Improved Structural Variation Detection Facilitates the Discovery of Somatic Mutations in Single Post-Mitotic Neurons

Gregory Gerard Faust
Charlottesville, Virginia

Bachelor of Science
Computer and Information Science
University of Michigan
Dearborn, Michigan
May 1977

Master of Science
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts
February 1981

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
May 2015

Abstract

The extent and origin of somatic cell genome diversity is a question of central importance to human biology and disease, and one about which surprisingly little is known. Somatic mutations are involved in tumor formation, are implicated in many developmental and neurological diseases, and have been suggested as a mechanism driving the vast diversity of morphology and stochastic interconnections exhibited by neurons. Conventional genome-wide methods applied to bulk tissue samples are ill suited for somatic variant detection. Such samples contain diverse cell types and intermixed lineages, making it difficult to distinguish somatic mutational patterns in a specific cell type, or the clonal prevalence of those mutations. Examination of single cell genomes avoids these problems, but current methods lack sensitivity.

We make two important methodological improvements to somatic mutation discovery, and use them to study somatic mutations in single post-mitotic neurons. First, we utilize a novel experimental design that allows deep sequencing of single cell genomes by forming clonal cell populations derived by somatic cell nuclear transfer and enculturation. The resultant sequencing data allows investigation of single cell somatic mutations with unprecedented resolution. Second, we improve bioinformatic methods for the detection of structural variation. Due to the complexity of structural variants, methods for their discovery have lagged behind those used to identify single nucleotide polymorphisms and small insertions and deletions. Improvements in these methods are useful in general. But they are particularly important for the study of post-mitotic neurons, as we wish to investigate the long proposed hypotheses that the diversity in neuronal morphology and connectivity patterns may be due to structural variations akin to V(D)J recombination in the immune system and/or high levels of mobile element transposition. By applying these two new methods, we find that each neuronal genome harbors hundreds of private somatic mutations that likely arose during late development or post-mitotic aging, and that many somatic structural variants are complex events defined by multiple clustered breakpoints. We also demonstrate that neither programmed or recurrent mutations, nor mobile element insertions, are likely to be a major mutational force shaping neuronal genome diversity.

Dedication

This dissertation is dedicated to the memory of my father, Vincent Faust. He instilled in me at an early age his desire to understand how everything around him worked. I hope I have been able to pass on this sense of curiosity and wonder to my own children.

Acknowledgments

I started my graduate studies at UVA in the Computer Science Department. I would like to thank Kevin Skadron, my CS thesis advisor who helped me through the process of passing my CS qualifying exam, was the inspiration for ArchFP, a coauthor on the ArchFP publication, and introduced me to Ira Hall. I would also like to thank Gabe Robbins, Marty Humphrey, and Wes Weimer for sitting on my CS qualifying exam committee. I would also like to thank Mark Sherriff who mentored me through teaching Introduction to Computer Programming to 80 students.

I would like to thank Anindya Dutta and especially Joel Hockensmith for making it easy for me to switch to the Biochemistry and Molecular Genetics Department. Ira Hall has been my de facto and ultimately my actual thesis advisor for all of my bioinformatics projects starting from my first semester at UVA. He has taught me almost everything I know about the field, and he has been the corresponding author on all related publications. I can't thank him enough for all his contributions towards my success. I would also like to thank Aaron Quinlan and Bill Pearson who both sat on my thesis committee and had many useful comments especially about YAHA, and Hui Zong who agreed to sit on my thesis committee as the GSAS representative.

I would like to thank members of Ira's lab for their camaraderie and interesting conversations and contributions on various projects. These include students and post-docs Mike Lindberg, Ryan Layer, Colby Chiang, Mitchell Leibowitz and Ankit Malhotra, and staff members Royden Clark and Svetlana Shumilina. I would like to thank all of the collaborators on the mouse neuron project from Scripps including Kristin Baldwin and especially Jennifer Hazen, and the rest of the Scripps team.

Finally, I would like to thank my wife Paulyn Heinmiller, and my two children Eric and Elaina Faust for their support throughout this often trying process, with an additional thanks to Eric for his insights on programming issues related to SAMBLASTER and YAHA.

1 Introduction and Background

The completion of a high quality draft human reference genome by the Human Genome Project in 2001 has ushered in a new era of discovery of all aspects of human genetics (Lander et al., 2001). One such area is the study of variation in genomes between individuals. Previously, it had been thought that the vast majority of genetic variation was in the form of single nucleotide polymorphisms (SNPs), the replacement of a single nucleotide in a DNA sequence with a different nucleotide. It has been found that we each differ from one another by approximately three million SNPs. However, a surprise finding since the completion of the Human Genome Project has been the number and complexity of structural variations (SVs), an insertion, deletion, duplication, or inversion of at least 50 consecutive nucleotides. Roughly 1,000 SVs distinguish the genomes of two normal humans, and within the human population these variants collectively affect more bases of the genome than SNPs (Conrad et al., 2010; Mills et al., 2011).

The idea that each cell in our body contains identical DNA is an oversimplification. It is estimated that there are over 10 trillion cells in the human body (Baserga, 1985; Bianconi et al., 2013), requiring at least as many cell divisions to produce. The eukaryotic DNA replication and repair machinery produces approximately one base error per billion base pairs replicated (McCulloch and Kunkel, 2008), which predicts an average of ~6 *de novo* mutations per diploid mammalian genome per cell cycle. Empirical studies are in close agreement, with estimates of the actual single nucleotide mutation rate ranging from 1 to 5 base substitutions per replication cycle for healthy somatic cells (Behjati et al., 2014; Holstege et al., 2014; Lynch, 2010; Welch et al., 2012). The mutation rate for small insertions and deletions in simple repeat regions of the genome may be as much 10X higher (Frumkin et al., 2005). Therefore, an adult human is necessarily composed of genetically distinct populations of somatic cells, a condition first termed somatic mosaicism by Cotterman in 1956 (Cotterman, 1956).

In addition to replication errors, there are many other mutagenic processes that can add to the somatic mutational burden, and therefore the overall genetic mosaicism of an organism. Reactive oxygen species (ROS) are a normal byproduct of oxidative phosphorylation in mitochondria, and are highly mutagenic (Dizdaroglu, 2012). Exposure to ionizing radiation can generate ROS, and also directly cause thymine-thymine and other nucleotide dimers and single and double strand breaks in DNA (Svobodova et al., 2012). People in industrialized societies are exposed to an ever increasing number of chemical. Tobacco smoke alone contains >60 mutagenic compounds (Pfeifer et al., 2002). Structural mutations can occur during replication and/or single or double strand break repair, many of which result in unique mutational signatures (**Figure 1.1**). These include mis-segregation or recombination of chromosomes (Youssoufian and Pyeritz, 2002), replication slippage, non-allellic homologous recombination (NAHR), breakage-fusion-bridge, nonhomologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), fork stalling and template switching (FoSTeS) and microhomology-mediated break induced replication (MMBIR) (Hastings et al., 2009). Mitochondrial DNA is subject to higher mutation rates than nuclear DNA, therefore single cells often contain mosaic mitochondria, a condition called heteroplasmy (Youssoufian and Pyeritz, 2002). Recently, circular extra-chromosomal DNA fragments 200-400bp in length and associated microdeletions have been found to be prevalent in many cells in a mosaic fashion (Shibata et al., 2012). Another source of somatic

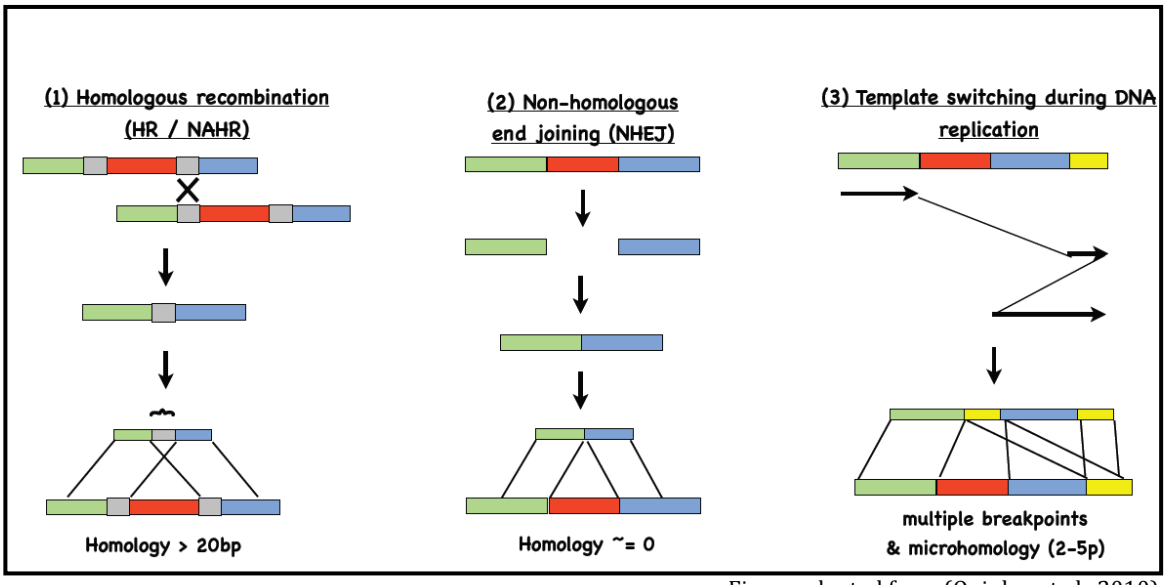

Figure adapted from (Quinlan et al., 2010)

Figure 1.1. Different mutational processes result in different signatures.
1) A deletion caused by homologous recombination results in >20bp of microhomology at the deletion breakpoint in the test sample. 2) NHEJ ligation after a deletion leaves no appreciable microhomology. 3) Template Switching leads to a complication genomic rearrangement involving a deleted region and a out of order duplication. Breakpoints have low microhomology.

mutations is the inclusion of DNA from retroviruses or endogenous retrotransposons into the genome (Kazazian, 2011). Finally, the adaptive immune system utilizes programmed mutations including V(D)J recombination and hypermutation of T-Cell receptors of immunoglobulins (Di Noia and Neuberger, 2007). As we will discuss in detail below, retrotransposon activity and programmed mutations have long been hypothesized to contribute to the diversity of neuronal morphology and stochastic interconnectivity.

In what follows, we will use the following abbreviations; base-pair (bp), thousand bp (Kbp), million bp (Mbp) and billion bp (Gbp).

There are a variety of types of mutations that can occur depending on a combination of their intrinsic and/or extrinsic causes. Mutations are categorized by the nature of the change to the DNA, and include *aneuploidy*, the addition or deletion of one of more whole chromosomes; *structural variation* (SV), the insertion, deletion, (tandem) duplication, inversion, or translocation of >50bp of contiguous DNA (**Figure 1.3**); *copy number variation* (CNV), a general term for an increase or decrease in the number of copies of a genomic region including both aneuploidy and unbalanced structural variants; *indels*, small insertions and deletions of between 1-50bp; and *single nucleotide variation* (SNV), the replacement of a single base-pair by another. Mitotic crossover can result in *loss of heterozygosity* (LOH). A particularly severe form of LOH occurs when two copies of one of a pair of sister chromosomes mis-segregate into a daughter cell. When this occurs during gametogenesis, it is called *uniparental disomy* (UPD). If this occurs as a somatic mutation during mitosis it is called *acquired UPD* (aUPD).

Mobile elements insertions (MEIs) are structural variants that are a particularly important category of somatic mutation leading to genetic diversity. There are several varieties of mobile elements that have over time contributed aproximately 45% of our genome. These include DNA transposons, mobile elements (MEs) that include long-terminal-repeat (LTRs), long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), remnants of which contribute approximately 3%, 8%, 17%, and 11% of our genetic material respectively (Cordaux and Batzer, 2009). Almost all of these insertions are millions of years old, have drifted through accumulated mutations, and are no longer active. L1 LINE elements are by far the most prominent ME, with ~80-100 active copies in the human genome, and ~3000 in the mouse genome (Muotri et al., 2005). L1 elements are self-contained as they encode the proteins needed for their own retrotransposition within the genome, and can also can transpose SINE elements. As these MEIs can occur essentially anywhere in the genome, they can cause the full range of effects on gene expression including disrupting exons, creating cryptic stop signals or alternate splice sites in introns, or alter gene regulatory regions (Cordaux and Batzer, 2009).

1.1 Significance of studying somatic mosaicism

Somatic mosaicism is a topic of growing interest, as there is increasing evidence for its frequency in apparently healthy tissue and its implications in disease states. In many cases, these somatic mutations are structural variants. In this section, we will briefly present the results from a variety of recent studies aimed at better characterizing the prevalence and phenotypic consequences of somatic mosaicism. We will then revisit these topics in greater detail in **Section 1.3** and **Section 1.4** after first discussing the experimental techniques available for studying somatic mutations in **Section 1.2**.

It is now clear that cancer is caused by one or more somatic mutations in a cell lineage leading to abnormal cell division and/or migration (Watson et al., 2013). In the two-hit model of cancer genesis, somatic structural variants often account for at least one of the causative mutations. Classic examples involve LOH either by mitotic crossover or deletion in retinoblastoma and BRCA-related breast cancers. Also, many cancers are caused by somatic gene fusions due to translocations, and 100s of such fusion genes have now been catalogued (Mitelman et al., 2007). Accurate characterization of the somatic mutations present in subclonal populations of tumor cells will be increasingly critical to precision treatment choices, especially given that relatively small populations of cells are often refractory to treatment or become metastatic.

More generally, the phenotypic changes that result from somatic mutations often depend on the genetic background of the individual (Gottlieb et al., 2001), which may account for some of the missing heritability of complex disorders (Manolio et al., 2009). Somatic CNVs have been implicated in the onset and severity of Alzheimer's disease (Beck et al., 2004; Freed et al., 2014), and a wide variety of other neurological disorders including autism spectrum disorder, bipolar disorder, attention-deficit-hyperactivity disorder, obsessive-compulsive disorder, depression, anxiety and panic disorder (Sebat et al., 2009; Weiss et al., 2008). Somatic MEIs have also been implicated in a number of neurological and neurodegenerative disorders including ALS (Douville et al., 2011), schizophrenia, bipolar disorder and major depression (Bundo et al., 2014), Rett syndrome (Muotri et al., 2010), Fragile X-associated tremor/ataxia syndrome (Tan et al., 2012), ataxia telangiectasia (Coufal et al., 2011), and others. That fact that somatic MEIs have also been associated with normal neuronal development suggests that there may be a sensitive balance between ME activity in healthy neurons, and ME over activity leading to pathology. Further studies are needed to understand under what circumstances that line is crossed.

Somatic mutations have been found in apparently normal tissue in healthy individuals (De, 2011). Studies of embryos show surprisingly high levels of mosaic aneuploidy (Bielanska et al., 2002; Kano et al., 2009; Vanneste et al., 2009). Blood samples show somatic CNV, which usually increasing with age (Jacobs et al., 2012; Laurie et al., 2012), and displays changing levels of clonal prevalence over time (Holstege et al.). Somatic mosaicism in monozygotic twins has similarly shown increased divergence with age that can lead to discordant phenotypes (Bruder et al., 2008; Forsberg et al., 2012). Behjati et al. used an experimental technique similar to ours to precisely characterize somatic mosaicism in clonal cell populations grown from gastro-intestinal stem cells, including SNVs (Behjati et al., 2014).

Somatic mosaicism in healthy neurons has been extensively studied due to the ongoing uncertainty concerning the importance of somatic mosaicism to neuronal diversity. Somatic aneuploidy, CNV, and MEIs have all been observed in healthy neurons in humans, mice, and flies. However, different studies provide a remarkably wide range of estimates for the prevalence of these somatic mutations, depending on the cell type being studied and the experimental technique used. Estimates for the percent of aneuploid neurons range from 3% to 35% (Rehen et al., 2001; Rehen et al., 2005; Yurov et al., 2007). Similarly, 16%, 69%, and 100% of neurons have been reported to be harboring somatic CNVs (Cai et al., 2014; Gole et al., 2013; McConnell et al., 2013). Estimates of the average number of somatic MEIs that occur in neurons differ by three orders of magnitude from 0.07 to ~80 (Baillie et al., 2011; Coufal et al., 2009; Evrony et al., 2012; Evrony et al.; Muotri et al., 2005; Muotri et al., 2009; Upton et al., 2015). These wide ranging estimates highlight the need for additional

studies with improved experimental design to clarify if this variability is due to differences in experimental technique, or represents actual biological complexity.

1.2 Somatic Mutation Detection

Somatic mutation detection is difficult for two separate but related reasons. First, the biology is complex. The prevalence of a somatic mutation in the body depends of several factors; 1) the time during morphogenesis the mutation occurred, 2) the cell type in which it occurred, 3) close intermixing of different cell types and lineages within tissues, 4) clonal expansion or death of cell lineages over time, 5) cell migration, and other factors (Freed et al., 2014; Youssoufian and Pyeritz, 2002). Second, available experimental techniques have limitations. Current genome-wide methods require a large number of cells to obtain sufficient DNA, typically from a blood or tissue sample that is often comprised of mixed cell types and lineages, making it difficult to detangle the biological complexity. To simplify things, studying mosaicism at the single cell level is often preferred. But available techniques limit experiments to the study of large effects such as aneuploidy or large CNVs (McConnell et al., 2013), studying only pre-identified target mutations such as retrotransposon insertions (Evrony et al., 2012), lack sensitivity and accuracy and/or have limited throughput thereby restricting the number of cells that can be studied. In addition, current single-cell techniques often consume all available DNA from the original cell therefore making it impossible to validate their specific mutations. As a result, the distribution of somatic mutations by cell type and mutational category remains poorly understood.

Here we present a novel experimental design for the study of somatic mosaicism in post-mitotic neurons that shares many of the advantages of both genome-wide and single-cell techniques. We use somatic cell nuclear transfer (SCNT) of DNA from post-mitotic neurons of known origin in the mouse olfactory bulb to seed clonal colonies of cells. We can then harvest sufficient DNA from these cultures to perform deep whole genome sequencing of each neuronal genome in order to detect the entire landscape of somatic mutations including SNVs, indels, SVs, CNV, and MEIs to single base-pair resolution with sensitivity and accuracy unprecedented for single-cell experiments. In addition, we validate every one of our SV and MEI calls, and a representative sample of our SNV and indel calls. Therefore we can correctly characterize the accuracy and sensitivity of our findings.

1.2.1 Somatic Mutation Detection without DNA Sequencing

Historically, the tools available for studying somatic mutations in single cells have been limited to the observation of chromosome-scale events (Bushman and Chun, 2013). Since the mid to late 1800s, researchers could directly observe the karyotype of single eukaryotic cells via direct observation with light microscopy of stained chromosomes during metaphase. Stains that distinguish between adjacent chromosomal regions (bands) can be used to distinguish some SVs down to ~5-10Mbp in length. By the early 1900s, long before the discovery of DNA as the molecular substrate of genes, researchers began to suggest that chromosomes might be the carriers of genetic material based on observation of the fate of aneuploid cells (Carlson, 2004), culminating in the seminal paper by Morgan et al. in 1915 (Brush, 2002). Remarkably, in that same timeframe, Theodor Boveri was among the first to suggest that tumors were created by *de novo* somatic mutations, based on his observations of karyotypes in tumor cells and his study of the creation of aneuploid cells by mis-segregation of chromosomes in multi-polar mitotic events (Boveri, 2008). Decades later, Barbara McClintock published her famous findings based on cytological studies of the mechanisms causing somatic mosaicism in maize. She found that many chromosomal

transposition are involved some caused by breakage-fusion-bridge cycles, while others involve elements that would later be recognized as retrotransposons (McClintock, 1951). Cytogenetic techniques are still in use today, however the number of cells that can be examined with these techniques is severely limited.

A more modern technique for the direct observation of aneuploidy in single cells is spectral karyotyping (SKY) which uniquely labels genomic fragments on each chromosome with distinct fluorochromes, thereby easing the identification of specific chromosomes (Schrock et al., 1996). For more specific chromosome loci detection, increasingly complex variations of fluorescent *in situ* hybridization (FISH) are being used to probe specific genomic regions of interest without requiring the cell to be in metaphase (Iourov et al., 2005). Depending on the choice of FISH probes used, CNVs as small as several Kbp can be detected, albeit at the cost of choosing predetermined loci to examine (Vorsanova et al., 2010). These techniques can now be combined with computerized analysis of the resultant images to improve throughput significantly.

Comparative genomic hybridization (CGH) is a related technique that attaches control and sample DNA to different labels, and co-hybridizes them to a collection of probes. Now, most such probes are placed in a large array in a technique called array CGH (aCGH). The difference in signal strength for the control and sample labels at specific probes identifies differences in DNA copy number at the associated loci. This technique can be used to compare different tissue from the same individual, for example a tumor/normal pair (Pinkel and Albertson, 2005). By carefully comparing the sample signal to a predetermined reference, somatic mutations appearing in as little as 10-20% of the cells from a blood sample can be detected (Ballif et al., 2006). Modern CGH arrays can detect CNVs down to ~50Kbp is size. SNP arrays (SNP-CHIPs) are closely related to CGH arrays, but use probes designed to detect common single nucleotide polymorphism (SNPs). Such an array can be used to test for CNVs as in CGH at least at the selected loci, but can in addition detect common disease causing SNPs and also loss of heterozygosity (LOH). The drawback of these techniques is that they use cell samples, and are difficult to use to study somatic mosaicism in single cells without DNA amplification.

Quantitative PCR (qPCR) is a technique that detects the amount of DNA amplified by specific primers selected to amplify genomic loci of interest for a given experiment. For example many experiments described below use qPCR with L1 specific primers to estimate somatic L1 insertion rates. Typically it is difficult to tell the absolute amount of the target DNA in the sample unless one also normalizes the results relative to a known quantity of DNA that is spiked into the mixture. This requires specific probes for both the target DNA and the reporter DNA that fluoresce at different frequencies. Like normal PCR reactions, qPCR is performed in a thermal cycler, but one modified to be able to measure the fluorescence after each cycle (Heid et al., 1996).

The total amount of DNA in each cell of a larger sample can be obtained using flow cytometry of cells stained with Propidium Iodide, Ethidium Bromide or other dyes. This has been used to distinguish cells in S and G2/M vs. other phases of the cell cycle to gauge mitotic activity (Laerum and Farsund, 1981), and has sufficient resolution to distinguish X vs. Y chromosome-bearing sperm prior to *in vitro* fertilization (Johnson, 1995). This technique has also been used to measure DNA content variation (DCV) in various cell types including neurons (Westra et al., 2010). The downside of this technique is that it cannot

detect balanced mutations such as conservative translocation or inversions, nor where in the genome the DCV occurs.

1.2.2 DNA Sequencing Technology

The advent of DNA sequencing technology revolutionized the study of many aspects of genetics and ushered in the closely related field of genomics. There are often four separate steps involved in identifying mutations from source DNA. First, the determination of the order of bases in a fragment of sample DNA is called *sequencing*. Second, *assembly* is the process used to create a single sequence of tens of millions of bases that appear in a chromosome of a reference genome from millions of much shorter 100s to 1000s of base-pair long sequences generated by available DNA sequencing technologies. Third, *alignment* is the process used to find the best match between two DNA sequences represented as strings of the letters ATCG, where one string is usually a set of one or more reference sequences representing the DNA for an organism, and the other is a DNA read (or "query") from a sequencing run. Finally, *variant detection* uses a number of bioinformatic methods to identify differences between two samples of DNA, or between a sample of DNA and a reference genome.

The invention of Sanger sequencing in 1977 simplified what had previously been an entirely manual biochemical process. The sequence of bases in fragments of DNA of approximately 100-1000bp in length is determined by a process called *chain termination*. The fragments are replicated by DNA polymerase using a mixture of the four normal triphosphate bases (dNTPs) that act as monomers for DNA replication, and a much smaller concentration of radiolabeled variants of one of these bases that has been chemically altered to prohibit further processivity of the replication process. With sufficient input DNA, the stochastic incorporation of the terminating bases produce an admixture of fragments of different lengths, which are then separated by gel electrophoresis and used to expose a photographic plate. By running four such gels, one for each radiolabeled terminating base, the entire sequence of the original DNA fragment could be ascertained by visual inspection (Sanger et al., 1977). Although this process was much faster and worked on longer DNA sequences than any previous technology, it was still required many manual steps. Over the next 20+ years, many improvements of this general process were incorporated into ever faster, cheaper and more reliable methods (Mardis, 2013). In 1986, radiolabeled terminating bases were replaced with ones that fluoresced at four different frequencies. This allowed all four bases to be run in the same mixture, and removed the need for drying gels, exposing film, and visual inspection to read the base sequences. Instead, the sequence could be read directly by a computer (Smith et al., 1986). Combined with the use of much thinner gels, the sequence of many input DNA fragments could now be read in parallel. This was the state of the art at the onset of the publically funded Human Genome Project to assemble the first human reference genome (Lander et al., 2001). By 1998, at the start of the privately funded parallel effort to assemble a human reference genome (Venter et al., 2001), the process had been further refined. In particular, the electrophoresis gels were replaced by small capillary tubes whose location were more fixed than columns in a gel, and could be loaded mechanically (Marsh et al., 1997). This added another dramatic increase in throughput, allowing both efforts to complete drafts of the human genome by 2001. Sanger sequencing, now often called *capillary* sequencing, is still is use today especially for 500bp-1000bp segments of DNA from a few loci. It is very reliable, especially when the input DNA fragments are redundantly sequenced from both strands in opposite directions.

With the advent of new sequencing technologies discussed below, the cost and speed of DNA sequencing has continued to improve at an exponential rate, which has made feasible many new genomics projects. Reference genomes have been produced for other species, especially so-called model organisms, including mouse (Waterston et al., 2002), fruit fly (Celniker and Rubin, 2003), yeast (Cherry et al., 2012), and dozens of other species (See (Mardis, 2011) for a review). In addition, many projects are now mapping the genetic diversity within human and other animal populations. Here we mention just a small sampling of some of the international efforts. The 1000 Genomes Project (1000 Genomes Project Consortium et al., 2012) has already exceeded its goal of sequencing the genomes of 1000 individuals from diverse genetic backgrounds. The HapMap project is identifying and cataloging SNPs from individuals across ethnic backgrounds, and assessing how they are linked into common haplotypes (International HapMap Consortium, 2003). Efforts are also underway to map structural variation across many genomes (Mills et al., 2011). The ENCODE project is adding many functional annotations to regions of the human reference genome (ENCODE Project Consortium, 2012). COSMIC is cataloging somatic mutations found in cancer (Forbes et al., 2011). The Cancer Genome Atlas (TCGA) project is collecting in-depth information about cancer genomes including full sequencing data, as well as somatic mutations, gene expression profiles and epigenetic markers (Cancer Genome Atlas Research Network, 2008). The Personal Genome Project is collecting full genomes of patients with publically annotated medical histories and symptomologies to aid in precision medicine efforts (Ball et al., 2012). Many countries also have internal initiatives to map diversity within their own populations. Of particularly interest to our own study of somatic mutations in mouse neurons, common SNPs and small indels (Keane et al., 2011) and structural variants (Quinlan et al., 2010) have been identified across mouse strains.

Next-generation or *massively-parallel* sequencing (NGS) dramatically both improves the biochemical preparation protocols and increases overall throughput of sequencing technologies. All NGS technologies share many features. The sample DNA is first broken up into random fragments using mechanical agitation followed by selection of fragments lengths suitable for the particular technology; a technique called *whole genome shotgun* (WGS) sequencing. General attachment and primer sequences are then ligated to the ends of the fragments. The fragments are then isolated on a surface via annealing of the attachment sequence. Each fragment is then amplified via PCR to ensure sufficient DNA to provide enough signal during sequencing. Sequencing itself is then a cycle of several steps. First there is a replication step, followed by a detection step, then a cleanup step to prepare for the next cycle (Mardis, 2013; Pettersson et al., 2009).

Three sequencing technologies utilize a similar technique to capture DNA fragments on beads before amplification. Roche 454 technology can sequence fragments up to 500bp in length (Margulies et al., 2005). It isolates each bead in a water bubble suspended in oil and places each bubble into a separate well. In each replication cycle, only one type of base is added to the mixture with no attached moiety to block replication. Therefore the signal, measured by the amount of pyrophosphate released, is proportional to the number of consecutive bases at the replication point that match the added NTP. The sensing of homopolymers is error prone, and the technology tends to incorporate small insertions and especially deletions into the reported sequence compared to the actual sample. Ion Torrent places each bead into a well on a semiconductor chip that acts as a pH meter, measuring the H+ ions released. It also uses non-terminating NTP one base at a time and therefore also suffers from indel errors reading homopolymers. It is able to produce reads of ~200bp long with an error rate of ~1% (Rothberg et al., 2011). The SOLiD technology ligates 2 bases in

each replication cycle using 4x4 or 16 different flourophores in such as way that each base is read twice, leading to the lowest error rates. The relatively long 200-500bp read length of 454 and Ion Torrent sequencing is favorable for many applications such as SV detection.

By far the most common and now most advanced NGS technology is from Illumina (Bentley et al., 2008). It anneals 400-500bp to a glass slide, and increases the copies of each fragment by bridge-amplification (Adessi et al., 2000). In each replication cycle, all four bases with different attached flourophores are added to the mixture, and exactly one incorporated into each fragment. The base added to each locus is read via a CCD camera. Once the sequence from one end of the fragment is read, the sequencing starts over at the other end of the bridge, allowing for the sequencing of paired-end reads. In 2005, the error rate of Illumina sequencing became quite high after about 30-35bp, but now at least 100bp can be reliable read from each end of the fragment with error rates below 0.5%.

All of the above NGS technologies suffer from artifacts from the biased PCR reactions used to increase the local copies of a DNA fragment prior to sequencing. In particular, regions of low and high GC content are not well amplified, and fragments from such areas tend to be underrepresented in the output of the sequencing run, leading to uneven coverage of the reads over the sampled DNA (Aird et al., 2011).

Both the assembly of genomes and the identification of structural variation are hampered by short read lengths in regions that contain highly repetitive sequence such as simple sequence repeats (SSRs), tandem duplications, and mobile element insertions (MEIs). There are two strategies to help alleviate these difficulties. The first is to use short reads in pairs that are separated by a gap of known length, resulting in a longer effective length for the reads. Such pairs can be constructed by forming circular DNA, cleaving it near the discontinuity on both sides, and sequencing the resultant fragment. This will form a *mate-pair* of short reads separated by the length of the original circle. This technique can separate the short reads by as much as 5-20Kbp using biochemical methods, or up to 150Kbp or more using bacteria to clone the fragment. Mate-pairs were used extensively in the human reference genome projects using Sanger sequencing, and can be used with any of the NGS technologies described above. In addition, Illumina NGS sequencing inherently sequences from both sides of a 400-500bp fragment, creating *paired-end* reads.

The second strategy is to use sequencing technology that is capable of reading much longer sequence fragments. Illumina has recently developed a technology called TruSeq that creates synthetic long reads using the same sequencing machines described above (McCoy et al., 2014). To achieve this, the DNA is first sheared into ~10Kbp fragments, which are then places in wells such that no well contains more than ~200 fragments. These are then amplified by PCR and fragmented to 500bp size. Finally, an independent barcode of a short unique DNA sequence is added to each well. After normal sequencing, the barcodes are then used to reconstruct the 10kb fragments via *de novo* assembly of the resultant paired-end reads. This technique may still suffer from GC bias and is susceptible to fragment assembly errors if the fragments in any one well are overly repetitive. However, the consensus sequence built from the *de novo* assembly can have error rates are low as 0.02%. This technology has been used to more accurately characterize repeat sequences in the drosophila genome, especially MEIs (McCoy et al., 2014).

The long-term best solution to these issues is to sequence single long DNA fragments without the need for PCR amplification or other biased processing. One technology for

doing this is to read single bases while they pass from one chamber to another through a nanopore (Feng et al., 2015). While many such pore configurations have been suggested, the only one is use today was developed by Oxford Nanopore (Clarke et al., 2009). In this method, a single-stranded DNA molecule is digested by an exonuclease near the entrance to the pore in such a way that the released nucleotides exit into the second chamber in the order is which they are released. Each nucleotide is recognized by the current across the pore when it flows past. In theory this technique can produce long reads with low error rates, but currently available technology can produce 5-10Kbp reads with significant error rates. This technology was recently used to *de novo* assemble an E. Coli genome (Quick et al., 2014, 2015). Pacific Biosciences has developed a system in which the DNA polymerase is attached at the bottom of a well, while a DNA fragment is replicated with fluorescing dNTPs. All the chemicals are in the bath at once, and no cycling is done. The optics and related computer processing must keep pace with the rate of polymerization. The current error rate of this process is ~18% mostly comprised of small indels. However, the system is capable of reads up to 10s of thousands of base pairs. To reduce error rates at the expense of read length, the DNA fragment can be formed into a circle of up to 2kb in length, and each base is sampled multiple times, reducing the per base error rate to <3%. This technology has been used in conjunction with Illumina short reads and special bioinformatic processing to further reduce the error rate to <0.1%. This combined sequencing and analysis method has shown promise in *de novo* assembly of several genomes, including the identification of the length and structure of repeat regions (Roberts et al., 2013; Shin et al., 2013).

1.2.3 Three Strategies for Somatic Mutation Discovery

All of the above sequencing technologies as well as aCGH and SNP-CHIP require the input of more DNA than can be directly acquired from a single cell. Therefore, to detect somatic mutations, one must use one of three techniques. First, one can use whole genome amplification (WGA) techniques to increase the amount of DNA from a few or even one cell. Second, one can identify somatic mutations in cells that have clonally expanded to the point where they comprise a significant percentage of a bulk tissue sample, and are therefore detectable in the sample albeit at low variant allele frequencies. Third, one can compare bulk test and control samples from the same individual or members of a family cohort in order to eliminate germline alleles that vary from the reference genome, but are not somatic mutations. We now discuss each of these technique in turn using cancer studies as instructive examples.

There are three WGA techniques in use today. They all use PCR amplification and largely differ in the primers and polymerases that are used. Degenerative Oligonucleotide Primed PCR (DOP-PCR) uses a single primer structure (Telenius et al., 1992). A few rounds of PCR are done at relatively low temperature to allow the primer to saturate all its binding sites. Then ~25 rounds of PCR are performed at a higher temperature to further amplify only those regions created in the earlier rounds, reducing noise in genome coverage. However, the resultant fragment size of ~100-1000bp is relatively small. The second technique is called multiple displacement amplification (MDA) (Dean et al., 2002). The unique features of MDA stem from the use of Φ29bacteriaphage polymerase that has a very low base replacement error rate, and is highly processive. It can displace the tail of double stranded DNA that it encounters while replicating, leading to branching replication structures. This eliminates the need for thermo-cycling, and can produce fragments of >10Kbp. However it provides very uneven genomic coverage, resulting in allelic dropout rates of up to 60% in low coverage areas. A modification to MDA called MIDAS uses a restricted amount of MDA amplification in 12-nl micro-wells, followed by linearization of the product with POL I.

