distinct mutational profiles. Several recent studies have applied advanced sequencing technologies and bioinformatics methods to detect and characterize somatic SNVs in the human brain. For example, Wang et al. presents a unified set of best practices to detect somatic SNVs in human brain tissue. They identified 43 bona fide somatic SNVs that range in variant allele fractions from ~ 0.005 to ~ 0.28 (Wang et al., 2021). Luquette et al. analyzed whole-genome sequencing data from 52 primary template-directed amplification

(PTA) amplified single neurons to identify SNVs and small indels. Their analysis confirms an increase in non-clonal somatic mutation in single neurons with age, but revises the estimated rate of this accumulation to 16 SNVs per year (Luquette et al., 2022). Miller et al. analyzed single-cell whole-genome sequencing data from 319 neurons from the prefrontal cortex and hippocampus of individuals with Alzheimer's disease and neurotypical control individuals. They found that somatic SNVs increase in individuals with Alzheimer's disease, with distinct molecular patterns (Miller et al., 2022). These studies demonstrate the power and potential of investigating somatic SNVs in the human brain to uncover novel insights into its development, diversity and disease.

Somatic mobile element insertions (MEIs) are a source of genomic variation in the human nervous system that may have implications for brain development and neuropsychiatric disorders (Bundo et al., 2014; Coufal et al., 2009; Evrony et al., 2012; Muotri et al., 2005). Somatic MEIs are generated by active retrotransposons, such as L1, Alu and SVA elements, that can transpose via an RNA intermediate in both germline and somatic cells. Several studies have used different approaches to identify and quantify somatic MEIs in human neurons. Erwin et al. reported that a subset of somatic L1-associated variants (SLAVs) comprises somatic deletions generated by L1 endonuclease cutting activity. SLAVs can present in crucial neural genes, and affect 44–63% of cells in the healthy brain (Erwin et al., 2016). Zhao et al. identified and validated somatic L1Hs insertions in both cortical neurons and non-brain tissues. They also explored the genomic patterns of somatic L1Hs insertions in neuronal and non-neuronal samples, and to investigate whether MeCP2 dysfunction could alter the distribution of L1Hs retrotransposition in patients with Rett syndrome (Zhao et al., 2019).

Somatic copy number variations (CNVs) are changes in the number of copies of a genomic region in a single cell, and they can range in size from hundreds to millions of base pairs (Gao et al., 2022; Knouse et al., 2016; Macaulay and Voet, 2014). They may have a significant impact on human disease and development, and may be caused by mechanisms such as retrotransposition (Sui and Peng, 2021; Turan et al., 2022). In recent years, there has been growing interest in studying somatic CNVs in human neurons, as they are believed to play a role in various aspects of neuronal function and behavior.

Researchers have used a variety of techniques, including single-cell sequencing and genomic arrays, to study the prevalence and characteristics of somatic CNVs in human neurons (Knouse et al., 2016; McConnell et al., 2013). These studies have revealed that somatic CNVs are relatively common (~8% - 41%) in human neurons and can vary greatly in size and genomic location. Some somatic CNVs have been shown to have significant effects on neuronal function and behavior, while others appear to be more benign (Zhang et al., 2009). Key findings from these studies include evidence of regional differences in the prevalence of somatic CNVs in human neurons (Turan et al., 2022), as well as a potential role for somatic CNVs in neurodevelopmental and neurodegenerative disorders (Maury and Walsh, 2021).

Regarding somatic CNVs, there are still several important questions that remain to be answered. For example, their frequency and distribution, and what factors influence them. In addition, a fundamental open question in neurodevelopmental genetics is whether and how somatic mosaicism may contribute to neuronal diversity within the neurotypical spectrum and in diseased brains. To study somatic mosaicism in human neurons, The National Institute of Mental Health (NIMH) has formed a network of 18 investigative teams representing 15 institutions called the Brain Somatic Mosaicism Network (BSMN) (McConnell et al., 2017). Researchers

can leverage various genomic technologies, including next-generation and long-read DNA sequencing technologies, single-cell genomics, and cutting-edge bioinformatics, to make it possible to determine the types and frequencies of somatic mutations within the human brain.

1.3 Single-cell DNA sequencing: technologies and applications

Before single-cell sequencing was invented, some early attempts on somatic mutation detection were performed on bulk tissue samples. Somatic mutation calling from bulk DNA sequencing data has been primarily studied in cancer research, where the sequencing data from tumor samples are compared to matched normal control samples to identify mutations. Tools such as Strelka (Saunders et al., 2012), VarScan2 (Koboldt et al., 2012), JointSNVMix (Roth et al., 2012), and MuTect (Cibulskis et al., 2013) have been developed to identify somatic mutations in these samples by comparing the mutant allele fractions between tumor and normal samples. These models also incorporate error filters to remove technical artifacts from sequencing data.

Targeted ultra-deep sequencing is a powerful technique to detect somatic mutations in cancer-related genes with high sensitivity. However, it also faces the challenge of distinguishing true mutations from technical artifacts that arise from various sources of error. Several methods have been developed to address this issue, such as RareVar (Hao et al., 2017) and RePlow (Kim et al., 2019). RareVar employs a position-specific error model to filter out false positives with a low allele fraction of 0.5%, while RePlow estimates the background error rate based on the distribution of sequencing depth and quality scores. These methods improve the accuracy of somatic variant detection in ultra-deep sequencing data, but they are mainly designed for bulk

sequencing. Bulk sequencing has inherent limitations in detecting rare or subclonal mutations due to the dilution effect and tumor heterogeneity (Ma et al., 2019). Therefore, alternative approaches such as single-cell sequencing may be needed to overcome these challenges and achieve more reliable somatic variant detection.

Single-cell sequencing is a cutting-edge technology that enables sequencing DNA at individual cell level. Compared to the bulk sequencing method, which sequences DNA from a bulk population of cells, single-cell sequencing can provide a deeper understanding of genetic heterogeneity and diversity within complex biological systems. Single-cell DNA sequencing has its origins in pioneering experiments that allowed the detection of gene expression in single cells by microarrays in the early 2000s (Tang et al., 2011). Since then, the single-cell technology has rapidly evolved. In 2009, Navin et al. developed a method to profile single-cell genomes using comparative genomic hybridization (CGH) to identify two types of genomic structural variations and infer pathways of cancer progression in human breast tumors (Navin et al., 2010). They also applied single-cell sequencing to investigate tumor population structure and evolution in two breast cancer cases (Navin et al., 2011). Single-cell sequencing examines the sequence information from individual cells with optimized next-generation sequencing technologies, providing a higher resolution than traditional array-based and bulk sequencing methods (Eberwine et al., 2014).

Single-cell DNA sequencing is limited by its DNA quantity, and it needs the whole genome amplification (WGA) to provide sufficient DNA before sequencing. There are several strategies for the WGA of single-cell DNA sequencing. Multiple Displacement Amplification (MDA), degenerate-oligonucleotide-primed PCR (DOP-PCR) and Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) are three of the most widely used methods for

single-cell DNA sequencing (Hou et al., 2015). MDA is a DNA amplification technique that can rapidly amplify minute amounts of DNA samples from a single cell (Dean et al., 2002; Spits et al., 2006). It is a non-PCR type of DNA amplification that can generate large amounts of DNA from very small starting material. MDA has several advantages, including high amplification efficiency and low input requirements. However, it is prone to amplification bias and errors, which can lead to inaccurate results (Marine et al., 2014). DOP-PCR is a PCR based WGA method, and it uses non-selective amplification to achieve whole genome sequencing (Cheung and Nelson, 1996; Telenius et al., 1992). DOP-PCR has been demonstrated to be a reliable technology to provide low-noise CNV profiles. However, the genomic coverage by DOP-PCR is relatively low, limiting its applications in SNV-related studies (Fu et al., 2019). MALBAC is a newer WGA technology, and it can complete the high-precision whole genome sequencing of a single cell (Lu et al., 2012; Zong et al., 2012). This technique has high sensitivity, and the amplification uniformity is better than other WGA techniques (He et al., 2018). A comprehensive comparison (Hou et al., 2015) of these three WGA methods revealed that the DOP-PCR method had the highest duplication ratio, but the read distribution was even, and it exhibited the best reproducibility and accuracy for CNV detection. In contrast, the MDA method had a significantly higher genome recovery sensitivity (~84%) compared to DOP-PCR (~6%) and MALBAC (~52%) at high sequencing depth. The efficiency of detecting single-nucleotide variations, the false-positive ratio, and the allele drop-out ratio were similar for MALBAC and MDA. Besides these 3 approaches, 10X Genomics developed a droplet-based method, Chromium Single Cell sequencing, with a high uniformity, low coverage, but high throughput, which is suitable for somatic CNV detection (10X Genomics, 2023). However, 10X Genomics discontinued this product after December 31, 2020.

Long-read sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have enabled the sequencing of individual DNA molecules with a length of up to tens of kilobases (Amarasinghe et al., 2020). They have already been applied to study multiple cancer genomics and transcriptomics (Porubsky et al., 2021; Sakamoto et al., 2020; Singh et al., 2019). One of the main advantages of long-read sequencing for somatic mutation detection is that it can overcome the limitations of short-read sequencing in detecting mutations in repetitive regions of the genome. In addition, long-read sequencing can also provide phasing information, which is important for studying allele-specific variants and for reconstructing haplotypes (Logsdon et al., 2020). However, there are still some challenges to overcome (Adewale, 2020). One of the main challenges is the low throughput of current long-read sequencing technologies, which limits the number of cells that can be sequenced in a single experiment. In addition, the high error rates of long-read sequencing can make it difficult to distinguish true somatic mutations from sequencing errors.

Single-cell DNA sequencing has revolutionized the field of genomics. This technology has numerous applications in various fields of biology and medicine, including somatic mosaicism, cancer research, microbiome, immunology, neurobiology, germline transmission, etc (Wang and Navin, 2015). One of the most exciting applications is in the study of somatic mosaicism. Single-cell DNA sequencing has been used to study somatic mosaicism in various tissues and diseases, including the brain (Cai et al., 2014; Wang et al., 2021), skin (Saini et al., 2016), and blood (Zhang et al., 2019). By sequencing individual cells, researchers can discover the distribution and frequency of somatic mutations, which can provide insights into disease mechanisms and potential therapeutic targets.

1.4 Single-cell CNV detection methods and challenges

Single-cell DNA-sequencing technologies provide a valuable opportunity to detect genetic variants in individual cells. It is necessary to develop computational tools to accurately identify such variants. While a number of methods have been developed for identifying SNVs from Single-cell DNA-sequencing data, there is a limited availability of CNV detection methods. Furthermore, several existing methods for CNV detection were originally designed for other types of data, including array-CGH and bulk next-generation sequencing data.

The main challenges for single-cell sequencing to identify CNV include low sequencing depth of coverage, low throughput, and amplification bias. The sequencing coverage determines the size range and the resolution of the CNV that can be detected. A higher coverage can lead to the CNV identification with smaller lower bounds and higher resolution of CNV boundaries. The sequencing throughput is defined as the number of cells that can be sequenced simultaneously and the speed of the turnaround time. Higher throughput enables the scalability of single-cell sequencing, and tends to have lower costs, as they require less labor and can thus accommodate the sequencing of thousands of cells for a single sample. The amplification bias can lead to the non-uniformity of the coverage, which will make false positive CNV calls as it is computationally challenging to distinguish whether read-count fluctuations are due to amplification biases or to true CNVs (Navin, 2014).

The general steps of single-cell CNV detection pipelines usually include genomic binning, GC correction, mappability correction, removal of outlier bins, removal of outlier cells, segmentation and calling the absolute copy numbers (Mallory et al., 2020a). In order to reduce the effects of inconsistent amplification and sequence sampling from single-cell sequencing, it is necessary to partition the genome into fixed- or variable-size bins (Garvin et al., 2015). By doing

so, the resolution of genome segmentation and copy number identification can be determined based on bins rather than the base pairs. GC correction is another essential step due to GC bias, and it would drop the read coverage at the regions with extreme GC contents (Yoon et al., 2009). The mappability is quantified by the percentage of the uniquely mappable positions within a genomic bin. Similar to GC correction, correction of the read counts based on mappability can follow the similar process by modeling the relationship between the read count and mappability of each bin, and subsequently normalizing the read count. Removing outlier bins can help reduce false-positive calls in single-cell sequencing analysis. Bins that are located at centromere or telomere regions, or have zero or extremely high read counts are often identified as outliers and removed from further analysis. Removing outlier cells is also important for further analysis as their coverage is lower or higher than expected. The above five steps are the data wrangling prior to CNV detection, and the following two steps are segmentation and absolute copy number calling. There are three approaches for segmentation: a sliding-window approach, an objective function-based approach, and an HMM-based approach. The sliding-window approach segments the genome by statistical testing and requires post-processing to calculate absolute copy number. The objective function-based approach models the read count by a piecewise constant function to minimize changes and approximate the data, while the HMM-based approach models absolute copy numbers as states and captures transitions between bins to identify breakpoints across multiple cells. After segmenting the genome, the next step is to determine the absolute copy number for each segment. If the DNA ploidy information is known, the absolute copy number can be calculated by scaling the read count of the segment with the genome-wide ploidy and dividing it by the average read count of the whole genome. In the absence of DNA ploidy

information, a copy number multiplier can be estimated to normalize the read count to its nearest integer value (Mallory et al., 2020a).

Currently, there are several computing tools for CNV detection from single-cell data. They can be divided into two categories: single cell-based methods (e.g. HMMcopy, Ginkgo and SCNV), which process a single cell at a time; and multiple cells-based methods (e.g. CHISEL and Alleloscope), which pool the information shared among all the cells. HMMcopy is a Hidden Markov Model (HMM) based approach (Shah et al., 2006). It models the read count distribution of each bin with a Gaussian mixture model and estimates the copy number using the Viterbi algorithm. It was originally designed for array CGH data, but has been widely applied to large-scale single-cell sequencing data (Knouse et al., 2016; Vitak et al., 2017). Ginkgo is a web platform for detecting single-cell CNVs, and it can also be used as a local application (Garvin et al., 2015). It applies a variable bin strategy to segment the genome into bins followed by GC correction. Ginkgo employs Circular Binary Segmentation (CBS) algorithm (Olshen et al., 2004) for genome segmentation and subsequent inference of integer absolute copy number values. Ginkgo has shown a higher accuracy than the other tools in an extensive assessment (Mallory et al., 2020b). SCNV is a bin-free single-cell CNV detector, which provides higher resolution boundaries for CNAs and generalizability to data at different sequencing depths, but it requires at least 20-30 normal cells in the pool of sampled single cells, which limits its usability (Wang et al., 2018). CHISEL is the first allele-specific and haplotype-specific single-cell CNV detection tool that uses a reference-based algorithm to phase blocks in each bin and an Expectation-Maximization algorithm to cluster bins and cells based on B-allele frequency (BAF) and read depth (Zaccaria and Raphael, 2021). CHISEL can identify complex genomic rearrangements and is suitable for analyzing cancer genomes. CHISEL has been shown to detect CNVs on 2 breast

cancer samples and reconstruct a more refined tumor evolutionary scenario that has been validated by SNVs. Alleloscope is another tool that estimates allele-specific copy number and is useful for analyzing single-cell DNA and ATAC sequencing data (Wu et al., 2021). This approach enables more precise identification of copy number states, detection of subclonal copy-neutral loss-of-heterozygosity, and mirrored copy number alteration events. Additionally, the tool allows for the integration of multi-omic analysis of allele-specific copy number and chromatin accessibility for the same cell in single-cell ATAC-seq data.

1.5 Single-cell DNA sequencing simulation methods

The generation of DNA sequences through in silico methods is a highly efficient and inexpensive technology for the evaluation and validation of bioinformatics tools and pipelines (Escalona et al., 2016). They are widely used in a wide range of bioinformatics applications, including genome assembly (Ono et al., 2013), variant calling (Sandmann et al., 2017), transcriptomics (Angly et al., 2012), metagenomics (Jia et al., 2013; Richter et al., 2008), etc.

Compared to the NGS data simulation, it is more necessary to develop a single-cell sequencing simulator as the replication of experiments from the same cells is not possible. The currently available single-cell DNA sequencing simulators, CellCoal, SCSsim, SCSIM, SimSCSnTree and SCSilicon, while able to simulate single-cell data, are not suitable for simulating CNVs in the non-cancer single-cell sequencing data. CellCoal is a computational tool that uses coalescent simulations to model the evolution of cells and infer clonal relationships from single-cell sequencing data (Posada, 2020). It can estimate the time of the most recent common ancestor and the clonal fraction of each cluster of cells. CellCoal was shown to accurately infer the clonal architecture of simulated and real datasets, but it only focuses on SNV

simulation. SCSsim is a MALBAC-based single-cell DNA sequencing simulator that integrates diverse sources of technical noise in scDNA-seq data, including SNVs, indels, and CNVs (Yu et al., 2020). The tool allows users to simulate different sequencing depths and cell coverages, and generates both read and variant files. SCSIM is designed to generate correlated single-cell and bulk DNA reads with SNVs (Giguere et al., 2020). SimSCSnTree generates evolutionary trees of cells that can be tuned via a Beta-splitting model to mimic clonality in cancer (Mallory and Nakhleh, 2022). Additionally, it generates both SNV and CNV, individually or simultaneously, and allows for the generation of ancestral data points to assess the performance of tools that infer CNVs and SNVs in ancestral cells. SimSCSnTree also generates both bulk and single-cell DNA sequencing data, mimicking technological artifacts like non-uniform coverage reflecting various library preparation technologies. SCSilicon is a simulation tool that can generate single-cell DNA reads with minimal manual intervention (Feng and Chen, 2022). It has the ability to automatically create a variety of genomic aberrations, such as SNV, indel, and CNV. Additionally, SCSilicon provides accurate information on CNV segmentation breakpoints and subclone cell labels, allowing researchers to make reliable and valid benchmarking in a controlled way.

While these simulators have been developed to simulate somatic variants based on the single-cell sequencing data, they focus more on the cancer sample and they may not be suitable for other samples such as brain tissue, where clonal evolution is not a major factor. In addition, the mutational landscape of different samples varies, and the performance of bioinformatics tools may depend on the type and frequency of somatic variants present. Therefore, there is a need for developing single-cell sequencing simulators that can generate for non-cancer samples.

1.6 Deep learning application in genomics research

Machine learning techniques are playing an increasingly important role in biomedical research, enabling researchers to analyze complex datasets and make new discoveries in areas such as drug discovery (Vamathevan et al., 2019), protein structural prediction (AlQuraishi, 2021), genomic/genetic data analysis (Libbrecht and Noble, 2015), etc. Machine learning algorithms are used to analyze large amounts of chemical compounds, predict therapeutic effects of compounds, analyze protein structures and identify potential drug targets. Through these approaches, researchers can identify promising drug candidates more efficiently than traditional methods. Due to the increasing availability of large-scale genomic datasets, it is possible to apply machine learning methods to a broad range of areas within genetics and genomics. Machine learning algorithms are useful in annotating various genomic sequence elements, including transcription start sites (TSSs), splice sites, promoters, and enhancers (Basith et al., 2021; Georgakilas et al., 2020; Oubounyt et al., 2019). Machine learning models can also be trained to recognize patterns in DNA sequences as well as input data generated by other genomic assays, such as RNA-seq data for gene expression, ATAC-seq for chromatin accessibility, and ChIP-seq data for histone modification or transcription factor binding (Libbrecht and Noble, 2015).

There are two primary types of machine learning methods: supervised learning and unsupervised learning. Supervised learning algorithms involve learning the relationship between a set of input variables and a designated dependent variable or labels from training instances. These algorithms can subsequently be used to predict the outcomes of new instances. Common supervised learning techniques include regression and classification. On the other hand, unsupervised learning algorithms infer patterns from data without a dependent variable or known

labels. Clustering is one of the popular unsupervised learning methods used to find patterns in high dimensionality data such as omics data.

Deep learning (artificial neural network) is a subtype of machine learning that has been inspired by the connectivity and behavior of neurons in the brain and was originally designed to learn about brain function (Greener et al., 2022). A key property of neural networks is that they are universal function approximators, which means that they can approximate any mathematical function with high accuracy. Deep learning models can be classified into various types, including fully connected networks, convolutional neural networks (CNNs), recurrent neural networks (RNNs), autoencoders, transformers, graph neural networks (GNNs), etc. The most basic form of neural network is a fully connected network, consisting of layers of artificial neurons connected in a dense fashion. These networks are trained by adjusting the weights between the neurons. Convolutional neural networks (CNNs) are ideal for image-like data that contain local structures. CNNs are composed of one or more convolutional layers, in which the output is the result of applying a small, one-layer fully connected neural network, called a 'filter' or 'kernel', to local groups of features in the input. Recurrent neural networks (RNNs) are best suited to sequential data, such as DNA or protein sequences. RNNs are a block of neural network layers that take each sequence entry or time step as input and produce an output that is dependent or correlated with the other entries. Autoencoders, transformers, and graph neural networks are other types of deep learning models that have shown great promise in specific applications.

Neural networks are complex and require specific concerns during training. The training process usually starts with training on a small training dataset to reveal programming errors, where the training loss function should quickly go to zero, indicating no errors. Once the network passes this test, training on the whole training set can proceed, where hyperparameters

such as the learning rate are tuned, and overfitting is prevented by early stopping, regularization of the model, or dropout techniques. Popular software packages used to train neural networks include PyTorch (Paszke et al., 2019) and Tensorflow (Abadi et al., 2016), which require a graphics processing unit or tensor processing unit with sufficient memory to train larger models on large datasets. However, running an already trained model is usually faster and feasible on a standard central processing unit. For those without access to a graphics processing unit, cloud computing solutions exist, and Google Colab allows Python-based deep learning code to be tested on graphics processing units or tensor processing units for small tasks free of charge, making it an excellent way to get started with Python-based deep learning.

Deep learning has been applied in many fields, largely driven by the massive increases in both computational power and big data. It can be both supervised and unsupervised and has revolutionized fields such as image recognition, natural language processing, etc. The promise of deep learning also extends to applications in genetics, especially for the genetic variation identification. Neural network-based methods for identifying SNVs and indels have been developed, such as DeepVariant and Clairvoyante. DeepVariant was developed by Google, and it utilizes a deep neural network to identify SNVs and indels from high-throughput sequencing data (Poplin et al., 2018). DeepVariant visualizes sequence reads in the forms of images. These images are then used to train a CNN model. The tool outperforms other methods, as it has been shown to have high accuracy and sensitivity. Clairvoyante also employs a CNN model to improve the identification of SNVs and indels in single molecule sequencing data (Luo et al., 2019). In addition to identifying small variations, deep learning models have also demonstrated the ability to detect larger structural variations and copy number variations. DeepSV is a deep learning-based tool that can accurately call genomic deletions from high-throughput sequencing

data (Cai et al., 2019). Similar to DeepVariant, it also visualizes mapped sequence reads as images and uses a CNN model to learn features from aligned reads and predict the presence or absence of deletions at each genomic position. DudeML is a deep learning approach that detects CNVs from low-coverage NGS data (Hill and Unckless, 2019). It uses relative coverage changes across genomic windows to classify the window copy number using different machine learning classifiers. DeepCNV is another deep learning-based tool that can authenticate CNVs from genomic data (Glessner et al., 2021). It uses a novel blended deep neural network structure that can exploit both image plots and summary statistics output from PennCNV, a hidden Markov Model based CNV caller (Wang et al., 2007). DeepCNV can replace human experts and reduce false positives in CNV detection. SVision is a CNN-based tool that can detect and characterize simple and complex structural variants from long-read sequencing data (Lin et al., 2022). It uses a novel image representation and a multi-object recognition framework to resolve CSVs with various structures.

However, there is still a lack of deep learning-based tools for calling CNVs from single-cell sequencing data. This is a challenging task due to the high noise and sparsity of single-cell data. Developing more accurate and robust deep learning methods for single-cell CNV detection is an important direction for future research.

1.7 Overview of dissertation research

In my dissertation, I present the development of computational methods for identifying somatic mosaic CNVs with single-cell sequencing. Mosaic CNVs are genomic alterations that occur in a subset of cells within an individual and can have significant implications for human health and disease. Single-cell sequencing is a powerful technology that can reveal the

heterogeneity and diversity of cell populations at high resolution. However, calling CNVs from single-cell sequencing data poses many challenges due to technical noise, data sparsity and complexity.

The main aim of this dissertation is to develop novel and robust methods for detecting and characterizing mosaic CNVs from single-cell sequencing data, as well as to demonstrate our findings regarding the somatic CNVs in the human brain sample. I outline these approaches in the following chapters:

- Chapter 2: My colleagues and I demonstrate our methods and results for identifying somatic CNVs in a human brain sample by single-cell sequencing data. We use allelic ratio to verify coverage-based single-cell CNV calls.
- Chapter 3: I develop a single-cell CNV simulator that can generate realistic synthetic datasets with various levels of mosaicism, coverage and noise. I use this simulator to assess the performance of different single-cell CNV calling tools and also generate the golden standard training data for single-cell CNV calling tool development.
- Chapter 4: I build a recurrent neural network model to call CNVs from single-cell sequencing data. I train and test the model on simulated and real datasets and show that it can achieve high performance compared to other tools.
- Chapter 5: I conclude the dissertation by summarizing the main findings, contributions and limitations of this work. I also discuss some future directions for improving and extending the methods developed in this dissertation.

This dissertation provides new insights into the detection and characterization of mosaic CNVs with single-cell sequencing, which can facilitate further studies on their biological functions and implications for human health.

In addition to the projects described above, another original published piece of work during my PhD study has led to one publication in which I was a first author:

Sun, C., Li, H., Mills, R.E. and Guan, Y., 2019. Prognostic model for multiple myeloma progression integrating gene expression and clinical features. Gigascience, 8(12), p.giz153.

Bibliography

10X Genomics, 2023. Single Cell CNV [WWW Document]. 10x Genomics. URL https://www.10xgenomics.com/products/single-cell-cnv (accessed 2.19.23).

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: a system for large-scale machine learning, in: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16. USENIX Association, USA, pp. 265–283.

Adewale, B.A., 2020. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? Afr. J. Lab. Med. 9, 5. https://doi.org/10.4102/ajlm.v9i1.1340

AlQuraishi, M., 2021. Machine learning in protein structure prediction. Curr. Opin. Chem. Biol., Mechanistic Biology * Machine Learning in Chemical Biology 65, 1–8. https://doi.org/10.1016/j.cbpa.2021.04.005

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 30. https://doi.org/10.1186/s13059-020-1935-5

Angly, F.E., Willner, D., Rohwer, F., Hugenholtz, P., Tyson, G.W., 2012. Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. 40, e94. https://doi.org/10.1093/nar/gks251

Basith, S., Hasan, M.M., Lee, G., Wei, L., Manavalan, B., 2021. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. Brief. Bioinform. 22, bbab252. https://doi.org/10.1093/bib/bbab252

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

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., Kato, M., Kasai, K., Kishimoto, T., Nawa, H., Okano, H., Yoshikawa, T., Kato, T., Iwamoto, K., 2014. Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. Neuron 81, 306–313. https://doi.org/10.1016/j.neuron.2013.10.053

Cagan, A., Baez-Ortega, A., Brzozowska, N., Abascal, F., Coorens, T.H.H., Sanders, M.A., Lawson, A.R.J., Harvey, L.M.R., Bhosle, S., Jones, D., Alcantara, R.E., Butler, T.M., Hooks, Y., Roberts, K., Anderson, E., Lunn, S., Flach, E., Spiro, S., Januszczak, I., Wrigglesworth, E., Jenkins, H., Dallas, T., Masters, N., Perkins, M.W., Deaville, R., Druce, M., Bogeska, R., Milsom, M.D., Neumann, B., Gorman, F., Constantino-Casas, F., Peachey, L., Bochynska, D., Smith, E.S.J., Gerstung, M., Campbell, P.J., Murchison, E.P., Stratton, M.R., Martincorena, I., 2022. Somatic mutation rates scale with lifespan across mammals. Nature 604, 517–524. https://doi.org/10.1038/s41586-022-04618-z

Cai, L., Wu, Y., Gao, J., 2019. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC Bioinformatics 20, 665. https://doi.org/10.1186/s12859-019-3299-y

Cai, X., Evrony, G.D., Lehmann, H.S., Elhosary, P.C., Mehta, B.K., Poduri, A., Walsh, C.A., 2014. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep. 8, 1280–1289. https://doi.org/10.1016/j.celrep.2014.07.043

Cheung, V.G., Nelson, S.F., 1996. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. Proc. Natl. Acad. Sci. 93, 14676–14679. https://doi.org/10.1073/pnas.93.25.14676

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., McConnell, M.J., 2019. Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell Rep. 26, 825-835.e7. https://doi.org/10.1016/j.celrep.2018.12.107

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219. https://doi.org/10.1038/nbt.2514

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., Gage, F.H., 2009. L1 retrotransposition in human neural progenitor cells. Nature 460, 1127–1131. https://doi.org/10.1038/nature08248

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore, S.F., Egholm, M., Lasken, R.S., 2002. Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. 99, 5261–5266. https://doi.org/10.1073/pnas.082089499

Definition of somatic mutation - NCI Dictionary of Cancer Terms - NCI [WWW Document], 2011. URL https://www.cancer.gov/publications/dictionaries/cancer-terms/def/somatic-mutation (accessed 2.12.23).

D'Gama, A.M., 2021. Somatic Mosaicism and Autism Spectrum Disorder. Genes 12, 1699. https://doi.org/10.3390/genes12111699

D'Gama, A.M., Walsh, C.A., 2018. Somatic mosaicism and neurodevelopmental disease. Nat. Neurosci. 21, 1504–1514. https://doi.org/10.1038/s41593-018-0257-3

Ding, J., McConechy, M.K., Horlings, H.M., Ha, G., Chun Chan, F., Funnell, T., Mullaly, S.C., Reimand, J., Bashashati, A., Bader, G.D., Huntsman, D., Aparicio, S., Condon, A., Shah, S.P., 2015. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. Nat. Commun. 6, 8554. https://doi.org/10.1038/ncomms9554

Dou, Y., Yang, X., Li, Z., Wang, S., Zhang, Z., Ye, A.Y., Yan, L., Yang, C., Wu, Q., Li, J., Zhao, B., Huang, A.Y., Wei, L., 2017. Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. Hum. Mutat. 38, 1002–1013. https://doi.org/10.1002/humu.23255

Eberwine, J., Sul, J.-Y., Bartfai, T., Kim, J., 2014. The promise of single-cell sequencing. Nat. Methods 11, 25–27. https://doi.org/10.1038/nmeth.2769

Erwin, J.A., Paquola, A.C.M., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I.A., Butcher, C.R., Herdy, J.R., Sarkar, A., Lasken, R.S., Muotri, A.R., Gage, F.H., 2016. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat. Neurosci. 19, 1583–1591. https://doi.org/10.1038/nn.4388

Escalona, M., Rocha, S., Posada, D., 2016. A comparison of tools for the simulation of genomic next-generation sequencing data. Nat. Rev. Genet. 17, 459–469. https://doi.org/10.1038/nrg.2016.57

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., Park, P.J., Walsh, C.A., 2012. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. Cell 151, 483–496. https://doi.org/10.1016/j.cell.2012.09.035

Feng, X., Chen, L., 2022. SCSilicon: a tool for synthetic single-cell DNA sequencing data generation. BMC Genomics 23, 359. https://doi.org/10.1186/s12864-022-08566-w

Fernández, L.C., Torres, M., Real, F.X., 2016. Somatic mosaicism: on the road to cancer. Nat. Rev. Cancer 16, 43–55. https://doi.org/10.1038/nrc.2015.1

Freed, D., Stevens, E.L., Pevsner, J., 2014. Somatic Mosaicism in the Human Genome. Genes 5, 1064–1094. https://doi.org/10.3390/genes5041064

Fu, Y., Zhang, F., Zhang, X., Yin, J., Du, M., Jiang, M., Liu, L., Li, J., Huang, Y., Wang, J., 2019. High-throughput single-cell whole-genome amplification through centrifugal emulsification and eMDA. Commun. Biol. 2, 1–10. https://doi.org/10.1038/s42003-019-0401-y

Gao, T., Soldatov, R., Sarkar, H., Kurkiewicz, A., Biederstedt, E., Loh, P.-R., Kharchenko, P.V., 2022. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. Nat. Biotechnol. 1–10. https://doi.org/10.1038/s41587-022-01468-y

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M., Schatz, M.C., 2015. Interactive analysis and assessment of single-cell copy-number variations. Nat. Methods 12, 1058–1060. https://doi.org/10.1038/nmeth.3578

Georgakilas, G.K., Perdikopanis, N., Hatzigeorgiou, A., 2020. Solving the transcription start site identification problem with ADAPT-CAGE: a Machine Learning algorithm for the analysis of CAGE data. Sci. Rep. 10, 877. https://doi.org/10.1038/s41598-020-57811-3

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N.Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C.R., Nohadani, M., Eklund, A.C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P.A., Swanton, C., 2012. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. N. Engl. J. Med. 366, 883–892. https://doi.org/10.1056/NEJMoa1113205

Giguere, C., Dubey, H.V., Sarsani, V.K., Saddiki, H., He, S., Flaherty, P., 2020. SCSIM: Jointly simulating correlated single-cell and bulk next-generation DNA sequencing data. BMC Bioinformatics 21, 215. https://doi.org/10.1186/s12859-020-03550-1

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H.G., de Vries, B.B.A., Kleefstra, T., Brunner, H.G., Vissers, L.E.L.M., Veltman, J.A., 2014. Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344–347. https://doi.org/10.1038/nature13394

Glessner, J.T., Hou, X., Zhong, C., Zhang, J., Khan, M., Brand, F., Krawitz, P., Sleiman, P.M.A., Hakonarson, H., Wei, Z., 2021. DeepCNV: a deep learning approach for authenticating copy number variations. Brief. Bioinform. 22, bbaa381. https://doi.org/10.1093/bib/bbaa381

Greaves, M., Maley, C.C., 2012. Clonal evolution in cancer. Nature 481, 306–313. https://doi.org/10.1038/nature10762

Greener, J.G., Kandathil, S.M., Moffat, L., Jones, D.T., 2022. A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 23, 40–55. https://doi.org/10.1038/s41580-021-00407-0

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E.E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D.P., Louis, D.N., Goldstraw, P., Nicholson, A.G., Brasseur, F., Looijenga, L., Weber, B.L., Chiew, Y.-E., deFazio, A., Greaves, M.F., Green, A.R., Campbell, P., Birney, E., Easton, D.F., Chenevix-Trench, G., Tan, M.-H., Khoo, S.K., Teh, B.T., Yuen, S.T., Leung, S.Y., Wooster, R., Futreal, P.A., Stratton, M.R., 2007. Patterns of somatic mutation in human cancer genomes. Nature 446, 153–158. https://doi.org/10.1038/nature05610

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M.C., Papaemmanuil, E., Brewer, D.S., Kallio, H.M.L., Högnäs, G., Annala, M., Kivinummi, K., Goody, V., Latimer, C., O'Meara, S., Dawson, K.J., Isaacs, W., Emmert-Buck, M.R., Nykter, M., Foster, C., Kote-Jarai, Z., Easton, D., Whitaker, H.C., Neal, D.E., Cooper, C.S., Eeles, R.A., Visakorpi, T., Campbell, P.J., McDermott, U., Wedge, D.C., Bova, G.S., 2015. The evolutionary history of lethal metastatic prostate cancer. Nature 520, 353–357. https://doi.org/10.1038/nature14347

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hao, Y., Xuei, X., Li, L., Nakshatri, H., Edenberg, H.J., Liu, Y., 2017. RareVar: A Framework for Detecting Low-Frequency Single-Nucleotide Variants. J. Comput. Biol. 24, 637–646. https://doi.org/10.1089/cmb.2017.0057

Harbst, K., Lauss, M., Cirenajwis, H., Isaksson, K., Rosengren, F., Törngren, T., Kvist, A., Johansson, M.C., Vallon-Christersson, J., Baldetorp, B., Borg, Å., Olsson, H., Ingvar, C., Carneiro, A., Jönsson, G., 2016. Multiregion Whole-Exome Sequencing Uncovers the Genetic Evolution and Mutational Heterogeneity of Early-Stage Metastatic Melanoma. Cancer Res. 76, 4765–4774. https://doi.org/10.1158/0008-5472.CAN-15-3476

He, F., Zhou, W., Cai, R., Yan, T., Xu, X., 2018. Systematic assessment of the performance of whole-genome amplification for SNP/CNV detection and β-thalassemia genotyping. J. Hum. Genet. 63, 407–416. https://doi.org/10.1038/s10038-018-0411-5

Hill, T., Unckless, R.L., 2019. A Deep Learning Approach for Detecting Copy Number Variation in Next-Generation Sequencing Data. G3 GenesGenomesGenetics 9, 3575–3582. https://doi.org/10.1534/g3.119.400596

Hou, Y., Wu, K., Shi, X., Li, F., Song, L., Wu, H., Dean, M., Li, G., Tsang, S., Jiang, R., Zhang, Xiaolong, Li, B., Liu, G., Bedekar, N., Lu, N., Xie, G., Liang, H., Chang, L., Wang, T., Chen, J., Li, Y., Zhang, Xiuqing, Yang, H., Xu, X., Wang, L., Wang, J., 2015. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. GigaScience 4, s13742-015-0068–3. https://doi.org/10.1186/s13742-015-0068-3

Huang, A.Y., Lee, E.A., 2022. Identification of Somatic Mutations From Bulk and Single-Cell Sequencing Data. Front. Aging 2.

Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B.K., Veeriah, S., Shafi, S., Johnson, D.H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D.A., Turajlic, S., Xu, H., Lee, S.-M., Forster, M.D., Ahmad, T., Hiley, C.T., Abbosh, C., Falzon, M., Borg, E., Marafioti, T., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S.M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Shah, R., Joseph, L., Quinn, A.M., Crosbie, P.A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D.A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J.F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Dentro, S., Taniere, P., O'Sullivan, B., Lowe, H.L., Hartley, J.A., Iles, N., Bell, H., Ngai, Y., Shaw, J.A., Herrero, J., Szallasi, Z., Schwarz, R.F., Stewart, A., Quezada, S.A., Le Quesne, J., Van Loo, P., Dive, C., Hackshaw, A., Swanton, C., 2017. Tracking the Evolution of Non–Small-Cell Lung Cancer. N. Engl. J. Med. 376, 2109–2121. https://doi.org/10.1056/NEJMoa1616288

Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L., Wei, C., 2013. NeSSM: A Next-Generation Sequencing Simulator for Metagenomics. PLOS ONE 8, e75448. https://doi.org/10.1371/journal.pone.0075448

Jia, P., Zhao, Z., 2017. Impacts of somatic mutations on gene expression: an association perspective. Brief. Bioinform. 18, 413–425. https://doi.org/10.1093/bib/bbw037

Kandel, E., Koester, J., Mack, S., Siegelbaum, S., 2021. Principles of Neural Science, Sixth Edition. ed. McGraw Hill.

