Multi-Sample Structural Variation Detection and Single Cell Analysis of Human Neurons

Michael Reed Lindberg
Alexandria, VA

B.S. Biochemistry and Molecular Biology,
Pennsylvania State University
University Park, 2009

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

Department of Biochemistry and Molecular Genetics

University of Virginia
December 2014

Ira M. Hall, Ph.D. (Mentor)

Aaron R. Quinlan, Ph.D.

P. Todd Stukenberg, Ph.D. (Chair)

Michael J. McConnell, Ph.D.

Table of Contents

List of Figures

List of Abbreviations

1KGP: The 1000 Genomes Project

Chr: Chromosome

CNV: Copy number variation

DNA: Deoxyribonucleic acid

FACS: Fluorescence activated cell sorting

FCTX: Human frontal cortex

FDR: False discovery rate

FNR: False negative rate

hiPSC: Human induced pluripotent stem cell

INDEL: Insertion-deletion

kb: Kilobase

MAD: Median absolute deviation

Mb: Megabase

MDA: Multiple displacement amplification

NPC: Neural precursor cell

PE: Paired-end

QC: Quality control

RD: Read-depth

SCS: Single cell sequencing

SNP: Single nucleotide polymorphism

SV: Structural variation

TCGA: The Cancer Genome Atlas

Abstract

Deoxyribonucleic acid (DNA) is the inheritance molecule, storing genetic information in its sequence of nucleic acids. The aggregate of an organism's genetic material, or DNA, is called its genome. The genome contains the complete set of instructions to create and maintain an organism. Initially, it was thought that genomes were static entities, changing slowly over very large timescales; however, insights into evolution and discoveries made by sequencing DNA have dispelled this notion. The current understanding asserts that genomes are highly plastic and dynamic structures. Genomic alterations are broadly classified as variations and can even be found occurring among cells in the same tissue. Given the importance of DNA in the functioning of a cell, detecting and characterizing DNA variation is paramount in understanding disease, especially cancer. In this bipartite dissertation, I will describe a population-based method for detecting variation and characterizing the prevalence of copy number variation in human neurons. In Chapter 1, I will provide a brief overview of genetics and variation followed by an explanation of current methods of detecting variation in the genome. In Chapter 2, I will detail a novel framework that increases the sensitivity and specificity of genomic structural variation detection by using multiple samples. In Chapter 3, I will describe how single cell sequencing has been used to uncover mosaic copy number variation in human neurons. Finally, in Chapter 4, I will conclude with a discussion of future directions and ongoing experiments.

CHAPTER 1:
GENERAL INTRODUCTION

Genetics and the Genome

The basics and historical context

An organism's genome encodes the instructions necessary for creating and maintaining life, and cataloging the differences between individual genomes is fundamental in the study of human disease, development, and evolution. Any alteration to a genome can be broadly categorized as a genetic variation, which may exist in only a single cell in one individual or conserved across an entire population or species. The effects of variations can range from silent to lethal, and it is the influence of variation on an organism's fitness and form, also known as phenotype, which is crucial to evolutionary processes. Observing these phenotypic differences and their inheritance is what led to key insights that gave rise to what is now modern genetics. The work presented here expands current scientific knowledge by characterizing genetic variations at two extremes, namely, at the population scale level and at the single cell level.

The scientific study of genetic variation is generally recognized to have started in the middle of the 19th century with Mendel's famous hybridizations of pea plants (Mendel, 1866). Mendel's quantitative study allowed him to arrive at a particulate theory of inheritance, whereby he believed that discrete particles were responsible for imparting inheritance to offspring. Soon after this, Miescher unknowingly isolated this hypothetical particle, DNA, which he described as nuclein to Wilhelm His (His, 1869). Unfortunately, Mendel's work, as well as the discovery of DNA, went largely unnoticed for years (Lander and Weinberg, 2000). In fact, Mendel's observations were in direct opposition to the prevailing model of blending inheritance that stated traits were continuous, not discrete. In this model, traits were randomly inherited from the range of parental phenotypes. At the same time, the attentions of the scientific community had also been fixed on the evolution controversy that had erupted the in the previous decade (Darwin and Wallace, 1858). Evolution had been widely accepted, but the process by which evolution occurs (natural selection vs. saltation) was hotly contested. Of the many extant explanations, Mendelian inheritance would not be considered as a potential contender until many years later. Furthermore, the marriage of Darwin's ideas of evolution and Mendel's ideas of inheritance would not occur until the middle of the 20th century with the advent of modern evolutionary synthesis (Huxley, 1942). A necessary condition for this union was the rediscovery of Mendel's work (Lander and Weinberg, 2000), occurring in 1900 by three independent researchers (Correns, 1990; Tschermak, 1900; de Vries, 1900).

The resurgence of Mendel's work prompted scientists to attempt to find a connection between inheritance and one of the many microscopic structures observed in the cell. This connection was established when the role of the nucleus was elucidated and then, more specifically, when it was determined that chromosomes inside the nucleus were responsible for carrying genetic material (Sutton, 1902; Boveri, 1902). Morgan's fruit fly experiments (Morgan, 1910) cemented this relationship and validated the Sutton-Boveri chromosome theory. One of Morgan's students then created the first genetic map of chromosomes (Sturtevant, 1913). These maps defined distances between genes with the eponymous centiMorgan measurement, but it was still unclear whether DNA or protein maintained genetic information. At the time, protein was thought to be the heritable material of a cell because it was considered to be a more complex macromolecule. This resulted in protein being given a misnomer meaning "primary" or "first." Before a resolution of the DNA vs. protein debate could be reached, the one gene, one enzyme hypothesis (Beadle and Tatum, 1941) arose, linking genes to protein function and to the phenotype of the organism. It was not until after a few well-devised experiments, that DNA was confirmed as the inheritance molecule (Griffith, 1928; Avery *et al.*, 1944). Following this confirmation, another milestone in understanding the function of DNA was made when Watson and Crick deduced its structure (Watson and Crick, 1953). Shortly thereafter, the triplet nature of codons and the genetic code was understood (Nirenberg and Matthaei, 1961). Amazingly, all of these discoveries were made before there was an effective means of analyzing DNA sequences; however, it became readily apparent that knowing the composition of nucleic acid sequences would be paramount in understanding biology.

Two sequencing methods debuted around the same time (Sanger and Coulson, 1975; Maxam and Gilbert, 1977). These first-generation DNA sequencing technologies allowed researchers to interrogate long DNA fragments. Improvements to the Sanger method (Sanger, Nicklen, Coulson, 1977) quickly made it the more favored technique. Further modifications to Sanger sequencing, using expressed sequence tags as well as shotgun and automated sequencing, allowed researchers to uncover many human genes and the full sequence of organisms that had comparatively small genomes (Fleischmannet *et al.*, 1995; The C. elegans Sequencing Consortium, 1998). The grand undertaking of sequencing the human genome was laboriously accomplished using the Sanger method and its modified versions, providing the first draft, or reference, of the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001).

The reference genome

The goal of the Human Genome Project was to establish the sequences that make up the human genome and to identify the location and function of all genes within the genome. The complete sequence, now known as the human reference genome, was the haploid version of human sequences from the 22 autosomes and 2 sex chromosomes produced by *de novo* assembly. Assembly is the process of merging sequenced DNA fragments together to reconstruct the initial sequence. The current reference, or assembly, is called hg 19, build 37, the 19th iteration, and it is a composite of several individuals. There are an estimated 3.2 billion nucleotide bases (A: adenine, T: thymine, G: guanine, and C: cytosine) in the reference genome. One of the primary utilities of the reference genome is that it serves as a common point between sample genomes once they are aligned. Genome alignment is the process of matching DNA fragments from sample genomes to the reference genome. Having the reference allows the comparison between genomes without requiring *de novo* assembly with each new sample. Alignment is often preferred over assembly, which can be arduous, time-consuming, and highly variable between methods. Using an aligned sample, genetic variations can be observed by comparing the differences in the sample genome to the reference genome and, transitively, to other sample genomes. Finally, the reference can then be annotated with these variations and saved in personal or shared databases for later comparisons.

Types of Genetic Variations

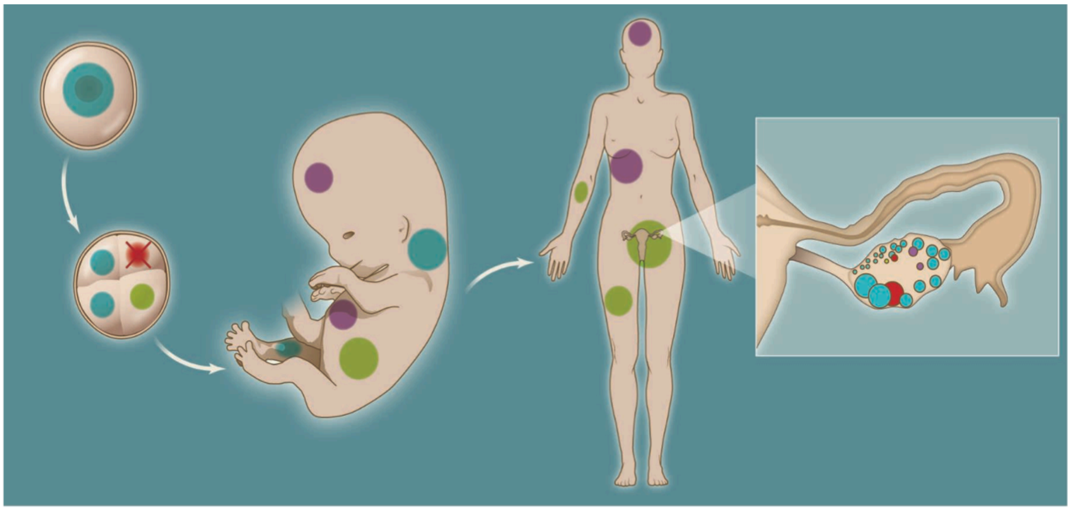
How do variations manifest?

The most common form of genetic variation in the human genome is also the smallest. Single nucleotide variations (SNVs) occur when a nucleotide at a distinct position, is converted to a different nucleotide relative to the reference genome. SNVs can result from mismatches during replication and/or chemical changes to a base. SNVs can affect the regulation of genes as well as the resulting gene products. When SNVs are present in at least 1% of the population, they are known as single nucleotide polymorphisms (SNPs). In any individual, there is on the order of one million SNPs (Sachidanandam *et al.*, 2001), equating to roughly one SNP every kilobase (kb). The next most frequent form of variation is small insertions and deletions (INDELs) of several bases (Mullikin *et al.*, 2000; Dawson *et al.*, 2001; Weber *et al.*, 2002). There has been a tremendous focus on SNPs and INDELs, as these smaller variations have been somewhat easier to characterize and detect compared to other variations. Larger variations will typically arise from more complex mechanisms, affecting larger segments of the genome, and can yield effects that may be more difficult to discern.

The class of variation that makes up larger events in the genome is called structural variation (SV). SVs are either copy neutral or change the number of copies of a particular DNA segment. Copy number variations (CNVs) are the class of SV, which are increases or decreases in the amount of DNA from a specific genomic locus (i.e., insertions, duplications, and deletions). Copy neutral forms of SV include inversions and balanced translocations, uniparental disomy, etc. (Feuk, Carson, Scherer, 2006). The largest variations consist of whole chromosomal alterations and can result in an aneuploidy state (e.g., non-disjunctions and unbalanced translocations). Aneuploidy is when the total number of chromosomes in a cell is not divisible by the haploid number of chromosomes for that organism. For instance, Trisomy 21 (or Down's syndrome) is an aneuploidy where an entire copy of chromosome 21 is gained due to missegregation in mitosis. Many mechanisms of SVs have been characterized, while others remain unknown (Onishi-Seebacher and Korbel, 2011). Generally, SVs are created by lapses in replication fidelity, mobile DNA elements, large-scale DNA damage, and the subsequent repair of that damage. These mechanisms and resulting SVs can be complex, most notably in cancer. When these variations are acquired during the lifetime of an individual, they are called somatic variations.

Somatic mosaicism

Variations of all sizes are classified as either germline or somatic (i.e., occurring in germline or somatic cells). Both germline and somatic variations can contribute to disease, phenotypic variability, and adaptation. Germline variation is present in every cell of an organism, while somatic variation occurs in non-gamete or non-reproductive cells and may not be seen in every cell (Youssoufian and Pyeritz, 2002; Lupski, 2013). An important distinction is that variations may arise by the same mechanism (i.e., homologous recombination) but will be either germline or somatic variations depending upon the context in which that variation occurs. Generally, germline variation is inherited and somatic variation is acquired. Germline variation must occur in gametes prior to fertilization and somatic variation occurs after the zygote, or fertilized egg, is formed. After the zygote is formed, all subsequent variants are somatic and will propagate to future daughter cells in a lineage-dependent pattern (Figure 1-1). Each round of mitotic division may contribute several somatic variations, creating multiple cell populations and lineages in a single organism. The presence of multiple cell populations in a tissue or



J.R. Lupski, *Science* (2013)

Figure 1-1

Figure 1-1. "Acquiring mosaicism." "Human development from a single fertilized cell to a multicellular organism requires many cell divisions and the genetic material to be replicated many times. Populations of cells (blue) can accumulate mutations at any stage in the life cycle (green, purple, and red). Some impair cellular fitness, and are consequently selected against (red cross); others survive and contribute to tissue mosaicism, which may serve physiological functions." (Lupski, 2013) individual is known as somatic mosaicism. These populations can arise early in development with estimates as great as 90% in some embryos (Mantikou *et al.*, 2012). While the incidence of mosaicism in a developing fetus may be high, it is unclear as to how many of these cells are viable because of cell competition (Figure 1-1). This is compounded by the fact that the biology of these early developmental variations is not evident, but they have the potential to influence phenotypes. While many of the causes and consequences of somatic mosaicism are unknown, there are two widely studied models of the phenomena: immune system adaptation and carcinogenesis.

The human immune system is a genetic mosaic. Adaptive immunity occurs through a process called V(D)J recombination (Brack *et al.*, 1978). In order to create diversity in immunoglobulin and T-cell receptors, immune cells induce somatic variations within their own genomes. This genetic shuffling permits the immune system to create new antibodies in order to recognize epitopes encountered on bacteria, fungi, viruses, and cancer cells so that they may be targeted for destruction. Ironically, cancer cells most often develop due to the accumulation of somatic variations. Outside of these two examples, the extent of somatic mosaicism is largely unstudied compared to germline variation. This has been primarily due to the fact that somatic variations are much more difficult to detect because they are usually present in only a subpopulation of cells analyzed.

Detecting Genetic Variations

The big and the small of it

Microscopic genetic variations have been widely characterized and routinely identified by karyotyping (Tjio and Levan, 1956; Jacobs *et al.*, 1959). While karyotyping methods have greatly improved since their inception, the resolution of detection has remained at larger than 3 Megabase (Mb) (Schaffer and Bejjani, 2004). This resolution is suitable for identifying the largest of genomic aberrations (e.g., non-disjunction and translocations events). An exception to this size cutoff may be made when using probes for known targets of specific sequence, but the sequence must be known *a priori*. Thus, other tools are required to detect events smaller than 3 Mb or at a submicrosopic level (Feuk, Carson, Scherer, 2006). As previously stated, first-generation sequencing technologies permitted detection of SNVs and small events (<1 kb). However, sample throughput was not sufficient to ascertain broad genomic structure in a meaningful way. The ability to analyze SVs was accomplished with the advent of second-generation sequencing technologies (Feuk, Carson, Scherer, 2006).

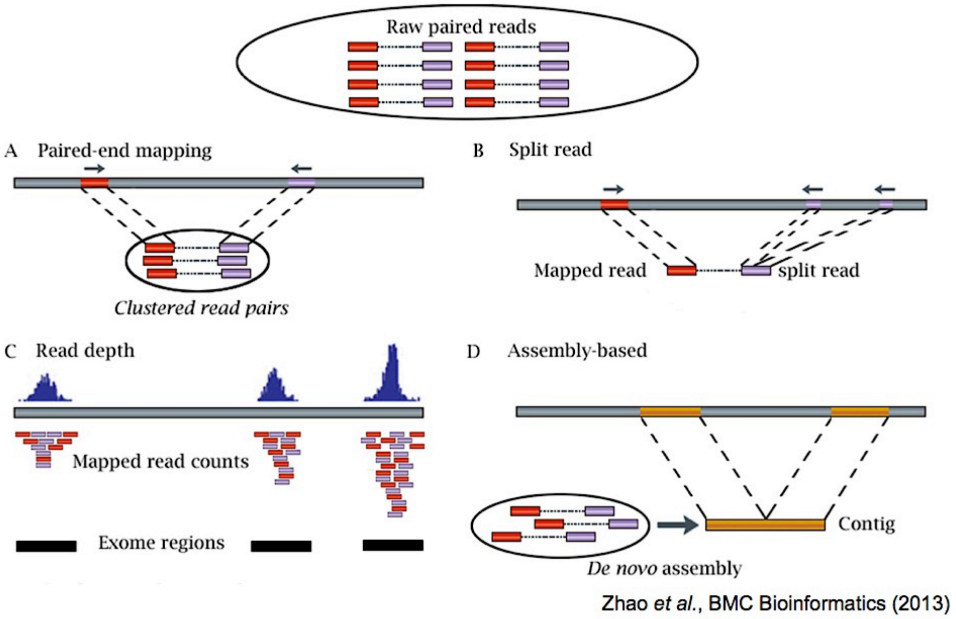
Lagom: in the middle

Second-generation sequencing platforms are capable of detecting variations of all sizes. These platforms derive their utility through excellent economy and high-throughput. Over the last decade, the cost of sequencing has continued to decrease at an increasing rate, and the volume of data produced in a single sequencing run is astronomical. This capacity for data production lends itself well to simultaneous whole genome sequencing, shedding light onto the prevalence of SVs in the human genome.

Today, the majority of second-generation sequencing data exists as paired-end (PE) reads. PE sequencing involves sequencing both ends of numerous DNA fragments from a collection of randomly generated fragments with an expected length. There are four prevailing computational SV detection strategies using PE reads: PE mapping (Figure 1-2A), split-read mapping (Figure 1-2B), read-depth (RD) analysis (Figure 1-2C), and local assembly (Figure 1-2D). SV detection tools use these strategies to find distinct sequence alignment "signatures," which then provide the location, as well as the type (e.g., deletion, inversion, translocation, etc.) of the rearrangement event (Alkan, Coe, and Eichler, 2011).

SV detection methods fundamentally rely on the reference genome. PE mapping analysis makes use of PE reads that have aberrant alignment configurations relative to the reference genome (Korbel *et al.*, 2007; Korbel *et al.*, 2009; Chen *et al.*, 2009). Alignments with unexpected mappings are called discordant read-pairs. For example, Figure 1-2A shows what could be inferred as a deletion in the sample genome based on reads mapping to the reference as discordant. Split-read mapping is the assessment of gaps in aligned reads, where there are subalignments of a read separated by large distances in the genome (Ye *et al.*, 2009; Abyzov and Gerstein, 2011; Wang *et al.*, 2011). Local assembly is the processes of creating longer reads by optimally combining treads to form "contigs." When used with the reference, assembly is particularly useful for refining SV locations (Quinlan *et. al.*, 2010; Malhotra *et al.*, 2011) and identifying novel sequences (Hajirasouliha *et al.*, 2010).

No single approach is comprehensive in its ability to detect SV. Each of the four approaches has distinct advantages and disadvantages, contextualized by the type of variation that one wishes to detect. As such, there are specific use cases for the application of each approach. In this work, principles of PE mapping, local assembly, and RD are used to detect variation in two specific situations. The first, found in Chapter 1, is a population-wide SV detection algorithm that incorporates data from many individuals to increase both sensitivity and specificity using PE mapping and local assembly. Second, RD is used in Chapters 3 and 4 to detect mosaic CNVs in single cells.



Zhao *et al.*, BMC Bioinformatics (2013)

Figure 1-2

Figure 1-2. Computational approaches to detect SV. In each approach, the grey bar (top) depicts the reference genome and the graphic (bottom) shows evidence of SV from the sample genome. **(A)** Paired-end (PE) mapping of aberrant or discordant read-pairs. The segment of the reference deleted compared to the sample, as the read-pairs align to positions in the reference greater than anticipated. **(B)** Split reads (purple) indicate a region was deleted from the sample genome. **(C)** The read depth (RD) approach detects CNVs by counting the number of reads mapped to each genomic region. In this instance, the third exon has been duplicated relative to the others because of the increased read counts observed. **(D)** The assembly-based method maps "contigs" to the reference genome. Shown is a portion of the sample genome, which has been deleted relative to the reference genome (Zhao *et al.*, 2013).

CHAPTER 2:
A MULTI-SAMPLE STRUCTURAL VARIATION DETECTION FRAMEWORK

This chapter is based on the following publications:
MR Lindberg, Hall IM, and Quinlan AR (in press). Population-based structural variation discovery with Hydra-Multi. *Bioinformatics.*

A Malhotra, Lindberg, MR, Faust, GG, Leibowitz, ML, Clark, RA, Layer, RM, Quinlan AR, and Hall IM (2013). Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Research*, *23* (5):762–776.

Abstract

Methods for SNP and INDEL discovery typically incorporate signals from multiple datasets in order to improve sensitivity and specificity. It is widely accepted that this approach could similarly enhance SV detection; however, its effective implementation has been stymied by the fundamental difficulties of SV calling (e.g., data volume, non-allelic states, and scalability). In recognizing this, we created a novel algorithm that efficiently integrates all available datasets and ameliorates the aforementioned issues. Here, we present the accuracy and speed of the *Hydra-Multi* algorithm using datasets from high-coverage tumor-normal pairs from The Cancer Genome Atlas (TCGA) and low-coverage individuals of diverse world populations from The 1000 Genomes Project (1KGP).

Introduction

The ever-increasing accuracy and affordability of sequencing technologies have brought forth a nascent understanding of human genomic structure (Durbin *et al.*, 2010). Interrogating chromosomal architecture in many individuals and samples has provided an operational map of SV in the human genome (Mills *et al.*, 2011), garnered insight into mechanisms of SV (Quinlan and Hall, 2012), uncovered intricacies of cancer genome structure (Stephens *et al.*, 2012), and challenged the canonical models of tumor development (Navin *et al.*, 2011; Stephens *et al.*, 2012).

Most SV discovery tools will compare a single sample genome to the human reference genome. However, tools that detect smaller genomic variations, namely SNPs and INDELs, have already shown improved sensitivity and specificity from including data from all available samples (Koboldt *et al.*, 2009; Lee *et al.*, 2010; McKenna *et al.*, 2010; Larson *et al.*, 2011). The superiority of this strategy is a marked decrease in the number of false negatives. This is accomplished by affirming weak evidence in one sample with supporting data from other individuals (Figure 2-1A and B). The strength of this approach is obvious, so why hasn't it been more widely adopted by SV detection tools? The primary limitations have been technical; SV calling is simply more complicated. As in the case of SNP detection, presence or absence of the event at a single base is the major concern. In detecting SVs, not only must the presence or absence be ascertained, but the type and size must also be distinguished. Additionally, the procedure for SNP calling is done by counting evidence for alleles in the alignment data at a particular locus. In contrast, SV calling must incorporate variability in both the alignment distances and sequencing libraries, as well as the primary alignment data when evaluating evidence. Not only do these difficulties make SV detection in a single individual more complicated, they tend to compound when incorporating multiple samples.

Despite the complexities of the approach, cancer and population genomics have a clear benefit from the ability to confidently assess SV across many samples. The typical cancer genome analysis only considers the genome of a single tumor and the matched normal tissue from the same individual when identifying SV. These standard tumor-normal comparisons are often fraught with somatic mutation false discoveries (i.e., a variant is predicted as somatic, when it is an inherited or germline variant). Somatic mutation false discoveries commonly occur when there is presence of SV evidence in the tumor and not the normal, but only because there was insufficient sequence coverage in the matched normal. Light physical coverage gained by common short-insert PE sequencing approaches can exacerbate the false discovery problem in the classical tumor-normal comparison. Therefore, a multi-sample SV discovery framework would be more scrupulous in its ability to avoid somatic false discoveries. A multi-sample framework can use the supporting alignments that exist in the genomes of other “normal” samples in the experiment when they do not exist in the matched normal (Figure 2-1C and D). Similarly, this approach can be used to reduce the false *de novo* mutation candidates when performing pedigree studies. Rather than incorporating sequence data from other matched normal, one would include other pedigrees or unrelated individuals.

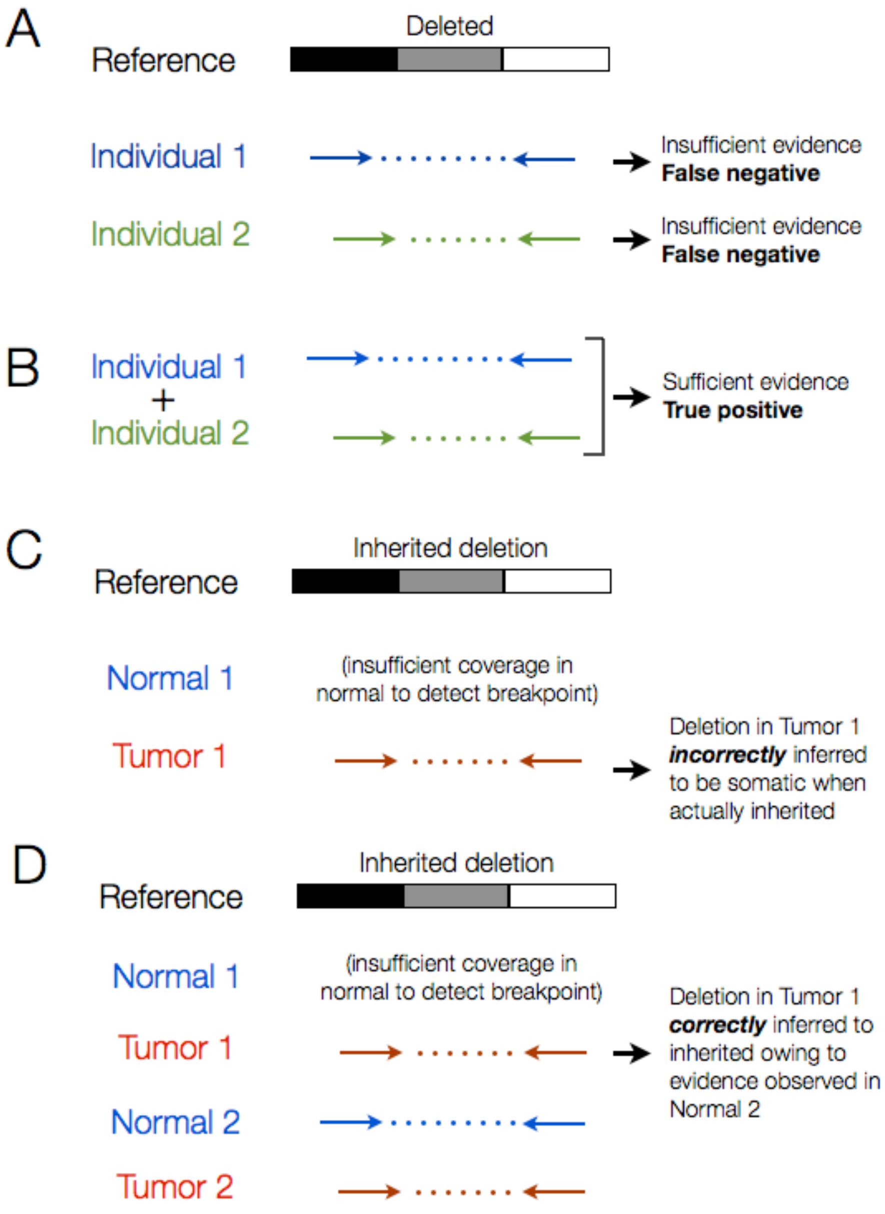

A
Reference
Deleted
Individual 1
Insufficient evidence
False negative
Individual 2
Insufficient evidence
False negative
B
Individual 1
+
Individual 2
Sufficient evidence
True positive
C
Reference
Inherited deletion
Normal 1
(insufficient coverage in normal to detect breakpoint)
Tumor 1
Deletion in Tumor 1 ***incorrectly*** inferred to be somatic when actually inherited
D
Reference
Inherited deletion
Normal 1
(insufficient coverage in normal to detect breakpoint)
Tumor 1
Deletion in Tumor 1 ***correctly*** inferred to inherited owing to evidence observed in Normal 2
Normal 2
Tumor 2

Figure 2-1

Figure 2-1. Conceptual overview of increased sensitivity conferred by multi-sample SV calling. (**A**) Owing to chimeric molecules and misalignments, all SV discovery algorithms set a support threshold dictating the minimum number of PE mapping alignments required to believe a putative event. As such, when SVs are identified individually in each sample, true SVs can be missed when there is insufficient evidence in a single sample. (**B**) However, when PE alignments are combined among samples, data can bed pooled to rescue real events that would otherwise have been missed. (**C**) When comparing a single tumor genome to its matched normal genome, insufficient evidence in the matched normal can cause what are actually inherited SVs to appear as somatic events. (**D**) Yet combining data from many tumor normal pairs can eliminate such somatic false discoveries by recognizing that, while missed in the matched normal, the event was observed in other normal samples indicating that the rearrangement is likely inherited.