These two techniques reduce the amplification of artifacts, thereby increasing the signal to noise ratio (Gole et al., 2013). The third technique is called multiple annealing looping-based amplification cycles (MALBAC) (Zong et al., 2012). The key feature that distinguishes this technique is the structure of the primers that form ssDNA loops of ~1Kbp in length after each replication cycle. This means that amplification is linear instead of exponential, leading to more even coverage over the genome with very low allelic dropout rates.

Using WGA techniques, CNV detection in single cells has been achieved with some success using aCGH or SNP-CHIPs. When used with *single-cell sequencing* (SCS), the uneven coverage can be overcome with binning techniques that identify relatively large CNVs. However, single-cell SNP calling is quite challenging due to allelic dropout and other artifacts. Therefore, strict calling criteria is required to avoid false positives, which in turn leads to higher false negative rates (Ning et al., 2014).

Cancer studies have utilized WGA in conjunction with SCS. A breast cancer study used DOC-PCR WGA and SCS to identify CNV in ~54Kbp bins (Navin et al., 2011). They found four subclones of uniform composition in one breast carcinoma, and a single aneuploid clone in a second breast carcinoma and its metastasis in the liver. Two studies used MDA WGA, whole-exome SCS, and SNP calling to identify subclones and potential causative variants. The first studied a case of essential thrombocythemia (Hou et al., 2012). After quality control, 58/90 cells exhibited an average ADO of 43%. To reduce false positives, they called somatic mutations that appeared in ≥5 cells, leading to a 90% validation rate by Sanger sequencing. They also identified four genes containing driver mutations present in most of their cells. The second used the same methodology to study a renal cell carcinoma (Xu et al., 2012). They found that 20 cancer cells contained ~4X more somatic mutations than 4 normal cells, and that none of the cancer cells shared enough mutations to be considered clonal.

Another cancer study used very deep (10,000X to 15,000X) 454 resequencing of the immunoglobulin heavy chain in blood samples from 20 patients with chronic lymphocytic leukemia in order to explore the complex etiology of cancer due to selection between subclones of cancer cells (Campbell et al., 2008). The ~250bp reads allowed bioinformatic haplotyping of the region to eliminate indels anomalies near homopolymers, and false positive SNPs. They showed that disease progression is often very complex. In the two patients exhibiting the most clonal mosaicism (19 and 7 subclones), the clonal phylogenetic trees placed the dominant clone at neither the trunk nor a leaf.

As NGS sequencing costs have dropped, studies of cancer now routinely perform whole genome or exome sequencing and analysis of paired tumor/normal tissue from the same individual. Since we also use paired test/normal tissue samples in our study of somatic mutation in post-mitotic neurons (**Chapter 5**), here we will limit our discussion to a few cancer studies that elucidate general mutational mechanisms or use novel techniques with relevance to our study. For example, as cancers have high SNV mutation rates, studies of tumors with different cause or primary tissue of origin such as ionizing radiation in melanoma (Pleasance et al., 2010a) or chemical carcinogens in lung cancer (Pleasance et al., 2010b) have shed light on signatures of general mutational processes and/or DNA damage repair mechanisms such as transcription-coupled repair and base-excision repair that also occur in healthy tissue or contribute to other diseases (Alexandrov et al., 2013; Lawrence et al., 2013). For their study of the prevalence of MEIs in 43 cancer genomes across 5 cancer types, Lee et al. devised a sensitive technique for the detection of MEIs that has now been adopted for many MEI detection studies (Lee et al., 2012a).

Finally, it has long been known that cancers contained many aneuploidies and large scale CNVs, but only recently has it become clear just how complex these events can be (Stephens et al., 2011). Complex genomic rearrangements (CGRs), the most extreme of which are called *chromothripsis*, involve several to dozens of deletions, inversions and translocations of fragments from one or more chromosomes joined together in an apparently random fashion. These appear to be caused by a single catastrophic event resulting in multiple double strand breaks ligated together by NHEJ (Maher and Wilson, 2012). The incidence of CGR varies markedly by tumor type (Malhotra et al., 2013), and is sometimes associated with mutations in p53 (Rausch et al., 2012a). CGRs have also been found in the germline (Chiang et al., 2012), and we found at least one somatic CGR in healthy post-mitotic neurons (**Chapter 5**).

1.2.4 Bioinformatic Techniques for Detecting Somatic Mutations in Sequencing Data

The study of somatic mutations in paired test/normal tissue samples involves separately identifying alleles that differ from the reference genome in each. Variants that appear in both are deemed to be in the germline, while those only in the test sample are putative somatic mutations. Therefore, we now discuss the tools used to identify alleles that differ from the reference genome across all major mutational categories. This usually involves *alignment* of each sequenced DNA fragment to the reference genome, and *variation detection* performed by the pileup or clustering of those alignments by genomic region in order to determine their common characteristics. We will postpone our discussion of the alignment task until after we discuss variation detection methods, which put strong requirements on aligners.

All the information used to detect SNPs and indels is local to a small genomic region around their occurrence. Therefore, techniques for detecting them are easiest to understand and tools for such detection are concomitantly more mature. While there are a large number of such tools available today, including SOAPsnp (Li et al., 2009b), VarScan (Koboldt et al., 2009), FreeBayes (Garrison and Marth, 2012), and GATK (DePristo et al., 2011; McKenna et al., 2010), they essentially all use some type of Bayesian model to calculate the probability of a particular mutation (genotype) at a locus given the input sequence reads and prior probabilities of each genotype. For example, all other things being equal, a heterozygous SNP on an autosome should appear in approximately half of the sequence reads at that loci. Prior probabilities can also be improved by using either a gold standard set of mutations and/or using input data from multiple samples as supported by the GATK UnifiedGenotyper. Tools typically try to reduce errors by reporting a lower probability for a given genotype if the region has low read coverage, the genotype is supported by reads in which the reported sequencer quality scores for the base is low, or the variant is disproportionally represented in reads from one strand over the other strand. In addition, some form of alignment normalization may be done in cases in which there is more than one possible alignment for a given mutation, for example the deletion of a nucleotide from within a homopolymer. Perhaps the best way to handle such ambiguous alignments and strand bias is to calculate longer local haplotype regions; a technique used by FreeBayes and the GATK HaplotypeCaller.

Once the SNPs and indels are called against the reference genome in this fashion, somatic mutations in the test sample are identified by removing the control sample calls, and usually also those in curated databases of known population alleles, as these are unlikely to have arisen *de novo* by chance. SomaticSniper (Larson et al., 2012) and MuTect (Cibulskis et al.,

2013) are tools that works directly on tumor/normal pairs to do this type of somatic mutation analysis. However, these tools cannot identify somatic mutations across multiple test samples. Therefore, to identify putative somatic mutations in post-mitotic neurons (**Chapter 5**), we used the GATK UnifiedGenotyper to make our initial SNP and indel calls, and custom Python scripts to filter out germline calls using criteria similar to those used by (Kong et al., 2012) to identify somatic mutations during gametogenesis using family pedigrees. Finally, to gauge false positive call rates, a subset of putative somatic mutations is usually verified by PCR using primers for the appropriate genomic region, followed by capillary sequencing. The resultant traces can be read visually for a small number of heterozygous calls. However, to reduce errors in this process for a large number of calls, or if the variant allele frequency may be different than 50%, a tool like SNPdetector can be helpful to read capillary sequence traces (Zhang et al., 2005).

Somatic SV events are harder to identify than SNPs and indels for several reasons (Alkan et al., 2011). First, by definition, SVs cannot be identified by looking at a small local region of the genome, and their variation in size and complexity limits the ability of tools to predict where to look for the information needed to reconstruct them. For example, an insertion from one genomic region to another is often detected as two separate breakpoints that then have to be later combined into the proper interpretation of a single mutational event. Second, informative sequencing reads for SV detection often fall into highly repetitive sequences or span an SV breakpoint, and are therefore hard to properly align to the reference genome.

There are two techniques used to discover SV using sequencing data. The first borrows the strategy used with aCGH and SNP-CHIP data, using sequencing read depth to detect CNV. The second is to use the subset of reads that align in an aberrant fashion to the reference genome to form clusters that can signal a SV breakpoint (**Figure 1.2**).

To my knowledge, all CNV detection algorithms use variations on a theme. First the genome is broken up into non-overlapping bins with varying sizes determined by GC content and the amount of unique sequence they contain. Reads are assigned to bins, and the average bin read depth is determined. Then z-scores are calculated for the difference in read depth between each bin and some aggregate average of all bins. Consecutive bins of similar z-scores are then combined to define larger genomic regions with the same copy number. Regions with large absolute z-scores represent CNV events. The Event-Wise Testing EWT method starts with very small bins compared to most approaches (Yoon et al., 2009). CNVnator is probably the most widely used, and utilizes a copy-number invariant method for calculating the aggregate mean (Abyzov et al., 2011). For finding CNV in post-mitotic neurons, we use a variant of this same binning approach that carefully calculates the mean aggregate copy number separately for bins with different GC content, and uses circular binary segmentation to combine bins into integral copy number (Malhotra et al., 2013).

The second approach to SV detection identifies pairs of clusters of aberrantly aligned reads as putative SV *breakpoints*, defined as adjacent locations in the test genome that map to discontinuous locations in the reference genome. SV breakpoints can be detected using discordant paired-end mapping (PEM), in which the two ends of a paired-end pair do not map to the reference genome with the expected distance or strand orientation, or split-read mapping (SRM), in which a portion of a longer read aligns to one region of the reference genome, and another portion aligns elsewhere (**Figure 1.3**). SRM is significantly more precise and less error prone in locating SV breakpoints than PEM, as the latter can only

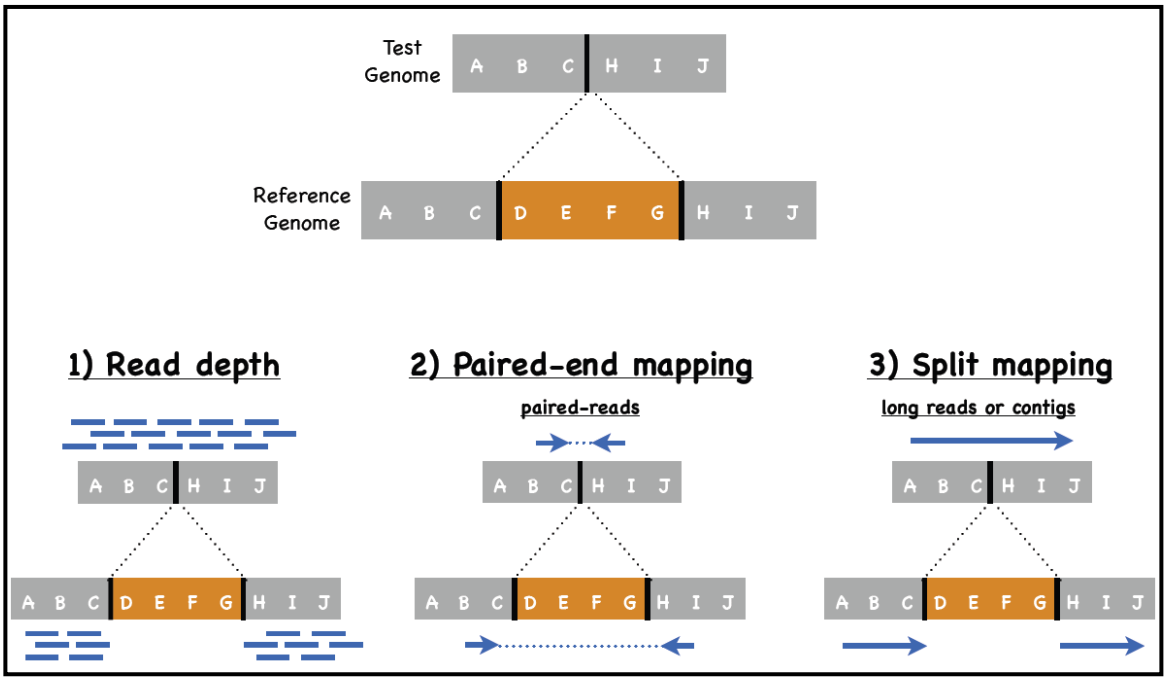

Figure courtesy of Ira Hall

Figure 1.2. Methods of SV detection.
Structural Variation, in this case a deletion, can be detected by 1) differences in read depth, 2) discordant paired-end mappings, and/or 3) split-read mappings.

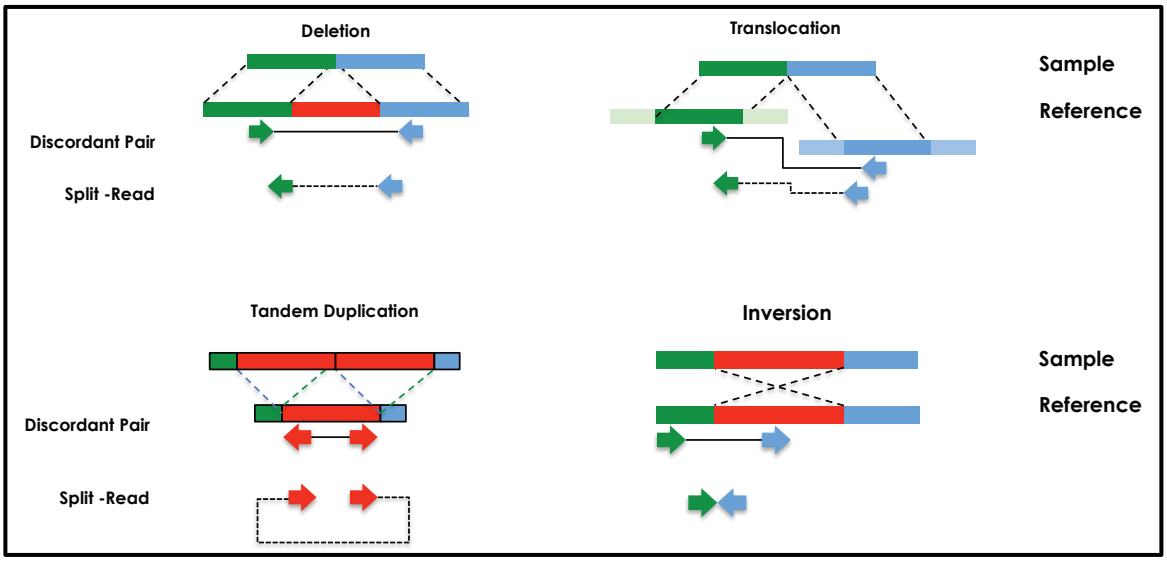

Figure adapted from (Layer et al., 2014)

Figure 1.3. SV breakpoint signatures from paired-end and split-read mappings. Different structural variant breakpoints are detected by different aberrant paired-end and split-read mappings. For a deletion, discordant PEMs have the expected strand orientation (+/-) but are farther apart on the reference than expected, while SRMs are also far apart and point in the same direction. For tandem duplications, PEMs have the wrong strand orientation (-/+) and in SRMs the beginning of the read maps after the end of the read. For inversions, PEMs have the wrong strand orientation, (+/+ or -/-), and SRMs have two alignments that point at each other. Translocations can have a variety of PEM and SRM mapping orientations depending on the location of the two adjoined regions. Note that insertions are not directly detected by these signatures, as they contain two breakpoints, one on each end of the insertion. Therefore, their architecture is constructed post hoc by combining two breakpoints into one event.

probabilistically determine breakpoint location based on the expected insert size between paired-end reads, while SRM can often locate SV breakpoints to single base-pair resolution.

When Illumina paired-end sequencing was first introduced, the ~30-35bp reads at each end were too short to use SRM, and several PEM aligners were created to map SV with this newly available sequencing data. These include PINDEL (Ye et al., 2009), BreakDancer (Chen et al., 2009) and HYDRA (Quinlan et al., 2010). However, each end of an Illumina pair is now ~100bp in length, longer read technologies are becoming more commonplace, and the techniques for *de novo* assembly of short reads into longer contig are more advanced. Therefore, the latest generation of SV detection tools uses a combination of PEM and SRM information to increase both precision and accuracy over using either alone. These include DELLY (Rausch et al., 2012b), GASVPro (Sindi et al., 2012) and LUMPY (Layer et al., 2014). Both DELLY and GASVPro internally perform all the isolation of discordant PEMs, make tentative breakpoint calls using them, then attempt SRM alignment to confirm those breakpoints. However, LUMPY takes the more flexible approach of accepting discordant PEMs and SRMs as input, and using both for breakpoint identification. Therefore, it is not biased against breakpoints that can only be found by SRMs. In addition, it can take advantage of best in breed aligners for doing each type of mapping, and also can be used in variant detection pipelines in more efficient ways.

1.2.5 Improved Tools and Methods to Study Structural Variation

Tools that detect structural variation with improved accuracy are generally useful in their own right, and can aid in the study of cancer and many other diseases, as well as improve our understanding of structural variation within the general population. However, they are particularly important for the study of post-mitotic neurons. A primary goal of our research is test the long proposed hypotheses that the great diversity in neuronal morphology and connectivity patterns are due to some form of recurrent or programmed mutations that commonly or requisitely take place during neurogenesis or post-mitotically while interconnections are formed. As most of the proposed mechanisms for these mutations are structural in nature, accurate SV detection tools are of critical importance to our endeavor.

SV detection algorithms such as LUMPY are more effective when presented with accurate SRM, yet most DNA aligners are not well designed for aligning breakpoint-containing query sequences, and are therefore only suitable for finding SV via PEM or read pile-up for CNV detection (**Figure 1.2**). Prior to 2010, the majority of DNA aligners fell into three categories. Many, for example FASTA (Pearson and Lipman, 1988), BLAST (Altschul et al., 1990) and SSAHA (Ning et al., 2001) looked for the best matching subsequence alignments for the query DNA in the reference library or genome of DNA sequences. Others, for example BLAT (Kent, 2002), were primarily aimed at aligning cDNA from mRNA, looking for a set of alignments for sequential portions in the query DNA consistent with a sequential set of exons along the reference genome with breakpoints at exon/intron boundaries. Such aligners can identify breakpoints for deletions, but not other forms of SV. Finally, with the advent in Illumina sequencing, efforts to produce new aligners focused on paired-end alignments of very short ~30-35bp reads. These aligners, for example SOAP (Li et al., 2008) and Bowtie (Langmead et al., 2009), often obtained increased speed by severely restricting the number of allowed mismatches and/or small indels tolerated in the alignments.

Conceptually alignment requires simple string matching between a relatively short query string and a much longer reference string, however in practice the problem is much harder. First, the human reference genome is long and contains many repeat sequences. Second, the

query sequence often does not exactly match the reference sequence due to both SNP and SV polymorphisms between the test subject and the reference genome, and sequencing errors in the reads. Therefore, algorithms targeted to finding exact string matches are not useful in this context. Dynamic Programming (DP) is a well known technique for finding the optimal match between two strings with inserts, deletes, and replacements. In particular, Smith-Waterman (SW) is an algorithm for finding the optimal "local alignment" of a shorter string against a longer one. However, for strings of length L_1 and L_2, DP approaches require time proportional to $L_1 \times L_2$. A typical aligner task is to align 100Ks to 100Ms of queries. It is not practical to use DP along the entire length of the reference genome for each query. Therefore, aligners usually use some combination of an index and heuristics to find regions of interest in the reference, possibly followed by either partial or full use of DP techniques to find the best alignment of the query against the reference in just those regions. In addition, some aligners take advantage of paired-end data, using the expected distance between the paired reads to help anchor reads that map to multiple regions in the genome using a mate with a unique mapping.

Two types of indexes are used by almost all modern aligners; a hash table or some form of suffix tree (Li and Homer, 2010). A hash table maps sequences of *k* base pairs in length, called a *k-mer*, to each place they occur as a subsequence of the reference genome. A query is then broken into a set of k-mers, the locations where those k-mers appear in the reference found from the hash table, and combined to find genomic loci in which the read may map. Such an approach has the advantage of simplicity, but has drawbacks. For short k-mers, especially ones appearing very often in the reference, the aligner may need to consider many places in the genome that will ultimately not be the best match for the query, and the k-mer combination step can be time consuming. YAHA (Faust and Hall, 2012), Novoalign (Hercus, 2009), MegaBLAST (Altschul et al., 1990), SSAHA2 (Ning et al., 2001), MOSAIK (Lee et al., 2014) and many other aligners use this indexing approach.

A suffix tree stores the starting position of every suffix in the reference genome. To find matches for a read, each suffix in the read is compared to the table to find all the genomic positions in which they occur. Such matches are not restricted to a fixed length; therefore it is easier to identify the best matches for a query more quickly. However, suffix trees also have their shortcomings. First, naïve implementations are very large, so aligners tend to only store portions of the tree using an FM-index on the Burrows-Wheeler transform of the tree, from which the portions of the tree required for a given query can be reconstructed (Ferragina and Manzini, 2000). This can be time consuming. In addition, suffix trees do not handle high error rates in the query very well, and can't be easily used to find all the suboptimal alignments for a query in the cases in which it is important. BWA uses this approach for paired-end alignment (Li and Durbin, 2009) and BWA-SW for split-read alignment on singleton data (Li and Durbin, 2010). Recently, BWA-MEM combined these algorithms to use paired-end information when available, and look for split-read mapping in either paired-end or singleton reads (Li, 2013).

In **Chapter 4** we present YAHA (Faust and Hall, 2012). It is a hash-table based aligner with a novel approach to finding split-read mapping in singleton reads of 100bp-32Kbp that optimizes the value of a biologically relevant objective function. It also uses a flexible scoring algorithm called Affine Gap Scoring that allows it to accommodate the varied error models of different sequencing technologies. As a result, it is a fast, flexible and effective all-purpose aligner that outperforms best-in-class tools for three very different tasks: 1) reporting all mappings per query; 2) reporting the single best mapping; and 3) identifying

split-read mappings that define one or more SV breakpoints within a query. In addition, LUMPY performs better with SRM input from YAHA than any other known aligner, helping to identify SV breakpoints at single base-pair resolution. We use YAHA in our study of post-mitotic neurons not only to find SV events, but also in the validation of SV and MEI events from Sanger sequencing of the region containing the breakpoint.

In addition, we use YAHA to assist in detection of MEIs, which are notoriously hard to detect by general methods. CNV detection algorithms are not sensitive enough to detect these relatively short (≤6Kbp) insertions. In addition, though YAHA and LUMPY are very good at detecting most forms of SV, all general SV detection tools have a hard time detecting MEIs. By definition, the ME is highly repetitive in the genome, and therefore reads from ME DNA do not tend to pile up in one place, but be scattered across many (nearly) identical copies. The result is that aligners typically report one of a number of equally good alignments. Therefore, the clustering algorithms used by SV detection tools cannot find a strong signal for the ME side of either breakpoint. Although YAHA is capable of reporting all the places such a read aligns in the genome, the flood of alignments this generates confound clustering algorithms which now have too much information. Therefore, we use YAHA to sensitively map reads to a separate mobile element library as part of an improved version of an MEI calling pipeline originally devised to study MEIs in cancer (Lee et al., 2012a).

The speed of bioinformatic pipelines to analyze genomic data is important both in the research lab and especially within a clinical setting. Such pipelines can be slow either because individual tools in the pipeline require a long time to run, or the organization of the pipeline as a whole requires some large sets of data to be handled multiple times (Chiang et al., 2014). One task that has historically been slow is the marking of duplicate sequences, which are artifacts of current sequencing technologies that need to be removed from bioinformatic analyses to avoid bias. A second source of unnecessary time consumption is the reprocessing of large alignment files to pull out discordant read-pairs and split-read mappings needed for SV calling tools, or the reads that didn't map well to the genome for use to a more sensitive split-read alignment tools such as YAHA. In chapter 3 we present SAMBLASTER (Faust and Hall, 2014). It is a tool that marks duplicates faster while using less memory than any current duplicate marking program. In addition, it pulls the reads pertinent to SV detection in the same step, thereby removing the need to revisit these large files to find those reads later in the pipeline (Chiang et al., 2014).

Finally, it is very difficult to debug and tune SV detection algorithms without a dataset that includes known structural variants at known genomic locations. SVsim is a flexible tool for generating a wide range of SV events built for precisely this purpose (**Chapter 2**).

YAHA, SAMBLASTER, and SVsim are all open source software projects freely distributed on github for use under the MIT license (https://github.com/GregoryFaust).

1.3 Functional Ramifications of Somatic Mosaicism

Somatic mutations have been observed in many disorders, but their prevalence and mechanism of action in disease states is not yet well understood. The phenotypic changes that result from somatic mutations often depend on several compounding factors, including the genetic background of the individual, perhaps accounting for some of the missing heritability of complex disorders (Manolio et al., 2009), and when and where the mutation occurs during development or aging (Gottlieb et al., 2001). It is now clear that cancer is caused by one or more somatic mutations in a cell lineage that disrupts its normal

association with its surroundings, leading to abnormal cell division and/or migration (Watson et al., 2013). Many other disorders, especially developmental and/or neurological ones, have also been associated with somatic mutations and mosaicism. Yet, many studies have also found somatic mutations in apparently normal tissue in healthy individuals (De, 2011).

1.3.1 Somatic Mutation Rates Vary by Age and Cell Type

Somatic mutations occur surprising often during embryogenesis. A recent study found that while 7/8 or 87.5% of *in vitro* fertilized (IVF) oocytes were chromosomally balanced, 16/23 or 70% of IVF embryos contained blastomeres with segmental imbalances (Vanneste et al., 2009). Further analysis showed a wide variety of abnormalities including chromosome arm imbalance, segmental duplications and LOH. In many cases, a deletion in one blastomere from an embryo was accompanied by a duplication of the same region in another blastomere from the same embryo, indicating mis-segregation events. A similar study showed 48% of 4-cell embryos from *in vitro* fertilized eggs were genetically mosaic, rising to as high as 90% by the blastocyst stage (Bielanska et al., 2002). Another recent study showed high levels of L1 retrotransposon activity during embryogenesis leading to chromosomal damage (Kano et al., 2009). Yet, as we will see, CNV of this severity is far less common in live births. This suggests that severely aneuploid embryos may die before birth, as 50% of spontaneous abortions exhibit chromosomal imbalances (Vanneste et al., 2009). Another explanation is that blastomeres with severe CNV are under selection pressure, and their cell lineage dies out. Selection pressure acting on somatic clones is a common theme in cancer (Campbell et al., 2008), and also acts to reduce the complexity of the immune system with age (Holstege et al., 2014).

Increased mosaicism with age was found in two recent studies by reexamining SNP-CHIP data from preexisting GWAS studies, including peripheral blood samples from 50,222 subjects from the GENEVA consortium (Laurie et al., 2012), and peripheral blood and buccal swaps from 13 separate cancer studies with a total of 37,717 cancer patients and 26,136 control subjects (Jacobs et al., 2012). Both could detect mosaic CNV events occurring in as few as ~5-10% of cells sampled, with size sensitivity of 50Kbp and 2Mbp respectively. The incidence of detected CNV in patients below the age of 50 was 0.23% and 0.5%, and increasing to 1.91% and 3% in patients over the age of 70. Identified mutations including deletions, duplications, and LOH associated with either parental or acquired uniparental disomy. Cancer patients showed ~1/3 higher incidence of CNV, and healthy patients with identified CNVs had an increased odds ratio of acquiring cancer of 1.27 overall, and 1.56 and 1.98 in two smoking related cancers. A third study of three specific loci that include inverted tandem repeats in eight healthy individuals showed noticeable levels of mosaicism for inversions at those sites, suggesting NAHR as the mutational mechanism, and the percentage of effected cells went up with age (Flores et al., 2007).

One way to study somatic mutations is to examine monozygotic (MZ) twins. A recent study of nucleated blood cells by aCGH and/or SNP-CHIP from 19 MZ twin pairs, nine of which showed phenotypic discordance for Parkinson's Disease (PD), and ten of which were apparently healthy, showed 5% of individuals with a mosaic somatic mutation, in close agreement with the much large cohorts discussed above. One twin from the first cohort showed 22Mbp and 85Mbp mosaic deletions in ~20% and 10-15% of blood cells. These appeared in genomic regions unrelated to PD, but known to be associated with Chronic Lymphocytic Leukemia, leading to the eventual diagnosis of the latter. A healthy MZ pair showed a 1.6Mbp mosaic deletion in 70-80% of cells (Bruder et al., 2008). An extension of

this study showed that 3.6% of 78 MZ twins and 108 single-born individuals had mosaic mega-base CNVs, all occurring in individuals over the age of 55. In addition, for those individuals sampled at more than one age, the percentage of cells showing the mutation went up and down with time, showing clonal expansion and contraction of the progenitor cells. The number of smaller scale (1Kbp) somatic CNVs went up in all cohorts with age (Forsberg et al., 2012).

Studies examining multiple tissues in healthy individuals have found different mutations in different tissues of the same individual. A study of 11-12 tissues from three individuals showed as many as four different CNVs in one individual (Piotrowski et al., 2008). Another study of six subjects found tens of events per subject, often appearing in only one tissue, and enriched for genic regions (O'Huallachain et al., 2012). This mutational diversity across tissue types in an individual has important ramifications for selection of donor tissue for iPSC therapies, suggesting the need for careful prescreening, especially since there is evidence that clonal expansion in culture of cells harboring CNVs can increase the percentage of cells in which they occur (Abyzov et al., 2012).

Sharing many experimental design features with our study of post-mitotic neurons, Behjati et al. used four stem cell lines from stomach, small intestine, large bowel, and prostate known to grow in culture without reprogramming (Behjati et al., 2014). Individual cells were used to seed 25 successful cultures, called organoids, from two mice each containing enough DNA for deep Illumina NGS sequencing. These were compared to tail-snip tissue acting as germline samples. They used phylogenetic analysis to determine the timing of SNV mutations during development, found the number of SNVs per originating cell to vary between 179-1190, and by tissue type from 274-916 (274 stomach, 289 prostate, 727 colon, 916 small intestine). Finally the SNV base conversion profiles differed per cell type, possibly indicating cell-type specific mutational processes and/or rate of repair.

Several studies have estimated the SNV mutation rate per cell division. Behjati et al. estimate the rate for small bowel stem cells to be ~1.1 based on the average number of SNVs per organoid, the age of the animal from which the founding cells were harvested, and an estimate of 21.5h per cell division. They also estimate ~1.5 mutations per cell division during embryogenesis based on 35 observed mutations during the first 23 cell divisions (Behjati et al., 2014). Welch et al. estimate 0.13 exonic mutations per year in hematopoietic stem cells (HSCs), which corresponds to ~5 per cell division assuming the exome is 2% of the full genome, and that HSCs divide ~1.3 times a year (Welch et al., 2012). Holstege et al. found 450 SNVs in HSCs of a 115-year-old woman, leading to an estimate of ~3 mutations per cell division again assuming HSCs divide ~1.3 times per year (Holstege et al., 2014). In a review article, Lynch reports three studies of retinoblastoma, three studies of the APC gene in intestinal epithelial cells, two studies of specific loci in cell cultures, and three studies of fibroblast cultures that all yield similar per base mutation rates of 0.99E-9, 0.27E-9, 1.47E-9 and 0.34E-9 per cell division respectively. These average to 0.77E-9 per base per cell division, or ~2.3 SNVs per genome per cell division (Lynch, 2010). Overall, these studies estimate between ~1.1 to ~5 new SNVs per cell division, and are in close agreement to our own estimate of ~6 new SNVs per cell division in neurons (**Chapter 5**).

However, the observed rate of inherited germline mutations is noticeably lower. Kong et al. found germline SNPs inherited from the father to be dependent on the age of the father, and to accrete at the rate of ~2 new SNVs per year or ~60 new germline SNPs per generation (Kong et al., 2012). After correcting for their SNV detection sensitivity, and assuming an

average father's age of 30 years, this study yields a per base mutation rate of 1.2E-8 per generation. Many other studies have produced similar results of 1.1E-8 to 3E-8 per base mutation rate per generation (Conrad et al., 2011). These estimates equate to ~0.2 mutations per cell cycle in the maturation of sperm. The reason for the discrepancy between this mutation rate and that for somatic cells is unclear. However, the most likely answer is that there is significantly higher selection pressure against gametes than somatic cells, and that these numbers only account for live births.

1.3.2 Somatic Mutation in Non-neurological Disorders

It is increasingly clear that somatic mosaicism is more prevalent in non-cancerous disorders than previously thought (Erickson, 2010; Freed et al., 2014; Youssoufian and Pyeritz, 2002). For example, surprisingly, a recent study of fifteen patients that suffered atrial fibrillation showed that three had a functional gene mutation in their cardiac myocytes that was not found in their blood (Gollob et al., 2006). We briefly examine some additional examples relevant to an understanding of the mechanisms at work. A given case of a genetic disease is called *sporadic* when the parents and siblings of the affected individual do not have the disease. Genetic variants that are highly penetrant for severe disease states often arise in this manner, as carriers rarely live to reproductive age. Therefore the causative mutations are under severe selection pressure. In such cases, the genetic cause either occurred as a somatic mutation in the germline of a parent, or early during embryogenesis of the afflicted child. Many of the diseases discussed below are often sporadic, and the *de novo* mutations show high penetrance. In others, somatic mosaicism in the affected individual can increase or reduce the severity of a phenotype of the genetic disorder. Also, many of the mutations involved are structural variations, highlighting the need for the best possible tools to detect such variants to further investigate the full spectrum of mutations affecting disease states.

Several genetic disorders alter the rate of somatic mutation in the body, usually resulting in premature aging and susceptibility to mutagens (Freed et al., 2014). Progeria is the result of mutant lamin protein that disturbs normal mitosis and causes premature aging. Cockayne syndrome is caused by a defect in the transcription-coupled repair (TCR) mechanism, allowing SNVs to collect in genes. Werner syndrome and ataxia-telangiectasia are the result of defects in the repair mechanism for DNA double strand breaks (DSBs). Xeroderma Pigmentosum is caused by a defect in the base-excision repair mechanism, causing patients to be particularly susceptible to skin damage from UV radiation. Bloom syndrome results in an abnormally large number of mitotic recombinations between sister chromosomes. Ironically, this can often lead to the mosaic repair of the mutant gene by a fortuitous crossover event.