Kazazian, H.H., Moran, J.V., 2017. Mobile DNA in Health and Disease. N. Engl. J. Med. 377, 361–370. https://doi.org/10.1056/NEJMra1510092

Kazazian, H.H., Moran, J.V., 1998. The impact of L1 retrotransposons on the human genome. Nat. Genet. 19, 19–24. https://doi.org/10.1038/ng0598-19

Kim, J., Kim, D., Lim, J.S., Maeng, J.H., Son, H., Kang, H.-C., Nam, H., Lee, J.H., Kim, S., 2019. The use of technical replication for detection of low-level somatic mutations in next-generation sequencing. Nat. Commun. 10, 1047. https://doi.org/10.1038/s41467-019-09026-y

Knouse, K.A., Wu, J., Amon, A., 2016. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome Res. 26, 376–384. https://doi.org/10.1101/gr.198937.115

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576. https://doi.org/10.1101/gr.129684.111

Larson, P.A., Moldovan, J.B., Jasti, N., Kidd, J.M., Beck, C.R., Moran, J.V., 2018. Spliced integrated retrotransposed element (SpIRE) formation in the human genome. PLOS Biol. 16, e2003067. https://doi.org/10.1371/journal.pbio.2003067

Law, S.-M., Akizuki, S., Morinobu, A., Ohmura, K., 2022. A case of refractory systemic lupus erythematosus with monocytosis exhibiting somatic KRAS mutation. Inflamm. Regen. 42, 10. https://doi.org/10.1186/s41232-022-00195-w

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332. https://doi.org/10.1038/nrg3920

Lin, J., Wang, S., Audano, P.A., Meng, D., Flores, J.I., Kosters, W., Yang, X., Jia, P., Marschall, T., Beck, C.R., Ye, K., 2022. SVision: a deep learning approach to resolve complex structural variants. Nat. Methods 19, 1230–1233. https://doi.org/10.1038/s41592-022-01609-w

Logsdon, G.A., Vollger, M.R., Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nat. Rev. Genet. 21, 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A.R., Zhu, P., Hu, X., Xu, L., Yan, L., Bai, F., Qiao, J., Tang, F., Li, R., Xie, X.S., 2012. Probing Meiotic Recombination and Aneuploidy of Single Sperm Cells by Whole-Genome Sequencing. Science 338, 1627–1630. https://doi.org/10.1126/science.1229112

Luo, R., Sedlazeck, F.J., Lam, T.-W., Schatz, M.C., 2019. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat. Commun. 10, 998. https://doi.org/10.1038/s41467-019-09025-z

Luquette, L.J., Miller, M.B., Zhou, Z., Bohrson, C.L., Zhao, Y., Jin, H., Gulhan, D., Ganz, J., Bizzotto, S., Kirkham, S., Hochepied, T., Libert, C., Galor, A., Kim, J., Lodato, M.A., Garaycoechea, J.I., Gawad, C., West, J., Walsh, C.A., Park, P.J., 2022. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat. Genet. 54, 1564–1571. https://doi.org/10.1038/s41588-022-01180-2

Ma, X., Shao, Y., Tian, L., Flasch, D.A., Mulder, H.L., Edmonson, M.N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L.L., Levy, S., Easton, J., Zhang, J., 2019. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 20, 50. https://doi.org/10.1186/s13059-019-1659-6

Macaulay, I.C., Voet, T., 2014. Single Cell Genomics: Advances and Future Perspectives. PLOS Genet. 10, e1004126. https://doi.org/10.1371/journal.pgen.1004126

Malinverno, M., Maderna, C., Abu Taha, A., Corada, M., Orsenigo, F., Valentino, M., Pisati, F., Fusco, C., Graziano, P., Giannotta, M., Yu, Q.C., Zeng, Y.A., Lampugnani, M.G., Magnusson, P.U., Dejana, E., 2019. Endothelial cell clonal expansion in the development of cerebral cavernous malformations. Nat. Commun. 10, 2761. https://doi.org/10.1038/s41467-019-10707-x

Mallory, X.F., Edrisi, M., Navin, N., Nakhleh, L., 2020a. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol. 21, 208. https://doi.org/10.1186/s13059-020-02119-8

Mallory, X.F., Edrisi, M., Navin, N., Nakhleh, L., 2020b. Assessing the performance of methods for copy number aberration detection from single-cell DNA sequencing data. PLOS Comput. Biol. 16, e1008012. https://doi.org/10.1371/journal.pcbi.1008012

Mallory, X.F., Nakhleh, L., 2022. SimSCSnTree: a simulator of single-cell DNA sequencing data. Bioinformatics 38, 2912–2914. https://doi.org/10.1093/bioinformatics/btac169

Manders, F., van Boxtel, R., Middelkamp, S., 2021. The Dynamics of Somatic Mutagenesis During Life in Humans. Front. Aging 2.

Marine, R., McCarren, C., Vorrasane, V., Nasko, D., Crowgey, E., Polson, S.W., Wommack, K.E., 2014. Caught in the middle with multiple displacement amplification: the myth of

pooling for avoiding multiple displacement amplification bias in a metagenome. Microbiome 2, 3. https://doi.org/10.1186/2049-2618-2-3
Martincorena, I., Campbell, P.J., 2015. Somatic mutation in cancer and normal cells. Science 349, 1483–1489. https://doi.org/10.1126/science.aab4082
Mathelier, A., Lefebvre, C., Zhang, A.W., Arenillas, D.J., Ding, J., Wasserman, W.W., Shah, S.P., 2015. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. Genome Biol. 16, 84. https://doi.org/10.1186/s13059-015-0648-7
Maury, E.A., Walsh, C.A., 2021. Somatic copy number variants in neuropsychiatric disorders. Curr. Opin. Genet. Dev., Molecular and genetic basis of disease 68, 9–17. https://doi.org/10.1016/j.gde.2020.12.013
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., Gage, F.H., 2013. Mosaic Copy Number Variation in Human Neurons. Science 342, 632–637. https://doi.org/10.1126/science.1243472
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., Ganz, J., Jaffe, A.E., Kwan, K.Y., Kwon, M., Lodato, M.A., Mills, R.E., Paquola, A.C.M., Rodin, R.E., Rosenbluh, C., Sestan, N., Sherman, M.A., Shin, J.H., Song, S., Straub, R.E., Thorpe, J., Weinberger, D.R., Urban, A.E., Zhou, B., Gage, F.H., Lehner, T., Senthil, G., Walsh, C.A., Chess, A., Courchesne, E., Gleeson, J.G., Kidd, J.M., Park, P.J., Pevsner, J., Vaccarino, F.M., 2017. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science 356, eaal1641. https://doi.org/10.1126/science.aal1641
Merlo, L.M.F., Pepper, J.W., Reid, B.J., Maley, C.C., 2006. Cancer as an evolutionary and ecological process. Nat. Rev. Cancer 6, 924–935. https://doi.org/10.1038/nrc2013
Miller, M.B., Huang, A.Y., Kim, J., Zhou, Z., Kirkham, S.L., Maury, E.A., Ziegenfuss, J.S., Reed, H.C., Neil, J.E., Rento, L., Ryu, S.C., Ma, C.C., Luquette, L.J., Ames, H.M., Oakley, D.H., Frosch, M.P., Hyman, B.T., Lodato, M.A., Lee, E.A., Walsh, C.A., 2022. Somatic genomic changes in single Alzheimer's disease neurons. Nature 604, 714–722. https://doi.org/10.1038/s41586-022-04640-1
Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., Kazazian, H.H., 1996. High Frequency Retrotransposition in Cultured Mammalian Cells. Cell 87, 917–927. https://doi.org/10.1016/S0092-8674(00)81998-4
Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J.V., Gage, F.H., 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903–910. https://doi.org/10.1038/nature03663
Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W.R., Hicks, J., Wigler, M., 2011. Tumour evolution inferred by single-cell sequencing. Nature 472, 90–94. https://doi.org/10.1038/nature09807
Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Månér, S., Zetterberg, A., Hicks, J., Wigler, M., 2010. Inferring tumor progression from genomic heterogeneity. Genome Res. 20, 68–80. https://doi.org/10.1101/gr.099622.109
Navin, N.E., 2014. Cancer genomics: one cell at a time. Genome Biol. 15, 452. https://doi.org/10.1186/s13059-014-0452-9

Nenclares, P., Harrington, K.J., 2020. The biology of cancer. Medicine (Baltimore) 48, 67–72. https://doi.org/10.1016/j.mpmed.2019.11.001
Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M., 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557–572. https://doi.org/10.1093/biostatistics/kxh008
Ono, Y., Asai, K., Hamada, M., 2013. PBSIM: PacBio reads simulator—toward accurate genome assembly. Bioinformatics 29, 119–121. https://doi.org/10.1093/bioinformatics/bts649
Oota, S., 2020. Somatic mutations – Evolution within the individual. Methods, RNA-Seq: Methods and Applications 176, 91–98. https://doi.org/10.1016/j.ymeth.2019.11.002
Ostertag, E.M., Kazazian, H.H., 2001. Twin Priming: A Proposed Mechanism for the Creation of Inversions in L1 Retrotransposition. Genome Res. 11, 2059–2065. https://doi.org/10.1101/gr.205701
Ostertag, E.M., Luning Prak, E.T., DeBerardinis, R.J., Moran, J.V., Kazazian Jr, H.H., 2000. Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res. 28, 1418–1423. https://doi.org/10.1093/nar/28.6.1418
Oubounyt, M., Louadi, Z., Tayara, H., Chong, K.T., 2019. DeePromoter: Robust Promoter Predictor Using Deep Learning. Front. Genet. 10.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.
Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., Gross, S.S., Dorfman, L., McLean, C.Y., DePristo, M.A., 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987. https://doi.org/10.1038/nbt.4235
Porubsky, D., Ebert, P., Audano, P.A., Vollger, M.R., Harvey, W.T., Marijon, P., Ebler, J., Munson, K.M., Sorensen, M., Sulovari, A., Haukness, M., Ghareghani, M., Lansdorp, P.M., Paten, B., Devine, S.E., Sanders, A.D., Lee, C., Chaisson, M.J.P., Korbel, J.O., Eichler, E.E., Marschall, T., 2021. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat. Biotechnol. 39, 302–308. https://doi.org/10.1038/s41587-020-0719-5
Posada, D., 2020. CellCoal: Coalescent Simulation of Single-Cell Sequencing Samples. Mol. Biol. Evol. 37, 1535–1542. https://doi.org/10.1093/molbev/msaa025
Proukakis, C., 2020. Somatic mutations in neurodegeneration: An update. Neurobiol. Dis. 144, 105021. https://doi.org/10.1016/j.nbd.2020.105021
Richter, D.C., Ott, F., Auch, A.F., Schmid, R., Huson, D.H., 2008. MetaSim—A Sequencing Simulator for Genomics and Metagenomics. PLOS ONE 3, e3373. https://doi.org/10.1371/journal.pone.0003373
Roerink, S.F., Sasaki, N., Lee-Six, H., Young, M.D., Alexandrov, L.B., Behjati, S., Mitchell, T.J., Grossmann, S., Lightfoot, H., Egan, D.A., Pronk, A., Smakman, N., van Gorp, J., Anderson, E., Gamble, S.J., Alder, C., van de Wetering, M., Campbell, P.J., Stratton, M.R., Clevers, H., 2018. Intra-tumour diversification in colorectal cancer at the single-cell level. Nature 556, 457–462. https://doi.org/10.1038/s41586-018-0024-3

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., Hirst, M., Turashvili, G., Oloumi, A., Marra, M.A., Aparicio, S., Shah, S.P., 2012. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 28, 907–913. https://doi.org/10.1093/bioinformatics/bts053

Roy, J.C.L., Vitalo, A., Andrew, M.A., Mota-Silva, E., Kovalenko, M., Burch, Z., Nhu, A.M., Cohen, P.E., Grabczyk, E., Wheeler, V.C., Mouro Pinto, R., 2021. Somatic CAG expansion in Huntington's disease is dependent on the MLH3 endonuclease domain, which can be excluded via splice redirection. Nucleic Acids Res. 49, 3907–3918. https://doi.org/10.1093/nar/gkab152

Saini, N., Roberts, S.A., Klimczak, L.J., Chan, K., Grimm, S.A., Dai, S., Fargo, D.C., Boyer, J.C., Kaufmann, W.K., Taylor, J.A., Lee, E., Cortes-Ciriano, I., Park, P.J., Schurman, S.H., Malc, E.P., Mieczkowski, P.A., Gordenin, D.A., 2016. The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. PLOS Genet. 12, e1006385. https://doi.org/10.1371/journal.pgen.1006385

Sakamoto, Y., Sereewattanawoot, S., Suzuki, A., 2020. A new era of long-read sequencing for cancer genomics. J. Hum. Genet. 65, 3–10. https://doi.org/10.1038/s10038-019-0658-5

Sandmann, S., de Graaf, A.O., Karimi, M., van der Reijden, B.A., Hellström-Lindberg, E., Jansen, J.H., Dugas, M., 2017. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci. Rep. 7, 43169. https://doi.org/10.1038/srep43169

Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J., Cheetham, R.K., 2012. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 28, 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Shah, S.P., Xuan, X., DeLeeuw, R.J., Khojasteh, M., Lam, W.L., Ng, R., Murphy, K.P., 2006. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22, e431–e439. https://doi.org/10.1093/bioinformatics/btl238

Singer, T., McConnell, M.J., Marchetto, M.C.N., Coufal, N.G., Gage, F.H., 2010. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci. 33, 345–354. https://doi.org/10.1016/j.tins.2010.04.001

Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J.M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Giang Phan, T., Junankar, S., Jackson, K., Goodnow, C.C., Smith, M.A., Swarbrick, A., 2019. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat. Commun. 10, 3120. https://doi.org/10.1038/s41467-019-11049-4

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663. https://doi.org/10.1016/S0959-437X(99)00031-3

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., Sermon, K., 2006. Whole-genome multiple displacement amplification from single cells. Nat. Protoc. 1, 1965–1970. https://doi.org/10.1038/nprot.2006.326

Stosser, M.B., Lindy, A.S., Butler, E., Retterer, K., Piccirillo-Stosser, C.M., Richard, G., McKnight, D.A., 2018. High frequency of mosaic pathogenic variants in genes causing epilepsy-related neurodevelopmental disorders. Genet. Med. 20, 403–410. https://doi.org/10.1038/gim.2017.114

Sui, Y., Peng, S., 2021. A Mechanism Leading to Changes in Copy Number Variations Affected by Transcriptional Level Might Be Involved in Evolution, Embryonic Development, Senescence, and Oncogenesis Mediated by Retrotransposons. Front. Cell Dev. Biol. 9.

Tang, F., Lao, K., Surani, M.A., 2011. Development and applications of single-cell transcriptome analysis. Nat. Methods 8, S6–S11. https://doi.org/10.1038/nmeth.1557

Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjo¨ld, M., Ponder, B.A.J., Tunnacliffe, A., 1992. Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. Genomics 13, 718–725. https://doi.org/10.1016/0888-7543(92)90147-K

Turan, Z.G., Richter, V., Bochmann, J., Parvizi, P., Yapar, E., Işıldak, U., Waterholter, S.-K., Leclere-Turbant, S., Son, Ç.D., Duyckaerts, C., Yet, İ., Arendt, T., Somel, M., Ueberham, U., 2022. Somatic copy number variant load in neurons of healthy controls and Alzheimer's disease patients. Acta Neuropathol. Commun. 10, 175. https://doi.org/10.1186/s40478-022-01452-2

Valind, A., Jin, Y., Baldetorp, B., Gisselsson, D., 2013. Whole chromosome gain does not in itself confer cancer-like chromosomal instability. Proc. Natl. Acad. Sci. 110, 21119–21123. https://doi.org/10.1073/pnas.1311163110

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S., 2019. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. 18, 463–477. https://doi.org/10.1038/s41573-019-0024-5

Veeriah, S., Taylor, B.S., Meng, S., Fang, F., Yilmaz, E., Vivanco, I., Janakiraman, M., Schultz, N., Hanrahan, A.J., Pao, W., Ladanyi, M., Sander, C., Heguy, A., Holland, E.C., Paty, P.B., Mischel, P.S., Liau, L., Cloughesy, T.F., Mellinghoff, I.K., Solit, D.B., Chan, T.A., 2010. Somatic mutations of the Parkinson's disease–associated gene PARK2 in glioblastoma and other human malignancies. Nat. Genet. 42, 77–82. https://doi.org/10.1038/ng.491

Vijg, J., 2014. Somatic mutations, genome mosaicism, cancer and aging. Curr. Opin. Genet. Dev., Molecular and genetic bases of disease 26, 141–149. https://doi.org/10.1016/j.gde.2014.04.002

Vitak, S.A., Torkenczy, K.A., Rosenkrantz, J.L., Fields, A.J., Christiansen, L., Wong, M.H., Carbone, L., Steemers, F.J., Adey, A., 2017. Sequencing thousands of single-cell genomes with combinatorial indexing. Nat. Methods 14, 302–308. https://doi.org/10.1038/nmeth.4154

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., Bucan, M., 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17, 1665–1674. https://doi.org/10.1101/gr.6861907

Wang, X., Chen, H., Zhang, N.R., 2018. DNA copy number profiling using single-cell sequencing. Brief. Bioinform. 19, 731–736. https://doi.org/10.1093/bib/bbx004

Wang, Y., Bae, T., Thorpe, J., Sherman, M.A., Jones, A.G., Cho, S., Daily, K., Dou, Y., Ganz, J., Galor, A., Lobon, I., Pattni, R., Rosenbluh, C., Tomasi, S., Tomasini, L., Yang, X., Zhou, B., Akbarian, S., Ball, L.L., Bizzotto, S., Emery, S.B., Doan, R., Fasching, L., Jang, Y., Juan, D., Lizano, E., Luquette, L.J., Moldovan, J.B., Narurkar, R., Oetjens, M.T., Rodin, R.E., Sekar, S., Shin, J.H., Soriano, E., Straub, R.E., Zhou, W., Chess, A., Gleeson, J.G., Marquès-Bonet, T., Park, P.J., Peters, M.A., Pevsner, J., Walsh, C.A., Weinberger, D.R.,

Vaccarino, F.M., Moran, J.V., Urban, A.E., Kidd, J.M., Mills, R.E., Abyzov, A., Brain Somatic Mosaicism Network, 2021. Comprehensive identification of somatic nucleotide variants in human brain tissue. Genome Biol. 22, 92. https://doi.org/10.1186/s13059-021-02285-3

Wang, Y., Navin, N.E., 2015. Advances and Applications of Single-Cell Sequencing Technologies. Mol. Cell 58, 598–609. https://doi.org/10.1016/j.molcel.2015.05.005

Wu, C.-Y., Lau, B.T., Kim, H.S., Sathe, A., Grimes, S.M., Ji, H.P., Zhang, N.R., 2021. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. Nat. Biotechnol. 39, 1259–1269. https://doi.org/10.1038/s41587-021-00911-w

Yates, L.R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L.B., Larsimont, D., Davies, H., Li, Y., Ju, Y.S., Ramakrishna, M., Haugland, H.K., Lilleng, P.K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L.J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M.R., Sotiriou, C., Richardson, A.L., Lønning, P.E., Wedge, D.C., Campbell, P.J., 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat. Med. 21, 751–759. https://doi.org/10.1038/nm.3886

Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19, 1586–1592. https://doi.org/10.1101/gr.092981.109

Yu, Z., Du, F., Sun, X., Li, A., 2020. SCSsim: an integrated tool for simulating single-cell genome sequencing data. Bioinformatics 36, 1281–1282. https://doi.org/10.1093/bioinformatics/btz713

Zaccaria, S., Raphael, B.J., 2021. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nat. Biotechnol. 39, 207–214. https://doi.org/10.1038/s41587-020-0661-6

Zaidi, S.H., Harrison, T.A., Phipps, A.I., Steinfelder, R., Trinh, Q.M., Qu, C., Banbury, B.L., Georgeson, P., Grasso, C.S., Giannakis, M., Adams, J.B., Alwers, E., Amitay, E.L., Barfield, R.T., Berndt, S.I., Borozan, I., Brenner, H., Brezina, S., Buchanan, D.D., Cao, Y., Chan, A.T., Chang-Claude, J., Connolly, C.M., Drew, D.A., Farris, A.B., Figueiredo, J.C., French, A.J., Fuchs, C.S., Garraway, L.A., Gruber, S., Guinter, M.A., Hamilton, S.R., Harlid, S., Heisler, L.E., Hidaka, A., Hopper, J.L., Huang, W.-Y., Huyghe, J.R., Jenkins, M.A., Krzyzanowski, P.M., Lemire, M., Lin, Y., Luo, X., Mardis, E.R., McPherson, J.D., Miller, J.K., Moreno, V., Mu, X.J., Nishihara, R., Papadopoulos, N., Pasternack, D., Quist, M.J., Rafikova, A., Reid, E.E.G., Shinbrot, E., Shirts, B.H., Stein, L.D., Teney, C.D., Timms, L., Um, C.Y., Van Guelpen, B., Van Tassel, M., Wang, X., Wheeler, D.A., Yung, C.K., Hsu, L., Ogino, S., Gsur, A., Newcomb, P.A., Gallinger, S., Hoffmeister, M., Campbell, P.T., Thibodeau, S.N., Sun, W., Hudson, T.J., Peters, U., 2020. Landscape of somatic single nucleotide variants and indels in colorectal cancer and impact on survival. Nat. Commun. 11, 3644. https://doi.org/10.1038/s41467-020-17386-z

Zhang, F., Gu, W., Hurles, M.E., Lupski, J.R., 2009. Copy Number Variation in Human Health, Disease, and Evolution. Annu. Rev. Genomics Hum. Genet. 10, 451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217

Zhang, L., Dong, X., Lee, M., Maslov, A.Y., Wang, T., Vijg, J., 2019. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes

across the human lifespan. Proc. Natl. Acad. Sci. 116, 9014–9019. https://doi.org/10.1073/pnas.1902510116

Zhao, B., Wu, Q., Ye, A.Y., Guo, J., Zheng, X., Yang, X., Yan, L., Liu, Q.-R., Hyde, T.M., Wei, L., Huang, A.Y., 2019. Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. PLOS Genet. 15, e1008043. https://doi.org/10.1371/journal.pgen.1008043

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

Zong, C., Lu, S., Chapman, A.R., Xie, X.S., 2012. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. Science 338, 1622–1626. https://doi.org/10.1126/science.1229164

Chapter 2 Mapping the Complex Genetic Landscape of Human Neurons

2.1 Introduction

It is inaccurate to view an individual's genome as invariant from organ to organ, or from cell to cell within an organ. For example, somatic mosaicism among lymphocytes has been recognized since the 1970's with the discovery of somatic gene rearrangement at T cell receptor and immunoglobulin loci (Hozumi and Tonegawa, 1976). Recurrent somatic mutations also underlie the pathology of many cancers (Hanahan and Weinberg, 2011). Recent advances in single-cell and bulk DNA sequencing approaches have revealed abundant somatic mosaicism throughout the human body (Abascal et al., 2021; Bizzotto et al., 2021; Coorens et al., 2021; Moore et al., 2021; Mustjoki and Young, 2021; Spencer Chapman et al., 2021). Associated studies have linked environmental mutagens to somatic mutations in the skin, bladder, and other exposed cells (Lawson et al., 2020; Lee-Six et al., 2019; Moore et al., 2021). Rapidly dividing stem cell populations also incur somatic mutations due to DNA replication errors. Clonal expansion of variant genomes can, in turn, shape mosaicism among an individual's somatic cells (Martincorena, 2019). Somatic mutations, accompanied by cell death, set the stage for somatic selection during the lifespan of an individual.

Brain somatic mosaicism is associated with neurodevelopmental disorders, especially epilepsy (Jansen et al., 2015; Lee et al., 2012; Lim et al., 2015; Møller et al., 2016; Muotri et al., 2010; Poduri et al., 2012; Rodin et al., 2021; Shirley et al., 2013). Unlike other organs, cerebral cortical neurons arise in utero and are not replaced during normal human lifespan (Bhardwaj et al., 2006). Neural stem and progenitor cells proliferate rapidly during human cortical

development; these progeny overpopulate the developing cerebral cortex. Somatic selection is one means by which some progeny may thrive as cortical neurons while other progeny perish (Blaschke et al., 1998; McConnell et al., 2009; Rakic and Zecevic, 2000; Wong and Marín, 2019). The genomes of mature cortical neurons contain hundreds of single nucleotide variants (SNVs), some of which mark clonal lineages (Bae et al., 2018; Breuss et al., 2022; Fasching et al., 2021; Wang et al., 2021). LINE-1 mobile elements retrotranspose during neurogenesis and contribute to brain somatic mosaicism in a small subset of neurons (Baillie et al., 2011; Erwin et al., 2016; Evrony et al., 2012; Muotri et al., 2005; Zhu et al., 2021). Although SNVs are numerous and accumulate throughout life, relatively few are predicted to have protein-coding mutations with obvious consequences for affected neurons (Lodato et al., 2018; Miller et al., 2022; Wang et al., 2021). Megabase (Mb)-scale copy number variants (CNVs) - typically sub-chromosomal deletions - also contribute to brain somatic mosaicism (Cai et al., 2014; Knouse et al., 2016; McConnell et al., 2013).

In non-diseased (neurotypical) brains, dozens of genes are impacted in CNV neurons with substantial inter-individual variation in the frequency of CNV neurons among individuals (Chronister et al., 2019). CNV neurons are more prevalent in the frontal cortex of young individuals (n=4 individuals <30 years old; 28.5% CNV neurons, 75/263) than in aged individuals (n=5 individuals >70 years old; 7.3% CNV neurons, 26/354) (Chronister et al., 2019). However, small sample sizes (<100 neurons / individual) have limited the ability of these studies to find patterns of recurrent rearrangement (e.g., CNV hotspots) among neuronal genomes. If present, such recurrent sites of neuronal genome rearrangement may provide insight into the mechanisms and/or consequences of brain somatic mosaicism. Recurrent sites of neuronal genome rearrangement could be influenced by common fragile sites that are

predisposed to genome rearrangements (Glover et al., 2017; Lehman et al., 2017) and may reflect neurodevelopmental somatic selection. Neither mechanism is mutually exclusive.

We reasoned that if recurrent brain CNVs exist, hotspots would be found among neurons in any millimeter-scale cortical biopsy from a single individual. Using a commercial droplet-based whole genome amplification (WGA) method, we generated Illumina sequencing libraries from 2,125 frontal cortical nuclei isolated from a previously characterized neurotypical individual (Chronister et al., 2019; Wang et al., 2021). Read-depth analysis of each cell was coupled with phased germline single nucleotide polymorphisms (SNPs) to develop a single-cell sequencing coverage and allele-based approach (SCOVAL) that restricted read-depth based deletion calls using concordant, phased, loss-of-heterozygosity (LOH) information. In total, 2097 single neuron libraries passed quality controls (QC) and 10.8% (226/2097) contained at least one Mb-scale CNV. An unexpected subpopulation of these CNV neurons (65/226, 25%) have highly aberrant karyotypes wherein multiple chromosomes harbor multiple deletions, including 6 aneusomic neurons. When compared to a random model, CNVs are depleted in gene-dense genomic regions. However, frequent neuronal genome rearrangements are more common in genomic regions that contain genes encoded by more than 100 kilobases (kb) of genomic sequence (herein defined as long genes).

2.2 Methods

2.2.1 Sample and sequencing library preparation

We examined human neurons dissected from the dorsolateral prefrontal cortex (DLPFC) of a neurotypical individual (postmortem, 49-year-old male individual, ID: Br5154) used as the common reference brain in a previous study (Wang et al., 2021). Neuronal nuclei (NeuN+) were

isolated as in (Chronister et al., 2019). We then applied 10X Genomics Chromium Single Cell sequencing that ligated barcodes on the DNA in single cells within a Cell Bead Gel and the barcoded fragments are then pooled for library production, which can profile thousands of cells. We sequenced 2,125 neurons in two batches with mean coverage 0.114X (Figure 2.1A). We further applied 10X Genomics Chromium Linked-Read sequencing to dural fibroblast tissue with very high sequencing coverage (52.7X) from the same individual to identify and phase germline SNPs by isolating and fragmenting long DNA segments into barcoded short reads that could be used to reconstruct underlying haplotypes using Long Ranger v2.2 (https://github.com/10XGenomics/longranger).

2.2.2 Optimization of Ginkgo for single-cell CNV identification

The final CNV call set was generated using a combination of read-depth and phased loss-of-heterozygosity (LOH)-based validation. First, we processed read alignments from 2,125 single-cells using an adapted version of Ginkgo (Garvin et al., 2015) to arrive at our unvalidated call set. The call set was then filtered via empirical P-value selection using information pertaining to loss of a particular haplotype, obtained by aligning sample reads to the (diploid) phased genome for this individual. The resultant calls were then filtered using a Bayesian classification model to arrive at the final CNV call set, which was further classified by CNV type (heterozygous deletions, homozygous deletions, and duplications) because the strength of support is different for these different CNVs, and the ensuing permutation testing (using heterozygous deletions alone) became more regularized. Only CNV calls in autosomes were included in the final CNV call set. We will now describe the generation pipeline, similar to (Chronister et al., 2019), in some detail.

Setting CNV calling cutoffs in Ginkgo via Gaussian Mixture Model

Ginkgo was optimized by resetting default copy-number cutoffs that determine whether a segment detected by circular binary segmentation (CBS) will be called a CNV. To this end, we processed single-cell BAM files from 585 cells obtained from the five control individuals studied in (Chronister et al., 2019) using the CBS implementation DNACopy (https://bioconductor.org/packages/release/bioc/html/DNAcopy.html). Aligned reads from each single cell were separately processed into 5,067 autosomal bins across the hg19 human reference genome delineated by Ginkgo, which were then normalized to obtain an average copy number of two for the cell. These individual bins were then grouped contiguously into segments based on similarity of their read coverage using DNACopy. We then fit a Gaussian Mixture Model (GMM) to the distribution of the median copy number of all segments from all cells using an "undoSD" of three, whereby two putative segments had to be more than three times the standard deviation in "intra-segment" copy number to be actually written as separate segments, and alpha=.01. From this fit, the two-tailed probability for the Gaussian curve centered at CN=1 and the one at CN=2 was calculated to be 1.63 (Figure 2.2B). This became the new copy-number cutoff for Ginkgo to call deletions. As seen in Figure 2.2B, there were not many candidate duplications to yield a proper fit, but the duplication cutoff was set at 2.43.

Filtering to remove outlier bins via Tukey's rule

Next, the raw bin CN data were filtered for the presence of uniform outlier bins across all cells (e.g., due to data-specific genomic regions uniformly subject to overamplification or underamplification, regions of poor mappability in the genome, etc). The median of copy numbers of 2,125 cells for each of the 5,067 autosomal bins was first plotted. Tukey's rule was

then applied to tag all bins whose median copy number exceeded Q3 + 1.5* IQR, or was below Q1-1.5*IQR, where the interquartile range IQR is Q3-Q1 and Q1 and Q3 are the first and third quartiles, respectively, of all the median copy numbers. Three hundred and eight outlier bins were identified in addition to Ginkgo's original list containing 29 (Figure 2.2C). These bins were simply removed from the genome by Ginkgo prior to segment processing while other bins (retaining their genomic coordinates) were merged. For reference, the genomic bin size used for benchmarking Ginkgo was 500 Kb. Thus, in this work, as in (Chronister et al., 2019), we used Ginkgo settings pertaining to an approximate variable bin size of 500 Kb ("variable_500kb_101_bowtie") and only considered large (> 1 Mb) CNVs. Gingko reported a final mean bin size of 569 Kb, with bins ranging in size from 501 to 2812 Kb.

Filtering of irregular cells

For all cells, the mean absolute deviation (MAD) of bin copy numbers was calculated and fit to a Gaussian distribution. The mean (mu) and standard deviation (sigma) were .253 and .111, respectively. CNV calls from 19 cells (MAD > mu + 3* sigma) were removed before processing the data further (Figure 2.2A). The total number of reads for all remaining cells ranged uniformly from 580,809 to 8,983,573. However, one cell contained an inordinate proportion of reads (> 80%) aligned to just one of the chromosomes and was removed. Further, eight cells that were not filtered by the above methods were manually curated from the data set based on unlikely copy-number patterns, leaving a total of 2,097 good neurons (see Figure 2.2D).

2.2.3 Assessing the coverage-based single-cell CNV call set

To differentiate between bona fide CNVs and potential false-positives due to coverage fluctuations, we leveraged the long-range haplotype information obtained from the 10X linked-

read sequences generated from bulk analysis of matched dural fibroblast tissue. We made use of identified heterozygous SNPs (het-SNPs) and initially segmented the genome using phase blocks of heterozygous SNPs as identified by the linked-read data so that each segment would contain SNPs with consistent haplotype labeling. We then binned these segments further into windows of 20-100 SNPs based on empirical observations of SNP and read coverages. For each window in each cell, we then identified reads that overlapped het-SNPs (herein termed "informative reads") and noted the allele present on the read. Notably, the coverage in each single cell resulted in a sparse number of informative reads per SNP window, typically resulting in 5-15 reads with specific allele information. Using the inferred haplotype of each overlapped het-SNP, we counted the number of reads present on each of the two haplotypes and calculated the absolute log2 ratio between the read counts if the total number of reads on each haplotype was larger than three. We used this log2 ratio to filter the CNV call set from the previous stage. First, we calculated the median log2 ratio of the windows within the CNV regions in the cells with those CNVs and the median log2 ratio of the windows within the CNV regions but in the cells without those CNVs as a background null model. From these data, we derived an empirical p-value for the observed log2 ratio in the sample with the CNV. We then collated the p-values for each individual CNV to derive a p-value distribution and selected a set of candidate CNVs with a p-value < 0.05.

Next, we randomly permuted 100 sets of "non-CNVs" size-matched to these candidate calls to build a GMM from the underlying median log2 ratios of each CNV/non-CNV region, with the assumption that the two distributions followed two distinct Gaussian distributions. Using the median absolute log2 ratios of the two datasets as the training data, we estimated the parameters of the Gaussians and predicted the posterior probability that the CNV belonged to the

CNV distribution using a naive Bayesian classifier. Calls with posterior probability > .99 were selected to process further.

As allele imbalance cannot support the homozygous deletions, we implemented a read-depth ratio measurement to add additional support on the calls. We calculated the read-depth ratio for each bin in every cell based on the bulk sequencing from the same tissue (Wang et al., 2021). The read-depth ratio $RDR_{b,i}$ of bin b and cell i can be calculated as

$$RDR_{b,i} = \frac{C_{b,i}R_B}{B_bR_i}$$

Where $C_{b,i}$ is the number of reads in bin b of cell i, B_b is the number of the reads in bin b of bulk sequencing, R_B is the total number of reads of bulk sequencing, and R_i is the total number of reads of cell i. To distinguish between homozygous and heterozygous deletions, we applied a GMM on read-depth ratio to calculate the posterior probability for the homozygous deletions, and set the cutoff as >0.99 for posterior probability. The final call set for heterozygous deletions was obtained by adjudicating the above calls by requiring the CNV region to have an empirical median log2-ratio p-value (as described above) to be less than .01 (thus ensuring that only calls in regions showing the highest relative allelic preference were selected).

2.2.4 Benchmarking CNV detection

We applied CHISEL (Zaccaria and Raphael, 2021) to our single-cell sequencing data with its default parameters (max balanced ploidy=4); however, it reported unrealistic results. Only 8.16% of all 5MB windows were reported as normal diploid regions with haplotype copy number ‘1|1’, with most windows (77.83%) indicating the max balanced ploidy with haplotype copy number ‘2|2’. We adjusted the max balanced ploidy setting to 2, resulting in 98.15% of the

windows now indicated as normal diploid regions. We combined neighboring CNV windows within the same cell to calculate the overlap percentage with our final call set.

2.2.5 Clonal cells and recurrent CNVs

To detect the clonal structure of neurons based on CNVs, we designed a very conservative method to identify clonal events. We first found all the CNVs that shared the same start and end breakpoints, then we marked these loci as CNVR. With the haplotype information, we could identify whether these loci were clonal events or the recurrent events that existed on the different haplotypes. For each bin covered by the CNVR, we took the maximum log2 ratio and minimum log2 ratio of the cells with the CNVR and calculated the delta log2 ratio using maximum minus minimum. Next, we calculated the median delta log2 ratio across the bins for each CNVR and observed two distinct distributions, one representing potential clonal events (low delta log2 ratio; CNVs are on the same haplotype) and the other indicating likely independent events (high delta log2 ratio; CNVs are on the different haplotypes).

2.2.6 Characterizing CNV Neurons

Neuronal distribution of CNVs

The raw distribution of the number of CNVs per neuron is shown as a histogram (Figure 2.3A) on a log scale, along with a null model based on a uniform random distribution of all CNVs in the final call set across all good neurons. Thus, a Poisson curve with mean = (# final CNVs) / (# good cells), scaled up by the total number of good neurons, was superimposed on the

first plot to assess whether the final call set contained more CNV-rich neurons than expected by a uniform distribution.

Hierarchical clustering and complex karyotypes

The 2,097 good neurons were ordered based on the number of total base pairs affected by heterozygous deletions in descending order. A heat map of all cells was generated showing the percentage of base pairs affected by heterozygous deletions in each autosome (see Figure 2.3C), Neurons were sorted and numbered in reverse order of % base pairs affected. Those cells affected more than 5% were termed complex neurons and numbered 1-65 in our call set. All good neurons were clustered using hierarchical clustering using each autosome as an independent dimension and the percentage of base pairs affected as the distance measure. Thus, cells with chromosomes that were similarly affected by heterozygous deletions clustered together (Figure 2.3D). Some cells with possibly multiple recurrent events were identified (Figure 2.3E), and some seemingly clonal cells were analyzed to be technical replicates.

Identifying CNV hotspots and cold spots via permutation testing

The final heterozygous deletion call set was "shuffled" using bedtools (Quinlan and Hall, 2010) to arrive at 10,000 unique synthetic permutations (Figure 2.4A). In each permutation, CNVs in each cell were permuted uniformly at random in the autosomes while prohibiting collision ("noOverlapping" option) and then assembled together. The process was repeated 10,000 times without genomic constraints, as unmappable regions were a priori removed (refer to subsection Optimization of Ginkgo for single-cell CNV identification), and calls "straddling" such regions commonly occurred in the final call set.

Each autosome was divided into contiguous 5Mb regions (remaining smaller tails of chromosomes were not considered). The number of unique hits (defined as simple overlap) of each region with synthetic CNVs from all 10,000 permutations was recorded, resulting in a CNV distribution profile for the synthetic data. For each 5Mb region, a P-value was assessed for the number of CNV hits in real data among the 10,000 hit-values in the region's synthetic CNV profile. For our purposes, we define P-value to be the fraction of simulated instances that were at least as high as the real number of CNV hits to the 5Mb region. Given that CNV hits are discrete-valued, and we are using the same definition of P-value for cold spots and hotspots, we impose a more stringent cutoff for cold spots to account for the inherent liberal treatment of data values on the lower extreme (which may lead to an overabundance of cold spots). Regions with a P-value < .05 (i.e., where hits were among the top 5 % of synthetic hit-values for that region) were termed "hotspots" and those with P-value > .99 were termed "cold spots." Regional significance (defined as 1 - P-value) was plotted against the autosomal genome on the x-axis (Figure 2.5). The distribution of the raw number of CNV hits in 5Mb regions is shown in Figure 2.4B. Cold spots were screened for aberrant genomic blocks that might hamper CNV calling or regions a priori neglected. To this end, cold spot regions were coordinate-merged (via "bedtools merge") and compared to all a priori removed bad bins as well as blacklisted regions (Amemiya et al., 2019) by means of a relative permutation analysis. A merged cold spot that overlapped more with blacklisted regions and bad bins as appropriately compared to 1,000 randomly selected non-cold spot intervals was removed from the list of final cold spots (the cutoff chosen was p > .05) (Figure 2.6A). Each merged cold spot was mapped to 1,000 randomly selected regions other than existing cold spots, and its overlap with bases contained in bad bins and blacklisted regions, respectively, were calculated in each instance in order to assign it a p-value.