Materials and Methods

The following methods are identical to those found in the manuscript accepted:
MR Lindberg, Hall IM, and Quinlan AR. (in press). "Population-based structural variation discovery with Hydra-Multi" *Bioinformatics*.

Experiments and analyses were developed and/or performed by MR Lindberg under the guidance of AR Quinlan and IM Hall unless otherwise noted. AR Quinlan wrote most of the actual software source code and accompanying scripts.

Data processing and implementation (performed with A Malhotra)

The data was obtained from TCGA and 1KGP available on dpGAP and the EBI/NCBI FTPs. In the analysis of TCGA datasets, data were processed slightly different than the standard *Hydra-Multi* analysis; a detailed explanation of how this was done can be found in Malhotra *et al*. and assumes that bam files may contain multiple sequencing libraries. Under this assumption, the sample statistics that are used to determine which read-pairs are proper pairs will be more accurate because they are evaluated by library. This is accomplished by several adjustments to parameters and changes to the overall workflow compared to the typical use of *Hydra-Multi*. 1KGP datasets analyzed separate from TCGA were processed using the more standard, user-friendly implementation, as described on the website: https://github.com/arq5x/Hydra, which maintains that bam files contain a single sequencing library.

1KGP accuracy performance analyses

From the 1KGP datasets, the NA12878 dataset (50x coverage from EBI) of the CEU family cohort and 64 random bam files (~5x coverage from NCBI) were used for performance analyses. The 50x NA12878 dataset was randomly subsampled by read-pairs to roughly 5x coverage, simulating a comparable 5x low coverage dataset for this individual. The 64 datasets were of approximately similar size, between 12 and 27 Gb, to prevent large differences in coverage from dramatically changing runtimes or SV support. A necessary step was the realignment of both NA12878 datasets (5x and 50x) using *BWA* (Li and Durbin, 2009) and duplicate removal with *SAMBLASTER* (Faust and Hall, 2014) with default parameters. Realignment of these datasets was necessary due to the different reference genome versions used for original alignment of 1KGP and Illumina Platinum datasets. These datasets were then used to compare the relative performance of *Hydra-Multi, GASVPro* v2.0 (Sindi. *et al.,* 2009), and an unpublished multi-sample version of *DELLY* v0.5.3 *(*Rausch *et al.,* 2012, https://github.com/tobiasrausch/delly). In these comparisons, the number of true positives and false positives were measured for each tool in three scenarios: 1) analyzing NA12878 at 5x coverage by itself, 2) NA12878 at 50x coverage by itself, and finally 3) NA12878 (5x) with the 64 random datasets. To measure the true and false positive rates, we used a truth set consisting of 3,077 non-overlapping validated deletions in NA12878 from the Mills *et al.* study.

Putative calls from *Hydra-Multi*, *GASVPro*, and *DELLY* were generated in each of the analyses using similar parameters for each tool. In all analyses, a minimum mapping quality of 20 was required for each read. The input deviation parameter values across all tools were made to be equivalent. As such, the deviation parameters for *Hydra-Multi* and *DELLY* were set to 5 and 8 MADs for each of the respective analyses. The corresponding computed *Hydra-Multi* values at these settings (5 and 8 median absolute deviations or MADs) were used as the input *GASVPro's "*LminLxmax" values. This was necessary because *GASVPro* calculates fragment size distributions using standard deviations from the input parameter, while the other two tools use MADs. Standard deviations can be significantly larger than MADs, making the comparison incongruous. The tool specific "punt" parameter for *Hydra-Multi* was set to a value corresponding to 5 times the summed mean dataset coverage (e.g., 25 for one 5x dataset, or 250 for ten 5x datasets). At the outset of variant calling, a minimum support of two read-pairs (read-pairs and/or split-reads for *DELLY*) was required for discovery. Because all three tools report breakpoints in three different output formats, the outputs for all tools were converted into BEDPE format using zero-based, half-open arithmetic to make a fair comparison. The *GASVPro* regions file reports the breakpoint boundary points of the final polygons constructed. The midpoint the left and right boundaries were calculated and padded with 100 bp in both directions to create a set of two 200 bp intervals. Next, the single base start and end coordinates reported by *DELLY* in VCF format were given 100 bp of slop in both directions to make two intervals of 200 bp. The final reported *Hydra-Multi* breakpoint footprints were made into breakpoint intervals by drawing two 200 bp intervals inward from the read-mapping "footprints" (i.e., end1 + 200 bp and start2 – 200bp). In each conversion to BEDPE intervals, the number supporting read-pairs were tracked; however, this was not possible in the analysis using 64 datasets and NA12878 at 5x. First, *GASVPro* is not currently a multi-sample caller and therefore could not be evaluated in a multi-sample analysis. For *DELLY*, the presence or absence of a call in NA12878 from the multi-sample analysis was assessed by the reported genotype (GT) or whether NA12878 contributed at least one high-quality variant pair (DV) to the breakpoint. GT is equal to DV in a single sample with support of 4 – 10 read pairs, thus only GT was reported For *Hydra-Multi*, a call was considered present in NA12878 when at least one original read-pair from NA12878 was used in making the final breakpoint call. Since both *DELLY* and *Hydra-Multi* report the total read support across all datasets, this was used as a filtering criterion in the ROC curves.

All putative breakpoint interval sets and the truth set were filtered to remove GL, MT, and Y chromosome calls, as well as any interval that overlapped with a set of exclude regions (https://github.com/cc2qe/speedseq/blob/master/annotations/ceph18.b37.lumpy.exclude.2014-01-15.bed). These exclude regions are comprised of very high sequence coverage in the 17-member CEPH 1463 pedigree. The RD at these locations was greater than 2*mode + 3 standard deviations, as found by aligning with *BWA MEM* (Li, 2013) and measuring the depth with *BEDTools* (Quinlan and Hall, 2010), whereby the autosomal and sex chromosomes were analyzed separately. The filtered call sets from each of the three tools (both 5 and 8 MADs) were evaluated at varying levels of total support (4 to 10 read-pairs) in three scenarios. True positive were calculated by finding the number of uniquely identified truth set deletions. An identified truth set deletion was one with which the two truth set breakpoint intervals intersected with any pair of breakpoint intervals reported in a given comparison. A false positive was defined as the number of SVs reported by a tool that did not identify a truth set deletion.

TCGA tumor-normal false discovery rate estimations

Given c = n choose k, false discovery rates were calculated by using combinations of the total number of tumor-normal pairs, n, as more pairs, k, are considered across the set of high-confidence breakpoints found in the TCGA datasets. For practical purposes, c was limited to 1000 random combinations due to it being very large for most values k. Therefore, c = 64 choose k for k = 1 to 64, where c ≤ 1000, which limits the number of combinations explored to 1000. Thus, for each c, false discoveries were determined by the presence or absence of somatic and germline in all breakpoints. A breakpoint is said to be a false discovery for any c if it is seen in a normal sample genome of any one tumor-normal pair and it is not in it's matched tumor sample or another tumor-normal pair. Each breakpoint was tested to be a false discovery in all combinations c and the false discovery rate at each value k was calculated. The false discovery rate was the ratio of the total private normal sample breakpoints to the total private tumor sample breakpoints in each k.

1KGP speed and scalability performance analyses

Maximum memory and runtimes were determined with the runit utility (https://github.com/lh3/misc/tree/master/sys/runit). The three previously described deletion detection scenarios were included in the speed and scalability performance analyses. The maximum memory usage and total runtime for the three scenarios with 8 MADs were recorded. An additional analysis to simulate a large number of input datasets for *Hydra-Multi* and *DELLY* consisted of repeating instances of the NA12878 datasets. In all speed and scalability measurements, *Hydra-Multi* was allocated 8 processors, *DELLY* was permitted up to 32 threads, and *GASVPro* was given 20 Gb for the Java Virtual Machine. *Hydra-Multi*'s punt parameter was also adjusted to 5 times the summed mean input dataset coverage. Runtime measurement of each processor usage versus the number of datasets as well as discordant read-pairs was done using 1KGP datasets. The previous 64 datasets were sequentially subsampled at random 3 times (e.g., n = 1, 2, 4, 8, 16, and 32) for each benchmark.

Hardware specifications and utilities

All analyses were done using a single compute node running CentOS 2.6.32-358.2.1.el6.x86_64 on 16 Intel Xeon E5-2670 CPUs with 128 Gb random access memory and an array of twelve 4 Tb hard disks spinning at 7200 rpm. All interval intersections were done with *BEDTools* (Quinlan and Hall, 2010).

Hydra-Multi Algorithm (development by AR Quinlan)

Set up and configuration

A configuration file is necessary to provide a unique identifier and file path defining the appropriate alignment file for each DNA library (or sample) interrogated for SV breakpoints. Before running *Hydra-Multi*, it is imperative that these bam files have duplicate reads (molecules) removed. The configuration file must define i) a central tendency statistic for the insert size for the library – *mean* or *median* are recommended, ii) a standardized measure of the variance in the library – typically the *standard deviation* or *median absolute deviation*, and iii) the number of units of variance (e.g., “6” for six standard deviations) that should be used to define a proper pair or *concordant* alignment when pairs align in the +,- (forward, reverse) orientation within expected genomic space.

Routing discordant alignments by genomic coordinates and strand

The algorithm begins by routing similar *discordant* alignments (i.e., PE mappings with either an unexpected insert size (greater than the central statistic + number of variance metrics) or an aberrant strand combination (-,+;+,+;-,- orientations), from all libraries (or samples) defined in the configuration file to common alignment files using the hydra-router tool). In effect, this step consolidates all mappings from each of the libraries (or samples) into discrete files that are likely to support the same breakpoints. For example, all PE mappings that indicate deletions on chromosome 10 would be place in a file called *chr10.chr10.+.-*. The benefit of this routing step is that each discrete file of discordant mappings can be processed independently, since the mappings therein can only support specific types of SV breakpoints involving the defined chromosomes. In turn, this independence facilitates a high degree of parallelism when screening for candidate breakpoints.

Identifying candidate breakpoint "clusters"

Each routed file of discordant alignments is then sorted by each mapping's "leftmost" start position (that is, the position with the lowest start coordinate). Sorting in this manner organizes the mappings from each library (or sample) such that alignments supporting the same breakpoint are aggregated. In order to maximize sorting efficiency, we developed a custom C++ implementation (https://github.com/arq5x/kway-mergesort) of the k-way, memory assisted merge-sort algorithm. Importantly, as sorting the discordant mappings is one of the more computationally intensive steps of our approach, this approach combines the benefits of a disk-based merge-sort algorithm with the speed of in-memory sorting. This allows the user to define precisely how much memory should be allocated to the sorting step and avoids prohibitive memory consumption while providing fast sorting times.

Greedy breakpoint reconstruction

After sorting mappings based on their start position, it is common for mappings supporting different rearrangement events to be sorted together, especially when integrating data from hundreds of samples and/or from highly rearranged genomes. *Hydra-Multi* addresses this in two phases. First, it scans each sorted file to build clusters with mappings that have the potential to support the same breakpoint based on the supplied variance statistics. A cluster is terminated once an a mapping is encountered whose start coordinate is to the "right" of the current cluster's rightmost end coordinate; by definition, such a mapping cannot support the same breakpoint as the mappings already in the cluster. The clusters are then sorted by relative support and assembled into "contigs" which represents a larger interval of assembled space in the genome. Once assembled, the contributing cluster is removed from the pool of usable clusters that can corroborate subsequent breakpoints in further assemblies. A cluster may also be terminated in regions of overly complex genomic rearrangements. These regions can cause excessive runtimes and may be averted by "punting" once a certain number of mappings have been attributed to a putative cluster. A reasonable heuristic for this parameter is 5 multiplied by the summed mean coverage of the datasets to be analyzed (e.g., 250 for ten 5x genomes)

Punt parameter

The purpose of punting is to avoid needless increases in runtimes caused by the unnecessary analysis of regions that appear to have highly rearranged segments. These regions are often artifacts native to the genome assembly, but they may also be bona fide highly complex genomic rearrangements. As such, misassembles inherent to the genome and overly complex rearrangements will often lack the need and/or the ability to be interpreted. Under the provided usage parameters, nearly all germline variants will be included, but some somatic variants will be susceptible to being discounted. We stress that the punt parameter is tunable to the user's experimental design; however, the recommended value is viable in most cases. The user also has the option to perform an exhaustive analysis by trading for an increase in total runtime. We recommend 5 times the summed mean coverage of all datasets in an analysis. For example, in an analysis of 10 cancer genomes sequenced to 50x coverage, the punt parameter would be 2500. This setting should be able to detect an average (depending on ploidy) of about 10 copies per genome or roughly 100 copies in total. In the unlikely event that all genomes in the analysis contain a particular segment amplified more than 10 times, this variant will be punted and go undetected. However, it is more commonly seen that a single dataset or collection of datasets will contain the highly amplified region, while others do not. Therefore, in a more likely scenario where two genomes have 40 copies of a region and all other genomes contain 2 copies, this amplification will be detected. Thus, the punt parameter, like many other features of *Hydra-Multi*, benefits from the population-based framework, that is adding more datasets will temper the effects of any single dataset on incorrectly punting a bona fide variant.

Results

Here, we introduce *Hydra-Multi*, SV discovery software that simultaneously incorporates DNA sequence alignments from many individuals. *Hydra-Multi* is capable of determining the presence or absence of a genomic rearrangement in individual samples with high sensitivity. We illustrate a large reduction in somatic misclassifications in 64 tumor genomes from TCGA and we show benchmark comparisons of *Hydra-Multi* to other state-of-the-art algorithms.

Origins of the approach

Hydra-Multi is an extension of our PE mapping SV caller, *Hydra*, which was designed to detect SV in a single genome (Quinlan *et al.*, 2010). Seeing the limitations of the single genome approach, *Hydra* was first used in a 2011 study of genome instability in mouse induced pluripotent stem cell lines (Quinlan *et al*., 2011). While bearing the conceptual framework for multi-sample SV calling, the implementation of *Hydra* in that study was not able of analyzing hundreds to thousands of samples nor capable of incorporating variably sized DNA libraries within or between samples. *Hydra-Multi* is specifically designed to ameliorate these weaknesses and integrate data from multiple individuals, conferring increased sensitivity for SV detection and sample genotyping.

Other investigators have since recognized the advantages in multi-sample SV calling. In 2011, Handsaker and colleagues developed a framework which genotypes deletions. Deletions are discovered in one or more samples and then subsequently examining in all other samples in order to infer genotypes, rather than at the outset of SV calling (Handsaker *et al.*, 2011). This approach can suffer from a lack of sensitivity when there is insufficient coverage in any one sample and the threshold for confidently detecting is not met. This *post hoc* assessment of breakpoints does not perform well on rearrangements arising from repetitive or duplicated DNA because the alignments present in one sample may differ from those present in another sample. Since *Hydra-Multi* has the same design principles of *Hydra*, it can evaluate multiple alignments from all samples, thereby improving sensitivity for both repetitive and unique genomic elements. Hormozdiari and colleagues developed a new multi-sample version of *VariationHunter* later that year. This version yielded increased sensitivity, but it is designed for a small set of related genomes; therefore, it does not scale well to many genomes (Hormozdiari *et al.*, 2011). Lastly, the best contender is an unpublished version of *DELLY* (Rausch, T. *et al.*, 2012), which operates on several datasets and performs quite well, but it is eclipsed by the scalability performance of *Hydra-Multi.*

Overview of computational framework

Hydra-Multi is intended to be an easy-to-use and effective software package for SV discovery among many samples in order to improve SV detection sensitivity. First, *Hydra-Multi* consults a configuration file, indicating the DNA fragment library statistics (e.g., median and median absolute deviation) for all of the sample alignment files. *Hydra-Multi* is then capable of recognizing corroboration between and within samples for the same SV, despite variability in absolute mapping distances of discordant alignments.

The algorithm then separates all of the alignments from each sample by the chromosome and alignment orientation observed from each of the pairs (e.g.*,* all paired alignments with +/- orientation on chromosome 1) based on the sample statistics.

This will cull the sets of alignments that have the potential to support each rearrangement class (e.g. deletions, inversions, etc.) on a given chromosome (or pair of chromosomes) (Figure 2-2). Alignments in each chromosome/orientation group are subsequently sorted by their chromosomal coordinate. To remain within the memory constraints of typical commodity computing hardware, we implemented a k-way merge-sort algorithm to permit population-scale SV discovery whereby hundreds to thousands of samples may be analyzed. By sorting discordant alignments by chromosome coordinate, the discovery algorithms can then search for clusters of aberrant alignments that support a common SV breakpoint by performing a "sweep" from the beginning to the end of a chromosome. After identifying clusters, we use a greedy algorithm to "assemble" a single breakpoint call, whilst keeping track of the sample and library of each supporting alignment. *Hydra-Multi* can then report the number of supporting alignments observed in each sample for each breakpoint call.

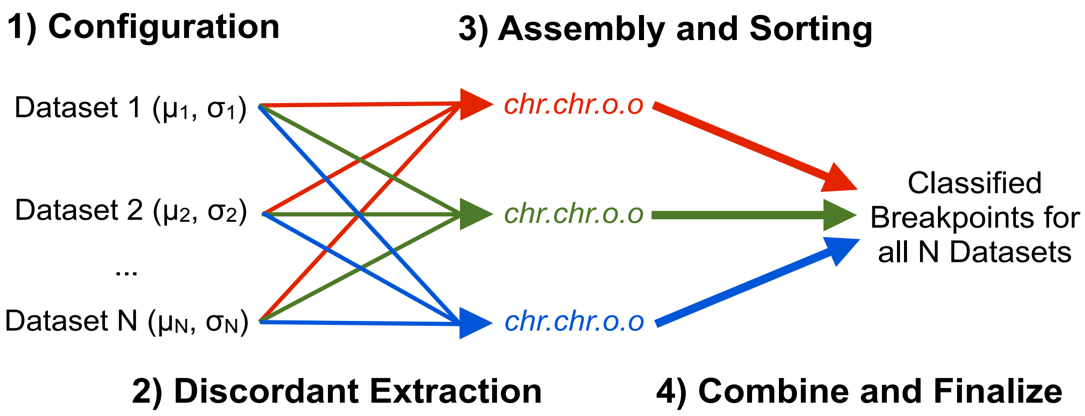


Figure 2-2

Figure 2-2. Hydra-Multi workflow. The algorithm consists of the following steps: configuration, discordant alignment extraction, assembly and sorting, and combining for SV breakpoint finalizing. The extraction of discordant reads, as well as assembly and sorting occur in parallel.

Somatic mutation discovery in 129 datasets from TCGA

One of the main objectives of cancer genomics is to find somatic mutations that either initiated the carcinoma or developed during tumor progression. Thus, tools that discriminate somatic from inherited mutations are valuable, due to the cost and time associated with pursuing spurious calls; higher fidelity predictions can provide increased focus on the set of variants that are more likely to be phenotypically relevant and/or clinically actionable.

The canonical method for studying somatic mutations in tumor genomes is to do a one to one comparison of the mutations found in the tumor genome to those observed in normal tissue from the same individual. Unfortunately, mutations could be present in the normal genome and may have been missed because of insufficient coverage. Many studies have tried to address this weakness by deeply sequencing the matched normal. Here, we show that despite deep coverage for matched normal sample, inherited mutations are still frequently misinterpreted as somatic tumor mutation. Further, the *Hydra-Multi* framework can be used to substantially reduce the somatic false discovery rate (FDR) by incorporating data from other pairs of tumor and matched normal genomes.

Hydra-Multi was used to interrogate the mechanisms that drive complex genomic rearrangements (Malhotra *et al.*, 2013). This study was made up of 12 invasive breast cancers, 3 colon adenocarcinomas, 18 glioblastomas, 6 lung adenocarcinomas, 13 lung squamous cell carcinomas, 11 ovarian cancers, and 2 renal adenomas. This study totaled 129 PE sequencing whole genome datasets (64 tumors and 65 matched normal tissues) from TCGA. The discordant alignments were between 1 and 3 percent of datasets and were processed as described previously (Quinlan *et al.*, 2011; Malhotra *et al.*, 2013). We applied standard filtering measures, found in Malhotra *et al.*, excluding breakpoints in highly repetitive or misassembled regions of human genome. The final set of high-confidence SV breakpoints for the 129 genomes was 33,218 in total.

Over 80% (27,039) of these breakpoints form Malhotra *et al.* were classified as germline meaning they were found in at least one of the matched normal genomes and most likely did not arise in a tumor genome. Owing to the fact that each tumor genome evolved from the normal somatic genome from the same individual, our expectation is that the distance between all 129 tumor and normal genomes would find that the tumor and normal genomes from the same individual are most closely related. To test this anticipation, we used previously described clustering strategy on 11,944 high-quality germline deletions and duplications that are smaller than 1 Mb in size. For each germline SV, we included 129 columns reflecting the presence or absence of the breakpoint in each tumor or normal sample. Meeting our expectation, each tumor-normal pair is most closely related to one another when using germline breakpoints as a measure of genetic distance (Supplementary Figure S2-1).

Estimating the somatic mutation false discovery rate using 64 tumor-normal pairs

There were 6,502 structural rearrangements private to a single sample in Malhotra *et al.* study. Furthermore, over 95% (6,179) of these breakpoints were isolated events in an individual tumor genome, confirming the notion that solid tumor genomes are known harbor many rearrangements (Gandi *et al.*, 2010). The other 323 (5%) breakpoints were unique in the genome of a single matched normal sample. Using the assumption that all variants private to a normal genome are spurious, we can infer that the FDR for mutations specific to a single tumor was 5.2% (323 / 6,179) as found in Malhotra *et al.*. This estimation may be high, as the fraction of the normal only variants are probably real due to a loss of heterozygosity in the tumor. Furthermore, approximately half of the 5.2% of false positives are relatively small; these variants are more than likely misclassified due to differential resolution amongst samples from varying insert size distributions. However, a prediction of 5.2% is only possible by integrating data from all tumor and normal pairs simultaneously. If we had predicted somatic SVs using the common practice of comparing each tumor individually to its matched normal, about 89.1% of the predictions would have been false. Moreover, performing single-sample variant calling separately on all 129 genomes, 21.9% of somatic SV calls would have been incorrect (versus 5.2% for joint calling). The rate of somatic FDR is greatly decreased as more tumor-normal pairs in discovery (Figure 2-3). This argues that cancer genomics studies can reduce erroneous calls by adopting this variant detection strategy or directly using *Hydra-Multi*.

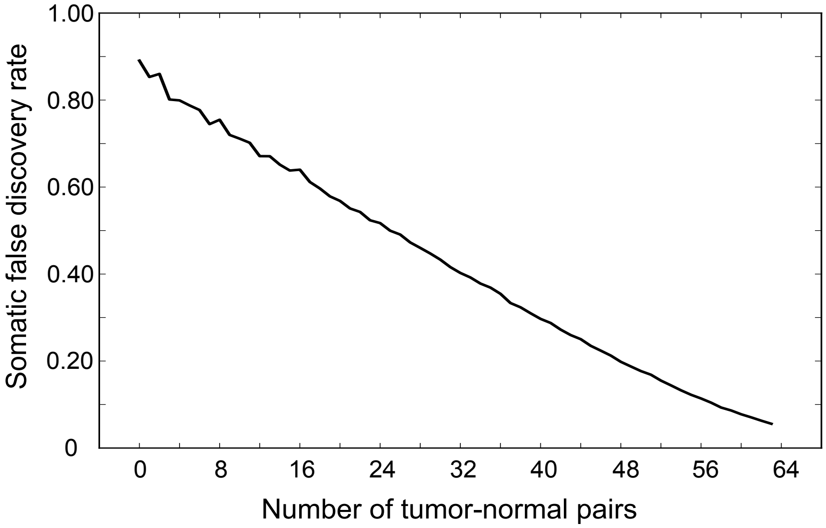


Figure 2-3

Figure 2-3. Reduction in the somatic SV FDR for tumor-specific mutations by simultaneously integrating data from 128 TCGA samples. The somatic FDR is the predicted rate at which somatic SV breakpoints are false, either due to false positive SV calls or due to inherited germline SVs that have been misclassified as somatic due to false negatives. For this experiment, we identify false somatic calls by their presence in a single normal genome but not in the paired tumor genome or any of N additional tumor-normal pairs (X-axis).