Several X-linked genetic disorders show patterns of somatic mosaicism (Youssoufian and Pyeritz, 2002). Incontinentia Pigmenti and Rett syndrome are prenatally fatal to most 46,XY males, but some survive if they are mosaic for 47,XXY. Turner syndrome due to 45,X is also 98% fatal unless the fetus is mosaic for 46,XX or 46,XY. Similarly, Alport syndrome is usually very severe in males, but some males show a much milder phonotype presumably associated with mosaicism. Trisomy 8 is not X-linked, but is also almost always fatal *in utero* unless it appears only in a subset of cells.

Some genetic diseases show common mosaic reversion to the wild-type phenotype in some somatic cells (Erickson, 2010). Included are Wiskott-aldrich syndrome, epidermolysis bullosa, Franconi anaemia, and tyrosinaemia. The most common reversion mechanism is mitotic recombination, but complex compensatory mutations have been observed. In one

study, a patient with a 1bp frame-shift insertion was compensated by a 5bp insertion, and in another, a 1bp deletion was compensated by two additional 1bp deletions, in both cases the reading frame was restored (Waisfisz et al., 1999).

Several other diseases are often sporadic or show mosaicism. Of 113 patients with McCune-Albright syndrome tested, 90% of affected tissue showed the causative mutation, but was present in the blood of only the 46% of patients with the most severe phenotype (Erickson, 2010). Sturge-Weber syndrome may be similar, although the incidence of somatic mutations in affected tissue appears to be variable, with different studies estimating the appearance of the causative somatic mutation in 1-47% of cells (Freed et al., 2014). Both Ollier disease and Maffucci syndrome may similarly be sporadic mosaic disorders (Freed et al., 2014). Neurofibromatosis has been reported to be mosaic for specific deletions of 1.2Mbp and 1.4Mbp in 25-40% of cases, possibly formed by NAHR between tandem repeats (Erickson, 2010).

1.3.3 Somatic Mutations in Neurological Disorders

As a major goal of this research is to study the extent of somatic mosaicism in healthy neurons, it is instructive to review the prevalence of somatic mosaicism in neural diseases. A sporadic case of Creutzfeldt-Jakob's disease was shown to be mosaic in the brain of the affected individual, appearing in ~97% of both brain and blood samples, indicating a somatic mutation occurring early in embryogenesis (Alzualde et al., 2010). A sporadic case of early-onset Alzheimer's disease (AD) was similarly shown to be mosaic in 14% of cerebral cortex cells, but only 8% of lymphocytes (Beck et al., 2004). The variability of the onset of AD has led many researchers to suspect a somatic contribution to many cases (Freed et al., 2014). Three cases of sporadic hemimegalenchephaly were shown by SNP-CHIP analysis to be caused by somatic mutations involving AKT3, a know driver of the disorder (Poduri et al., 2012). In two of the cases, a mosaic trisomy of the region containing AKT3 was present with a copy number of ~2.7X. As this is a non-integral value, mosaicism is indicated. In addition, one case showed no trisomy in blood cells. The third showed a point mutation in AKT3 in ~35% of brain cells, but not in leukocytes. This subject was also shown by single-cell sequencing to have the mutation in 39% of neurons, and 27% of non-neuron brain cells (Evrony et al., 2012).

CNVs have been associated with a wide range of other neurological disorders including autism spectrum disorder (ASD), bipolar disorder, attention-deficit-hyperactivity disorder, obsessive-compulsive disorder, major depressive disorder, anxiety and panic disorder. The same genomic regions, for example a ~600Kbp deletion at 16p11.2, or genes, for example neurexin-1 (NRXN1), have been associated with many of these disorders, indicating the variability of phenotypic expression and complex nature of these diseases (Sebat et al., 2009; Weiss et al., 2008). Many of the cases involving rare CNVs are sporadic, and some show mosaicism. These cases act as counter-examples to the "common disease-common allele" model previously proposed for these diseases.

It is a common finding in these studies that the overall mutational burden is similar in cases vs. controls, but the mutations in cases have more severe predicted functional consequences. A recent study of CNVs in 179 patients with Tourette syndrome, including 2 sporadic cases, stratified CNVs into 4 size categories (Nag et al., 2013). They found that only the largest category (>500Mbp) was enriched in cases vs. controls, including 2 cases that include the NRXN1 gene. A study of CNVs in 242 patients with schizophrenia showed that cases were 3-4X more likely to harbor CNVs that affected genes than controls. Another such

study of 359 cases showed that *de novo* CNVs are 8X more prevalent in sporadic cases than in controls (Xu et al., 2008). A study of ASD patients detected a ~593Kbp deletion at location 16p11.2 that occurred *de novo* in five cases, but was inherited in none (Weiss et al., 2008). They also found duplication of the region in six cases in which it was inherited, and one in which it was *de novo*. This region is flanked by a 147Kbp segmental duplication, leading to the conclusion that the CNV was caused by NAHR (**Figure 1.1**). In a second CNV study of 118 sporadic cases of ASD, they found 14 *de novo* SNVs in cases, but only 2 in controls (Sebat et al., 2007).

Four additional ASD studies used WGS (1) or exome sequencing (3) to more fully characterize the causes and heritability of mutations (Iossifov et al., 2012; Michaelson et al., 2012; Neale et al., 2012; O'Roak et al., 2012). All four studies report higher point mutation rates for both control and affected children with older parents, especially older fathers, consistent with the findings of (Kong et al., 2012). In addition, they found that while synonymous and missense *de novo* mutations were fairly evenly distributed between cases and controls, *de novo* nonsense, frame shift, and splice site mutations occurred disproportionately more often in cases. In addition, the genes affected by these disruptive mutations are enriched for *gene ontology* (GO) terms associated with neuronal function. Two studies estimate the total number of ASD related genes to be several hundred (Iossifov et al., 2012; Neale et al., 2012). One study found enrichment for genes associated with Fragile-X disorder (Iossifov et al., 2012). Another study included affected MZ twins, and found that of 29 *de novo* mutations in genes, 7 additional hits in 5 of the genes were found by other studies, while none were found in controls from any of the reported studies. From this they infer that *de novo* mutations found in sporadic cases with both MZ twins affected are highly likely to be causative (Michaelson et al., 2012).

There may be a sensitive balance between mobile element (ME) activity in healthy neurons, and an overabundance of ME insertions or transcripts leading to pathology. As we will discuss in the next section, mobile element insertion (MEI) has been observed in healthy brains, and has been suggested as one mechanism that increases neuronal diversity (Muotri and Gage, 2006). Yet increased MEIs have also been associated with several neurological and neurodegenerative disorders. Increased L1 insertions have been measured by qPCR in the prefrontal cortex of patients suffering from schizophrenia, bipolar disorder, and major depression (Bundo et al., 2014). WGS of schizophrenics vs. controls did not show increased L1 insertions in cases, suggesting that the insertions were somatic mosaics, and therefore not detectable in bulk tissue samples. However, the WGS data did show that L1 insertions in cases were significantly enriched in genes associated with neuronal GO terms. Also, patients suffering from Rett syndrome that had loss of function mutations in MeCP2 had a 2X increased susceptibility of L1 insertions (Muotri et al., 2010). MeCP2 normally suppresses L1 transcription via methylation of the L1 promoter region, a process that is suppressed by the dysfunctional MeCP2 mutant. Similarly, patients with ataxia telangiectasia with a mutated ATM gene had increased levels of L1 insertions in their hippocampus relative to controls. In this case, the posited mechanism of action is the complication of the repair of the L1 transposition, perhaps leading to tandem insertions (Coufal et al., 2011).

An increase in ME transcripts has also been noted in several disorders. A study of ALS patients showed increased mRNA transcripts of the HERV-K LTR mobile element in the prefrontal, sensory, occipital, and especially the motor cortex, but not in controls including Parkinson disease patients. The increased transcription was only in neurons, but the insertion of the transcripts into neuronal DNA was not measured (Douville et al., 2011). In a

similar way, HERV transcripts were found to be higher in cell-free spinal fluid of sporadic patients with Creutzfeldt-Jakob disorder vs. unaffected controls and patients with other neurological disorders (Jeong et al., 2010). Geographic atrophy is an advanced form of age-related macular degeneration in which the DICER1 complex is inhibited. In turn, this results in an overabundance of Alu transcripts flooding retinal pigmented-epithelial cells, causing cell death perhaps by triggering apoptosis (Kaneko et al., 2011). Similarly, TDP-43 binds to ME transcripts, is down regulated in certain neurodegenerative diseases including ALS and frontotemporal lobar degeneration, leading to an over accumulation of ME transcripts, and neuron death (Li et al., 2012). There is evidence for a similar mechanism in Fragile X-associated tremor/ataxia syndrome (Tan et al., 2012).

1.4 Somatic Mosaicism in Healthy Neurons

The diversity of morphology and apparently stochastic interconnectivity of the ~100 billion neurons in a human brain is quite staggering, exhibiting a level of complexity not seen in any other organ of the body. This has led to wide speculation about possible mechanisms that drive this diversity.

The discovery of V(D)J recombination and hypermutation as the source of diversity in T and B cells in the adaptive immune system prompted the hypothesis that a similar mechanism may also be involved in creating somatic diversity in neurons (Chun and Schatz, 1999). This idea gained support when it was discovered that RAG1, one of two necessary proteins for V(D)J recombination, is expressed in mouse brains precisely where new neurons are being formed (Chun et al., 1991). Several years later, the absence of either DNA ligase IV or XRCC4 was found to disrupt V(D)J recombination, retard development of B and T cells, increase sensitivity of fibroblasts to ionizing radiation, and most intriguingly, is embryonic lethal in mice starting around E14 specifically due to widespread apoptosis of newly differentiated neurons (Frank et al., 1998; Gao et al., 1998). These and other facts support the hypothesis that XRCC4 and DNA ligase IV are critical to NHEJ of DSBs in DNA, leading some to speculate that DSBs are associated with a necessary element of neurogenesis and diversity; for example to repair DSBs which occur during V(D)J-like recombination.

More specifically, it had long been postulated by analogy to the active immune system that the wide diversity of odorant receptors observed in olfactory epithelial neurons may be generated by programmed genomic rearrangement, as each such neuron shows one or at most a few distinct receptors in its cell membrane (Ferreira et al., 2014; Mombaerts, 2004; Young and Trask, 2002). However, it is now known that there are up to ~900 olfactory receptor genes in humans and ~1500 in mice. They are dispersed throughout the genome in clusters of several up to a 100 genes presumably produced by tandem duplication followed by differentiation over time due to random mutations. The expression of which receptor(s) are expressed is apparently determined stochastically in each neuron, followed by epigenetic suppression of the expression of the others. Thus, it appears that olfactory receptors are not generated by programmed mutation.

Cadherin-related receptors are neuronal cell membrane proteins that also show a remarkable diversity. Protocadherin genes appear in three clusters each containing 40-60 genes, and their genomic organization shared many features in common with the T-Cell receptors and immunoglobulin gene clusters (Yagi, 2003). Intriguingly, they localize to the synaptic regions of neural membranes, suggesting a role in the stochastic nature of synaptic connections between neurons. The expression of each protocadherin gene can be separately regulated. In addition, the expression of a small number of different receptors selected from

a much larger group could lead to a nearly unique receptor pattern in each neuron in the brain without the need for programmed mutation. Yet unusually high levels of mutations in protocadherin transcripts have been observed. Therefore, it remains unclear whether recombination or some other form of programmed mutation occurs in protocadherins.

Nor is it known if such mechanisms otherwise occurs requisitely or frequently during neurogenesis, what mechanisms are involved, or what adaptive advantage they provide. We have seen that recurrent aneuploidy and specific CNVs and other mutations are associated with neurological disease states. Using a variety of bulk tissue and single-cell approaches, many recent studies have investigated the mutational mechanisms that may be involved during normal neurogenesis, including somatic aneuploidy, CNV and MEIs (See (Richardson et al., 2014) for a review). As these studies are directly relevant to our own study of somatic mosaicism in neurons, we will discuss them in some depth. These studies report a wide range of estimates for the prevalence of these events, but no evidence for recurrent somatic mutations in healthy neurons.

We start with the prevalence of mosaic aneuploidy in neurons. A study of human neurons from six individuals ranging in age from 2 to 86 showed using FISH that 4% of NeuN+ cells from the cerebral cortex and hippocampus were aneuploid for chromosome 21, compared to 0.6% for lymphocytes (Rehen et al., 2005). Assuming chromosome 21 is representative would lead to an overall estimate of ~50% aneuploid cells! However, naïve interpretation of FISH results has been shown to overestimate the loss of chromosomes due to the coalescence of signal from two chromosomes that are collocated (Yurov et al., 2007). Using an improved technique called QFISH, the level of aneuploidy in human fetal brain has been estimated at ~30% compared to estimates for chorionic tissue and skin of 24% and 19% respectively, consistent with prior studies of aneuploidy during embryogenesis (Yurov et al., 2007). Using SKY for full karyotyping, and FISH of XY chromosomes, another study determined that ~33% of mice neuroblasts were aneuploid compared to only 3.4% of lymphocytes. Chromosome loss varied from 1.8-8% per chromosome, while gains were <2% for every chromosome. However, culturing neuroblasts in growth medium reduced the level of aneuploidy, consistent with the hypothesis that aneuploid cells have a higher mortality rate or lower fecundity than euploid cells. Also, the measured rate of aneuploidy in adult cortex was 6X lower than in neuroblasts, or about 5% (Rehen et al., 2001). Overall these studies of neuronal aneuploidy show a very high level of aneuploidy in embryonic and fetal brain tissue, with a lower and less certain level of aneuploidy in adult human neurons. In addition, as described below, wide scale aneuploidy has not been confirmed in recent SCS studies of CNVs in post-natal human neurons.

A study from our own lab examined CNV in single NeuN+ neurons from adult frontal cortex and hiPSC-derived neurons (McConnell et al., 2013). Neurons were derived from three separate hiPSC cell lines, DNA was amplified by MDA and copy number was measured via SNP-CHIP. Somatic CNVs were detected in 13/40 analyzed neurons, including seven whole-chromosome gains, four whole-chromosome losses, and 12 sub-chromosomal CNVs, each appearing in a single neuron. This data showed a high aneuploidy rate of ~25%. However, 110 adult frontal cortex neurons were also obtained post-mortem from three individuals in their early to mid twenties, their DNA amplified via GenomePlex, an amplification strategy similar to DOP-PCR, followed by low coverage Illumina sequencing. One or more unique somatic CNVs were identified in 45/110 or 41% of neurons. Of these, three covered large areas of a chromosome, and were deemed potential aneuploid events, including one duplication and two deletions. The other CNVs ranged in size between 2.9-75Mbp, with 2X

as many deletions as duplications. After adjusting for an estimated 17% false-negative rate, the estimated rate of aneuploidy observed in the SCS data was ~3-4%, far lower than the SNP-CHIP analysis of hiPSC-derived neurons. This may be due to false positives caused by MDA and SNP-CHIP biases, and/or higher prevalence of aneuploidy in pluripotent cells (Peterson et al., 2011) that is later selectively reduced *in vivo* via apoptosis during or shortly after differentiation.

A second study also used a combination of MDA and GenomePlex for WGA of single NeuN+ neurons, followed by low coverage Illumina sequencing (Cai et al., 2014). Of 82 cortical neurons from healthy individuals, 78/82 or 95% were euploid, and none showed a full chromosome gain or loss, in spite of 100% sensitivity for detection of aneuploidy in trisomy-18 control samples. However, 19 euploid neurons tested for CNV showed that 13/19 or 69% had at least one large CNV, and that these 13 had an average of 3.4 CNVs, ranging in size from 1.7-17Mbp. Two of these CNVs were shared between two neurons apiece, indicating the mutation occurred in the cell lineage before entering the post-mitotic state. Their estimated sensitivity for ~2Mbp CNV was ~62%, so their actual estimates for CNV prevalence may be too low.

A third study sequenced six adult neurons, two from a healthy patient and four from a patient with trisomy-21 (Gole et al., 2013). They used MIDAS WGA, followed by Nextera library prep, and low coverage Illumina sequencing. This resulted in much less noise in copy number estimates in ~60Kbp bins than traditional MDA or even MALBAC. The extra copy of chromosome 21 was easily detected in all 4 trisomy-21 samples. Yet they detected no other aneuploidy, and found ~2 >2Mbp CNVs in each neuron, as well as 9-18 <1Mbp CNVs for which they were unable to calculate a false positive rate.

Taken together, the SCS-based estimates for post-natal aneuploidy are ~3-5% vs. ~25-35% from the previous studies. Some of this difference can be attributed to the inclusion of pre-natal tissue in the higher estimates, and the outlier estimate reported by (Rehen et al., 2005). However, it is also possible that SKY, FISH, and array hybridization tend towards false positives created by failure to detect a chromosome, and/or the SCS approaches have higher false negative rates. Yet the latter studies have all bioinformatically estimated FNRs of 17-40%, leading to an upper estimate of <10% aneuploidy rate. In our study, we find 6/6 neurons are euploid. This is consistent with the SCS studies, as using 10% as the aneuploidy rate, one would expect to see 6/6 euploid cells ~50% of the time by chance. However, our detected CNV rates of ~1-2 per neuron are below the estimated rates for all of the above experiments, and the CNVs we detected are smaller in size, perhaps in part due to the finer resolution of our detection pipeline.

MEIs have also been suggested as a mechanism contributing to somatic mosaicism in neurons. Muotri et al. found that rat hippocampus neural stem cells (HNSC) that were in the act of committing to neuronal lineage had 1.5-2X higher L1 transcripts than the remainder of the HNSC population as measured by qPCR (Muotri et al., 2005). This L1 expression occurred in a very specific time course during the transition to neural progenitor cells (NPCs), and this effect was highly correlated to reduced levels of Sox2 expression. A sharp decline in Sox2 expression in day one, led to a marked increase in L1 expression, which then dropped back to near baseline levels within 4 days. Cells from multiple lineages were transfected with an L1 cassette that expresses EGFP only upon ME insertion. After 7 days, EGFP was detected in ~1% of HNSC and NPCs, but not in neurons, astrocytes, fibroblasts or lymphocytes. Transfected NPCs showed continued EGFP expression upon differentiation to

neurons, but not to astrocytes or oligodendrocytes. PCR and subsequent sequencing of the EGFP integration sites indicated that integration often occurred in genes and particularly in neuronally expressed genes. Finally, they formed a knock-in mouse containing the L1-EGFP construct to test MEI events *in vivo* in mouse brains. By FACS, EGFP+ cells in E10.5 embryos and adult brains collocated with a neuronal marker (NeuN), but not with markers for oligodendrocytes or astrocytes, and sequencing of the EGFP+ cells confirmed the integration of the EGFP construct into genomic DNA.

In a follow on study, the same L1-EGFP knock-in mouse was used to study the effects of exercise on neurogenesis and L1 transposition in adult mice (Muotri et al., 2009). Fourteen mice were divided evenly into two groups, one with access to a treadmill, and the other not. The runners showed 3X higher EGFP+ cells in the dentate gyrus of their hippocampus, perhaps indicating that they were forming new neurons from NPCs, opening a window for L1 insertions. This was corroborated by a similar study from the same lab with mice injected with retrovirus associated GFP construct and BrdU that showed neurogenesis in the dentate gyrus of runners.

Coufal et al. extended this study to humans (Coufal et al., 2009). They transfected fetal brain stem cells with the same L1-EGFP construct and found 8-12 L1 MEIs per 100,000 cells, or ~1/10,000 cells, confirming that L1 activity can occur during embryogenesis. The EGFP+ cells could be differentiated into both neuronal and glial lineages. Transfection of human ESC lines with this construct showed similar results to (Muotri et al., 2005); EFGP+ cells showed neuron markers, and could be differentiated into both neuronal and glial lineages. Nineteen MEIs were sequenced and found to occur at known TTTT/A cleavage sites, and 16 were within 100Kbp of a gene, some of which were neuronally expressed. Target site duplications (TSD) were found in 5/8 fully characterized insertions. By qPCR they found that L1s were more prevalent in hippocampal tissue than matched cerebellum, heart or liver control tissues from three adults. The same assay applied to ten brain regions from three other individuals showed statistically more L1 prevalence in the grouped brain samples than in control heart or liver. Finally, in what is perhaps the most controversial finding of this study, they estimated the number of somatic L1 insertions per neuron by comparing qPCR results of DNA samples from ~12 cells of hippocampal and cerebellum neurons to ~12 cells worth of heart or liver DNA spiked with L1 plasmids at different ratios. From this, they estimated that the hippocampal sample had ~1000X more L1s than heart or liver, and therefore ~80 *de novo* L1 insertions per cell. This estimation technique is of uncertain accuracy, and the results are not entirely in keeping with later studies.

Baillie et al. recently measured MEI activity in the brain using a novel experimental technique (Baillie et al., 2011). Fragmented DNA from bulk tissue was hybridized to a custom capture array to enrich ME fragments, which were then deeply sequenced. The enrichment step helped to ensure that the collected DNA was from a ME insertion site, and to reduce the number of PCR rounds during library preparation in order to increase the signal of rare insertions vs. the large number of germline insertions. A sensitive ME detection pipeline was used to cluster paired-end read alignments based on their genomic location, strand orientation, and ME family. Concordant pairs were called as germline variants. Discordant pairs were considered potential somatic insertions if they did not appear in brain or blood control tissue from any individual, or match a previously annotated MEI polymorphism. This pipeline shares many features with the one used by Lee et al. to detect somatic MEIs in cancer, and our pipeline for detection of somatic MEIs in mouse neurons (Lee et al., 2012a).

Using this approach, they tested neurons from the hippocampus and caudate nucleus of three individuals. Of the 25,229 initial calls, only 8.4% of Alu insertions and 1.9% of L1 insertion were deemed germline. Almost all of the MEIs called as somatic were supported by a single read-pair, and a few by two read-pairs. While the low read count compared to germline insertions is evidence of a rare somatic event occurring in one or a few cells in the sample, it also raises issues about potential false positives. Therefore, a subsample of calls was stringently validated using PCR and capillary sequencing. All 35/35 germline calls, 14/14 somatic L1s and 12/15 somatic Alus were confirmed. Valid PCR primers for 3' ends of calls were difficult to identify, therefore they were unable to determine the existence of TSD via capillary sequence. However, 2/3 putative somatic calls with paired-end reads that fortuitously straddled the insertion breakpoint showed TSD. The ratio of somatic L1 insertions in the hippocampal vs. the caudate nucleus for the three individuals was 1.3, 0.5, and 2.2, and they were enriched in genes, and those genes were enriched for neuronal GO terms. This experimental design did now allow a per cell estimate of MEI load.

Several studies have used single-cell sequencing to estimate per neuron MEI loads. Evrony et al. studied L1 insertions in 300 individual neurons, 50 each from the cerebral cortex and caudate nucleus of three neurologically normal human (Evrony et al., 2012). Neurons were obtained by FACS (NeuN+), and DNA amplified by MDA. L1 3' regions were further amplified by PCR using a combination of random and L1 3' specific primers. DNA was deep sequenced and aligned, followed by read-depth pileup. Sensitivity was high in identifying known reference insertions, 81% and 75% with bulk and single-cell samples respectively. Insertions within 20Kbp of known reference insertions and those annotated in other studies were removed from the set of putative somatic mutations. Even setting a threshold that would identify only 50% of L1 insertions from previous studies, the validation rate for putative somatic mutations was very low. Initial putative somatic mutation rate estimates were 1.1 total and 0.6 unique somatic insertion of per neuron. To estimate false positive rates, the top 16 candidate somatic mutations from each sample were selected for validation (96 total). Unfortunately, only 5/81 calls for which primers could be designed passed PCR validation, implying a false positive rate of 94%. This dropped the estimates L1 insertion rate to 0.07 total and 0.04 unique per neuron. For the five fully validated somatic insertions, one was full-length, showed TSD and a poly-A tail, and appeared in two neurons. The other four appeared in a single neuron each. A follow up study using SCS on 16 neurons from the cerebral cortex and a MEI calling pipeline similar to (Baillie et al., 2011) found 12/16 or 75% of neurons with no MEI insertion, and two MEI insertions overall, each occurring in two of the remaining four cells (Evrony et al., 2015).

The most recent SCS study of L1 insertions adds additional uncertainty in estimates of L1 activity in neurons (Upton et al., 2015). After FACS sorting NeuN+ cells, MALBAC WGA was performed followed by L1 specific capture to increase signal to noise ratio similar to (Baillie et al., 2011). The L1 enriched DNA was then sequenced. Putative somatic MEI calls detected with a calling pipeline similar to the one used by (Baillie et al., 2011), (Lee et al., 2012a), and us. Performing this procedure on 92 hippocampal neurons, they estimate an average per neuron L1 insertion rate between 9.9 and 55.8, depending on the validation estimation technique used, with 13.7 stated as their most confident estimate. As with some previous studies, they found that L1s were preferentially inserted in genes associated with neuronal GO terms and in genes expressed in the hippocampus. In addition, using large read-depth bins, they found 5/92 hippocampal neurons contained CNVs of >5Mbp in size. They performed the same analysis on 35 Neun+ cortical neurons, with a best estimate of an average 16.3 MEIs per neuron. This study is unique in estimating a higher rate of MEIs in

cortical neurons than hippocampal neurons. They also had a hard time accurately estimating their true positive rate. This highlights the need for enough DNA left over after MEI detection to do full validation of all calls as we do in our study.

MEIs have also been studied in drosophila. Using techniques such as qRT-PCR to measure transposon expression levels, and transfection with vectors that express GFP after a ME integration, MEI have been shown to increase with age, to adversely affect performance in long term memory tests, and be associated with relaxation of the Piwi siRNA system for eliminating such transcripts (Li et al., 2013). Using qRT-PCR it was discovered that Piwi is uniquely constitutively turned off in alpha-beta neurons of the mushroom body responsible for olfactory memory, leading to a high level of ME transcription from birth. From 3.1X sequencing data, and Monte Carlo simulations, an estimate of ~200 unique *de novo* MEIs appear in alpha-beta neurons per fly and ~100 per neuron, indicating many shared mutations likely arising early in development (Perrat et al., 2013). Such MEI loads did not appear in other neurons, implicating MEI in memory formation. Finally, low expression levels of Hsp90 has been shown to down regulate the Piwi complex, allowing for increased organism-wide MEIs, perhaps as a mechanism to stimulate genetic diversity in a population of flies under environmental stress (Specchia et al., 2010).

Several lines of evidence suggest that there may be a narrow window during the transition from NSC to NPC to neuron during which L1 expression and transposition are activated, and that this activation is instrumental to neurogenesis and neuronal diversity (Richardson et al., 2014). Previously we discussed that both Sox2 and MeCP2 repress L1 transposition through enhanced methylation in L1 promoter regions. Improper MeCP2 down regulation is associated with higher L1 activity and abnormal neurogenesis in patients with Rett syndrome (Muotri et al., 2010). In addition, L1 transcripts have been observed to cause DSBs even when they don't insert (Cordaux and Batzer, 2009), suggesting their involvement in the defects in neurogenesis observed in the context of compromised DSB repair mechanisms (Frank et al., 1998; Gao et al., 1998). Sox2 is down regulated in a narrow window during normal neurogenesis (Muotri et al., 2005). Recently, it has been conjectured that Wnt signaling is responsible for down regulating Sox2 and up regulating NeuroD1, resulting in both neurogenesis and increased L1 transcription (Kuwabara et al., 2009).

This hypothesis makes an accurate determination of the number of somatic MEI in functioning adult neurons a matter of some importance. However, various studies report disparate estimates of ~80 MEIs per hippocampal neuron to ~0.07 per cortical neuron; a range of three orders of magnitude. All these studies have strengths and weaknesses making it difficult to determine where the truth lies. Projections based on qPCR measurements may be more error prone than previously thought, as recent evidence in yeast shows that L1 retrotransposition events may often fail to integrate into the host genome, and instead form episomal DNA circles (Han and Shao, 2012). Single-cell sequencing studies have the potential to make quite accurate predictions, but have difficulties in validating their calls due to insufficient left-over DNA from the sequenced cells, leading to uncertainty in their estimates.

1.5 Conclusion

In **Chapter 5** we use a novel experimental design for the study of somatic mosaicism in post-mitotic neurons that shares many of the advantages of both genome-wide and single-cell techniques. We use somatic cell nuclear transfer (SCNT) of DNA from post-mitotic neurons of known origin in the mouse olfactory bulb to seed clonal cell colonies. We then

harvest sufficient DNA from these cultures to perform deep WGS of each neuron genome in order to detect all categories of somatic mutations including SNVs, indels, SVs, CNV, and MEIs to single base-pair resolution with sensitivity and accuracy unprecedented for single-cell experiments. Somatic mutations are distinguished from germline mutations using bulk tissue control samples from the same mice. In addition, we validate every one of our SV and MEI calls, and a representative sample of our SNV and indel calls.

We detect variants using bioinformatic pipelines that combine the best available third party tools with our own innovations. For example, YAHA (**Chapter 4**) is used to provide all split-read mappings for SV and MEI detection and call validation (Faust and Hall, 2012). We use GATK (DePristo et al., 2011; McKenna et al., 2010) to detect SNVs and indels, followed by filtering to identify somatic mutations using a strategy similar to Kong et al. (Kong et al., 2012). We map structural variation (SV) using a sensitive pipeline which utilizes multiple tools including Novoalign for sensitive paired-end mapping (Hercus, 2009), YAHA for split-read mapping of unmapped and clipped reads and LUMPY to integrate read-pair and split-read analysis (Layer et al., 2014) to detect SV breakpoints at high resolution. We detect CNV with a read-depth analysis pipeline that detects relatively large (>15Kbp) CNVs (Malhotra et al., 2013). Finally, we detect MEIs using a pipeline that is similar to the approach first proposed by Lee et al. (Lee et al., 2012a), improved by the use of YAHA for finding split-read mappings to define the insertion breakpoints to single base pair resolution. For each mutation class, we estimated variant detection sensitivity via comparison to the Mouse Genomes Project (Keane et al., 2011; Yalcin et al., 2012).

We compare our findings of somatic mutations in neurons to the results from previous studies. We find ~1 L1 insertions per neuron, which falls between the extremes described above, no recurrent mutations or aneuploidy, and other than one quite complex CGR, only 1-2 SVs per neuron. Together, this argues strongly against these mechanisms as required in neural development. We also find ~86 SNVs per neuron, corresponding to ~6 per cell division, in close agreement with studies discussed above. In addition, the SNVs show a base conversion distribution similar to iPSCs derived from mouse fibroblasts (Young et al., 2012), and human hematopoietic stem cells (Welch et al., 2012), but different than those from human germline (Kong et al., 2012) or mouse endodermal cell types (Behjati et al., 2014). Similar to some L1 studies described above, we find that neuronal SNVs are more prevalent in genes expressed in post-mitotic neurons, and that the genes they fall within are enriched for GO terms associated with neuronal development and activity. Also, the lack of shared mutations within neurons from the same animal, or clear correlation between the number of mutations found in neurons and the amount of time they have been in a post-mitotic state, is consistent with a burst model of higher mutational rate occurring around the time of neurogenesis.

2 SVsim

SVsim is a flexible structural variation simulation tool written in Python by the author. SVsim has not been published, but is open source software that can be downloaded from https://github.com/GregoryFaust/SVsim.

A structural variant (SV) is an unexpected juxtaposition of two regions of DNA sequence in a test sample of DNA that are not continuous in the reference genome. Therefore, it is harder to detect than single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) that can be found in local sequencing and alignment information. In addition, some types of SV, particularly deletions, are easier to detect than other forms of SV that have more complex structures. It is estimated that roughly 1,000 SVs distinguish the genomes of two healthy humans, and in the human population these variants collectively affect more bases of the genome than SNPs (Conrad et al., 2010; Mills et al., 2011). However, there is no simple consensus sequence for genomes that differ by these non-local structural events. Therefore, reference genomes for humans, mice and other species that are constructed using DNA from multiple individuals are likely to contain errors in these regions. In addition, SV annotation databases are likely more biased, less complete, less well validated, and contain less well-defined variant locations than corresponding SNP and indel databases.

Yet it is difficult to design, develop, debug and evaluate SV detection algorithms without using sample datasets that contain SV events with known architecture and breakpoint locations to act as a truth set from which to estimate the sensitivity and accuracy of such algorithms. The easiest way to break the circular dependency between SV annotation databases and improved SV detection algorithms is to use synthetic datasets during algorithm development. Ergo, we have constructed SVsim, a flexible tool for simulating SV events of all varieties including insertions, deletions, duplications, or inversions of specified lengths using a declarative language.

SVsim was used to generate simulated SV events which aided in the development of YAHA (Faust and Hall, 2012), LUMPY (Layer et al., 2014), and other tools. Each declarative command for an event can specify whether the event will be placed at a specified location in the genome, or whether SVsim should choose a random location. The source for insertion events can also be chosen randomly, or the location specified or the inserted bases can be included within the command itself. This latter capability allows for spiking in DNA from foreign organisms, such as viruses or bacteria, into the test genome, and was used for Alu insertions into the MEI test dataset for YAHA. SVsim can also be used to create random complex genomic rearrangements (CGRs) with many clustered breakpoints while maintaining user-specified minimum and maximum segment sizes, and ratios for deletions and duplications. Uniquely among SV simulators, SVsim supports simulating SV events in both haploid and diploid genomes.

The output from SVsim includes either a full-length mutated genome, or small contigs around the SV breakpoints, as a fasta file. Simulated sequence reads can then be generated from these files using `wgsim` from the samtools toolkit, which allows the specification of various error models and read average read depth (Li et al., 2009a). Also, a bedpe file is output for each of the SV breakpoints created. This is a standard file format for describing SV breakpoints, is output by many SV detection tools, and can easily be generated from split-read mappings. Therefore, one can directly compare the output of the SV tool one is trying to debug/tune with the expected SV breakpoints output by SVsim to calculate both

accuracy and sensitivity of SV detection. This can be done by using, for example, `bedtools pairtopair` from the bedtools suite (Quinlan and Hall, 2010).

SVsim was developed in 2009 and 2010 primarily for internal lab usage and has not been published. However, since then it has been made open source on github and used by other groups. Several other SV simulators are also now available. For example, svsim is another unpublished open source Python SV simulator in which a very simple set of SV events can be specified using a declarative language. However it cannot simulate complex genomic rearrangements. But it does have some additional features not directly related to SV simulation (https://github.com/mfranberg/svsim). RSVsim is another SV simulator written for R (Bartenhagen and Dugas, 2013). Unlike SVsim's model for generating SV events in random or user specified locations, it attempts to use a biologically derived simulation technique that will, for example, simulate deletions or tandem duplications in genomic regions that it selects as likely places for NAHR to occur. While this is an admirable goal, it is unclear how well this can be done in practice without adding bias. Therefore, we specifically decided against using this approach in SVsim in favor of simplicity and user control. Both of these other SV simulators also output fasta files for mutated genomes, and RSVsim can also output bedpe.