For additional relevant detail, some genomic heat maps of copy number of CNV neurons are shown in Figure 2.6C, D along with merged cold spots and bad bins. For rigor, cold spots were analyzed for the presence of deduplicated germline structural variants from 1,000 individuals from FusorSV (Becker et al., 2018), the cold spots had a larger SV coverage (11.4) than the unremarkable regions (7.25), further supporting that CNVs are callable in these regions.

Hotspots and cold spots are shown throughout the genome in a Circos (Krzywinski et al., 2009) plot along with 33 regions of the genome where germline CNVs are associated with neurodevelopmental phenotypes (Birnbaum et al., 2022) to assess any possible correlation between the two (Figure 2.4C). The distribution of the number of genes in 5Mb regions was also plotted for hotspots, cold spots and unremarkable regions as control (Figure 2.4D). Similar distributions were plotted (with assigned p-values) for long genes and different expression levels (Figure 2.7).

In a complementary assessment, the above permutation analysis was repeated for genes instead of 5Mb genomic regions. To profile gene expression, histograms of p-values for genes were shown for different gene expression categories (Figure 2.8) to assess/confirm general prevalence of hotspots and cold spots in each expression category.

Recurrent CNV breakpoint analysis

To assess the impact of different Ginkgo bin sizes on the CNV breakpoint distribution, we used the previously described 10K permuted CNV sets to determine the relationship between the number of breakpoints and Ginkgo bin size. We calculated the mean of the number of breakpoints from all permuted CNVs and compared this to the size of the Ginkgo bin in which they fell. We then normalized the number of breakpoints by the Ginkgo bin size and compared this normalized number of observed breakpoints within CNVB regions with those in permuted

regions using a one-sided t-test with the alternative hypothesis that observed > permuted. We then calculated the normalized number of long genes (>100K) overlapped with CNVB bins and compared against the permuted regions using the same strategy. The gene expression analysis was conducted by calculating the transcript per million (TPM) values for the longest gene observed in each of the CNVB and permuted regions and assessing whether they were significantly different using a one-tailed t-test.

2.3 Results

2.3.1 Determining the genetic architecture of individual neurons

When CNVs are clonal or recurrent, as in populations of cancer cells, read-depth based single-cell genomic approaches can accurately reconstruct clonal cell lineages (Garvin et al., 2015; Lim et al., 2020; Navin et al., 2011). However, neuronal CNVs are rarely clonal (Chronister et al., 2019) precluding validation in lineage-derived "sister" neurons. SCOVAL combined read-depth and phased LOH metrics (Figure 2.1A) to determine the prevalence of CNVs in single neurons isolated from the post-mortem brain of a neurotypical 49-year-old male. Samples from the same neurotypical individual were analyzed in previous studies conducted by the Brain Somatic Mosaicism Network (McConnell et al., 2017; Wang et al., 2021) and a directly relevant small study (i.e., consisting of 99 neuronal nuclei, 26 non-neuronal nuclei) that identified 11 CNV neurons (~11%) and two non-neuronal nuclei (7.6%) containing CNVs > 2Mb (Chronister et al., 2019).

Briefly, we isolated >50,000 human frontal cortical neurons using fluorescence-activated nuclei sorting of NeuN-positive nuclei. Two DNA libraries were then prepared in separate lanes on the 10X Genomics Chromium platform (Figure 2.1A); each lane obtained ~1,000 single

neuronal genomic libraries with unique barcodes. The resultant libraries (2125 total) were combined into one pool that was sequenced in two batches on an Illumina NovaSeq platform, achieving an average of 2.83 +/- 1.22 million reads per neuron. Following our previous approach (Chronister et al., 2019), we mapped reads to 5067 variable sized autosomal bins, each containing 500kb of uniquely mappable sequence (mean bin size = 569kb, range = 501 to 2812kb). Our quality control (QC) filters excluded 28 single neurons with aberrant bin-to-bin variance [i.e., Median Absolute Deviation (MAD), 2097 (>95%) libraries passed QC] and masked 308 genomic bins that were outliers in global read coverage across all neurons (Figure 2.2A-C). We adapted Ginkgo (Garvin et al., 2015) to call CNVs larger than 1Mb, defined copy number (CN) state thresholds (see Methods), and identified 2,564 putative autosomal CNVs (2,401 deletions and 163 duplications) in 469 different neurons (Figure 2.1B).

In parallel, we sequenced dural fibroblast DNA from the same individual at high coverage (~52.7X) to identify and phase germline SNPs using 10X Genomics linked-read sequencing (Weisenfeld et al., 2017). Briefly, this approach isolated and fragmented long DNA segments into barcoded short reads that could be used to reconstruct underlying haplotypes into 2548 phased genomic blocks (mean 1178kb +/- 2034kb, median 234kb). Within each of these phased blocks, we further segmented the genome into windows of 20-100 phased heterozygous germline SNPs (mean = 107kb, range = 0.687 to 1470kb) that arbitrate predicted somatic deletions with phased LOH. For each window of each cell, we counted the number of informative reads (e.g., reads that intersect with phased heterozygous SNPs) on each haplotype. We then calculated the absolute log2 ratio of the number of reads on each haplotype and integrated this ratio into the filtering models (Figure 2.1C). The application of our naïve Bayesian-based pipeline (see Methods, Figure 2.9) identified 1,985 regions with both sequence

coverage and phased LOH support consistent with heterozygous deletions in 231 neurons. We excluded Gingko deletion calls where more than 75% of internal phased SNP windows contained fewer than 3 informative reads and arrived at a call set of 1,853 heterozygous somatic deletions in 226 neurons.

Other candidate neuronal CNVs (i.e., duplications and homozygous deletions) were more challenging to validate using SCOVAL. Previous studies using read-depth alone reported more than two-fold fewer duplications than deletions (Chronister et al., 2019; McConnell et al., 2013). Using SCOVAL, we measured allelic ratios between haplotypes to assess the 163 Ginkgo duplication calls. The log2 ratios of haplotype-resolved alleles for each duplication were not significantly different from randomly sampled euploid regions of that particular cell (one-tailed t-test, p-value = 0.998, Figure 2.10A). These findings suggest that greater single-cell sequencing coverage may be required for SCOVAL to assess duplications in single neuron WGA data, although phased LOH may also allow us to filter regions where Ginkgo reports false positives (Figure 2.1F, green arrow). Nevertheless, although some of these regions may represent bona fide duplications, set we opted to exclude putative duplications with only Gingko support from further analysis in the interest of evaluating a conservative call.

Homozygous deletions have been uncommon in previous datasets and have distinct properties compared to heterozygous deletions. Specifically, these deletions are not directly amenable to allelic modeling as both haplotypes are absent and any observed non-zero allele ratios likely would be derived from mis-mapped reads. Thus, we developed an additional filter to reduce the false positive rate for 106 putative homozygous deletions with read-depth support. We calculated a read-depth ratio for each Ginkgo window by comparing the read-depth in every cell with the read-depth from bulk sequencing (Wang et al., 2021) and derived a Gaussian mixture

model to calculate the posterior probability for putative homozygous deletions using these values from our initial heterozygous and homozygous deletion calls (see Methods, Figure 2.10B) This strategy found additional support for 86/106 putative homozygous deletions (posterior probability > 0.99, Figure 2.10C). These 86 regions were included in our final deletion call set for subsequent analyses of CNV locations. Importantly, homozygous deletions are only found in neurons with highly aberrant karyotypes and all flank a heterozygous deletion (Figure 2.1F, red arrow), indicating that they are likely the result of two independent and overlapping heterozygous deletions. Further, we identified 8 Ginkgo-called homozygous deletions that exhibited a read depth and allele ratio profile consistent with heterozygous deletions and reclassified them as such (Figure 2.11).

SCOVAL produced a final deletion CNV set comprising 1,957 somatic CNV calls (13.95 Mb +/- 17.47 Mb) among 226 CNV neurons (~11%). These represent 76.3% of the initial 2,564 read depth predictions. CNV neuron prevalence (226/2097 neurons) is in good agreement with previous read-depth based CNV detection from this individual (~11% of 99 neurons) (Chronister et al., 2019). Although the nature of single-cell DNA sequencing prohibits the direct validation of identified CNVs, manual, subjective inspection of read-depth and allele ratios are strikingly concordant.

SCOVAL was designed to identify idiosyncratic CNVs in human neurons. Another single-cell CNV caller, CHISEL, was designed to study tumor evolution and intra-tumor heterogeneity (Zaccaria and Raphael, 2021). CHISEL and similar approaches (Wu et al., 2021) assume a higher frequency of tumor subclones (>5-10% (Dentro et al., 2021)) than has been observed in CNV neurons (Chronister et al., 2019) 39. When we tested CHISEL using our single neuron data, almost all reported CNVs (21,906) clustered collectively within 12 genomic loci

(99.25% of CHISEL calls) and were reported in more than 50% of neurons (Figure 2.12). Notably, 11 of the 12 loci overlapped with SCOVAL outlier bins that were associated with WGA artifacts (see Methods and (Chronister et al., 2019)). We compared the remaining 165/21906 CHISEL CNV calls with our final call set. These 165 calls were reported in only three neurons, but 39 CHISEL CNV calls overlapped with 15 SCOVAL CNV calls. Manual inspection of read-depth and LOH at the other 126 CHISEL CNV calls found no subjective support. Consistent with reports attempting to apply similar cancer-oriented approaches for identifying somatic CNVs in neurons (Wang et al., 2021), we conclude that CHISEL and other cancer-oriented approaches are not appropriate to study brain somatic mosaicism.

2.3.2 Some CNV Neurons have highly aberrant karyotypes

SCOVAL identified 226 CNV neurons with at least one deletion. These deletions ranged in size from 1Mb to entire chromosomes. We also observed that when neurons harbored multiple deletions, many clustered on single chromosomes. In contrast to a uniform background model (see Methods and below), CNVs did not appear to be distributed randomly among CNV neurons (Figure 2.3A). Forty-six CNV neurons contained a single deletion, but five contained greater than 30 deletions. Apparent chromosomal monosomies (i.e., where all genomic bins reported a copy number (CN) state = 1) were observed in six different neurons. One neuron (#1) was monosomic for Chr5, another (neuron #7) was monosomic for Chr9, two neurons (#2, 3) were monosomic for Chr13, and two other neurons (#4, 46) were monosomic for Chr18 (Figure 2.3B, C). All monosomic neuronal genomes were highly aberrant and harbored many additional deletions affecting 40 – 98% of other chromosomes (Figure 2.3C). Among 65 CNV neurons with

deletions affecting >5% of their genome, 48 contained at least one chromosome that was >50% monosomic.

We evaluated CNV locations in CNV neurons based on the percentage of each chromosome affected by CNVs (Figure 2.3C) and found two pairs of neurons (#17, #19 and #154, #155) that were nearly identical in their genomic read-depth patterns and could, in principle, represent clonal "sister" neurons that arose from a common progenitor cell during neurodevelopment (Figure 2.13). However, each of these pairs arose from the same 10X Genomics Chromium lane; therefore, we cannot exclude the possibility that one nucleus may have paired with two 10X GEM beads in a single droplet. Subsequent analyses assume that these two pairs are highly concordant technical replicates.

Hierarchical clustering (Figure 2.3D) identified three other neurons (cells #32, #33, and #47) with similar karyotypes that could, in principle, share identity by descent (Figure 2.3E). Thus, we investigated whether these deletions occurred on the same chromosomal phase block (i.e., haplotype). Multiple deletions in cells #32, #33, and #47 mapped to Chr3; however, read-depth alone cannot assess whether these deletions occur on the same physical chromosome. Also, 10X linked-read haplotyping identifies phased SNPs with Mb-scale resolution, as described above. To determine phasing at a chromosome level, we generated extended phase blocks using three CNV neurons (cells #33, #10, and #5) that contained overlapping deletions accounting for the full-length of Chr3 (Figure 2.14). Although CNV locations overlapped among these three neurons (Figure 2.3F), the Chr3 CNVs were constrained to one haplotype in two neurons (cells #32 and #47) but occurred on the other haplotype in the third neuron (cell #33). The presence of other idiosyncratic CNVs suggest that these three neurons arose in distinct neurodevelopmental lineages. The possible ontogeny of these chromosomes might include

chromosome mis-segregation, micronucleus formation, and a chromothripsis-like event (Cortés-Ciriano et al., 2020; de Pagter et al., 2015; Hatch et al., 2013; Shoshani et al., 2021; Zhang et al., 2015). In any case, the strikingly similar patterns of loss observed in these three neurons likely represent recurrent rather than clonal events.

2.3.3 CNVs are not randomly distributed in neuronal genomes

The similar patterns of chromosomal loss observed in subsets of CNV neurons led us to hypothesize that, in contrast to what has been reported in other tissue types (Liu et al., 2022), neuronal CNV locations may not arise randomly. Thus, we generated a control dataset of randomly placed deletions and explored whether neuronal genomes accumulate CNVs in "hotspots" or are protected from CNVs in "cold spots." Briefly, the empirical call set was randomly rearranged, without collision, while keeping the size and abundance of CNVs constant on a per neuron basis. We reasoned that randomly, and reiteratively, placing the "real" CNVs throughout the genome would effectively generate a "random" CNV landscape (Figure 2.4A); we then performed 10,000 synthetic iterations of real data to generate a null model. For analysis, the genome was segregated into 567 contiguous 5Mb regions and the number of simulated CNVs that overlapped each 5Mb genomic region (i.e., hits) were counted to generate a null model.

A Gaussian-shaped distribution of CNVs / 5Mb region was observed in the null model, but empirical data was enriched for observations at the extremities (Figure 2.4B). Specifically, when empirical P-values were calculated for each 5Mb region, we found eighty-three 5Mb regions (14.6%) where observed CNVs occurred more frequently than in the random model ("hotspots," P-value <0.05) and fifty-six 5Mb regions (9.9%) where empirical CNVs overlapped less frequently than in the null model ("cold spots," P-value >0.99) (see Methods for P-value

determinations). For example, fourteen 5Mb regions were hit at least 24 times by real CNVs, however this frequency (⩾24 hits in a 5Mb region) occurred in only 0.5% of null model permutations. Importantly, no CNV-free region was observed in null model perturbations, but seven CNV-free cold spots were found in empirical data.

CNV hotspots and cold spots are also clustered in several semi-contiguous stretches of the genome (Figure 2.4C). Eighty-three 5Mb hotspots clustered into 47 distinct contiguous regions, whereas the 56 cold spots clustered into 22 distinct contiguous regions. Surprisingly, individual chromosomes also clustered as either hot or cold with respect to CNV presence or absence. For example, 9/83 (~11%) and 15/83 hotspots (~18%) clustered on chromosomes 18 and 5, respectively, whereas 12/56 cold spot regions (21%) clustered on chromosome 1. Thirteen highly aberrant neuronal genomes (containing ⩾25 CNVs in empirical data) all had a CNV(s) that intersected hotspots, whereas only nine had CNVs intersecting cold spots. Similarly, of the 112 CNV neurons that contained between 1-5 CNVs, fifty-four had CNVs intersecting hotspots and only seven had CNVs intersecting cold spots. Overall, 163 neuronal genomes had a CNV(s) overlapping a hotspot, whereas only 50 CNV neurons overlapped cold spots.

Because a depletion of CNVs in some regions could artificially increase the detection of CNVs elsewhere, putative CNV cold spots and hotspots may have a technical explanation. Thus, we also functionally assessed observed cold spots for overlap with 33 germline CNVs (fifty-six 5Mb regions) that are associated with adverse neurodevelopmental phenotypes (Birnbaum et al., 2022). One third (11/33) of these germline CNVs were in cold spots. By comparison, none (0/33) of the germline CNVs overlapped hotspots. The probability that a neuropathogenic germline CNV occurs in any 5Mb genomic region by chance is approximately 33/567 (5.8%); however, empirical overlap was observed in 11/56 (19.6%) of 5Mb cold spot regions. Gene content further

distinguished hotspots and cold spots from other control regions of the genome (Figure 2.4D). Cold spots typically were gene dense (64.7 +/- 56.2 genes per 5Mb region) and were not distributed uniformly when compared to control regions of the genome. By comparison, hotspots typically were gene-sparse relative to cold spots (32.6 +/- 15.2 genes per 5Mb region).

2.3.4 Recurrent regions of neuronal genome rearrangement

The observation that neuronal deletions cluster in genomic hotspots suggested that local genomic instability could, in principle, lead to recurrent mosaicism among neurons. To explore this hypothesis, we examined CNV start or end locations (i.e., breakpoints) that were shared amongst CNV neurons. Breakpoints are defined by one of the 5067 variably sized Gingko bins that each include 500kb of mappable sequence. Among these bins, 857 accounted for two or more CNV breakpoints (termed CNVBs) (Figure 2.15 A, B), many of which (220/851; ~26%) fell within previously identified hotspots.

We next sought to determine whether the number of bins containing more than 2 breakpoints was significantly different from a random CNV distribution (i.e., the control set of CNV permutations). Given variably sized Gingko bins (Methods), we first assessed whether Ginkgo bin size impacted breakpoint frequency. While bin size scaled linearly with CNVB frequency in random permutations, this linear relationship was not observed with empirical CNVBs (Figure 2.15C). When breakpoint counts are normalized by bin size, observed CNVBs cluster more frequently in common bins than random CNVBs (one-sided t-test, P-value: 2.08*e-134), suggesting that CNVBs likely originate from a non-random process (Figure 2.15D).

Empirical CNVBs were further assessed for properties that might suggest mechanisms of CNV formation. As the endpoints of each CNV are imprecise within the Ginkgo bins, we were unable to use typical approaches that examine sequence context around precise structural

breakpoints (Lam et al., 2010); thus, we restricted our analysis to larger genomic features. Recent studies have indicated that somatic CNV hotspots in non-cancer systems are localized around large (>500kb) transcriptional units that form due to replication stress by a mechanism termed transcription-dependent double-fork failure (Wang et al., 2020; Wilson et al., 2015). To test if the CNVBs in our empirical dataset were consistent with this mechanism, we examined gene content in CNVB regions relative to random CNV permutations. Intriguingly, we observed a significant enrichment of empirical CNVBs within long genes (which we define as >100kb, one-sided t-test, P-value: 1.32*e-5), suggesting possible support for the hypothesis (Glover and Wilson, 2016; Wei et al., 2016, p. 164; Weissman and Gage, 2016) that longer genes may incur an increased frequency of DNA double strand breaks (DSBs) and, in turn, lead to neuronal CNVs (Figure 2.15E). However, the size of our detected CNVs and corresponding breakpoint windows are large and CNVB locations were not enriched for expression level (Figure 2.15F). Thus, neuronal CNVs could arise by related, but perhaps different, mechanisms.

Among 98 of the 226 CNV neurons, we observed 73 CNVs that shared both 3' and 5' CNVBs. These may be recurrent CNVs (CNVRs). Haplotype information was then used to determine if CNVRs support a clonal relationship among neurons. Briefly, we used phased allele ratios to compare whether CNVRs shared haplotypes by determining the median of the differences between the minimum and maximum log2 allele ratios observed in each SNP window within the CNVR across all cells where it was identified, reasoning that lower log2 allele ratio values would represent CNVRs on a shared haplotype (Methods, Figure 2.16A). These calculations resulted in two apparent distributions of both lower (32/73) and higher (41/73) delta log2 ratio values. The lowest delta log2 ratio cluster contained the two pairs of technical replicates, indicating the veracity of our approach. The remaining CNVRs exhibited a

delta median log2 ratio larger than 5, suggesting that these CNVs occurred on opposite haplotypes (Figure 2.16B). However, all CNV neurons harboring CNVRs had complex karyotypes with divergent CNV patterns across the genome (e.g., Figure 2.17). These findings suggest that shared CNVs are not necessarily clonally-derived, but, instead, likely represent recurrent events (Figure 2.16 C, D). Of note, similar CNVRs were observed in the analysis of cancer genomes and are referred to as "mirrored-subclonal" CNVs (Masoodi et al., 2019; Zaccaria and Raphael, 2021).

2.4 Discussion

The genetic landscape of human neurons is a mosaic of the individual's germline genome; it is likely that every human neuron accumulates more than a thousand somatic variants over a person's lifetime (Costantino et al., 2021; Jourdon et al., 2020; McConnell et al., 2017; Miller et al., 2021). Specific somatic mutations have been linked to overgrowth phenotypes in patients with hemimegalencephaly and focal cortical dysplasia (Baldassari et al., 2019; D'Gama et al., 2017; Lee et al., 2012; Lim et al., 2017). Other studies report differential somatic mutation burden in subsets of patients with autism and schizophrenia (Bundo et al., 2014; Muotri et al., 2010; Rodin et al., 2021). Furthermore, mosaic SNVs mark neural cell lineages within brain regions and some neuronal SNVs trace their origin prior to neuroectodermal specification (Bizzotto et al., 2021; Breuss et al., 2022; Wang et al., 2021). Somatic SNVs acquired in early human development can be shared among progeny in multiple germ layers. Mosaic SNVs contribute to intra-individual genetic variation, but mosaic Mb-scale CNVs alter the neurogenetic landscape in dramatic ways. However, it is unknown whether some genomic regions are more, or less, prone to CNV occurrence than other regions. The identification of CNV-prone genomic

loci, if they exist, could indicate mechanisms for somatic CNV formation, and, possibly, reveal a role for CNV neurons in brain function and disease.

Here we employed a droplet-based WGA approach to map CNVs in 2097 frontal cortical neurons from a single individual. Technical barriers have limited previous studies to extrapolation from fewer than 100 neurons per individual and reported a total of 129 CNV neurons out of 879 frontal cortical neurons examined among 15 individuals (Chronister et al., 2019. We developed SCOVAL to add veracity to read-depth based CNV detection through an analysis of haplotype drop out. We showed high concordance between heterozygous deletions identified by read-depth and by phased LOH in single neuronal nuclei. In this sample, we found that 226/2097 (10.8%) of neurons harbor at least one Mb-scale CNV, and that 2% of CNV neurons exhibited aneuploidies. Moreover, we found that 65/226 CNV neurons contained many deletions across multiple chromosomes leading to highly aberrant karyotypes.

By combining haplotype and read-depth approaches, we have strong confidence that neuronal genomes contain large segments of chromosomes that are not sampled using single-cell sequencing approaches. This finding is consistent with previous reports that have examined a limited number of cells from neuronal and non-neuronal tissues using multiple technologies. Although we posit that the assayed sequence is missing because the corresponding segments have been deleted in vivo, unexpected technical or biological factors may yet contribute to the loss of signal. For example, neuronal preps exclude micronuclei (Ye et al., 2019) however, the appreciable occurrence of micronuclei in neuronal tissue would still reflect an underlying alteration in genome content in the brain. Similarly, the lack of validated duplications in single-cell neuronal sequencing is striking. Further study is required to develop mechanistic or technical explanations for this disparity.

Our finding of a nonrandom distribution of 1,861 deletions among 226 CNV neurons also allays concerns of random technical artifacts in neuronal CNV detection. Spurious WGA events, such as uneven genome amplification, are expected to occur randomly across the genome and are physically limited in size by the processivity of the polymerase (<20kb). Multiple whole genome amplification (WGA) approaches have been performed on single human neurons; all of these reported Mb-scale CNVs (Cai et al., 2014; Knouse et al., 2016; McConnell et al., 2013). This technical concern was addressed previously (Chronister et al., 2019; Rehen et al., 2005) wherein a similar prevalence of CNV neurons was observed in two samples from the same individual (26 year-old), subjected to different WGA approaches. In Chronister, et al., parameter optimization on synthetic datasets limited read-depth based CNV detection to false positive rates <5%. Here, we provide additional lines of evidence that single-cell approaches for neuronal CNV detection are robust to technical artifacts. First, we showed that SCOVAL finds haplotype allele-level support for 76% of read-depth based deletion calls. Importantly, 99% of >10 Mb heterozygous deletions received orthogonal support via phased LOH. Second, when SCOVAL was applied to 2,097 neurons, the fraction of CNV neurons observed (10.8%) was concordant with the fraction (11.1%) identified using different chemistry on a smaller (99 neuron) sample from the same brain region. Perhaps most strikingly, we identified CNV hotspots and cold spots that were inconsistent with a random distribution of technical artifacts. Moreover, these data resolved disparate reports regarding aneuploid human neurons. Approaches that measured single (or few) chromosomes in each neuron suggested that >10% of neurons were aneuploid (Rehen et al., 2005; Yurov et al., 2007). Extrapolations based on these data did not account for unmeasured chromosomes in the same neuron, implicitly assuming that every measured aneusomy was unique. We identified 6 aneuploid neurons (2.7%), consistent with other reports (Knouse et al.,

2014; van den Bos et al., 2016). These and other CNV neurons harbored additional deletions that covered >50% (52/2095) of chromosomes and could be scored by traditional hybridization-based approaches as aneuploid on multiple chromosomes.

In addition to finding a nonrandom distribution of CNVs among CNV neurons, we identified genomic hotspots that were impacted by neuronal CNVs more often than expected by chance; the same approach identified genomic cold spots. Further analysis of these regions found high gene density in cold spots (64.7 +/- 56.2 genes per 5Mb region), but a lower gene density (32.6 +/- 15.2 genes per 5Mb region) in hotspots. Complementary analysis identified 851 regions with 2 or more CNV breakpoints (i.e., CNVBs), and found that 220 of these refined previously defined 5Mb hotspots to +/- 0.5Mb. Hotspot CNVBs were enriched for long (>100Kb) genes, consistent with the paucity of genes found in these regions. In some cases, the functional consequences of the CNVs are also suggested by associations between long gene expression, neuronal development, and neuropathologies (Gabel et al., 2015; King et al., 2013). For example, we identified seven neurons with distinct CNVs sharing a breakpoint region within KCNT2, a long (~380kb) gene that encodes an outward-rectifying potassium channel. KCNT2 is important for neuron function and has been linked to several developmental pathologies (Ambrosino et al., 2018; Gururaj et al., 2017; Mao et al., 2020) (Fig. 4B). KCNT2 exhibited a TPM of 7.30, which falls within the expected range when considering the expression of all long genes in this tissue (mean TPM 9.56 +/- 19.82).

Our study shows that CNV neurons with highly aberrant karyotypes populate neurotypical human frontal cortex. Although their impact on neural circuits and behavior remain unknown, cross-sectional studies indicate that CNV neurons are selectively vulnerable to aging-related loss (Chronister et al., 2019). The extent to which recurrent CNV sites are shared among

individuals is not yet known; neither is it known if cold sites are refractory to CNV formation or are detrimental to neuronal survival during development. Nevertheless, we report candidate genomic regions that incur frequent neuronal gene rearrangement provides a rationale for tractable and scalable targeted single-cell sequencing. Many interesting questions follow from this study, including whether cold spots in neurotypical individuals are instead aberrant in individuals with neurological disease.

2.5 Data and materials availability

Data and call sets have been deposited in the NIMH Data Archive (NDA Study ID 1680, http://dx.doi.org/10.15154/1527774) and can be accessed as part of the NIMH Data Archive permission groups: https://nda.nih.gov/user/dashboard/data_permissions.html. The workflow to generate the final call set is available at https://github.com/mills-lab/Scoval.

2.6 Notes and acknowledgements

This work was previously published in biorxiv in March 2023 (doi: https://doi.org/10.1101/2023.03.07.531594). This work represents a group effort. I did most of the computational jobs for developing the SCOVAL pipeline and following analysis. Kunal Kathuria performed the Ginkgo adaptation and permutation test for the cold and hot spots analysis. Ryan Mills and Michael McConnell supervised the study. Other contributors include Sarah Emery, ByungJun Kim, Ian Burbulis, Joo Shin, Daniel Weinberger, John Moran, Jeffrey Kidd. We thank Drs. Thomas Wilson and Fred Gage for essential insight and helpful critique throughout the study, and ML Gage for editorial assistance. We also thank M Wolpert and M

Haakenson for technical assistance. This research is also a part of the Brain Somatic Mosaicism Network consortium.

Figures

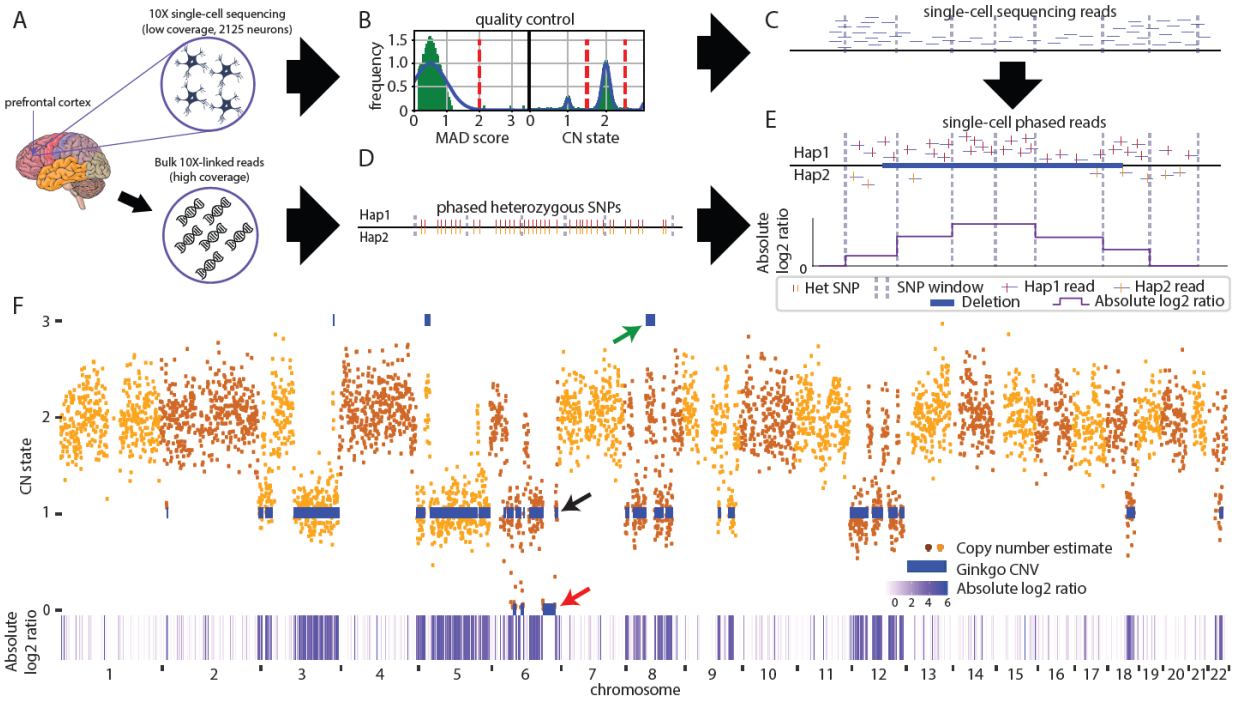


Figure 2.1 SCOVAL: identification of copy number variation using read-depth and allele imbalance.

(A) Single nuclei and bulk dural fibroblast DNA were analyzed using 10X platforms. (B) Single nuclei library quality is assessed based on median absolute deviation (MAD) and copy number thresholds are established using population statistics. Graphs depict schematized data; vertical red lines illustrate threshold strategy. (C) Candidate CNVs are identified based on altered read depth across consecutive genomic bins. (D) Heterozygous SNPs are phased using bulk linked-reads in chromosomal segments ("hap 1" or "hap 2"). (E) Absolute log2 ratios derived from "hap1" / "hap 2" are calculated across ~100 SNP windows (see text). A deletion with concordant loss of heterozygosity (log2 ratio <> 0) is illustrated. (F) A highly aberrant CNV neuron (#5) shows representative Gingko calls (blue bars), duplications (e.g., green arrow), heterozygous deletions (e.g., black arrow), and homozygous deletions (e.g., orange arrow) and qualitatively concordant increases in absolute log2 ratio (white<purple). The genome is plotted from left to right on the x axis, read-depth is in the upper panel (CN state on the Y axis) and absolute log2 ratios are reported in the lower panel.

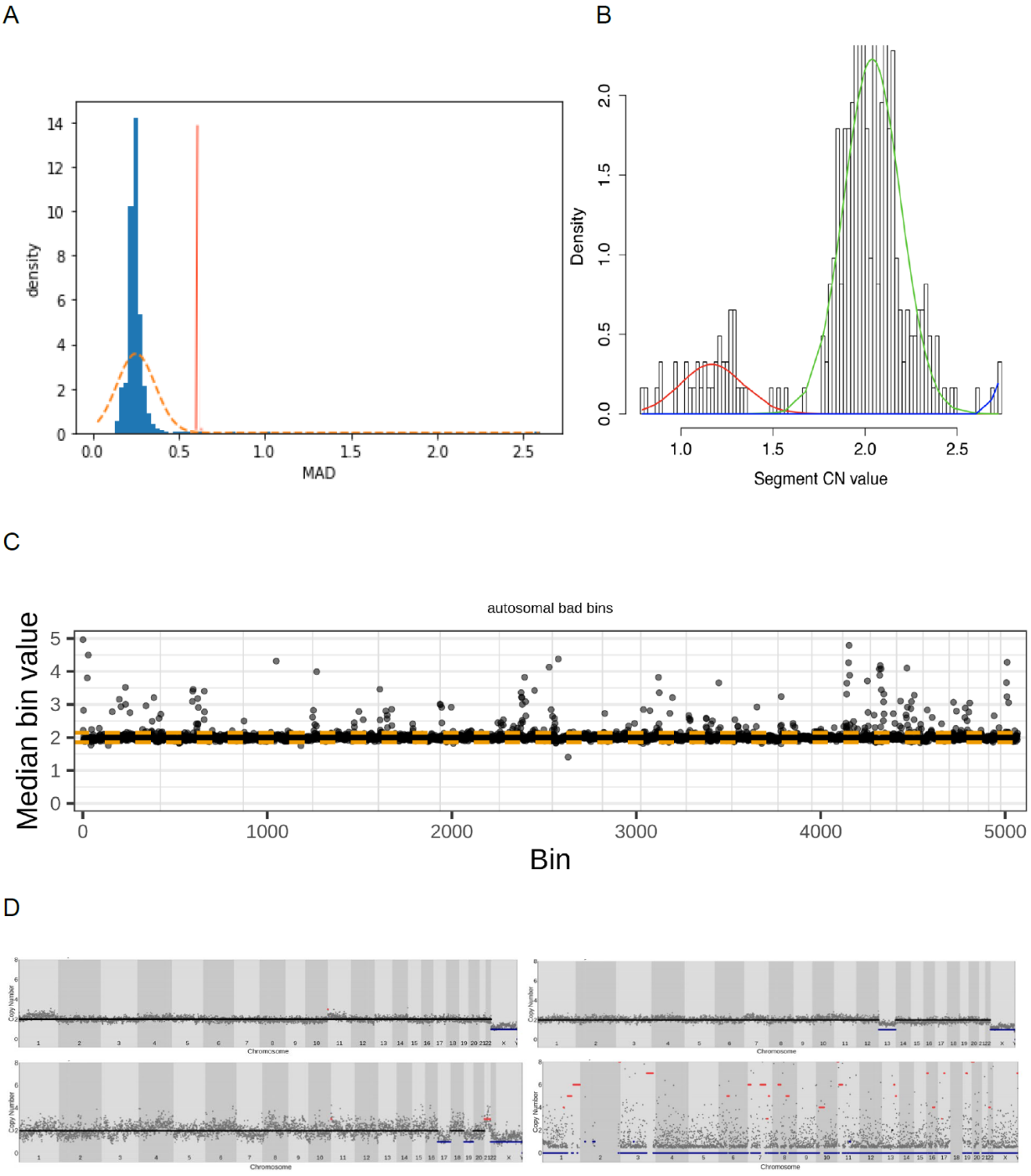


Figure 2.2 Optimization of Ginkgo for read-depth-based CNV calls.

(A) Mean absolute deviation (MAD) score distribution (based on bin copy numbers) excluded 19 of 2,125 neurons (with MAD > 3 standard deviations away from mean). (B) Thresholds for calling putative CNVs were set using a GMM based on 585 cells obtained from the 5 control

individuals studied in (20) at 1.63 for deletions and 2.43 for duplications. (C) Tukey's rule was applied to median copy numbers for all genomic bins across all neurons in our dataset to yield 308 additional outlier bins in addition to Ginkgo's original 29 that were excluded from further analysis. (D) 4 additional cells that passed the MAD cutoff but were curated manually due to unlikely copy-number patterns, including 1 that did not pass the read-count filter due to concentration of reads on Chr2 (bottom right).

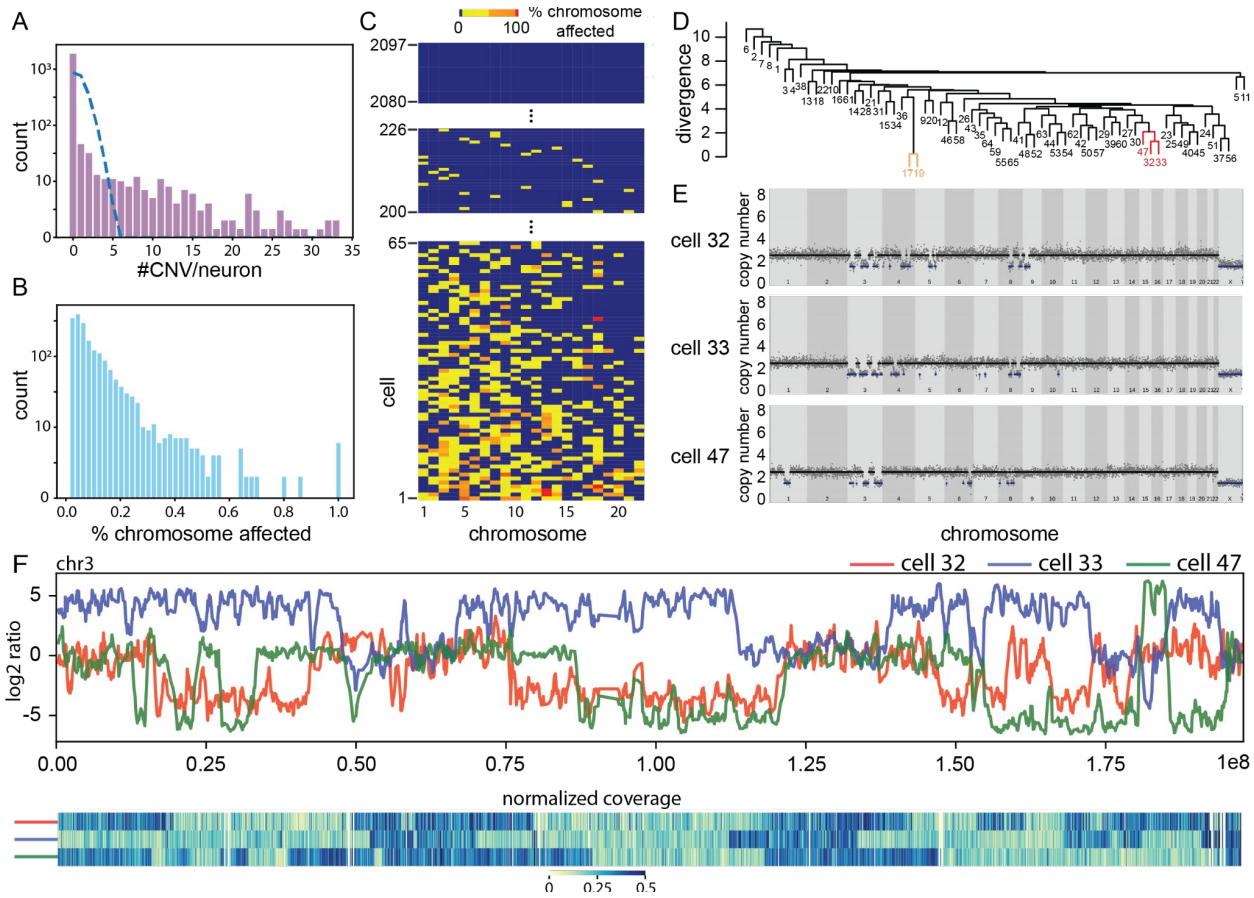


Figure 2.3 CNV neurons can have highly aberrant karyotypes.