Deletion detection accuracy from NA12878

We compared two widely-used SV discovery tools, *GASVPro* (Sindi *et al.*, 2009) and *DELLY* (Rausch *et al.*, 2012), against *Hydra-Multi* to evaluate relative accuracy and SV detection performance. These two methods have been used in the analysis of large-scale datasets from TCGA and 1KGP and outperform other extant methods. Multi-sample variant calling is a relatively new and unpublished feature of *DELLY*, and the current version of *GASVPro* is not capable of multi-sample calling. Each tool's ability to detect deletions was measured by analyzing NA12878 from the 1KGP CEPH (or Utah residence with Northern and Western European ancestry) population in three typical situations (Figure 2-4). The analysis was limited to detecting deletions in NA12878 as there is not a reliable truth set for hundreds to thousands of samples. As such, a trusted set of 3,077 validated, non-overlapping deletions in NA12878 were used (Mills *et al.*, 2011).

Overall *DELLY* was found to have the highest sensitivity and specificity when analyzing a single dataset alone (Figure 2-4A and 2-4B); however, *Hydra-Multi* has the best performance in a multi-sample analysis (Figure 2-4C). *DELLY's* marginal superiority in analyzing a single dataset is not surprising as it utilizes several signals during SV discovery. In contrast, *Hydra-Multi* and *GASVPro* use PE alignments alone. Further, *Hydra-Multi* was specifically designed to do multi-sample analyses, and it has a greatly improved sensitivity in this use case. Nevertheless, *Hydra-Multi* has competitive performance in single dataset usage scenarios, attaining near parity with *DELLY* and besting *GASVPro* in most cases. For the single dataset comparisons (Figure 2-4A and 2-4B), the true positive rates were fairly consistent across all tools and performance primarily deviated by the number of false positives. A high false positive rate was seen for all tools using the minimum evidence parameters, but this permitted much higher sensitivity. False positive ranges under stricter settings suggest that this can be assuaged through filtering and parameter tuning. This underscores the well-known problem of performing both sensitive and accurate SV detection using short-read sequencing data. However, the 1KGP truth set from Mills *et al.* is incomplete and the number of false positives is therefore an upper bound estimate.

Co-analyzing multiple samples is clearly beneficial, as there is dramatic improvement in SV detection sensitivity for both *Hydra-Multi* and *DELLY* when 5x NA12878 data is jointly analyzed with 64 additional 5x genomes (Figure 2-4C) when compared to the 5x NA12878 in isolation (Figure 2-4A). *Hydra-Multi* has substantially higher sensitivity than *DELLY* in this comparison, but this also comes with a tolerable increase in the number of false positives. In summary, *Hydra-Multi* is competitive with other best-in-class SV detection tools when run on a single dataset by itself, and *Hydra-Multi* eclipses all other tools in multi-sample SV calling.

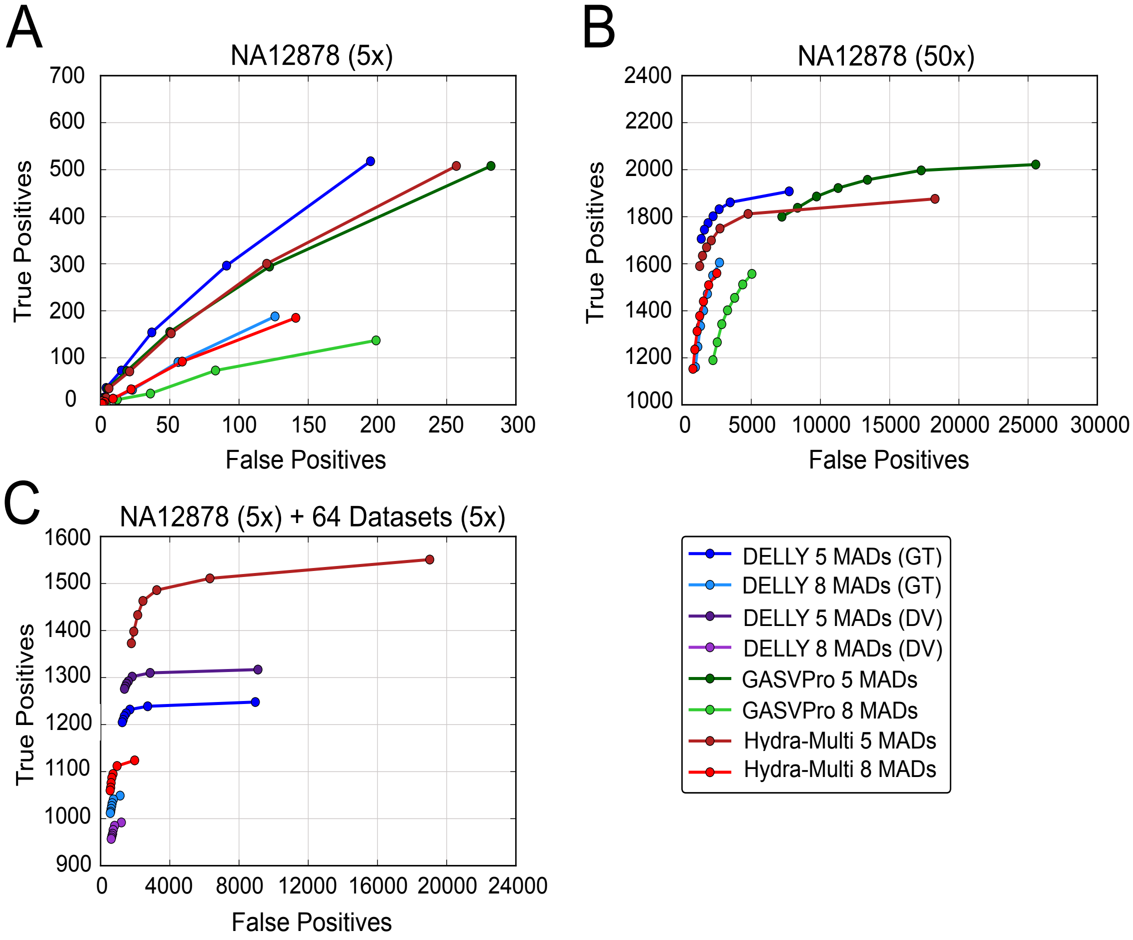

A
NA12878 (5x)
True Positives
0
100
200
300
400
500
600
700
0
50
100
150
200
250
300
False Positives
B
NA12878 (50x)
True Positives
1000
1200
1400
1600
1800
2000
2200
2400
0
5000
10000
15000
20000
25000
30000
False Positives
C
NA12878 (5x) + 64 Datasets (5x)
True Positives
900
1000
1100
1200
1300
1400
1500
1600
0
4000
8000
12000
16000
20000
24000
False Positives
DELLY 5 MADs (GT)
DELLY 8 MADs (GT)
DELLY 5 MADs (DV)
DELLY 8 MADs (DV)
GASVPro 5 MADs
GASVPro 8 MADs
Hydra-Multi 5 MADs
Hydra-Multi 8 MADs

Figure 2-4

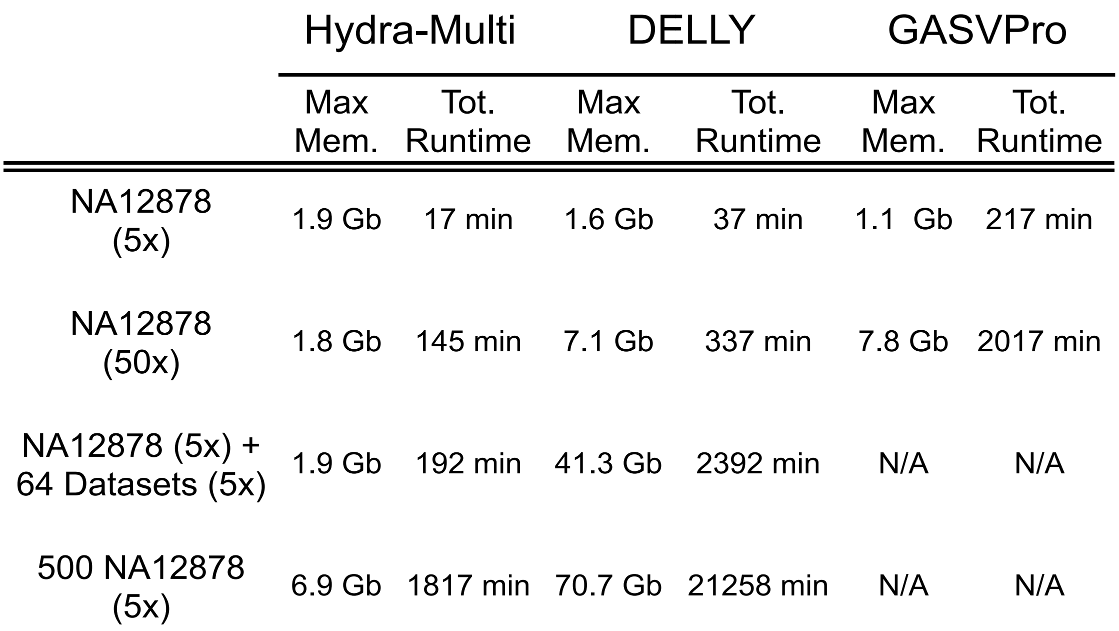
Figure 2-4. Receiver operating characteristic (ROC) curves describing deletion detection in NA12878 from three scenarios. The relative accuracy of *Hydra-Multi* (red) was compared to both *DELLY* (blue; GT or purple; DV) and *GASVPro* (green) in three analyses that each compared fragment size parameters of 5 and 8 MADs (See Methods and Materials). Each plot displays the relationship between the number of true and false positives at varying levels of minimum alignment support (4 -10 read-pairs). A true positive was defined as detection of one of the 3,077 non-overlapping truth set deletions where both intervals from a predicted deletion breakpoint intersected with both of the truth set deletion breakpoint intervals. In order to make a fair comparison across all tools, each predicted breakpoint was represented as two 200bp intervals that faithfully represent the region implicated by the original SV call. A list of regions to exclude based on excessively high read-depth were used on both the truth set and putative call sets. The three situations used to assess the three tools are as follows: **(A)** The 50x NA12878 dataset was subsampled to 5x and analyzed. **(B)** The 50x NA12878 data was analyzed. **(C)** The subsampled 5x NA12878 dataset was analyzed concurrently with 64 randomly selected datasets of ~5x coverage from 1KGP. Total support was evaluated as the total number of read-pairs across all datasets analyzed. The presence of a deletion in NA12878 by *DELLY* was inferred by the reported genotype (GT) and/or by observing at least one high-quality variant pair (DV) in NA12878. Only GT was reported in the single dataset analyses, as GT and DV are functionally the same when requiring 4 – 10 read-pairs of support. In both single and joint analyses using *Hydra-Multi*, the contribution of at least one read-pair by NA12878 was required. **Note:** *GASVPro* does not simultaneously run on multiple datasets.

Scalability and performance

Hydra-Multi was designed around having fast runtimes and scalable performance, and it significantly outperforms other tools in these measures. In the same scenarios presented in Figure 2-4, *Hydra-Multi* ran 2-13x (2.2, 2.3, and 12.5x) faster than *DELLY* and 12-14x (12.8 and 13.9x) faster than *GASVPro*, and required merely 3.2 hours to analyze the set of 65 5x datasets (Table 2-1), whereas DELLY required 39.9 hours. *Hydra-Multi* is capable faster runtimes while simultaneously using a substantially smaller memory footprint than the other tools: for example, in the 65 dataset comparison (Table 2-1), *Hydra-Multi* used merely 1.9 Gb of memory while *DELLY* used 41.3 Gb, which is nearly a 22-fold difference. *Hydra-Multi's* lesser resource requirements also allows for a much larger number of datasets to be jointly analyzed on a single machine. Generally, each additional sample included into analysis improves the overall variant detection sensitivity. A simulation of a large input dataset experiment was done using 500 repeated inputs of the 5x NA12878 dataset, yielding tractable runtime (~30 hours) and memory usage (6.9 Gb) for *Hydra-Multi* on a single commodity server with 128 Gb of RAM. In comparison, *DELLY* requires more than two weeks and >70Gb of RAM to analyze the 500 NA12878 datasets (Table 2-1).

The small memory footprint incurred by *Hydra-Multi's* is garnered mostly by using a memory assisted, k-way merge sorting algorithm and its speed is obtained largely through parallelization at the discordant extraction and assembly steps (Figure 2-2). Coarse parallelization is used for extraction and assembly (i.e., one processor for each dataset and chromosome/orientation set, respectively). With recommended parameters, discordant PE extraction predominates algorithm and scales near linearly with input data when given a single processor (Supplementary S2-2). Supporting this assertion is the direct relationship between the number of discordant PE reads and runtime (Supplementary S3-2). The cost of examining additional data is ameliorated by parallelism. Scalability is a central strength of *Hydra*-Multi, which is gained through the disk-based sort and parallelization, enabling incorporation of an extremely large number of datasets for SV discovery.

With the ever-increasing number of genomes sequenced and the rapid accumulation of whole-genome sequencing data, the benefits of joint sample variant discovery will become evermore evident. *Hydra-Multi* is perfectly poised for use on very large-scale projects.



	Hydra-Multi		DELLY		GASVPro	
	Max Mem.	Tot. Runtime	Max Mem.	Tot. Runtime	Max Mem.	Tot. Runtime
NA12878 (5x)	1.9 Gb	17 min	1.6 Gb	37 min	1.1 Gb	217 min
NA12878 (50x)	1.8 Gb	145 min	7.1 Gb	337 min	7.8 Gb	2017 min
NA12878 (5x) + 64 Datasets (5x)	1.9 Gb	192 min	41.3 Gb	2392 min	N/A	N/A
500 NA12878 (5x)	6.9 Gb	1817 min	70.7 Gb	21258 min	N/A	N/A

Table 2-1

***Table 2-1. Memory usage and runtime performance from four scenarios*.** The relative speed and scalability of *Hydra-Multi* was compared to the other tools by measuring the maximum memory used per process and runtime with Runit (https://github.com/lh3/misc/tree/master/sys/runit). *Hydra-Multi* (8 processors) and *DELLY* were parallelized (32 threads). *GASVPro* ran as a single process/thread, never exceeding the Java Virtual Machine allocation of 20 Gb. From top, we analyzed the following datasets: a 5x NA12878 dataset obtained by subsampling the 50x NA12878 dataset; the 50x NA12878 dataset; the 5x NA12878 dataset combined with 64 additional ~5x datasets from 1KGP; 500 copies of the 5x NA12878 dataset. **Note**: *GASVPro* cannot jointly analyze multiple datasets (indicated by "N/A").

Discussion

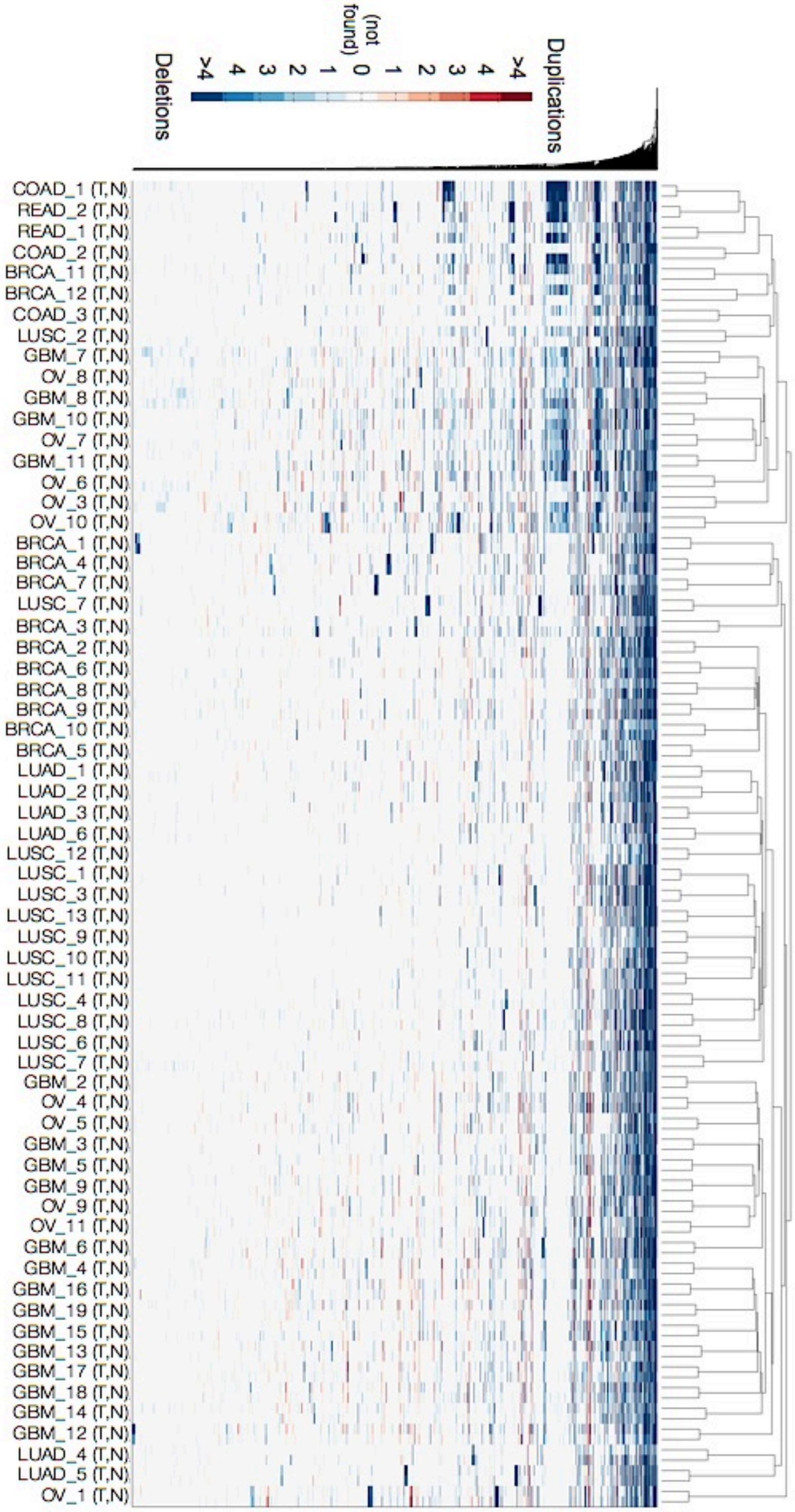
Hydra-Multi is the first SV discovery framework to employ the strategies used by SNP and INDEL discovery tools such as the *GATK* (McKenna *et al.*, 2010), *SAMTools* (Li, 2009), *MoGUL* (Lee *et al.*, 2010), *VarScan* (Koboldt *et al.*, 2009), and *FreeBayes* (Garrison *et al.*, unpublished). The software maximizes discovery sensitivity by combining PE sequencing alignments from hundreds of individual genomes. While *Hydra-Multi* solely examines PE alignment evidence for SVs, it maintains high sensitivity by combining information across many samples. Therefore, it is tailored to perform well on many samples with low genome coverage, whereas poor sensitivity would be garnered when identifying SVs in each sample individually.

We have demonstrated the strength of the approach in analyzing 64 tumor-normal genomes whereby the number of somatic misclassifications was dramatically reduced. Additionally, *Hydra-Multi* is sensitive and fast when compared to current tools such as *DELLY* (Rausch *et al.*, 2012) and *GASVPro* (Sindi, S. *et al.*, 2009), showing its aptness and scalability for many samples. Finally, *Hydra-Multi* can be used to detect SVs in tumor datasets, as well as discern the whether the variant is somatic or germline. This utility is advantageous when identifying spontaneous mutations in disorders within a familial study or one of a large population.

Acknowledgements

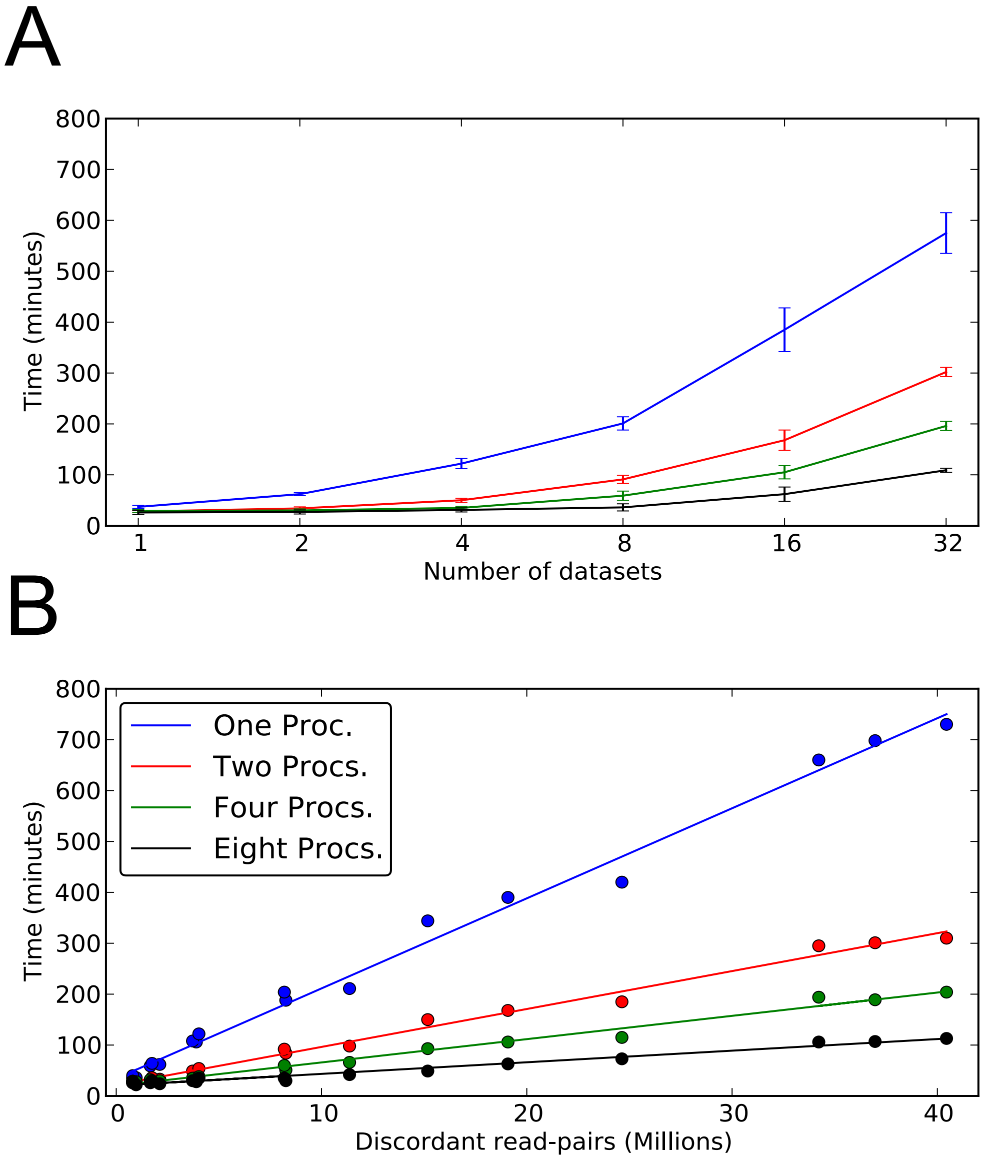
We would like to thank G Faust and R Layer for useful discussions and technical assistance. *Funding*: This work was supported by an NIH/NHGRI grant (1R01HG006693-01) to ARQ.; and an NIH New Innovator Award (DP2OD006493-01) and a Burroughs Wellcome Fund Career Award to IMH.

Supplementary Data



Supplementary Figure S2-1

Supplementary Figure S2-1. Clustergram of tumor and matched normal germline breakpoints. Hierarchical clustering of 64 tumor-normal genome pairs from The Cancer Genome Atlas based on 11,994 high-quality germline deletions (blue) and duplications (red), ≤ 1 Mb in size, made with the Matlab Clustergram function using Spearman correlation distance and Ward linkage. The y-axis is the 11,994 breakpoints, where the number of reads supporting the event is the value, indicated by the intensity of the cell color. Along the X-axis are the samples used in this study: 12 invasive breast cancers (BRCA), 3 colon adenocarcinomas (COAD), 18 glioblastomas (GBM), 6 lung adenocarcinomas (LUAD), 13 lung squamous cell carcinomas (LUSC), 11 ovarian cancers (OV), and 2 renal adenomas (READ) where tumor-normal pairs (denoted by T_N). The tumor-normal pairs are seen to cluster together given their breakpoints, as they are most genetically similar to each other.



Supplementary Figure S2-2

Supplementary Figure S2-2. Runtimes and speed-up with respect to input size and processor usage with 32 1KGP samples. Random sequential subsets of 1 to 32 (e.g., n = 1, 2, 4, 8, 16 and 32 dataset(s) were analyzed 3 times to create each of the dataset benchmarks). Runtimes were determined with an increasing number of processors (1:blue, 2:red, 4:green, and 8:black). *Hydra-Multi* was executed using the specified number of processes spawned on each subset and the runtime was measured in minutes. **(A)** The average runtime (minutes) across 3 random samplings at each dataset benchmark subset. Error bars represent 95% confidence intervals. **(B)** All runtimes (minutes) of the random samplings, 3 at each dataset benchmark, plotted against the number of discordant read-pairs (Millions) analyzed and a least-squares regression line.

CHAPTER 3:
GENOME-WIDE SINGLE CELL ANALYSIS OF HUMAN NEURONS UNCOVERS MOSAIC COPY NUMBER VARIATION

This chapter is based on the following manuscript:
MJ McConnell, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Shumilina S, Lasken R, Vermeesch, J, Hall IM, and Gage F. “Mosaic copy number variation in human neurons.” *Science* (2013), 342 (6158):632-637.

Abstract

The extent of endogenous somatic variation in the human body remains an open question. Here, we sought to evaluate and characterize the prevalence of somatic variation in human neurons. We employed newly developed single cell genomic approaches to detect copy CNVs in neurons obtained from human induced pluripotent stem cell (hiPSC) lines and postmortem human frontal cortex (FCTX). We identified numerous subchromosomal CNVs by single cell sequencing (SCS) of endogenous human frontal cortex neurons revealed that roughly 13 to 41% of neurons have at least one Mb-scale de novo CNV, that deletions occur twice as often as duplications, and that a portion of neurons have highly altered genomes marked by many CNVs. This work shows that mosaic CNVs are detected by SCS in human neurons.

Introduction

Investigators have postulated that somatic DNA variation in neurons may be a source of cellular diversity in the human brain (Ostertag *et al.*, 2005; Martin, 2009; Bushman and Chun, 2013). Such speculation of genetic variation among individual neurons arose from several studies reporting a higher incidence of retrotransposition (Muotri *et al.*, 2005; Baillie *et al.*, 2011) and aneuploidy (Rehen *et al.*, 2001; Rehen *et al.*, 2005; Yurov *et al.,* 2007) in neuronal genomes compared to other cell types. However, accurately assessing somatic variation is difficult. Somatic variants exist in only a small subset of cells within the whole population, making their identification in bulk tissue exceedingly error-prone, time-consuming, and expensive. Fortunately, two single cell methods have been developed to investigate genome-wide somatic variation: analysis of multiple displacement amplification (MDA) products on microarrays (Vanneste *et al.,* 2009) and single cell sequencing (Navin *et al.*, 2011). These methods assuage the aforementioned inhibitory factors when analyzing bulk tissue by isolating, amplifying, and analyzing copy number variations across the whole genome of a single cell. While these methods were developed for other cell types, we made adjustments and improvements to investigate somatic variation in neurons.