SVsim has another very important set of features that to the best of our knowledge are not available in any other generally available SV simulation tool. These features were added and used by the author to perform Monte Carlo simulations, incrementally adding many clustered SV events in a random order at specified genomic locations in such a way that the resultant CGR would have all of the same breakpoint signatures detected in a CGR from a cancer genome. After each such simulation, the resultant count of distinct copy number states was noted. This helped to prove that the CGRs found in certain cancers in the cancer genome atlas had far fewer copy number states than those occurring in the simulations, making it very unlikely that the observed CGRs could have occurred by a sequence of simple events, and were almost certainly the result of a single complex event such as chromothripsis (Malhotra et al., 2013).

3 SAMBLASTER

This chapter contains a slightly edited version of material published in 2014 (Faust and Hall, 2014) plus the author's online supplemental material (**Section 3.5** and **Table 3.2**) that was not included in the original manuscript due to size limitations.

SAMBLASTER is written in C++ and is open source software that can be downloaded from https://github.com/GregoryFaust/samblaster.

The ongoing rapid cost reduction of Illumina paired-end sequencing has resulted in the increasingly common use of this technology for a wide range of genomic studies, and the volume of such data is growing exponentially. Generating high quality variant calls from raw sequence data requires numerous data processing steps using multiple tools in complex pipelines. Typically, the first step in this analysis is the alignment of the sequence data to a reference genome, followed by the removal of duplicate read-pairs that arise as artifacts either during PCR amplification or sequencing. This is an important pipeline step, as failure to remove duplicate measurements can result in biased downstream analyses.

3.1 Common usage scenario: piped SAM input

Existing popular duplicate marking programs such as PICARD *MarkDuplicates* (http://picard.sourceforge.net/) and SAMBAMBA *markdup* (http://github.com/sambamba) require position-sorted SAM or BAM (Li et al., 2009a) as input, and perform two passes over the input data, thereby requiring that their input file be stored on disk. Instead, SAMBLASTER marks duplicates in a single pass over a SAM file in which all alignments for the same read-id are grouped together. This allows the SAM output of paired-end alignment tools such as NOVOALIGN (Hercus, 2009) or BWA-MEM (Li, 2013) to be piped directly into SAMBLASTER, which marks duplicates and outputs read-id grouped SAM, which in turn is piped to SAMTOOLS or SAMBAMBA for sorting and/or compression, without the need to store any intermediate files. This saves one compression-write-read step in the common case in which a duplicate marked position-sorted file is needed later in the pipeline, and two such cycles if a duplicate marked read-id grouped file is also needed (**Figure 3.1**). The elimination of each such cycle is a significant cost savings of both disk space and runtime. For example, using ~50X-coverage whole genome sequence data for NA12878 from the Illumina Platinum Genomes (ENA Accession: ERP001960), each compressed BAM file consumes over 100 GB of space and requires 7+ hours to compress with SAMTOOLS, and 8.5 hours of CPU time in 1.5 hours elapsed time with SAMBAMBA using 10+ threads on a server class machine. An advantage of the two-pass duplicate marking strategy is that one can retain the "best" read-pair of a set of duplicates, while SAMBLASTER always keeps the first such pair found in its input. However, we will show that the quality of the resultant duplicate marking is nearly identical using several metrics, and in practice we find this has negligible impact on variant detection.

SAMBLASTER will mark as duplicate any secondary alignments associated with a duplicate primary alignment, and thus works particularly well with BWA-MEM output. Currently, neither SAMBAMBA nor PICARD has this functionality.

3.2 Extracting reads for structural variation detection

Structural Variation (SV) is a major source of genome diversity but is difficult to detect relative to other forms of variation. SV detection algorithms typically predict SV breakpoints based on the distribution of discordant paired-end alignments, in which the paired reads

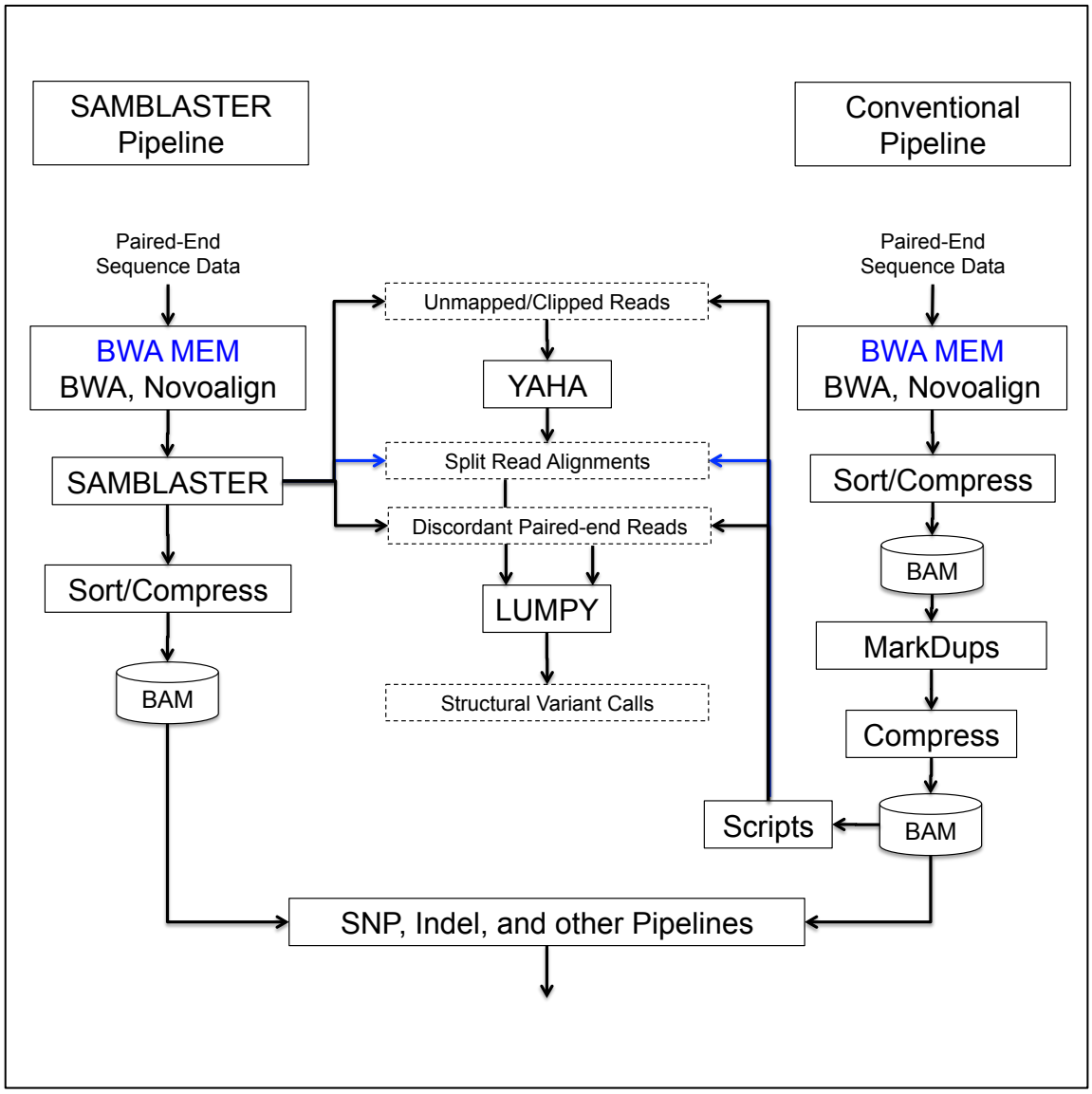

Figure 3.1. SAMBLASTER vs. conventional pipeline for marking duplicate read-pairs. Comparison of SAMBLASTER (left) vs. conventional (right) pipeline for marking duplicates and isolating discordant read-pairs, split-read mappings (BWA-MEM only), and/or unmapped and clipped reads for realignment with a sensitive split-read aligner such as YAHA. Note that use of SAMBLASTER saves a compression cycle, produces fewer intermediate files, and avoid another decompression step to feed scripts used to extract the proper reads for SV detection.

map to either end of an SV breakpoint, and/or split-read alignments where reads align across an SV breakpoint. Many SV detection algorithms require long runtimes due to the overhead associated with searching for and extracting these alignments from large BAM files comprised predominantly of uninformative read-pairs. However, some SV detection algorithms, including HYDRA (Quinlan et al., 2010) and LUMPY (Layer et al., 2014), are able to input files comprised solely of discordant and/or split-read mappings, which are typically >100-fold smaller in size. This presents an opportunity to greatly increase pipeline efficiency by extracting discordant and split-read mappings during a prior pipeline step that already requires reading through the entire dataset. SAMBLASTER is able to extract such reads directly from the SAM output of an aligner such as BWA-MEM that can detect both discordant read-pairs and split read mappings. In addition, when used with other popular paired-end aligner such as BWA-ALN or NOVOALIGN, which do not identify split-read mappings, SAMBLASTER can extract unmapped and clipped reads for realignment with a sensitive split-read alignment tool such as YAHA (Faust and Hall, 2012) for later use to detect SV. By including these capabilities directly in a tool that also marks duplicates, several SV detection pipeline steps can be eliminated (**Figure 3.1**).

3.3 Custom data structures

SAMBLASTER utilizes custom data structures that use significantly less memory than other duplicate marking programs. Two or more read-pairs are considered duplicates when they have the same *signature*, defined as the combination of the sequence, strand, and reference position of both reads in the pair. To most accurately define the reference positions, it parses the CIGAR string to calculate where the 5' end of each read would align to the reference genome under the assumption that the entire read is mapped. This is similar to the strategy used by PICARD. To detect duplicates, it builds a set of such signatures, marking a read-pair as duplicate if its signature has previously appeared in the input.

To avoid storing a structure with all of this information, the signature is broken into pieces. Each unique combination of sequence1, strand1, sequence2 and strand2 maps to its own position in an array in which a set of the associated position pairs is stored as a hash table. The hash tables are optimized to store 64-bit integers; 32 bits for each reference position. SAMBLASTER thus has low memory requirements relative to other tools, ~20 bytes per read pair, which frees it from the need to use temporary intermediate files. See **Figure 3.2** for details. In addition, SAMBLASTER does not allocate/free any per-read memory structures for reading/writing SAM records, thereby increasing I/O throughput.

3.4 Performance evaluation

To evaluate the speed, memory and disk usage of SAMBLASTER as a stand-alone duplicate marking algorithm vs. PICARD *MarkDuplicates* and SAMBAMBA *markdup*, we used the NA12878 dataset aligned via BWA-MEM as our input source. All timings were performed on a server class machine with 128 GB of RAM and two 8-core (16 thread) Intel Xeon E5-2670 processors running at 2.6GHz. To make the comparison of SAMBLASTER to PICARD as similar as possible, we ran both using SAM for both the input and the output format. SAMBAMBA *markdup* does not support SAM format for either input or output. To make the test as comparable as possible, we used uncompressed BAM for both, even though such files are still much smaller than SAM. While SAMBLASTER is single threaded, to show best possible PICARD and SAMBAMBA runtimes, each were allocated 32 threads, and SAMBAMBA single-threaded statistics are also shown. The results of the comparison test are shown in **Table 3.1**. SAMBLASTER outperforms the other duplicate marking programs in terms of CPU seconds, wall time, disk IO, and memory usage.

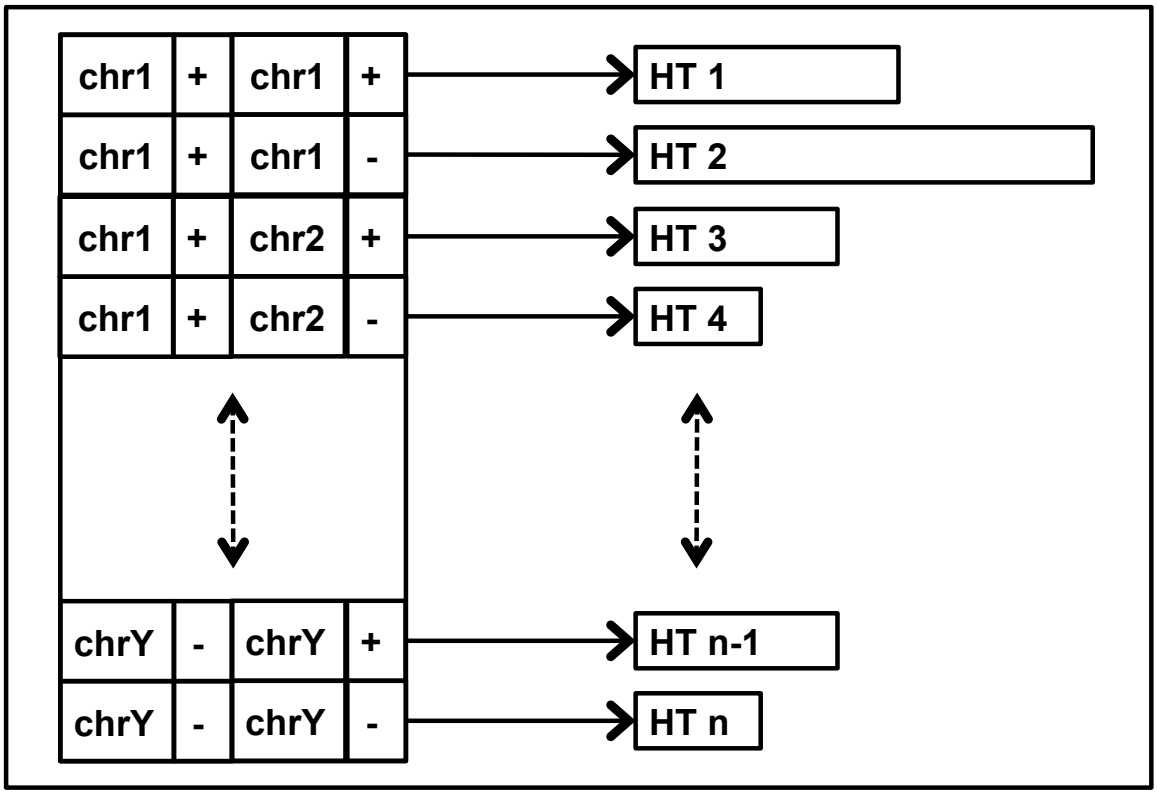

Figure 3.2. Custom data structure in SAMBLASTER.
SAMBLASTER uses a separate set of reference-offset pairs, stored as a hash table, for each combination of sequence1, strand1, sequence2, and strand2. The hash tables are optimized to store 64-bit integers.

Table 3.1. **Comparative performance of SAMBLATER vs. PICARD and SAMBAMBA.** Runtime, memory usage, and disk usage statistics for SAMBLASTER 0.1.14, PICARD *MarkDuplicates* 1.99 and SAMBAMBA *markdup* 0.4.4 as stand-alone duplicate marking tools, and in a common pipeline that produces a duplicate marked position-sorted BAM file as its final output. In the pipeline, SAMBAMBA sort and compression are used. There is also a control pipeline run without duplicate marking, which demonstrates that SAMBLASTER adds little overhead. SAMBAMBA *markdup* times are shown for both an uncompressed and compressed position-sorted intermediate file. These tests were run using local RAID storage with fast read/write times. In a more common scenario using networked disk access, SAMBLASTER's reduced IO results in greater runtime savings vs. the other tools.

	Mark Dups Threads	Extra Disk Space (GB)	Total Disk IO (G ops)	CPU Time (seconds)	Wall Time (minutes)	Mem Usage (GB)
Stand Alone Mark Duplicates Function						
SAMBLASTER	1	-	1.863	2077	43	~15
SAMBAMBA	1	-	2.285	6338	75	~32
SAMBAMBA	32	-	2.285	6603	54	~43
PICARD	32	-	3.056	63160	302	~30
Mark Duplicates – Sort – Compress Pipeline						
No duplicate marking	-	-	1.954	51819	117	~19
SAMBLASTER	1	0	1.987	52767	118	~23
SAMBAMBA compressed	32	108	2.455	86512	154	~43
SAMBAMBA uncompressed	32	391	3.997	61321	163	~43

3.5 Duplicate Marking Quality

Two-pass algorithms have the advantage that they can choose to keep from amongst a set of duplicates, the read pair with the "best" score using some metric that differentiates sequence and/or alignment quality. In contrast, SAMBLASTER's one pass approach can only keep the first pair from a set of duplicate pairs. Given the high quality of Illumina paired-end sequencing, we find this makes little difference in practice. We now analyze the quantity and quality of reads marked as duplicates by SAMBLASTER and PICARD, the generally agreed upon gold standard for quality duplicate marking. In particular, we count the number of reads that each tools marks as duplicate that fall into various categories, note the percent of the total represented by that category, and report the mean alignment score (MAS), mean number of alignment mismatches (MNM) and the mean base sequencing quality scores (MBQS) as measures of alignment and/or sequence quality. The results are summarized in **Table 3.2**. Although there are some notable differences in the number of duplicates in various read categories, PICARD and SAMBLASTER find almost the same total number of duplicates. The resulting non-duplicate reads have almost identical MAS, MNM, and MBQS statistics.

Among the duplicates, by far the largest and the most interesting group are the doubly mapped (DM) pairs in which an alignment is found for both reads in the pair. Because PICARD and SAMBLASTER use the same calculation to locate the 5' end of reads used to identify duplicates, both find the identical number of duplicate DM pairs. In addition, they agree on which of the DM pairs to mark as duplicate ~80% of the time. Assuming that PICARD has a metric that distinguishes all of these DM pairs, we would expect that SAMBLASTER could choose to keep the better scoring pair by chance only 50% the time. Therefore, to explain this 80% concordance, two things must be true. First, at least 60% of the duplicate DM pairs must be considered by Picard to be of the same quality, and therefore it has no way to choose between them. We call these "don't cares". Second, PICARD and SAMBLASTER must pick the same DM pair to mark as duplicate for these "don't cares" a disproportionate percentage of the time. This latter condition is likely caused by the fact that the input file to PICARD was position sorted using Novosort (http://www.novocraft.com/Novosort), which is known to use a stable sort algorithm. Therefore, many of the reads with the same nominal genomic position as reported in the SAM file will be in the same read-id order in the input to PICARD as they were in the SAM file used as input to SAMBLASTER. The high percentage of agreement on "don't care" pairs can then be explained if PICARD chooses to keep the first of these "don't care" cases as the non-duplicate pair, thereby picking the same one as SAMBLASTER. For the remaining 20% of the DM pairs in which PICARD and SAMBLASTER disagree (0.43% of the total reads), the duplicates marked by SAMBLASTER have slightly better MAS, MNM, and MBQS statistics than those marked by PICARD, with the concomitant result that the corresponding non-duplicate pairs kept by SAMBLASTER have worse scores for these pairs.

For the remainder of the read categories, it is clear that PICARD and SAMBLASTER are using different strategies to identify duplicates. Read pairs in which one read is mapped and the other unmapped are called "orphans". SAMBLASTER compares orphans only to other orphans to find duplicates, and always marks both reads in an orphan pair as either duplicate or not duplicate. PICARD marks many more mapped reads in orphans as duplicate than SAMBLASTER, and marks no unmapped reads in orphans as duplicates. One possible explanation for this large number of mapped orphan duplicates is that Picard compares the mapped orphan reads to all mapped reads to determine if it is a duplicate. This could also account for the better MAS, MNM, and MBQS scores for PICARD mapped orphan duplicates

when compared to either the DM pairs, or SAMBLASTER orphan statistics. Finally, SAMBLASTER marks as duplicate any secondary alignments associated with primary duplicates, while PICARD currently does not. By definition, these are the result of a split mapping of the read, are therefore shorter alignments, and have correspondingly much lower MAS and MNM statistics. The lower MAS and MNM scores for the SAMBLASTER duplicate secondary alignments and mapped orphans are partially compensating for the better scores for duplicate mismatched DM pairs, leading to a final total non-duplicate MAS and MNM for SAMBLASTER that is very close to that of PICARD.

Table 3.2. Duplicate marking quality statistics.
Statistics for the number and quality of duplicates and non-duplicate reads for PICARD and SAMBLASTER runs on NA12878. Statistics include the mean alignment score (MAS), mean number of mismatches (MNM), and mean base quality scores (MBQS). The MAS and MNM numbers exclude unaligned reads.

	PICARD					SAMBLASTER				
Read Category	Reads	%Total	MAS	MNM	MBQS	Reads	%Total	MAS	MNM	MBQS
Total Reads	1,578,585,456	100.00	96.95	0.56	36.35	1,578,585,456	100.00	96.95	0.56	36.35
Total Non-duplicates	1,542,595,943	97.72	97.52	0.47	36.35	1,543,282,023	97.76	97.51	0.48	36.34
Total Duplicates	35,989,513	2.28	72.85	4.19	36.44	35,303,433	2.24	72.63	4.23	36.94
Total DM Pairs	34,690,470	2.20	72.60	4.28	36.60	34,690,470	2.20	72.98	4.24	37.05
DM Matching	27,877,300	1.77	72.57	4.31	36.72	27,877,300	1.77	72.57	4.31	36.72
DM Mismatching	6,813,170	0.43	72.74	4.16	36.10	6,813,170	0.43	74.66	3.99	38.42
Orphans, mapped	1,299,043	0.08	79.34	1.96	32.15	190,924	0.01	54.33	3.63	29.70
Orphans, unmapped		0.00				190,924	0.01	NA	NA	24.44
Secondary Alignments		0.00				231,115	0.01	34.76	3.16	36.46

4 YAHA

This chapter contains a slightly edited version of material published in 2012 (Faust and Hall, 2012) plus a modest update of more recent findings (**Section 4.5** and **Figure 4.4**).

YAHA is written in C and C++ and is open source software that can be downloaded from https://github.com/GregoryFaust/yaha.

Structural variation (SV) is a major source of diversity in germline and cancer genomes, but is difficult to map relative to other forms of variation. Since 2008, most sequence-based studies of SV have used paired-end mapping (PEM), which relies upon clustering of discordant paired-end reads that map to either side of an SV breakpoint. Now, with the rapid improvement of short-read assembly algorithms and the development of third generation long-read sequencing technologies, split-read or split-contig mapping (we refer to both as SRM) will soon be the preferred method. SRM is significantly more precise and less error prone than PEM. Yet, current read mappers are not well designed for aligning breakpoint-containing query sequences. Here, we present YAHA, a flexible hash-based aligner that is explicitly designed for optimal SV breakpoint detection from long query sequences (100-32Kbp).

4.1 Introduction

To accurately determine SV breakpoints using SRM, an aligner must do four things well. First, it must accurately determine the best set of alignments that cover the length of the query; the *Optimal Coverage Set* (OCS). This is best accomplished by using an algorithm that provides provably optimal results given some objective function. Our use of a best-path algorithm on a directed acyclic graph (DAG) of alignments does just that. The objective function is specifically tuned to finding SV events by taking into account the length and quality of alignments, the number of alignments in the OCS and the genomic distance between those alignments. Second, it must be able to report alignments similar to those in the OCS in order to allow for the use of combinatorial breakpoint detection algorithms that cluster multiple mappings per read (Hormozdiari et al., 2009; Quinlan et al., 2010). YAHA's use of an optimal DAG algorithm for discovery of the OCS and its ability to find collections of alignments similar to the OCS are completely novel. Third, it must be able to generate a large number of viable alignments to feed the above two algorithms. Long-read aligners such as BWA-SW (Li and Durbin, 2010) and AGILE (Misra et al., 2011) severely restrict the number of alignments under consideration early in query processing. While this improves speed, it reduces the likelihood of finding the OCS and precludes finding alignments similar to them. YAHA can produce the required large number of alignments. Optionally, the user can choose to output all of them. Other aligners such as MegaBLAST (Altschul et al., 1990) and SSAHA2 (Ning et al., 2001) can also produce numerous alignments, but have no notion of an OCS. Fourth, the aligner must be able to run in a reasonable amount of time. YAHA uses a unique combination of heuristics and optimizations to accomplish this. We use a hashing scheme similar to SSAHA, but with a considerably faster approach for sorting hash table seeds. We use banded Smith–Waterman (SW) and a modified version of MegaBLAST's X-Dropoff heuristic for extensions. Finally, we calculate the OCS without unduly impacting performance by using a time and space optimized DAG algorithm. YAHA is the only aligner that does all four of these things well, and therefore is uniquely well suited to SV breakpoint detection. In addition, it is important to score alignments using a metric that is capable of accommodating a wide range of error profiles in order to perform well on queries from diverse sources, including existing and future long-read sequencing technologies (Mardis,

2013). To accomplish this, YAHA utilizes Affine Gap Scoring (AGS) with user specified cost/reward parameters.

4.2 Methods

YAHA uses a seed and extend strategy for DNA alignment. Alignments are output in SAM format (Li et al., 2009). YAHA breaks the alignment process into six stages. Steps 5, *Optimal Query Coverage* (OQC) and 6, *Filter By Similarity* (FBS), are not included in any other DNA aligner. Although many of the basic algorithms used by YAHA are not novel, their inclusion in a DNA alignment tool is.

4.2.1 Find seed matches

A base-pair sequence of fixed-length k is called a *k-mer*. YAHA uses a hash table index to locate the set of locations (seeds) where each k-mer in the query sequence appears as a subsequence of the reference. There are three parameters that control the creation of the index; seed length (k), the *skip-distance* between the starting locations of seeds in the reference, and the maximum allowed hits for a k-mer before it is considered too repetitive to be useful (*maxHits*). Typical values of k range from 8 to 15. The skip-distance can range from 1 (max overlap) to k (no overlap). YAHA builds an index once per desired combination of reference genome and index parameters and stores it in a file. While performing alignments, the index file is accessed via memory-mapped IO as if it were stored in RAM.

For mammalian genomes, and k≤15, a very large percentage of all unique k-mers will appear at least once in the genome. Therefore, a natural way to form a hash key is to compress the k-mer using 2 bits per base, then use it as an offset into a table with an entry for each k-mer (**Figure 4.1A**). The list of reference hits for all keys are concatenated in one large array called the *Reference Offset Array* (ROA). This index structure is the one used in YAHA and was taken directly from SSAHA. Similar indexing strategies are used by MegaBLAST, BLAT (Kent, 2002), and others. For mammalian genomes, we find that k=15 and skip-distance=1 (a *15/1* index) performs quite well, and we use such an index for all YAHA test runs discussed below. Like MOSAIK (Lee et al., 2014) YAHA allows for sampling of hits in the reference down to the specified maxHits parameter setting. This can have a dramatic impact on the trade-off between sensitivity and run-time, as we later show in **Section 4.3**.

4.2.2 Combine seed matches into *fragments*

Next, seeds are joined together to form extended seeds or *fragments* of contiguous matching bases between the query and the reference (**Figure 4.1B**). Seeds that can be strung together in this way appear on the same *diagonal* in a plot of Reference Offset (RO) versus Query Offset (QO) (**Figure 4.1C**). The query length (QL) determines the number of k-mers that appear in the query. Let N equal the sum over QL of the number of reference hits for the k-mer starting at each query offset. To find extended seeds, many aligners collect all n seeds for a query into an array and use an $\boldsymbol{O}(n\log n)$ sort to collocate seeds to be placed in a fragment. However, since the seeds for each k-mer are presorted by RO in the ROA, YAHA instead performs a QL-way merge of the presorted ROA regions. A Priority Queue (Binary Heap) is used to aid the merging process. The process proceeds in two phases. First, the queue is initialized with the first reference hit for each query offset. Then, we repeatedly extract the minimum value of the heap, and replace it with the next reference hit for the retrieved query offset (if any). The fragments are collected into an array as the seeds are extracted from the queue. This phase continues until the heap is empty. This approach reduces the complexity to an upper bound of $n\log QL$ because the heap will contain a

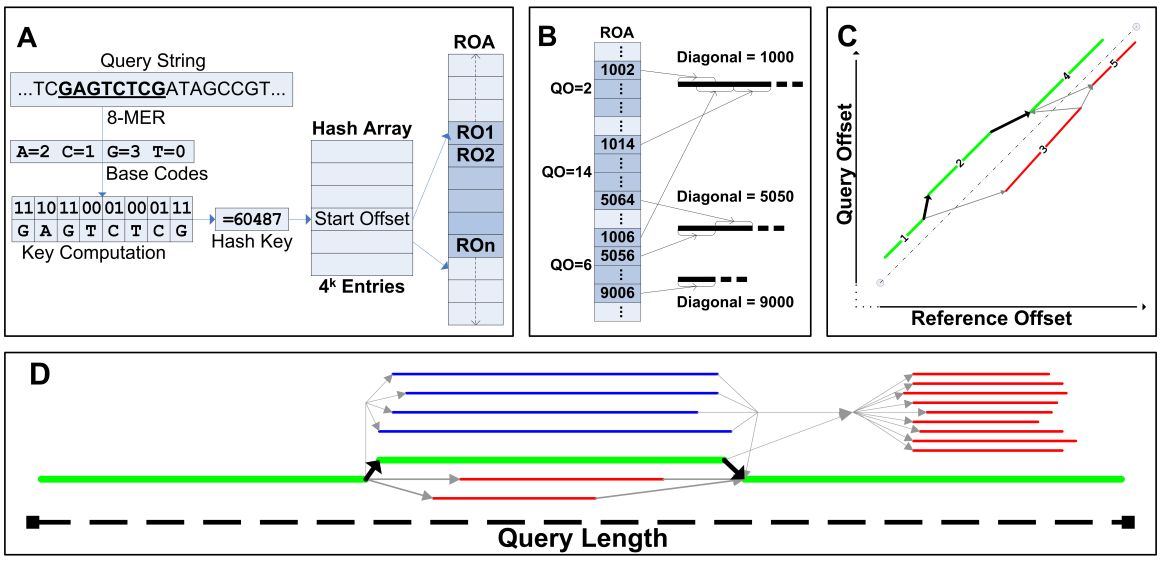

Figure 4.1. YAHA methods.
A) Starting at each location in the query we form a k-mer that is then converted to a hash key by compressing the bases in the k-mer using 2-bits per base. That hash key is then used to directly index into the Hash Array, giving the starting offset and length of the subset of the ROA that contains the collection of reference locations for that k-mer. B) Next, seed matches from the query and reference that fall along the same diagonal are collected into extended seeds called "fragments" by merging the pre-sorted ROA regions for each query location using a Binary Heap. C) In any given region of the reference, many fragments can be included in a potential alignment. YAHA uses a graph algorithm to find the set that maximizes the estimated score. In this example, fragments 1, 2, and 4 form the best alignment. D) During the Optimal Query Coverage algorithm, we will find the best collection of "primary" alignment (green lines) that has the highest non-overlapping sum of scores. Filter By Similarity is then used to determine the remaining "secondary" alignments (blue lines) that are highly similar to any primary alignment. The remaining alignments (red lines) are not included in the output for the query.

maximum of QL entries and will become smaller over time as loci on the query exhaust their lists of seed matches. In addition, we do not need to create the large n element array, saving memory footprint and possibly improving cache locality. To the best of our knowledge, no other hash based aligner uses this optimization to reduce the cost of sorting the seed matches.

4.2.3 Combine fragments into an alignment

YAHA next finds the best potential alignment in each region of the reference by combining the fragments that contribute to the highest estimated alignment score in that region. Selecting fragments can be difficult in regions with tandem repeats as there may be numerous overlapping fragments with various distances between their diagonals. We calculate the estimated score for each possible collection of fragments in the region using the AGS parameters; fragments are scored as matches, while differences between fragment diagonals are scored as a single indel. YAHA uses a graph algorithm that finds the path with the maximum estimated score (**Figure 4.1C**). The nodes of the graph (colored lines) represent fragments, and the edges (gray lines) represent the cost or benefit of one fragment succeeding another in the alignment. Since fragments earlier in the query can only be succeeded by fragments later in the query, the graph is directed and acyclic (a DAG), with a maximum of $n^2/2$ edges for n fragments. In DAGs, such min/max path algorithms need visit each edge only once in the proper (topological sort) order. By placing the nodes in an array and presorting them by starting QO using a conventional $\boldsymbol{O}(n\log n)$ sort, we perform the graph algorithm without ever forming the edges. This saves space and improves cache behavior. Each node is visited sequentially while checking against all nodes above it in the sorted array. If an edge is allowed between two nodes, we immediately score and relax the edge, resetting the best score and best-path back-pointer in the later node when appropriate.

Traditionally, the task of selecting the best seed matches to include in an alignment has been performed by using Dynamic Programming (DP) (Pearson and Lipman, 1988). Straightforward DP implementations require time and space proportional to n^2. However, the Hirschberg algorithm (Myers and Miller, 1988), reduces the DP space requirement to $\boldsymbol{O}(n)$, but approximately doubles the runtime. The graph algorithm used in YAHA also uses time proportional to n^2, and space proportional to n, but without this added complexity. We reuse this graph algorithm in the OQC phase described below.

Once a best set of fragments is found for a reference region, it is placed into a potential alignment to be completed as described below. If any of the remaining fragments from this reference region do not overlap on the query with any potential alignments already found in this region, they are used in another run of the graph algorithm. This process continues until there are no remaining fragments in an uncovered portion of the query. We next discard all potential alignments that contain a number of seed matches that falls below a user specified threshold (minMatch). It is common for aligners to define this threshold in terms of the number of seed hits from the index. As the seeds can be overlapping, YAHA instead uses a threshold for the total number of non-overlapping bases that appear in seeds. We believe such a threshold is both more accurate and easier for the user to manage.

4.2.4 Complete alignments using DP

YAHA now takes each potential alignment from above, and completes the calculation of the full alignment. It uses a modified version of SW only to find the portions of the alignment that fall between fragments, and to find the best forward and backward extensions for the

alignment. Our implementation of SW calculates AGS with a well-known strategy to reduce memory usage first proposed by (Gotoh, 1982). YAHA also uses a common heuristic called 'banding' that reduces costs by calculating only the DP values near the diagonal of the array. Because the endpoints of extensions are not known, we heuristically use twice the bandwidth during extension as used between fragments, and a simplified version of the well known 'X-Dropoff' heuristic (Zhang et al., 1998) which stops extending an alignment when the score for the current extension is more than X below the best score for a shorter extension. YAHA almost always finds the optimal local alignment. However, due to the use of various heuristics such as X-Dropoff and banding, this is not guaranteed.

4.2.5 Apply Optimal Query Coverage algorithm

Optionally, YAHA can report all alignments identified through the above steps. This feature is invaluable when it is important to gain knowledge about the uniqueness of a query sequence or the distribution of repeats in the reference genome. However, in order to define SV breakpoint locations, it is often preferable to ignore the potentially large numbers of irrelevant alignments that arise from repeats embedded within larger, more unique portions of the query. For this purpose we have devised an algorithm called OQC, which finds the set of alignments that cover the length of the query with the maximum coverage score. This *Optimal Coverage Set* (OCS) is composed of one or more *Primary Alignments*. This algorithm greatly aids in reconstructing breakpoint architecture and is a crucial, and novel, feature of YAHA.