(A) The observed CNV per neuron [(purple bars, counts (y axis), CNVs/neuron (x axis)] distribution deviates ($P < 0.0001$) from Poisson expectations (dashed blue line). (B, C) Deletions cluster in a subset of CNV neurons. (B) Counts (y axis) of the cumulative percent of each chromosome deleted (n = 2097 neurons * 22 autosomes) in CNV neurons. (C) Neuronal genomes (n=2097) are arranged in a cells-by-chromosome matrix, ranked by the total percentage of their genome containing deletions. Cell #226 is the first CNV neuron among 2097 total neurons with the smallest observed single deletion (blue = unaffected chromosome, yellow <50%, orange = 50 - 99%, red 100%). (D-F) Among 65 neurons with the most aberrant genomes, some have similar karyotypes. (D) Hierarchical clustering identifies two groups (yellow, red) with the least divergence from similarity (y axis). (E) Red cluster neurons [cells #32, 33, and 47 in (C)] have similar CNV profiles. Read-depth is plotted as in Fig. 1F. The yellow cluster (cells #17 and #19) is shown in Fig. S6B. (F) Concordant read-depth is observed on opposite haplotypes in the most similar pair [#32(red) and #33(blue)]. When overlapping, events on cell #47 (green) match the #32 haplotype, but never the #33 haplotype. Chromosome 3 is plotted from left to right. Haplotype log2 ratio (upper panel) and corresponding read-depth (lower panel, blue = diploid) plots show overlapping deletions and LOH for each haplotype.

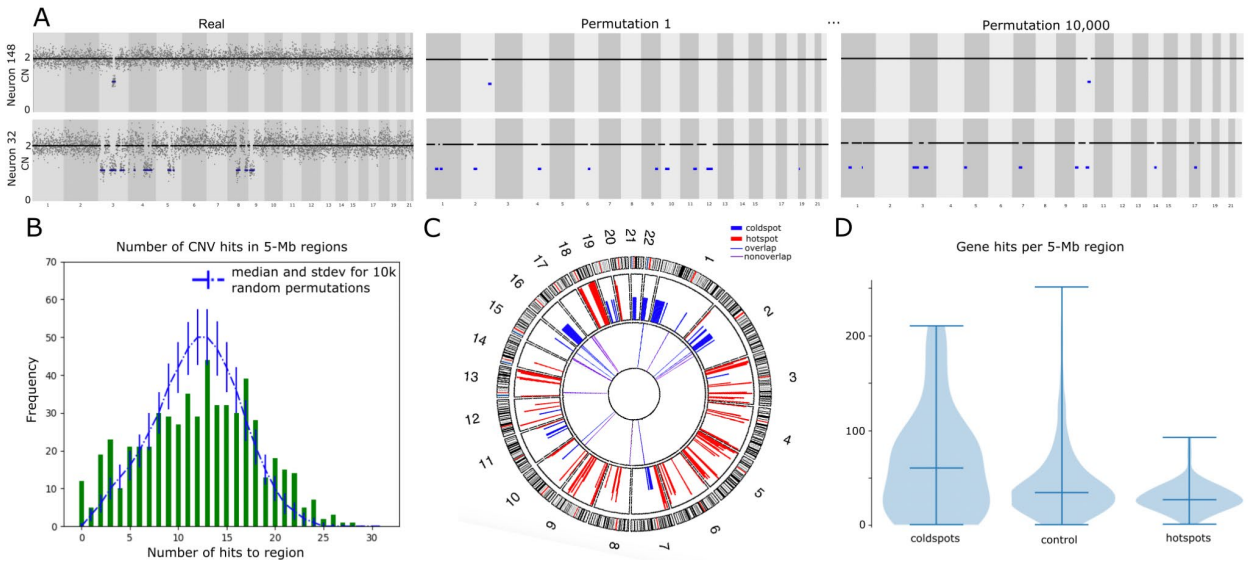


Figure 2.4 Analysis of CNV distribution relative to random null model.

(A) Empirical read-depth plots of two CNV neurons (left panels) and representative permutations (right two panels) are displayed as in Fig. 1F. (B) Relative to 10,000 permutations of real data (represented by blue dotted line and error bars), high and low CNV burden are enriched at the extremities of the Gaussian distribution (green bars). (C) Circos plot shows that hotspots (red, outer tier) and cold spots (blue, middle tier) cluster on distinct chromosomes. Thirty-three pathogenic CNVs (blue, purple, inner tier) never overlap hotspots. Eleven (blue) overlap cold spots. (D) Violin plot showing gene enrichment in cold spots (left) and depletion in hotspots (right) relative to other 5Mb regions ($P<0.001$).

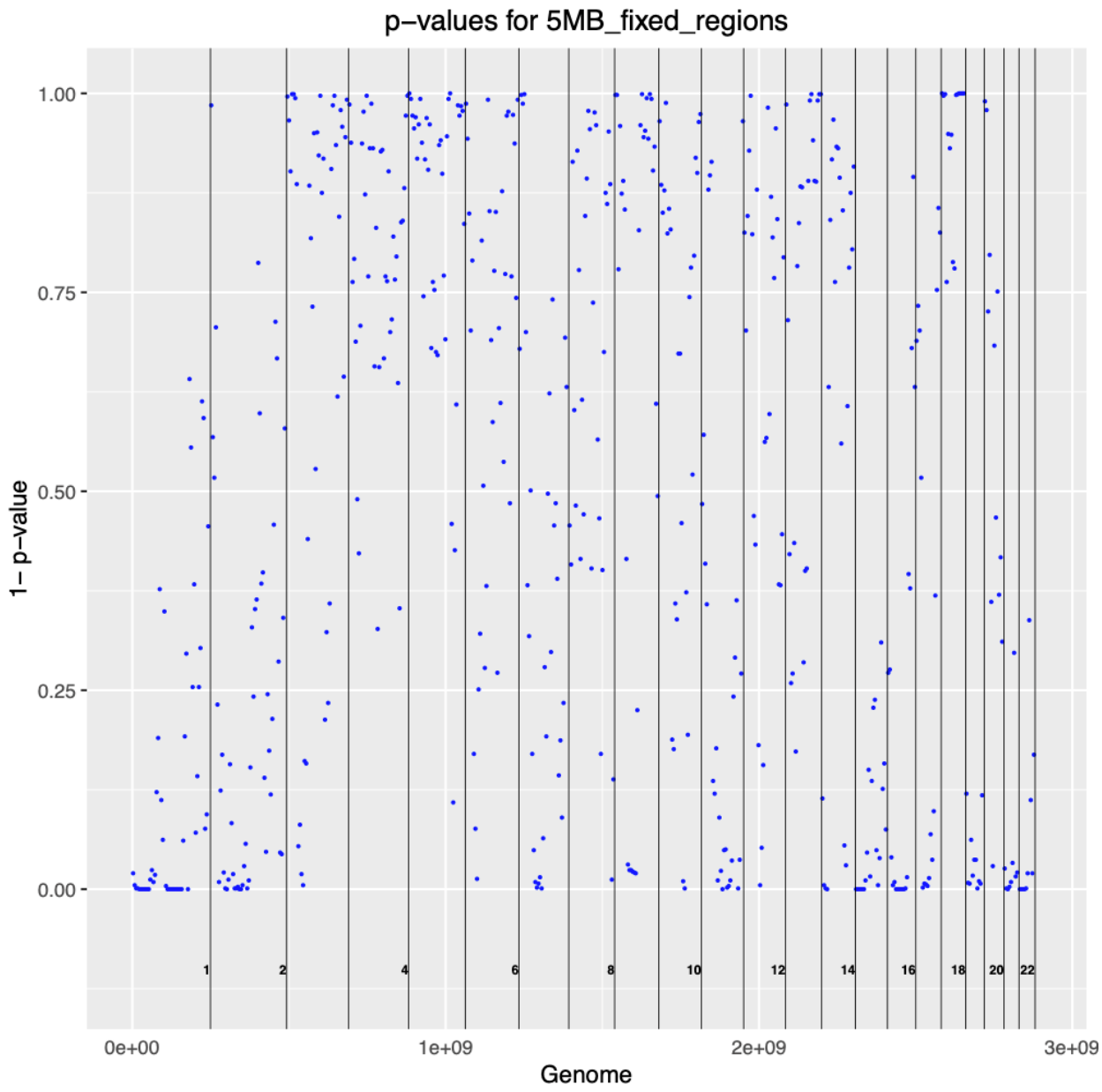


Figure 2.5 Genome-wide identification of hotspots and cold spots.

Regional significance (1 minus p-value of number of CNV hits in a 5Mb region compared to random synthetic data) is plotted for all 5Mb genomic regions. Hotspots and cold spots shown in Fig. 3C. were identified using values > .95 and < .01 respectively.

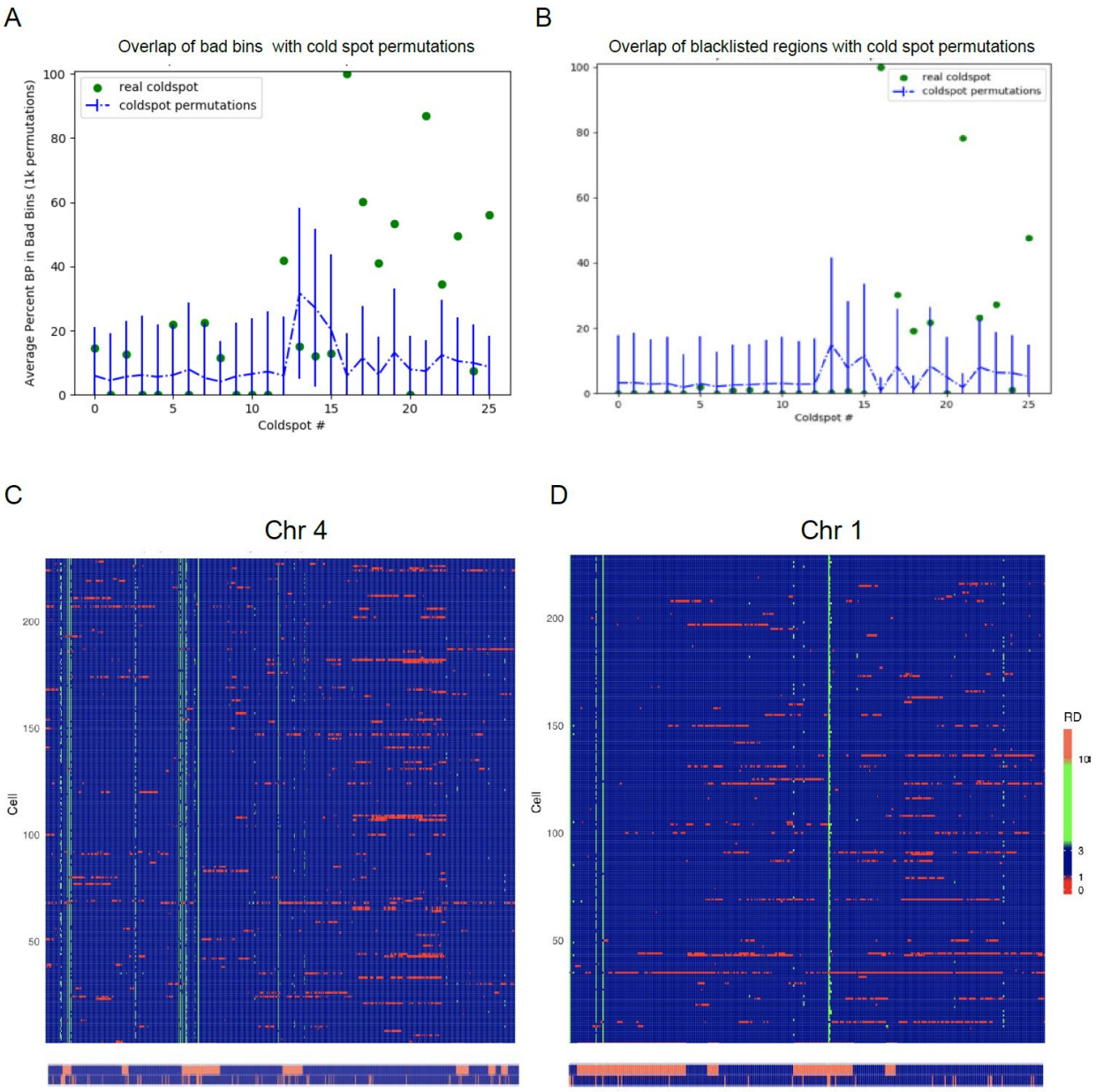
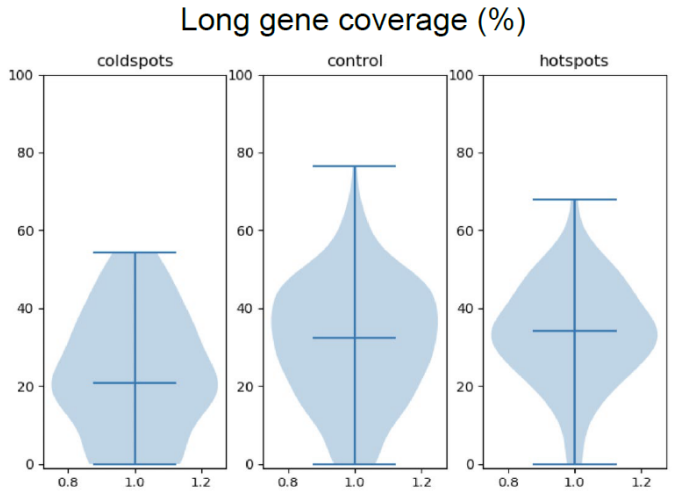


Figure 2.6 Filtering of cold spots based on unmappable genomic regions.

(A, B) Percentage of cold spots occupied by bad bins and by blacklisted regions identified by ENCODE respectively (green) compared to median of same quantity for cold spot permutations in control regions (blue). Cold spots registering high unmappable content (p-value < .05 cutoff) were filtered out (see Methods). (C, D) Schematic overview showing correlation of cold spots, bad bins and read depth for all CNV neurons across all genomic bins for 2 different chromosomes.

A



B

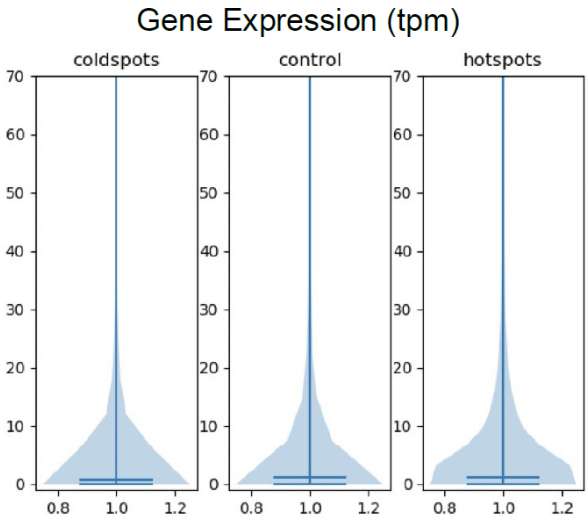


Figure 2.7 Comparison of hotspots and coldspots to control regions regarding long gene coverage and gene expression level.

Violin plot distribution showing (A) a depletion of long genes (> 100 Kb) covering the region and (A) overall gene expression of genes in the region relative to control (middle) and hotspots (right).

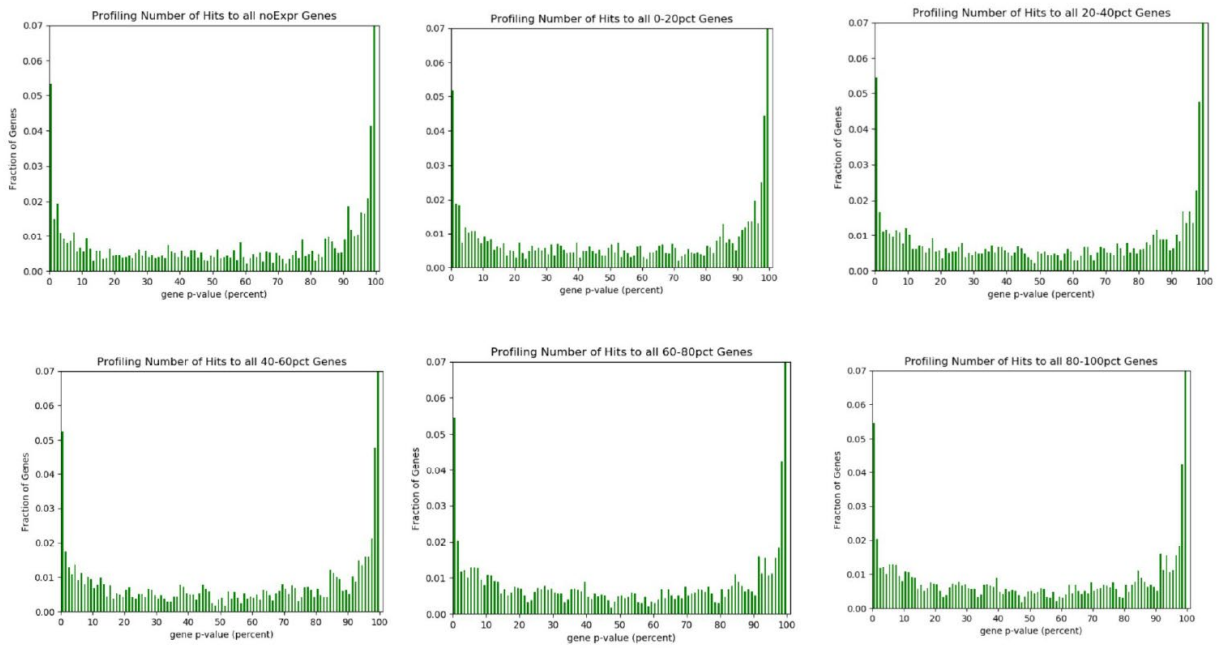


Figure 2.8 Complementary view of hotspots and cold spots in physical genes.

Depicted are the distribution of p-values (defined as in Supplementary Figure 9 but for physical genes) for genes showing the presence of hotspots (first 5 bins) and cold spots (last bin) in six expression categories (genes not expressed, and 5 quintiles of genes expressed in DLPFC). This analysis is complementary to that performed in 5Mb regions and shows the presence of hotspots and cold spots in genes expressed at various levels.

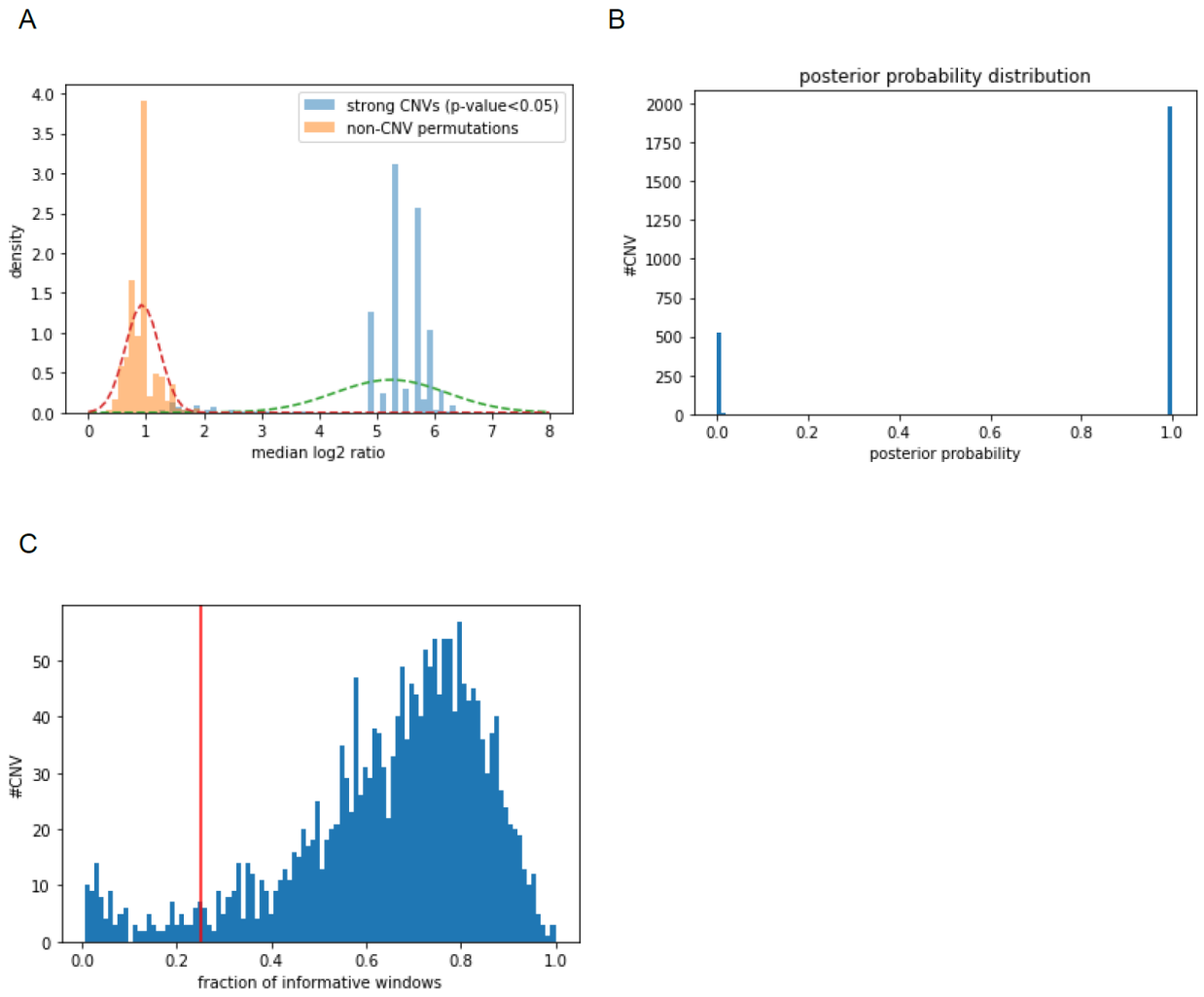


Figure 2.9 Naïve Bayesian-based pipeline to filter CNVs.

(A) We labeled CNV calls as "strong" based on an empirical p-value, which is derived from the median absolute log2 ratio of the windows within the CNV regions. Then we derived a Gaussian mixture model of strong calls and 100 non-CNV set permutations. (B) Using the median absolute log2 ratios of the two datasets as the training data, we estimated the parameters of the Gaussians and predicted the posterior probability that a candidate CNV belonged to a specific CNV distribution. (C) We filter out deletion calls where more than 75% of its het-SNP windows contained fewer than 3 informative reads which precludes an accurate haplotype assessment.

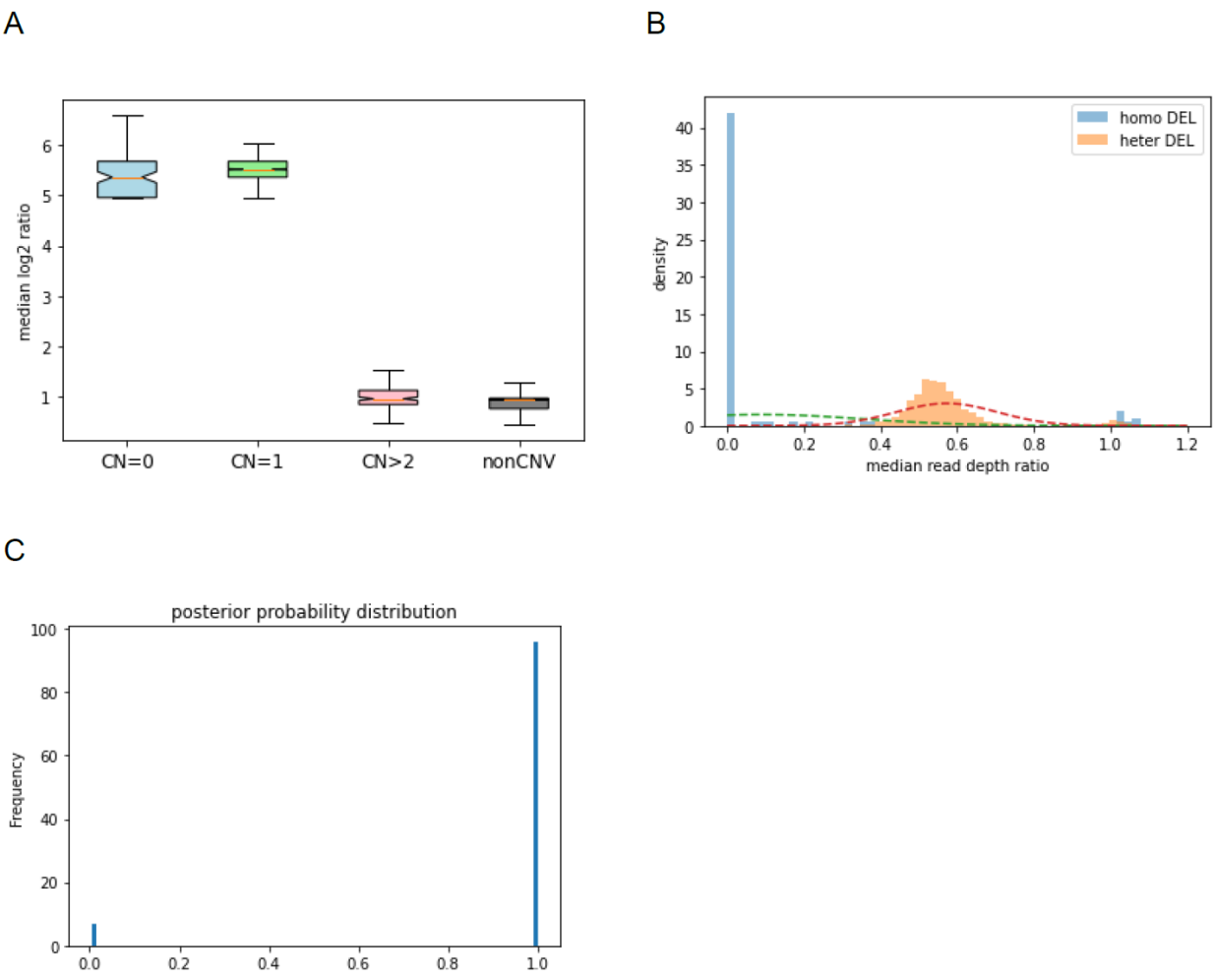


Figure 2.10 Homozygous deletions and duplications are more challenging to validate using SCOVAL.

(A) The median absolute log2 ratio of informative reads in candidate homozygous deletions informative reads are similar to heterozygous deletions. The median absolute log2 ratio of duplications are not significantly different from randomly sampled non-CNV regions. (B) Derived Gaussian mixture model from median read depth ratios between homozygous and heterozygous deletions. (C) Posterior probability for putative homozygous deletions using a naive Bayesian classifier on the Gaussian mixture model from the initial heterozygous and homozygous deletion calls.

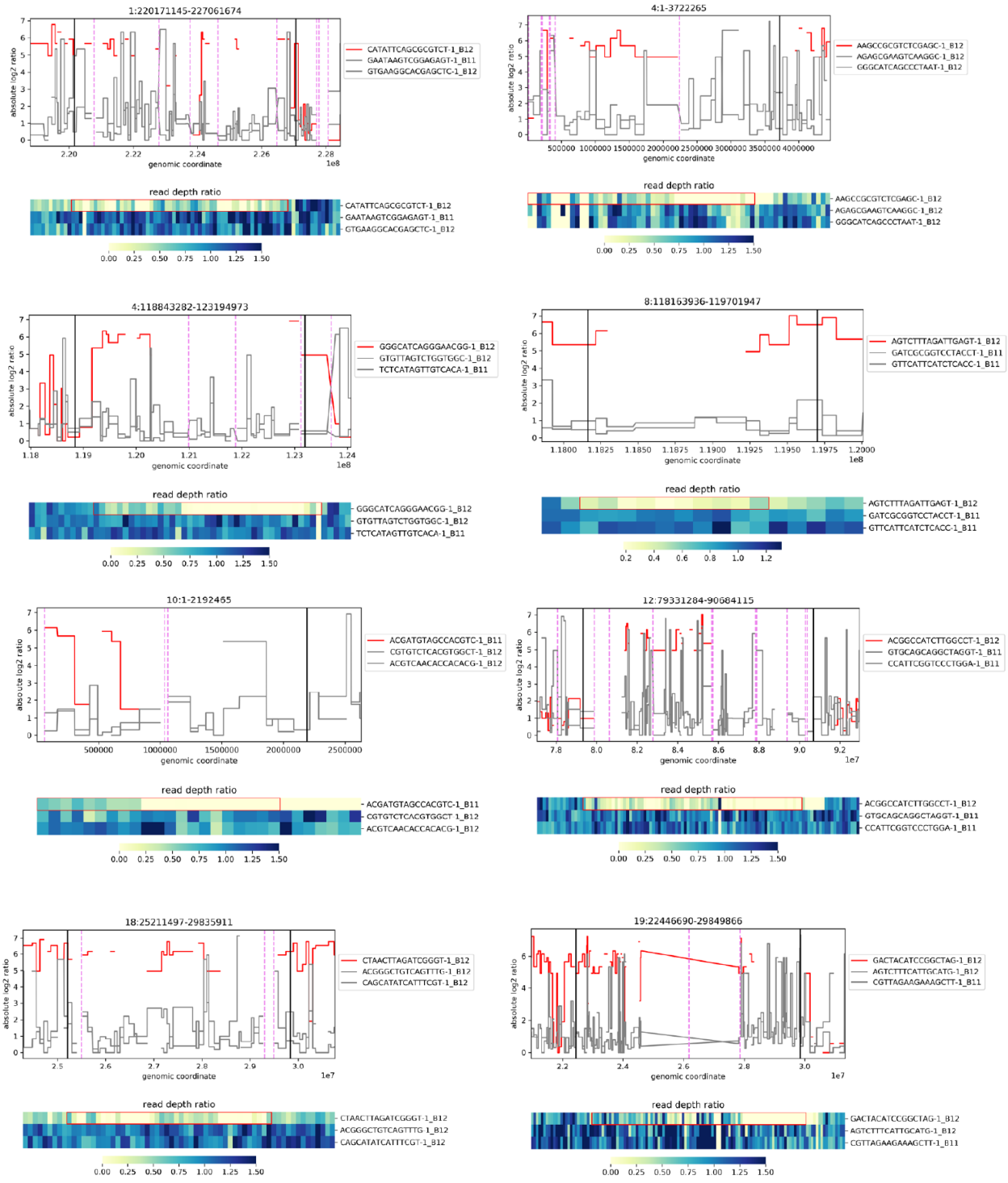


Figure 2.11 Heterozygous deletions miscalled as homozygous deletions.

We identified 8 homozygous deletion calls from Ginkgo with read depth and allele ratio characteristics consistent with heterozygous deletions. The upper panel for each figure is the absolute log2 ratio. Red line indicates the cell with CNV and the gray lines represent two random

background cells. The bottom panel is the read depth ratio. The first row is for the cell with the candidate CNV, supplemented in rows two and three with randomly chosen cells as background.

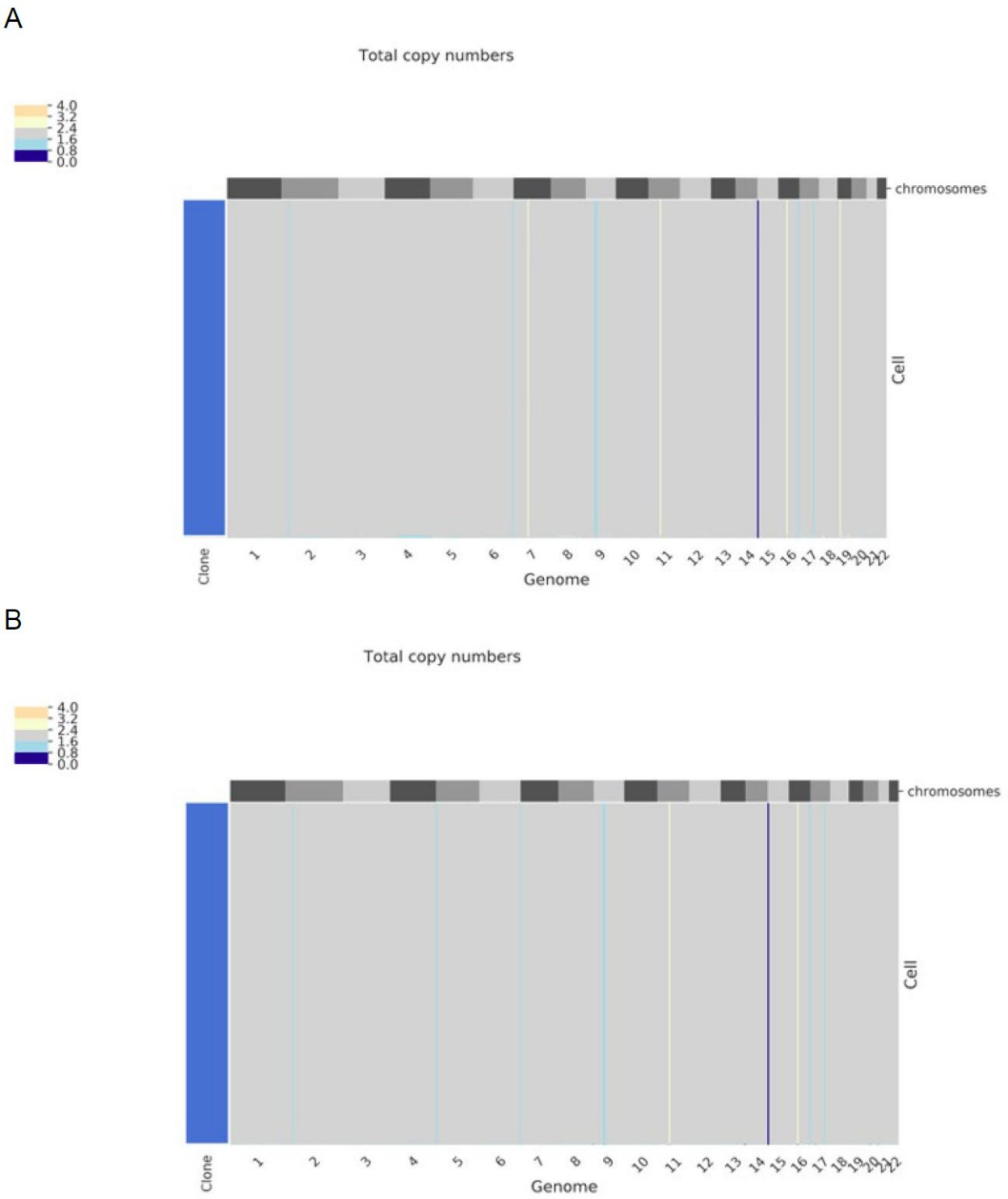


Figure 2.12 Benchmarking CNV detection with CHISEL.

Output of CHISEL to our single-cell sequencing data in (A) batch B11 and (B) batch B12 using diploid=2 parameters. The majority of CNVs were reported in all cells.

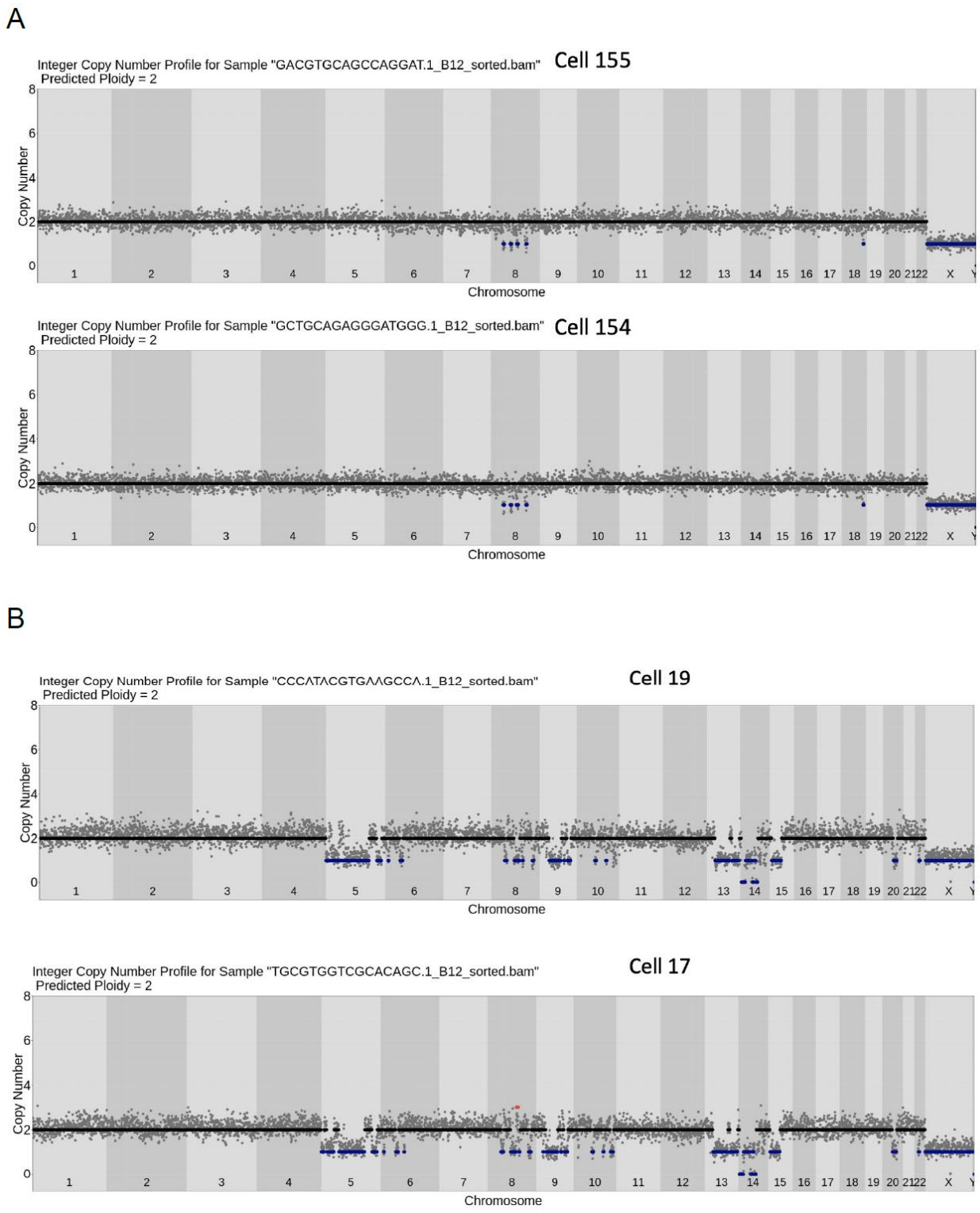


Figure 2.13 Putative clones that cannot be ruled out as technical replicates.