Researchers had already used single cell sequencing approaches to investigate somatic variation neuronal genomes. In 2012, Evrony *et al.* sought to investigate and confirm the reports of the increased retrotransposition in human neuronal genomes by mapping the frequency of L1 insertions. In this study, authors isolated single neurons and created sequencing libraries enriched for L1 insertions to assess the number of unique L1 retrotransposition events within each individual neuron. In doing so, the authors detected bona fide somatic variation and mosaicism, but reported fewer than 0.6 insertions per neuron (n=300) and concluded that L1 insertions are not a major generator of neuronal diversity in the cortex and caudate (Evrony *et al.*, 2012). However, this study was limited in its scope, as the authors only investigated L1 insertions, leaving the extent of genome-wide somatic variation unanswered.

Here, we demonstrate that human neuronal genomes are likely to exhibit somatic mosaicism at a relatively high incidence. We achieve this by improving two independent strategies of single cell analysis: the Vanneste *et al.* and Navin *et al.* methods. First, we show the robustness of our altered methods by identifying CNVs human fibroblasts. Next, we positively identify hemizygosity and trisomy in karyotypically confirmed in human trisomic fibroblasts. We then use our approaches to examining genome-wide somatic CNVs in two different sources of human neurons: human induced pluripotent stem cells (hiPSC) differentiated into neurons and human post-mortem frontal cortex (FCTX) neurons. Finally, we characterize these CNVs and illustrate their abundance at varying levels of experimental stringency. Concisely, these CNVs do not appear to be enriched in known genomic features and we only detected a single *in vivo* reoccurring (or clonal) CNV in the FCTX experiments.

Materials and Methods

The following methods are identical to those found in the manuscript:
MJ McConnell, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Shumilina S, Lasken R, Vermeesch, J, Hall IM, and Gage F. "Mosaic copy number variation in human neurons." *Science* (2013), 342(6158):632-7.

Experiments and analyses were developed and/or performed by MR Lindberg under the guidance of IM Hall and MJ McConnell unless otherwise noted. Co-authors provided the neuron and fibroblast sources, as well as microarray data. S Shumilina assisted in the creation of FCTX sequencing libraries alongside MR Lindberg. MR Lindberg performed the analysis of all data shown (microarray and sequencing), unless attributed.

Human cell culture (performed by co-authors)

The human fibroblasts, human induced pluripotent stem cell (hiPSC)-derived neural progenitor cells (NPCs), and hiPSC-derived neurons used in this study were parallel cultures of the neurotypic control lines reported previously in Brennand *et al.*. Reagents for cell culture were purchased from Life Technologies and their subsidiaries (San Diego, CA) unless noted otherwise. Human fibroblasts from AG09319 (referred to as “D” herein), AG09429 (referred to as “C” herein), AG03651 (referred to as “E” herein), and GM01920 (trisomic male) were obtained from the Coriell Institute (Camden, NJ) and grown in DMEM with Glutamax supplemented with 15% FBS (Atlanta Biologicals, Atlanta, GA).

Briefly, reprogramming was initiated using a cocktail of 5 tetracycline-inducible lentivirus (LV) vectors expressing human OCT4, SOX2, c-MYC, KLF4, and LIN28 cDNAs. Human fibroblasts were infected every day for five days. Following infection, fibroblasts were plated on a mouse embryonic fibroblast (MEF) feeder layer and switched to HUES media (KO-DMEM, 10% KO-Serum Replacement, 10% Plasminate, 1x Glutamax, 1x NEAA, 1x 2-mercaptoethanol and 20 ng/ml bFGF2 (Peprotech, Rocky Hill, NJ), supplemented with 1ug/mL doxycycline (Sigma, St. Louis, MO). Successful reprogramming was confirmed by human embryonic stem (ES) cell-like morphology, by expansion and maintenance of a euploid karyotype beyond 15 passages, by expression of endogenous pluripotency genes (e.g., OCT4, SOX2, NANOG, REX1, and CRIPTO mRNA) and proteins (OCT, SOX2, NANOG, and TRA-1-60), and, importantly, by repression of LV genes in the absence of doxycycline. Karyotypically normal hiPSCs were used to derive NPCs. hiPSCs were enzymatically dissociated from the MEF feeder layer using Collagenase type IV and grown in suspension as embryoid bodies (EBs) in N2/B27 media (DMEM/F12-Glutamax, 1X N2, 1XB27). After 1 week, EBs were transferred onto polyornithine (PORN)/laminin-coated plates in N2 media containing 1 μg/ml laminin. After an additional week of differentiation, neural rosettes formed; these were manually dissected, dissociated, and plated onto PORN/laminin-coated plates in NPC media (N2/B27 media with 1 μg/ml laminin and 20 ng/ml FGF-2) to expand NPCs. hiPSC-derived NPCs (passages 7 and 8) were differentiated into neurons in neural differentiation media (DMEM/F12-Glutamax, 1X B27-RA, 1X N2 with 20 ng/ml BDNF, 20 ng/ml GDNF (Peprotech), 1 mm dibutyrl-cyclicAMP (Sigma), 200 nm ascorbic acid (Sigma)) for 7 weeks. Karyotyping and FISH were performed by WiCell Cytogenetics (Madison, WI). FISH probes for chromosomes (Chrs) ChrX (Chr (Kallman probe set) were obtained from Abbott Laboratories (Abbott Park, IL). The ChrX p arm probe is specific for ChrXp22.3. The centromeric Chr20 probe is from Cytocell (Cambridge, UK). The Chr20 q arm probe is specific for Chr20q21 (RPCI-11 702M8-552, Empire Genomics, Buffalo, NY).

Isolation of single cells (developed and performed by co-authors, providing fibroblasts and neurons microarray data)

Confluent fibroblast cultures (passage 7 – 10) were serum-starved for 72 hours; G1 arrest was confirmed on a subset of this population using flow cytometry. NPCs (passages 9 and 10) were refractory to serum starvation; therefore, possible analysis of some S or G2 cells cannot be excluded. Single cells were picked by hand using a micropipette ("the Stripper") and 75 uM glass pipettes (Origio Midatlantic Devices, Mt. Laurel, NJ). Five-week-old hiPSC-derived neuronal cultures were infected twice with a LV construct (Brennand *et al.*, 2011) where GFP expression is driven by a synapsin promoter (Syn::GFP). Two weeks later, cells were dissociated using TrypLE and counterstained with 10 ug/mL propidium iodide (PI). GFP-positive, PI-negative cells were isolated via fluorescence activated cell sorting (FACS) on a FACS Aria II (BD Biosciences, San Jose, CA). Neurons were sorted into DMEM with 10% FBS and 10% DMSO and then frozen at -80C in Styrofoam. Frozen vials of hiPSC-derived neurons were thawed and individual cells isolated manually as before (2). Single cells were lysed and genomic DNA amplified via multiple displacement amplification (MDA) using phi29 polymerase (Genomiphi V2, GE Healthcare, Piscataway, NJ) as described (11). MDA products (5 ng) were examined for even amplification (e.g., +/- 5% of the Ct for 5 ng bulk genomic DNA) using qPCR (Applied Biosystems, San Diego, CA). To test for even amplification, we used a 10 locus subset of the 47 single copy loci used in Hosono *et al.* (34) (here, Chr1p, Chr2p, Chr3q, Chr7p, Chr10p, Chr11p, Chr14q, Chr17q, Chr19p, and Chr21q), similar to the approach employed previously for MDA QC (35, 36).

Detection of copy number variations from microarray data (data provided by co-authors)

MDA products passing qPCR quality control (QC) measures were analyzed on Affymetrix 250K NSP chips (Affymetrix, San Jose, CA). Partek Genomics Suite Software (version 6.6 beta, Partek, St. Louis, MO) was used to calculate predicted copy numbers for each probe set intensity. A custom copy number model composed of 161

MDA single cell experiments (from this and other studies) was generated to perform quantile normalization of the calculated copy numbers. The background-adjusted values were then subjected to GC correction in windows of 10 Mb, and artifact-prone probes were removed according to the Pugh *et al*. probe list. We then performed smoothing by taking the median copy number value in non-overlapping genomic windows composed of 100 probes. On average, each 100-probe bin corresponds to 666 kb of genome sequence. At this stage we also excluded 6 out of 107 samples that had excessively "noisy" copy number profiles, defined as having a median absolute deviation (MAD) greater than 0.7. To detect CNVs we used the circularly binary segmentation (CBS) algorithm (*Olshen et al.*, 2004) from the DNAcopy package in R, with the following parameters: *alpha=0.001, undo.splits="sdundo", undo.SD=1*. We defined CNVs as segments composed of 10 or more contiguous genomic windows whose copy number value differed from the dataset's median copy number by at least 1 MAD. We did not attempt to detect CNVs on the Y chromosome.

Isolation of post-mortem neuronal nuclei (protocol developed by MJ McConnell and J Piper; performed alongside MJ McConnell and J Piper)

Postmortem human frontal cortex from UMB#5125 (a neurotypic 24-year-old female, 9 hour post-mortem interval), UMB#1846 (a neurotypic 20-year-old female, 9 hour post-mortem interval) and UMB#1583 (a neurotypic 26-year-old male, 18 hour post-mortem interval) were obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland. Tissue samples were placed in nuclear isolation medium [(NIM) 25 mM KCl, 5 mM MgCl2, 10 mM Tris-Cl, 250 mM sucrose, 1mM dithiothreitol (DTT), and 1X protease inhibitor cocktail (Roche)] and homogenized with a polytron tissue homogenizer (Kinematica, Inc., Bohemia, NY). Homogenized tissue was supplemented with 0.1% TritonX-100, and further processed using a dounce homogenizer. Samples were centrifuged (1,000xg, 8 min) and the pellet was resuspended in 10:5:1 NIM:Iodixanol (Sigma):OptiPrep Diluent for Nuclei [(ODN) 150 mM KCl, 30 mM MgCl2, 60 mM Tris-Cl, 250 mM sucrose)]. Samples were layered onto a 29% Iodixanol in ODN cushion using a 1 mL syringe and centrifuged (10,300xg, 20 min, 4°C) in a Beckman L8-M ultracentrifuge with SW55 Ti rotor. Pellets were resuspended in nuclei storage buffer [(NSB), 5 mM MgCl2, 50 mM TrisCl, 166 mM sucrose, 1 mM DTT, and 1X protease inhibitor cocktail. Free nuclei and purity were confirmed visually by microscope.

Neuronal nuclei were purified from bulk brain nuclei using NeuN immunostaining (Spalding *et al.*, 2005; Westra *et al.*, 2010). Immunostaining was performed for 1 hour at 4°C with gentle agitation in PBS containing 5 ug/mL (1:2000). AF488-conjugated NeuN (Chemicon, Billerica, MA). Nuclei were then stained for DNA content with 10 ug/mL DAPI and analyzed by FACS (primarily performed by the University of Virginia Flow Cytometry Core). Single cells from the NeuN and DAPI positive population were sorted into 96 well plates alongside 1 water control per row. For benchmarking experiments, trisomic male fibroblasts were similarly sorted into 96 well plates based on size and their propidium iodide exclusion.

Single cell sequencing (protocol developed by S Shumilina, under IM Hall; performed alongside S Shumilina)

Isolated single nuclei or cells were lysed and amplified using the WGA4 GenomePlex Single Cell Whole Genome Amplification Kit (Sigma), using 15 cycles of PCR amplification. Subsequent WGA4 products were purified with Qiagen mini-elute columns (Qiagen, Germantown, MD). Illumina-compatible sequencing libraries were constructed using the Nextera Sample Prep (Epicentre Biotechnologies, Madison, WI and Illumina, San Diego, CA) according to the manufacturer's protocol, with the modification that we used a 1:200 dilution of the "transposome" enzyme complex in the "tagmentation" reaction (which helps control the fragment size distribution in single cell reactions). Tagmented DNA fragments were purified with mini-elute columns (Qiagen) and subjected to 12-15 cycles of PCR, during which barcodes were added to each library to facilitate pooled sequencing. The resulting barcoded sequencing libraries were purified with mini-elute columns (Qiagen). Each library was run on a 2% Low Range Ultra Agarose gel (Bio-Rad, Hercules, CA) with TAE and stained with SYBR Gold (Invitrogen) for 10-40 minutes. The 200-600 bp size fraction was isolated by gel extraction and purified with the QIAquick kit (Qiagen). Frontal cortex nuclei libraries were sequenced with PE sequencing on an Illumina GAIIx (performed by the University of Virginia DNA Sequencing Core) with 38-39 bp reads and fibroblast cell libraries were sequenced by single-end sequencing on an Illumina MiSeq with a read length of 59 bp.

Detection of copy number variations from single cell sequencing data

Copy number was assessed in dynamically sized genomic windows containing 500 kb of uniquely-mappable DNA sequence, as defined by the *wgEncodeCrgMapabilityAlign40mer* track from the UCSC Genome Browser (Meyer *et* *al.*, 2013). The mean absolute window size was 687 kb. PE reads were aligned to the human genome (NCBI Build 37) using BWA (version 0.5.10) with default settings (Li and Durbin, 2009) and duplicates were removed using *MarkDuplicates* from the Picard software suite (http://picard.sourceforge.net/). RD analysis was performed similarly to those described in previous works (Quinlan *et al.*, 2010; Quinlan *et al.*, 2011; Malhotra i, 2013). RD was assessed using *coverageBed* from the BEDTools software suite (Quinlan and Hall, 2010). Since Illumina sequence coverage is known to vary due to GC content, to obtain the predicted copy number of each genomic window we divided the RD of that window by the genome-wide median RD of all windows with similar GC content, as measured in 1-3% intervals, then multiplied by 2. CNVs were identified using the CBS algorithm (Olshen *et al.,* 2004) with the aforementioned parameters. We defined CNVs as segments composed of 5 or more contiguous genomic windows whose copy number value differed from the dataset's median copy number by at least 2 MADs. CNVs were not called on the Y chromosome. For putative CNVs on the X chromosome in the male sample, the median and MAD of the X chromosome were used to filter CNV calls.

The final CNV callset only includes datasets that passed the following QC criteria: 1) the dataset contained more than $5x10^5$ reads following duplicate removal; 2) the median absolute deviation of predicted copy number values in autosomal genomic windows was not more than 0.35; and 3) the dataset had a confidence score of at least 0.85, as defined below. In total, 110 of 208 datasets passed all of these QC filters. The confidence score, *S,* is a measure of the extent to which a given datasets adheres to the expectation of integer-like copy number measurements. The rationale for this QC measure is that we are using a digital technology (DNA sequencing) to measure copy number in single cells, and thus there is a strong expectation that copy number profiles should display approximately integer values. Non-integer copy number values may potentially occur due to regional variability in DNA amplification efficiency or flow-sorting errors that result in multiple nuclei being deposited into a single well.

Confidence Score, *S*: $S = 1 - 2\,\frac{\sum_{i=0}^{n} \min(\lceil C_i \rceil - C_i, C_i - \lfloor C_i \rfloor)}{n}$

C: the median predicted copy number of a given genomic interval (*i*) after copy number segmentation

n: the total number of genomic windows in the dataset

This score is the average distance between the predicted absolute copy number of each genomic segment in the dataset (as defined by the CBS algorithm) to the nearest integer value. This computed average is then multiplied by a factor of two in order to compare the actual distances to the worst-case distance (0.5) for every interval. This actual to worst-case ratio is then subtracted from 1 to yield a score between 0 and 1, where the more digital the dataset, the closer this score is to 1. Therefore, a dataset with a score of 0.85 or higher is very close to the assumed model of containing integer copy number values.

Single cell sequencing and microarray clustering

Both SCS and microarray analysis were performed for 7 of the hiPSC-derived neurons. To enable straightforward comparison of these two data types across the same genomic intervals, for this analysis we aggregated SNP array data in the same genomic windows as SCS data (rather than 100-probe windows). The mean window size is 687 kb, and the windows contained a mean of 57.8 probes (median 58). Only 6 of the windows had zero probes and these were assumed to have a copy number of 2. The microarray data processed in this manner are somewhat more noisy than those analyzed with a 100-probe window, but overall data quality is similar. To assess the concordance between SCS and microarray methods, the raw per-window copy number values of these 14 datasets were subjected to unsupervised clustering using the pvclust package (http://www.is.titech.ac.jp/~shimo/prog/) in R, using default parameters: *distance=correlation, linkage=average*.

Enrichment Analyses

For enrichment analyses we used the BITS algorithm (Layer *et al.*, 2013) to count the observed number of overlaps between CNVs and various genome annotations. The fragile sites track was obtained from Fungtammasan *et al.*, while all other tracks were downloaded from the UCSC Genome Browser (Meyer *et al.*, 2013). For these analyses we used CNVs less than 20 Mb in size, which reduces the total callset from 148 to 133. We then conducted Monte-Carlo simulations to find the expected number of intersections by shuffling both the CNVs and annotation track 1000 times. The log2 enrichment ratio was caclulated as the observed number of overlaps divided by the median (or mean if the median was 0) number of intersections observed in simulations. Analyses of telomeric enrichment were performed in a similar way, however, only the CNVs were shuffled for the 1000 iterations.

Estimating the false discovery rate of copy number variant detection by read-depth analysis

To estimate the FDR for CNV detection by RD analysis, we performed Monte-Carlo simulations in which the relative order of genomic windows was shuffled 1000 times for each dataset. Shuffled datasets were subjected to copy number segmentation and filtering exactly as for real data, with the caveat that we excluded the X and Y chromosomes from these analyses to avoid sex-related effects. The FDR was calculated as the mean number of CNVs detected in simulated data, adjusted for the exclusion of sex chromosomes (based on their size). This FDR estimation strategy specifically measures the specificity of CNV detection with respect to random sources of noise, however, it does not account for potential systematic or regional effects and therefore should be considered a lower bound.

Estimating the false negative rate of copy number variations detection by read-depth analysis

It is difficult to estimate the false negative rate (FNR) because our CNV size detection limits (~3.4 Mb) greatly exceed the size of known germline CNVs, and therefore we do not have access to a set of true CNVs with which to measure sensitivity. However, for the 41 cells derived from male individual 1583 we expect to detect the X and Y chromosomes as single copy "aberrations" relative to autosomes. We exploited this feature to develop a simulation-based approach to measure FNR in these 41 datasets. For example, to simulate a single deletion comprising 5 genomic windows, we randomly selected 5 contiguous genomic windows from the X chromosome, extracted their predicted copy number values, and used these values to replace the copy number values of 5 contiguous windows from a randomly selected autosomal location. To simulate duplications we used a similar approach, but instead of replacing the 5 autosomal copy number values, we simply added the autosomal values to the values extracted from the X chromosome. The resulting simulated dataset was then subjected to copy number segmentation and CNV filtering precisely as for the real data. To calculate the FNR for CNVs of a given size (e.g., 5 windows) in a given dataset, we simulated 1000 CNVs of that size and assessed the fraction of simulations in which we detected the synthetic CNV. Detection was defined as a reciprocal overlap of 50% between the simulated and detected genomic segment.

Results

We investigated human neurons from hiPSC-derived neurons and FCTX neurons (Figure 3-1). We used fluorescence activated cell sorting (FACS) on neuronogenic hiPSCs expressing synapsin::GFP and post-mortem tissue based on NeuN Immunostainin (Spalding *et al.*, 2005). Single hiPSC-derived neurons were subjected to multiple displacement amplification (MDA) (Dean *et al.*, 2002) and hybridized to Affymetrix 250K SNP arrays (Vanneste *et al.,* 2009). Nuclei from post-mortem tissue were subjected to Illumina DNA sequencing using a custom version of the protocol developed by

(Navin*, et al.*, 2011) that combines the GenomePlex whole-genome amplification method with Nextera-based library preparation (Adey *et al.*, 2010). Stringent QC measures were developed to ensure that only the highest quality amplification reactions and datasets were included in downstream analyses (see Materials and Methods).

Design and detection

We detected CNVs by partitioning the genome into intervals 10 to 100 times larger than the local amplification biases reported for single cell DNA amplification (Lasken and Stockwell, 2007; Lasken, 2009). In the SNP microarray data, median copy numbers of every 100 consecutive probes, with a mean genomic interval of 666 kb, was calculated. For the sequencing data, we created bins of 500 kb of uniquely mappable sequence, with a mean size of 687 kb, and counted RD. Circular binary segmentation (Olshen *et al.*, 2004) was used to produce distinct segments for CNV detection.

Rigorous filtering based on the amplitude of the segments relative to the noise (median absolute deviation) of each dataset, the number of reads, and the overall quality

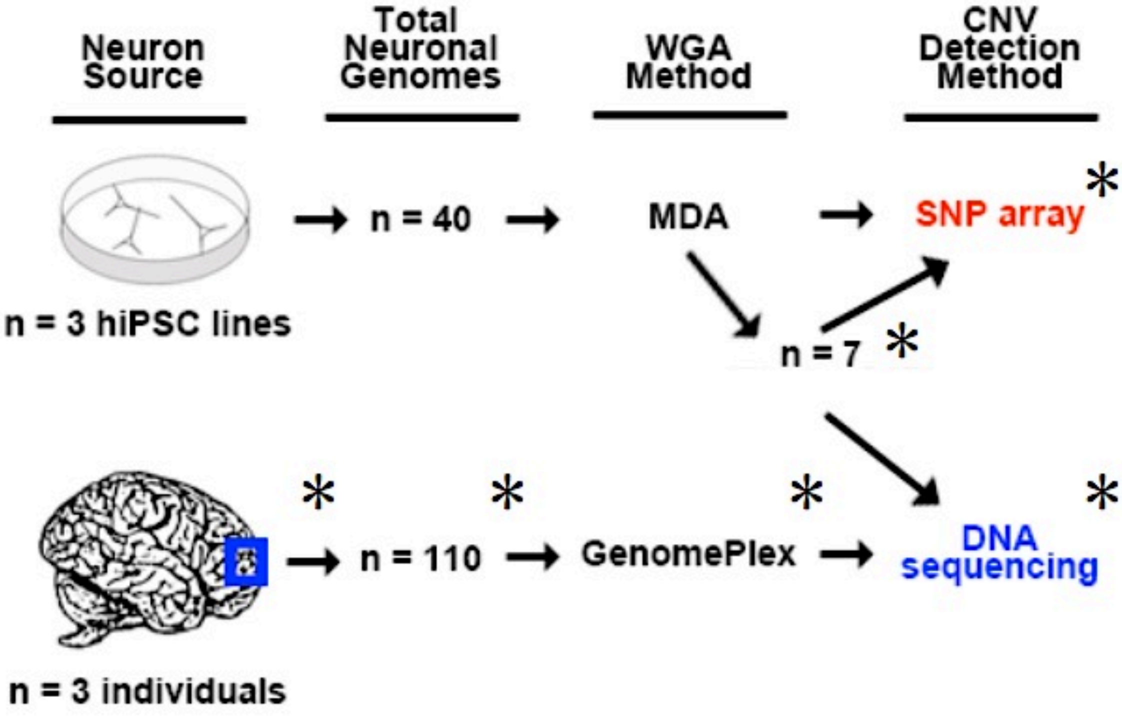


Figure 3-1

Figure 3-1. Single cell analysis by SNP array and DNA sequencing. Summary of the single cell approaches used.

*Indicates personal contribution to McConnell *et al.* (detailed in Materials and Methods) determined by a confidence score were applied. Furthermore, a minimum consecutive bin threshold for putative CNVs (see Materials and Methods). The mean CNV detection resolution minimum of 6.7 Mb for SNP array data and 3.4 Mb for sequencing data.

In order to establish concordance between the two methods, 7 MDA-amplified hiPSC-derived neurons were analyzed by both SNP array and sequencing (Figure 3-1, Figure 3-2, Supplementary Figure S3-1 and Supplementary Figure S3-2). Sub-chromosomal deletions (Figure 3-2A and 3-2C) and duplications (Figure 3-2B and 3-2D) were found using both methods and both sources of neurons.

In vitro hiPSC microarray analysis

We analyzed 40 neurons from three human hiPSC lines, known hereafter as C (n=21), D,(n=6) and E (n=13). These cell lines originated from three different individuals that are regarded as neurotypic controls for a human hiPSC-based disease model (Brennand *et al.*, 2011). In examining bulk DNA from the C and D line donor fibroblasts or human hiPSC-derived NPCs, we did not observe genomic aberrations of clonal origin. Additionally, we found that 27 of the 40 hiPSC-derived neurons had copy number profiles similar to that of bulk DNA. In contrast, we observed 13 unique genomes that had the following genomic rearrangements: four whole chromosome losses, seven whole chromosome gains, and 12 sub-chromosomal CNVs, ranging from 7.0 Mb to 156 Mb (Figure 3-3A and 3-3B). All CNVs were unique in each of the neurons, which would imply that the CNVs are likely not early clonal events, but instead developing in later lineages or are private to individual cells.