To find the OCS, we use a max-path DAG algorithm similar to that described in **Section 4.2.3** above. The nodes now represent the alignments, and the edges represent one alignment being included with another in the OCS (**Figure 4.1D**). We again presort the alignments by starting QO to avoid creating the edges. In cases where two alignments overlap at the breakpoint, as occurs when structural variants are generated by homology dependent mechanisms, the score of the better alignment in the overlap region is used. In order to avoid an overly fractured OCS, a penalty is applied for each split between adjacent alignments. This penalty is the product of two factors. The first is a user-supplied parameter called the *Breakpoint Penalty* (BP). The second is a *Genomic Distance Penalty* (GDP) calculated as $\log_{10}$ of the number of base pairs along the reference genome between the two alignments. The user can specify a maximum GDP (maxGDP). Alignments on separate chromosomes always incur the maxGDP. Through these two parameters, the user can control how sensitive the query coverage score is to genomic distance, and how large the non-overlapping portion of an alignment must be before it is included in the OCS. With a relatively high maxGDP, collections of alignments near each other on a reference chromosome will be favored, helping to identify deletions, tandem duplications and inversions. A low maxGDP will be more neutral to genomic distance. A higher BP will favor alignment sets with fewer, larger, alignments.

We believe that this OQC calculation allows for the discovery of biologically meaningful collections of alignments. We show below that YAHA's OQC algorithm is better at discovering SV events than the heuristic approach used by BWA-SW for finding split alignments.

4.2.6 Apply Filter By Similarity algorithm

Optionally, we next perform the FBS step to identify *Secondary Alignments* that have a high length overlap and score agreement with a primary alignment (**Figure 4.1D**). This allows the user to gain knowledge of repetitive mappings specifically for those sections of the

query that comprise a primary alignment. This is required for clustering algorithms designed to identify breakpoints in repetitive genomic regions, and may be useful for characterizing the repetitive structure of fully sequenced reference genomes (Bailey et al., 2002). This novel feature of YAHA combines the utility of finding large numbers of alignments with the advantage of defining the optimal collection along the query.

4.3 Results

To demonstrate YAHA's power and flexibility, we measure its performance in three test scenarios. First, we show that YAHA is sufficiently sensitive to find large numbers of alignments for queries with repeated (sub)sequences. Second, we measure YAHA's ability to accurately find alignments when using the OQC algorithm for non-chimeric queries. Third, we test the OQC and FBS algorithms by measuring YAHA's ability to detect SV breakpoints in chimeric queries. For each test, we compare YAHA to what we believe to be best of breed among commonly used aligners for that specific task. In the sensitivity test, we compare against MegaBLAST because it is generally considered one of the most sensitive heuristic aligners for finding a large number of alignments in a practical amount of time. Because we use the same indexing strategy as SSAHA2, we also include it in this test. BWA-SW only reports primary alignments so cannot be included in the sensitivity comparison. We do not include MegaBLAST or SSAHA2 in the accuracy or SV detection tests because neither has any strategy for finding an OCS. For these tests, we compare our results to BWA-SW, which is the most widely used long-read aligner, and the most challenging competitor to YAHA for finding primary alignments on either chimeric or non-chimeric queries. In particular, it has already been shown that BWA-SW outperforms SSAHA2 and BLAT on non-chimeric reads (Li and Durbin, 2010). In all these tests, CPU time is an important metric, as any alignment task is easy to perform by brute force if an aligner is given unlimited computer resources. Finally, we note that it is difficult to compare results from different aligners because most are highly parameterizable, but do not share all the same parameters. We have made a considerable effort to select the most effective parameters to use for YAHA and the other aligners, but we cannot exclude the possibility that untested parameter combinations might produce superior results to those we present here.

The data for the accuracy test was generated using WGSIM (Li et al., 2009a) to sample reads from the hg18 reference genome with the lengths and error rates shown in **Table 4.2**. For the sensitivity test, we focus on the first of these datasets; 100,000 queries of length 100 with a 2% error rate. For the SV detection test, we used our own tool, SVsim, to simulate SV events of various types.

All tests were run on a server class machine with 4 Xeon X7350 processors, 128 GB of shared RAM, running CentOS 5.5. YAHA's 15/1 index and compressed reference total 15.5 GB, SSAHA's total 22.3 GB and BWA-SW's index and reference total 7.4 GB. However, we believe that index size is a minor concern given modern computing environments.

4.3.1 Sensitivity test

To test sensitivity, we ran YAHA bypassing the OQC and FBS algorithms and output all alignments that pass applicable thresholds. An issue arises in trying to compare results from YAHA, MegaBLAST, and SSAHA2 because they do not have the same threshold parameters, and SSAHA2 does not support AGS. To equalize results, we applied an external filter to keep only alignments ≥50bp in length, and ran each aligner with its thresholds set so that such alignments were not filtered out internally. Also, each aligner uses different criteria to determine if two alignments are 'distinct' enough to report separately. To account

for this, we filtered out alignments that are overlapping on the reference with another alignment for the same query. We refer to the final set of filtered alignments as *GE50U* alignments. As there is no practical way to determine every location in the reference that can be aligned to a given query, we measure relative sensitivity in this test. Version 2.2.19 of MegaBLAST was run using parameter settings that are as sensitive as possible with a seed length of 15. The result is the baseline against which we compare the sensitivity of YAHA and SSAHA2.

Seven aligner runs are used in this test (**Table 4.1**). We study four YAHA runs with parameter settings representing different points along the sensitivity spectrum. For three of the runs, we used a minMatch of 15 (one seed hit) and a maxHits of 65,525 sampled (Y1), 10,000 sampled (Y2) and 10,000 unsampled (Y3). The fourth run (Y4) used faster but less sensitive parameters; maxHits of 650 unsampled and minMatch of 20. We also study two runs of version 2.5.1 of SSAHA2. The first (S1) used SSAHA2's built-in 454 mode, which implies many SSAHA2 parameters, including ones for the Crossmatch back-end, and a 13/3 index. SSAHA2's default and solexa modes do not perform well in this test and are not included. The second (S2) also used the 454 mode Crossmatch parameters, but with a 15/1 index, SSAHA2 default value of 10,000 for maxHits, and a minMatch of 1. These parameters make this run directly comparable to the Y3 run.

Table 4.1 shows the test results. The percentage of queries with the same number of GE50U alignments as M is similar across runs. In fact, 64% of the queries produce the same number of GE50U alignments across all seven runs. Of these, 97.7% produce a single GE50U alignment. This indicates there is high agreement between aligners for queries that map to unique locations on the reference. In addition, all of the YAHA runs produce significantly more total and GE50U alignments per second than any other aligner runs.

Y1 uses the most sensitive YAHA parameters possible for a 15/1 index, and produces more total alignments, more GE50U alignments and more queries with a greater number of GE50U alignments, in less runtime than MegaBLAST. This is a striking result. We analyze it further by expanding the last three columns of the Y1 row of **Table 4.1** into a histogram of the difference in the number of GE50U alignments in the Y1 run versus M (**Figure 4.2**). This shows that the two aligners can differ in the number of alignments for queries with highly repetitive (sub)sequences by five orders of magnitude. Yet, the graph is highly skewed in Y1's favor, showing that YAHA is significantly more sensitive than MegaBLAST at identifying large numbers of alignments for such queries, while using less runtime.

YAHA greatly outperforms SSAHA2. For example, Y3 and S2 use comparable parameters, yet Y3 reports ~297X more GE50U alignments at ~12X greater speed (GE50U/s). The algorithmic basis for this dramatic difference in sensitivity is unclear. While S1 fared somewhat better, we note that this disparity persists across a wide range of SSAHA2 parameters (data not shown).

Y2 and Y3 agree in all parameter settings except Y2 uses an index with random sampling of k-mers that appear more than 10,000 times in the hg18 genome. More than 99.9996% of all 15-mers appear fewer than 10,000 times in hg18, yet Y2 requires ~4X the runtime and produces ~4.5X the number of GE50U alignments. This shows that the very few highly repetitive k-mers greatly impact queries that contain them. It also shows, together with the use of sampling in Y1, the improvement in sensitivity derived from sampling such k-mers instead of excluding them.

Table 4.1. Results of the sensitivity test.
The first two columns give the name and aligner parameters, column 3 gives the runtimes, columns 4-7 contain the total alignments, GE50U alignments, total alignments/second, and GE50U alignments/second, and the last 3 columns show the number of queries with >, =, or < the number of alignments as the MegaBLAST run.

Run	Aligner and Parameters	CPU Secs	Alignments				Versus MegaBLAST		
			Total	GE50U	A/Sec	U/Sec	> M	= M	< M
M	MegaBLAST: wordLen=15, score=15	190,773	4,012,294,854	1,827,862,215	21,032	9,581	0	100,000	0
Y1	YAHA: minMatch=15, maxHits=65525S	160,501	6,085,988,010	2,343,744,189	37,919	14,603	30,638	68,357	1,005
Y2	YAHA: minMatch=15, maxHits=10000S	91,097	3,403,790,544	1,470,115,221	37,364	16,138	23,789	68,387	7,824
Y3	YAHA: minMatch=15, maxHits=10000	22,385	950,852,793	327,644,121	42,477	14,637	20,021	68,371	11,60
Y4	YAHA: minMatch=20, maxHits=650	284	11,680,597	6,796,153	41,129	23,930	716	69,536	29,74
S1	SSAHA2: 454 mode	1850	6,066,013	5,488,465	3,279	2,967	834	66,634	32,53
S2	SSAHA2: minMatch=1, maxHits=10000	937	2,633,833	1,101,352	2,811	1,175	120	65,622	34,25

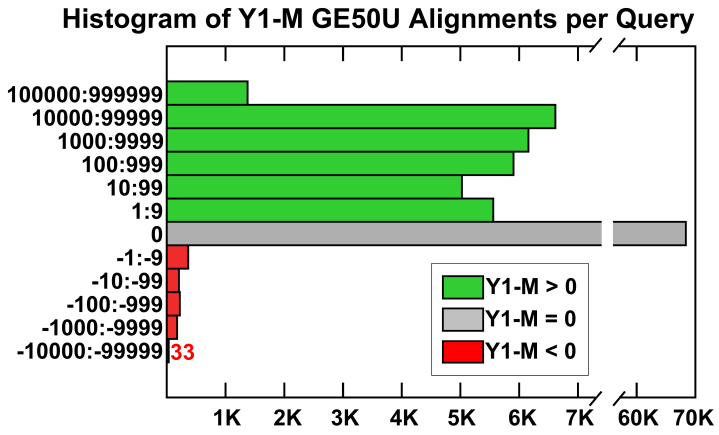

Figure 4.2. Alignment statistics in run Y1 vs. M.
Histogram of the number of queries in the Y1 YAHA run with varying numbers of greater, equal, and fewer GE50U alignments than MegaBLAST (M). Note the $\log_{10}$ scale bucket sizes. The total number of queries >0 is 30,638 and <0 is 1005 as in **Table 4.1**.

Over 99.99% of all possible 15-mers appear fewer than 650 times in hg18. Therefore, even without using sampling, this acts as a reasonable maxHits cutoff for relatively fast runs. Y4 uses this maxHits threshold, and further reduces runtimes by using a minMatch of 20 instead of 15. These prove to be effective settings, as Y4 produces ~1.6X as many GE50U alignments per CPU second as the other YAHA runs, and ~2.5X as many as M. Given these results, we use 650 as the maxHits threshold for YAHA in the accuracy test below. The nearly 11.7 million total alignments in the Y4 run act as the input to the OQC algorithm across the 100,000 queries in the first dataset of the accuracy test as we discuss next.

4.3.2 Accuracy test

We now compare the accuracy of YAHA to BWA-SWin finding primary alignments. We use the same process for generating synthetic queries as used in the accuracy test in the BWA-SW paper. However, we use slightly different accuracy metrics. In that study, they determined the false positive rate using 'mapping quality', a heuristically determined measure of an aligner's confidence in the uniqueness of its alignments. Instead, we use as the benchmark the optimal alignment and score of each generated read at the source reference location found by SSEARCH, a tool from the FASTA suite (Pearson and Lipman, 1988) that uses full SW to find the best local alignment. For each aligner, we place each query into one of four categories. If no alignment was generated for a query, it is a false negative. If a primary alignment matches the optimal alignment found by SSEARCH, it is a 'match'. For each remaining query, we independently calculated the best non-overlapping score of the alignment(s), called the Coverage Score (CS). If the CS is less than the SSEARCH score, it is a false positive. If the CS equals or exceeds the SSEARCH score, the alignment(s) produced are at least as viable as the one from the source location. Such queries are not real false positives, and are reported in their own category. We believe that the use of externally verified alignment scores is a far less biased and more precise metric of aligner accuracy because it isolates the effects of alignment heuristics from the mapping quality heuristics.

BWA-SW version 0.5.8 was run with default settings. YAHA was run with OQC turned on, BP=5, maximum GDP (maxGDP)=5, maxHits.650, and varying values for minMatch of 20, 26, 38, 100 and 500 for the different QLs, respectively. **Table 4.2** shows the results of the BWA-SW and YAHA runs using these 15 datasets. The aligners differ most on the 100-mer queries with 5% and 10% error rates, and the 200-mer queries with 10% error rate. These three datasets are the most challenging for both aligners, but YAHA has a significantly lower false positive and especially false negative rate, accounting for most of the large difference in these metrics shown in the Totals column. YAHA has a lower false negative rate for six of the datasets, versus five for BWA-SW. However, for three of the datasets in which BWA-SW has a lower false negative rate, YAHA merely fails to align a single query. YAHA has a lower false positive rate for eleven of the datasets, versus two for BWA-SW. YAHA has a lower sum of error rates for ten of the datasets, versus two for BWA-SW. The aggregation of results by query and by dataset both contain biases, albeit different ones. The former is biased by the fact that the datasets do not all contain the same number of queries, while the latter is biased by datasets with a small number of queries that differ. Nonetheless, YAHA achieves better results than BWA-SW for both aggregation strategies.

As a further test of accuracy, we compare all matching alignment scores against the optimal scores determined by SSEARCH. For reasons discussed in **Section 4.2**, YAHA produced sub-optimal scores for 20/522,244 matching alignments (0.0038%). BWA-SW produced a sub-optimal score for 1/483,786 matching alignments (0.0002%). All the sub-optimal alignments from both aligners were from datasets with a 10% error rate.

Table 4.2. Results of accuracy test.
Accuracy comparison of YAHA to BWA-SW over 15 datasets generated in a similar fashion as those in the BWA-SW paper. Each query is put into one of four categories depending on the accuracy of the alignment (see text for details). The CPU time in seconds, and total error rate for each run are also shown. The right-most column shows the aggregate runtimes and category percentages.

		100K 100bp Reads Error Rate			50K 200bp Reads Error Rate			20K 500bp Reads Error Rate			10K 1000bp Reads Error Rate			1K 10,000bp Reads Error Rate			
	Metric	2%	5%	10%	2%	5%	10%	2%	5%	10%	2%	5%	10%	2%	5%	10%	TOTALS
BWA-SW	CPU Secs	160	135	102	220	186	140	259	194	154	219	193	142	155	146	129	2534
	% False Negatives	0.44	5.21	27.4	0.00	0.13	5.44	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	6.61
	% Matching	96.0	89.3	64.0	98.2	97.5	89.3	98.9	98.9	98.2	99.3	99.2	99.2	99.8	99.5	98.1	89.10
	% CS ≥ SSEARCH	2.96	2.92	2.69	1.74	1.71	1.66	1.09	1.02	1.13	0.68	0.74	0.66	0.20	0.50	0.90	2.21
	% False Positives	0.56	2.53	5.85	0.11	0.70	3.63	0.01	0.12	0.56	0.02	0.02	0.12	0.00	0.00	1.00	2.09
	% Total Error	1.00	7.74	33.3	0.11	0.83	9.08	0.01	0.12	0.66	0.02	0.02	0.12	0.00	0.00	1.00	8.70
YAHA	CPU Secs	284	241	176	212	171	109	245	188	112	108	86	58	81	79	66	2216
	% False Negatives	0.32	0.12	0.55	0.03	0.00	0.02	0.00	0.01	0.00	0.01	0.01	0.07	0.00	0.00	0.00	0.19
	% Matching	96.3	95.5	91.8	98.1	97.9	97.2	99.0	98.8	98.8	99.3	99.2	99.1	99.9	99.7	99.2	96.18
	% CS ≥ SSEARCH	2.83	3.03	3.72	1.71	1.77	1.87	1.04	1.18	1.06	0.71	0.81	0.76	0.10	0.30	0.80	2.42
	% False Positives	0.55	1.31	3.97	0.16	0.34	0.93	0.02	0.02	0.11	0.00	0.01	0.03	0.00	0.00	0.00	1.21
	% Total Error	0.87	1.43	4.52	0.19	0.35	0.96	0.02	0.03	0.11	0.01	0.02	0.10	0.00	0.00	0.00	1.40

Summed over the datasets, YAHA uses less CPU time than BWA-SW. This is impressive given that YAHA considers many more alignments. For example, in **Table 4.1** the Y4 run on 100-mers at 2% error rate produces ~11.7 million alignments, from which the OQC algorithm selects 99,696 primary alignments. In contrast, BWA-SW severely restricts the number of potential alignments early during query processing. By considering many more alignments, YAHA achieves greater accuracy. The advantages of using many alignments as input to OQC becomes more apparent in our test of SV breakpoint detection in the next section.

4.3.3 SV detection test

Finally, we compare YAHA to BWA-SW in their ability to correctly identify SV breakpoints with split-read mappings. This is an important criterion for evaluating long read aligners, because as read lengths grow, split-read mapping is rapidly replacing PEM as the method of choice for SV detection. We constructed three simulated datasets using SVsim, a tool we devised for this purpose (Faust and Hall, Chapter 2). First, we simulated 10,000 SV events with lengths from 100-10Kbp in random genome locations with equal numbers of deletions, tandem duplications, and inversions, as well as insertions from a random distant genome location. For events of length ≤500, we generated a single 'contig' spanning the event, with 500 flanking bases on each side. For larger events, we generated a contig for only the left breakpoint, with 500 flanking bases. We then generated 500-mer reads by sampling these contigs with WGSIM using a 2% error rate and 5X coverage. We examined only a single breakpoint for each variant, yielding 10,000 total breakpoint calls. BWA-SW was run with default settings except the z parameter was set to 1, 2, 5 and 10 to investigate the trade-off between runtime and sensitivity. YAHA was run with similar parameters as in the accuracy test, using a minMatch of 25 with increasing values of maxHits. We measured the percentage of queries with a split alignment that verified the correct SV breakpoint (within 5bp) and the total number of verified breakpoints (**Figure 4.3A**).

Both aligners perform very well on this dataset, identifying ~98% of the simulated breakpoints. However, YAHA verifies breakpoints in more queries than BWA-SW at comparable runtimes. This test shows that both aligners are quite effective at identifying isolated SV breakpoints in random (mostly non-repetitive) genomic regions.

To investigate performance at breakpoints involving repetitive sequences, we simulated 10,000 Alu insertion events. We randomly selected 1000 intact Alu elements with minimal divergence to the canonical active elements (milliDev ≤10 and length ≥300) from the UCSC RepeatMasker annotation track, and injected each into 10 random genome locations. Read simulations and performance metrics are as above. This is a challenging test, because to detect an Alu insertion by split-read mapping, the breakpoint-containing read must be aligned correctly not only to the flanking sequence at the recipient locus (where the Alu inserted), but also to the Alu element at the correct donor locus. This is difficult given the extremely large number of Alu elements in the reference genome, the high DNA sequence similarity shared between them, and the simulated error rate in the reads.

As a result, both aligners identify fewer breakpoints from far fewer queries for Alu insertions than for the standard SVs. Yet, YAHA again identifies slightly more breakpoints (not shown) from more queries (**Figure 4.3B**). For example, using ~50K CPU seconds YAHA identifies 79.6% of breakpoints from 24.5% of queries, while BWA-SW finds 76.8% of breakpoints from 21.7% of queries. In addition, this test shows the utility of YAHA's FBS algorithm. When both primary and secondary alignments are taken into account, YAHA identifies 96.3% of breakpoints from 43.2% of queries. These additional alignments

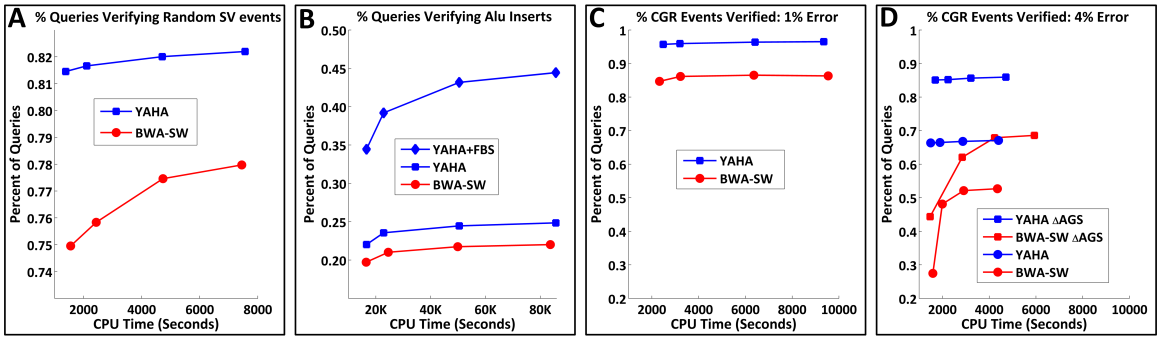

Figure 4.3. Results of SV detection test.
Shown are graphs of the percentage of queries with which each aligner correctly verified an SV breakpoint for various types of SV events vs. the amount of CPU time consumed. Note the large improvement with the inclusion of YAHA's secondary alignments in the Alu dataset (B). Also note the marked improvement for both BWA-SW and YAHA in the CGR dataset with 4% error rate by changing the AGS parameters to lower the penalty for indels relative to replacements (D). Still, YAHA outperforms BWA-SW with both sets of AGS parameters. Graphs C and D are shown with the same axes to ease comparison.

enable the discovery of repetitive element insertion events using combinatorial clustering algorithms (Hormozdiari et al., 2009; Quinlan et al., 2010).

Recent evidence indicates that complex genomic rearrangements (CGRs) are a common form of SV in both normal and cancer genomes (Malhotra et al., 2013; Quinlan and Hall, 2012). The most extreme example of this is chromothripsis (Stephens et al., 2011), where chromosome regions are extensively rearranged due to the repair of chromosome shattering events involving hundreds of breakpoints. CGR events pose a unique challenge for breakpoint discovery because, with long-reads or assembled contigs, numerous breakpoints may be present on a single query. YAHA's OQC algorithm is designed to select the optimal collection of alignments and should handle such situations better than heuristic strategies. To test this, we simulated 1500 chromosome shattering events each with a total length of ~30Kbp. Of these, 1000 involve a single random genomic location, and 500 combine fragments from two different genomic locations. Fragments were generated from random locations within the selected regions, with an average size of 300bp and a minimum size of 50bp. Of these fragments, 30% were deleted, 10% duplicated, and 50% inverted. The resulting collection of fragments was then randomly shuffled and 'ligated' into a single contig. This generated a total of 129,915 CGR breakpoints. The contigs were used directly as long reads after impressing two different error profiles. The first models a contig reconstructed via *de novo* assembly of short reads, and has a 1% error rate, 10% of which are indels. The second models a single long read from third generation sequencing technology, such as the Oxford Nanopore (Clarke et al., 2009) instrument, and has a 4% error rate, 90% of which are indels.

YAHA greatly outperforms BWA-SW with the long CGR contigs. In the 1% error profile data, YAHA finds ~96% of the breakpoints, versus ~86% for BWA-SW (**Figure 4.3C**). In the 4%error profile data, using the default AGS parameters, YAHA finds ~67%of the breakpoints regardless of the maxHits setting, while BWA-SW finds from 27.5% to 52.7% of the breakpoints as the z parameter is increased. BWA-SW does not perform well on this test with its default z=1, and requires ~2X the runtime of YAHA to approach its sensitivity asymptote (**Figure 4.3D**). Both aligners use the same default AGS parameter settings (Match=1, Mismatch=-3, GapOpen=-5, GapBase=-2). However, these parameter settings are tuned for low error rates and especially low indel rates, and are not optimal for the high-indel CGR dataset with a 4% error rate. Thus, we re-ran both aligners against the 4% error rate CGR dataset with AGS parameters that increase the relative penalty for replacements versus indels (Mismatch=-5, GapOpen=-2, GapBase=-1). While both aligners now do significantly better, YAHA still far outperforms BWA-SW (**Figure 4.3D**). BWA-SW now finds between 44% and 69% of the breakpoints, while YAHA finds ~85%. This shows the importance of using an alignment scoring strategy, such as parameterized AGS, to handle the high error/indel rates that exist in current and future third generation sequencing technologies (Mardis, 2013).

The inclusion of a genomic distance penalty in the objective function of YAHA's OQC algorithm undoubtedly aids its performance in these tests, as it allows YAHA to favor collections of alignments for the OCS that are near each other in the genome.

4.4 Conclusion

We have shown that YAHA is a fast and effective all-purpose aligner that outperforms best-in-class tools for very three different tasks: (i) reporting all mappings per query; (ii)

reporting the single best mapping and (iii) identifying split-mappings that define one or more SV breakpoints within a query. YAHA's main strength as a general alignment tool is that it simply attempts to identify all possible matches according to the parameters set by the user. YAHA is able to explore many possible alignments without sacrificing speed through the use of a number of pre-existing and novel heuristics, as well as optimized implementations of computationally intensive procedures such as seed-match sorting, banded SW and max-path graph algorithms.

The most important and novel feature of YAHA is that it determines the set of the alignments that cover a query using an algorithm that provably optimizes a biologically relevant objective function tuned to SV breakpoint detection. This capability, as well as the ability to report secondary alignments using FBS, will be invaluable for SV mapping experiments that rely on long reads or assembled contigs. As we have shown, these methods are especially powerful for defining breakpoints caused by repetitive elements, and for reconstructing highly complex genome rearrangements.

4.5 Update

Since the original publication of YAHA, a new aligner BWA-MEM has been developed with the combined features of BWA and BWA-SW (Li, 2013). As each end of an Illumina paired-end read is now ~100bp, it has become possible to find split-read mappings from these reads. In addition, such an aligner has the advantage of using information from the pairs to locate potential split-read mappings. YAHA still beat BWA-MEM in the accuracy test, which uses singleton reads (data not shown), albeit by a narrower margin. In addition, in a recent test of SV detection using LUMPY (Layer et al., 2014), use of the split-read mappings from YAHA had better sensitivity than those from BWA-MEM, with equally low false discovery rate (**Figure 4.4**). However, BWA-MEM is now much faster than YAHA due primarily to the use of SSE vector instructions for Smith-Waterman calculations; an upgrade that would also improve YAHA's performance (Farrar, 2007).

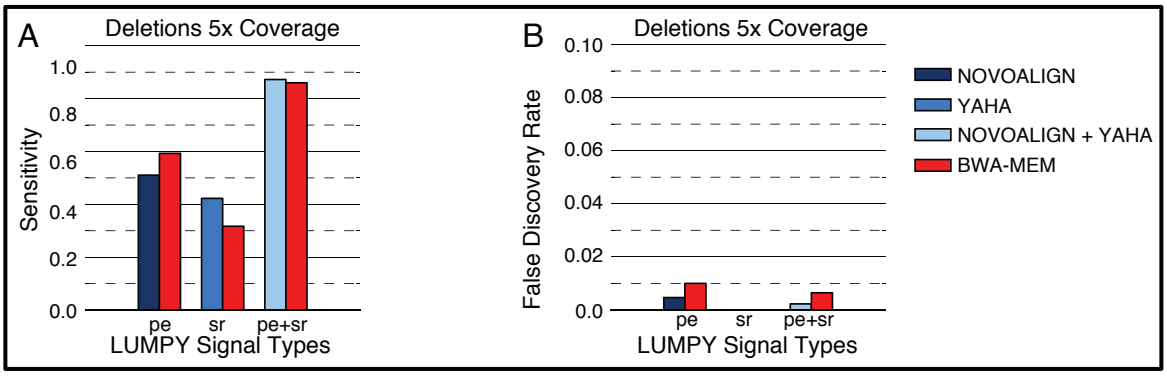

Figure adapted from (Layer et al., 2014)

Figure 4.4 LUMPY sensitivity use split-read mappings from YAHA vs. BWA-MEM. A) YAHA provided better sensitivity than BWA-MEM. B) Both show equally low false discovery rate.

5 Mutational Landscape of Single Post-Mitotic Neurons

The material for this chapter is largely from an manuscript produced in preparation for publication (Hazen et al., 2015). This research was performed as a large collaborative effort between Kristin Baldwin's Lab at the Scripps Research Institute in La Jolla CA and Ira Hall's Lab at the University of Virginia. The original manuscript had two first authors, Jennifer Hazen from the Baldwin Lab, and myself, and two corresponding authors, Kristen Baldwin and Ira Hall. Several other authors performed experiments and other work and made experimental suggestions, but were not primary contributors to the manuscript itself. **Section 5.2.1** was written at Scripps. All other sections have had drafts and rewrites contributed by all four primary authors. I have added and changed text, especially in **Section 5.2.3** and the Discussion (**Section 5.3**), for the version contained herein, and combined main and extended figures and tables from the original manuscript into a sequential numbering scheme consistent with their references in the text. As these tables and figures are referenced both in this chapter and **Appendix 1**, they are inserted at the end of this chapter in numeric order with all tables before the first figure. In addition, there are 11 supplemental tables (**Tables S1-S11**) prepared by me that appeared in the original manuscript in an excel spreadsheet that is an adjunct to this document.

The creation and validation of the knock-in mouse, derivation of the clonal neuronal cell lines by SCNT, the cloning of live mice by TEC, extraction of DNA for DNA sequencing, RNA Sequencing, and PCR validations of SNV and indels were performed at Scripps. Bioinformatic analysis of all DNA and RNA-Seq data as well as functional enrichment studies and statistical tests were performed in the Hall Lab by myself under Ira's guidance, although suggestions for analyses to perform and their biological interpretation were made by all four primary authors. In addition, most SV and MEI call validation was performed in Hall Lab by research staff. The detailed method descriptions for this research in **Appendix 1** make very clear which lab performed which tasks.

Neurons exhibit remarkable diversity and persist without cell replication for the lifetime of an individual. The extent to which mature neurons preserve their epigenetic plasticity and genome integrity is unclear. Here we apply nuclear transfer to reprogram and amplify the genomes of neurons from adult mouse brains. We show that mature neurons can retain sufficient plasticity and genomic integrity to produce fertile adult mice. However, genome sequencing uncovered extensive genomic diversity among neurons. Each neuron harbors ~110 unique somatic mutations including structural variants (0-3), transposable element insertions (0-4), indels (12-34) and single nucleotide variants (SNVs) (62-142). While we did not detect recurrent rearrangements, most neurons have acquired gene-disrupting mutations (0-3). Furthermore, neuronal SNVs are enriched in genic regions and in neuronally expressed genes relative to other lineages. These results predict that somatic mutations can impact neuronal function, particularly if they continually accumulate in post-mitotic neurons during maturation and aging.

5.1 Introduction

The human nervous system contains more than one billion neurons that are divided into a large but undefined number of subtypes based on features such as morphology, connectivity, location, and their patterns of gene expression. Production of distinct neuronal subtypes occurs at defined stages, typically during embryonic development when they exit the cell cycle, differentiate, integrate into circuits and then maintain their cell identity without cell division or replacement for the lifetime of an organism. These key features of

neurons – their diversity, their longevity and their irreversible post-mitotic status – pose a series of intriguing and unresolved questions regarding neuronal genomes.

One such question is the extent to which neuronal genomes maintain epigenetic plasticity over the lifetime of an organism. Given that neurons neither divide nor serve as precursors of other cell types, and that they must maintain a stable cellular identity over many decades, one might predict that neuronal epigenetic plasticity might be limited relative to other cell types. One means to interrogate the developmental potency of neuronal nuclei is to reprogram them by somatic cell nuclear transfer (SCNT). The first reports of mouse cloning by SCNT suggested that neurons were resistant to reprogramming compared to other lineages (Wakayama et al., 1998). However, in 2004, two groups used SCNT to reprogram post-mitotic olfactory sensory neurons (OSNs) and produce fertile adult mice (Eggan et al., 2004; Li et al., 2004). These experiments demonstrated that the genomes of at least some post-mitotic neurons retain sufficient plasticity and integrity to produce all cell lineages in an adult organism. An important caveat to these results is that OSNs differ from most cortical neurons in that they are produced constantly throughout the life of an animal and have an average lifetime of only 6 weeks. Therefore, it is possible that the mice cloned from OSNs were derived from recently born, undifferentiated neurons, while more mature neurons remain refractory to reprogramming due to irreversible epigenetic or genetic changes. To address this possibility, a series of additional studies tested the developmental potency of various populations of cortical neurons (Kawase et al., 2000; Makino et al., 2005; Osada et al., 2002; Osada et al., 2005; Yamazaki et al., 2001). However, the only neural populations that proved successful at producing mouse clones by SCNT included cell types that were not mature post-mitotic neurons, leaving the authors to conclude that it is still unclear whether adult post-mitotic neurons harbor epigenetic or genetic changes that preclude reprogramming or impair developmental potency.

A second question that has been raised regarding neuronal genomes is whether they undergo programmed genomic rearrangements as a mechanism for generating neuronal cell type diversity (Chun and Schatz, 1999). Organisms ranging from yeast to lampreys to mammals are known to use programmed DNA rearrangements to generate cellular diversity. Perhaps the most well characterized example is the mammalian immune system, in which site-specific DNA rearrangements and somatic hypermutation generate antibody and T-cell receptor diversity. The similarities in neuronal and immune system cellular diversity, and the shared expression of some immune system proteins involved in DNA recombination led to the suggestion that neurons may undergo programmed DNA rearrangements during their development (Chun et al., 1991). Furthermore, neurons are affected by defects in DNA repair pathways that also impair the development of B and T cells, for unknown reasons (Frank et al., 1998; Gao et al., 1998). One hypothesis to explain this phenomenon is that DNA repair pathways might be important in neurons to generate cellular diversity arising from irreversible genomic changes that accompany differentiation. One way to test this hypothesis is to sequence the genomes of multiple post-mitotic neurons with sufficient resolution to detect DNA rearrangements or somatic hypermutation of neuronal genes.