(**A**) Putative clone-pair 1 (**B**) Putative clone-pair 2. Cells #17 and #19 show some deviances in overall bin copy number variances, but cannot conclusively be established as independently amplified neurons.

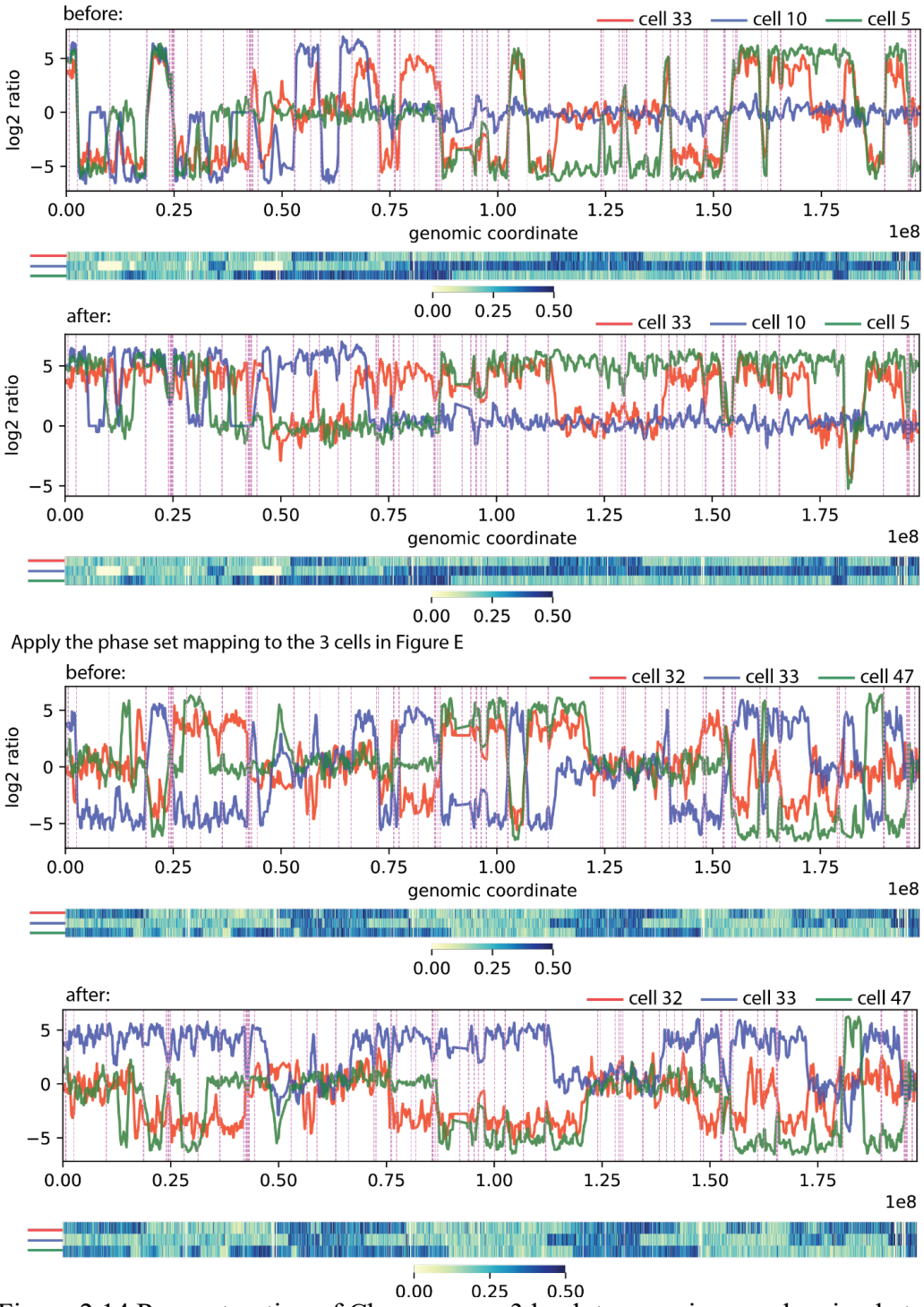


Figure 2.14 Reconstruction of Chromosome 3 haplotypes using overlapping heterozygous deletions in 3 cells.

We generated extended phase blocks using three CNV neurons (cells #33, #10, and #5) that contained overlapping deletions that in aggregate cover the full-length of Chromosome 3 in order to determine phasing at chromosome level. These were used to reconstruct the haplotype of 3 cells reported in Figure 2.3E.

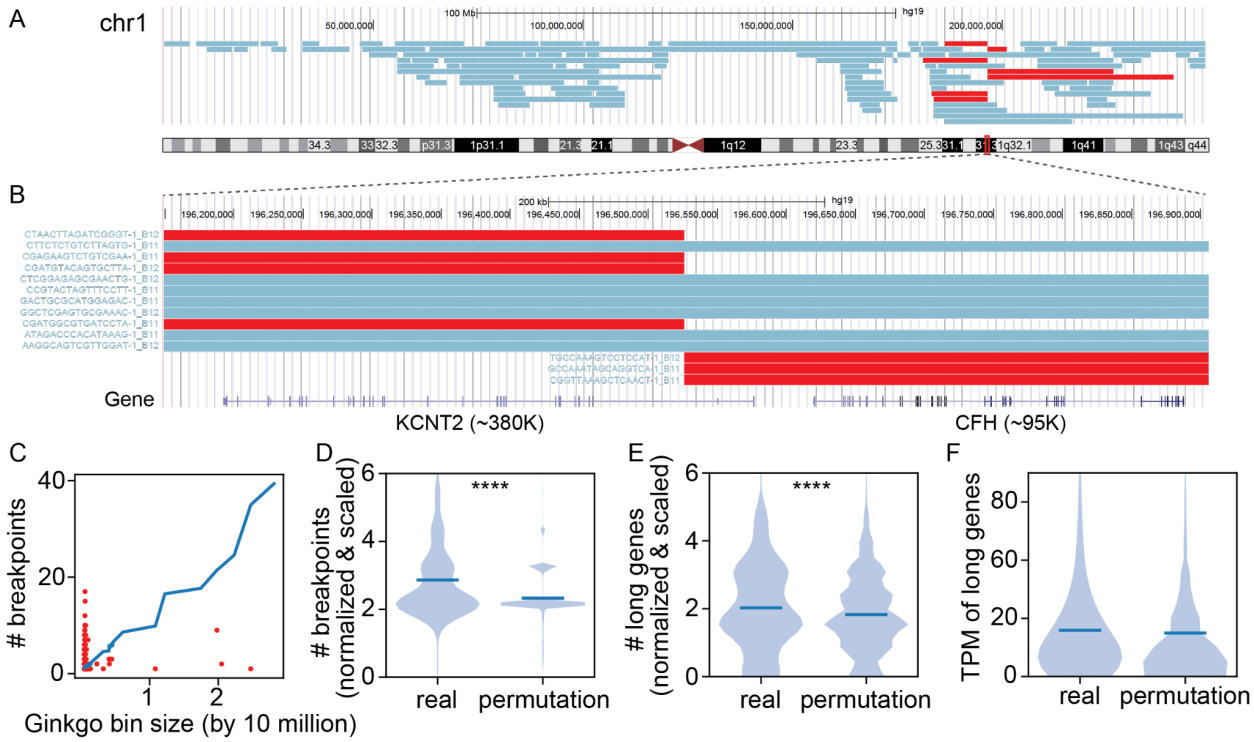


Figure 2.15 Recurrent CNV breakpoints across multiple neurons.

(A) UCSC Genome Browser view of all CNVs detected on Chromosome 1 (47 neurons, rows). Seven neurons (red) contain CNVs that share a breakpoint region (CNVB). (B) Representative CNVB (red) on Chromosome 1 overlaps (+/- 250kb) two genes (lower panel). (C) Number of breakpoints identified in each Ginkgo bin (y axis) relative to bin size (x axis), shown for bins containing two or more CNVs (red) and averaged across all permutations in control set (blue line) (D-F) Violin plots show real and permuted data sets, normalized by bin size, when examined for (D) number of breakpoints, (E) number of long (>100k) genes (**** $p<0.0001$ for one-tailed t-test), and (F) transcripts per million bp (TPM) values of the longest gene in each bin.

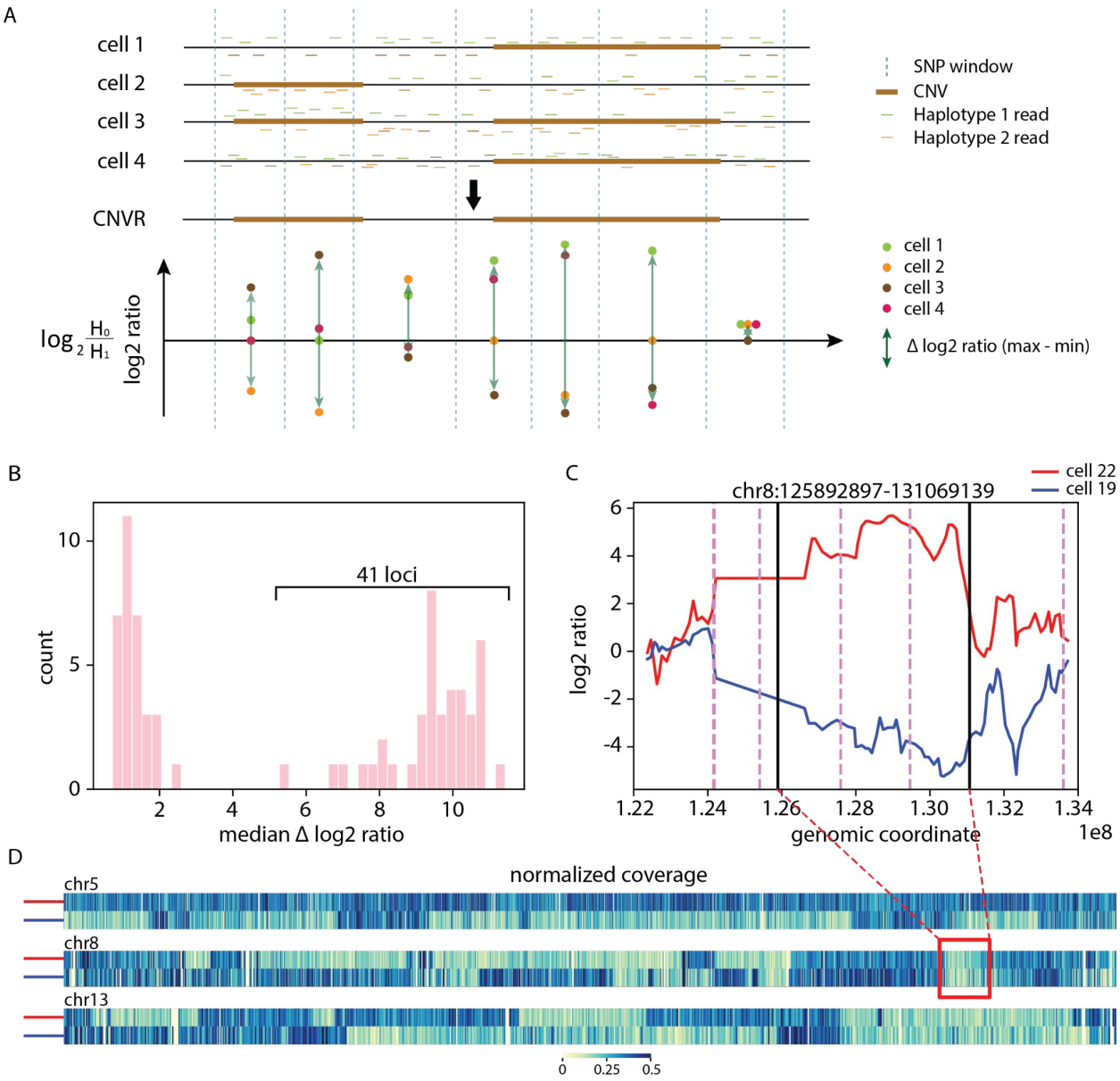
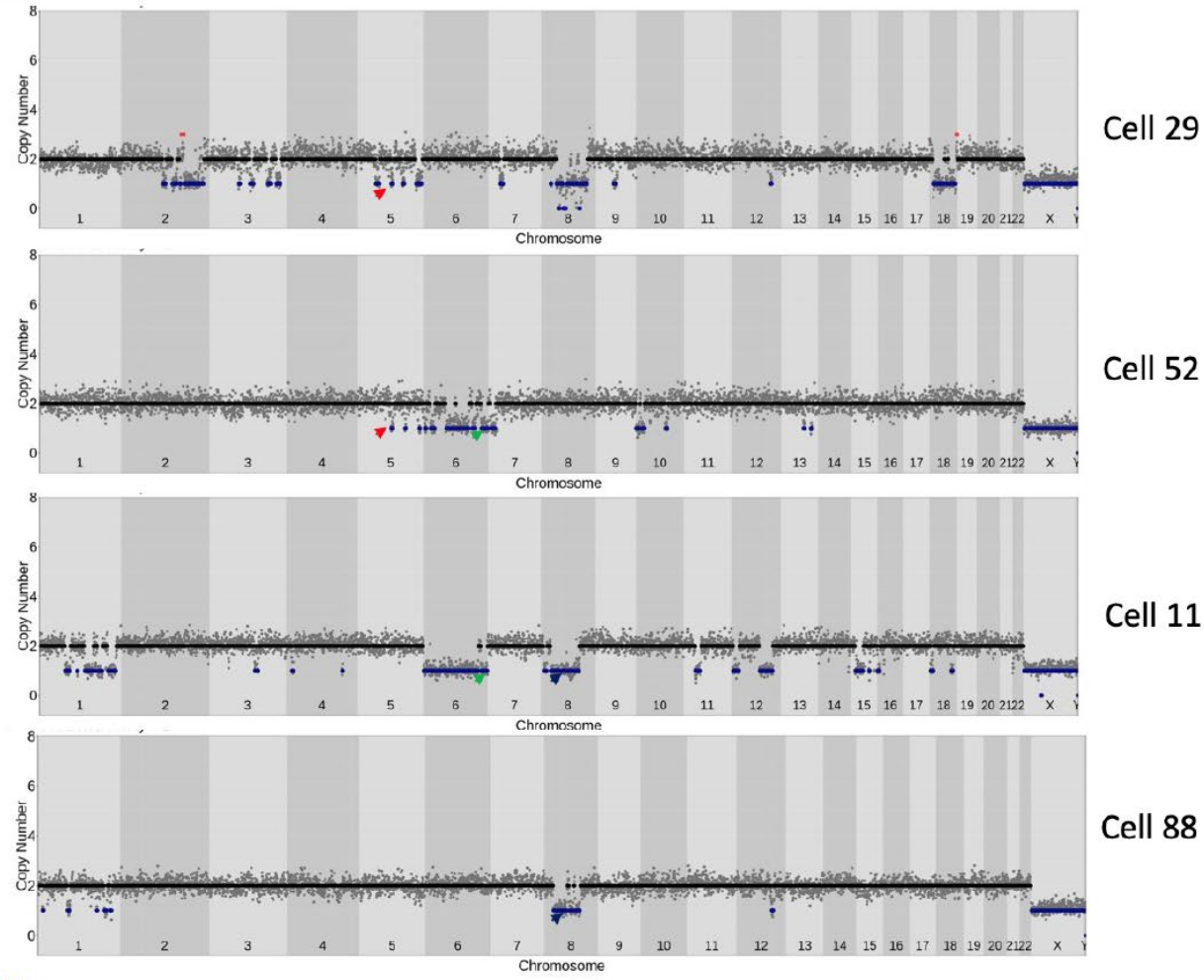


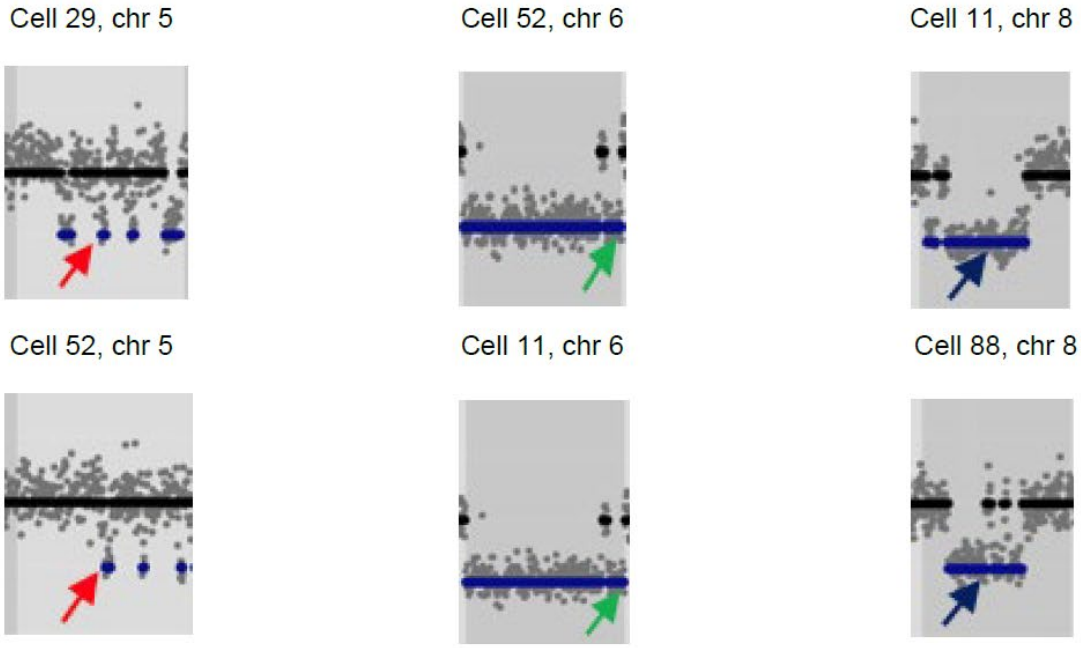
Figure 2.16 CNVs sharing the same location are on different haplotypes.

(A) We derived a min-max median delta log2 ratio to determine whether CNVRs likely reside on the same haplotype. (B) There are two apparent distributions of delta log2 ratio values. CNVs from 41 CNVRs with higher median delta log2 ratio likely occurred on different haplotypes. (C) Two cells (#22 and #19) both exhibit CNVs with the same location on Chr8, but show allelic ratios consistent with residing on different haplotypes. (D) An examination of CNVs on other chromosomes in these cells further indicates that these shared CNVs are not clonally derived.

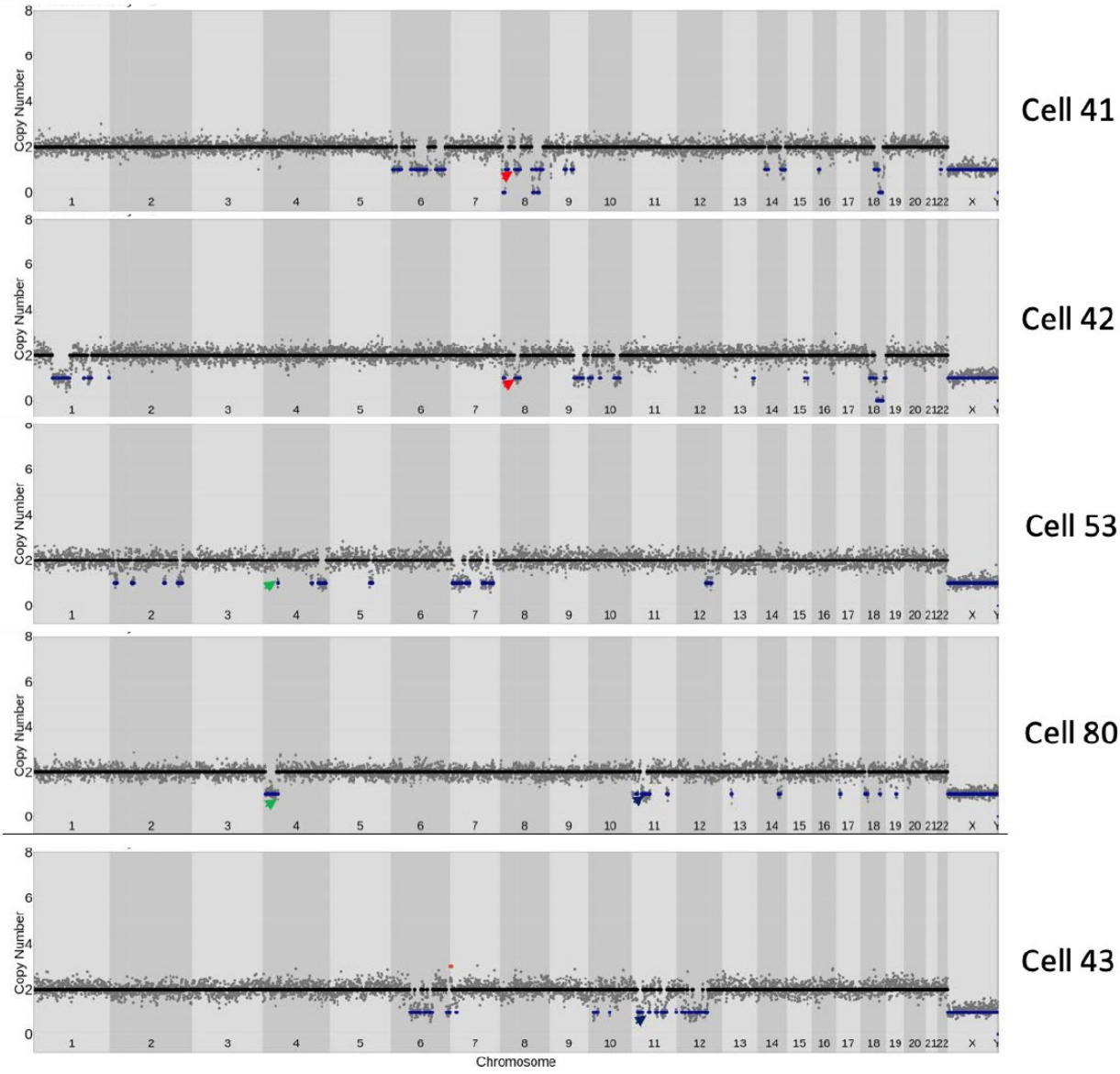
A


Copy Number
Chromosome
Cell 29
Cell 52
Cell 11
Cell 88

B


Cell 29, chr 5
Cell 52, chr 6
Cell 11, chr 8
Cell 52, chr 5
Cell 11, chr 6
Cell 88, chr 8

C


Copy Number
0
2
4
6
8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
Cell 41
Cell 42
Cell 53
Cell 80
Cell 43

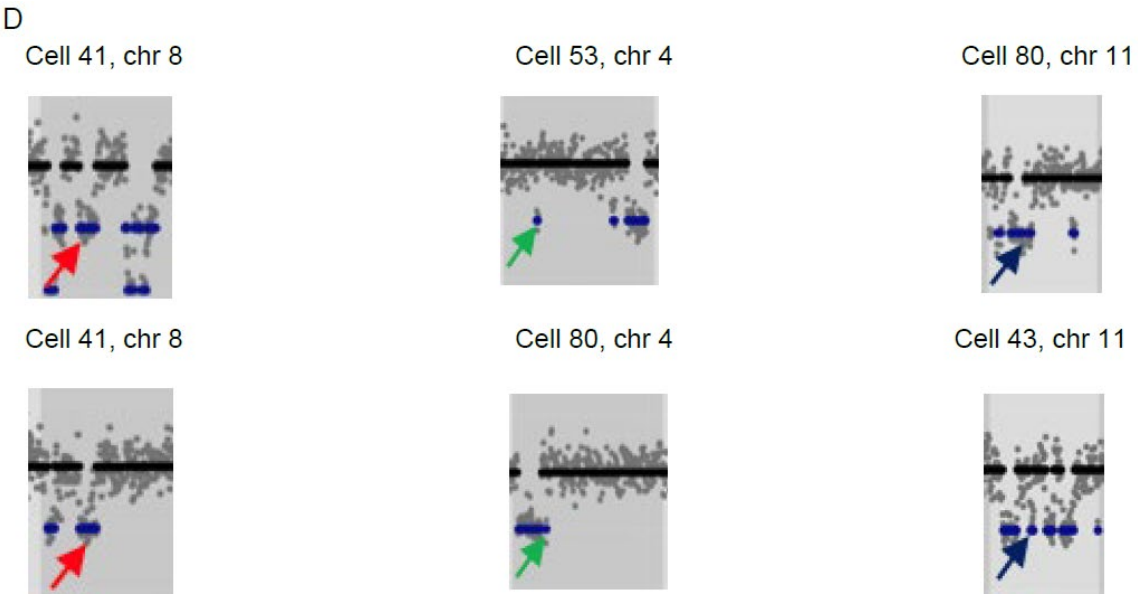


Figure 2.17 Cells showing single shared events among complex karyotypes.

(A, B) Three pairs of CNVs in 4 cells (shown by red, green and blue arrows respectively) are shared/recurrent. The shared CNVs are magnified in the lower panel. None of the other CNVs are shared. This indicates that the recurring CNVs are not necessarily clonal. (C) Same as above for another group of 5 cells. (D) The lower panel shows a magnified view of shared CNVs in (C)

Bibliography

Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S.V., Ellis, P., Russell, A.J.C., Alcantara, R.E., Baez-Ortega, A., Wang, Y., Kwa, E.J., Lee-Six, H., Cagan, A., Coorens, T.H.H., Chapman, M.S., Olafsson, S., Leonard, S., Jones, D., Machado, H.E., Davies, M., Øbro, N.F., Mahubani, K.T., Allinson, K., Gerstung, M., Saeb-Parsy, K., Kent, D.G., Laurenti, E., Stratton, M.R., Rahbari, R., Campbell, P.J., Osborne, R.J., Martincorena, I., 2021. Somatic mutation landscapes at single-molecule resolution. Nature 593, 405–410. https://doi.org/10.1038/s41586-021-03477-4

Ambrosino, P., Soldovieri, M.V., Bast, T., Turnpenny, P.D., Uhrig, S., Biskup, S., Döcker, M., Fleck, T., Mosca, I., Manocchio, L., Iraci, N., Taglialatela, M., Lemke, J.R., 2018. De novo gain-of-function variants in KCNT2 as a novel cause of developmental and epileptic encephalopathy. Ann. Neurol. 83, 1198–1204. https://doi.org/10.1002/ana.25248

Amemiya, H.M., Kundaje, A., Boyle, A.P., 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 9, 9354. https://doi.org/10.1038/s41598-019-45839-z

Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., Pletikos, M., Pattni, R., Chen, B.-J., Venturini, E., Riley-Gillis, B., Sestan, N., Urban, A.E., Abyzov, A., Vaccarino, F.M., 2018. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science 359, 550–555. https://doi.org/10.1126/science.aan8690

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., Talbot, R.T., Gustincich, S., Freeman, T.C., Mattick, J.S., Hume, D.A., Heutink, P., Carninci, P., Jeddeloh, J.A., Faulkner, G.J., 2011. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534–537. https://doi.org/10.1038/nature10531

Baldassari, S., Ribierre, T., Marsan, E., Adle-Biassette, H., Ferrand-Sorbets, S., Bulteau, C., Dorison, N., Fohlen, M., Polivka, M., Weckhuysen, S., Dorfmüller, G., Chipaux, M., Baulac, S., 2019. Dissecting the genetic basis of focal cortical dysplasia: a large cohort study. Acta Neuropathol. (Berl.) 138, 885–900. https://doi.org/10.1007/s00401-019-02061-5

Becker, T., Lee, W.-P., Leone, J., Zhu, Q., Zhang, C., Liu, S., Sargent, J., Shanker, K., Mil-homens, A., Cerveira, E., Ryan, M., Cha, J., Navarro, F.C.P., Galeev, T., Gerstein, M., Mills, R.E., Shin, D.-G., Lee, C., Malhotra, A., 2018. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. Genome Biol. 19, 38. https://doi.org/10.1186/s13059-018-1404-6

Bhardwaj, R.D., Curtis, M.A., Spalding, K.L., Buchholz, B.A., Fink, D., Björk-Eriksson, T., Nordborg, C., Gage, F.H., Druid, H., Eriksson, P.S., Frisén, J., 2006. Neocortical neurogenesis in humans is restricted to development. Proc. Natl. Acad. Sci. 103, 12564–12568. https://doi.org/10.1073/pnas.0605177103

Birnbaum, R., Mahjani, B., Loos, R.J.F., Sharp, A.J., 2022. Clinical Characterization of Copy Number Variants Associated With Neurodevelopmental Disorders in a Large-scale Multiancestry Biobank. JAMA Psychiatry 79, 250–259. https://doi.org/10.1001/jamapsychiatry.2021.4080

Bizzotto, S., Dou, Y., Ganz, J., Doan, R.N., Kwon, M., Bohrson, C.L., Kim, S.N., Bae, T., Abyzov, A., NIMH Brain Somatic Mosaicism Network, Park, P.J., Walsh, C.A., 2021. Landmarks of human embryonic development inscribed in somatic mutations. Science 371, 1249–1253. https://doi.org/10.1126/science.abe1544

Blaschke, A.J., Weiner, J.A., Chun, J., 1998. Programmed cell death is a universal feature of embryonic and postnatal neuroproliferative regions throughout the central nervous system. J. Comp. Neurol. 396, 39–50. https://doi.org/10.1002/(sici)1096-9861(19980622)396:1<39::aid-cne4>3.0.co;2-j

Breuss, M.W., Yang, X., Schlachetzki, J.C.M., Antaki, D., Lana, A.J., Xu, X., Chung, C., Chai, G., Stanley, V., Song, Q., Newmeyer, T.F., Nguyen, A., O'Brien, S., Hoeksema, M.A., Cao, B., Nott, A., McEvoy-Venneri, J., Pasillas, M.P., Barton, S.T., Copeland, B.R., Nahas, S., Van Der Kraan, L., Ding, Y., Christopher K. Glass, Gleeson, J.G., 2022. Somatic mosaicism reveals clonal distributions of neocortical development. Nature 604, 689–696. https://doi.org/10.1038/s41586-022-04602-7

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., Kato, M., Kasai, K., Kishimoto, T., Nawa, H., Okano, H., Yoshikawa, T., Kato, T., Iwamoto, K., 2014. Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. Neuron 81, 306–313. https://doi.org/10.1016/j.neuron.2013.10.053

Cai, X., Evrony, G.D., Lehmann, H.S., Elhosary, P.C., Mehta, B.K., Poduri, A., Walsh, C.A., 2014. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep. 8, 1280–1289. https://doi.org/10.1016/j.celrep.2014.07.043

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., McConnell, M.J., 2019. Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell Rep. 26, 825-835.e7. https://doi.org/10.1016/j.celrep.2018.12.107

Coorens, T.H.H., Moore, L., Robinson, P.S., Sanghvi, R., Christopher, J., Hewinson, J., Przybilla, M.J., Lawson, A.R.J., Spencer Chapman, M., Cagan, A., Oliver, T.R.W., Neville, M.D.C., Hooks, Y., Noorani, A., Mitchell, T.J., Fitzgerald, R.C., Campbell, P.J., Martincorena, I., Rahbari, R., Stratton, M.R., 2021. Extensive phylogenies of human development inferred from somatic mutations. Nature 597, 387–392. https://doi.org/10.1038/s41586-021-03790-y

Cortés-Ciriano, I., Lee, J.J.-K., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D., Klimczak, L.J., Zhang, C.-Z., Pellman, D.S., Park, P.J., 2020. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat. Genet. 52, 331–341. https://doi.org/10.1038/s41588-019-0576-7

Costantino, I., Nicodemus, J., Chun, J., 2021. Genomic Mosaicism Formed by Somatic Variation in the Aging and Diseased Brain. Genes 12, 1071. https://doi.org/10.3390/genes12071071

Dentro, S.C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Rubanova, Y., Macintyre, G., Demeulemeester, J., Vázquez-García, I., Kleinheinz, K., Livitz, D.G., Malikic, S., Donmez, N., Sengupta, S., Anur, P., Jolly, C., Cmero, M., Rosebrock, D., Schumacher, S.E., Fan, Y., Fittall, M., Drews, R.M., Yao, X., Watkins, T.B.K., Lee, J., Schlesner, M., Zhu, H., Adams, D.J., McGranahan, N., Swanton, C., Getz, G., Boutros, P.C., Imielinski, M., Beroukhim, R., Sahinalp, S.C., Ji, Y., Peifer, M., Martincorena, I., Markowetz, F., Mustonen, V., Yuan, K., Gerstung, M., Spellman, P.T.,

Wang, W., Morris, Q.D., Wedge, D.C., Loo, P.V., Dentro, S.C., Leshchiner, I., Gerstung, M., Jolly, C., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A.G., Yu, K., Gonzalez, S., Rubanova, Y., Macintyre, G., Demeulemeester, J., Adams, D.J., Anur, P., Beroukhim, R., Boutros, P.C., Bowtell, D.D., Campbell, P.J., Cao, S., Christie, E.L., Cmero, M., Cun, Y., Dawson, K.J., Donmez, N., Drews, R.M., Eils, R., Fan, Y., Fittall, M., Garsed, D.W., Getz, G., Ha, G., Imielinski, M., Jerman, L., Ji, Y., Kleinheinz, K., Lee, J., Lee-Six, H., Livitz, D.G., Malikic, S., Markowetz, F., Martincorena, I., Mitchell, T.J., Mustonen, V., Oesper, L., Peifer, M., Peto, M., Raphael, B.J., Rosebrock, D., Sahinalp, S.C., Salcedo, A., Schlesner, M., Schumacher, S.E., Sengupta, S., Shi, R., Shin, S.J., Stein, L.D., Spiro, O., Vázquez-García, I., Vembu, S., Wheeler, D.A., Yang, T.-P., Yao, X., Yuan, K., Zhu, H., Wang, W., Morris, Q.D., Spellman, P.T., Wedge, D.C., Loo, P.V., 2021. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell 184, 2239-2254.e39. https://doi.org/10.1016/j.cell.2021.03.009

de Pagter, M.S., van Roosmalen, M.J., Baas, A.F., Renkens, I., Duran, K.J., van Binsbergen, E., Tavakoli-Yaraki, M., Hochstenbach, R., van der Veken, L.T., Cuppen, E., Kloosterman, W.P., 2015. Chromothripsis in Healthy Individuals Affects Multiple Protein-Coding Genes and Can Result in Severe Congenital Abnormalities in Offspring. Am. J. Hum. Genet. 96, 651–656. https://doi.org/10.1016/j.ajhg.2015.02.005

D'Gama, A.M., Woodworth, M.B., Hossain, A.A., Bizzotto, S., Hatem, N.E., LaCoursiere, C.M., Najm, I., Ying, Z., Yang, E., Barkovich, A.J., Kwiatkowski, D.J., Vinters, H.V., Madsen, J.R., Mathern, G.W., Blümcke, I., Poduri, A., Walsh, C.A., 2017. Somatic Mutations Activating the mTOR Pathway in Dorsal Telencephalic Progenitors Cause a Continuum of Cortical Dysplasias. Cell Rep. 21, 3754–3766. https://doi.org/10.1016/j.celrep.2017.11.106

Erwin, J.A., Paquola, A.C.M., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I.A., Butcher, C.R., Herdy, J.R., Sarkar, A., Lasken, R.S., Muotri, A.R., Gage, F.H., 2016. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat. Neurosci. 19, 1583–1591. https://doi.org/10.1038/nn.4388

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., Park, P.J., Walsh, C.A., 2012. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. Cell 151, 483–496. https://doi.org/10.1016/j.cell.2012.09.035

Fasching, L., Jang, Y., Tomasi, S., Schreiner, J., Tomasini, L., Brady, M.V., Bae, T., Sarangi, V., Vasmatzis, N., Wang, Y., Szekely, A., Fernandez, T.V., Leckman, J.F., Abyzov, A., Vaccarino, F.M., 2021. Early developmental asymmetries in cell lineage trees in living individuals. Science 371, 1245–1248. https://doi.org/10.1126/science.abe0981

Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., Greenberg, M.E., 2015. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature 522, 89–93. https://doi.org/10.1038/nature14319

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M., Schatz, M.C., 2015. Interactive analysis and assessment of single-cell copy-number variations. Nat. Methods 12, 1058–1060. https://doi.org/10.1038/nmeth.3578

Glover, T.W., Wilson, T.E., 2016. Breaks in the brain. Nature 532, 46–47. https://doi.org/10.1038/nature17316

Glover, T.W., Wilson, T.E., Arlt, M.F., 2017. Fragile sites in cancer: more than meets the eye. Nat. Rev. Cancer 17, 489–501. https://doi.org/10.1038/nrc.2017.52

Gururaj, S., Palmer, E.E., Sheehan, G.D., Kandula, T., Macintosh, R., Ying, K., Morris, P., Tao, J., Dias, K.-R., Zhu, Y., Dinger, M.E., Cowley, M.J., Kirk, E.P., Roscioli, T., Sachdev, R., Duffey, M.E., Bye, A., Bhattacharjee, A., 2017. A De Novo Mutation in the Sodium-Activated Potassium Channel KCNT2 Alters Ion Selectivity and Causes Epileptic Encephalopathy. Cell Rep. 21, 926–933. https://doi.org/10.1016/j.celrep.2017.09.088

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hatch, E.M., Fischer, A.H., Deerinck, T.J., Hetzer, M.W., 2013. Catastrophic Nuclear Envelope Collapse in Cancer Cell Micronuclei. Cell 154, 47–60. https://doi.org/10.1016/j.cell.2013.06.007

Hozumi, N., Tonegawa, S., 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. Proc. Natl. Acad. Sci. U. S. A. 73, 3628–3632.