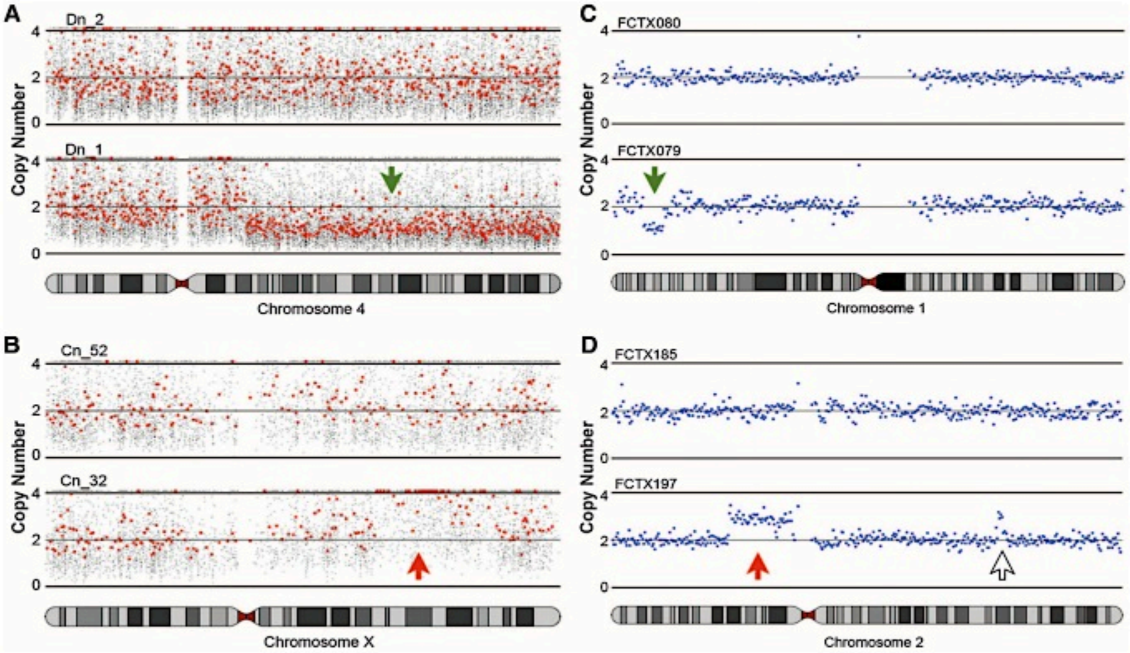


Figure 3-2; A and B data provided by co-authors

***Figure 3-2. Mosaic CNV detected in human neurons*. (A and B)** Subchromosomal deletions (green down arrow) and duplications (red up arrow) are observed in hiPSC-derived neurons. **(A)** Neuron Dn_1 has a deletion on chromosome (chr) 4q (bottom); neuron Dn_2 has no CNV on Chr4 (top). Small gray dots show the predicted copy number at individual SNPs; red dots show every 30th SNP. **(B)** Neuron Cn_32 has a duplication on ChrXq (bottom); neuron Cn_2 does not (top). **(C and D)** Single-cell sequencing reveals subchromosomal deletions (green down arrow) and duplications (red up arrow) in FCTX neurons. **(C)** FCTX079 has a deletion on Chr1p (bottom); FCTX080 does not (top). Blue dots show raw copy number predictions obtained by read-depth analysis (mean window size ~687 kb; see methods) **(D)** Neuron FCTX197 has a duplication on Chr2p (bottom), whereas FCTX185 does not (top). There is another possible duplication on Chr2q in FCTX197 (white arrow), which is comprised only four consecutive bins and therefore failed our five-bin confidence threshold.

CNVs seen in the C and D line human hiPSC-derived neurons were different than from those in C and D line fibroblasts or NPCs (Figure 3-3A). Among the 29 fibroblasts, one was aneuploid, missing Chr22 and ChrX, and six had single CNVs, ranging from 5.2 Mb to 27.7 Mb (Figure 3-3A). In the 19 hiPSC-derived NPCs, six duplications were observed. Furthermore, deletions were only seen in hiPSC-derived neurons and not in hiPSC-derived NPCs. Cumulative distributions of CNVs in the three cell types (Figure 3-3B) showed that distribution of neurons are significantly different than that of fibroblasts (Kolmogorov-Smirnov test, $P< 0.001$).

We wanted to confirm that CNVs and basal aneuploidy occurred in single fibroblasts. To do this, we used limiting dilution to seed single fibroblasts for expansion. Over seven days, the fibroblasts were allowed to expanded to roughly 20 sister cells and sister clones were isolated. One clonal expansion underwent a chromosome missegregation of Chr2, where one cell had an additional copy and a loss was observed in the sister cell (Supplementary Figure S3-3A). We also observed non-clonal CNVs and decided to use fluorescence *in situ* hybridization (FISH) for a common hiPSC CNV on Chr20 (Laurent *et al*., 2011) and for ChrX to corroborate these findings. In performing metaphase spreads, 20 karyotyped as euploid, 13/200 were aneuploid for ChrX (Supplementary Figure S3-3B), and 26/200 nuclei had a Chr20 CNV (Supplementary Figure S3-3C). These data show that CNVs can be detected in single human cells in culture by SNP array and FISH.

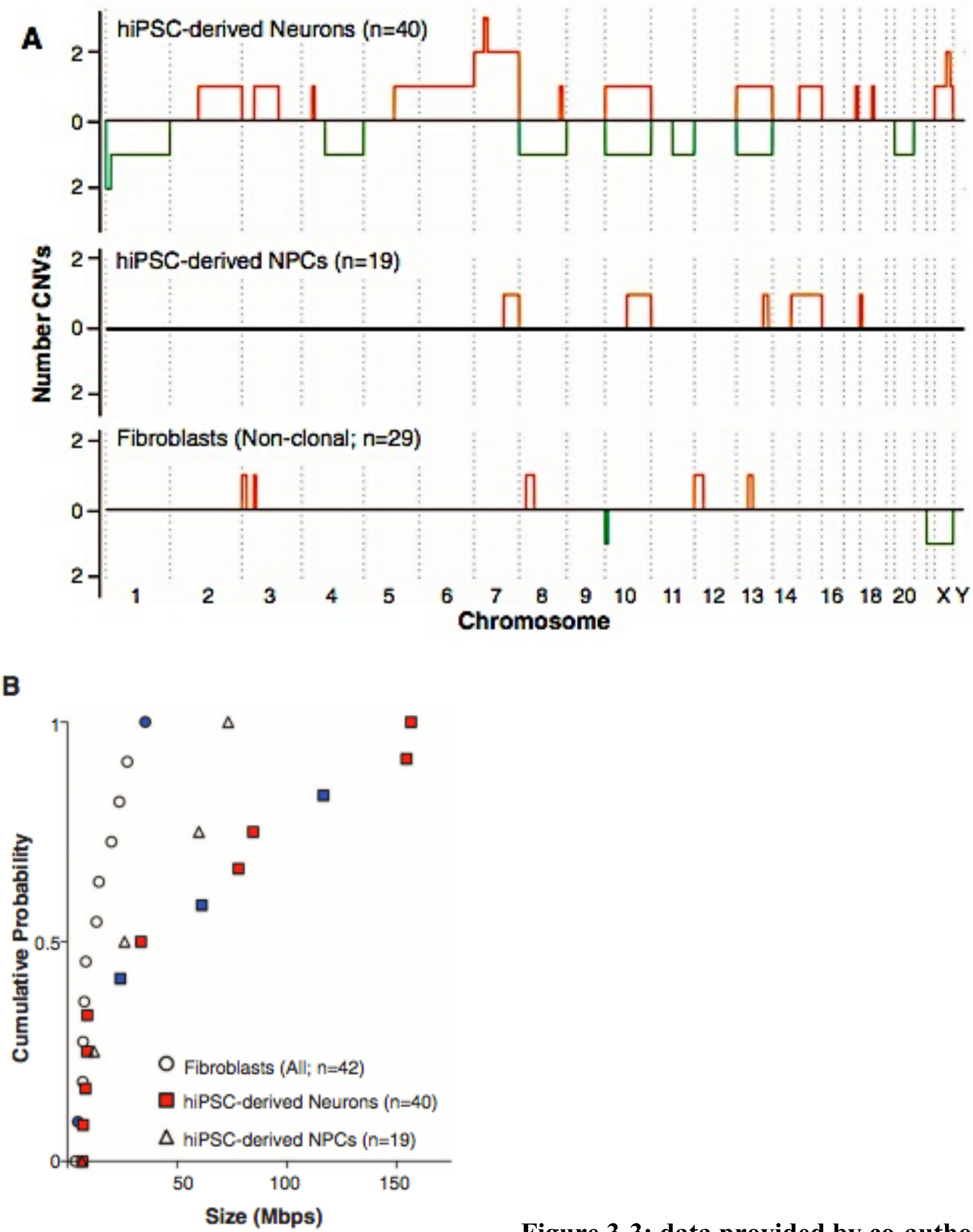


Figure 3-3; data provided by co-authors

Figure 3-3. Large CNVs are found in hiPSC-derived neurons. (**A**) Whole and subchromosomal duplications (red) and deletions (green) are summarized for 40 hiPSC-derived neurons (top). The *y* axis value represents the number of times each genomic interval was deleted (below in green) or duplicated (above in red). CNVs were detected in 9 out of 21 C neurons (Cn), 2 out of 6 D neurons (Dn), and 2 out of 13 E neurons (En). In donor hiPSC-derived NPC populations (middle), CNVs were detected in 1 out of 10 D NPCs (Dp) and 3 out of 9 C NPCs (Cp). In donor fibroblast populations (bottom), CNVs were detected in 7 out of 20 D fibroblasts (Df) and 0 out of 9 C fibroblasts (Cf). Note that chromosomes are not plotted to scale because data are summarized in 100-SNP bins. (**B**) The distribution of subchromosomal CNVs in fibroblasts were significantly different than in hiPSC-derived neurons (Kolmogorov-Smirnov test, $P < 0.001$). No deletions were observed in NPCs. Deletions are denoted with blue markers; all other markers indicate duplications. Aneuploidies are not included in this plot. For completeness, subchromosomal CNVs from clonal fibroblasts (**Supplemental Figure S3-4)** were included in this plot, bringing the total *n* to 42 fibroblasts.

In vivo neuron single cell sequencing

We then wanted to determine if mosaic CNVs were seen *in vivo* and were not simply some unforeseen artifact of cell culture. We adapted the SCS method (Navin *et al.*, 2011) to use on FCTX neurons because the digital readout of DNA sequence data offers superiority over the noisiness of microarrays (Navin *et al.*, 2011; Baslan *et al.*, 2012). We benchmarked our sequencing method with trisomic, male fibroblasts (47XY21+) (Supplementary Figure S3-3D; Supplementary Figure S3-4; Supplementary Figure S3-5). Subsequently, we sequenced 110 FCTX neurons originating from three different individuals [(a 24 year-old female (NICHD Brain Bank ID#5125; n = 19), a 26 year-old male (ID#1583; n = 41), and a 20 year-old female (ID#1846; n = 50)] (Supplementary Figure S3-6). Genomic profiles of some of these cells can be viewed in Figure 3-4. We imposed a strict set of filtering criterion to identifying high confidence CNVs (see Materials and Methods), where each CNV had to be comprised of five or more consecutive bins. In the 41 male neurons, our protocol positively identified 100% monosomy X and Y and identified 100% three copies of 21 and a single copy of X in the fibroblasts (Supplementary Figure S3-7). Furthermore, simulation experiments present us with a predicted mean FNR of 17% and a predicted mean FDR of 0.6% (Supplementary Figure S3-8; see Materials and Methods) illustrating that our methods can detect CNVs at a relatively high sensitivity and specificity. We cannot know our true FNR and FDR rates, but our predicted FNR rate indicates that we may be too conservative and underestimating the true incidence of CNVs.

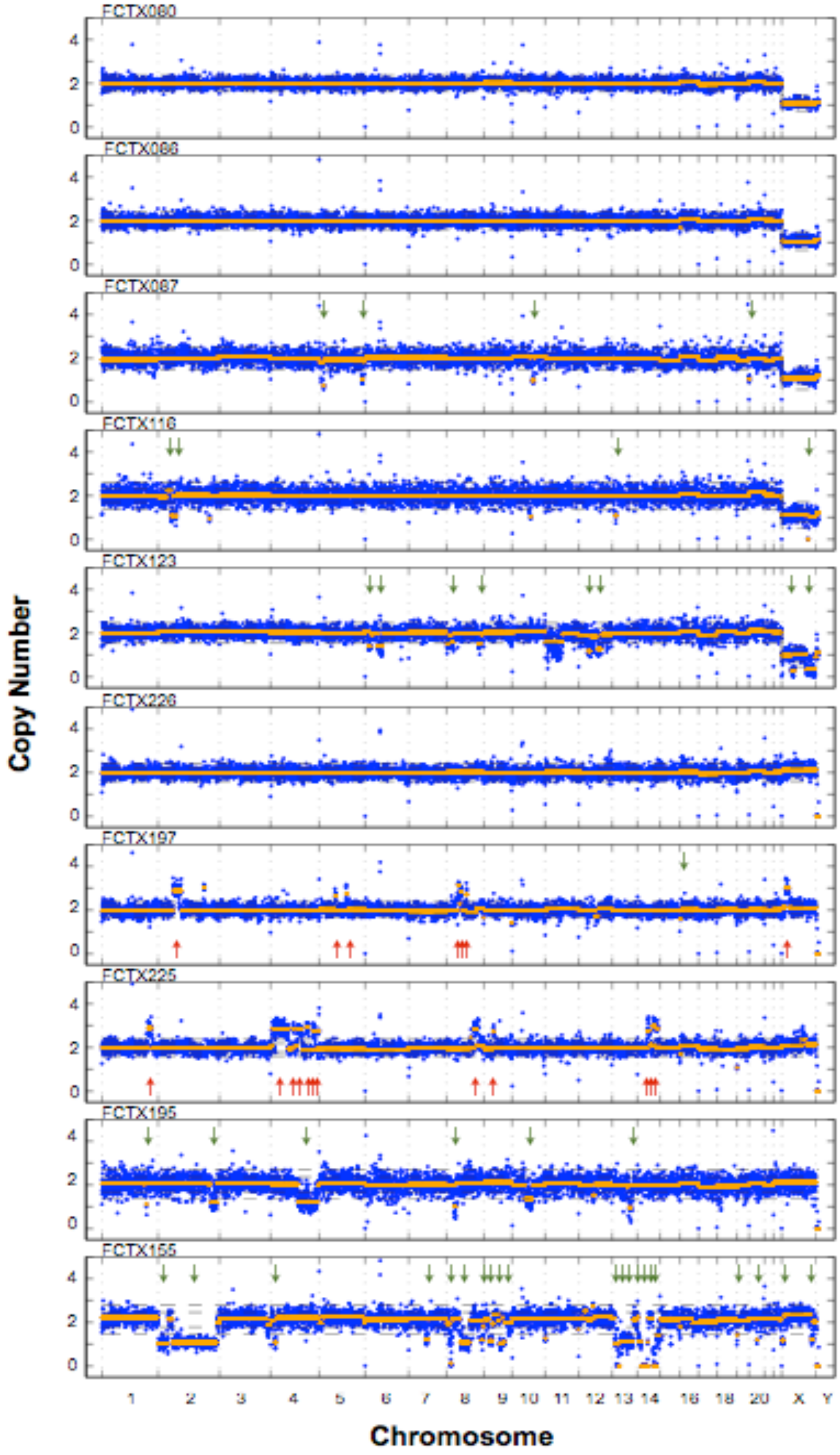


Figure 3-4; S Shumilina assisted in FCTX library preparation

Figure 3-4. Identification of CNVs in postmortem neurons using single cell sequencing. Genome-wide copy number profiles of five male (top) and five female (bottom) neurons from two individuals, no. 1583 and no. 1846, respectively. DNA copy number (*y* axis) was calculated by RD analysis of variably sized genomic windows containing 500 kb of uniquely mappable sequence (blue), and CNVs were detected by circular binary segmentation (orange). Green (down) and red (up) arrows denote deletions and duplications, respectively, that were identified by segmentation and passed filtering criteria. Reported CNVs comprise five or more consecutive bins and exceed two median absolute deviations (MADs). Dotted gray lines show 1 and 2 MADs from the median copy number of each data set. Arrows denote deletions (green, down and at an angle in FTCX195 and 155) and duplications (red, up) that were identified by copy number segmentation and passed filtering criteria. Note that single-copy "losses" of ChrX in cells from male individual no. 1583 are not indicated by arrows, but were identified in 100% of cells. All neuronal genomes can be found in McConnell *et al.*.

Of the 110 FCTX neurons we sequenced, 45 or roughly 41% harbored one or more somatic CNVs (Figure 3-5A, see Supplementary Figure S3-5A and S3-5D for different stringencies). The overwhelming majority of somatic CNVs were subchromosomal, ranging in size from 2.9 Mb to 75 Mb (Figure 3-5B); however, three CNVs affected >50% of the chromosome (e.g., FCTX155, Figure 3-4) and may be of a different mechanistic origin than the subchromosomal CNVs. Only in one instance did two CNVs share the same breakpoints (a 3 Mb subtelomeric deletion on Chr16 in FCTX198 and FCTX224. CNVs were predominantly smaller in size, less than 20Mb (n=133), and often (23.3%) occurred at telomeres (Figure 3-5C, 2067-fold enrichment by Monte-Carlo simulation experiments; see Materials and Methods). These small CNVs are not enriched at features known to affect genome stability such as transposons, segmental duplications or fragile sites. Additionally, these relatively small somatic CNVs were not enriched with germline CNVs or known genes (Supplementary Figure S3-9). On average, subchromosomal deletions were twice as common as duplications across the three individuals, suggesting a bias towards DNA loss in post-mitotic neurons. However, #1846 showed far more duplications than the other two individuals (Figure 3-5B and 3-5C, Supplementary Figure S3-6E). These results illustrate that somatic CNVs are a feature of neuronal genomes, but the CNV type and rate may be variable among individuals.

The seemingly high mutational burden reported is due to few cells with highly rearranged genomes. FCTX neurons typically exhibited 0 (59%) or 1-2 CNVs (25%), while 17 cells (15%) exhibited 108 of the 148 total CNV calls (73%) and seven cells exhibited nearly half (49%) of all CNVs (Figure 3-5A). These highly aberrant cells contained multiple discrete copy number oscillations between altered and unaltered segments on the same Chr. As expected, these switches maintained integer-like copy number values, an attribute of digital DNA sequencing technology. This was also observed, to a lesser degree, in hiPSC neurons, where a number of cells contained multiple CNVs on the same chromosome. This phenomenon of a few cells harboring many CNVs is illustrated by the fact that two of the FCTX neurons had more than 10 events. FCTX155 has nearly all of Chr2 affected, along with a single duplication and 18 deletions (Figure 3-4). In sequencing the 16 control fibroblasts, we did not observe this phenomenon where a few cells were highly aberrant, accounting for many of the CNVs calls (Supplementary Figure S3-5 and S3-7E). This may be due to insufficient sampling of the fibroblasts. Should more fibroblast have been sequenced, it is possible that we may have observed; however, these findings suggest that a subset of neurons may be more susceptible to (or undergo more) CNVs.

While single cell genome analysis is attractive and technically impressive, the approach makes orthogonal experimental validation impossible because the initial state of a single cell's genome cannot be knowable once it is amplified. However, we maintain that our bona fide set of CNVs is conservative and is most likely real due to the following lines of reason. First, we based our methods in those, which had been validated on clonally related cell populations: eight cell embryos (Vanneste *et al.*, 2009) and tumors (Navin *et al.*, 2011). Second, the CNVs we report are several orders of magnitude larger than the amplicons produced by whole genome amplification. Previous studies have reported amplification artifacts to be small (less than 10 kb) and located relatively uniformly throughout the genome (Laskin and Stockwell, 2007; Laskin, 2009); therefore, biases in amplification cannot account for the large-scale CNVs we see. Furthermore, simple amplification artifacts should not create both gains and losses of DNA at integer copy number values when sequenced. Third, DNA degraded during the post-mortem interval, the length of time elapsed from death to freezing of brain tissues, could not generate duplications. Additionally, the large deletions seen in FCTX were also observed in hiPSC-derived neurons. Fourth, Monte-Carlo simulation experiments demonstrated that our CNV detection strategy effectively identified single copy gains and losses at high sensitivity in control experiments. Fifth, enacted conservative QC measures to exclude datasets that may have exhibited uneven or suboptimal amplification. We also employed scrupulous filtering reflecting inherent physical constraints imposed by the nature sequencing data, thereby requiring integer-like copy number profiles (see Materials and Methods). Lastly, the CNV calls appear to be of very high quality based on their size, amplitude (deviation from the MAD), and integer-like properties (Figure 3-4). Additionally, when increase the stringency of our CNV detection requirements, a portion (30-56%) of the CNVs do not change (Supplementary Figure S3-7D). Even at the highest level of stringency, the central findings remain: somatic CNVs exist in a portion of human neurons (13-24%), there are more deletions than duplications (Figure 3-5C and Supplementary Figure S3-7A), and a small number of neurons marked copy number switches account for much of the observed variation (Figure 3-4 and Supplementary Figure S3-7D). Although it is still possible that some unknown single cell amplification or technical artifacts may confound aspects of this study, the aforementioned justifications and rationale underscore the strength of our core results and bolster our conclusions.

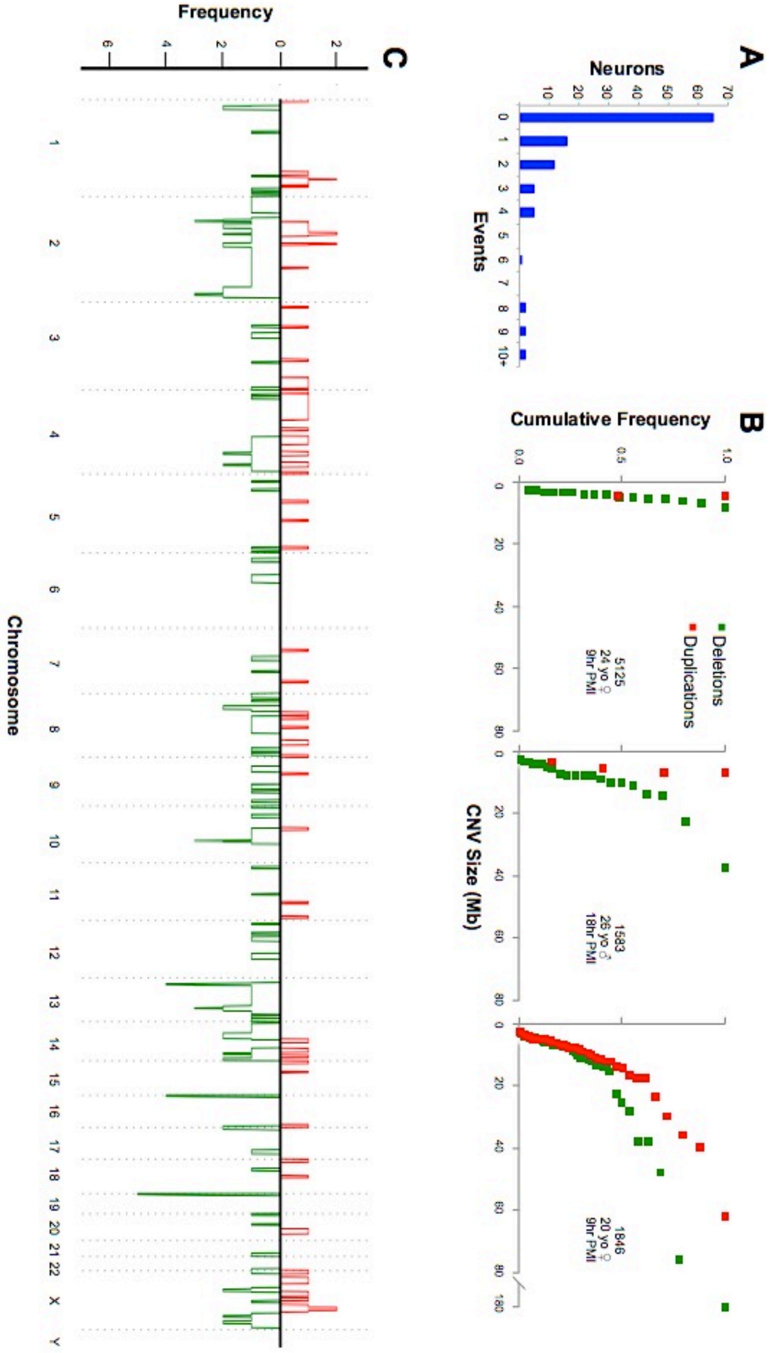


Figure 3-5; S Shumilina assisted in FCTX library preparation

Figure 3-5. Characterization of CNVs in postmortem neurons using single cell sequencing. **(A)** The number of individual neurons (Y-axis) that exhibited a given number of CNVs (X-axis). **(B)** Cumulative frequency of CNV sizes found per individual (deletions in green, duplications in red). **(C)** Whole and subchromosomal duplications (red) and deletions (green) are summarized for the 110 FCTX neurons.

Discussion

Through single cell genomic analysis of human neurons, we have characterized somatic mosaicism in the nervous system and extended its characterization to the single cell level. Previous studies using bulk DNA from somatic tissues, including brain, identified CNVs among monozygotic twins (Bruder *et al.*, 2008) and in different organs or brain regions from the same individual (O'Huallachain *et al*., 2012; Piotrowski *et al.*, 2008). These studies were limited in their ability to detect CNVs because they used bulk tissue. The authors reported less than 10% of cells harboring CNVs and only provided a rough assessment of somatic mosaicism, while our study shows that mosaic copy number variation is prevalent in human neurons. Further work is required to explore the mechanism and function of somatic mutation in neurons, as well as other cell lineages. For instance, neuronal lineages develop genomic instability during development and propagate them or individual neurons may become prone to CNVs due to persistent DNA damage. Reports have implicated electrophysiological activity as a source of DNA double stranded-breaks in neurons (Suberbielle *et al.*, 2013), and small circular DNAs caused by excision have been seen in multiple somatic cell types, including neurons (Shibata et al., 2012; Maeda et al., 2004). Furthermore, sub-chromosomal deletions and other rearrangements in human cells can be caused by retrotransposition (Gilbert, Lutz-Prigge, and Moran, 2002; Callinan *et al.*, 2005; Gilbert *et al.*, 2005; Symer *et al.*, 2002) whereby the increased rates of retrotransposon during human neurogenesis (Baille *et al.*, 2011; Coufal *et al.*, 2009) could lead to the preponderance of CNVs observed in neuronal genomes, thus supporting our observed ratio of deletions to duplications.

The patterns of CNVs observed in neurons were different than those observed in fibroblasts. Using three independent single cell approaches (SNP array, sequencing, and FISH) we detected Mb-scale CNVs in human cultured fibroblasts. Previous studies estimated CNVs (no larger than 1 Mb) to occur in skin fibroblasts at a frequency of 30% (Abyzov *et al.*, 2008). In order to study single cells, Abyzov *et al.* used stem cell reprogramming on single fibroblasts and then performed deep whole-genome sequencing on the group of cells from the hiPSC cell lines. This differs from our approach because they used a population of cells in their analysis, where as we used single cells. Their method afforded high resolution (2-5 kb); however, the reprogramming process may have served as a bottleneck event. This is because cells harboring many large CNVs may not be efficiently reprogrammed or clonally expanded, thus becoming underrepresented genomes. This leads us to believe that our findings may not be in contradiction with theirs. However, both studies did not see a subset of fibroblasts with highly aberrant genomes like those in neuronal genomes.