A third question regarding neuronal genomes is the extent to which they differ from other genomes in the organism, termed somatic mosaicism. Recent advances in single cell whole genome sequencing and cellular reprogramming have begun to reveal that individual somatic cells harbor significant numbers of unique mutations, while cancer cells typically display even higher degrees of genomic individuality (Abyzov et al., 2012; Burrell et al.,

2013; Gore et al., 2011; Ji et al., 2012; Lupski, 2013; Young et al., 2012). Some somatic mutations are found in multiple cells in a lineage, while others appear to be rare or unique to an individual cell (Behjati et al., 2014). Previous studies have reported that neurons may have increased levels of aneuploidy or DNA content (Kingsbury et al., 2005; Rehen et al., 2001; Rehen et al., 2005; Westra et al., 2008; Westra et al., 2010), and recent single cell genome sequencing of human neurons revealed frequent, large-scale DNA copy number variants (CNVs) (Cai et al., 2014; Gole et al., 2013; McConnell et al., 2013). However, these studies were only able to detect megabase scale CNVs, and lacked sufficient resolution to identify the most common classes of mutation (single nucleotide variants and indels), as well as smaller or complex structural variants. At present, the extent to which neurons harbor these types of mutations remains unclear.

Another potential source of somatic mosaicism in neurons is mobile element insertion (MEI). In recent years, somatic MEIs have been detected in various regions of mouse (Muotri et al., 2005; Muotri et al., 2009), human (Baillie et al., 2011; Cordaux and Batzer, 2009; Coufal et al., 2009; Evrony et al., 2012), and fly brains (Li et al., 2013; Perrat et al., 2013). While the functional significance of somatic MEIs remains unclear, some have proposed that transposons may alter the expression of nearby genes to generate neuronal diversity (Muotri et al., 2005). Somatic MEIs have also been proposed play a role in disease. De-repression of transposable element transcripts has been implicated in a broad range of neurodevelopmental and neurodegenerative disorders, including ALS, Alzheimer's disease, frontotemporal lobar degeneration (Douville et al., 2011; Li et al., 2012), prion disorders (Jeong et al., 2010; Lathe and Harris, 2009), Fragile X Syndrome (Tan et al., 2012), and age-related macular degeneration (Kaneko et al., 2011). Similarly, elevated levels of somatic integration have been linked to neurodevelopmental disorders such as Ataxia Telangectesia (Coufal et al., 2011), schizophrenia (Bundo et al., 2014), and Rett Syndrome (Muotri et al., 2010). Preliminary estimates from studies of bulk tissue or groups of neurons predicted as many as 80 somatic MEIs per human neuron (Coufal et al., 2009) and 129 per drosophila neuron (Perrat et al., 2013). However, two more recent single cell sequencing based studies found less than one transposon insertion per human neuron (Evrony et al., 2012; Evrony et al., 2015), but a third found ~13.7 per hippocampal neuron and ~16.3 per cortical neuron (Upton et al., 2015). The discrepancy between these estimates highlights the importance of additional studies to understand the impact that transposons may have on neuronal development, aging, and neurologic disease.

Here, we report results from experiments designed to address these critical questions using a combination of two technologies: SCNT and whole genome sequencing. Using SCNT, we established seven somatic cell nuclear transfer embryonic stem (SCNT-ES) cell lines derived from MT neurons harvested from adult mouse brains. Several of these SCNT-ES cell lines produced fertile adult mice without obvious neurologic or other abnormalities. These findings demonstrate for the first time, that the cytoplasm of the egg can reverse the epigenetic changes that arise in adult post-mitotic neurons and show that a subset of these neurons preserves sufficient genomic plasticity and integrity to produce fertile adult mice.

In parallel, we performed whole genome sequencing on each of the neuron-derived SCNT-ES cell lines (called MCNT-ES cell lines) along with non-neural tissue from the same donor animals. These analyses reveal that each neuronal genome is unique, harboring on average ~110 *de novo* somatic mutations of all classes. However, we did not find evidence for recurrent or programmed DNA rearrangements, somatic hypermutation or excessive mobile element transposition. Surprisingly, we also find that neuronal SNVs are enriched in

genes that are linked to neuronal function and which are highly expressed in MT neurons, which suggests that somatic mutation in neurons differs from other somatic cell types.

5.2 Results

5.2.1 Adult post-mitotic neurons retain developmental pluripotency

In order to establish the epigenetic plasticity and mutational spectra of mature post-mitotic neurons, we wished to reprogram a neuronal subtype that exits the cell cycle early in development, and which is neither lost nor replaced during the lifetime of the animal. In addition, because previous studies have suffered from a lack of specificity in the genetic marking strategies, we aimed to identify a neuronal population for which we had a tightly regulated Cre-loxP system to definitively establish the identity of the donor nucleus.

Accordingly, for this study we elected to reprogram the mitral and tufted neuronal subtype of the olfactory bulb (MT neurons). MT neurons of the olfactory bulb are among the earliest born neurons in the brain. The majority of MT neurons are generated in the early embryo between embryonic days 9 and 13 (Imamura et al., 2011), and no MT neurons are produced postnatally (Blanchart et al., 2006; Hinds, 1968a, b). MT neurons are active and functional at birth and they exhibit spontaneous and evoked electrical activity throughout the lifetime of an animal (Mair and Gesteland, 1982). In addition, recent single neuron tracing studies performed at Scripps and other laboratories have predicted that MT neurons employ stochastic mechanisms to produce their complex patterns of axonal branching and synaptic connectivity (Ghosh et al., 2011; Miyamichi et al., 2011; Sosulski et al., 2011). Other cell lineages that exhibit stochastic diversification, such as B and T cells, use genomic rearrangements to generate diversity, raising the question of whether genomic changes might be incurred during MT development or maturation. Thus, MT neurons are representative of the morphological and functional complexity present in many other types of CNS neurons.

SCNT typically produces embryonic stem (SCNT-ES) cell lines with a frequency of 1-10% when applied to somatic lineages, making it difficult to exclude the possibility of reprogramming a rare non-neuronal cell type from mixed cell populations (Ogura et al., 2013). Therefore we devised an irreversible genetic marking strategy to label MT neurons (**Figure 5.1a**). In the olfactory bulb, the Pcdh21 gene is expressed strongly and exclusively in MT neurons (Boland et al., 2009; Nagai et al., 2005). We generated a knock-in mouse line in which Cre recombinase is co-expressed with the Pcdh21 gene (Pcdh21/Cre). By crossing the Pcdh21/Cre line to the Ai9 Cre-reporter mouse line (Madisen et al., 2010), we produced the Pcdh21/Cre-Ai9 mouse line. In the olfactory bulbs of these mice, Cre expression in MT neurons excises a stop cassette and activates constitutive expression of the red fluorescent protein, tdTomato (**Figure 5.1a**). Therefore SCNT derived blastocysts, SCNT-ES cell lines and mice cloned from neurons will exhibit uniform red fluorescence and carry a small genomic deletion. This line and similar genetic marking strategies are described in previous publications (Boland et al., 2009; Ghosh et al., 2011).

To establish the specificity of tdTomato expression we performed immunohistochemistry on brain sections from mice of the same ages as those that served as donors for cloning experiments. As expected, in tissue sections of the Pcdh21/Cre-Ai9 olfactory bulb, tdTomato is present in the mitral and tufted cell layers (**Figure 5.1b)** and overlaps with a marker of MT neurons, Tbr2 (**Figure 5.1c**)**.** We also performed immunostaining for markers of astrocytes (S100b), oligodendrocytes (Olig2), microglia (Iba1), and cell division (Ki67) (**Figure 5.1d-g**). For each marker, we counted over 1,000 tdTomato positive cells and did

not observe any co-labeling with markers of glia or cell division (**Figure 5.1h-k**). In the tissue sections, we detected sparse tdTomato positive cells in the granule cell layer of the olfactory bulb (<0.1% of tdTomato positive cells), which likely results from rare ectopic recombination in granule cells. However, given that the efficiency of generating SCNT-ES lines is 1-10%, this minor population is highly unlikely to be the source of more than one SCNT-ES cell line.

To determine whether the genomes of mature post-mitotic MT neurons are sufficiently intact to re-enter the cell cycle and direct pre-implantation development, we performed SCNT using dissociated tdTomato positive MT neurons from adult mice (aged 3 weeks to 6 months). We harvested MT neurons from adult olfactory bulbs using an optimized method that preserves neuronal viability and morphology (**Figure 5.2a**). To improve cloning efficiency we included the histone deacetylase inhibitor Trichostatin A (Kishigami et al., 2006a). We performed 624 nuclear transfers, resulting in 35 morula/blastocyst stage embryos and seven SCNT-ES cell lines (**Figure 5.2b-c, Table 5.1**). These MCNT-ES cell lines morphologically resemble ES cells and express appropriate markers of pluripotency based on immunostaining (**Figure 5.3**). The efficiency of reprogramming MT neurons (~1%) is similar to the efficiencies reported for other terminally differentiated cell types such as tail tip fibroblasts (Wakayama et al., 2005), olfactory sensory neurons (Eggan et al., 2004), and B and T lymphocytes (Hochedlinger and Jaenisch, 2002). This suggests that mature neurons, which have exited the cell cycle and maintained their genomes without cell division, retain epigenetic plasticity that is similar to that of other differentiated cell types.

To establish the developmental potential of these lines we performed tetraploid embryo complementation (TEC) assays. In TEC only fully pluripotent cells can produce a viable mouse (Nagy et al., 1990; Nagy et al., 1993). Three MCNT-ES cell lines exhibited full pluripotency based on the production of fertile adult mice, while another line produced full term pups that died shortly after birth (**Figure 5.2d-f, Table 5.2**). The remaining two lines produced embryos that died in early or mid-gestation. To confirm that these MCNT-mice were derived entirely from the injected MCNT-ES cell lines, we performed immunofluorescence analyses of multiple tissues and showed that they were uniformly tdTomato positive (**Figure 5.2g**). In addition, to rule out minor contribution of tetraploid blastocyst cells we used PCR assays for differences in microsatellite DNA lengths that vary between the MCNT-ES cell lines and the tetraploid host cells, as described previously (Boland et al., 2009) (**Figure 5.2h, Figure 5.4**)**.** These data demonstrate that MT neurons that have been post-mitotic and undergone terminal differentiation, synaptic refinement, and persistent activity for up to 4.5 months can maintain sufficient epigenetic plasticity to produce all tissues required for survival to adulthood and reproduction.

5.2.2 Identifying mutations in MT neurons

Next, to generate a high-resolution picture of the somatic mutations occurring in neuronal genomes, we performed whole genome sequencing (WGS) on MCNT-ES cell lines and control tissue (thymus or spleen) from each donor animal (**Figure 5.5a**). WGS was performed using Illumina paired-end sequencing to a mean coverage of 32X-59X (**Table 5.3**). To identify mutations, we performed sensitive read alignment with Novoalign (Hercus, 2009) and YAHA (Faust and Hall, 2012), and employed a suite of variant detection pipelines. We used GATK (DePristo et al., 2011; McKenna et al., 2010) to detect SNVs and indels and assign them to high and low confidence categories, followed by filtering to identify somatic mutations using a similar strategy as in Kong et al. (Kong et al., 2012). We mapped structural variation (SV) using a sensitive pipeline which utilizes multiple tools including

Novoalign for paired-end mapping, YAHA for split-read mapping of unmapped and clipped reads, and LUMPY (Layer et al., 2014) to integrate read-pair and split-read analysis to detect SV breakpoints at high resolution (**Figure 5.6**); a read-depth analysis pipeline that detects relatively large (>15Kbp) CNVs (Malhotra et al., 2013; Quinlan et al., 2010; Quinlan and Hall, 2012); and an improved custom MEI detection pipeline that maps SINE, LINE and LTR insertions using an approach similar to Lee et al. (Lee et al., 2012a) (**Figure 5.7, Appendix 1, Table S1**). For each class, we estimated variant detection sensitivity via comparison to the Mouse Genomes Project (Keane et al., 2011; Yalcin et al., 2012) (**Appendix 1, Table S2).**

MCNT-ES cell genomes may contain variation from several sources: "germline variants" that differ among donor mice due to mouse strain variation or recent *de novo* germline mutations; "culture mutations" incurred during reprogramming or expansion; and true "somatic mutations" that arose in neurons during development or aging. Germline variants were excluded by comparing MCNT-ES cell lines to donor mouse control tissues (thymus or spleen) and to a database of mouse strain polymorphisms (Keane et al., 2011). We note that thymus and spleen are strong controls because the majority of their cells derive from the endodermal and mesodermal embryonic germ layers, while MT neurons derive from ectoderm. Therefore, variants shared between MT neurons and thymus/spleen represent germline variants or potentially somatic mutations that arose prior to gastrulation. For each variant class, we estimated detection accuracy by validating a subset of calls by PCR and capillary sequencing (**Table S3).**

To distinguish somatic mutations from culture associated mutations, we reasoned that somatic mutations should be heterozygous, present in 100% of cells in an MCNT-ES line, and exhibit a variant allele frequency (VAF) of ~50%, while culture associated mutations will be mosaic in the line resulting in lower VAFs. Therefore for SNVs and indels, we used alignment-based VAF estimates of >30% to define candidate somatic mutations (**Figure 5.8**). For SVs, whose VAFs are difficult to estimate directly from sequencing data, we generated single cell subclones from each line and assessed whether SVs were present in all subclones, indicating a somatic origin. SVs present in only subsets of subclones were deemed culture associated and eliminated from further analyses (**Figure 5.5b**). The only class of culture associated mutation that we could not exclude are mutations that were acquired on both strands of the neuronal genome prior to the first S-phase following nuclear transfer (**Figure 5.9**), which are expected to be exceedingly rare (Li et al., 2014; Ma et al., 2014).

5.2.3 The landscape of genome variation in MT neurons

These analyses identified 87 (68-139) somatic mutations per genome, comprising 69 (50-112) SNVs, 17 (9-24) indels, 1.5 (0-3) SVs, 0.7 (0-2) MEIs (**Tables S4-7**). Taking variant detection sensitivity into account, these data predict a true mutational burden of 112 (89-181) mutations per genome comprising 86 (62-142) SNVs, 23 (12-34) indels, 1.7 (0-4) SVs, and 1.3 (0-4) MEIs (**Table 5.4, Table S8**). While every neuron harbors a unique and significant number of mutations, the mutational load per neuron can vary considerably (**Figure 5.5c-d**). For example, two neurons from 3-week-old mice differed in their mutational burden by a factor of ~2 (89 vs. 181), despite having nearly identical sequencing depth and variant detection sensitivity (**Table 5.4)**. These data are consistent with the extreme variability in mutational burden previously observed for large-scale CNVs (>5Mbp) using single neuron genome sequencing (Cai et al., 2014; Gole et al.; McConnell et al., 2013). In addition, we detected no aneuploidy, which given our sample size is also consistent with

the relatively low ~5-10% prevalence of aneuploidy found in these studies in post-natal neurons. However, in contrast to these studies, we did not observe any highly aberrant neuronal genomes marked by multiple large CNVs. It is not clear what accounts for this discrepancy. Estimates for the prevalence of large scale CNVs found in neurons by the previous SCS studies range widely from 16% to over 60%. In addition, these SCS studies have a hard time validating their mutational calls, and might be subject to higher than expected false positive rates. However, it is also possible that our experimental design that obtains samples from SCNT-based cloning may preclude isolation of MCNT-ES lines from neurons harboring numerous large-scale mutations, thereby introducing an ascertainment bias. Further work will be required to address this question.

To gain insight into the mutational processes that shape neuronal genomes, we assessed their SNV base conversion profiles. The SNV spectrum in neurons closely resembles those described in induced pluripotent stem cell lines derived from mouse fibroblasts (Young et al., 2012), and broadly resembles those reported for human blood stem cell progenitors (Welch et al., 2012), but differs from the human germline (Kong et al., 2012) and clonal organoids formed from mouse endodermal cell types (stomach, intestine and prostate) (Behjati et al., 2014) (**Figure 5.10a**). These results suggest the genomes of different cell types may be shaped by different molecular processes or by differential exposure to mutagens. We examined the prevalence of loss of heterozygosity as an indication of mitotic crossover events by looking for unexpected regions of homozygous SNPs, but found no evidence of any such events anywhere in MT neuron genomes. We did detect three clusters of 2-3 SNVs; two were multiple nucleotide polymorphisms affecting adjacent nucleotides, and may have been caused by error prone polymerases used during base excision repair (Di Noia and Neuberger, 2007), while the third impacted three nucleotides spanning 249bp in the Atxn7l1 gene, suggesting a mutational event similar to kataegis (Nik-Zainal et al., 2012).

Interestingly, we found a significant enrichment of C→T conversions that appear in a TpCpN context; ~44% for somatic mutations versus ~25% for germline SNPs ($p<0.0001$, Fisher's Exact) (**Figure 5.10b**). To our knowledge, the only mutational process that favors this sequence context is cytosine deamination by members of the APOBEC family (Beale et al., 2004; Lawrence et al.). Curiously, we also found that C→T conversions preferentially occur on the transcribed strand in genes highly expressed in MT neurons, and on the untranscribed strand in genes with no or low expression in MT neurons ($p=0.0046$, Fisher's Exact). The reason for this is not clear. Transcription coupled repair (TCR) and C→T conversion by APOBEC deaminases often result in a strand bias in expressed genes (Alexandrov et al.; Pleasance et al., 2010a). But the strand bias in unexpressed genes is hard to explain unless it occurred during a different gene expression environment before neuronal differentiation.

Structural variants are of special interest in neurons due to their potential to cause large phenotypic effects, and because several lines of evidence have suggested that neurons may be especially prone to double-strand breaks (Frank et al., 1998; Gao et al., 1998; Suberbielle et al., 2013). We identified nine SVs among the six neuronal genomes, with a range of 0-3 per cell (**Figure 5.5d**). Remarkably, three of the nine SVs that we identified in neurons were complex rearrangements involving multiple breakpoints. One is a 21Kbp deletion with an additional 17bp inversion at the breakpoint junction, and a second comprises two adjacent deletions (1.7Kbp and 1.2Kbp) affecting the Pkd212 gene, one of which deletes an entire exon (**Figure 5.11b-c)**. The third event is a remarkably complex rearrangement (**Figure 5.11a**) that resulted in the non-duplicative transposition of a 7.3Kbp segment to a location

10.4Mbp downstream, inversion of a 1.4Mbp segment (disrupting the Aven gene), and deletions of 3.7Kbp, 78bp, 41bp, 16bp and 10bp. Strikingly, each of the small deletions arose at the junction of a larger rearrangement, and none of the breakpoints show more than 1bp of microhomology. Thus, this variant was likely caused by simultaneous formation of 6-10 double strand breaks within a 10.4Mbp region, and is best explained by a mechanism similar to chromothripsis involving DNA breakage and error prone NHEJ (Quinlan and Hall, 2012). These data represent the first observation of a chromothripsis-like complex genomic rearrangement in non-cancer somatic cell genome, and suggest that complex rearrangements may play an unanticipated role in neuronal genome diversification.

5.2.4 Neuronal genomes contain few *de novo* mobile element insertions

The extent to which somatically acquired mobile element insertions contribute to neuronal genome diversity is a major unresolved question in the field. Our whole genome sequencing data provides a unique opportunity to measure the MEI landscape in single neurons at high sensitivity and accuracy. In total, we predicted five MEIs, of which four were validated by PCR. Individual neurons carried 0-2 MEIs (**Figure 5.5d**). Our most conservative estimate for MEI detection sensitivity is 52% (**Table S2**), which predicts an average of at most 1.3 new MEI insertions per neuronal genome. Thus, our results are most consistent with recent single cell sequencing experiments (Evrony et al., 2012; Evrony et al., 2015), and suggest that most MT neurons have a relatively low MEI burden.

5.2.5 Each MT neuron has a unique genome without recurrent genomic changes

Mutations can be private to a single neuron or can be shared with other neurons from the same individual due to early arising mutations, mutational hotspots, or programmed rearrangements (Evrony et al., 2015). Here, we have analyzed three MCNT-ES lines from one donor mouse (B) and two from another (C), although we note one of the latter datasets was excluded from other analyses due to culture-derived aneuploidy and a known population bottleneck. None of the mutations we detected were shared, suggesting they arose late in development, possibly even after neuronal differentiation. Although the sample size is limited, these results are consistent with two prior single cell studies in which the vast majority (>99%) of mutations were detected in a single cell (Cai et al., 2014; McConnell et al., 2013), and with the observation that only 0.54% of mutations discovered in mouse organoid cell lines arose during early development (Behjati et al., 2014).

One longstanding hypothesis is that neurons may exploit programmed DNA rearrangements such as those seen in the immune system to generate diversity in gene expression. None of our high confidence somatic mutations were shared by any subset of neurons, and attempts to validate 17 low confidence shared SV and MEI calls were unsuccessful (**Table S3**). Further, none of the SVs we detected in individual MCNT-ES cell lines bear hallmarks of programmed rearrangement, such as joining of alternative exons, or the generation of novel open reading frames. We also visually inspected WGS data covering the protocadherin gene clusters, which have been proposed as candidate loci for programmed rearrangements in neuronal genomes (Yagi, 2003), and found no evidence of rearrangements. Therefore, 100% of the validated somatic mutations identified in this study were restricted to a single neuron. Taken together, these results strongly argue that MT neurons do not require DNA rearrangements at defined loci for their function or maturation. However, they do not exclude the possibility that other neuronal subtypes may exploit this strategy.

5.2.6 Functional consequences of somatic variation in neurons

We discovered 10 somatic mutations that alter the coding sequence of known genes (**Table 5.5**). These include missense mutations in Cdc40, Tas2r113, Klf16, Dhx37, and Tekt5, a single codon deletion in Gpr44, an exon deletion in Pkd2l2 (**Figure 5.11b**), an exon duplication in Atp10b, a deletion encompassing the Zic1 and Zic4 genes, and disruption of the Aven gene (**Figure 5.11a**). These results demonstrate for the first time that individual neuronal genomes often carry one or more newly mutated genes. We next compared the distribution of the 395 high confidence autosomal somatic SNVs to various genome annotations. Intriguingly, while somatic mutations are distributed randomly with respect to most genomic features (**Table S9**), they show a significant enrichment in evolutionarily conserved elements (1.6-fold, p=0.01 by Monte Carlo simulation, **Figure 5.10c**). However, our power to discern subtle effects is limited by the relatively small number of mutations discovered in this study.

To determine whether MT neurons exhibit mutational distributions that could depend on their cell type or post mitotic status, we compared our data to a recent study that identified somatic mutations arising in individual cells from mouse endodermal cell types (Behjati et al., 2014). This analysis reveals that neuronal SNVs are 1.22 fold more prevalent in genic regions than SNVs from endodermal cell types (p=0.0039, Fisher's Exact, **Figure 5.10d**). This suggests that neurons differ from stomach, intestine, and prostate cells in the means by which they acquire or repair DNA damage.

Next, we assessed whether MT neuronal mutations were enriched in genes that could impact neuronal function. DAVID analysis (Huang da et al., 2009b), which assays for functional enrichment within lists of genes, showed that somatic mutations in neurons are often found in genes with neuronal function, with the top six and 19% of all enriched gene ontology (GO) terms related to neurobiology (**Figure 5.10e, Table S10**). In contrast, DAVID analysis of intestine-derived somatic mutations identified zero neuron-related enriched GO terms among the top ten, the top neuron-related term occurring 14th on the list, and only 11% of all enriched GO terms related to neurobiology (**Figure 5.10e, Table S11**).

Finally, to more directly assess the likelihood that somatic mutations impact MT neuronal function, we performed RNA-Seq on flow sorted MT neurons from the Pcdh21-Cre/Ai9 mouse strain. We find that MT neuron-derived SNVs are enriched in genes that are highly expressed (top 50%) in MT neurons compared to endodermal cell SNVs (Behjati et al., 2014) (p=0.025 by Fisher's Exact Test, **Figure 5.10f**). To ask how this compares to mutations in endodermal cell types, we used a recently published RNA-Seq data set for Lrg5 expressing small intestine stem cells (Sheaffer et al., 2014), the same stem cells used to generate small intestine derived organoids sequenced by Behjati et al. (Behjati et al., 2014) In contrast to MT neuron derived SNVs, mutations from the small intestine are depleted in genes highly expressed in small intestine compared to MT neuron SNVs (Behjati et al., 2014) ($p=7.06 \times 10^{-4}$ by Fisher's Exact Test, **Figure 5.10g**).

Here we provide several independent lines of evidence suggesting that mutations in MT neurons may preferentially accumulate in functionally relevant genomic regions. These include evolutionarily conserved elements, genic regions, genes with neuronal function and, intriguingly, genes that are highly expressed in MT neurons themselves. This is surprising given that cancer genomes, and indeed small intestinal stem cells (**Figure 5.10g**), exhibit a relative depletion of mutations in expressed genes, which in cancer genomes has been attributed to transcription coupled repair (Alexandrov et al., 2013; Lawrence et al., 2013).

One possible explanation for the bias we detect is that neurons enact less efficient transcription coupled repair than other cell types. Alternatively, neurons could have an increased mutation rate at actively transcribed loci, either due to the transcription process itself, increased reactive oxygen species exposure resulting from the high metabolic rate of neurons, or to chromatin alterations that accompany gene activation. L1 insertions in neurons have also been observed to occur preferentially in neuronally active gene transcripts further suggesting increased mutation rates in euchromatin (Baillie et al.; Upton et al.). Future work will be necessary to resolve this question.

5.3 Discussion

Neurons are diverse and irreversibly post-mitotic. As such, information on neuronal genomes has been limited to studies of bulk samples of mixed cell types or to relatively low resolution single cell studies. Here, we amplified genomes of single neurons using SCNT and applied whole genome sequencing to assess their genomic plasticity and mutational burden. By cloning mice from adult post-mitotic neurons we showed that at least half of the neurons amenable to reprogramming also maintain sufficient genomic plasticity and integrity to serve as a template for all tissues required to generate a fertile adult mouse. However, despite their developmental potency, neurons exhibit striking levels of genomic individuality. On average, MT neurons harbor more than 100 unique mutations of multiple classes. Of these, relatively few are SVs and MEIs (0-3), but several SVs exhibit unusual complexity and impact genes expressed in MT neurons. The neuronal burden of the most abundant mutational class, SNVs (~86), is lower than mouse fibroblasts (190-698) (Young et al., 2012) and endodermal cell types (274-916) (Behjati et al., 2014). However, it is more similar to male germline cells (~55 per haploid genome) (Kong et al., 2012), than to oocytes (~14 per haploid genome) (Kong et al., 2012). This result is somewhat surprising given that neurons and oocytes exit mitosis at similar times in embryonic development and are maintained without cell division for many years, while male germline cells divide throughout the lifetime of an individual. Yet there is some evidence that oocytes can harbor more mutations than previously thought (Conrad et al., 2011).

Closer inspection of these SNVs reveals another difference between neuronal mutations and those of other cell types. Neuronal SNVs appear to be biased towards genomic regions likely to have functional significance to MT neurons as compared to SNVs derived from endodermal cells. However, the different sequencing approaches and bioinformatic methods employed by various studies may account for some of these differences between cell types. Future studies with larger sample sizes and additional cell types will more precisely establish the magnitude of our initial finding and whether the bias toward expressed genes we observe in MT neurons is unique, or a feature shared with other neuronal subtypes or post-mitotic cells in general. Nevertheless, these preliminary findings underscore the importance of examining somatic mutation with single cell resolution in different cell types and raise intriguing questions regarding the source and impact of neuronal genome diversity.

A key question regarding neuronal mutations is when do they arise? One possibility is that most mutations arise early in development, in the dividing precursors of the neurons. While the precise number of cell divisions that precede MT neurons has not been reported, MT neuron production peaks around embryonic day 11.5 (e11.5) (Imamura et al., 2011), a time at which MT precursors would be predicted to have undergone ~22 cell divisions based on embryologic studies (Imamura et al., 2011). However, we required that putative neuronal mutations be absent from thymus or spleen, which segregate from ectoderm at gastrulation

(e6.5) or slightly before (e4.5) (**Figure 5.12**). A conservative estimate of the number of divisions between e4.5 and e11.5 is 14 (~2 per day). For the range of SNVs we detect ~86 (62-142), these calculations predict a mutation rate of ~6 (4.4-10) SNVs per somatic cell division, somewhat higher than 1.1 SNVs per somatic division reported for mouse organoid cell lines (Behjati et al., 2014), a mean of 2.3 SNVs per somatic division reported by several prior single gene studies in human (Lynch, 2010), and ~3 to ~5 per somatic cell division reported for hematopoietic stem cells in humans (Holstege et al., 2014; Welch et al., 2012).

However, neurons spend the vast majority of their "lifespan" in a non-dividing state. Therefore, it is tempting to speculate that this apparently high per division mutation rate and the discrepancy between oocyte and MT neuron mutational burden noted above instead reflects post-mitotic mutation. The fact that we observe no shared mutations among pairs or trios of neurons from the same animal is also consistent with this model. If we assume that neuronal SNVs arise at the lowest levels reported for other cell types during development (~1 per cell division) (Behjati et al., 2014), MT neurons should harbor ~14 SNVs. Subtracting this from the observed number of SNVs in neurons would result in a maximum average of ~72 (48-128) SNVs that could be attributed to post-mitotic mutation. Dividing the ~72 SNVs by the ages of the donor neuron (~30-190 days) predicts a remarkably high post-mitotic mutation rate of 0.38-2.4 SNVs per day. Continuing this logic, if human neurons parallel mouse neurons and accumulate ~1 SNV mutation per day, a 50-year-old brain would harbor neurons with ~18,000 SNVs. This mutational load is on par with highly mutated cancer genomes, which might impact neuronal function broadly throughout the brain. In addition, if aging associated mutations preferentially accumulate in expressed genes, their potential impact on neuronal function could be significant, leading to altered function or degeneration. Such high post-mitotic mutation rates would surely have functional relevance, particularly in aged individuals.

An alternative explanation for the apparently high somatic SNV burden is an "early burst" model in which many mutations arise near the time of neuronal cell cycle exit and terminal differentiation and then taper off. This model is also consistent with a higher mutation burden in neurons and the lack of shared mutations between neurons from the same mouse. It is additionally supported by the fact that we do not observe a consistent increase of mutational load with neuronal age predicted by the hypothesis of post-mitotic accumulation of mutations. The SNV burst model parallels the observed burst of L1 activity during neurogenesis caused by repression of Sox2 and other factors (Kuwabara et al., 2009; Muotri et al., 2005; Richardson et al., 2014). However it is unclear what would cause such a burst of SNVs, although it has been suggested that mutation prone mechanisms may be involved in the changing of cytosine methylation states observed during neuronal differentiation (Guo et al., 2014; Guo et al., 2011).

How might somatic mutation in neurons impact brain function? Somatic mutations that arise early in development have been found broadly distributed in particular brain regions (Evrony et al., 2012; Evrony et al., 2015) and have been shown to impact brain function in a number of human diseases (Jamuar et al., 2014; Lee et al., 2012b; Poduri et al., 2012; Riviere et al., 2012). Burst-model and post-mitotic mutations, in contrast, would be predicted to impact only one neuron. The relative contribution of these mechanisms may differ by individual or even by neuron, perhaps explaining some of the variability of onset and severity of age-related neurological conditions. At present our sample size is too small to distinguish between the contributions of these various models (**Table 5.4**), and more study is needed particularly with neurons from older mice.

Finally, it is important to note that our study is specifically designed to provide a "best-case scenario" estimate of genomic mosaicism in neurons. Cloning from neurons using SCNT may select against the most highly mutated neurons, leading us to underestimate the true scope of neuronal mutation. Furthermore, MT neurons are exposed to limited environmental stressors relative to sensory and peripheral neurons. Differences in other physiological properties such as the metabolic demands dictated by individual synaptic firing rates, as well as the dynamics of chromatin remodeling established by differences in gene expression, may also influence somatic mutation in a neuronal subtype specific manner. Therefore, we believe these studies underscore the importance of using comprehensive genome-wide approaches to evaluate somatic mutation in neurons of diverse subtypes and ages, as well as in the context of neurodegenerative disease.

Methods Summary:

All methods are detailed in **Appendix 1**.

Table 5.1. Efficiency of SCNT using MT neuron nuclei.
Results of 13 SCNT experiments in which individual MT neuron nuclei were transferred into enucleated oocytes, allowed to develop to the blastocyst stage and then used to produce SCNT-ES cells.

Donor age	Oocytes activated	2-cell embryos (% oocytes activated)	Morula/blastocysts (% oocytes activated)	MCNT-ES cell lines (% oocytes activated)	Independent experiments
3 wks	297	137 (46%)	20 (7%)	3 (1%)	7
4.5-6 mos	327	253 (77%)	15 (5%)	4 (2%)	6

Table 5.2. MCNT-ES cell development in the TEC assay.
Perinatal pups, juvenile, and adult animal are defined as those which survived to post natal day four, weaning, and two months of age respectively. The single pup generated from MCNT-ES cell line D4 that was able to survive to perinatal stages was moribund and euthanized at postnatal day 4. This pup was later determined to have high contribution of tetraploid blastocyst host DNA **(Figure 5.4e)**, and was likely a diploid chimera, which can result from rare failed fusion events during tetraploid blastocyst generation.

MCNT-ES cell line	Age of donor	Tetraploid blastocysts Injected	Alive at term (% injected)	Breathing normally (% injected)	Perinatal pups (% injected)	Juvenile animals (% injected)	Adult animals (% injected)
C1	3 wks	152	15 (10%)	10 (7%)	8 (5%)	8 (5%)	8 (5%)
C5	3 wks	140	8 (6%)	5 (4%)	3 (2%)	2 (1%)	2 (1%)
D4	3 wks	214	15 (7%)	6 (3%)	1 (0.5%)	0	0
B2	4.5 mos	140	32 (23%)	26 (19%)	20 (14%)	19 (14%)	19 (14%)
B3	4.5 mos	140	0	0	0	0	0
B4	4.5 mos	150	0	0	0	0	0

Table 5.3. Cell line sequencing statistics.
Table showing information about the four mice used in this study, along with the tissue source of the control sample, passage number of MCNT-ES cells at the time of Whole Genome Sequencing (WGS), and statistics concerning the WGS runs associated with each sample. The WGS Illumina paired-end sequencing resulted in two paired reads ~100bp in length, encompassing an outer template length of ~474bp. The Median Genome Coverage is a measure of the median number of 100bp reads covering each base in the genome, while the Median Physical Coverage also includes those bases in the insert between the two 100bp reads.