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., Rivière, J.-B., St-Onge, J., Ojemann, J.G., Shendure, J., Hevner, R.F., Dobyns, W.B., 2015. PI3K/AKT pathway mutations cause a spectrum of brain malformations from megalencephaly to focal cortical dysplasia. Brain 138, 1613–1628. https://doi.org/10.1093/brain/awv045

Jourdon, A., Fasching, L., Scuderi, S., Abyzov, A., Vaccarino, F.M., 2020. The role of somatic mosaicism in brain disease. Curr. Opin. Genet. Dev., Molecular and Genetic Bases of Disease 65, 84–90. https://doi.org/10.1016/j.gde.2020.05.002

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., Chamberlain, S.J., Philpot, B.D., Zylka, M.J., 2013. Topoisomerases facilitate transcription of long genes linked to autism. Nature 501, 58–62. https://doi.org/10.1038/nature12504

Knouse, K.A., Wu, J., Amon, A., 2016. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome Res. 26, 376–384. https://doi.org/10.1101/gr.198937.115

Knouse, K.A., Wu, J., Whittaker, C.A., Amon, A., 2014. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. Proc. Natl. Acad. Sci. 111, 13409–13414. https://doi.org/10.1073/pnas.1415287111

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: An information aesthetic for comparative genomics. Genome Res. 19, 1639–1645. https://doi.org/10.1101/gr.092759.109

Lam, H.Y.K., Mu, X.J., Stütz, A.M., Tanzer, A., Cayting, P.D., Snyder, M., Kim, P.M., Korbel, J.O., Gerstein, M.B., 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat. Biotechnol. 28, 47–55. https://doi.org/10.1038/nbt.1600

Lawson, A.R.J., Abascal, F., Coorens, T.H.H., Hooks, Y., O'Neill, L., Latimer, C., Raine, K., Sanders, M.A., Warren, A.Y., Mahbubani, K.T.A., Bareham, B., Butler, T.M., Harvey, L.M.R., Cagan, A., Menzies, A., Moore, L., Colquhoun, A.J., Turner, W., Thomas, B., Gnanapragasam, V., Williams, N., Rassl, D.M., Vöhringer, H., Zumalave, S., Nangalia, J., Tubío, J.M.C., Gerstung, M., Saeb-Parsy, K., Stratton, M.R., Campbell, P.J., Mitchell, T.J., Martincorena, I., 2020. Extensive heterogeneity in somatic mutation and selection in the human bladder. Science 370, 75–82. https://doi.org/10.1126/science.aba8347

Lee, J.H., Huynh, M., Silhavy, J.L., Kim, S., Dixon-Salazar, T., Heiberg, A., Scott, E., Bafna, V., Hill, K.J., Collazo, A., Funari, V., Russ, C., Gabriel, S.B., Mathern, G.W., Gleeson, J.G., 2012. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat. Genet. 44, 941–945. https://doi.org/10.1038/ng.2329

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., Robinson, P., Coorens, T.H.H., O'Neill, L., Alder, C., Wang, J., Fitzgerald, R.C., Zilbauer, M., Coleman, N., Saeb-Parsy, K., Martincorena, I., Campbell, P.J., Stratton, M.R., 2019. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 574, 532–537. https://doi.org/10.1038/s41586-019-1672-7

Lehman, C.E., Dillon, L.W., Nikiforov, Y.E., Wang, Y.-H., 2017. DNA fragile site breakage as a measure of chemical exposure and predictor of individual susceptibility to form oncogenic rearrangements. Carcinogenesis 38, 293–301. https://doi.org/10.1093/carcin/bgw210

Lim, B., Lin, Y., Navin, N., 2020. Advancing Cancer Research and Medicine with Single-Cell Genomics. Cancer Cell 37, 456–470. https://doi.org/10.1016/j.ccell.2020.03.008

Lim, J.S., Gopalappa, R., Kim, S.H., Ramakrishna, S., Lee, M., Kim, W., Kim, J., Park, S.M., Lee, J., Oh, J.-H., Kim, H.D., Park, C.-H., Lee, J.S., Kim, S., Kim, D.S., Han, J.M., Kang, H.-C., Kim, H. (Henry), Lee, J.H., 2017. Somatic Mutations in TSC1 and TSC2 Cause Focal Cortical Dysplasia. Am. J. Hum. Genet. 100, 454–472. https://doi.org/10.1016/j.ajhg.2017.01.030

Lim, J.S., Kim, W., Kang, H.-C., Kim, S.H., Park, A.H., Park, E.K., Cho, Y.-W., Kim, S., Kim, H.M., Kim, J.A., Kim, J., Rhee, H., Kang, S.-G., Kim, H.D., Kim, D., Kim, D.-S., Lee, J.H., 2015. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. Nat. Med. 21, 395–400. https://doi.org/10.1038/nm.3824

Liu, L., Chen, H., Sun, C., Zhang, J., Wang, Juncheng, Du, M., Li, J., Di, L., Shen, J., Geng, S., Pang, Y., Luo, Y., Wu, C., Fu, Y., Zheng, Z., Wang, Jianbin, Huang, Y., 2022. Low-frequency somatic copy number alterations in normal human lymphocytes revealed by large-scale single-cell whole-genome profiling. Genome Res. 32, 44–54. https://doi.org/10.1101/gr.275453.121

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., Yang, P., Chittenden, T.W., Hatem, N.E., Ryu, S.C., Woodworth, M.B., Park, P.J., Walsh, C.A., 2018. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555–559. https://doi.org/10.1126/science.aao4426

Mao, X., Bruneau, N., Gao, Q., Becq, H., Jia, Z., Xi, H., Shu, L., Wang, H., Szepetowski, P., Aniksztejn, L., 2020. The Epilepsy of Infancy With Migrating Focal Seizures: Identification of de novo Mutations of the KCNT2 Gene That Exert Inhibitory Effects on the Corresponding Heteromeric KNa1.1/KNa1.2 Potassium Channel. Front. Cell. Neurosci. 14.

Martincorena, I., 2019. Somatic mutation and clonal expansions in human tissues. Genome Med. 11, 35. https://doi.org/10.1186/s13073-019-0648-4

Masoodi, T., Siraj, A.K., Siraj, S., Azam, S., Qadri, Z., Parvathareddy, S.K., Al-Sobhi, S.S., AlDawish, M., Alkuraya, F.S., Al-Kuraya, K.S., 2019. Evolution and Impact of Subclonal Mutations in Papillary Thyroid Cancer. Am. J. Hum. Genet. 105, 959–973. https://doi.org/10.1016/j.ajhg.2019.09.026

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., Gage, F.H., 2013. Mosaic Copy Number Variation in Human Neurons. Science 342, 632–637. https://doi.org/10.1126/science.1243472

McConnell, M.J., MacMillan, H.R., Chun, J., 2009. Mathematical modeling supports substantial mouse neural progenitor cell death. Neural Develop. 4, 28. https://doi.org/10.1186/1749-8104-4-28

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., Ganz, J., Jaffe, A.E., Kwan, K.Y., Kwon, M., Lodato, M.A., Mills, R.E., Paquola, A.C.M., Rodin, R.E., Rosenbluh, C., Sestan, N., Sherman, M.A., Shin, J.H., Song, S., Straub, R.E., Thorpe, J., Weinberger, D.R., Urban, A.E., Zhou, B., Gage, F.H., Lehner, T., Senthil, G., Walsh, C.A., Chess, A., Courchesne, E., Gleeson, J.G., Kidd, J.M., Park, P.J., Pevsner, J., Vaccarino, F.M., 2017. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science 356, eaal1641. https://doi.org/10.1126/science.aal1641

Miller, M.B., Huang, A.Y., Kim, J., Zhou, Z., Kirkham, S.L., Maury, E.A., Ziegenfuss, J.S., Reed, H.C., Neil, J.E., Rento, L., Ryu, S.C., Ma, C.C., Luquette, L.J., Ames, H.M., Oakley, D.H., Frosch, M.P., Hyman, B.T., Lodato, M.A., Lee, E.A., Walsh, C.A., 2022. Somatic genomic changes in single Alzheimer's disease neurons. Nature 604, 714–722. https://doi.org/10.1038/s41586-022-04640-1

Miller, M.B., Reed, H.C., Walsh, C.A., 2021. Brain Somatic Mutation in Aging and Alzheimer's Disease. Annu. Rev. Genomics Hum. Genet. 22, 239–256. https://doi.org/10.1146/annurev-genom-121520-081242

Møller, R.S., Weckhuysen, S., Chipaux, M., Marsan, E., Taly, V., Bebin, E.M., Hiatt, S.M., Prokop, J.W., Bowling, K.M., Mei, D., Conti, V., Grange, P. de la, Ferrand-Sorbets, S., Dorfmüller, G., Lambrecq, V., Larsen, L.H.G., Leguern, E., Guerrini, R., Rubboli, G., Cooper, G.M., Baulac, S., 2016. Germline and somatic mutations in the MTOR gene in focal cortical dysplasia and epilepsy. Neurol. Genet. 2. https://doi.org/10.1212/NXG.0000000000000118

Moore, L., Cagan, A., Coorens, T.H.H., Neville, M.D.C., Sanghvi, R., Sanders, M.A., Oliver, T.R.W., Leongamornlert, D., Ellis, P., Noorani, A., Mitchell, T.J., Butler, T.M., Hooks, Y., Warren, A.Y., Jorgensen, M., Dawson, K.J., Menzies, A., O'Neill, L., Latimer, C., Teng, M., van Boxtel, R., Iacobuzio-Donahue, C.A., Martincorena, I., Heer, R., Campbell, P.J., Fitzgerald, R.C., Stratton, M.R., Rahbari, R., 2021. The mutational landscape of human somatic and germline cells. Nature 597, 381–386. https://doi.org/10.1038/s41586-021-03822-7

Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J.V., Gage, F.H., 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903–910. https://doi.org/10.1038/nature03663

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

Mustjoki, S., Young, N.S., 2021. Somatic Mutations in "Benign" Disease. N. Engl. J. Med. 384, 2039–2052. https://doi.org/10.1056/NEJMra2101920

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W.R., Hicks, J.,

Wigler, M., 2011. Tumour evolution inferred by single-cell sequencing. Nature 472, 90–94. https://doi.org/10.1038/nature09807

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., Barry, B.J., Bourgeois, B.F.D., Riviello, J.J., Barkovich, A.J., Black, P.M., Ligon, K.L., Walsh, C.A., 2012. Somatic Activation of AKT3 Causes Hemispheric Developmental Brain Malformations. Neuron 74, 41–48. https://doi.org/10.1016/j.neuron.2012.03.010

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rakic, S., Zecevic, N., 2000. Programmed cell death in the developing human telencephalon. Eur. J. Neurosci. 12, 2721–2734. https://doi.org/10.1046/j.1460-9568.2000.00153.x

Rehen, S.K., Yung, Y.C., McCreight, M.P., Kaushal, D., Yang, A.H., Almeida, B.S.V., Kingsbury, M.A., Cabral, K.M.S., McConnell, M.J., Anliker, B., Fontanoz, M., Chun, J., 2005. Constitutional Aneuploidy in the Normal Human Brain. J. Neurosci. 25, 2176–2180. https://doi.org/10.1523/JNEUROSCI.4560-04.2005

Rodin, R.E., Dou, Y., Kwon, M., Sherman, M.A., D'Gama, A.M., Doan, R.N., Rento, L.M., Girskis, K.M., Bohrson, C.L., Kim, S.N., Nadig, A., Luquette, L.J., Gulhan, D.C., Peter J. Park, Walsh, C.A., 2021. The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. Nat. Neurosci. 24, 176–185. https://doi.org/10.1038/s41593-020-00765-6

Shirley, M.D., Tang, H., Gallione, C.J., Baugher, J.D., Frelin, L.P., Cohen, B., North, P.E., Marchuk, D.A., Comi, A.M., Pevsner, J., 2013. Sturge–Weber Syndrome and Port-Wine Stains Caused by Somatic Mutation in GNAQ. N. Engl. J. Med. 368, 1971–1979. https://doi.org/10.1056/NEJMoa1213507

Shoshani, O., Brunner, S.F., Yaeger, R., Ly, P., Nechemia-Arbely, Y., Kim, D.H., Fang, R., Castillon, G.A., Yu, M., Li, J.S.Z., Sun, Y., Ellisman, M.H., Ren, B., Campbell, P.J., Cleveland, D.W., 2021. Chromothripsis drives the evolution of gene amplification in cancer. Nature 591, 137–141. https://doi.org/10.1038/s41586-020-03064-z

Spencer Chapman, M., Ranzoni, A.M., Myers, B., Williams, N., Coorens, T.H.H., Mitchell, E., Butler, T., Dawson, K.J., Hooks, Y., Moore, L., Nangalia, J., Robinson, P.S., Yoshida, K., Hook, E., Campbell, P.J., Cvejic, A., 2021. Lineage tracing of human development through somatic mutations. Nature 595, 85–90. https://doi.org/10.1038/s41586-021-03548-6

van den Bos, H., Spierings, D.C.J., Taudt, A., Bakker, B., Porubský, D., Falconer, E., Novoa, C., Halsema, N., Kazemier, H.G., Hoekstra-Wakker, K., Guryev, V., den Dunnen, W.F.A., Foijer, F., Colomé-Tatché, M., Boddeke, H.W.G.M., Lansdorp, P.M., 2016. Single-cell whole genome sequencing reveals no evidence for common aneuploidy in normal and Alzheimer's disease neurons. Genome Biol. 17, 116. https://doi.org/10.1186/s13059-016-0976-2

Wang, M., Wei, P.-C., Lim, C.K., Gallina, I.S., Marshall, S., Marchetto, M.C., Alt, F.W., Gage, F.H., 2020. Increased Neural Progenitor Proliferation in a hiPSC Model of Autism Induces Replication Stress-Associated Genome Instability. Cell Stem Cell 26, 221-233.e6. https://doi.org/10.1016/j.stem.2019.12.013

Wang, Y., Bae, T., Thorpe, J., Sherman, M.A., Jones, A.G., Cho, S., Daily, K., Dou, Y., Ganz, J., Galor, A., Lobon, I., Pattni, R., Rosenbluh, C., Tomasi, S., Tomasini, L., Yang, X., Zhou, B., Akbarian, S., Ball, L.L., Bizzotto, S., Emery, S.B., Doan, R., Fasching, L., Jang, Y.,

Juan, D., Lizano, E., Luquette, L.J., Moldovan, J.B., Narurkar, R., Oetjens, M.T., Rodin, R.E., Sekar, S., Shin, J.H., Soriano, E., Straub, R.E., Zhou, W., Chess, A., Gleeson, J.G., Marquès-Bonet, T., Park, P.J., Peters, M.A., Pevsner, J., Walsh, C.A., Weinberger, D.R., Vaccarino, F.M., Moran, J.V., Urban, A.E., Kidd, J.M., Mills, R.E., Abyzov, A., Brain Somatic Mosaicism Network, 2021. Comprehensive identification of somatic nucleotide variants in human brain tissue. Genome Biol. 22, 92. https://doi.org/10.1186/s13059-021-02285-3

Wei, P.-C., Chang, A.N., Kao, J., Du, Z., Meyers, R.M., Alt, F.W., Schwer, B., 2016. Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. Cell 164, 644–655. https://doi.org/10.1016/j.cell.2015.12.039

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., Jaffe, D.B., 2017. Direct determination of diploid genome sequences. Genome Res. 27, 757–767. https://doi.org/10.1101/gr.214874.116

Weissman, I.L., Gage, F.H., 2016. A Mechanism for Somatic Brain Mosaicism. Cell 164, 593–595. https://doi.org/10.1016/j.cell.2016.01.048

Wilson, T.E., Arlt, M.F., Park, S.H., Rajendran, S., Paulsen, M., Ljungman, M., Glover, T.W., 2015. Large transcription units unify copy number variants and common fragile sites arising under replication stress. Genome Res. 25, 189–200. https://doi.org/10.1101/gr.177121.114

Wong, F.K., Marín, O., 2019. Developmental Cell Death in the Cerebral Cortex. Annu. Rev. Cell Dev. Biol. 35, 523–542. https://doi.org/10.1146/annurev-cellbio-100818-125204

Wu, C.-Y., Lau, B.T., Kim, H.S., Sathe, A., Grimes, S.M., Ji, H.P., Zhang, N.R., 2021. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. Nat. Biotechnol. 39, 1259–1269. https://doi.org/10.1038/s41587-021-00911-w

Ye, C.J., Sharpe, Z., Alemara, S., Mackenzie, S., Liu, G., Abdallah, B., Horne, S., Regan, S., Heng, H.H., 2019. Micronuclei and Genome Chaos: Changing the System Inheritance. Genes 10, 366. https://doi.org/10.3390/genes10050366

Yurov, Y.B., Iourov, I.Y., Vorsanova, S.G., Liehr, T., Kolotii, A.D., Kutsev, S.I., Pellestor, F., Beresheva, A.K., Demidova, I.A., Kravets, V.S., Monakhov, V.V., Soloviev, I.V., 2007. Aneuploidy and Confined Chromosomal Mosaicism in the Developing Human Brain. PLOS ONE 2, e558. https://doi.org/10.1371/journal.pone.0000558

Zaccaria, S., Raphael, B.J., 2021. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nat. Biotechnol. 39, 207–214. https://doi.org/10.1038/s41587-020-0661-6

Zhang, C.-Z., Spektor, A., Cornils, H., Francis, J.M., Jackson, E.K., Liu, S., Meyerson, M., Pellman, D., 2015. Chromothripsis from DNA damage in micronuclei. Nature 522, 179–184. https://doi.org/10.1038/nature14493

Zhu, X., Zhou, B., Pattni, R., Gleason, K., Tan, C., Kalinowski, A., Sloan, S., Fiston-Lavier, A.-S., Mariani, J., Petrov, D., Barres, B.A., Duncan, L., Abyzov, A., Vogel, H., Moran, J.V., Vaccarino, F.M., Tamminga, C.A., Levinson, D.F., Urban, A.E., 2021. Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia. Nat. Neurosci. 24, 186–196. https://doi.org/10.1038/s41593-020-00767-4

Chapter 3 Synthetic Assessment of Single-cell CNV Detection Tools

3.1 Motivation

Single-cell DNA sequencing (scDNA-seq) is a powerful technique to study genetic variations at the cellular level, such as copy number variations (CNVs), which are not rare in somatic cells (Cai et al., 2014; D'Gama and Walsh, 2018; McConnell et al., 2013; Perez-Rodriguez et al., 2023). However, scDNA-seq data poses several challenges for CNV analysis, such as low depth of sequencing coverage, high noise from sequencing errors and library preparation, and potential bias from whole genome amplification techniques. Therefore, more accurate and robust methods for CNV detection and quantification from scDNA-seq data are needed.

Currently, researchers have developed several state-of-the-art single-cell CNV detection tools. To evaluate and compare different single-cell CNV callers, it is essential to have a reliable benchmark dataset with known ground truth. However, such a dataset is difficult to obtain experimentally due to the fact that it is very difficult to replicate the exact same genomic variations in different cells by biological means. An alternative method would thus be to simulate sequence data with known single-cell parameters. However, most methods in this area have been designed specifically on cancer cells (Feng et al., 2021; Feng and Chen, 2022; Giguere et al., 2020; Posada, 2020), which have high levels of genomic instability and complex subclonal structures that are not present in other tissues like the brain, which have lower levels of genomic

variation and more homogeneous cell populations. In addition, other single-cell simulators such as SCSsim cannot support CNVs (Yu et al., 2020).

In addition to differences in approaches due to genomic features, there are other limitations that current single-cell simulators exhibit. Fully synthetic data simulators generate reads from scratch based on predefined models and parameters, such as mutation rates and sequencing errors. Sampling-based methods, on the other hand, use existing data as a reference and sample reads from it according to some criteria, such as the read count in a genomic region. Fully synthetic data simulators face challenges in modeling complex biological phenomena accurately and realistically. They may introduce biases or artifacts that do not reflect the true characteristics of single-cell data. Moreover, fully synthetic data simulators require extensive validation and calibration using real data to ensure their reliability and applicability. Sampling-based methods can overcome some of these limitations by leveraging existing data as a source of information and variation. They can capture more realistic features of single-cell data, such as technical noise patterns. They can also benefit from the availability and diversity of single-cell datasets in different domains and contexts.

In this chapter, we present a novel sampling-based single-cell CNV simulator that can generate realistic synthetic data for assessing the performance of single-cell CNV callers. Our simulator has two main components: (1) a CNV profile generator that simulates the location, size and copy number of CNVs for each cell based on empirical distributions derived from real scDNA-seq data; (2) a read simulator that creates BAM files for each cell by introducing CNVs according to predefined parameters. Our simulator also outputs the ground truth of CNV segments and breakpoints for each cell.

We demonstrate the utility of our simulator by generating synthetic scDNA-seq data for brain tissue and comparing several state-of-the-art single-cell CNV callers on it. We show that our simulator can produce realistic data that mimics the characteristics of real scDNA-seq data in terms of coverage distribution, allelic ratio and noise level. We also show that our simulator can reveal the strengths and weaknesses of different single-cell CNV callers and help identify areas for improvement.

3.2 Data and methods

3.2.1 Simulation framework

Our simulator is designed for the benchmark of CNV detection tools using scDNA-seq data (Figure 3.1). The inputs consist of single-cell bam files of the non-CNV cells and the CNV profile, including CNV size, location and cell. The outputs are the simulated single-cell bam files with CNV. From the 2,097 cells we sequenced from the human neurons dissected from the dorsolateral prefrontal cortex (DLPFC) of a neurotypical individual described in Chapter 2, we exclude the 475 cells with CNVs called by Ginkgo (Garvin et al., 2015). Within the 1,628 non-CNV cells, we randomly select 1,000 cells as the base samples for the simulator. We also randomly determine the CNV cell, genomic location and CNV size based on the distribution of the final call set described in Chapter 2.

Since we already have the phased heterozygous SNPs (het-SNPs) for this sample, we first bin the whole genome into 100 phased het-SNP windows. If users need to apply our simulation method on the samples without phased het-SNPs, we develop a pseudo-bulk SNP calling and phasing pipeline. Users can merge all the single-cell sequences together and use GATK toolkit (McKenna et al., 2010) to call the germline SNPs. Then users can leverage the Michigan

imputation server (Das et al., 2016) or other similar imputation and phasing tools to impute and phase SNPs. We applied this pipeline to the common control sample we described in Chapter 2 and compared phased het-SNPs with het-SNPs called from 10X linked reads. More than 97% of het-SNPs generated from this pipeline overlapped with those from 10X linked reads and shared the same phasing information, which shows our method is reliable. (Table 3.1)

Next, we build het-SNP window-based statistical distributions for the normalized total read count, informative read count (reads covered by the het-SNP), and allelic ratio from the highly conservative CNV call set described in Chapter 2. We then count the total read count and informative read count in the original samples and generate simulated read counts for each window. Based on the differences in read counts, we add or delete reads to match the read counts to the simulated counts and generate simulated bam files.

3.2.2 CNV profile simulation

In the 1,000 randomly selected non-CNV cells, we randomly select 300 cells as the simulated CNV cells to achieve the frequency of somatic mosaicism at 30%. We designed 4 simulated CNV sets with 0%, 10%, 20% and 30% subclonal frequency respectively, which means 0, 100, 200, and 300 cells share the same CNV set. Then we build the distribution for the size of CNVs in the final call set described in Chapter 2 and fit it with the exponential distribution using a fixed location parameter as 1M, as all the CNV calls are larger than 1Mb (Figure 3.2). Next, for each CNV, we randomly select a cell, a chromosome number and a genomic location without overlapping with other CNVs. Some CNVs would fall into a single het-SNP window, which is the basic unit for the simulator. Therefore, these CNVs would be filtered out.

3.2.3 Derive statistical distributions from the highly conservative CNV call set

We build the statistical distributions for the normalized total read counts, normalized informative read counts and log2 ratio of informative read count between two haplotypes from all the het-SNP windows. The read count normalization is as follow:

$$normalized\ count = \frac{10^{10} * raw\ read\ count}{read\ count\ of\ this\ cell * window\ size}$$

For heterozygous deletions and homozygous deletions, we build the log normal distributions from the CNV regions in the final call set directly. For CN-LOH, we can build the distributions based on the non-CNV regions as CN-LOHs have the same copy number with non-CNV regions. For duplications, we have to infer the location parameter of log normal distribution based on the difference between means of heterozygous deletions and non-CNV regions. We add this difference to the location parameter of the non-CNV distributions. The other two parameters would be the same with the non-CNV distributions.

In order to reduce the noise of the simulation data, we also set the cutoffs for the original distributions before fitting to the log normal distributions. For non-CNV regions, we filter out the windows with the normalized total reads count higher than top 5% quantile or lower than bottom 5% quantile, and also filter out the windows with informative reads count lower than 10 % quantile (Figure 3.3 A, B). For heterozygous deletions, we filter out the windows with the normalized total reads count higher than top 5% quantile or lower than bottom 5% quantile, and also filter out the windows with absolute log2 ratio lower than 5 and windows with informative reads count lower than 15% quantile (Figure 3.3 C, D). For homozygous deletions, we filter out

the windows with the normalized total reads count higher than bottom 5% quantile of distribution for heterozygous deletions, and also filter out the windows with informative reads count higher than the bottom 15% quantile of distribution for heterozygous deletions (Figure 3.3 G, H). Next, we fit these filtered data with log normal distributions, and fit the absolute log2 ratio of heterozygous deletions with an empirical distribution (Figure 3.3 E, F).

3.2.4 Count the current reads and generate simulated read counts to manipulate the bam files

We count the reads for each het-SNP window and each cell using SAMtools (Li et al., 2009). We also count the informative reads for each het-SNP window and each cell on the two haplotypes using 4th and 5th steps of the SCOVAL pipeline (Sun, 2023) described in the Chapter 2. Next, we sample the read counts and informative read counts from the pre-built log normal distributions for each window as the target read counts, and we also sample the absolute log2 ratio from the pre-built empirical distribution for the heterozygous deletion windows. We first transform the sampled normalized count into the real count using the formula:

$$read\ count = \frac{normalized\ count * read\ count\ of\ this\ cell * window\ size}{10^{10}}$$

Second, We assign the informative reads count to each haplotype based on the different copy numbers: 1) for the homozygous deletion window, we take the half of the informative reads count for each haplotype; 2) for the heterozygous deletion window, we calculate the informative reads count based on the simulated absolute log2 ratio; 3) for the CN-LOH window, one haplotype takes all the informative reads and the other one takes 0; 4) for the duplication window, we divide the informative reads count by the simulated copy number for one haplotype, and subtract this number from total informative reads count for the other haplotype.

Then for each window and each haplotype, we compare the target reads count and current reads count. If the current count is lower, we copy the reads from the other cells in the same window and on the same haplotype to this cell. If the target count is lower, we randomly delete the reads for the particular haplotype. Similarly, we also compare, and add or delete the non-informative reads based on the difference between target and current reads count.

3.2.5 Benchmark other single-cell CNV tools using simulation data

Over the decades, several scDNA-seq CNV callers have been proposed. There are two types: single cell-based tools, such as Ginkgo (Garvin et al., 2015) and SCOVAL (Sun et al., 2023), and multiple cells based tools, such as CHISEL (Zaccaria and Raphael, 2021), SCYN (Feng et al., 2021) and Alleloscope (Wu et al., 2021). Here we use our simulation data to benchmark single-cell CNV callers, and evaluate 3 state-of-the-art CNV callers: Ginkgo, CHISEL and SCYN. Ginkgo is a web platform for the automated and interactive analysis of single-cell CNVs, and it can also be used as a local application. It applies a variable bin strategy to segment the genome into bins followed by corrections of GC contents and other amplification artifacts. Ginkgo employs Circular Binary Segmentation (CBS) algorithm for genome segmentation and subsequent inference of integer absolute copy number state. CHISEL is the first method for allele-specific copy number estimation with scDNA-seq data. It can detect allele-specific CNVs at single-cell resolution, but it requires high sequencing depth and external phasing haplotypes. SCYN is a CNV segment method and it is based on another single-cell CNV calling method, SCOPE (Wang et al., 2020). SCOPE detected CNV by a Poisson latent factor model. SCYN adopts SCOPE's normalization method and uses dynamic programming to conduct CNV segmentation.

We evaluate the performance of these 3 tools for detecting single-cell CNVs using simulation data with varying subclonal frequencies. We compare their CNV call set with the ground truth CNV set we generated. We use two levels of comparison: CNV level and window level. At the CNV level, we consider a CNV call to be true positive if it overlaps with at least 80% of any ground truth CNV. At the window level, we assign the called copy number to each genomic window and compare it with the corresponding ground truth. We then treat this as a multiclass classification problem and use metrics such as precision, recall and F1-score for each copy number class to assess the quality of the CNV calls. The formulas are:

$$Precion = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = \frac{2 * Precison * Recall}{Precison + Recall}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TP}{TN + FP}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Here we use the one-vs-rest method to calculate TP, FP, FN and TN for each class separately.

3.3 Results

3.3.1 Simulation result

We performed four simulations of CNV call sets with different subclone frequencies: 0%, 10%, 20%, and 30%. The number of simulated CNVs for each frequency was: (1) 1,174 CNVs for 0%; (2) 1,200 CNVs for 10%; (3) 1,196 CNVs for 20%; and (4) 1,200 CNVs for 30%. (Table 3.2) To evaluate the quality of the simulated CNVs, we compared them with the very conservative CNV call set obtained in Chapter 2. We computed the number of informative read counts, the median absolute log2 ratio, and the median read depth ratio for both the simulated and the conservative CNVs. (Figure 3.4) We also visualized the distribution of the absolute log2 ratio and the read depth ratio for each simulation. (Figure 3.5) These analyses demonstrate that our simulated CNVs are realistic and representative of the data.

3.3.2 Evaluation of current single-cell CNV tools

We conducted a comparative analysis of Ginkgo, CHISEL and SCYN on four simulated CNV datasets with known ground truth. These datasets had different subclone frequencies ranging from 0% to 30% (Table 3.3). In general, Ginkgo typically calls less CNVs and SCYN calls more CNVs than the ground truth. In the 30% dataset, however, neither Ginkgo and SCYN could recall the CNVs from the simulation set. In addition, CHISEL calls an extremely large number of CNVs, which is not reliable. These results suggest that Ginkgo is more accurate and robust than CHISEL and SCYN in detecting single-cell CNVs in the lower subclone frequencies, with none performing well when the subclone frequency is high.

We assessed the performance of each tool at both the CNV and window levels using precision, recall and F1-score metrics (Figure 3.4). Ginkgo achieved the highest performance across all four subclone frequency scenarios at the CNV level, with an F1-score ranging from 0.8 (no subclones) to 0.7 (20% subclones). However, none of the tools could detect CNVs reliably

when the subclone frequency is 30%, resulting in near-zero F1-scores. At the window level, Ginkgo maintained its superiority for subclone frequencies of 0%, 10% and 20%, while CHISEL surpassed the other tools for subclone frequency of 30% (Figure 3.5). The confusion matrices are shown in Figure 3.6.

3.4 Conclusion and Discussion

In this chapter, I have presented a novel single-cell CNV simulator that can generate realistic and diverse datasets of single-cell DNA copy number profiles. The simulator can capture the complex and heterogeneous nature of single-cell DNA copy number alterations. Unlike previous simulators that either focused on cancer samples or could not simulate CNVs, the simulator can produce datasets with different levels of complexity and heterogeneity to challenge existing single-cell CNV tools and to facilitate the development of new methods. The simulator is a powerful tool for benchmarking and evaluating the performance of different single-cell CNV analysis methods. Researchers can use the simulator to generate synthetic datasets with known ground truth CNV profiles, enabling them to compare and assess the accuracy, sensitivity, and specificity of different methods. It can also aid in the development of new single-cell CNV analysis methods. Researchers can use the simulated datasets to test and optimize new algorithms, potentially leading to more accurate and efficient methods for identifying CNVs in single cells.

I have also benchmarked three current single-cell CNV tools, Ginkgo, CHISEL and SCYN, using simulation data to evaluate their performance and limitations. These tools represent different approaches to infer single-cell DNA copy number profiles from scDNA-seq data. I have compared their performance using simulated data sets with varying subclone frequencies and

evaluated them using five metrics: precision (the proportion of true positives among all predicted positives), recall or sensitivity (the proportion of true positives among all real positives), specificity (the proportion of true negatives among all real negatives), and F1-score (the harmonic mean of precision and recall). My results show that Ginkgo has the highest precision, recall and F1-score among the three tools when the subclone frequency is low. SCYN has a lower performance than Ginkgo in all metrics. These findings suggest that Ginkgo is more sensitive and reliable in detecting CNVs from scDNA-seq data, especially for low-frequency subclones, while CHISEL may perform better when the subclone frequency is higher. Therefore, depending on the research question and the quality of the data, different tools may be more suitable for single-cell CNV identification.

There are some limitations and challenges that need to be addressed in future work. First, the simulation dataset relies on the single-cell sequencing data from 1,000 brain cells from a single individual to generate CNV profiles, which may introduce biases or errors if our data are not representative. Though we provide the methods to generate simulation data from other datasets, a better solution is to develop a self-learning algorithm that can adaptively adjust the parameters of the simulator based on the input data. Second, the simulator does not account for other types of somatic variations in scDNA-seq data, such as SNVs and SVs. The simulator may reflect more real simulation when incorporating them into the simulation model. Third, the simulator has not been validated using real single-cell CNV data from different organisms or tissues. Such validation is necessary to assess the generalizability and applicability of the simulator in different biological contexts. A possible solution is to collect more single-cell CNV data from various sources and compare them with the simulated data using appropriate metrics. We can leverage some ongoing efforts such as the Somatic Mosaicism Across Human Tissues

(SMaHT) network (https://commonfund.nih.gov/smaht) at NIH to implement this. The SMaHT Network aims to understand how somatic mosaicism influences biology and disease by cataloging the extent of somatic mosaicism in different cell types, disease states, and life stages. They plan to achieve this by systematically documenting DNA sequence variants within personal genomes using state-of-the art sequencing technologies and spurring technological development that will enable researchers to detect different types of variation.

One of the main challenges in studying single-cell CNVs is the lack of reliable and realistic simulation data that can capture the complexity and diversity of genomic aberrations in single cells. Our simulator would enable researchers to evaluate single-cell CNV calling tools under various scenarios and settings, and to identify the optimal parameters and strategies for detecting CNVs from single-cell sequencing data. In Chapter 4, we introduce the development of a deep learning based single-cell CNV calling tool, which uses our simulation data set as the training data.

Another future direction for research in this area is to develop a new single-cell CNV simulator for long-reads single-cell data. Current methods for single-cell CNV calling are mostly designed for short-reads data, which have limited resolution and specificity for detecting complex and subtle CNVs. Long-reads data offer the potential to overcome these limitations by providing more information about breakpoints, copy number states, allele frequencies, and haplotype phasing. However, long-reads data also pose new challenges such as higher error rates, lower coverage, and more difficult to extract the feature distributions. Therefore, new algorithms and models are needed to address these challenges and to exploit the full potential of long-reads data for single-cell CNV analysis. Such tools would enable researchers to uncover novel insights

into the mechanisms and consequences of somatic CNVs at unprecedented resolution and accuracy.

In addition, our single-cell CNV simulator can be generalized to the data generated from other whole genome amplification (WGA) methods or sequencing technologies. The method is not limited to the specific WGA method or sequencing platform used in our study, as it can accommodate different sources of noise and biases that are inherent in different technologies. This feature makes the simulator a flexible and versatile tool that can generate simulated datasets that are tailored to specific experimental conditions and requirements by using the base datasets from other platforms. For example, our simulator can be used to generate datasets that mimic other types of WGA or sequencing technologies, such as MALBAC and MDA, to compare and benchmark the performance of different methods under different experimental conditions. By doing so, the simulator can provide insights into the strengths and limitations of different technologies for single-cell CNV analysis, and guide the development of more robust and accurate methods for detecting CNVs in single cells.

Figures

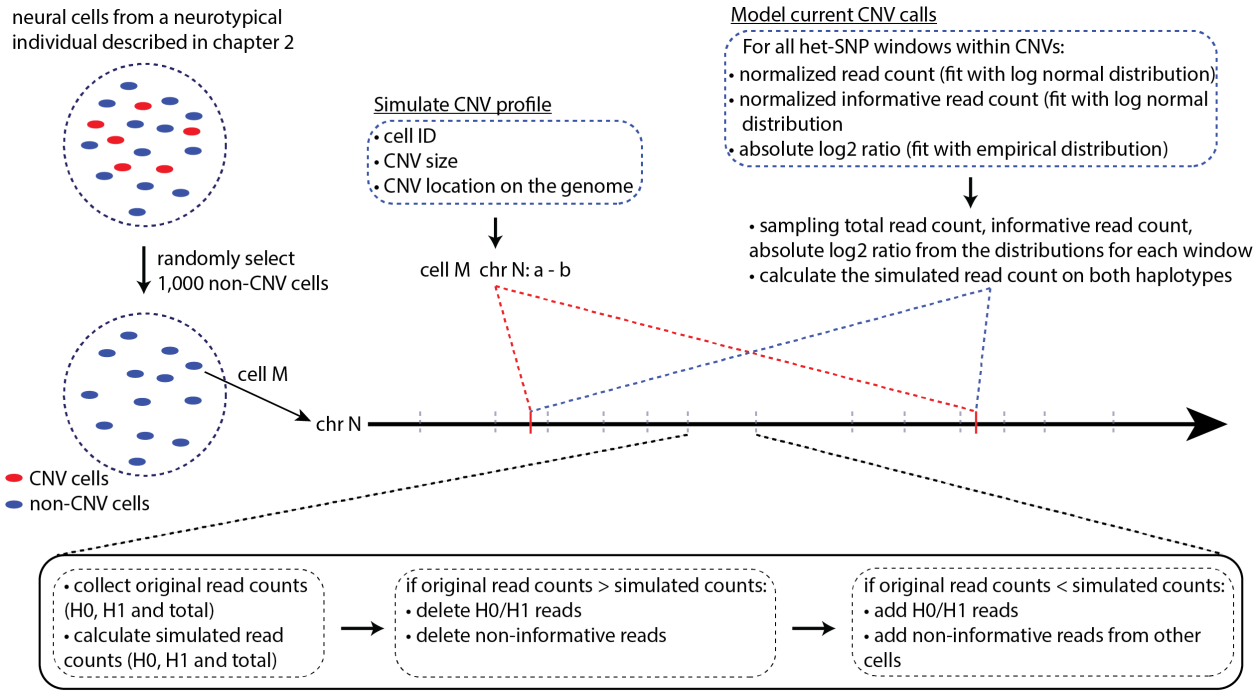


Figure 3.1 Overview of the simulator workflow.