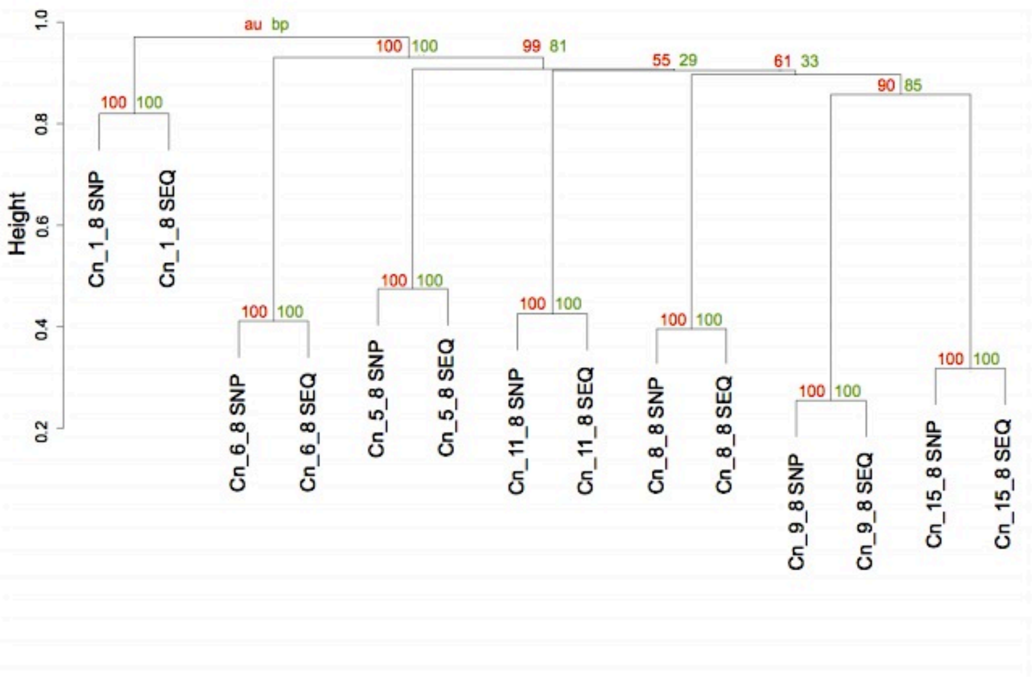
The consequence of somatic mosaicism on the function of neurons is not understood. The simplest hypothesis is that neurons will have well defined cellular phenotypes with distinct transcriptional or epigenetic programs based on their differing genomes. Future advances in single cell technologies should allow for this hypothesis to be tested by concomitantly analyzing the genome, epigenome, transcriptome, and the proteome of a single neuron. Moreover, somatic mutations and mosaicism in neurons has been shown to cause neurological disorders (Fishler, K. and Koch, R., 1991; Poduri *et al.*, 2012; Lee *et al.*, 2012). Existence or effects of somatic variation and mosaicism in neurons could explain complex neurological disorders (e.g., autism, schizophrenia, and Alzheimer's).

Acknowledgements

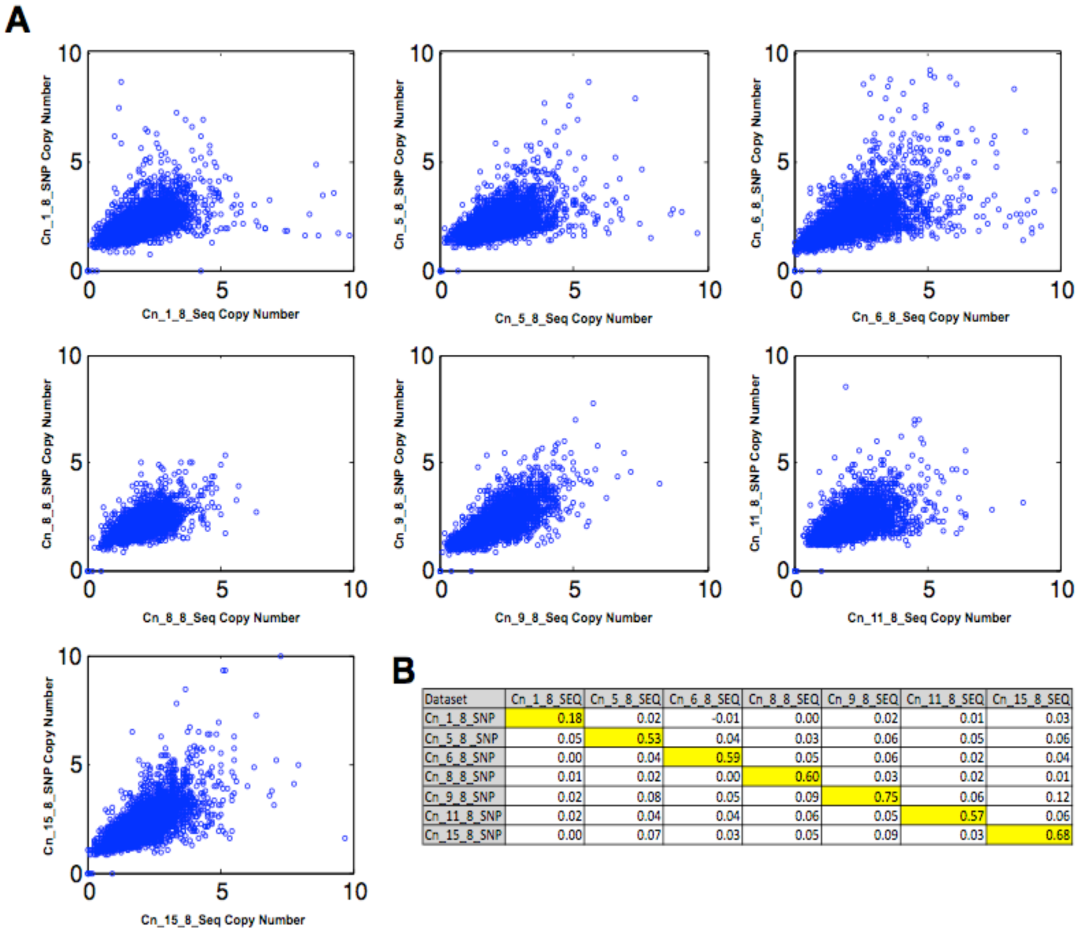
We thank D Husband (Salk), L Moore (Salk), and S Jackmaert (KU Leuven), R Layer and R Clark for technical assistance, A Prorock and Y Bao (UVA Sequencing Core) for DNA sequencing, M Solga and S Coquery (University of Virginia Flow Cytometry Core), and all members of the Gage laboratory for critical feedback on the project. This work was supported by a Crick-Jacobs Junior Fellowship to MJM; a Mather's Family Foundation Grant, an NIH TR01 (R01 MH095741) and a Helmsley Foundation grant to FHG; and an NIH New Innovator Award (DP20D006493-01) and Burroughs Wellcome Fund Career Award to IMH. Human tissue was obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD, contract HHSN2752009000011C, Ref. No. N01-HD-9-011. Microarray data have been deposited to GEO Omnibus under accession GSE51538, and DNA sequence data have been deposited to the NCBI Short Read Archive under accession SRP030642

Supplementary Data

Supplementary Figure S3-1. Single cell analysis using SNP array and sequencing. A cluster dendrogram shows concordance in copy number profiles for seven neurons mapped by SNP array hybridization intensity ("SNP") or sequencing read depth ("SEQ"). Numbers at tree nodes reflect the significance values reported by the R pvclust package for bootstrap resampling (1000 iterations) and can be interpreted as the percentage of simulated trees with the observed topology



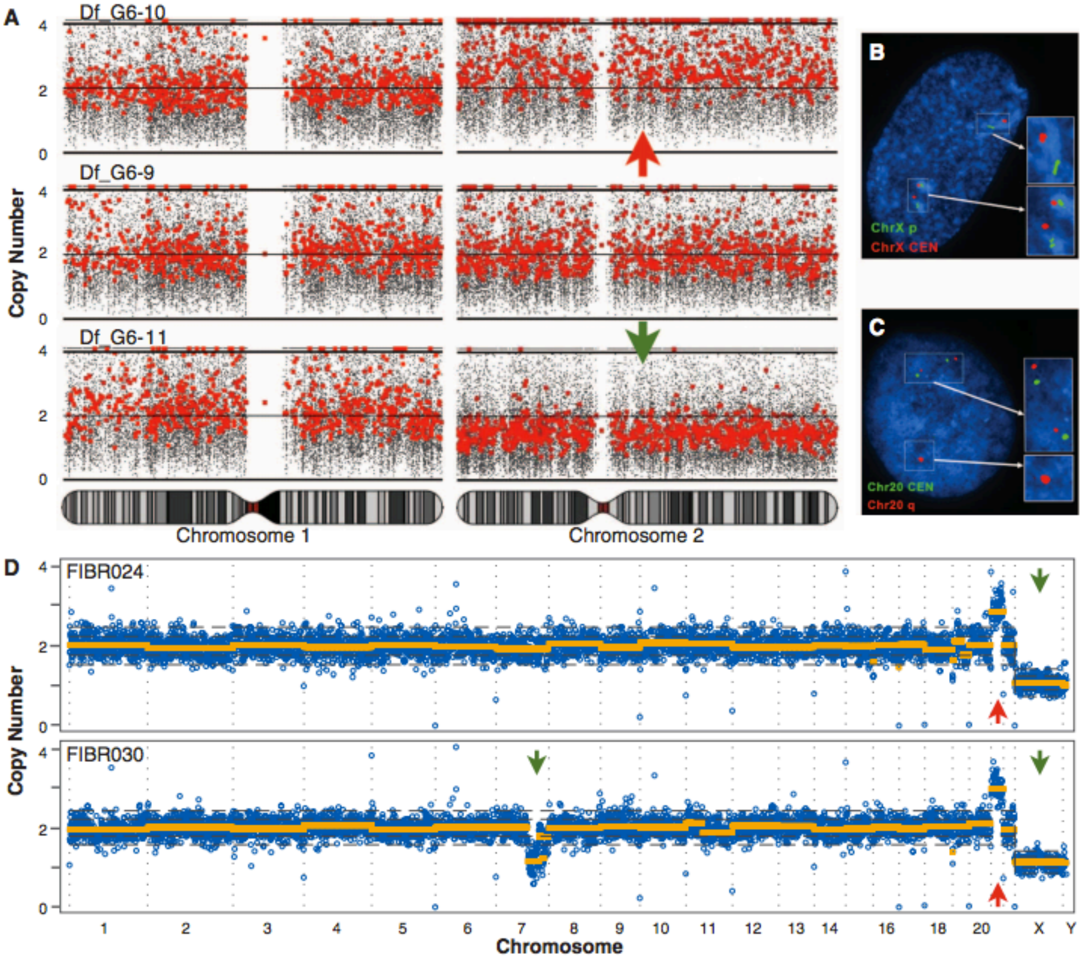
Supplementary Figure S3-1; data provided by co-authors



Dataset	Cn_1_8_SEQ	Cn_5_8_SEQ	Cn_6_8_SEQ	Cn_8_8_SEQ	Cn_9_8_SEQ	Cn_11_8_SEQ	Cn_15_8_SEQ
Cn_1_8_SNP	0.18	0.02	-0.01	0.00	0.02	0.01	0.03
Cn_5_8_SNP	0.05	0.53	0.04	0.03	0.06	0.05	0.06
Cn_6_8_SNP	0.00	0.04	0.59	0.05	0.06	0.02	0.04
Cn_8_8_SNP	0.01	0.02	0.00	0.60	0.03	0.02	0.01
Cn_9_8_SNP	0.02	0.08	0.05	0.09	0.75	0.06	0.12
Cn_11_8_SNP	0.02	0.04	0.04	0.06	0.05	0.57	0.06
Cn_15_8_SNP	0.00	0.07	0.03	0.05	0.09	0.03	0.68

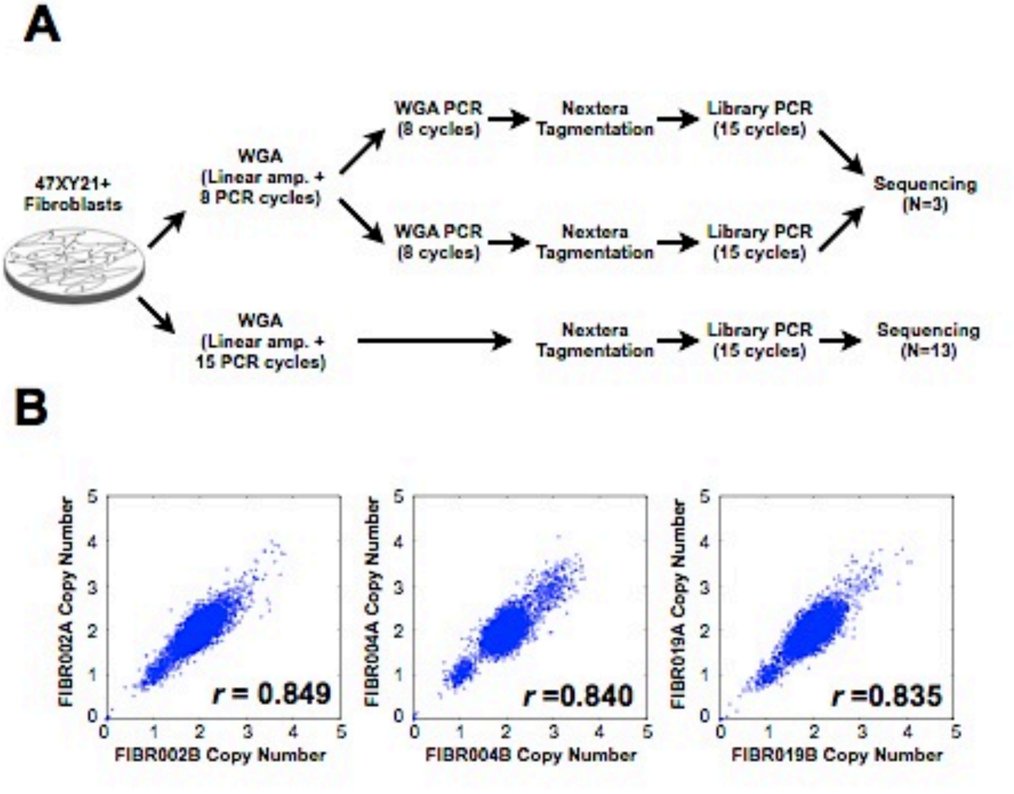
Supplementary Figure S3-2; data provided by co-authors

Supplementary Figure S3-2. Concordance between SNP array and DNA sequencing. **(A)** Scatter plots comparing raw copy number values between the seven neurons subjected to MDA-based whole-genome amplification followed by both SNP array analysis ("SNP") and DNA sequencing ("SEQ"). Copy number values were directly compared using the same ~687 kb windows used to measure read-depth (see methods). **(B)** Correlation matrix reporting pairwise Pearson correlation coefficients for every "SNP" and "SEQ" combination. Note that replicate SNP/SEQ experiments have dramatically larger correlation coefficients than non-replicate combinations.



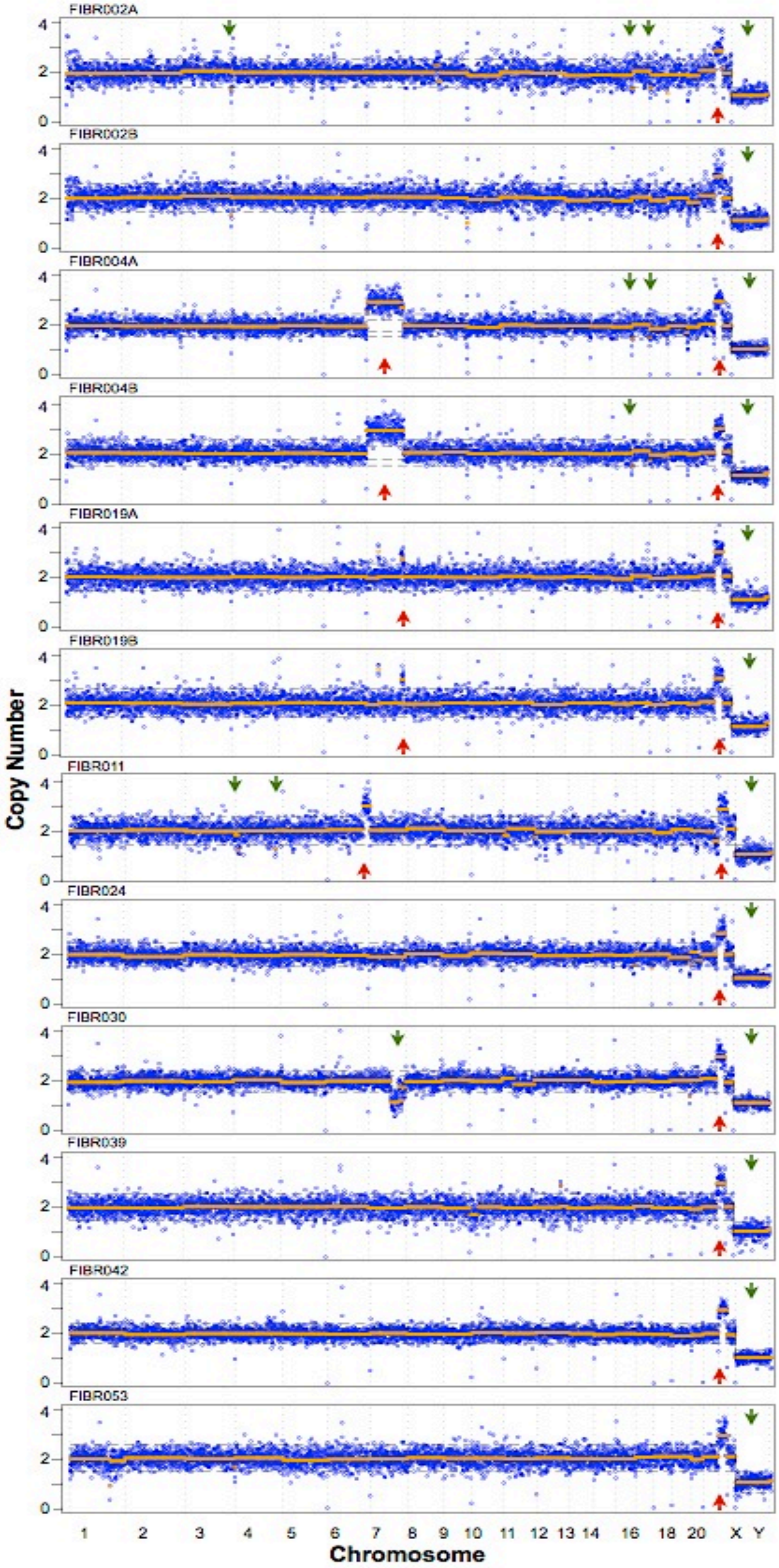
Supplementary Figure S3-3; A, B, ad C provided and analyzed by co-authors

Supplementary Figure S3-3. Large CNVs are found in cultured fibroblasts. (**A**) Single fibroblasts obtained by limiting dilution were expanded to a population of ~20 clonal fibroblasts after 7 days in vitro (DIV). In one clonal population, a reciprocal chromosome missegregation event was detected. One fibroblast was trisomic for Chr2 (top) and a sister was monosomic for Chr2 (bottom). Chromosome 1 is shown alongwith the third euploid cell. (**B** and **C**) Two groups of Df (passages 7 and 8) were summarized in (Fig. 2A); a parallel culture of the p7 group was sent for karyotyping and FISH. Out of 20 metaphase chromosome spreads, 20 were euploid. (B) FISH was performed for a ChrX p arm telomere (green) and ChrX centromere (red). Out of 200 nuclei, 13 were aneuploid. (C) FISH was performed for the Chr20 centromere (green) and Chr20 CNV (red). Out of 200 nuclei, 26 had the CNV. (**D**) Single-cell sequencing of two male fibroblasts with karyotypically defined trisomy 21. Genome-wide copy number profiles show that, in both cells, most of the genome is present at two copies, Chr21 is present at three copies, and ChrX is present at one copy. In addition, we identified a large deletion on Chr7q in FIBR030. DNA copy number (*y* axis) was calculated by read-depth analysis of variably sized genomic windows containing 500 kb of uniquely mappable sequence (blue), and CNVs were detected by circular binary segmentation (orange). Green (down) and red (up) arrows denote deletions and duplications, respectively, that were identified by segmentation and passed filtering criteria. Reported CNVs comprise five or more consecutive bins and exceed two median absolute deviations (MADs). Dotted gray lines show 1 and 2 MADs from the median copy number of each data set.



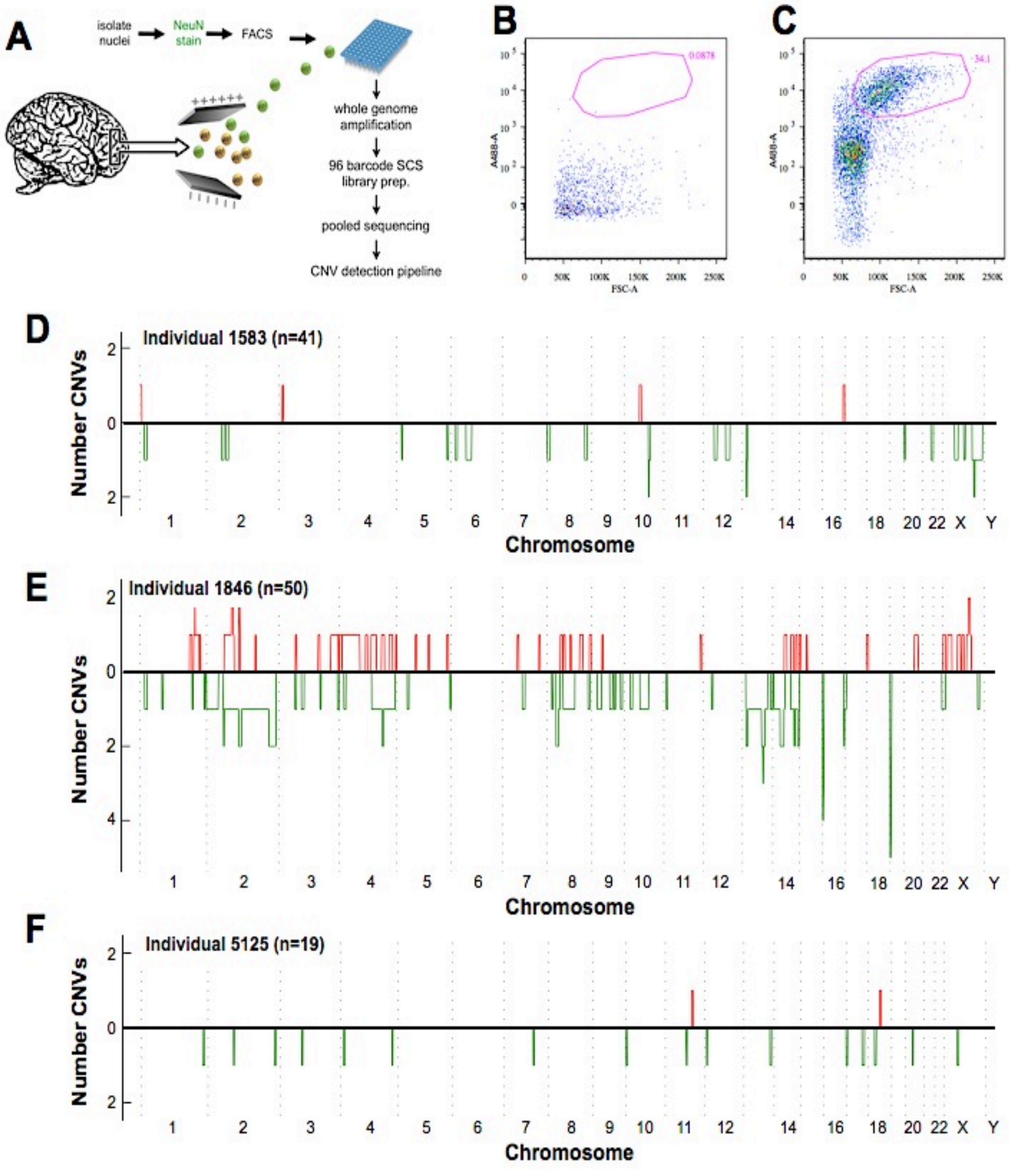
Supplementary Figure S3-4

Supplementary Figure S3-4 Single cell sequencing of trisomy 21 human fibroblasts. **(A)** Flow chart of the single fibroblast sequencing experiment. The top section shows the protocol for sequencing 3 cells in replicate (corresponding to **B** and the top half of **Supplementary Figure S3-5**), accomplished by splitting each sample after 8 cycles of the whole-genome amplification (GenomePlex WGA) PCR step. The standard protocol used for neurons and the 13 single fibroblasts (corresponding to the bottom half of **Supplementary Figure S3-5**) is shown in the bottom section. QC filtering was performed exactly as for neurons. **(B)** Scatter plots comparing concordance between replicate experiments (corresponding to the top half of **Supplementary Figure S3-5**), where each data point represents the predicted copy number of a single genomic window.



Supplementary Figure S3-5

Supplementary Figure S3-5. Identification of CNVs in male trisomy 21 fibroblasts using single cell sequencing. Genome-wide copy number profiles of the three replicate fibroblasts (top) and six single fibroblast cells (bottom). RD analysis, copy number segmentation and CNV filtering were performed exactly as for neurons. These plots follow the conventions of **Figure 3-4**. Blue dots represent the predicted copy number (Y-axis) of each individual genomic window, and orange lines show the results of copy number segmentation. Dotted gray lines show 1 and 2 MADs from the median copy number of each dataset. Reported CNVs comprise five or more consecutive bins and exceed two MADs. Arrows indicate CNV calls that passed filtering criteria (deletions in green and duplications in red).



Supplementary Figure S3-6; S Shumilina assisted in FCTX library preparation

Supplementary Figure S3-6. Single cell analysis of FCTX neurons. **(A)** Flow chart of the protocol. **(B, C)** FACS-based identification of large nuclei that stain positive for NeuN **(C)**, relative to unstained controls **(B)**. Sorted nuclei are gated from the pink circle. **(D, E, F)** Summary of duplications and deletions for each individual (number indicated) plotted as in **Figure 3-5C**. The Y-axis represents the number of times each genomic interval was deleted (below in green) or duplicated (above in red).