Sample	Source	Passage Number	Median Genome Coverage	Median Template Length	Median Outer Span Physical Coverage
C Mouse: 3 week old female					
C0	Spleen	n/a	32	471	78
C1	SCNT	21	34	464	81
C5	SCNT	7	32	479	81
D Mouse: 3 week old male					
D0	Spleen	n/a	34	481	85
D4	SCNT	7	33	486	85
B Mouse: 4.5 month old male					
B1	Thymus	n/a	38	465	90
B2	SCNT	4	59	477	146
B3	SCNT	4	59	464	143
B4	SCNT	4	58	470	149
E Mouse: 6 month old female					
E0	Thymus	n/a	34	469	83
E1	SCNT	7	36	484	88
Range for all samples					
Min		4	34	464	83
Max		7	59	484	149

Table 5.4. Somatic mutation discovery statistics.
Summarizes mutations predicted by variant detection pipelines (mutation calls), the results of subsequent validation experiments, and an estimate of the true number of mutations present in neuronal genomes. Note that mutation calls and false discovery rate (FDR) are omitted for SV and MEIs because all mutation calls were tested by PCR and definitively determined to be valid or invalid. For SNVs and indels, the false discovery rate (FDR) was estimated by PCR validation of a subset of calls. 69 of 69 SNVs and 22 of 23 indels validated (**Table S3**). Estimated mutation counts take into account the FDR for SNVs and indels, and the false negative rate (FNR) for all mutations (**Table S2-3, Appendix 1**). See also **Tables S5-S8.**

		C5	D4	B2	B3	B4	E1	Mean
SNVs	Mutation Calls	112	50	50	68	70	61	68.5
	%FDR (n = 69)	0.0	0.0	0.0	0.0	0.0	0.0	--
	%FNR	21.4	20.8	19.0	18.9	19.2	22.8	--
	Estimated Mutations	142	63	62	84	87	79	86.2
Indels	Mutation Calls	25	19	16	9	17	18	17.3
	%FDR (n = 23)	4.3	4.3	4.3	4.3	4.3	4.3	--
	%FNR	28.7	25.2	24.2	24.0	24.7	28.5	--
	Estimated Mutations	34	24	20	12	21	24	22.5
SVs	Validated Breakpoints	3	0	7	1	0	3	2.3
	Validated Events	2	0	3	1	0	3	1.5
	%FNR	13.5	12.4	13.1	8.6	8.4	13.4	--
	Estimated Breakpoints	3	0	8	1	0	3	2.5
MEIs	Validated Mutations	1	1	0	2	0	0	0.7
	%FNR	48.2	45.8	47.6	48.3	47.7	47.0	--
	Estimated Mutations	2	2	0	4	0	0	1.3
	Total Estimated Mutations	181	89	90	101	108	106	112.5

Table 5.5. Genes effected by known mutations.
Genomic location and coding changes associated with all known validated somatic mutations in MCNT-ES cell lines, as well as the developmental potential of the associated mouse. Mutations shown in red are in genes expressed in MT neurons. Note there are many other known somatic mutations that fall within non-exonic regions of gene transcripts (**Tables S4-7**). These may also have effects on gene expression.

Type	Chrom	Start	End	Line	Potency	Gene Effect	Gene(s)
SNV	chr10	40577358	40577359	B4	early gest/0	MisSense:Thr->Ala	Cdc40
SNV	chr6	132843921	132843922	D4	term	MisSense:Phe->Leu	Tas2r113
SNV	**chr10**	**80031966**	**80031967**	**D4**	**term**	**MisSense:Gly->Val**	**Klf16**
SNV	chr5	125909673	125909674	E1	nd	MisSense:Arg->Trp	Dhx37
SNV	**chr16**	**10358345**	**10358346**	**C5**	**full**	**MisSense:Phe->Tyr**	**Tekt5**
INDEL	chr19	11015537	11015541	C5	full	CodonDeletion:Leu	Gpr44
CGR	**chr2**	**107930138**	**118338090**	**B2**	**full**	**1Exon(2)Deletion**	**Aven**
CGR	**chr2**	**107930138**	**118338090**	**B2**	**full**	**MultiExonInversion**	**Aven**
Deletion	**chr18**	**34573248**	**34574435**	**C5**	**full**	**1Exon(4)Deletion**	**Pkd2l2**
Duplication	chr11	42981066	42998077	E1	nd	1Exon(4)Duplication	Atp10b
Deletion	chr9	91183064	91411016	E1	nd	2GeneDeletion	Zic1,Zic4

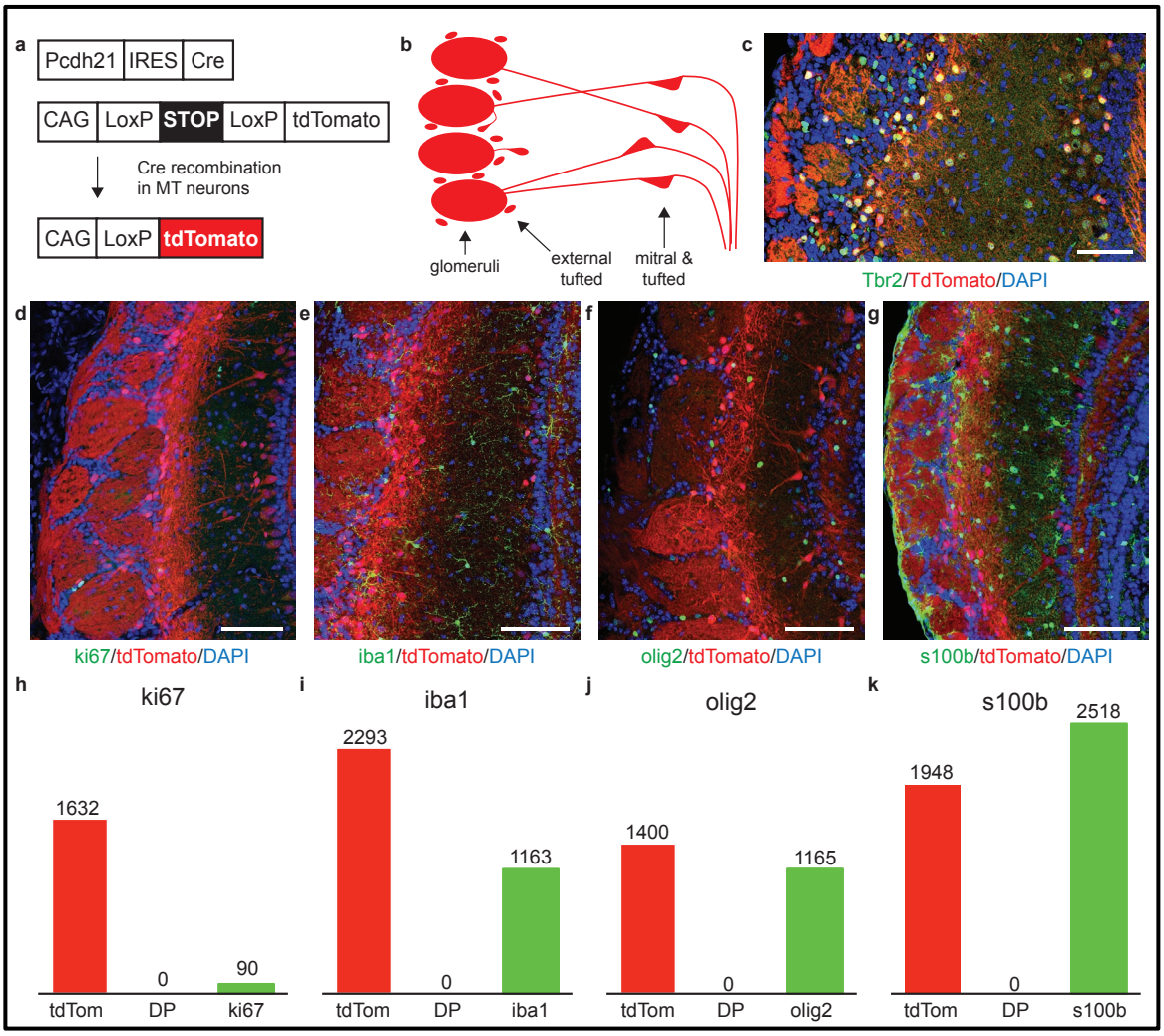

Figure 5.1. Genetic labeling of mitral and tufted (MT) neurons.
a, Donor animals carry one Pcdh21/Cre allele (top) and one copy of the Ai9 Cre reporter transgene (middle). Cre expression in MT neurons excises the STOP cassette within the Ai9 transgene, resulting in specific tdTomato expression and genetic labeling of MT neurons (bottom). **b**, Schematic representation of the MT neuron localization and morphology within the olfactory bulb. Mitral and tufted cells in the mitral and tufted cell layer, as well as external tufted cells send their dendrites into spherical structures known as glomeruli, where they synapse with olfactory sensory neurons. **c-g**, Immunostaining of Pcdh21/Cre-Ai9 mouse olfactory bulb sections for markers of MT neurons, glia and dividing cells. Blue, DAPI nuclear stain; red, endogenous tdTomato fluorescence; green, antibody staining for MT neuron marker Tbr2. **d**, Dividing cell marker Ki67. **e**, Microglia marker Iba1. **f**, Oligodendrocyte marker Olig2. **g**, Astrocyte and olfactory ensheathing cell marker S100b. **h-k**, Quantification of the absence of co-expression of tdTomato with glial and dividing cell markers. DP: double positive for tdTomato and glial/dividing cell maker. Scale bar in c, 15 μ. Scale bars in d-g 100 μ.

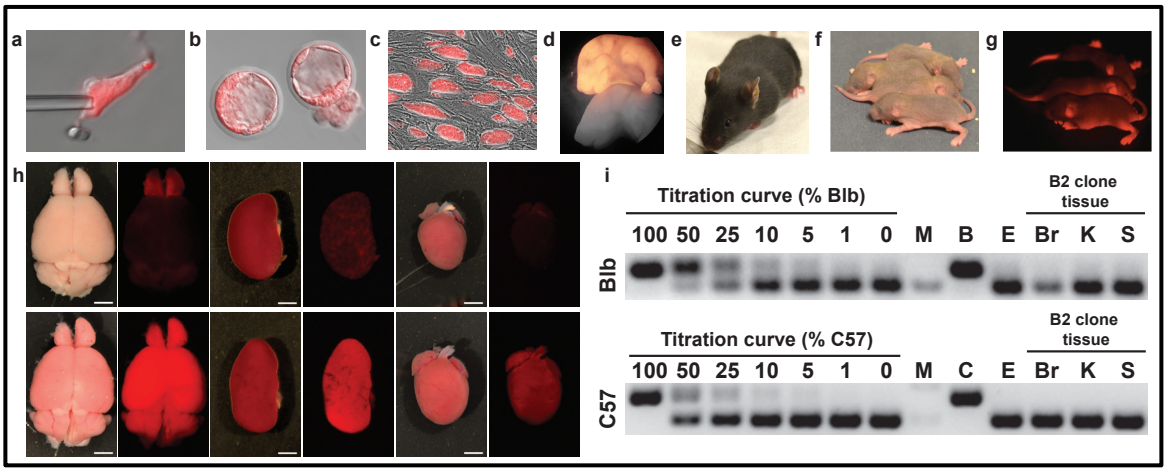

Figure 5.2. ES cells and mice derived from MT neurons.
a, Representative dissociated MT neuron used as nuclear donor in SCNT experiments, shown with SCNT injection pipette. **b,** tdTomato positive blastocysts generated from MT neurons. **c**, tdTomato positive MCNT-ES cells derived from MT neurons. **d**, Newborn and **e**, adult clones generated from MCNT-ES cells. **f,** standard and **g**, fluorescence images of offspring of MCNT-mice. Transmission of the tdTomato transgene demonstrates MCNT-ES cell derived cells are able to differentiated into functional germ cells. **h**, Alternating standard and fluorescent images of brain, kidney, and heart dissected from Pcdh21/Cre-Ai9 control mice (top row) and MCNT-mice (bottom row). Organs from MCNT-mice demonstrate uniform tdTomato expression. **i**, Sample microsatellite PCR assay for tetraploid host blastocyst contribution to MCNT-mice. Band size distinguishes cells derived from MCNT-ES cell line B2 from the tetraploid host strains C57 (C57BL/6J-Tyrc-2J) and Blb (Balb/cByJ). DNA titration curve demonstrates 5% detection limit. Analysis of DNA from B2 clone tissues demonstrates no detectable tetraploid host DNA. See **Figure 5.4** for analysis of mice from other MCNT-ES cell lines. M, molecular weight; E, B2 ES cell DNA; Br, brain; K, kidney; S, spleen.

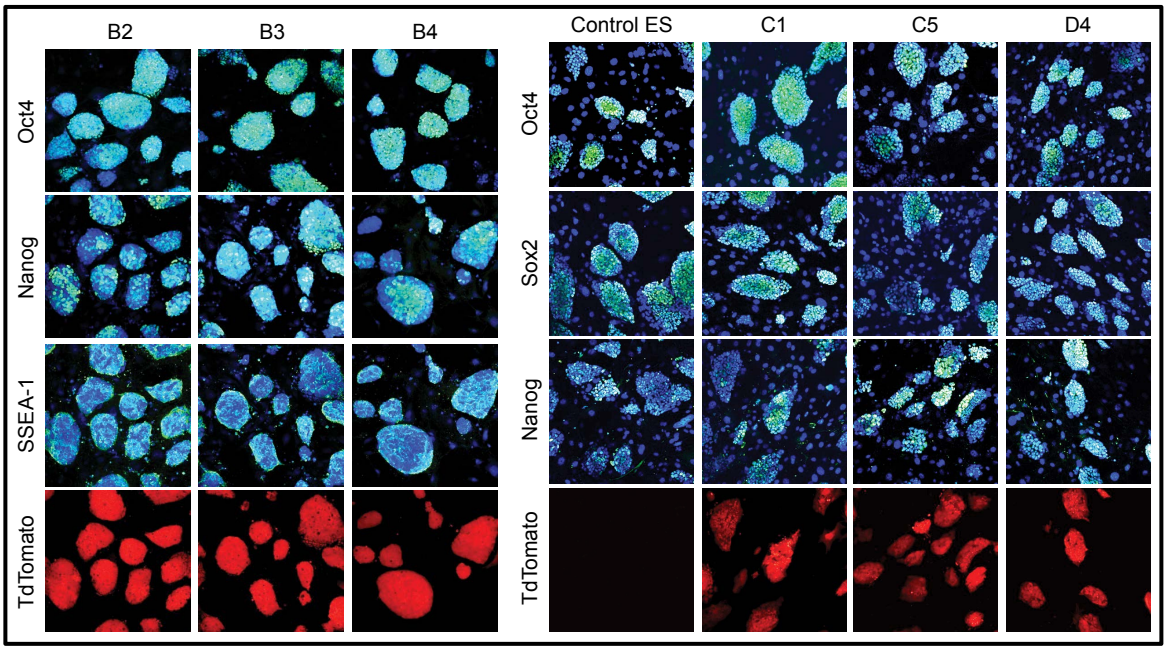

Figure 5.3. **Pluripotency marker staining in MCNT-ES cells**.
MCNT-ES cells display endogenous tdTomato fluorescence (red) and stain positively for the pluripotency genes Oct4, Sox2, Nanog, and SSEA-1 (green). Nuclei are counter stained with DAPI (blue).

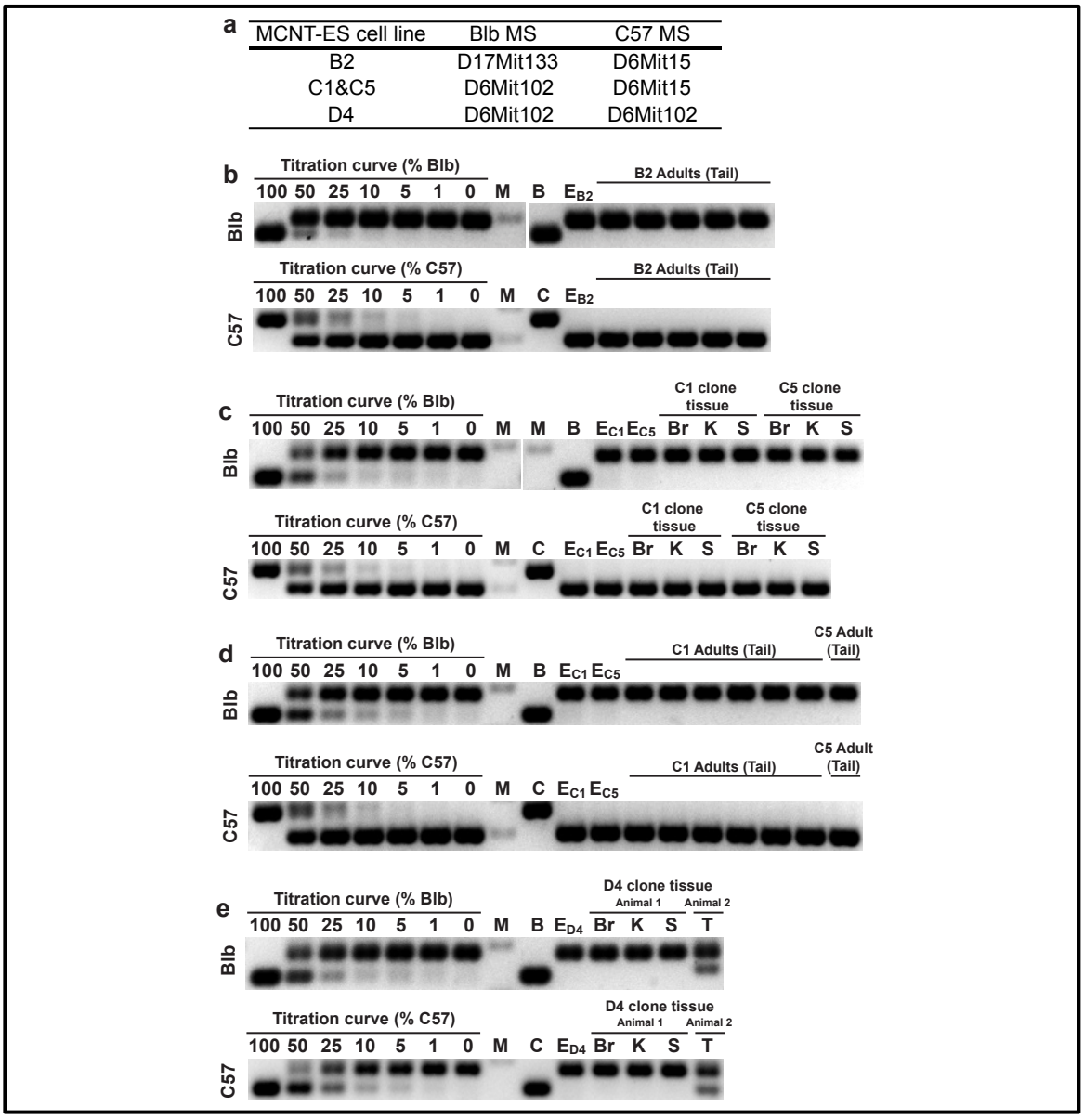

Figure 5.4. **Microsatellite PCR assay**
This assay demonstrating lack of TEC host blastocyst contribution in MCNT-mice. **a**, Diagnostic microsatellites used to distinguish MCNT-ES cell DNA from tetraploid host strains Balb/cByJ (Blb) and C57BL/6J-Tyrc-2J (C57). Primary data for mice derived from MCNT-ES cell lines B2 (**b**), C1 and C5 (**c-d**), and D4 (**e**). For each line, analysis was performed on various organs from a newborn animal and on tails from several different adult animals, with the exception of D4, which only produced a single perinatal animal. This single D4 perinatal animal was the only MCNT-mouse to show detectable tetraploid host strain contribution (**e**), which may explain why it was able to survive longer than littermates displaying no tetraploid host contribution. For all primer pairs, DNA titration curves demonstrate a 5-10% detection limit. M, molecular weight; B, Blb; C, C57; E_{B2}, E_{C1}, E_{C5}, E_{D4}, DNA from B2, C1, C5, and D4 MCNT ES cells respectively; Br, brain; K, kidney; S, spleen; T, tail.

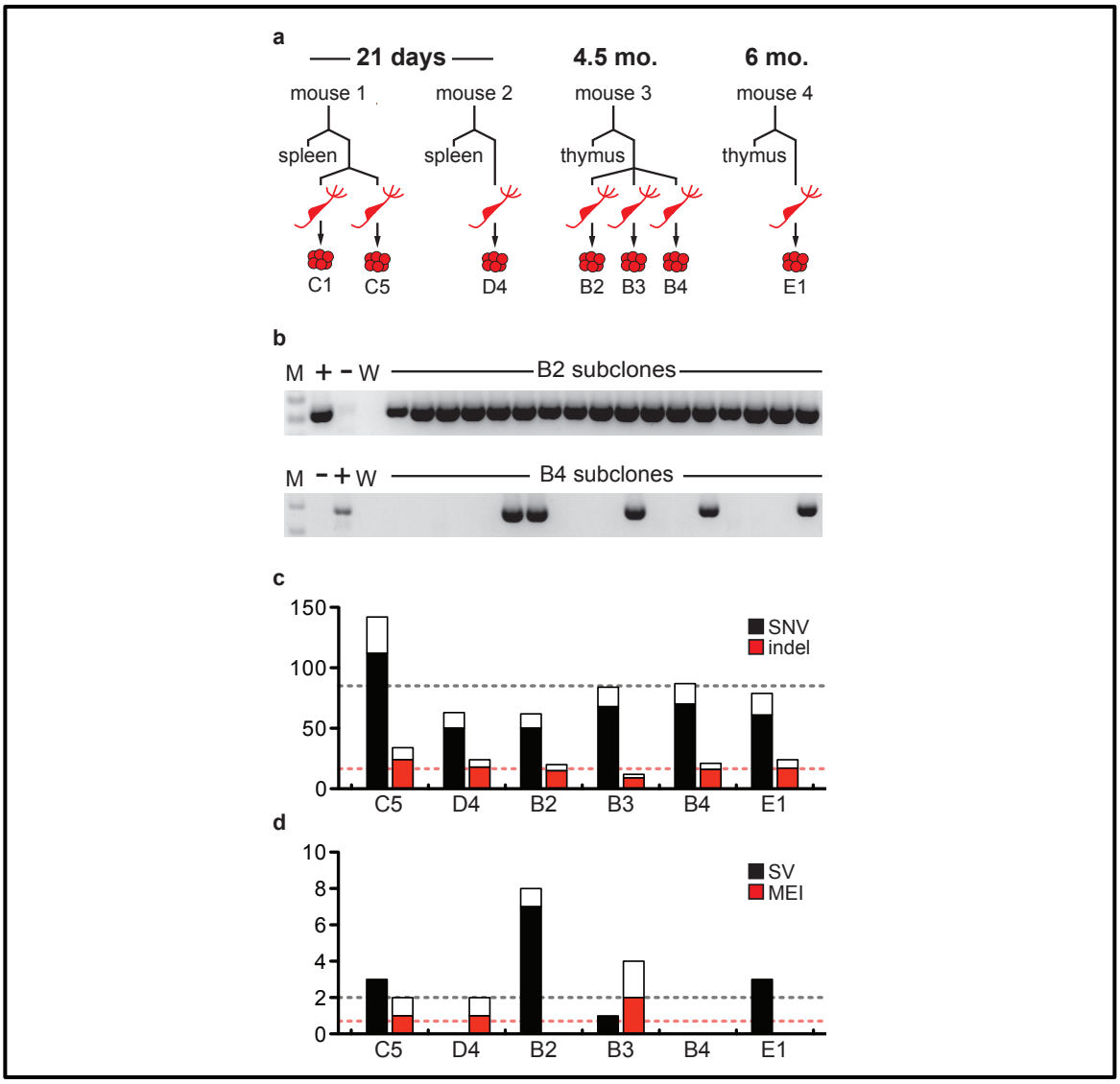

Figure 5.5. Whole genome sequencing of MCNT-ES cells.
a, Schematic overview of Pcdh21/Cre-Ai9 donor animals and the MCNT-ES cell lines and control tissues sequenced from each animal. **b**, Representative PCR subclone validation for two structural variants (SVs). PCR primers flank the SV breakpoint, and are diagnostic for the presence of the SV mutation. One SV is predicted to be somatic by its presence in all early passage subclones (top). The other SV arose during culture or reprogramming, as it is present in only some subclones (bottom). Images are cropped to the region of diagnostic band size. M, molecular weight. +, positive control for diagnostic band, which was B2 (top) or B4 (bottom) MCNT-ES cell DNA of similar passage to DNA used in WGS. –, negative control for diagnostic band, which was thymus DNA from the original Pcdh21/Cre-Ai9 donor animal. **c** and **d**, Observed mutations (black/red bars) and estimated true mutational burden based on the false negative rate (FNR; colored plus white bars). For SVs, the FNR is calculated for breakpoints, rather than for mutational events, which may contain multiple breakpoints. Therefore, observed and predicted values for breakpoints are plotted above.

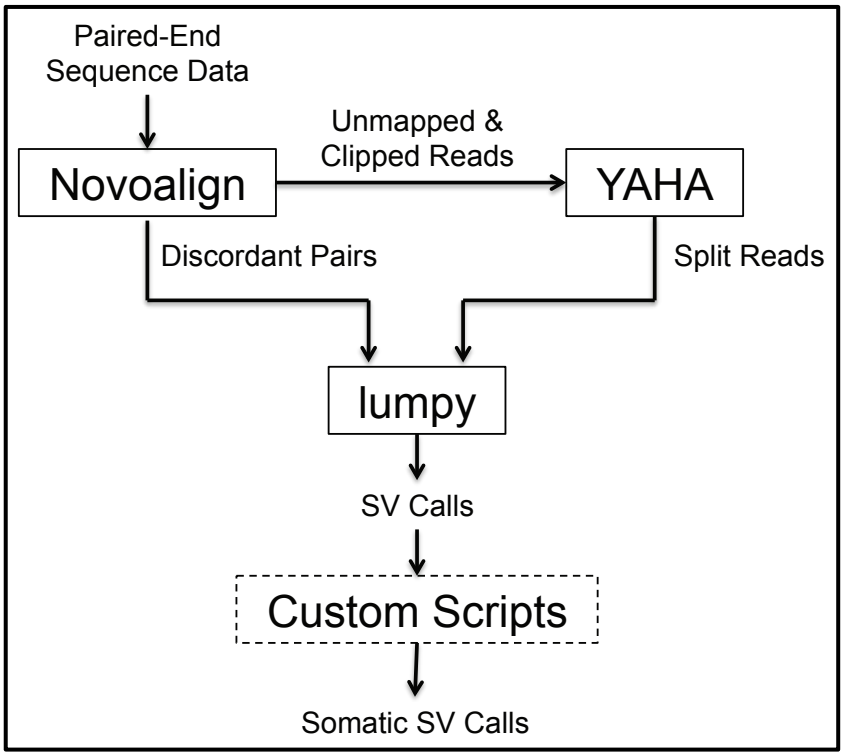

Figure 5.6. Schematic overview of Structural Variant calling pipeline
We use a custom pipeline for SV detection. Novoalign is used for initial paired-end mapping of the sequencing data. YAHA is then used to realign all unmapped and clipped reads to find possible split-read mappings. Discordant read-pairs and split-read mapping are fed to LUMPY to make initial SV calls. Custom scripts then filter those calls to find one that are putative *de novo* somatic calls. See **Appendix 1** for details.

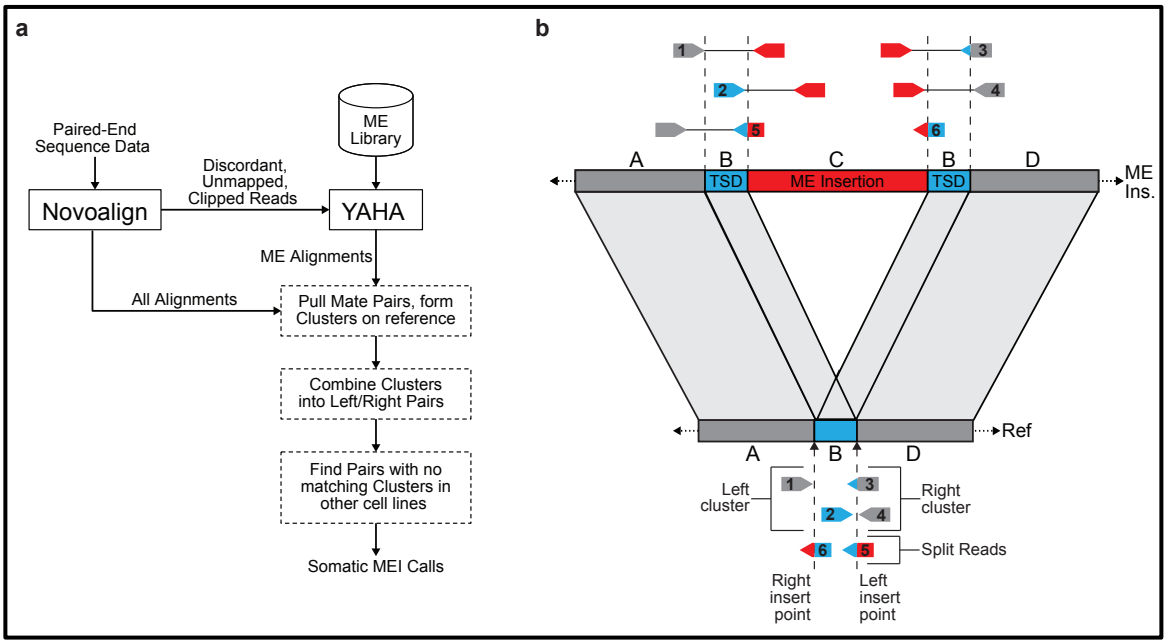

Figure 5.7b adapted from Lee et al. (Lee et al., 2012a).

Figure 5.7. Schematic overview of MEI calling pipeline.
a, Flowchart depicting processing steps in MEI calling pipeline. **b**, Schematic depiction of the structure of a ME insertion event. The ends of paired-end reads that fall within the ME insertion (red) are difficult to map to the reference genome. Therefore, all discordant, unmapped and clipped reads are first aligned to a ME library (**Table S1**). The mates of reads that map well to the ME library (1,2,3 and 4) are clustered by their reference coordinates. Left/right clusters that form properly oriented pairs define a possible MEI event. Further supporting evidence for the call is gathered from split-reads in which one end of the read maps well to the reference adjoining an insertion point, while the other maps well to the ME library, thereby spanning an insertion breakpoint (5 and 6). In addition, we determine if a Target Site Duplication (TSD) has occurred by checking if the right insertion point falls before the left insertion point on the reference. Such a TSD is further confirming evidence for a MEI event. See **Appendix 1** for details.

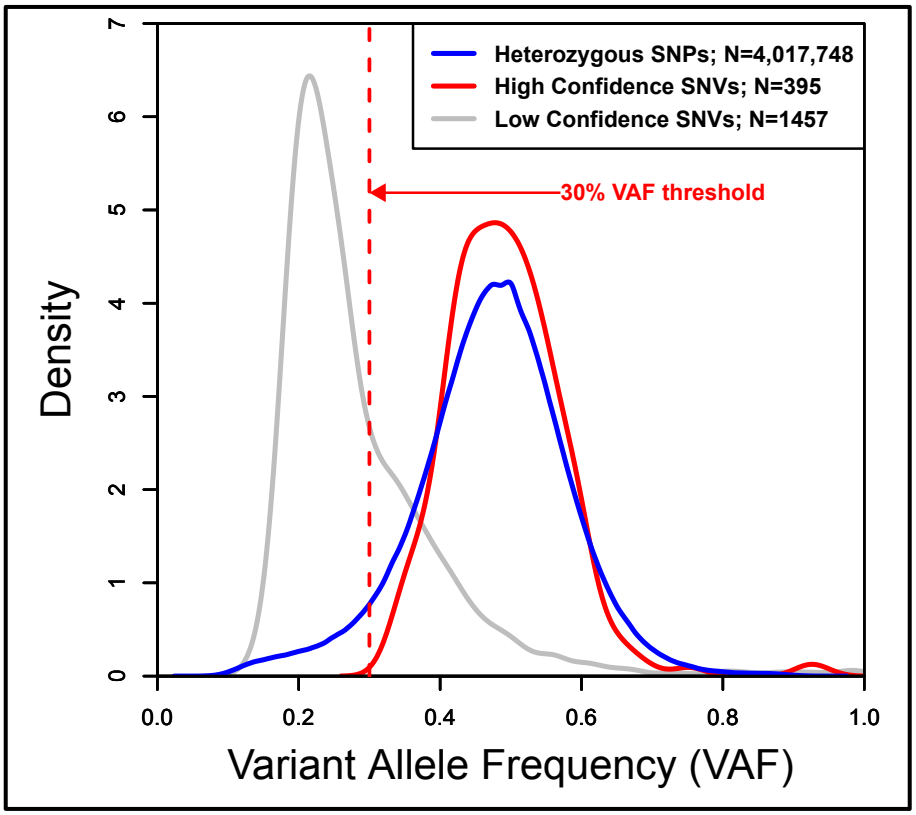

Figure 5.8. VAF distribution of autosomal SNV mutations.
The distribution of Variant Allele Frequency (VAF), defined as the number of reads containing the alternate allele divided by total read depth, for three different categories of single nucleotide mutations. Note that, as expected, heterozygous autosomal SNPs have a VAF distribution that is roughly normal with a mean of 50%. This distribution is very closely matched by the high confidence (HC) SNVs (as defined by GATK) that have an estimated FDR of 0% based on our PCR validation experiments. In contrast, low confidence (LC) SNVs have a much lower mean VAF and the distribution is heavily skewed to the left. This is an indication of possible mutations that arose during clonal expansion, or other contamination, and not from the original neuron used during SCNT. The vertical line at 30% VAF demarcates the threshold we applied to putative SNVs above which they were considered candidate neuronal somatic mutations. This threshold is just over two standard deviations from the SNP and HC SNV 50% mean, and as can be seen from the graph eliminates almost no HC calls, but most of the LC calls.