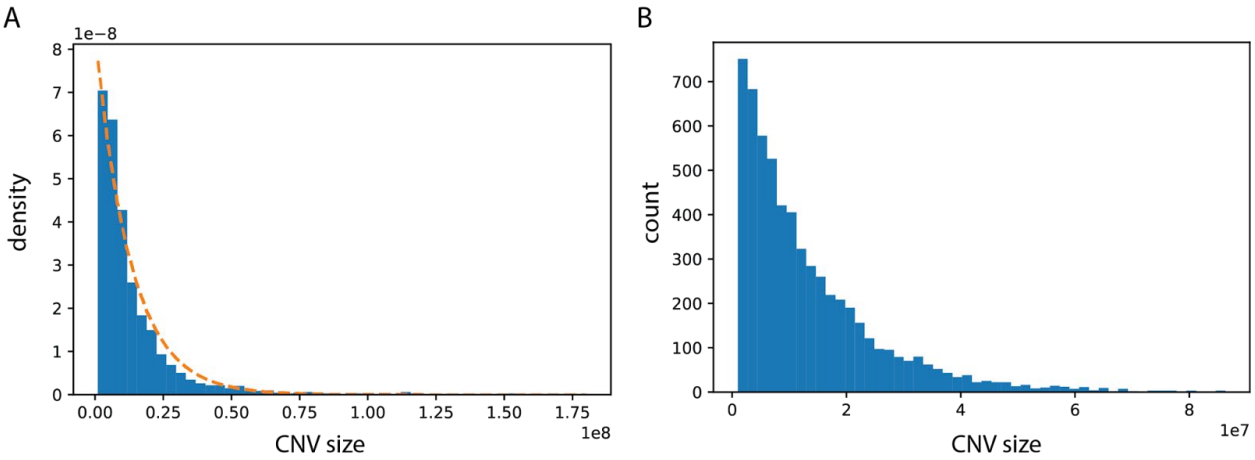


Figure 3.2 CNV size distribution.

(A) fit validated call set CNV size with an exponential distribution (orange dotted line). (B) sample 6,000 CNV sizes from the fitted exponential distribution.

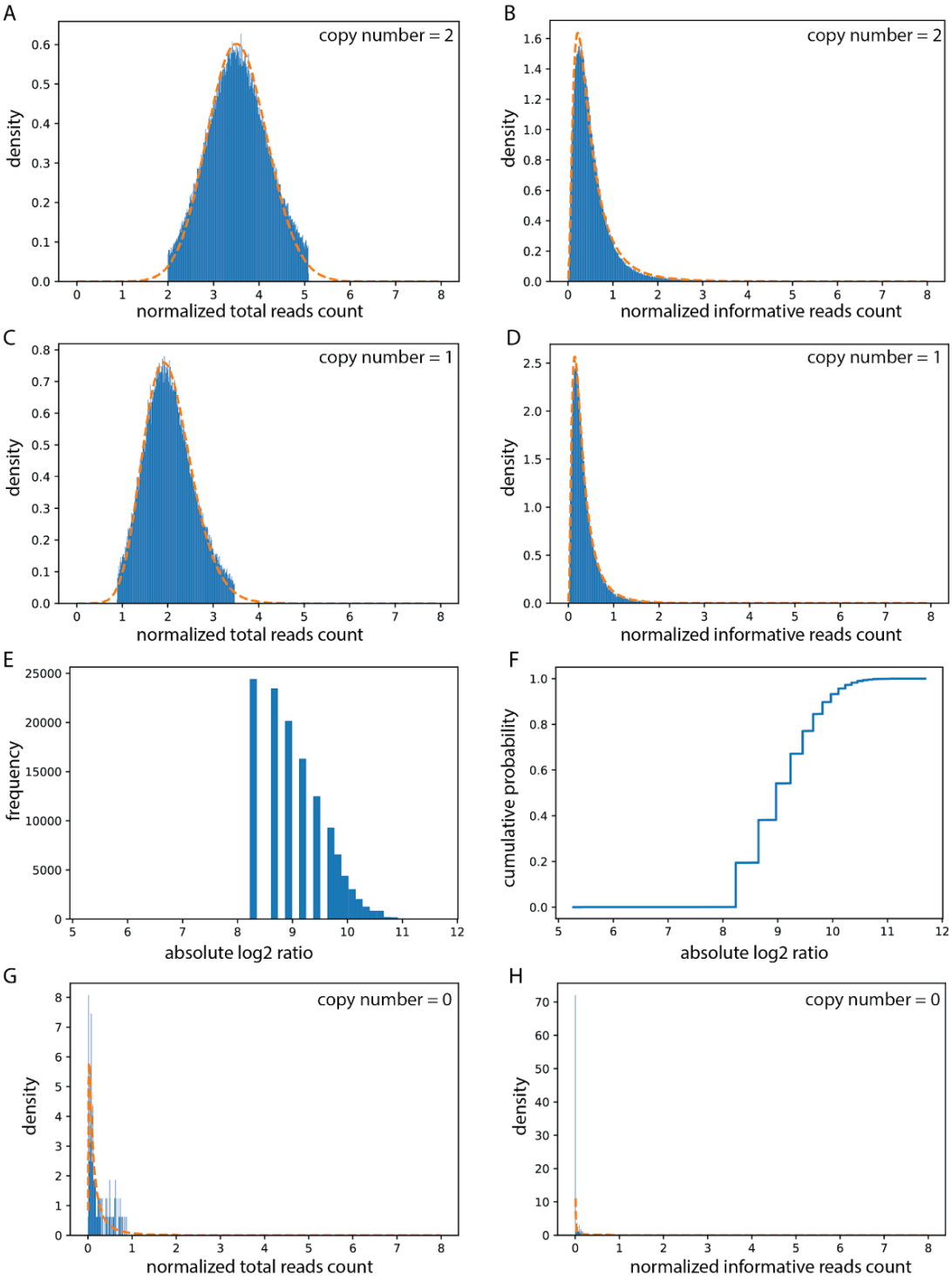


Figure 3.3 Distributions of normalized total reads count and informative reads count for different types of CNVs and distribution of absolute log2 ratio for heterozygous deletions.

(A-D, G, H) normalized total and informative reads count for the windows with copy number 0, 1 and 2. (E, F) filtered absolute log2 ratio distribution from heterozygous deletions. Dashed lines are the fitted log normal distribution.

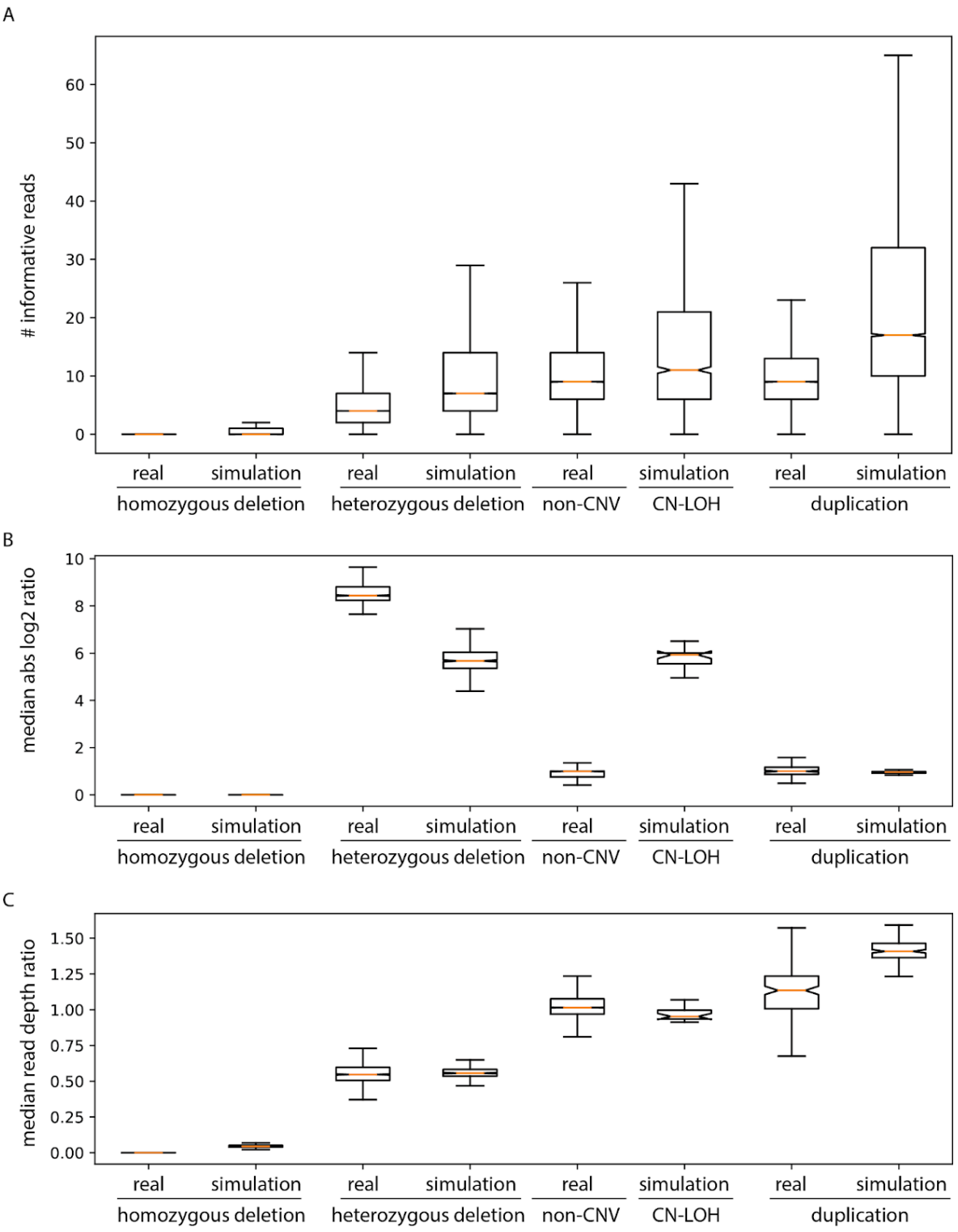


Figure 3.4 Comparison of informative reads count, median absolute ratio and median read depth ratio for the real and simulation CNVs.

The real homozygous and heterozygous deletions are from the conservative calls described in Chapter 2. The real duplications are from Ginkgo duplication calls described in Chapter 2. The simulation CNVs are from the 0% subclone frequency simulation set.

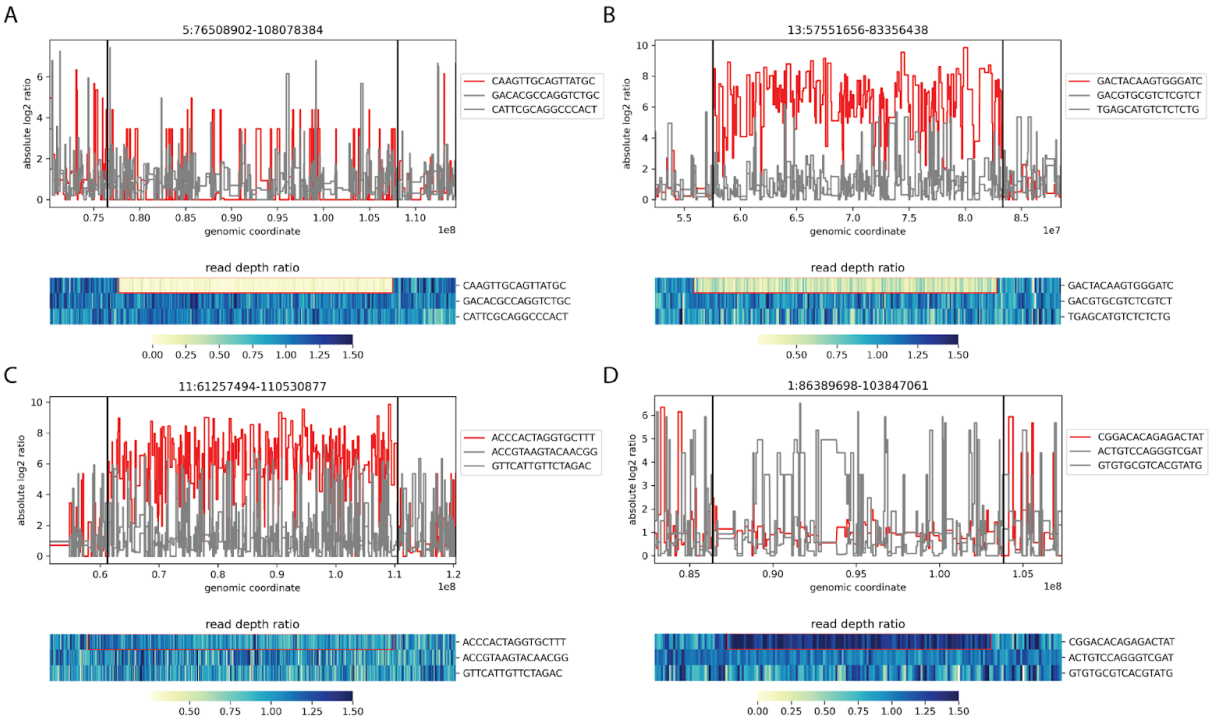


Figure 3.5 Examples of the simulated 4 types of CNVs.

(A) homozygous deletion. (B) heterozygous deletion. (C) CN-LOH. (D) duplication.

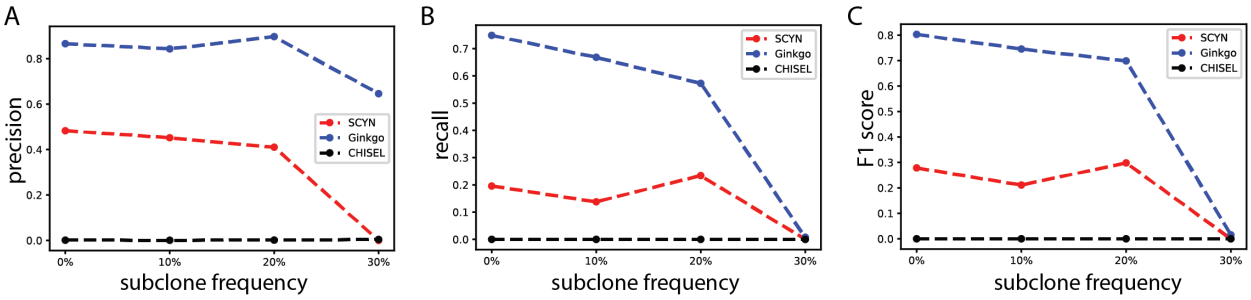


Figure 3.6 CNV level performance of 3 tools on the different subclone frequencies.

(A) precision. (B) recall. (C) F1 score.

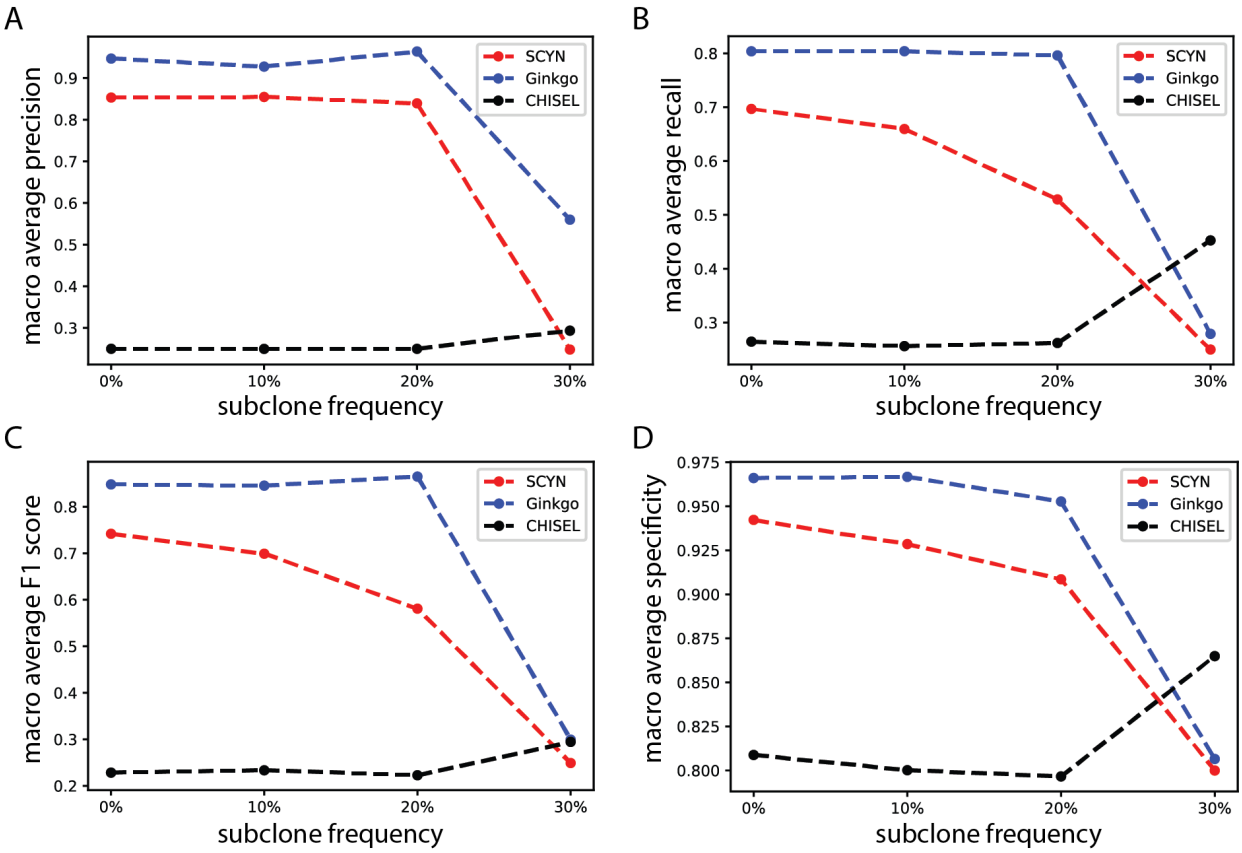


Figure 3.7 Window level performance comparison of 3 tools on the different subclone frequencies.

(A) macro average precision of all the classes. (B) macro average recall of all the classes. (C) macro average F1 score of all the classes. (D) macro average specificity of all the classes.

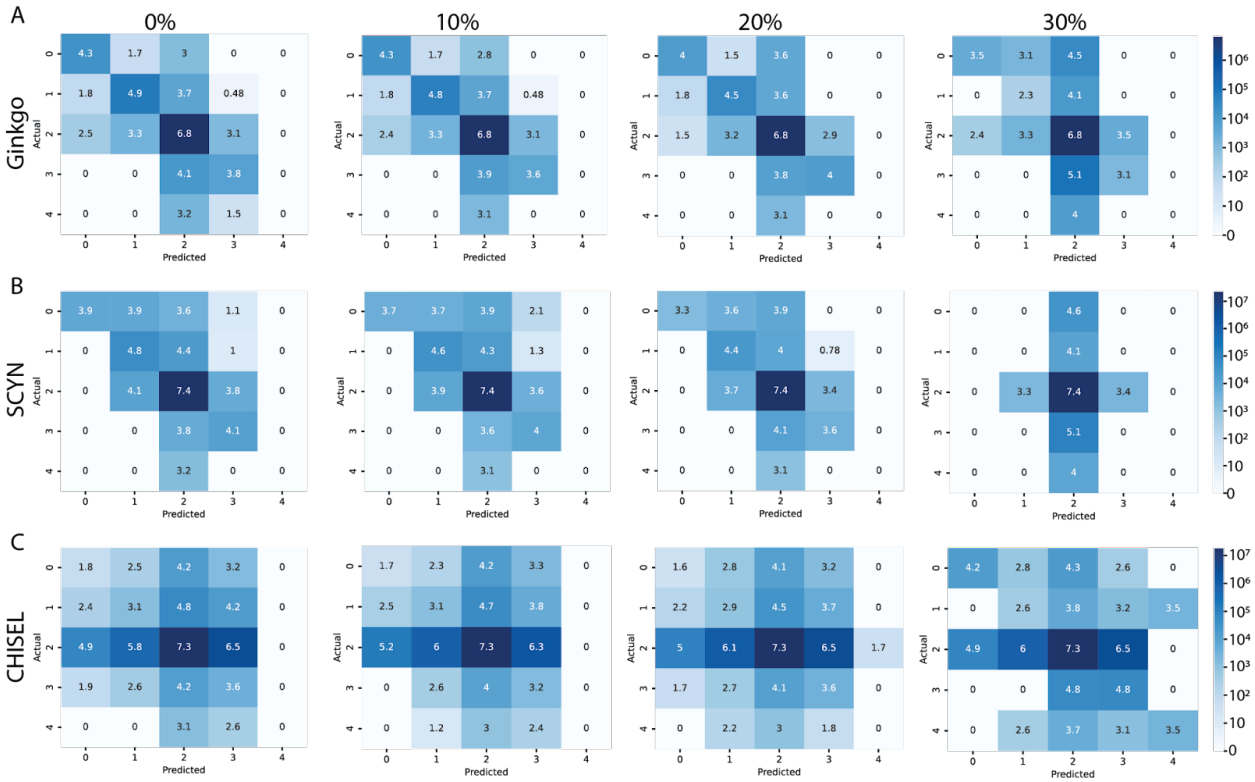


Figure 3.8 Heatmaps of the confusion matrices for Ginkgo, SCYN and CHISEL copy number estimations on the 4 simulation data sets with different subclone frequencies.

The labels 0, 1, 2, 3, and 4 represent the homozygous deletion, heterozygous deletion, normal, duplication and CN-LOH.

Tables

Table 3.1 Comparison of phased heterozygous SNPs called from 10X linked reads and our pipeline.

Phased het-SNPs from 10X linked reads	2,299,568
Phased het SNPs after imputation	1,951,416
Overlapped SNPs with phased 10X SNP calls	1,900,860 (97.41%)
Overlapped SNPs with GATK SNP calls	1,925,987 (98.70%)

Table 3.2 Four simulation CNV set with different subclone frequencies.

Subclone frequency	0%	10%	20%	30%
# CNV	1,174	1,200	1,196	1,200
# CNV cell	300	300	300	300
# Homo-DEL(CN=0)	196	224	258	300
# Homo-DEL cell (CN = 0)	134	183	238	300
# Het-DEL(CN=1)	785	848	663	300
# Het-DEL cell (CN = 1)	279	284	288	300
# CN-LOH(CN=2)	20	17	17	300
# CN-LOH cell (CN = 2)	18	15	15	300
# DUP (CN > 2)	173	111	258	300
# DUP cell (CN > 2)	123	79	239	300

Table 3.3 Comparison of 3 tools for the CNV detection on the different subclone frequency.

Subclone frequency	0%	10%	20%	30%
# simulated CNV	1,174	1,200	1,196	1,200
# simulated CNV window	122,824	94,799	68,098	184,200
# Ginkgo CNV	961	923	931	144
# Ginkgo CNV window	105,691	82,824	54,656	11,462
# SCYN CNV	3,299	2,477	2,081	829
# SCYN CNV window	105,938	74,288	44,948	4,990
# CHISEL CNV	142,211	164,814	153,456	150,451
# CHISEL CNV window	3,909,743	2,994,778	4,642,915	3,981,416

Bibliography

Cai, X., Evrony, G.D., Lehmann, H.S., Elhosary, P.C., Mehta, B.K., Poduri, A., Walsh, C.A., 2014. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep. 8, 1280–1289. https://doi.org/10.1016/j.celrep.2014.07.043

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W.G., Swaroop, A., Scott, L.J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G.R., Fuchsberger, C., 2016. Next-generation genotype imputation service and methods. Nat. Genet. 48, 1284–1287. https://doi.org/10.1038/ng.3656

D'Gama, A.M., Walsh, C.A., 2018. Somatic mosaicism and neurodevelopmental disease. Nat. Neurosci. 21, 1504–1514. https://doi.org/10.1038/s41593-018-0257-3

Feng, X., Chen, L., 2022. SCSilicon: a tool for synthetic single-cell DNA sequencing data generation. BMC Genomics 23, 359. https://doi.org/10.1186/s12864-022-08566-w

Feng, X., Chen, L., Qing, Y., Li, R., Li, C., Li, S.C., 2021. SCYN: single cell CNV profiling method using dynamic programming. BMC Genomics 22, 651. https://doi.org/10.1186/s12864-021-07941-3

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M., Schatz, M.C., 2015. Interactive analysis and assessment of single-cell copy-number variations. Nat. Methods 12, 1058–1060. https://doi.org/10.1038/nmeth.3578

Giguere, C., Dubey, H.V., Sarsani, V.K., Saddiki, H., He, S., Flaherty, P., 2020. SCSIM: Jointly simulating correlated single-cell and bulk next-generation DNA sequencing data. BMC Bioinformatics 21, 215. https://doi.org/10.1186/s12859-020-03550-1

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

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., Gage, F.H., 2013. Mosaic Copy Number Variation in Human Neurons. Science 342, 632–637. https://doi.org/10.1126/science.1243472

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110

Perez-Rodriguez, D., Kalyva, M., Santucci, C., Proukakis, C., 2023. Somatic CNV Detection by Single-Cell Whole-Genome Sequencing in Postmortem Human Brain, in: Chun, J. (Ed.), Alzheimer's Disease: Methods and Protocols, Methods in Molecular Biology. Springer US, New York, NY, pp. 205–230. https://doi.org/10.1007/978-1-0716-2655-9_11

Posada, D., 2020. CellCoal: Coalescent Simulation of Single-Cell Sequencing Samples. Mol. Biol. Evol. 37, 1535–1542. https://doi.org/10.1093/molbev/msaa025

Sun, C., 2023. mills-lab/Scoval [WWW Document]. Scoval Pipeline. URL https://github.com/mills-lab/Scoval (accessed 3.7.23).

Sun, C., Kathuria, K., Emery, S.B., Kim, B., Burbulis, I.E., Shin, J.H., Network, B.S.M., Weinberger, D.R., Moran, J.V., Kidd, J.M., Mills, R.E., McConnell, M.J., 2023. Mapping

the Complex Genetic Landscape of Human Neurons. https://doi.org/10.1101/2023.03.07.531594
Wang, R., Lin, D.-Y., Jiang, Y., 2020. SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. Cell Syst. 10, 445-452.e6. https://doi.org/10.1016/j.cels.2020.03.005
Wu, C.-Y., Lau, B.T., Kim, H.S., Sathe, A., Grimes, S.M., Ji, H.P., Zhang, N.R., 2021. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. Nat. Biotechnol. 39, 1259–1269. https://doi.org/10.1038/s41587-021-00911-w
Yu, Z., Du, F., Sun, X., Li, A., 2020. SCSsim: an integrated tool for simulating single-cell genome sequencing data. Bioinformatics 36, 1281–1282. https://doi.org/10.1093/bioinformatics/btz713
Zaccaria, S., Raphael, B.J., 2021. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nat. Biotechnol. 39, 207–214. https://doi.org/10.1038/s41587-020-0661-6

Chapter 4 Development of a Neural Network-based Single-cell CNV Caller

4.1 Background

Copy number variations (CNVs) are genomic alterations that result in gains or losses of DNA segments (Zarrei et al., 2015). CNVs can affect gene expression, function, and interactions, and are associated with various diseases, such as cancer and neurodevelopmental disorders (Grayton et al., 2012; Henrichsen et al., 2009; Shlien and Malkin, 2009). Detecting CNVs at single-cell resolution can reveal the heterogeneity and dynamics of genomic instability within a cell population, which is crucial for understanding the mechanisms and consequences of somatic mosaicism (Mallory et al., 2020; Ning et al., 2014).

Deep learning is a subfield of machine learning that employs artificial neural networks to learn from large and complex data. Deep learning has been applied to various problems in genomics research, such as gene expression analysis, variant calling, regulatory motif discovery, and disease diagnosis (Eraslan et al., 2019; Zou et al., 2019). Some examples of deep learning applications in genomics are: DeepSEA: a deep convolutional neural network that predicts the chromatin effects of sequence alterations with single-nucleotide sensitivity (Zhou and Troyanskaya, 2015); DeepVariant: a deep neural network that converts high-throughput sequencing data into a list of variants (Poplin et al., 2018); DeepBind: a deep learning framework that learns the sequence preferences of DNA- and RNA-binding proteins from high-throughput binding assays (Alipanahi et al., 2015); DeepGestalt: a deep facial analysis framework that identifies rare genetic syndromes from facial images (Gurovich et al., 2019). Although more and

more deep learning-based approaches are applied to genomics research, there is no tool developed for single-cell CNV calling using such powerful techniques.

There are various deep learning approaches that can be used to model CNVs in single cells. For example, we can use a 1-dimension convolutional neural network (CNN) to apply sliding convolutional filters to the input. We can also use Long Short-Term Memory (LSTM), which is a type of recurrent neural network (RNN) that can learn from sequential data such as text, speech, or video (Hochreiter and Schmidhuber, 1997). Compared with 1D CNN, LSTM networks are better suited for tasks that require modeling long-term dependencies in sequential data, while 1D CNNs are better suited for tasks that require feature extraction from local patterns in the input sequence (Kiranyaz et al., 2019). Unlike other machine learning models that assume the input data are independent and identically distributed, LSTM can capture the temporal dependencies and long-term patterns in the data. Compared with other RNN models, LSTM has a special mechanism called the memory cell, which consists of three gates: input gate, forget gate, and output gate. These gates can regulate the flow of information in and out of the cell, allowing LSTM to selectively remember or forget previous states. This helps LSTM to overcome the problem of vanishing or exploding gradients that often occurs in standard RNNs when dealing with long sequences. Compared with the Hidden Markov Model (HMM), LSTM can learn complex nonlinear relationships between the input and output sequences, while HMM can only model linear or simple nonlinear dependencies (Rabiner, 1989). Therefore, LSTM is a powerful and versatile tool for various sequence learning tasks such as natural language processing, speech recognition, genomic sequence analysis, and more.

Bidirectional LSTM models are a type of recurrent neural network that can process sequential data in both forward and backward directions (Schuster and Paliwal, 1997). Unlike

standard LSTM models, which only use the previous information to make predictions, bidirectional LSTM models can also leverage the posterior information to improve the accuracy and robustness of the model. Bidirectional LSTM models consist of two parallel LSTM layers, one for each direction, and a merge layer that combines the outputs of both layers. The merge layer can use different strategies, such as concatenation, summation, or averaging, to fuse the information from both directions. Bidirectional LSTM models are especially useful for sequence prediction tasks, such as sentiment analysis, machine translation, and named entity recognition, where the context from both sides of a word or a sentence can provide valuable clues for the task (Graves et al., 2013; Lample et al., 2016; Long et al., 2019; Sundermeyer et al., 2014). Therefore, we can also apply bidirectional LSTM models on the CNV calling problem, which is the prediction of copy number on genomic regions.

We present a novel method, ScovalNN, for single-cell CNV calling based on deep learning. Our method uses both sequencing coverage and allelic ratio information as features to train a bidirectional long short-term memory (LSTM) neural network that can predict the copy number variation on each genomic window. We use our simulation dataset, which mimics realistic scenarios of CNV events, as the training set for our model. We compare our method with three existing tools: Ginkgo (Garvin et al., 2015), SCYN (Feng et al., 2021) and CHISEL (Zaccaria and Raphael, 2021), which are based on circular binary segmentation, dynamic programming and probabilistic modeling, respectively. We show that our method outperforms these tools in terms of accuracy, sensitivity and specificity on both CNV and window level. Moreover, our method can leverage allelic ratio information to detect copy-neutral loss of heterozygosity (CN-LOH), which is a challenge for Ginkgo and SCYN. Our method is also faster and more scalable than CHISEL, which requires haplotype phasing and multiple iterations.

Therefore, we propose that our deep learning-based method is a powerful and efficient tool for single-cell CNV calling.

4.2 Data and methods

In order to acquire the training data, we use the single-cell CNV simulator, which is described in Chapter 3, to randomly simulate 6,000 CNVs in the 1,000 randomly selected non-CNV cells, 1,500 CNVs for each type respectively (homozygous deletion, heterozygous deletion, copy-neutral loss of heterozygosity (CN-LOH), and duplication). Then we use the phased heterozygous SNPs (het-SNPs) to make windows, and each window contains 100 phased het-SNPs (see details in Chapter 2, 3, and Figure 4.1). Some CNVs would fall into a single het-SNP window, which is the basic unit for the simulator and ScovalNN. Therefore, these CNVs would be filtered out. In the end, we have 5,842 CNVs from 998 cells. There are 1,469 homozygous deletions, 1,460 heterozygous deletions, 1,453 CN-LOHs, and 1,460 duplications.

Then we extract all the 5,032 chromosomes with CNVs as the model input sequences. We split these 5,032 sequences into training, validation and testing sets, 3,522, 755, 755 for each. For each window, we generate 5 features: log2 ratio between informative reads on the two haplotypes, number of total informative reads, normalized number of total informative reads, number of total reads, normalized number of total reads. The read number normalization method is the same as that in Chapter 2 and 3.

$$normalized\ count = \frac{10^{10} * raw\ read\ count}{read\ count\ of\ this\ cell * window\ size}$$

The target label for each window would be converted into 0, 1, 2, 3, and 4, which represent homozygous deletion, heterozygous deletion, normal, duplication and CN-LOH.

However, as most of the het-SNP windows are still normal (non CNV) windows, the class ratio among these 5 classes (normal and 4 types of CNV) is still very high (~150:1:1:1:1). In order to keep a balanced ratio among classes for the training data, we extract the CNV and its two-sides flanking regions (⅓ size of CNV). These subsequences from the same chromosome would be combined into a new sequence. The ratio is decreased to ~1.5:1:1:1:1 for the normal windows and other 4 types of CNV windows. For the validation and testing sequences, we keep the original sequences to make them similar to the real case.

We implemented a bidirectional LSTM neural network for genomic region copy number prediction with 5 features as input and 5 types of classes as output. The network consists of seven layers: one input layer, one fully connected layer, one bidirectional LSTM layer, three fully connected layers, and one output layer. The input layer takes the 5 features as input. The second layer transforms the low-dimensional (5) data into high-dimensional (128) space using a linear transformation. The bidirectional LSTM layer has 64 units for each direction and learns long-term dependencies from both the forward and backward sequences. The three fully connected layers have 128 units each and use Leaky ReLU activation functions to introduce non-linearity and avoid gradient vanishing. The output layer uses a softmax function to produce the class probabilities. We use cross entropy as the loss function to measure the discrepancy between the predicted and true labels. To optimize the network parameters, we used Adam optimizer with a learning rate of 1e-3 and a batch size of 64. We also applied early stopping to prevent overfitting and terminate the training process if the validation performance does not improve for more than 50 epochs.

We applied the model on different test sets and obtained the initial CNV calls. Then, we would smooth and filter the CNV calls using a pipeline that consists of six steps. First, we

adjusted the copy number of a small number (<5) of CNV windows to match their neighboring copy number, and smoothed the copy number for each chromosome. Second, we merged two CNVs with the same copy number if their distance is less than 500K. Third, we discarded CNVs smaller than 500K. Fourth, we excluded CNVs that overlap more than 20% with the centromere or telomere of each chromosome. Fifth, we removed CNVs that overlap more than 20% with ENCODE blacklist regions (Amemiya et al., 2019). Sixth, we eliminated CNVs that have fewer than 5 informative windows or less than 30% of windows are informative. An informative window is a het-SNP window with more than 3 het-SNPs covered by the reads.

4.3 Results

In this study, we developed ScovalNN, a bidirectional LSTM neural network for single-cell copy number prediction with 5 features as input and 5 types of classes as output. As described in the method section, the network consisted of seven layers, including one input layer, one fully connected layer, one bidirectional LSTM layer, three fully connected layers, and one output layer. (Figure 4.1) The input layer received the 5 features as input, which were transformed into a high-dimensional space of 128 units using a linear transformation in the second layer. The bidirectional LSTM layer, with 64 units for each direction, learned long-term dependencies from both the forward and backward sequences. The three fully connected layers, each with 128 units, introduced non-linearity and avoided gradient vanishing using Leaky ReLU activation functions. The output layer used a softmax function to produce the class probabilities.

To evaluate the performance of the proposed model during the training, we use cross entropy to calculate the training and validation loss. (Figure 4.2) We use accuracy and F1 score

to evaluate the performance of the model. The model achieves a high accuracy of 0.9936 and a F1 score of 0.9556 on the validation set, indicating its effectiveness and robustness.

To further demonstrate the applicability of ScovalNN, we apply it to two different datasets: a simulated dataset of 300 cells with CNVs (the same dataset described in Chapter 3 with no subclonal frequency simulation), and a real dataset of 2,097 cells (the common control sample in Chapter 2). We perform CNV calling on both datasets and compare the results with the ground truth at both CNV level and window level. For the simulated dataset, we use the simulated CNV set as the ground truth; for the real dataset, we use our conservative CNV calls as the ground truth. We report the comparison results before and after applying a filtering pipeline to remove low-confidence CNV calls.

On the CNV level comparison, the criteria of the same CNV between ground truth and prediction is that the CNV in one set is 80% overlap with the CNV in another set, and the CNVs have the same copy number. Based on this criterion, we can calculate the precision, recall and F1-score for ScovalNN on each data set (Table 4.1). Precision is the proportion of predicted CNVs that are true positives, and recall is the proportion of true CNVs that are detected by the method. Compared with other methods benchmarked in Chapter 3 on the CNV level, ScovalNN outperforms SCYN and CHISEL, and is comparable with Ginkgo (Figure 4.3). Ginkgo has a little higher precision as there is a training bias since the simulation data was based on Ginkgo's initial results. However, Ginkgo cannot identify CN-LOH. On the window level comparison, we use a different criterion to match CNVs between the ground truth and the prediction sets. We consider a window as a true positive if it has the same copy number as the ground truth. We then calculate the confusion matrices for each method on each data set (Figure 4.4), which show the number of windows that are correctly or incorrectly classified by each method. From the

confusion matrices, we can also calculate the precision, recall, F1-score and specificity for each method on each data set (Table 4.2 & 4.3). The F1-score is the harmonic mean of precision and recall, and specificity is the proportion of windows with other copy numbers that are correctly classified by the method. A higher F1-score and specificity indicate a better balance between accuracy and sensitivity of the method. Even though the ground truth set does not include duplications and CN-LOH and it is very conservative, ScovalNN still performs very well on the real data set. This demonstrates that ScovalNN is a robust and reliable method for CNV detection from single-cell data.

To further evaluate the performance of our method, we compared the median absolute log2 ratio and median read depth ratio of the predicted CNVs in the real data set. (Figure 4.5) These metrics reflect the characteristics of different types of CNVs, and can be used as additional filters to refine the predictions. As shown in Figure 4.5, our method produces distinct clusters for each type of CNVs. Moreover, our method can accurately identify the boundaries of CNVs, which is crucial for downstream analysis. (Figure 4.6) In contrast, Ginkgo and SCOVAL tend to overestimate or underestimate the size of CNVs, leading to false positives or false negatives.

4.4 Discussion

In this study, we proposed a novel method called ScovalNN for single-cell CNV calling based on deep learning. Our method uses a bidirectional LSTM neural network to predict the copy number variation on each genomic window using both sequencing coverage and allelic ratio information as features. We compared our method with three existing tools: Ginkgo, SCYN, and CHISEL and showed that our method outperforms these tools in terms of accuracy, sensitivity, and specificity on both CNV and window level.

Our results demonstrate that ScovalNN is a robust and reliable method for CNV detection from single-cell data. It outperforms several other methods and can accurately identify the boundaries of CNVs, which is crucial for downstream analysis. Moreover, our method can leverage allelic ratio information to detect copy-neutral loss of heterozygosity (CN-LOH), which is a challenge for Ginkgo and SCYN. Our method is also faster and more scalable than CHISEL, which requires haplotype phasing and multiple iterations. Overall, these results suggest that ScovalNN is a robust and reliable method for CNV detection from single-cell data.