A

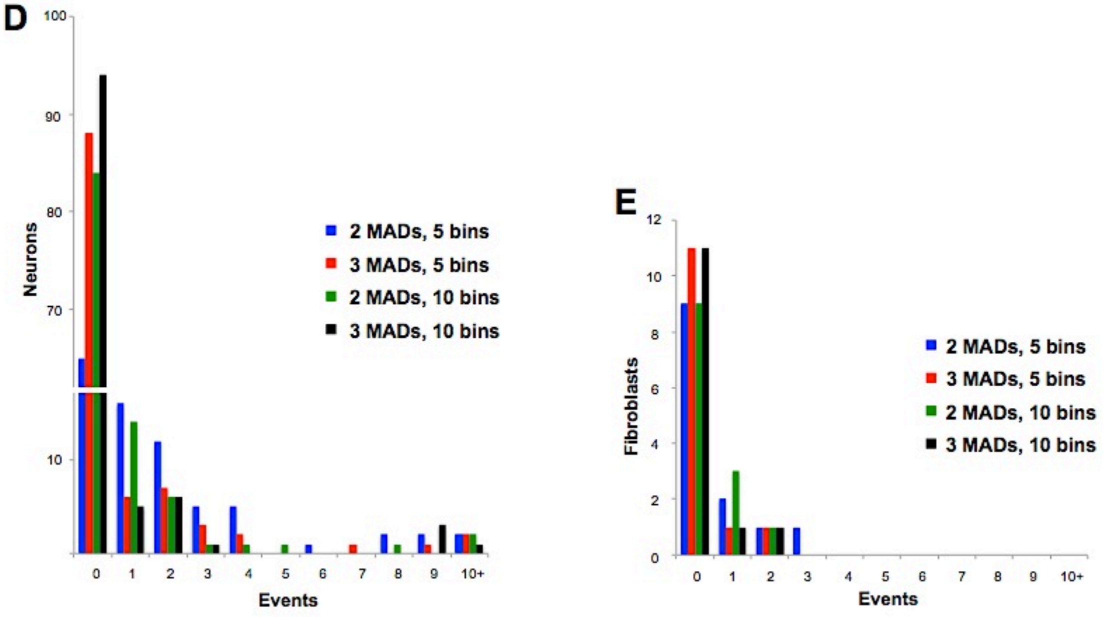
Frontal Cortex Neurons (110 cells)						
Stringency	Total CNV Calls	Deletions	Duplications	Cells w/ >=1 CNV	Predicted FNR	Monosomy X detected
2 MADs, 5 bins	148	98	50	45 (41%)	17%	41 (100%)
3 MADs, 5 bins	83	50	33	22 (20%)	22%	37 (90%)
2 MADs, 10 bins	73	47	26	26 (24%)	7%	41 (100%)
3 MADs, 10 bins	45	29	16	14 (13%)	15%	37 (90%)

B

Fibroblasts - Single Datasets (13 cells)						
Stringency	Total CNV Calls	Deletions	Duplications	Cells with >=1 CNV	Monosomy X detected	Trisomy 21 detected
2 MADs, 5 bins	7	6	1	4 (31%)	13 (100%)	13 (100%)
3 MADs, 5 bins	3	2	1	2 (16%)	10 (77%)	10 (77%)
2 MADs, 10 bins	5	4	1	4 (31%)	13 (100%)	13 (100%)
3 MADs, 10 bins	3	2	1	2 (16%)	10 (77%)	10 (77%)

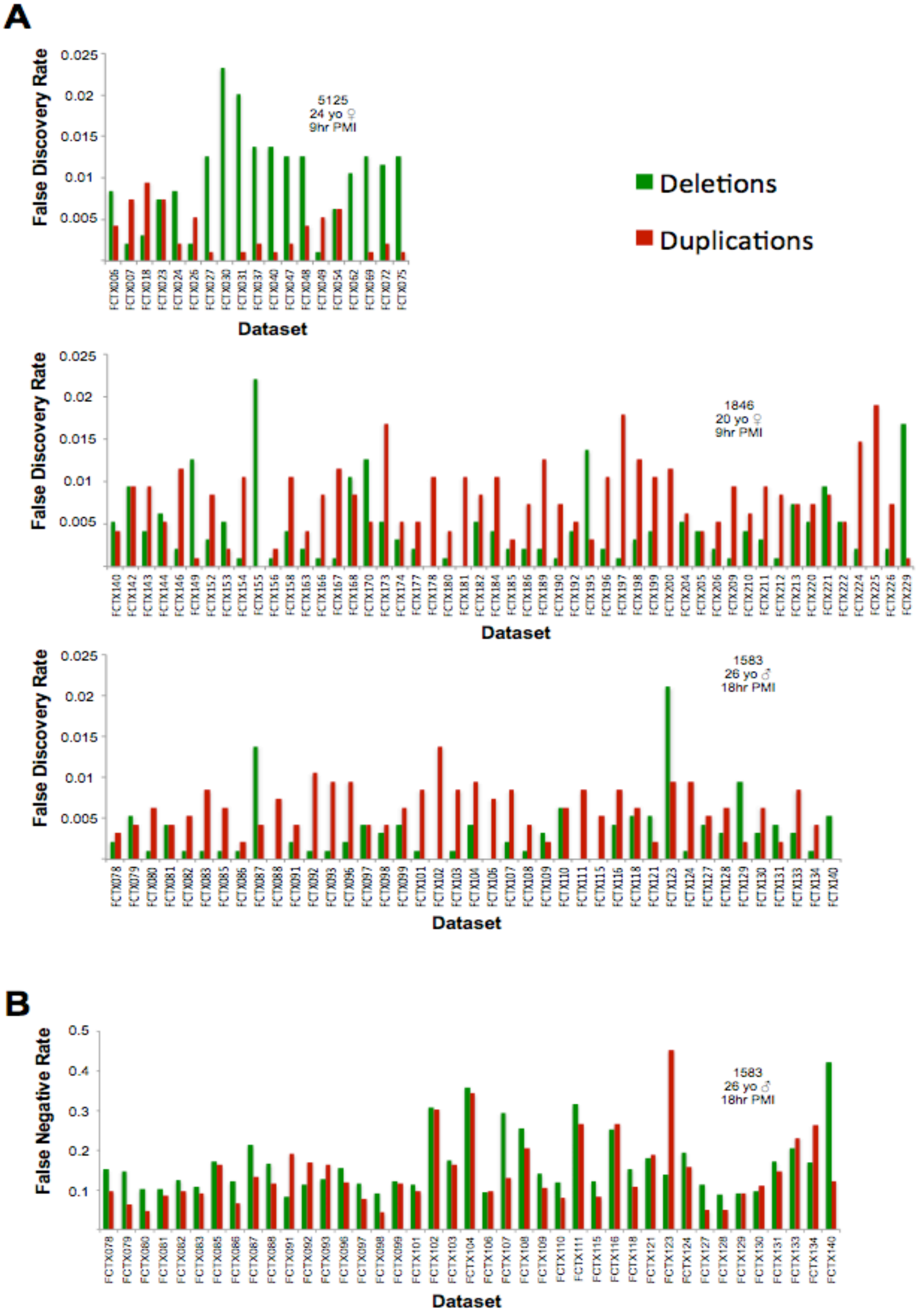
C

Fibroblasts - Replicate Datasets (3 cells, 6 datasets)								
Stringency	Total CNV Calls	Deletions	Duplications	Cells with >=1 CNV	Concordant CNVs	Discordant CNVs	Monosomy X detected	Trisomy 21 detected
2 MADs, 5 bins	7 (11)	2 (5)	5 (6)	3 (100%)	7	4	3 (6)	3 (6)
3 MADs, 5 bins	2 (4)	2 (4)	0 (0)	2 (67%)	3	1	3 (6)	3 (5)
2 MADs, 10 bins	2 (5)	2 (5)	0 (0)	2 (67%)	5	0	3 (6)	3 (6)
3 MADs, 10 bins	2 (4)	2 (4)	0 (0)	2 (67%)	3	1	3 (6)	3 (5)



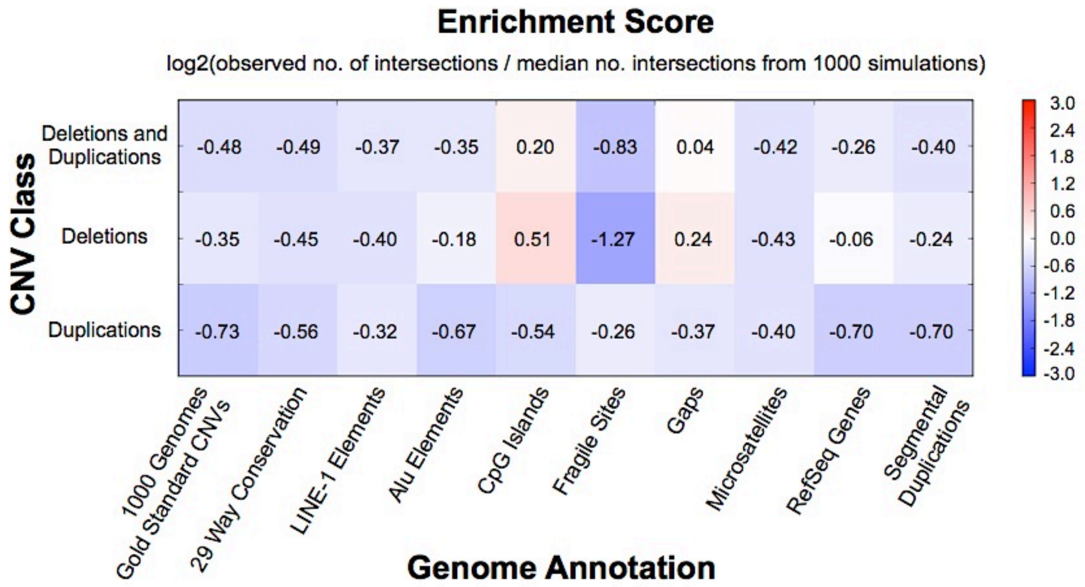
Supplementary Figure S3-7; S Shumilina assisted in FCTX library preparation

Supplementary Figure S3-7. Effect of increased CNV calling stringency. **(A)** Table showing the effect of increasingly stringent CNV detection thresholds on the level and types of CNVs found in FCTX neurons. From left we show the total number of CNV calls detected, the number of deletions and duplications, the number and percentage of cells that were found to have at least 1 CNV, the predicted FNR as calculated in the same manner as for **Supplementary Figure S3-8** (see Materials and Methods), and the fraction of male neurons, which monosomy X was detected at the given thresholds. Note that the false negative rate is calculated using simulated CNVs that are the same size as the minimum number of bins that could be detected according to the bin thresholds at far left (either 5 or 10), and therefore FNR actually decreases with the 10-bin threshold because larger CNVs are easier to detect. **(B)** The effect of increased stringency on the 13 control fibroblast cells. **(C)** The effect of increased stringency on the 3 single fibroblasts subjected to the replicate single cell sequencing experiment. In addition to the columns described above, this table includes the number of concordant and discordant CNVs detected at each indicated threshold. Concordant CNVs are defined as those detected in both replicate cells; discordant CNVs are those detected in merely one replicate cell, according to the filtering thresholds shown at left. In one case two CNV calls in one replicate dataset were concordant with a single call in the pair, hence the odd number of concordant calls. **(D)** Bar chart showing the number of individual neurons (Y-axis) that exhibited a given number of CNVs (X-axis) at the four CNV detection thresholds indicated in the legend. **(E)** Bar chart showing the number of fibroblasts (Y-axis) that exhibited a given number of CNVs (X-axis) at the four CNV detection thresholds indicated in the legend.



Supplementary Figure S3-8

Supplementary Figure S3-8. Estimated FDR and FNR for single cell sequencing experiments. In each case, CNVs were identified using precisely the same methods and criteria as for real data (see Materials and Methods), and the FDR or FNR shown is the mean value obtained from 1000 simulation experiments. Deletions are shown in green and duplications in red. **(A)** FDR for each dataset, as determined by randomly shuffling copy number values across all autosomal bins and then calling CNVs. **(B)** FNR for all cells derived from the male individual (1583). FNR was calculated by randomly selecting 5 contiguous bins from the X chromosome and either replacing (deletion) or adding (duplication) the copy number values from the these bins at a randomly chosen genomic location.



CNV Class	1000 Genomes Gold Standard CNVs	29 Way Conservation	LINE-1 Elements	Alu Elements	CpG Islands	Fragile Sites	Gaps	Microsatellites	RefSeq Genes	Segmental Duplications
Deletions and Duplications	-0.48	-0.49	-0.37	-0.35	0.20	-0.83	0.04	-0.42	-0.26	-0.40
Deletions	-0.35	-0.45	-0.40	-0.18	0.51	-1.27	0.24	-0.43	-0.06	-0.24
Duplications	-0.73	-0.56	-0.32	-0.67	-0.54	-0.26	-0.37	-0.40	-0.70	-0.70

Supplementary Figure S3-9; S Shumilina assisted in FCTX library preparation

Supplementary Figure S3-9. Enrichment of CNV calls at various genome annotations. Monte-Carlo simulations were used to determine whether CNVs identified in post-mortem neurons preferentially overlapped various genomic features. Enrichments are displayed as the log2 ratio of the observed number of intersections between each CNV class (X-axis) and each genome annotation (Y-axis), relative to the expected number of random intersections calculated by the simulations. A positive correlation between CNVs and a given annotation will result in a red-colored positive value; a negative correlation will result in a blue-colored negative value. The highest level of enrichment observed was between deletions and CpG islands, whereas the lowest level of enrichment observed was between deletions and fragile sites.

CHAPTER 4:
FUTURE DIRECTIONS AND ONGOING EXPERIMENTS

The Role of Neuronal Phenotype in Genetic Mosaicism

Background and proposed research

Human neurons that originate from post-mortem brains of healthy individuals harbor somatic DNA CNVs and have been characterized at the single cell level (McConnell *et al.*, 2013; Cai *et al.*, 2014). The mechanisms and effects underlying somatic CNVs in neuronal genomes are still outstanding. While many have offered explanations (Gilbert, Lutz-Prigge, and Moran, 2002; Callinan *et al.*, 2005; Gilbert *et al.*, 2005; Symer *et al.*, 2002; Shibata *et al.*, 2012; Maeda *et al.*, 2004; Shibata *et al.*, 2012; Suberbielle *et al.*, 2013), there is very little agreement or evidence to conclusively identify the causes and consequences of the phenomenon. One way to begin addressing this gap in knowledge would be to analyze neuronal genomes from a wide array of sources. The rationale being that a broader and more developed investigation of neuronal genomic variability could not only serve to uncover incidences somatic variation within individuals, but also between individuals. A characterization of differential or recurrent CNVs in neurons originating from a myriad of sources may elucidate the process and/or biological function of neuronal somatic variation.

In order to make such an undertaking possible, it would be imperative to improve the economy and accuracy of our SCS method. After doing so, an exhaustive study of mosaicism in a cohort of normal individuals could be performed. First, it is possible there is a neurotransmitter-specific and/or spatial influence on the patterns of neuronal genomic mosaicism. In our previous study, we had investigated a paucity (n = 110) of FCTX neurons of unknown neurotransmitter-type. Second, the location (e.g., frontal cortex, cerebellum, hippocampus, etc.) or layer of origin for a given neuron may be related to the prevalence or pattern of somatic CNVs. Current models of corticogenesis state that neurons radially migrate with defined laminar fates (Thomson and Bannister, 2003; Kreigstein and Alvarez-Buylla, 2011). It is then reasonable to believe that the CNVs in each neuron could follow a genetic lineage-specific inheritance pattern. Another, but related, hypothesis is that the microenvironment could dictate neuronal specificity, as well as the expressed neurotransmitter profile. This would then be reflected by the genomic, or perhaps epigenomic, landscape observed in each neuron. The added benefit by performing this experiment is that these data could be recycled as controls for investigating the role of somatic CNVs and mosaicism in complex phenotypes and diseases if no specific enrichment or pattern emerges from the various neuronal types.

Mosaic CNVs have been found to cause rare neurological phenotypes with quantifiable clinical presentations, such as mosaic Down syndrome (Fishler, K. and Koch, R., 1991) and hemimegalencephaly (Poduri *et al.*, 2012; Lee *et al.*, 2012). Accordingly, somatic mosaicism may also be involved in other more prevalent complex neurological conditions. This hypothesis is bolstered by a growing body of evidence implicating somatic variation as the underlying cause of many diseases other than cancer (Poduri *et al.*, 2013). Rare germline and *de novo* CNVs have already been associated diseases such as epilepsy (Olson, H. *et al.*, 2014), autism (Sebat *et al.* 2007; Szatmari *et al.* 2007; Christian *et al.* 2008; Sanders *et al.*, 2012; Iossifov *et al.*, 2014), and schizophrenia (International Schizophrenia Consortium 2008; Kumar *et al.* 2008; Marshall *et al.* 2008; Stefansson *et al.* 2008; Walsh *et al.* 2008; Weiss *et al.* 2008; Kirov *et al.* 2009; Mefford *et al.* 2009; Stefansson *et al.*, 2014; Bundo *et al.*, 2014); however, the precise role and abundance of somatic mosaicism in the development and progression of these complex neurological disorders has not been fully answered.

Recently, a cursory assessment of somatic deletions in bulk brain tissue has been performed to characterize the differences in regional (prefrontal cortex vs. cerebellum) and schizophrenic neuronal genomes (Kim *et al.*, 2014). The authors identified 106 putative somatic deletions by sequencing bulk tissues to very high-depth. The study was perfunctory and inconclusive to many questions raised here; yet, the basic experimental design of Kim *et al.* is useful as a model. The experimental design could be modified to employ single cell methods, rather than bulk tissue, and then adapted to other disorders such as age-related neurodegenerative diseases. Diseases with an age onset dependency, may relate the effect of aging on the incidence of somatic CNVs and disease. Currently, it is not definitively known as to whether the number of somatic CNVs increases with age. Our previous study, McConnell *et al.*, comprised of "normal" or "neurotypic" individuals ranging in age 20 to 26. The youngest individual, unexpectedly, exhibited the most CNVs and most CNVs per neuron. While this observation may have been due to a sampling bias (i.e., this individual had the most neurons sequenced making it possible to discover the population of cells with highly aberrant genomes), but there may be genuine biology underpinning this observation. Genome instabilities, particularly in microsatellites (Eshleman, J. R. and Markowitzm, S. D., 1995), have been found to be inherited and will often manifest themselves sporadically during the lifetime of an individual. Therefore, it is plausible that there is an undiscovered relationship between age and the abundance of somatic CNVs in neurons.

Materials and methods

The following methods are modified from those found in the manuscript:
MJ McConnell, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Shumilina S, Lasken R, Vermeesch, J, Hall IM, and Gage F. "Mosaic copy number variation in human neurons." *Science* (2013), 342(6158):632-7.

Experiments and analyses were developed and/or performed by MR Lindberg under the guidance of IM Hall and MJ McConnell unless otherwise noted.

A modified *Isolation of post-mortem neuronal nuclei* (see Chapter 3: Materials and methods, pg. 62; McConnell *et al.*, 2013) was performed on FCTX tissue from one of the three used in the McConnell *et al.* study, UMB#1846 (a neurotypic 20-year-old female, 9 hour post-mortem interval). The isolation procedure was performed identically until FACS (performed by the University of Virginia Flow Cytometry Core). At this stage, single nuclei from the NeuN and DAPI positive population were sorted based on their relative DNA content as reported by DAPI. The cells from the "left" and "right" tails of the distribution were gated deposited into 96 well plates alongside 1 water control per row. The "left" and "right" tails were the extremes of the DNA content distribution, determined by the percentage (around 5-10%) of events relative to the whole population of cells. These cells would ostensibly have an overall DNA content less than or greater than the center of the distribution. Ian Blurbis provided MALBAC sequencing libraries. The SCS procedure and subsequent analyses were consistent with the McConnell *et al.* protocols in all experiments.

Preliminary results

In the McConnell *et al.* study, we reported 110 FCTX neuronal genomes; however, well over 400 sequencing libraries were prepared. This rate (nearly one in four) was predicated on the QC measures imposed on the sequencing libraries. The first QC requirement eliminated nearly one-third of sequencing libraries prepared, as they did not produce a high enough DNA concentration, and could not be sequenced. The remaining fraction failed because they did not pass the MAD and confidence score thresholds (see Chapter 3: Materials and methods, pg. 67; McConnell *et al.*, 2013). The low success rate was tolerable for this study because SCS is a nascent technology and the goal of our study was to simply detect the presence of somatic mosaicism in human neurons. However, a broader study with several hundred or even thousands of neurons derived from a variety of sources would be prohibitively expensive if one were to use those in employed in McConnell *et al.*.

Here, we have begun to explore two ways to potentially make SCS more affordable. Both approaches are based on the fact that minimizing the number of cells analyzed will substantially lower costs. The first approach identifies the fact that the most liberal CNV detection parameters putatively show that only 43% of neurons contained at least one CNV (McConnell *et al.*, 2013), or that zero CNVs were detected in over half of all neurons. This means that if a study were to be concerned with neurons with CNVs, not the rate of occurrence, then one could preferentially analyze the cells with the most aberrant genomes. We have made preliminary attempts to do exactly this by using FACS on FCTX neurons stained for NeuN and DAPI (see Methods and materials) to select ostensibly aneuploid neurons only. The basis of the approach comes from the fact that neurons with aberrantly measured FACS DNA content would likely be aneuploid. As such, we gated on the tails (left and right) or edges of the cell cycle normal distribution measuring DNA content (NeuN positive nuclei are in G_0 because the are derived from post-mitotic neurons). These tails were thought to reflect the relative extremes of the DNA content distribution as measured by DAPI binding. Previous reports have claimed that neuronal genome content measured in cells of the right tail of the distribution make up an enriched subpopulation of about 250 Mb extra DNA (Westra *et al.*, 2010), albeit we have yet to observe any enriched subpopulation or a non-normal neuronal DNA content distribution. Regardless, the left and right tails were sorted alongside neurons from the center of the distribution and sequencing libraries were prepared.

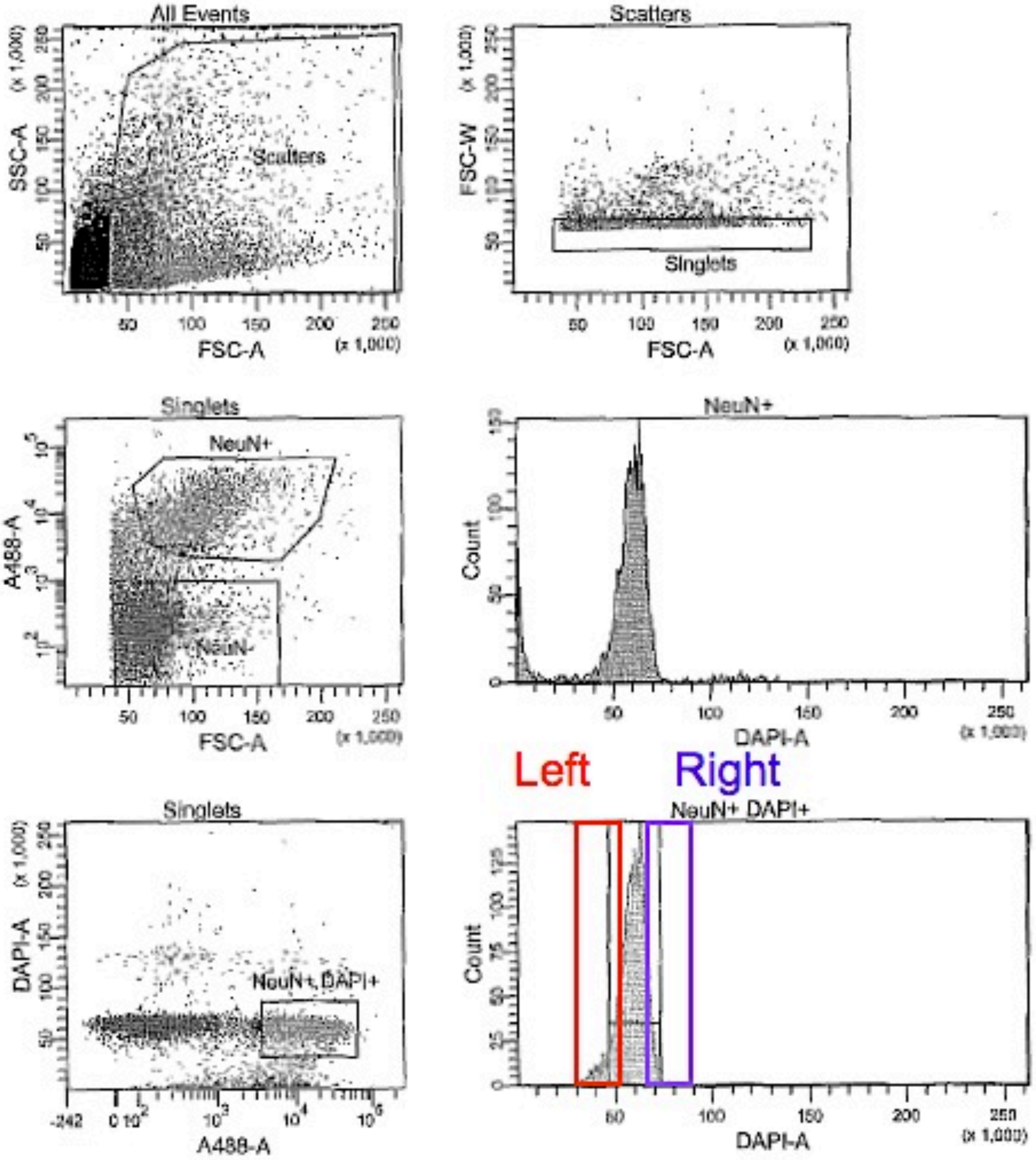


Figure 4-1

Figure 4-1. FACS sorting of FCTX neurons. The forward and side scatters of all events are shown and nuclei are selected in the singlet distribution. Nuclei populations are isolated by gating DAPI positive and NeuN positive events. Single nuclei are then selected from the left (red) and right (purple) populations of DNA content and then analyzed with SCS.

Unfortunately, we found that the libraries prepared from the left and right tails pass QC measures at a markedly lower rate than the standard SCS approach (less than 10%). The sequencing libraries from the left and right tail sequencing libraries were also abundant in duplicate sequencing molecules, which would require far deeper sequencing to obtain unique measurements for necessary CNV detection (data not shown). Moreover, the neurons sorted from the center of the distribution did not these exhibit similar characteristics and displayed typical duplicate and QC failure rates seen in previous experiments. Although there may be true biology underpinning these observations, using this approach would be more expensive than our current protocol. It is more likely that these differences between the tails and center of the cell cycle distribution technical in nature and this avenue may be cost-effective in the future if these technical artifacts could be understood and skirted.

A second strategy for increasing the economy of SCS is to directly improve the quality of sequencing libraries. An increase in the proportion of cells passing the imposed QC requirements would reduce the burden of library preparation and number of cells sequenced. Changing the step that introduces the most technical variation and errors will make for the most effective improvements of SCS, (i.e. initial amplification of genomes). Recently, the MALBAC method was used to amplify the genomes of single cancer cells and sperm with high fidelity (Zong *et al.*, 2012; Lu *et al.*, 2012). Not only do the authors boast a high signal to noise ratio, they also claimed to have robustly detected single nucleotide variation. MALBAC or another method could allow additional variant analysis and discovery using different signals and evidence. Our preliminary applications of MALBAC have not yet yielded results meeting the QC thresholds, but furthered fine-tuning of MALBAC will likely yield high-quality data, making it a promising prospect.

Discussion

While technically impressive, the SCS assay is expensive and still being developed. There are necessary advances in the protocol before scaling to a larger, more comprehensive study of neurons from a spectrum of sources. Since a concern of accomplishing any research project is the incurred cost, the two aforementioned avenues aim to improve SCS by reducing the number of sequencing libraries prepared and sequenced, thereby abating the total financial burden. The first approach, preferentially selecting for a higher putative mutational load or aneuploidy would be a more expensive method in its current practice. Although this approach may be detecting true genomic variation in the neuronal nuclei that appear to have higher and lower DNA content by FACS, it is more likely that technical limitations occlude a faithful and accurate analysis. Artifacts may originate from any step in the nuclear isolation or FACS itself and would require extensive investigation. Therefore, more robust initial whole-genome amplification (i.e., MALBAC) would likely be a more viable option for improvement. Not only would such a method drive down costs, it could also provide the ability to detect SNVs or obtain base-pair resolution of breakpoints. Conceivably, *Hydra-Multi*, or another SV caller, could be used to simultaneously analyze the genomes of neurons given sufficiently high data quality. Finally, the continued development of sequencing platforms may allow higher accuracy by not using a whole genome amplification and/or PCR-free preparation. With an improved SCS method or platform, a broad study on many individuals, including those with diseases, could be performed. The work proposed here will provide insight into the possible causes and effects of mosaic variation in human neurons.