One limitation of our study is that we only evaluated our method on simulated and real datasets with conservative and partial ground truth. While our method performed well on these datasets, it is possible that its performance may vary on other datasets with different characteristics. Further studies are needed to evaluate the performance of our method on other datasets with complete ground truth.

Another limitation is that our method only uses sequencing coverage and allelic ratio information as features for CNV calling. While these features are informative, there may be other types of data that could improve the performance of our method. For example, integrating gene expression or epigenetic data could provide additional information for CNV calling. Future studies could explore the potential of incorporating these types of data into our method.

In terms of future directions, there are several avenues for further research. One direction could be to extend our method to call CNV from single-cell ATAC data. Single-cell ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a powerful technique for profiling chromatin accessibility at single-cell resolution. By analyzing the patterns of chromatin accessibility in single cells, it may be possible to infer the presence of CNVs. Developing

methods for calling CNVs from single-cell ATAC data could provide a complementary approach to our current method and could improve the accuracy and sensitivity of CNV calling.

Another direction could be to extend our method to call CNV from single-cell long-read sequencing data. Long-read sequencing technologies, such as those developed by PacBio and Oxford Nanopore, can generate reads that are tens of kilobases in length. These long reads can span complex genomic regions and can provide more accurate information about the structure of the genome. Developing methods for calling CNVs from single-cell long-read sequencing data could provide a higher-resolution view of genomic variation in single cells and could improve our ability to detect complex CNVs.

In addition, we can generalize our method to the single-cell sequencing data generated from other sequencing technologies. Our model is currently trained on single-cell sequencing data generated by 10X Genomics. However, other WGA methods or sequencing techniques may generate data with different characteristics, such as different sequencing error rates, library preparation protocols, and read lengths. By adapting our method to these different sequencing technologies, it may be possible to develop new tools for CNV calling that are tailored to specific datasets. This could further improve the accuracy and sensitivity of CNV calling and provide a more comprehensive view of genomic variation in single cells. Overall, the potential for our method to be generalized to other sequencing technologies highlights the importance of developing flexible and adaptable deep learning-based tools for single-cell genomics.

In conclusion, our study presents a powerful and efficient tool for single-cell CNV calling based on deep learning. Our method outperforms several existing tools and has several advantages, such as the ability to detect CN-LOH and its scalability. We believe that our method will be a valuable addition to the toolkit of researchers working on single-cell genomics.

Figures

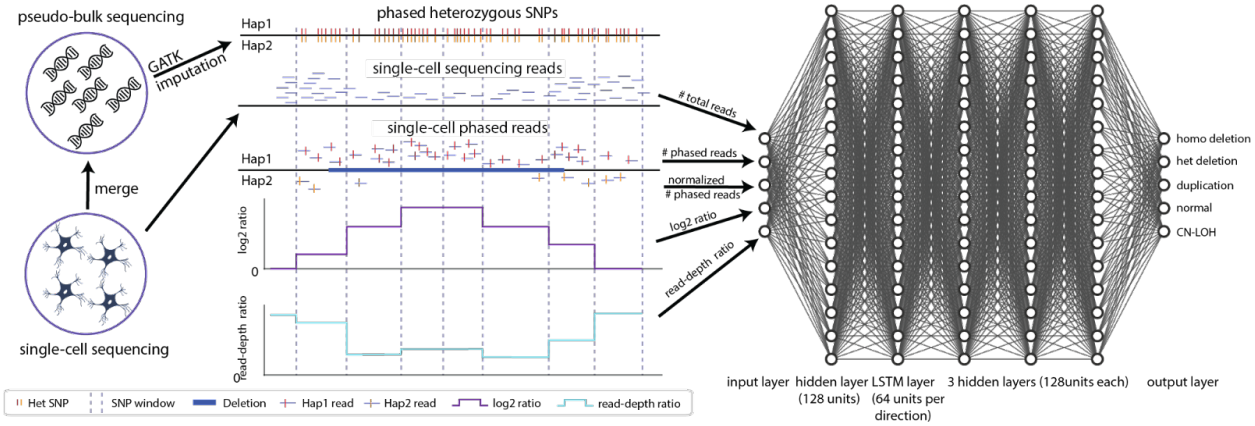


Figure 4.1 Overview of the ScovalNN framework.

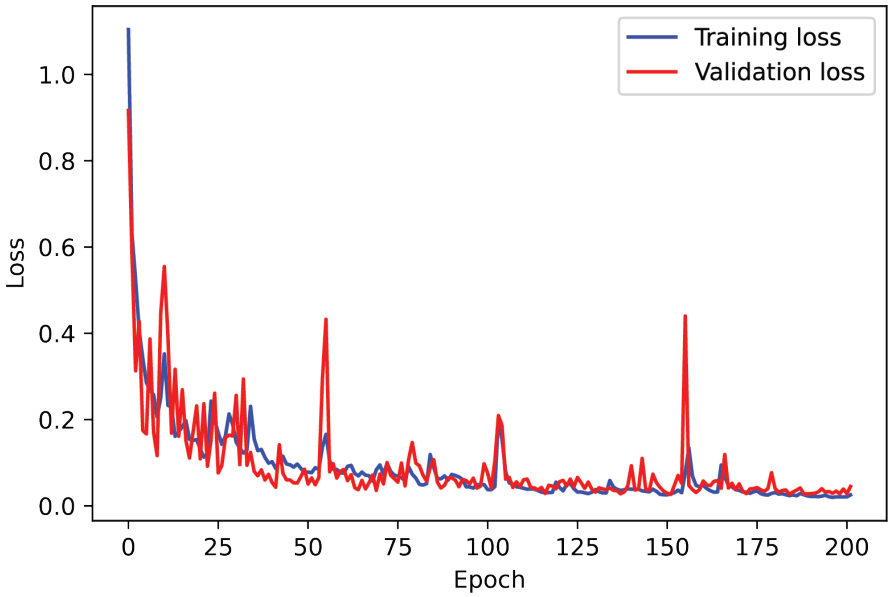


Figure 4.2 Training loss and validation loss during the training of ScovalNN.

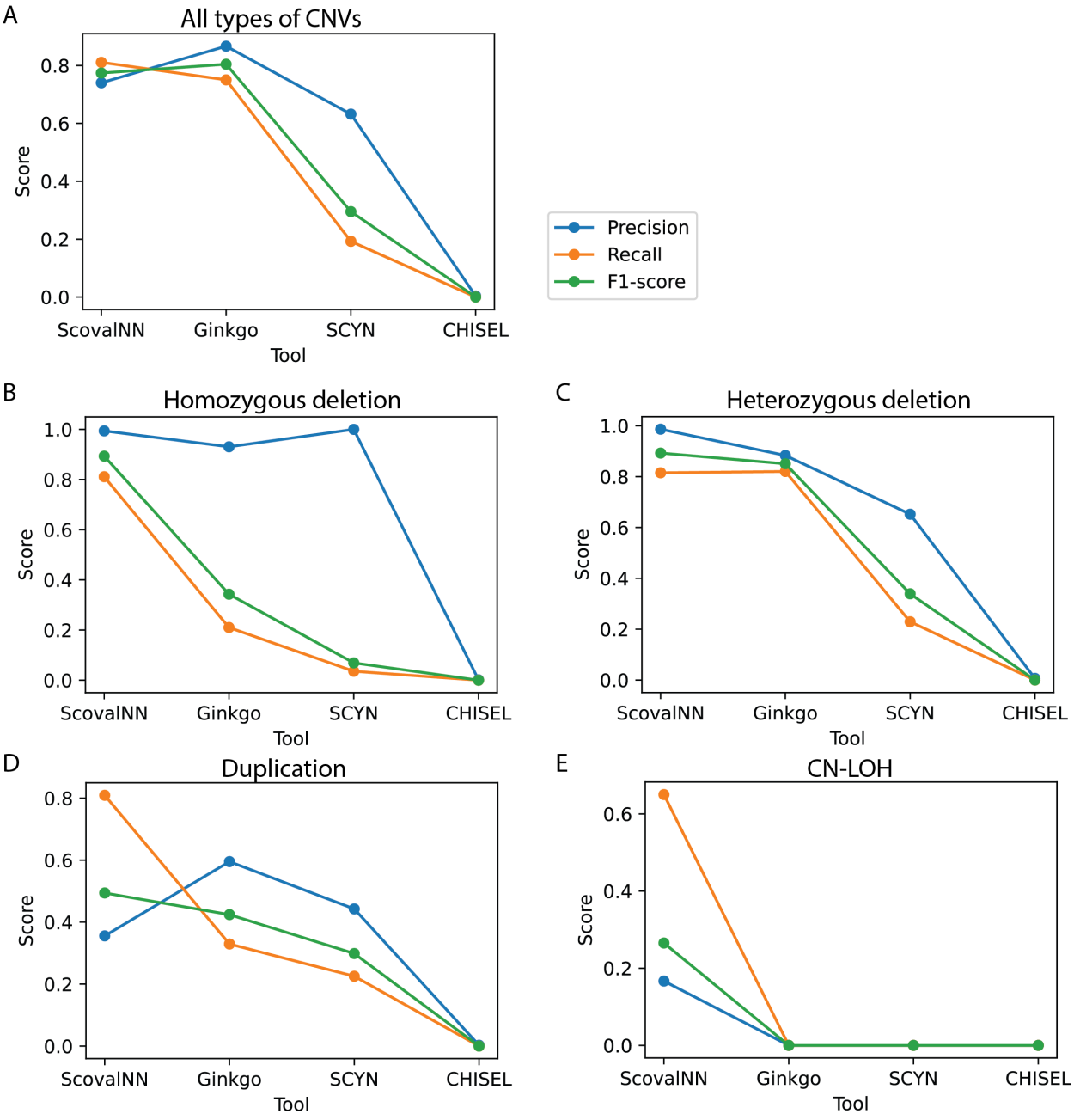


Figure 4.3 CNV level performance comparison of ScovalNN, Ginkgo, SCYN and CHISEL on the simulated data set.

(A) All types of CNVs. (B) Homozygous deletion. (C) Heterozygous deletion. (D) Duplication. (E) CN-LOH.

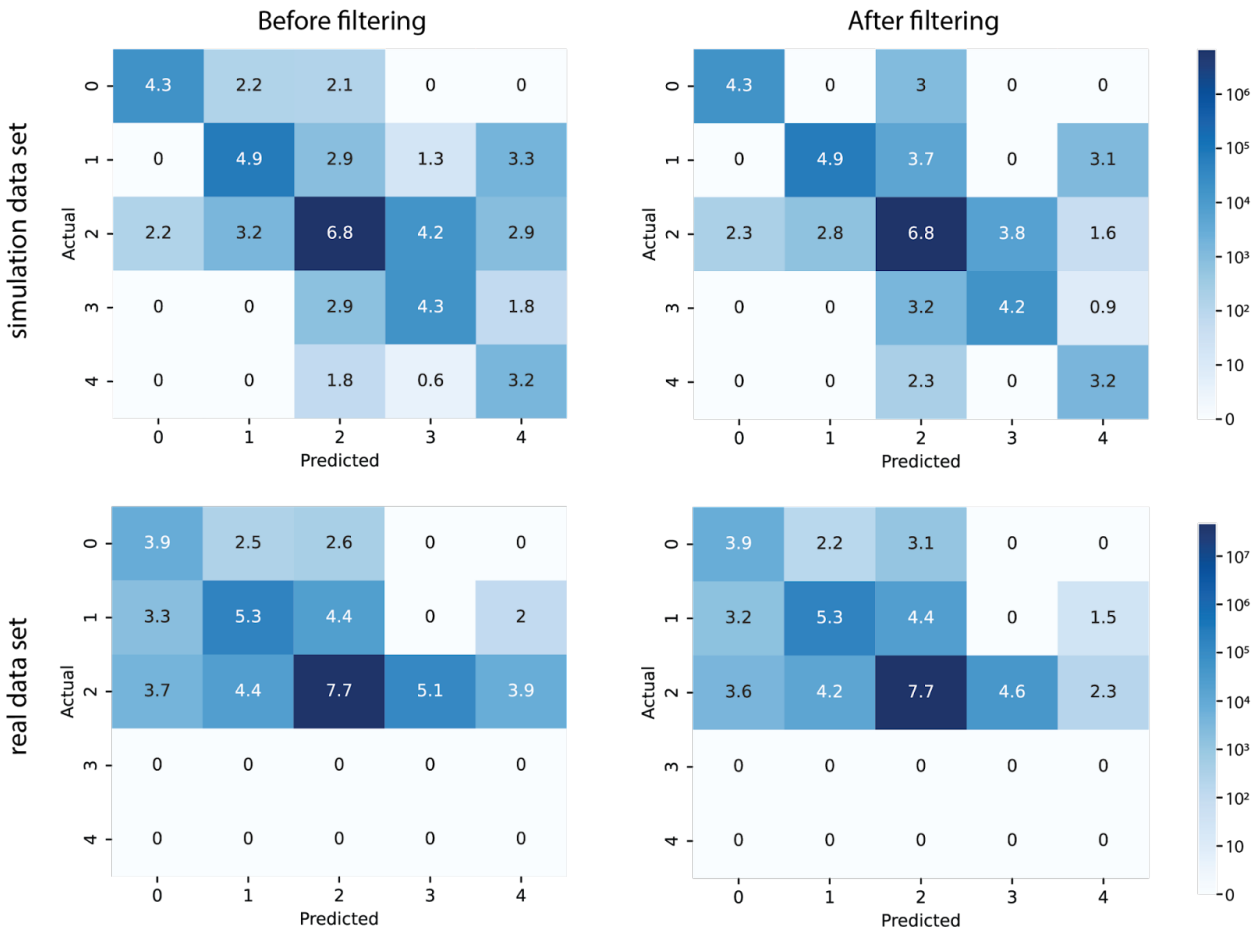


Figure 4.4 Heatmaps of the confusion matrices (log scale) for ScovalNN prediction on the window level on simulation data and real data sets.

The ground truth of the simulation set is the simulated CNV profile. The ground truth of the real data set is the very conservative CNV call set described in Chapter 2, and it does not include duplications and CN-LOHs.

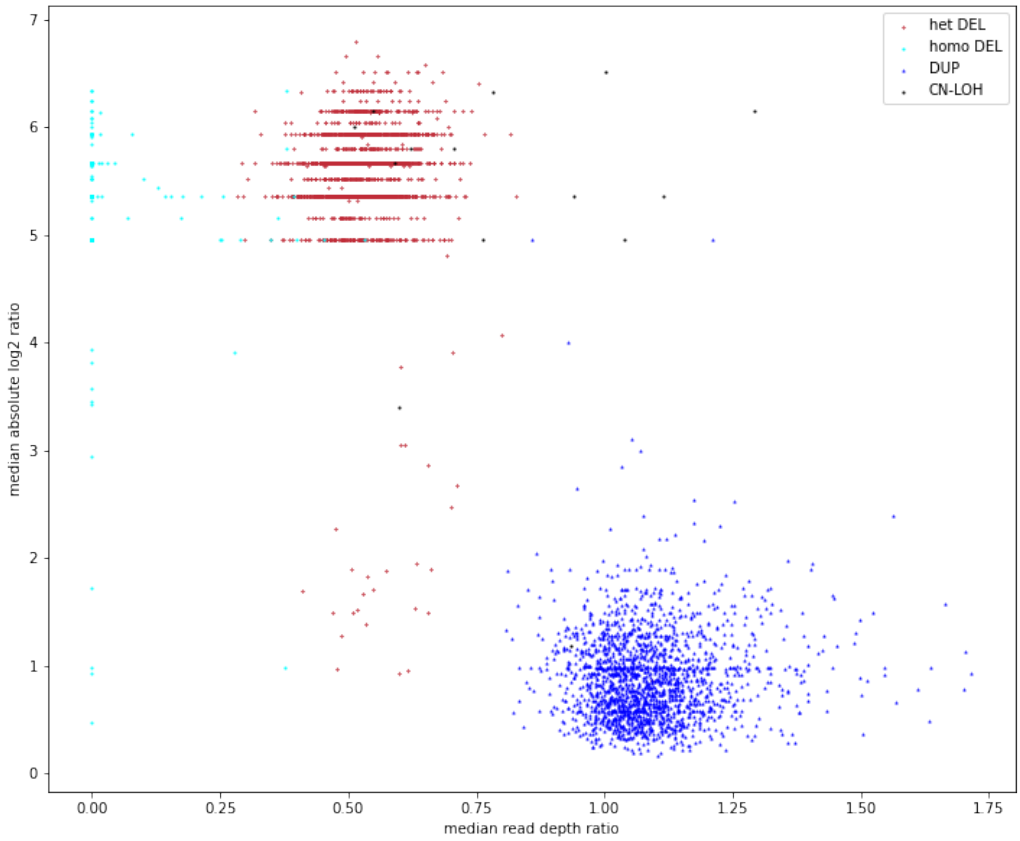
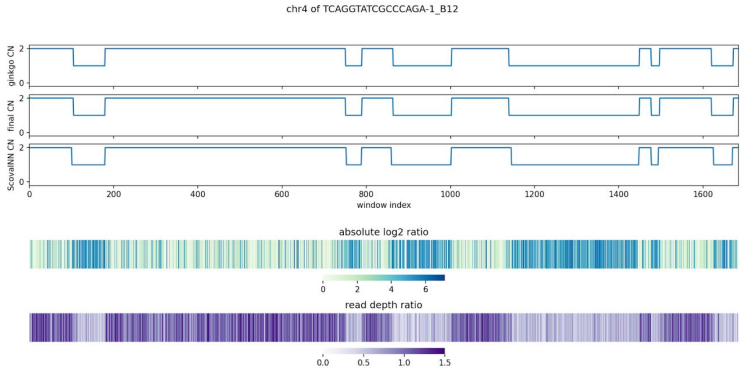
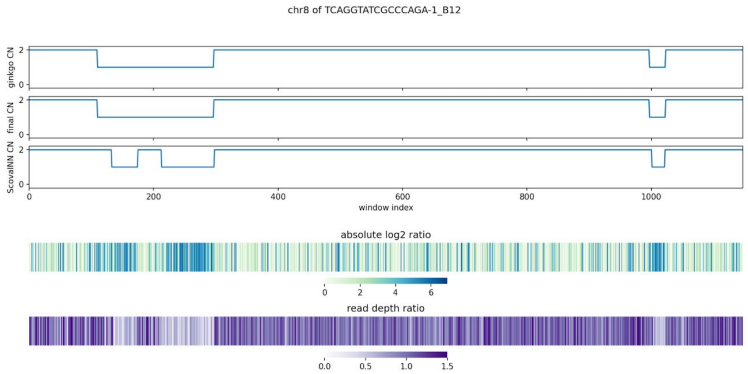


Figure 4.5 Median absolute log2 ratio vs. median read depth ratio for each predicted CNV on the real data set.

A



B



C

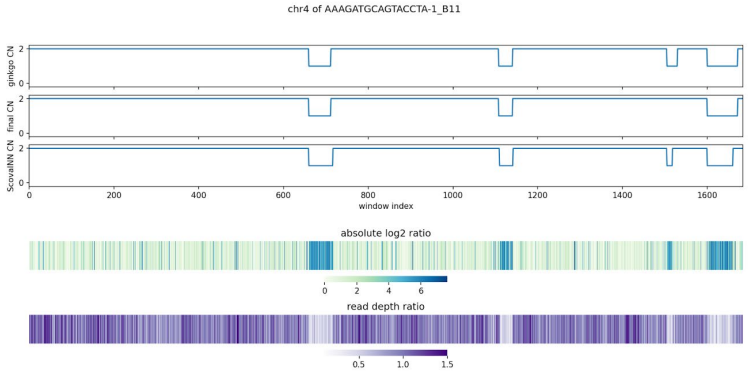


Figure 4.6 Example chromosomes comparing copy number profiles from Ginkgo, SCOVAL and ScovalNN.

Tables

Table 4.1 Comparison of precision and recall on the CNV level of the before and after filtering ScovalNN CNV calls for both simulation and real data sets.

		Precision	Recall	F1 score
Simulation data	Before filtering	0.1423	0.7104	0.2371
	After filtering	0.7300	0.7998	0.7633
Real data	Before filtering	0.0808	0.3199	0.1290
	After filtering	0.7618	0.7363	0.7489

Table 4.2 Window level performance of after filtering SocvalNN CNV calls for the simulation data set.

	Precision	Recall	F1 score	Specificity
Homo-DEL(CN=0)	0.9898	0.9462	0.9675	0.9999
Het-DEL(CN=1)	0.9912	0.9256	0.9573	0.9999
Normal (CN=2)	0.9989	0.9989	0.9989	0.9369
DUP(CN > 2)	0.7390	0.9158	0.8180	0.9991
CN-LOH(CN=2)	0.5421	0.8811	0.6712	0.9998
Macro average	0.8522	0.9335	0.8826	0.9871

Table 4.3 Window level performance of after filtering SocvalNN CNV calls for the real data set (ground truth does not include duplication and CN-LOH).

	Precision	Recall	F1 score	Specificity
Homo-DEL(CN=0)	0.5886	0.8567	0.6978	0.9999
Het-DEL(CN=1)	0.9262	0.8711	0.8978	0.9997
Normal (CN=2)	0.9994	0.9987	0.9991	0.8788
DUP(CN > 2)	0	0	0	0.9991
CN-LOH(CN=2)	0	0	0	0.9999
Macro average	0.5029	0.5453	0.5189	0.9755

Bibliography

Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J., 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. 33, 831–838. https://doi.org/10.1038/nbt.3300

Amemiya, H.M., Kundaje, A., Boyle, A.P., 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 9, 9354. https://doi.org/10.1038/s41598-019-45839-z

Eraslan, G., Avsec, Ž., Gagneur, J., Theis, F.J., 2019. Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. 20, 389–403. https://doi.org/10.1038/s41576-019-0122-6

Feng, X., Chen, L., Qing, Y., Li, R., Li, C., Li, S.C., 2021. SCYN: single cell CNV profiling method using dynamic programming. BMC Genomics 22, 651. https://doi.org/10.1186/s12864-021-07941-3

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M., Schatz, M.C., 2015. Interactive analysis and assessment of single-cell copy-number variations. Nat. Methods 12, 1058–1060. https://doi.org/10.1038/nmeth.3578

Graves, A., Jaitly, N., Mohamed, A., 2013. Hybrid speech recognition with Deep Bidirectional LSTM, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Presented at the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278. https://doi.org/10.1109/ASRU.2013.6707742

Grayton, H.M., Fernandes, C., Rujescu, D., Collier, D.A., 2012. Copy number variations in neurodevelopmental disorders. Prog. Neurobiol. 99, 81–91. https://doi.org/10.1016/j.pneurobio.2012.07.005

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P.M., Kamphausen, S.B., Zenker, M., Bird, L.M., Gripp, K.W., 2019. Identifying facial phenotypes of genetic disorders using deep learning. Nat. Med. 25, 60–64. https://doi.org/10.1038/s41591-018-0279-0

Henrichsen, C.N., Chaignat, E., Reymond, A., 2009. Copy number variants, diseases and gene expression. Hum. Mol. Genet. 18, R1–R8. https://doi.org/10.1093/hmg/ddp011

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., Gabbouj, M., 2019. 1-D Convolutional Neural Networks for Signal Processing Applications, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8360–8364. https://doi.org/10.1109/ICASSP.2019.8682194

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C., 2016. Neural Architectures for Named Entity Recognition. https://doi.org/10.48550/arXiv.1603.01360

Long, F., Zhou, K., Ou, W., 2019. Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention. IEEE Access 7, 141960–141969. https://doi.org/10.1109/ACCESS.2019.2942614

Mallory, X.F., Edrisi, M., Navin, N., Nakhleh, L., 2020. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol. 21, 208. https://doi.org/10.1186/s13059-020-02119-8

Ning, L., Liu, G., Li, G., Hou, Y., Tong, Y., He, J., 2014. Current Challenges in the Bioinformatics of Single Cell Genomics. Front. Oncol. 4.
Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., Gross, S.S., Dorfman, L., McLean, C.Y., DePristo, M.A., 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987. https://doi.org/10.1038/nbt.4235
Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286. https://doi.org/10.1109/5.18626
Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45, 2673–2681. https://doi.org/10.1109/78.650093
Shlien, A., Malkin, D., 2009. Copy number variations and cancer. Genome Med. 1, 62. https://doi.org/10.1186/gm62
Sundermeyer, M., Alkhouli, T., Wuebker, J., Ney, H., 2014. Translation Modeling with Bidirectional Recurrent Neural Networks, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Presented at the EMNLP 2014, Association for Computational Linguistics, Doha, Qatar, pp. 14–25. https://doi.org/10.3115/v1/D14-1003
Zaccaria, S., Raphael, B.J., 2021. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nat. Biotechnol. 39, 207–214. https://doi.org/10.1038/s41587-020-0661-6
Zarrei, M., MacDonald, J.R., Merico, D., Scherer, S.W., 2015. A copy number variation map of the human genome. Nat. Rev. Genet. 16, 172–183. https://doi.org/10.1038/nrg3871
Zhou, J., Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods 12, 931–934. https://doi.org/10.1038/nmeth.3547
Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A., 2019. A primer on deep learning in genomics. Nat. Genet. 51, 12–18. https://doi.org/10.1038/s41588-018-0295-5

Chapter 5 Conclusion

5.1 Overview

Somatic mutations are genetic variations that occur in a subset of cells during an individual's lifespan, accumulating during the developmental and aging processes. While these mutations have long been associated with cancer, recent studies suggest their involvement in a range of non-cancer diseases, such as neurological diseases. High-throughput sequencing technology has made it possible to profile these mutations genome-wide; however, detecting somatic mutations from bulk tissue samples poses significant challenges as they occur only in a subset of cells. An alternative approach is to sequence single-cell genomes after whole-genome amplification, but this method presents technical challenges due to error-prone and uneven genome amplification. To overcome these challenges, many bioinformatic tools have been developed. In this dissertation, we discuss our latest advancements in somatic copy number variation (CNV) detection using single-cell DNA sequencing and the challenges that still need to be addressed.

In this dissertation, I aimed to develop and evaluate novel computational methods for detecting somatic CNVs using single-cell DNA sequencing data. Somatic CNVs can have crucial implications for various biological mechanisms and diseases. Single-cell DNA sequencing is a powerful technique that can reveal the presence and extent of somatic CNVs at the resolution of individual cells. Nonetheless, the analysis of somatic CNVs through single-cell DNA sequencing data presents significant challenges, including inadequate sequencing coverage, high noise,

amplification bias, and artifacts. Therefore, the development of robust and precise methodologies capable of overcoming these challenges and accurately identifying somatic CNVs from single-cell DNA sequencing data is urgently required.

In this dissertation, I present three main contributions that advance the state-of-the-art of somatic CNV detection using single-cell DNA sequencing data. First, I developed SCOVAL, a method that integrates read-depth and phased loss-of-heterozygosity information to identify somatic CNVs in single cells. I applied SCOVAL to 2,125 frontal cortical neurons from a neurotypical human brain and discovered 226 CNV neurons, including a novel class of neurons with complex karyotypes characterized by whole or substantial losses of multiple chromosomes. Second, I developed a single-cell CNV simulator that can generate realistic single-cell DNA sequencing data with predefined somatic CNVs. I used the simulator to benchmark existing tools for somatic CNV detection and to provide a ground truth dataset for the development and evaluation of new methods. Third, I developed a state-of-the-art deep learning-based approach, ScovalNN, which leverages Long short-term memory (LSTM) neural networks for identifying somatic CNVs from single-cell DNA sequencing data. The performance of ScovalNN was rigorously assessed on both simulated and real-world data, exhibiting superior results compared to existing methodologies across various performance metrics.

The methods and results presented in this dissertation provide new insights into the occurrence and characteristics of somatic CNVs in human neurons and other cell types. They also demonstrate the potential of single-cell DNA sequencing and computational analysis for uncovering the genomic diversity and complexity of somatic cells.

5.2 Limitations and challenges

In this dissertation, we investigated the genomic architecture of neuronal cells and developed new methods for detecting copy number variations (CNVs) at the single-cell level. While the findings and tools presented in this study offer promising results, there are still limitations and challenges that need to be addressed in future studies.

In Chapter 2, our approach identified 1,957 CNVs from 226 CNV neurons, as well as a class of CNV neurons with complex karyotypes containing whole or substantial losses on multiple chromosomes. Moreover, we found that CNV location appears to be nonrandom and recurrent regions of neuronal genome rearrangement containing fewer, but longer, genes. However, this research has some limitations that need to be addressed in future studies. First, we did not use any orthogonal methods to validate the existence of CNV in our data. Therefore, though our simulation experiments in Chapter 3 suggest that false positives will be less common, it is still possible that some of the CNV signals we detected are false positives or artifacts. Second, we only analyzed the data from neurons of a single individual, which may not be representative of the whole population of neurons in the brain region of interest. To draw a more general and robust conclusion, we need to repeat the experiments in multiple samples and compare their CNV patterns across different conditions.

In Chapter 3, we developed a novel sampling-based single-cell CNV simulator that can generate realistic synthetic data for assessing the performance of single-cell CNV callers. It shows that the simulator can produce realistic data that mimics the characteristics of real scDNA-seq data in terms of coverage distribution, allelic ratio and noise level. It also shows that the simulator can reveal the strengths and weaknesses of different single-cell CNV callers and help identify areas for improvement. Although our simulator can overcome the limitations in the fully

synthetic simulators by leveraging existing data as a source of information and variation, it still has the bias from the single individual as the only source sample, including the feature distributions, technical noise patterns, etc. One possible way to overcome the limitation in the future is to use multiple individuals with different genetic backgrounds to create a more diverse and representative source dataset. Additionally, incorporating more comprehensive technical noise patterns and feature distributions from different sequencing technologies may further enhance the simulator's accuracy and reduce any potential bias.

In Chapter 4, we proposed ScovalNN, a deep learning based single-cell CNV caller. ScovalNN can leverage both read coverage and allelic ratio information to detect different types of CNVs, including copy-neutral loss of heterozygosity (CN-LOH), which is a challenge for some other tools. Our results show that ScovalNN outperforms other tools on both CNV and window level. However, there are several limitations for this tool. We only consider the sequencing coverage and allelic ratio as the input features. It may improve the performance if we add other features such as epigenetic data. In addition, we only evaluated it on the simulated dataset and the real dataset with conservative and partial ground truth, which may not be representative of all datasets.

Overall, further validation and analysis across multiple samples and datasets, as well as incorporation of additional features, will be necessary to fully understand the genomic landscape of neuronal cells and improve the accuracy of single-cell CNV detection tools.

5.3 Implications for neuropsychiatric disease

The findings of my dissertation have important implications for our understanding of the role of somatic CNVs in neuropsychiatric diseases. In particular, my work can shed light on how

single-cell DNA sequencing can be used to investigate the genomic diversity and complexity of somatic cells in the context of neurological diseases.

Neuropsychiatric disorders, such as autism spectrum disorder (ASD), schizophrenia, and bipolar disorder, are complex and heterogeneous conditions that affect brain function and behavior. The genetic basis of these disorders is not fully understood, but it is known that some individuals carry large germline CNVs or microdeletions that disrupt genes involved in neuronal development, synaptic transmission, or neuroplasticity (Cook Jr and Scherer, 2008; Fanciulli et al., 2010). These germline variants can affect brain structure and connectivity, and increase the risk for developing neuropsychiatric symptoms (Nakatochi et al., 2021). However, not all individuals with germline CNVs or microdeletions develop neuropsychiatric disorders, suggesting that other factors may modulate the phenotypic outcome (Kirov, 2015). One of these factors could be the presence of somatic mutations that arise during brain development or later in life, resulting in mosaicism. Somatic mutations can act as a second hit that worsens the effects of the germline variant, or confers additional risk for developing neuropsychiatric disorders (Poduri et al., 2013).

One type of somatic mutation that has been associated with various diseases is the double-hit somatic variant, which involves the simultaneous occurrence of two or more mutations in a single cell (Knudson, 1971). Double-hit somatic variants can have a greater impact on gene expression or function than single-hit variants, and may lead to more severe phenotypes (Pelorosso et al., 2019; Ye et al., 2019; Zeng et al., 2021). Single-cell DNA sequencing is a powerful technique that can detect and characterize these double-hit somatic variants, as well as other types of somatic mutations, at high resolution and sensitivity (Bizzotto and Walsh, 2022). By applying single-cell DNA sequencing to brain tissue samples from

individuals with neuropsychiatric disorders, we can investigate the role of double-hit somatic variants in these conditions, and to elucidate the genetic and molecular mechanisms underlying their development. My dissertation provides new insights into the contribution of somatic CNVs to neuropsychiatric disorders, and will demonstrate the utility of single-cell DNA sequencing for studying the genomic diversity and complexity of somatic cells in the brain.

5.4 Future directions

This dissertation presented three novel methods, SCOVAL, single-cell CNV simulator and ScovalNN, for detection and simulation of somatic CNVs from single-cell DNA sequencing data. We applied our methods to a brain tissue sample from a neurotypical individual and revealed the complex and diverse CNV landscape of the human brain. However, there are still many directions for future research.

One direction is to extend our analysis to other samples, both neurotypical and disease-related. By comparing the CNV profiles of different samples, we could identify the CNVs that are associated with neurological disorders such as autism, schizophrenia, and Alzheimer's disease. We could also investigate how the CNVs affect the gene expression and function of different cell types in the brain. Furthermore, we could use the new data to improve our single-cell CNV simulator and generate more realistic and diverse training data for ScovalNN. This would enhance the generalizability and robustness of our deep learning model across different samples and sequencing platforms.

Another direction is to adapt our methods to long-read sequencing data. Our methods are designed for short-read sequencing data and cannot be directly applied to long-read data. Long-read sequencing technologies have become increasingly popular as they can capture the long-

range information to resolve complex structural variations, and do not need whole genome amplification. One potential technology that could enable the direct sequencing of long molecules without whole genome amplification is Nanopore sequencing (Deamer et al., 2016; Jain et al., 2016; Lebrigand et al., 2020; Wen and Tang, 2022). Nanopore sequencing utilizes nanopores embedded in a membrane to directly read DNA molecules as they pass through the pore. This technology has the advantage of being able to sequence long DNA molecules without the need for amplification, allowing for more accurate detection of CNVs and other genetic variations. Furthermore, the development of new algorithms and tools specifically designed for analyzing Nanopore sequencing data could facilitate the identification and genotyping of CNVs and other genomic alterations with greater accuracy and efficiency. Overall, the utilization of Nanopore sequencing coupled with new analysis tools and algorithms holds great potential for advancing the field of single-cell CNV detection and genomic research.

The other promising direction for future research is to investigate the simultaneous measurement of DNA and RNA in the same single cell (Zhu et al., 2020). Current technologies such as single-cell ATAC-seq and RNA-seq can provide valuable information on the chromatin accessibility and gene expression of individual cells, but they do not provide direct information on the relationship between genome sequencing and transcriptome sequencing within the same cell. However, recent advances in single-cell multiomics technologies offer the possibility of profiling both DNA and RNA from the same cell, providing a more comprehensive view of the genomic and transcriptomic landscapes of individual cells. One such technology that holds promise for this purpose is single-cell combinatorial indexing of chromatin accessibility and gene expression (sci-CAR), which uses combinatorial barcoding to simultaneously profile chromatin accessibility and gene expression from the same single cell (Cao et al., 2018). In

addition, Hou et al. developed scTrio-seq, which can be used to simultaneously analyze the genomic copy-number variations (CNVs), DNA methylome, and transcriptome of an individual mammalian cell (Hou et al., 2016).

In summary, this dissertation made significant contributions to the field of single-cell CNV analysis, but there are still many challenges and opportunities for future work. We hope that our methods will facilitate the discovery of novel insights into the genetic architecture and function of the human brain.

Bibliography

Bizzotto, S., Walsh, C.A., 2022. Genetic mosaicism in the human brain: from lineage tracing to neuropsychiatric disorders. Nat. Rev. Neurosci. 23, 275–286. https://doi.org/10.1038/s41583-022-00572-x

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., Steemers, F.J., Adey, A.C., Trapnell, C., Shendure, J., 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 361, 1380–1385. https://doi.org/10.1126/science.aau0730

Cook Jr, E.H., Scherer, S.W., 2008. Copy-number variations associated with neuropsychiatric conditions. Nature 455, 919–923. https://doi.org/10.1038/nature07458

Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524. https://doi.org/10.1038/nbt.3423

Fanciulli, M., Petretto, E., Aitman, T., 2010. Gene copy number variation and common human disease. Clin. Genet. 77, 201–213. https://doi.org/10.1111/j.1399-0004.2009.01342.x

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., Peng, J., 2016. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 26, 304–319. https://doi.org/10.1038/cr.2016.23

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

Kirov, G., 2015. CNVs in neuropsychiatric disorders. Hum. Mol. Genet. 24, R45–R49. https://doi.org/10.1093/hmg/ddv253

Knudson, A.G., 1971. Mutation and Cancer: Statistical Study of Retinoblastoma. Proc. Natl. Acad. Sci. 68, 820–823. https://doi.org/10.1073/pnas.68.4.820

Lebrigand, K., Magnone, V., Barbry, P., Waldmann, R., 2020. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat. Commun. 11, 4025. https://doi.org/10.1038/s41467-020-17800-6

Nakatochi, M., Kushima, I., Ozaki, N., 2021. Implications of germline copy-number variations in psychiatric disorders: review of large-scale genetic studies. J. Hum. Genet. 66, 25–37. https://doi.org/10.1038/s10038-020-00838-1

Pelorosso, C., Watrin, F., Conti, V., Buhler, E., Gelot, A., Yang, X., Mei, D., McEvoy-Venneri, J., Manent, J.-B., Cetica, V., Ball, L.L., Buccoliero, A.M., Vinck, A., Barba, C., Gleeson, J.G., Guerrini, R., Represa, A., 2019. Somatic double-hit in MTOR and RPS6 in hemimegalencephaly with intractable epilepsy. Hum. Mol. Genet. 28, 3755–3765. https://doi.org/10.1093/hmg/ddz194

Poduri, A., Evrony, G.D., Cai, X., Walsh, C.A., 2013. Somatic Mutation, Genomic Variation, and Neurological Disease. Science 341, 1237758. https://doi.org/10.1126/science.1237758

Wen, L., Tang, F., 2022. Recent advances in single-cell sequencing technologies. Precis. Clin. Med. 5, pbac002. https://doi.org/10.1093/pcmedi/pbac002

Ye, Z., McQuillan, L., Poduri, A., Green, T.E., Matsumoto, N., Mefford, H.C., Scheffer, I.E., Berkovic, S.F., Hildebrand, M.S., 2019. Somatic mutation: The hidden genetics of brain

malformations and focal epilepsies. Epilepsy Res. 155, 106161. https://doi.org/10.1016/j.eplepsyres.2019.106161
Zeng, B., Huang, P., Du, P., Sun, X., Huang, X., Fang, X., Li, L., 2021. Comprehensive Study of Germline Mutations and Double-Hit Events in Esophageal Squamous Cell Cancer. Front. Oncol. 11.
Zhu, C., Preissl, S., Ren, B., 2020. Single-cell multimodal omics: the power of many. Nat. Methods 17, 11–14. https://doi.org/10.1038/s41592-019-0691-5