Somatic Variation and Mosaicism in Human Cardiac Myocytes

Background and proposed research

The full extent of somatic variation in human tissues remains largely unknown. We have used SCS to show that normal human neuronal genomes contain mosaic CNVs (McConnell *et al.*, 2013); however, many other tissues have yet to be analyzed. We have begun the cursory evaluation of healthy human myocardium, or heart muscle, and have found that it is possible that the human heart may also harbor somatic variation and mosaicism. The investigation of the heart is an important step towards understanding the entire landscape of somatic variation in human tissues.

We had chosen to analyze cardiac myocytes because heart muscle cells bear some superficial similarities to neurons. Foremost, muscle cells in the myocardium undergo electrochemical action potentials like neurons in the brain. While cardiac potentials are different than neuronal, the persistent electrophysiological stress that both cells undergo may serve as a common source of DNA double-stranded breaks. This electrophysiological stress has already been identified in neurons (Suberbielle *et al.*, 2013). An additional commonality between the cell types is the turnover of cells in adult hearts and brains. Cardiac myocytes, like the neurons, experience limited replacement and regeneration over a lifetime (Bergmann, O. *et al.*, 2009; Gage, F. and Temple, S. 2013). This low rate of cellular replacement and relative longevity of the constituent cells in these organs may contribute to the accumulation of aberrant genomes in senescent cells. Somatic variation in affected tissue has been reported to be a common pathology between congenital heart defects (Reamon-Buettner, S. and Borlak, J., 2004; Reamon-Buettner, S. and Borlak, J., 2004; Reamon-Buettner, S. *et al.,* 2004, Reamon-Buettner and Borlak, J. 2006). It is therefore reasonable to hypothesize that age-related cardiomyopathies and cardiovascular disorders could be the result of acquired somatic variation.

Here, we demonstrate the beginnings of an investigation into somatic variation in the human myocardium and show evidence for their putative existence. We applied SCS methods to otherwise healthy human heart cells and found large-scale CNVs, affecting whole chromosomes or chromosome arms. These preliminary results show that tissues may contain a genetic mosaic and warrants deeper exploration. Future experiments would affirm the existence of somatic variation in human heart muscle cells. Following works would consist of a survey of various myocardial tissue sources originating from both healthy individuals and those with cardiomyopathies.

Materials and methods

The following methods are modified from those found in the manuscript:
MJ McConnell, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Shumilina S, Lasken R, Vermeesch, J, Hall IM, and Gage F. "Mosaic copy number variation in human neurons." *Science* (2013), 342(6158):632-7.

Experiments and analyses were developed and/or performed by MR Lindberg under the guidance of IM Hall and MJ McConnell unless otherwise noted. Co-authors provided isolated heart cells and S Shumilina assisted in sequencing library creation.

A modified *Isolation of post-mortem neuronal nuclei* (see Chapter 3: Materials and methods, pg. 62; McConnell *et al.*, 2013) was performed on human myocardial tissue from one of the three used in the McConnell *et al.* study, UMB#5125 (a 24-year-old female, 9 hour post-mortem interval). The isolation procedure was performed identically until the FACS. At this stage, single cells were instead stained with only Propidium Iodide and sorted solely on DNA content (performed by co-authors). Following the modified FACS, the SCS procedure and subsequent analyses were consistent with those in Chapter 3 and McConnell *et al.*. Additional MAD and confidence score calculations were used to reflect the ploidy differences in the heart cells. The MAD calculation was calculated by taking the weighted average of all segment MADs. A second confidence score was also used.

$$max(S) = 1 - 2\,\frac{\sum_{i=0}^{n} \min(\lceil C_i \rceil - C_i - j, C_i - \lfloor C_i \rfloor - j)}{n}$$

Confidence Score, *S*:

$$\text{for } j = 0.01 \ldots 1.00$$

C: the median predicted copy number of a given genomic interval (*i*) after copy number segmentation

n: the total number of genomic windows in the dataset

j: each iteration between j_1 and $j_{n,}$ (e.g., 0.01 to 1.00)

This confidence score reflects a step-wise approximation to determine the highest confidence score in an iterative process. The ploidy approximation per dataset is adjusted based on the maximum confidence score.

Preliminary results

We sought to perform a cursory analysis of the prevalence of somatic variation in the human heart. A similar SCS strategy to the one found in Chapter 3 and McConnell *et al.* was used to investigate the genome of single cardiac myocytes. The FACS isolation strategy and staining was altered slightly, consisting of only staining nuclei with Propidium Iodide and sorting on DNA content. Future work can easily improve the FACS single cell isolation step used in these preliminary analyses. After sequencing 76 cells from the heart, an abysmal 13 of 76 datasets passed using the previously circumscribed QC filters of a MAD less than or equal to 0.35 and a confidence score greater than or equal to 0.85. To improve this, bioinformatic corrections to the analysis would require changing the genomic ploidy assumption and making the QC filters reflect that change.

Currently, we see that these 13 cells passing previous QC did not exhibit a single somatic CNV. This low pass rate likely stems from the incongruous ploidy assumption that was applied when analyzing neurons. These 13 cells all had a ploidy of 2N, but many of the other datasets of the 76 had genomes that appeared to be other than 2N. These genomes contained very large alterations consisting of entire chromosomes and chromosome arms; these massive CNVs provide a total genomic content that was consistently fewer than 46 chromosomes. By removing the MAD filter, we see that 23 of the 76 datasets would have passed filtering criteria. This may be due to the large losses in these cells, which increase the value of the dataset autosomal MAD. Adjustments on how MAD threshold is calculated and/or using a different signal-to-noise metric will be needed for an accurate assessment. The global autosomal MAD is not an appropriate absolute measurement when the genomes are highly rearranged. A possible solution is to simply calculate the local signal-to-noise ratios using the appropriately weighted average of all segment MADs. This metric would not be influenced by global fluctuations between chromosomes and represents only local differences. Using this strategy, 22 datasets had a weighted MAD lower then 0.35. For example, HERT002 (Figure 4-2) is a characteristic dataset that has a high autosomal MAD of 1.66, but the average local-adjusted segment MAD is 0.28. The confidence score of HERT002 was also low (0.33), which is also due to the current ploidy estimation as the multiplier after GC normalization is 2 for 2N. To ameliorate this, increments of 0.01 were subtracted iteratively from the copy number values to find the maximum confidence score. The values at each genomic bin can then be adjusted by subtracting the step size with the highest confidence score. With the maximum confidence scores, 39 genomes had a confidence score over 0.85. This procedure may be used to explore the space of ploidy multipliers suitable for each dataset after GC normalization; which may even be more accurate than the previously described method in all analyses. Many of these genomes passing the new MAD and confidence score measures contained numerous whole-chromosome and arm deletions. Overall, we detected 3 copy number states, denoting loss of either one or both copies at many loci in heart cells. These aberrant copy number profiles are greatly enriched in heart cells compared to neuron

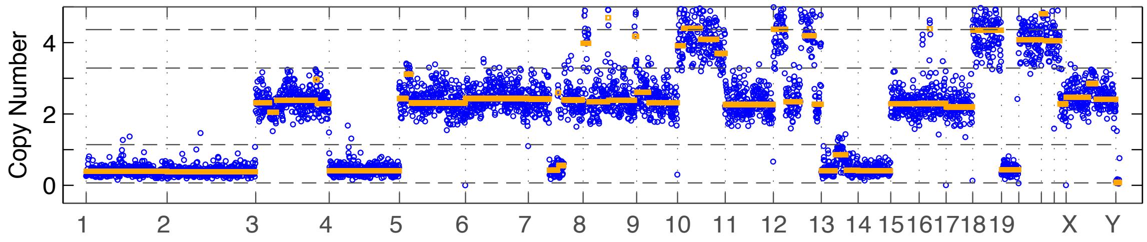


Figure 4-2.

Figure 4-2. Genome-wide CNV profile of a human heart cell. Genome-wide copy number profile of a typical human heart cell sequenced. Read-depth analysis, copy number segmentation and CNV filtering were performed exactly as for neurons. Blue dots represent the predicted copy number (Y-axis) of each individual genomic window, and orange lines show the results of copy number segmentation. Dotted gray lines show 1 and 2 MADs from the median copy number of each dataset. Reported CNVs comprise five or more consecutive bins and exceed two MADs.

It is important to confirm the origin of the nuclei with these numerous large-scale CNVs. In our preliminary experiments, we naively sorted nuclei from heart tissue, even though the heart is composed of multiple cell types (i.e., fibroblasts, epithelial, smooth and cardiac muscle cells). Given the frequency of aberrant cells and the abundance of cardiac myocytes in the heart, it is likely that at least some of these aberrant cells are in fact cardiac myocytes. To validate this one could refine the FACS by using cardiac myocyte-specific markers such as cTroponin T or I antibody (Bergmann, O. *et al.*, 2009) in combination with very conservative gating. This will provide verification of cell types because cTroponin positive events would be cardiac myocytes. We anticipate that these genomic profiles will be similar to the ones we have already observed; however, if cardiac myocyte nuclei do not produce these CNV profiles, and exhaustive attempt to discover the cell-type origin of aberrant nuclei, and their relationship to the heart will be necessary.

Having performed these essential experiments and confirmed the nuclei producing these profiles are indeed from cardiac myocytes, a larger study of somatic variation in the heart can be performed. Sequencing heart cell nuclei from 2 relatively young (<30 years) and 2 old (>70 years) individuals and identify CNVs. Sequencing such a large number of nuclei from 4 individuals will allow one to address a number of important questions regarding somatic variation in the heart. For example, are CNV-laden heart cells observed in all individuals, or is the first individual an exceptional case, or is it really just a technical artifact? Does the frequency of CNVs in heart cells vary among human individuals? Does the frequency of CNVs increase with age? Are certain chromosomes more prone to aneuploidy and CNVs than others? Are there defined “hotspots” in the genome, where CNVs arise at very high rates? How do the level and patterns of somatic CNV in the heart compare to other tissues such as neurons?

There are follow-up experiments that can confirm and extend the SCS observations. For example, given the frequency of aberrant cells in preliminary experiments, it should be possible to detect losses of the most variable chromosomes or chromosomal regions using simple FISH experiments. By probing for the copy number of 3 different chromosomes (or loci) using 3-color FISH, it will be possible to observe a frequency of loss that is similar to our SCS data, as the very large-scale CNVs should be easily ascertained by FISH. Additionally, differentiating hiPSC cells to be become cardiac myocyte-like cells, much like the hiSPC neurons previous described, can also enable validation of the SCS results.

Discussion

This unexpected result in the human mycardium is novel and very interesting, but it is also difficult to explain. Continuing forward, it will be necessary to rule out all possible artifacts. It is not possible to explain these large-scale copy number differences by known SCS artifacts, such as uneven amplification, since amplification artifacts produce much smaller-scale fluctuations. These CNVs cannot also not be due to DNA contamination, since we perform nuclei-free negative controls in each experiment, and these exhibit very few reads that align to the reference genome. Interestingly, genomes with aberrant CNV profiles could be multi-nucleated or have undergone polyploidization during cardiac myocyte development (Adler and Frideburg, 1986). For example, if CNV generation is related to these processes, CNV-laden nuclei will be enriched in certain cardiac myocytes, rather than being distributed randomly as expected under the null model.

Taken together, performing these experiments will lead to a better understanding of the levels and origin of somatically acquired copy number variation in the human heart, and will suggest future studies aimed at determining the presence and functional consequences of this extraordinary phenomenon.

REFERENCE LIST

Abyzov A. and Gerstein M.B. (2011) AGE: Defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics *27*, 595-603.

Abyzov, A. *et al.* (2012) Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature *492*, 438-442.

Adey, A. *et al.* (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol *11*, R119.

Adler, C.P. and Friedburg, H. (1986) Myocardial DNA content, ploidy level and cell number in geriatric hearts: post-mortem examinations of human myocardium in old age. J Mol Cell Cardiol *18*, 39-53.

Alkan, C., Coe, B. P. and Eichler, E. E. (2011) Genome structural variation discovery and genotyping. Nature reviews genetics *12*, 363-376.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. J. Exp. Med. *79*, 137–157

Baillie, J. K. *et al.* (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Baslan, T. *et al.* (2012) Genome-wide copy number analysis of single cells. Nature Protocols *7*, 1024-1041.

Bergmann O. *et al.* (2009) Evidence for cardiomyocyte regeneration in humans. Science *324*, 98-102.

Bergmann, O. and Frisen, J. (2013) Why adults need new brain cells. Science *340*, 695-696.

Boveri, T. (1902) Über mehrpolige mitosen als mittel zur analyse des zellkerns. Verh. phys.-med. Ges. *35*, 67-90.

Brack, C. *et al*. (1978) A complete immunoglobulin gene is created by somatic recombination. Cell *15*, 1-14.

Brennand, K. J. *et al.* (2011) Modelling schizophrenia using human induced pluripotent stem cells. Nature *473*, 221-225.

Bruder, C. E. *et al.* (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. American journal of human genetics *82*, 763-771.

Bundo, M. *et al.* (2014) Increased L1 retrotransposition in the neuronal genome in schizophrenia. Neuron *81*, 306-313.

Bushman, D. M. and Chun, J. (2013) The genomically mosaic brain: Aneuploidy and more in neural diversity and disease. Semin Cell Dev Biol *24*, 357-369.

Cai, X. *et al.* (2014) Single-cell, genome-wide sequencing identifies clonal somatic copy number variation in the human brain. Cell Reports *8*, 1280-1289.

Callinan, P. A. *et al.* (2005) Alu retrotransposition-mediated deletion. J Mol Biol *348*, 791-800.

Chen, K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods, *6*, 677–681.

Christian, S. L. *et al.* (2008) Novel submicroscopic chromosomal abnormalities detected in autism spectrum disorder. Biol Psychiatry *63*, 1111-1117.

Correns, K. G. (1900). Ueber levkojenbastarge. Ber. dtsch. bot. Ges. *18*, 158.

Coufal, N. G. *et al.* (2009) L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127-1131.

Darwin, C. R. and Wallace, A. R. (1858). On the tendency of species to form varieties ; and on the perpetuation of varieties and species by natural means of selection. Journal of the Proc. of the Linn. Soc. of London *3*, 45-62.

Dawson, E. *et al.* (2001) SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. Genome Res. *11*, 170–178.

Dean, F. B. *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America *99*, 5261-5266.

de Vries, H. (1900). Ber. dtsch. bot. Ges. *18*, 83.

Durbin, R. M. *et al.* (2010) A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Eshleman, J. R. and Markowitz, S. D. (1995). Microsatellite instability in inherited and sporadic neoplasms. Current Opinion in Oncology *7*, 83-89.

Evrony G. D. *et al.* (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483-496.

Faust, G. F. and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 30, 2303-2305.

Fishler, K. and Koch, R. (1991) Mental development in down syndrome mosaicism. American Journal of Mental Retardation *96*, 345-351.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006) Structural variation in the human genome. Nature reviews genetics *7*, 85-97.

Fleischmann, R. D. *et al.* (1995) Whole-genome random sequencing and assembly of haemophilus influenzae rd. Science *269*, 496-512.

Fungtammasan, A. *et al.* (2012) A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? Genome Research *22*, 993-1005.

Gage, F. and Temple, S. (2013) Neural Stem Cells: Generating and regenerating the brain Neuron *80*, 588-601.

Gandi, M. Evdokimamova V., and Nikiforov Y. E. (2010) Mechanisms of chromosomal rearrangement in solid tumors; the model of papillary thyroid carcinoma. Mol Cell Endorinol *28*, 36-43.

Gilbert, N., Lutz, S., Morrish, T. A., and Moran, J. V. (2005) Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol *25*, 7780-7795.

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

Griffith, F. (1928) The Significance of Pneumococcal Types. J Hyg (Lond). *27*, 113–159.
Handsaker, R.E., Korn, J. M., Nemesh, J. and McCarroll, S. A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nature Genetics *43*, 269-276.

His, W. et al., (1869). Die histochemischen und physiologischen arbeiten von friedrich miescher—aus dem wissenschaftlichen briefwechsel von miescher, 33-38.

Hormozdiari, F., Hajirasouliha, I., McPherson, A., Eichler, E. E. and Sahinalp, S. C. (2011) Simultaneous structural variation discovery among multiple paired-end sequenced genomes. Genome Research *21*, 2203-2212.

Hosono, S. *et al.* (2003) Unbiased whole-genome amplification directly from clinical samples. Genome Research *13*, 954-964.

Huxley, J. S. (1942) Evolution: the modern synthesis. London: Allen and Unwin.

International Human Genome Sequencing Consortium (2001) Nature *409*, 860-921.

International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature *455*, 237-241.

Jacobs, P. A. and Strong, J. A. (1959) A case of human intersexuality having a possible XXY sex-determining mechanism. Nature *183*, 302-303.

Kim, J. *et al.* (2014) Somatic deletions implicated in functional diversity of brains cells of individuals with schizophrenia and unaffected controls. Scientific Reports *4*, e3807.

Kirov, G. *et al.* (2009) Support for the involvement of large CNVs in the pathogenesis of schizophrenia. Hum Mol Genet *18*, 1497-1503.

Koboldt, D. C. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics *25*, 2283-2285.

Korbel, J. O. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science *19*, 420-426

Korbel, J. O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. *10*, R23.

Kreigstein, A. and Alvarez-Buylla, A. (2009) The glial nature of embryonic and adult neural stem cells. Ann Rev Neurosci *32*, 149-184.

Kumar, R. A. *et al.* (2008) Recurrent 16p11.2 microdeletions in autism. Hum Mol Gen *17*, 628-638.

Lander, E. S. and Weinberg, R. A. (2000). Journey to the center of biology. Science *287*, 1777-1782.

Larson, D. E. *et al.* (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*, 311–317.

Lasken, R. S. (2009) Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochem Soc Trans *37*, 450-453.

Lasken, R. S. and Stockwell, T. B. (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol *7*, 19.

Laurent, L. C. *et al.* (2011) Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. Cell Stem Cell *8*, 106-118.

Layer, R. M. *et al.* (2013) Binary Interval Search: a scalable algorithm for counting interval intersections. Bioinformatics *29*, 1-7.

Lee, J. H. *et al.* (2012) De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat. Genet. *44*, 941-94.

Lee, S. *et al.* (2010) MoGUL: detecting common insertions and deletions in a population. Proc RECOMB 2010 *6044*, 357-368.

Li, H. (2013) Aligning sequence reads, clone sequences, and assembly contigs with BWA-MEM. *arXiv*, 1303.3997, (http://arxiv.org/pdf/1303.3997v2.pdf)

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

Li, H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, H. and Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics *25*, 1754-1760.

Li, Y. *et al.* (2012) Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. GigaScience *1*, 12.

Lupski, J. R. (2013) Genome mosaicism – one human, multiple genomes. Science *341*, 358-359.

Lindberg, M. R., Hall, I. M., and Quinlan A. R. (2014) Population-based structural variation discovery with Hydra-Multi. Bioinformatics, *epub ahead of print.*

Lu, S. *et al.* (2012) Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. Science *338*, 1627-1630.

Maeda, T. *et al.* (2004) Somatic DNA recombination yielding circular DNA and deletion of a genomic region in embryonic brain. Biochem Biophys Res Commun *319*, 1117-1123.

Mantikou, E. *et al.* (2012) Molecular origin of mitotic aneuploidies in preimplantation embryos. Biochimica et Biophysica Acta *1822*, 1921-1930.

Malhotra, A. *et al.* (2013) Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. Genome Research *22*, 762-776.

Marshall, C. R. *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet *82*, 477-488.

Martin, S. L. (2009) Developmental biology: Jumping-gene roulette. Nature *460*, 1087-1088.

Maxam, A. M. and Gilbert W. (1977) A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America *74*, 560-564.

McConnell, MJ *et al.* (2013) Mosaic copy number variation in human neurons. Science *342*, 632-637.

McKenna, A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research *20*, 1297-1303.

Mefford, H. C. *et al.* (2009) A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. Genome Research *19*, 1579-1585.

Mendel, G. (1866). Versuche über pflanzen-hybriden. Brünn Natural History Society, *3*-47.

Meyer, L. R. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Research *41*, D64-D69.

Mills, R. E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59-65.

Morgan, T. H. (1910) Sex-limited inheritance in Drosophila. Science *32*, 120-122.

Mullikin, J. C. *et al.* (2000) A SNP map of chromosome 22. Nature *407*, 516-520.

Muotri, A. R. *et al*. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903-910.

Navin, N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. Nature *472*, 90-94.

O'Huallachain, M., Karczewski, K. J., Weissman, S. M., Urban, A. E., and Snyder, M. P. (2012) Extensive genetic variation in somatic human tissues. Proceedings of the National Academy of Sciences of the United States of America *109*, 18018-18023.

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

Olson, H. *et al.* (2014) Copy number variation plays an important role in clinical epilepsy. Ann Neurol, *epub ahead of print.*

Onishi-Seebacher, M. and Korbel, J. O. (2011) Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. Bioessays *33*, 840-850.

Ostertag E. M. and Kazazian H. H. (2005) Genetics: LINEs in mind. Nature *435*, 890-891.

Piotrowski, A. *et al.* (2008) Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat *29*, 1118-1124.

Poduri, A. *et al.* (2012) Somatic activation of AKT3 causes hemispheric developmental malformations. Neuron *74*, 41-48.

Poduri, A., Evrony, G. D., Cai, X., and Walsh, C. A. (2013) Somatic mutation, genomic variation, and neurological disease. Science *341*, 6141.

Pugh, T. J. *et al*. (2008) Impact of whole genome amplification on analysis of copy number variants. Nucleic acids research *36*, e80.

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

Quinlan, A. R. and Hall, I. M (2012) Characterizing complex structural variation in germline and somatic genomes. Trends in Genetics : TIG *28*, 43-53.

Quinlan, A. R. *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome research *20*, 623-635.

Quinlan, A. R. *et al.* (2011) Genome sequencing of mouse induced pluripotent stem cells reveals retroelement stability and infrequent DNA rearrangement during reprogramming. Cell Stem Cell *9*, 366-373.

Rausch, T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333-i339.

Reamon-Buettner, S. M. and Borlak J. (2004) Somatic NKX2-5 mutations as a novel mechanism of disease in complex congenital heart disease. J Med Genet. *41*, 684–690.

Reamon-Buettner S.M. and Borlak J (2004) TBX5 mutations in non-Holt-Oram syndrome (HOS) malformed hearts. Hum Mutat. *24*, 104.

Reamon-Buettner S.M. *et al.* (2004) Novel NKX2-5 mutations in diseased heart tissues of patients with cardiac malformations. Am J Pathol. *164*, 2117–2125.

Reamon-Buettner, S.M. and Borlak J. (2006) HEY2 mutations in malformed hearts. Hum Mutat. *27*, 118.

Rehen, S. K. *et al.* (2001) Chromosomal variation in neurons of the developing and adult mammalian nervous system . Proceedings of the National Academy of Sciences of the United States of America *98*, 13361-13366.

Rehen, S. K. *et al.* (2005) Constitutional aneuploidy in the normal human brain. J Neurosci *25*, 2176-2180.

Sachidanandam, R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928-933.

Sanger F. and Coulson A. R. (1977) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. *94*, 441-448.

Sanger, F., Nicklen, S., and Coulson A. R. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463–5467.

Shaffer L. G. and Bejjani B. A. (2004) A cytogeneticist's perspective on genomic microarrays. Hum Reprod Update *10*, 221-226.

Sebat, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. Science *316*, 445-449.

Shibata, Y. *et al.* (2012) Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. Science *336*, 82-86.

Sindi, S. *et al.* (2009) A geometric approach for classification and comparison of structural variants. Bioinformatics, *25*, 222-230.

Singer, T., McConnell, M. J., Marchetto, M. C., Coufal, N. G., and Gage, F. H. (2010) LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci 345-354.

Spalding, K. L., Bhardwaj, R. D., Buchholz, B. A., Druid, H., and Frisen, J. (2005) Retrospective birth dating of cells in humans. Cell *122*, 133-143.

Stefansson, H. *et al.* (2008) Large recurrent microdeletions associated with schizophrenia. Nature *455*, 232-236.

Stefansson, H. *et al.* (2014) CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature *505*, 361-366.

Stephens, P. J. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27-40.

Stephens, P. J. *et al.* (2012) The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400-404.

Sturtevant, A. H. (1913) The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. Journal of experimental zoology, *14*, 43-59

Suberbielle, E. *et al.* (2013) Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-beta. Nat Neurosci *16*, 613-621.

Sutton, W.S. (1902) On the morphology of the chromosome group in brachystola magna. Biol. Bull. *4*, 24-39.

Symer, D. E. *et al.* (2002) Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327-338.

Szatmari, P. *et al.* (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Gen *39*, 319-328.

The C. Elegans Sequencing Consortium (1995) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science *282*, 2012-2018.

Tjio, J. H. and Levan, A. (1956) The chromosome number of man. Heriditas *24*, 1-6.

Thomson, A. M. and Bannister, A. P. (2003) Interlaminar connections in the neocortex Cerebral cortex *13*, 5-14.

Tschermak von Seysenegg, E. (1900). Ber. dtsch. bot. Ges. *18*, 158.

Vanneste, E. *et al.* (2009) Chromosome instability is common in human cleavage-stage embryos. Nat. Med. *15*, 577-583.

Venter, J. C. *et al.* (2001) The sequence of the human genome. Science *291*, 1304-1351.

Wang, J. *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods *8*, 652–654.

Walsh, T. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science *320*, 539-543.

Weber, J. L. *et al.* (2002) Human diallelic insertion/deletion polymorphisms. Am. J. Hum. *71*, 854–862.

Weiss, L. A. *et al.* (2008) Association between microdeletion and microduplciation at 16p.11.2 and autism. N Engl J Med *358*, 667-675.

Westra, J. W. *et al.* (2010) Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. The Journal of Comparative Neurology *518*, 3981-4000.

Xu, X. *et al.* (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, 886-895.

Ye, K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Youssoufian, H. and Pyeritz, R. E. (2002) Human genetics and disease: Mechanisms and consequences of somatic mosaicism in humans. Nature Reviews Genetics *3*, 748-758

Yurov, Y. B. *et al.* (2007) Aneuploidy and confined chromosomal mosaicism in the developing human brain. PloS one *2*, e558.

Zhao, M. *et al.* (2013) Computational tools for copy number variation detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14:S1.

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