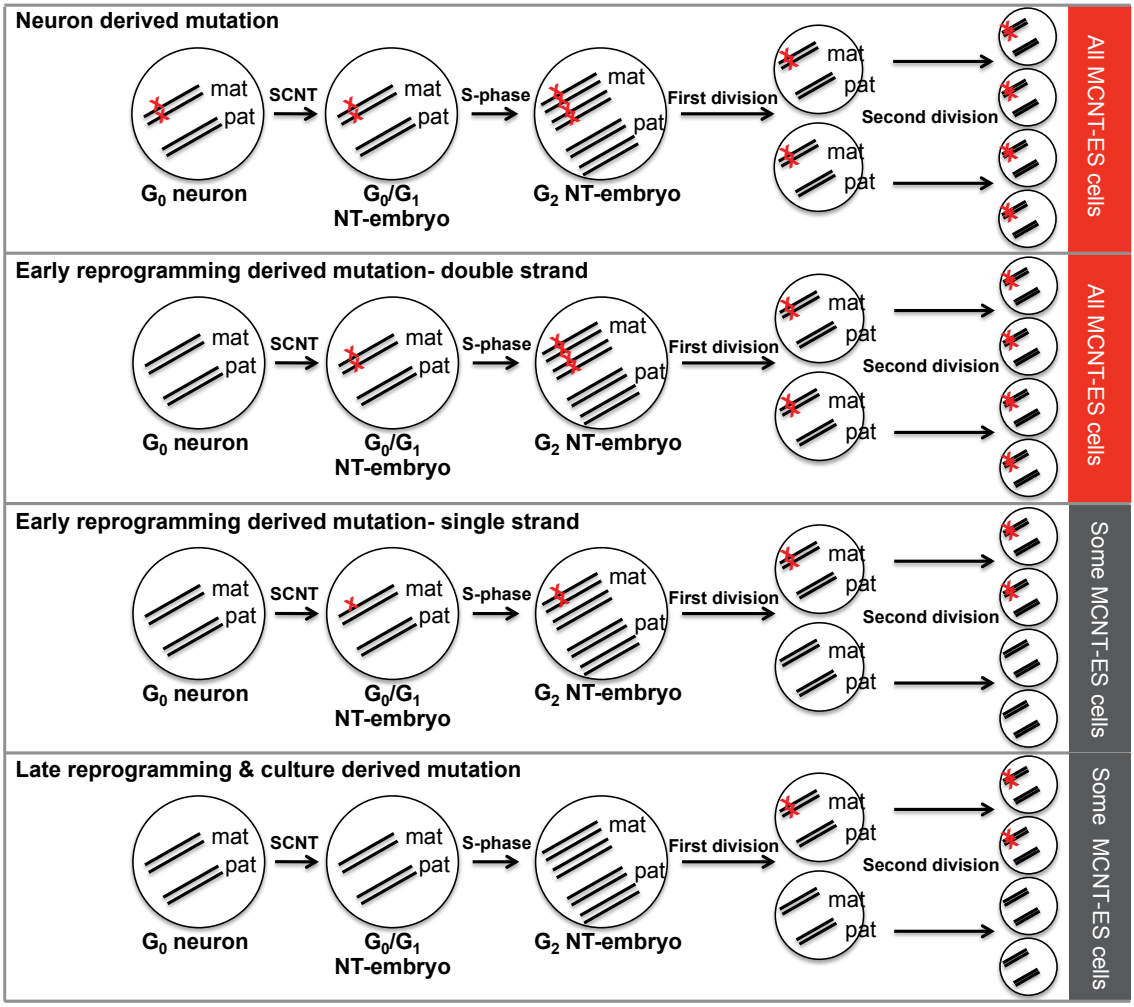

Figure 5.9. Sources of mutations in MCNT-ES cell.
Neuron derived mutations are present in all MCNT-ES cells and one of two homologous chromosomes. As a result, neuron derived mutations have an expected VAF of ~50% and appear in all subclones (top panel). Single-strand mutations occurring during early reprogramming, and all mutations occurring in culture are present in half or fewer cells and in one quarter or fewer of homologous chromosomes (bottom two panels). Therefore, for SNVs and indels, reprogramming and culture derived mutations are eliminated by requiring putative somatic mutations to have a variant allele frequency (VAF) of at least 30% (**Figure 5.8**). For SVs and MEIs reprogramming and culture derived mutations are eliminated by subclone analyses (**Figure 5.5b**). The only non-neuronal mutational category that can pass our calling filters and validation methods are mutations acquired on both strands before the first S-phase following SCNT (second panel). Such mutations are expected to be extremely rare.

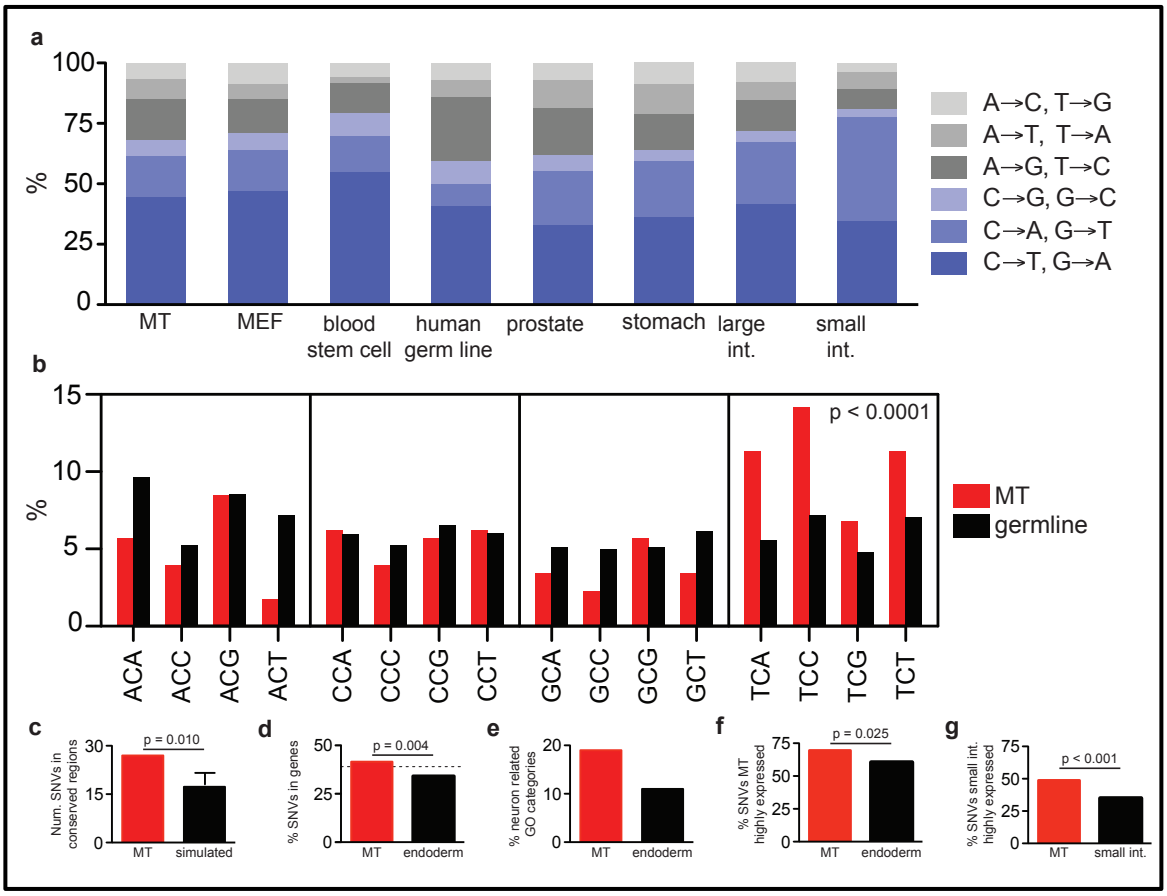

Figure 5.10. Features of SNVs in MT neurons and their genomic enrichment
a Stacked plot of SNV substitutions for MT neurons and other cell types. **b**, Bar chart comparing the percent of C→T conversions in MT neuron SNVs that appear in each 3bp context vs. germline SNPs. The MT neuron SNVs occur significantly more often in the TpCpN context (~44% vs. ~25%, p<0.0001, Fisher's Exact). **c**, The number of MT neuron SNVs appearing in evolutionarily conserved regions of the genome is significantly higher than expected by chance (27 actual vs. ~17 simulated with standard deviation = ~4, p=0.010, Monte Carlo). **d**, Percent of total genic MT neuron SNVs compared with SNVs from endodermal cell types. The dashed line indicates the percentage of the genome that falls into genes. MT neuron SNVs are enriched in genes relative to endodermal cell type SNVs (p=0.004, Fisher's Exact). **e**, Functional enrichment analysis of genes containing SNVs in MT neurons and in endodermal cell types by DAVID shows enrichment in a number of GO categories. The percentages of total GO categories related to neuronal function for each dataset are shown. **f**, RNA-Seq data from MT neurons allows us to define a set of expressed genes (top 50%). SNVs found in MT neurons are enriched in these genes compared to SNVs found in endodermal cell types. (p=0.025, Fisher's Exact). **g,** RNA-Seq data from Lgr5+ small intestine stem cells allows us to ask a similar question for SNVs detected in small intestine organoids. In contrast to MT neuron SNVs, small intestine mutations are depleted in highly expressed genes relative to SNVs found in MT neurons ($p=7.06 \times 10^{-4}$, Fisher's Exact).

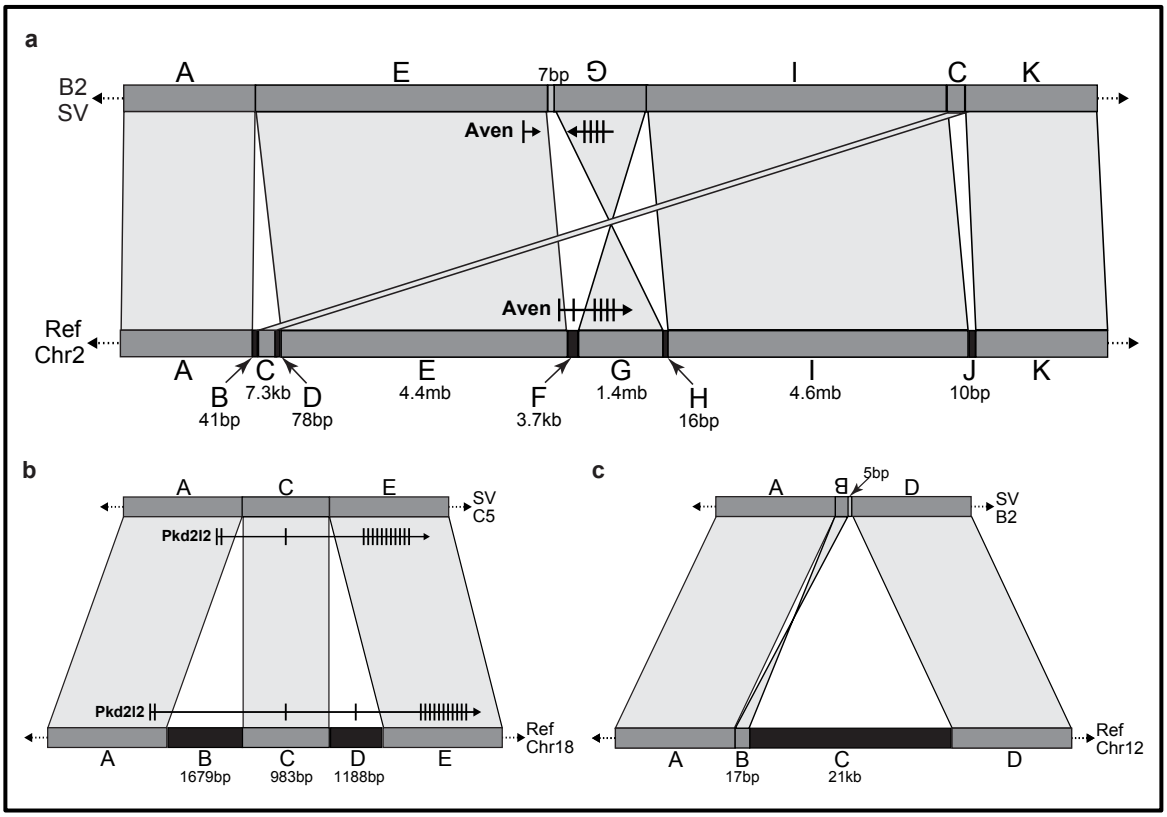

Figure 5.11. Structure of complex genomic rearrangements.
Several somatic structural variants in MT neurons exhibit notable complexity. **a**, Chromothripsis-like complex genomic rearrangement observed in MCNT-ES cell line B2 as the result of 6-10 double strand breaks. Bottom bar represents the wild type configuration, top bar represent the rearranged configuration in B2. Fragment C is transposed in a non-duplicative fashion 10.4Mbp downstream, between fragments I and K. Fragment F is deleted, which removes an exon from the Aven gene, and the inversion of Fragment G affects many of the remaining Aven exons. **b**, A complex variant on chromosome 18 in the C5 MCNT-ES cell line involves two deletions within 3Kbp. One deletes exon 4 of the pkd2l2 gene. The breakpoints show 4bp and 0bp of microhomology respectively. **c**, A single 21Kbp deletion on chromosome 12 in the B2 cell line. The 20bp region where the breakpoint occurs is comprised of Fragment B, a 17bp inversion with 2bp of microhomology with Fragment A, next to 5bp of DNA of unknown origin.

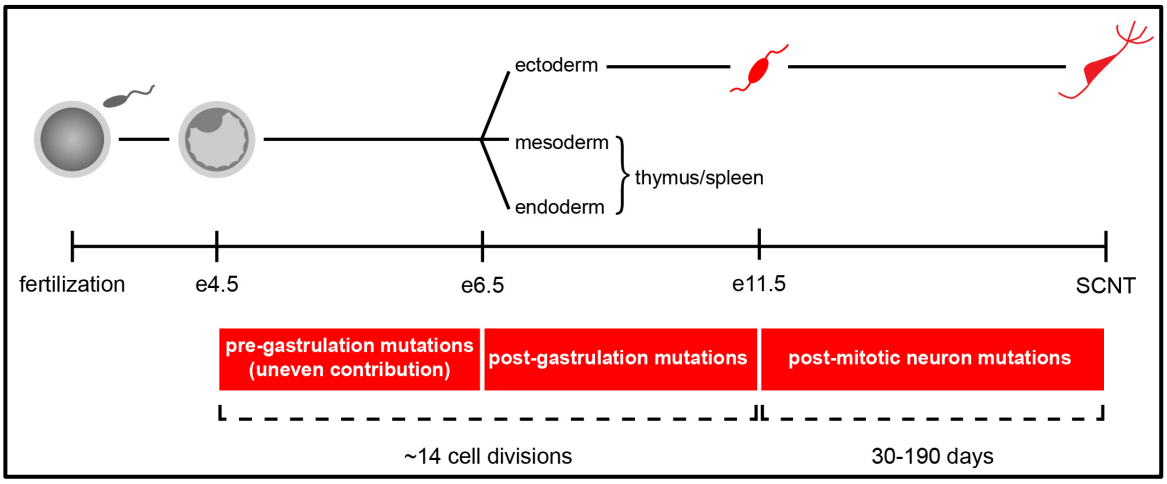

Figure 5.12. Developmental time periods when MT neuron mutations are acquired. MT neuron production occurs between embryonic day 9 (e9) and e18, and peaks around e11.5. We therefore approximate MT neuron progenitor cell cycle exit at e11.5. The beginning of our mutation detection window is defined by our filtering criteria designed to eliminate "germline mutations" (see text), which requires that putative MT neuron mutations be present in MCNT-ES cells, but absent from thymus/spleen control tissues. Most mutations acquired prior to gastrulation (e6.5) will be shared between MT neurons and thymus/spleen samples, and are therefore eliminated from our somatic mutation dataset. However, we conservatively extend our mutation detection window a few days before gastrulation (e4.5) as the small number of cells present in the early embryo can result in uneven distribution of early embryonic cells between germ layers. Therefore, some mutations arising prior to germ layer specification will be present at undetectable levels in the thymus/spleen. Assuming approximately 2 cell divisions per day over the 7 days between e4.5 and e11.5, leads up to predict that mitosis-associated mutations occur over 14 cell divisions. Post mitotically acquired mutations would start accumulating in the 9 embryonic days between e11.5 and birth, and end at the time of harvest for SCNT (between 3 weeks and 6 months of age).

Appendix 1 Supplemental Methods for Chapter 5

All methods and procedures in this chapter were performed by the author unless indicated otherwise in the section heading by appendage of "Scripps". In this latter case, the procedures were performed and the text written at the The Scripps Research Institute, La Jolla CA by Jen Hazen or other members of the lab of Kristin Baldwin in the Department of Molecular and Cellular Neuroscience, or in Scripps core facilities under their direction. In addition, SV and MEI PCR validations and subclone tests were performed in the Ira Hall Lab by research staff.

All animal procedures were approved by TSRI Institutional Animal Care and Use Committee.

A1.1 Immunohistochemistry and immunocytochemistry - Scripps

Newborn tissues were fixed at 4˚C overnight in PBS buffered 4% paraformaldehyde (PFA/PBS). Adult tissues were perfused with PFA/PBS, dissected, and fixed in PFA/PBS for 30 minutes on ice. After fixation, all tissues were sucrose protected in 30% sucrose at 4˚C overnight. Tissues were embedded in OCT and cryosectioned into 15 µ sections using a Leica CM3050S Cryostat. Sections were air-dried on superfrost slides for 40 minutes and fixed in PFA/PBS for 8 minutes. They were stained with primary antibodies against Iba1 (Wako, 019-19741, 1:1000), Ki67 (Acris, DRM004, 1:200), Olig2 (gift of Dr. Charles Stiles, Harvard Medical Center, 1:20,000), S100b (Abcam, ab868, 1:500), Sp8 (Santa Cruz, sc-104661, 1:500), and Tbr2 (Abcam, ab23345, 1:500). ES cells were stained as in Boland et. al. (Boland et al., 2009) MT neuron cell preparations for nuclear transfer (see below) were attached to glass slides using a cytospin cytocentrifuge and analyzed by immunocytochemistry. For cytospin, filters were wet by pre-spinning with 500 µL of PBS (6 minutes, 1,000 r.p.m.) followed by cellular attachment (6 minutes, 1,000 r.p.m.). The resulting "button" was fixed with room temperature PFA/PBS for 20 minutes and immunostained with Iba1, Ki67, Olig2, and S100b antibodies at the dilutions described above. Images were collected on a Nikon C2 or Nikon A1 confocal microscope and analyzed in Adobe Photoshop.

A1.2 Isolation of MT neurons for nuclear transfer - Scripps

MT neurons were dissociated and purified as in Brewer and Torricelli (Brewer and Torricelli, 2007) with the following modifications. We found it unnecessary to siliconize Pasteur pipettes to prevent cell loss and chopped olfactory bulbs using a scalpel rather than with a tissue slicer. We also used papain containing L-cysteine (Worthington Biochemical, PAP2 10 units/ml) as it has higher activity and allows shorter dissociation times (10 minutes total). We found it essential to add small amounts of DNase I (6.25 µg, Roche 10104159001) during papain treatment to prevent DNA related cell aggregation. After density gradient centrifugation, we found most MT neurons in the cell pellet fraction and a significant number in the 2 ml fraction immediately above the pellet. Cells from both fractions were combined and washed once in 10 mls of HAGB (Hibernate-A (Gibco A1247501), 1X B-27 supplement (Gibco 12587010), 500 µM GlutaMAX (Gibco 35050061)). After pelleting, cells were resuspended in 1 ml HAGB, transferred to a 1.5 ml centrifuge tube, pelleted again, resuspended in ~30 µls HAGB media, and stored on ice until nuclear transfer.

A1.3 Somatic cell nuclear transfer - Scripps

SCNT was performed as in Kishigami et al. (Kishigami et al., 2006b), except we extended the length of treatment with 5 nM Trichostatin A to 16 hours (6 hours during activation, 10 hours overnight) to improve efficiency of blastocyst and NT-ES cell generation (Kang and Roh, 2011).

A1.4 Derivation of MCNT-ES cell lines and MCNT-mice - Scripps

Embryos resulting from NT were cultured to blastocyst stage, then zonae pellucida were removed using a piezo-actuated drill needle (Nakayama et al., 1998). ES-cell lines were derived essentially as described in (Meissner et al., 2011), with some modifications in media composition. Briefly, zona-free embryos were cultured for 7-9 days on MEF feeder layer in ES-cell derivation medium (500 mls Knockout DMEM (Gibco 10829-018), 80 mls Knockout Serum Replacement (Gibco 10828-028), 6 mls MEM non-essential amino acids (Gibco11140-050), 6 mls Glutamax (Gibco 35050-079), 6 mls Pen/Step (Gibco 15140-122), 6ul B-Mercaptoethanol (Sigma M7522), 50 µm final concentration MEK1 Inhibitor PD98059 (Cell Signaling Technology 9900) and 2000 Units/ml LIF (Chemicon ESG1107)). Outgrowths of inner cell mass were picked and dissociated with 0.25% trypsin-EDTA (Gibco 25200-056). Cells were then expanded on a MEF feeder layer in ES-cell maintenance medium (500 mls Knockout DMEM (Gibco 10829-018), 80 mls Knockout Serum Replacement (Gibco 10828-028), 6 mls MEM non-essential amino acids (Gibco 11140-050), 6mls Glutamax (Gibco 35050-079), 6 mls Pen/Step (Gibco 15140-122), 6ul B-Mercaptoethanol (Sigma M7522) and 1000 Units/ml LIF (Chemicon ESG1107)). Tetraploid embryo complementation was performed as in our previous work (Boland et al., 2009).

A1.5 Microsatellite PCR assay to rule out host blastocyst contribution - Scripps

This assay was described by us previously (Boland et al., 2009), and relies on the detection of microsatellites that vary in length between MCNT-ES cells and tetraploid embryo cells. In these experiments, tetraploid embryos were F2 (BALB/cByJ X C57BL/6J-Tyr$^{c\text{-}2j}$). Therefore, to rule out trace contribution of tetraploid cells to MCNT-mice, we assayed for differences in microsatellite length diagnostic of both BALB/cByJ and C57BL/6J-Tyr$^{c\text{-}2j}$ strains. Genomic PCR was performed on DNA isolated from tissues of newborn and adult MCNT-mice for each TEC competent MCNT-ES cell line. Microsatellites assayed for each MCNT-ES cell line are listed in **Figure 5.4.** The following primers were used:

Microsatellite	Forward Primer	Reverse Primer
D17Mit133	TCTGCTGTGTTCACAGGTGA	GCCCCTGCTAGATCTGACAG
D6Mit102	CCATGTGGATATCTTCCCTTG	GTATACCCAGTTGTAAATCTTGTGTG
D6Mit15	CACTGACCCTAGCACAGCAG	TCCTGGCTTCCACAGGTACT

A1.6 Whole genome sequencing

Prior to sequencing, early passage MCNT-ES cells were separated from feeders by serial pre-plating on gelatin coated tissue culture dishes. DNA was isolated from MCNT-ES cells and thymus or spleen using standard phenol chloroform extraction and ethanol precipitation. Contaminating RNA was removed by RNase A digestion. Samples were sequenced by BGI (http://www.genomics.cn/en/index) using standard library prep for an Illumina Hi-Seq 2000. The target template length of approximately 500bp was chosen to give increased physical coverage to aid in accurate structural variant discovery. Each end of the paired-end data was 100bp in length. Quality control was performed on the output of the sequence run to eliminate reads with low base quality (≤5 ("A"-"E")) over at least 50%

of their length as well as reads with unknown nucleotides ("N") over at least 10% of their length.

A1.7 Initial alignment and post-processing

In these studies, default parameters were used for all bioinformatics software except as explicitly noted. We refer to an index with word length L and skip distance S as a L/S index. Mouse MCNT-ES cells and thymus/spleen control samples were sequenced using Illumina next-generation whole genome shotgun paired-end sequencing in which each read in the pair was approximately 100bp in length with a template length of approximately 475bp. Each sequencing lane was then separately aligned to the mm9 reference genome (July 2007 NCBI Build 37) using Novoalign v2.08.02 (Hercus, 2009) using a 14/1 index (-k 14, -s 1). Repetitive alignments were resolved using the random selection method (-r random).

GATK (DePristo et al., 2011; McKenna et al., 2010) (v2.5-2-gf57256b) and Picard Tools (Broad Institute) (v1.92) were used to further process alignments. Read group, library, platform, platform unit, and sample name information was added to the above alignments using Picard AddOrReplaceReadGroups. The BAM files for the various sequencing lanes for each cell line were then position sorted and merged using Picard ReorderSam and MergeSamFiles respectively. Duplicates were marked using Picard MarkDuplicates and removed with samtools view (Li et al., 2009a) (-F 0x400), resulting in a non-duplicate median per sample read-depth of approximately 32x-39x (**Table 5.3**).

A1.8 SNV and indel Detection

GATK and Picard Tools were further used for single nucleotide variant (SNV) and indel calling following the recommended best practices pipeline for GATK v2.0 (Van der Auwera et al., 2002). Here "indel" refers to any insertion or deletion of consecutive bases of less than 50bp in length. The GATK IndelRealigner was used to realign indel regions identified by RealignTargetCreator. Mate-pair information was cleaned by Picard FixMateInformation. We then used GATK BaseRecalibrator and PrintReads to recalibrate base quality scores. This step takes as input a set of known sites, which we created by selecting those single nucleotide polymorphisms (SNPs) marked as "High Confidence" by the Mouse Genomes Project (MGP) in the 129S1 mouse strain (Keane et al., 2011). The GATK UnifiedGenotyper was then run on all samples combined, calling indels and SNPs together, using per sample read-depth downsampling to a maximum read-depth of 500 (-glm BOTH –dt BY_SAMPLE –dcov 500).

GATK VariantRecalibrator and ApplyRecalibration steps were then run first on SNPs (--mode SNP), then on indels (--mode INDEL), to assign our calls into one of four sensitivity tranches. These steps require SNP and indel "truth" sets that were created as follows. For SNPs, we again started with the high confidence 129S1 SNP calls from the MGP, intersected these with our own autosomal GATK SNP calls from above, and selected the top 1 million such calls as ranked by the MGP variant quality score. For indel variant recalibration, we used all 129S1 indel calls from the MGP.

We then identified putative *de novo* somatic SNV and indel variants private to MCNT-ES cells lines using custom scripts that implement a modified version of the approach used by Kong et al. (Kong et al., 2012) Although we called variants in each donor mouse separately, we used information from the same locus across all samples to help reduce false positives. For a given mouse, the samples were partitioned into three sets; (1) the "control" sample of the thymus/spleen for that mouse, (2) the "MCNT-ES" cell line(s) for that mouse, and (3) the

"other" samples, comprised of all samples from the other mice. In what follows, "RR", "AR" and "AA" will refer to the genotype of a locus as homozygous for the reference allele (R), heterozygous, or homozygous for the alternate allele (A) respectively. The alternate allele genotype (AAG) of interest for the calling process depends on the chromosome and sex of the mouse. We used AR for all autosomes and for the X chromosome of female mice, and AA for the X/Y chromosomes of male mice. The variant allele frequency (VAF) is defined as the (alternate allele read-depth)/(reference allele read-depth + alternate allele read-depth). Phred likelihood scores for genotypes and per allele read-depth information are provided by GATK in the VCF output file.

To be called a putative somatic SNV in a particular MCNT-ES cell line, a SNV locus/allele pair was required to meet all of the following criteria:

1. The alternate allele is not reported as a variant at the same locus in any inbred mouse strain by the MGP at either high or low confidence.
2. The call appears in one of the 19 autosomes or the X or Y chromosome. No calls are made in "random" or "unknown" scaffolds. Mitochondrial variant calls were also excluded from the analysis because mitochondria in MCNT-ES cell lines are expected to originate from the oocyte used in nuclear transfer, not from the original neuron.
3. The control sample and the MCNT-ES cell line(s) from the mouse of interest each have a total read-depth between 10 and 250.
4. A control RR/AAG ratio of phred likelihood scores $\geq 10^5$, and a control VAF of at most 5%.
5. An MCNT cell line AAG/RR ratio of phred likelihood scores $\geq 10^{10}$, and VAF of at least 30% (95% for X/Y chromosomes in male mice).
6. A RR/AAG ratio of phred likelihood scores ≥1, and a VAF of at most 5% for all "other" samples.

Indel calling strategies are known to have higher false positive rates than SNV calling strategies. Therefore, we slightly modified the filtering criteria for indels to be more conservative as follows. In step 1, the variant is eliminated as a somatic call if it overlaps any indel reported by the MGP in an inbred mouse strain regardless of the type and size of the indel. In steps 4 and 6, the VAF for both the control sample and all samples from other mice are held to the stricter criteria that it be equal to zero.

We further categorized our SNV and Indel calls by the GATK VariantRecalibration assigned tranche annotation as high confidence (HC) if they fall into the two lowest sensitivity (highest specificity) tranches with an implied false discovery rate (FDR) threshold for the corresponding truth set of 1%. The remaining calls are categorized as low confidence (LC). As discussed below, our validation rates are markedly higher for the HC calls than for the LC calls. The resulting somatic SNV and Indel calls are in **Tables S4-5**.

A1.9 Structural variation breakpoint detection

We used a custom pipeline including Novoalign, YAHA, LUMPY and custom scripts to detect structural variant breakpoints (**Figure 5.6**). Here we define a structural variant (SV) as an apparent deletion, tandem duplication or inversion (as defined by relative read-pair orientation) of greater than 50bp in length, or an unexpected juxtaposition in the sample genome of two loci that appear far away from each other (>1Mbp) on the same or different chromosome(s) in the reference genome (which we refer to as "distant" rearrangements).

Insertions are not directly detected by LUMPY, but will instead be composed of two of the above event types (one for each of the two insertion breakpoints).

LUMPY can map SV breakpoints using evidence from both discordant paired-end reads ("read-pairs") and split-read mappings from multiple samples to find SVs (Layer et al., 2014). Informative discordant read-pairs were extracted from each BAM file as those read-pairs in which both reads were mapped, the mappings were either 1) on different chromosomes, 2) had improper strand orientation, or 3) had a template length that fell outside the mean length ±5 standard deviations (STDs). The insert size mean and STD was calculated for each dataset using custom scripts using properly paired alignments (samtools view –F 0x400 –f 0x2). In order to reduce the probability of false positive SV calls, we further filtered the set of input discordant read-pairs as follows. We first located collections of nearly duplicate pairs in which the corresponding mates of each pair mapped to the reference genome within ±3bp of each other. From such collections, we eliminated all but the pair aligned with the least edit distance from the reference genome. Discordant reads were converted to bedpe format using bedtools V2.16.2 (Quinlan and Hall, 2010) bamToBed and pairBedToBedpe, and additional duplicates were removed using dedupDiscordantsMultiPass.py (-s 3).

Separately, we extracted putative split-read alignments that were either unmapped or had a clipped region of ≥20bp on either end of the alignment. These were then realigned using YAHA version 0.1.78 (Faust and Hall, 2012) with an 13/1 index (-L 15 –S 1), and default alignment parameters except for maxHits of 2000, and minMatch of 15 (-H 2000 –M 15). From the resulting alignments, we selected for input to LUMPY reads that had a single split alignment (two mappings) in which each aligned portion involved ≥20bp of query sequence that was not included as part of the other aligned portion. We also required that split-read alignments suggesting a deletion variant had an implied deletion size ≥50bp (our definition of SV).

LUMPY was run on the above-described discordant read-pairs and split-read mappings from all eleven samples, requiring at least 4 confirming reads across 11 samples for a call, and a trimThreshold of 10^{-3} (-mw 4 –tt 1e-3), using a minimum alignment mapping quality of 10, and excluding all genomic regions in which any cell line had an aligned read-depth >500.

The resulting SV calls were filtered to find putative *de novo* somatic variants that appear in a single MCNT-ES cell line. We required such a call to meet all of the following criteria:

1. The SV call had at least 5 supporting discordant read-pairs and/or split-reads from one MCNT-ES cell line, and no supporting reads in any other MCNT-ES or control sample from any mouse.
2. The SV call was not previously reported as a germline polymorphism by MGP for any mouse strain. A LUMPY call was judged to correspond to an MGP call if the two were of the same variant type (e.g.,. deletion) and were at the same genomic location, as defined by 50% reciprocal overlap (bedtools intersect -r -f 0.5). Distant rearrangement involving >1Mbp of genomic sequence, or spanning multiple chromosomes, were not filtered in this manner since such variants were not reported by MGP.
3. The call appears in one of the 19 autosomes or the X or Y chromosome. No calls were made in unmapped contigs or in mitochondrial DNA.

A1.10 Mobile element insertion detection

Mobile element insertions (MEIs) pose a challenge for SV calling algorithms due to several factors including the fact that the mobile element (ME), or "transposon", is itself composed of repetitive sequence. Therefore, we have developed our own MEI calling pipeline based on the strategy used by Lee et al to study somatic retrotransposition in human cancers (Lee et al., 2012a).

The general approach is to start with all the reads that the aligner had difficulty aligning to the reference genome, and re-align them to a custom-built library of mobile element sequences. The mates to the reads that map well to this ME library are then used to identify regions of the sample genome in which to search for MEIs. In addition, we look for confirming evidence of MEIs using split-read mappings in which one side of the split maps to the ME library, and the other side to the reference genome next to the predicted ME insertion point (**Figure 5.7**).

The ME library is formed using both canonical sequences from version 18.02 of RepBase (Jurka et al., 2005) and their variants predicted to appear in the mm9 reference genome by RepeatMasker (Smit et al., 1996-2010) and included in the mm9 UCSC RepeatMasker annotation track (downloaded from http://genome.ucsc.edu/cgi-bin/hgTables/). From RepBase RepeatMaskerLib.embl and mousub.ref (downloaded from http://www.girinst.org/server/RepBase/) we found 120 LTR sequences labeled with "Species: Mus_musculus" and 6 SINE sequences, respectively. From the mm9 RepeatMasker annotation track we selected the genomic regions for all LINEs, SINEs, and LTRs with low sequence divergence (≤30 millidev) and length of at least 100bp, then extracted the corresponding DNA sequences from the reference genome using bedtools getfasta. We then removed duplicate sequences from the above, and appended multiple "N" bases to the ends of each sequence to aid in alignment. The final ME library contains 51,413 unique sequences. Detailed information about the composition of the ME library can be found in **Table S1**.

We selected reads to align to the ME library that met any of the following criteria:

1. It was the unmapped read of a pair in which one read is mapped and the other unmapped.
2. It was either read of a discordant read-pair in which either the reads were aligned to separate chromosomes, or the reads were aligned at least 100Kbp apart from each other.
3. Any mapped read not in the above two categories whose alignment was clipped by at least 20bp.

The above reads were then aligned to the ME library using YAHA. Since the ME library is highly repetitive, we used very sensitive alignment parameters: an 11/1 index, maxHits of 9000, minMatch of 15, and a maxGap of 20 (-H 9000 –M 15 –G 20).

We then formed clusters separately for each ME subtype as follows. From the ME library alignments, we selected ones matching the current ME type and subtype that were from a discordant or unmapped read and had a good alignment to the ME library, defined as at least 50bp in length and clipped by no more than 3bp on at least on end. We then extracted the aligned coordinates for their mate in the reference genome, and formed clusters from those reads that were aligned to the same strand and fell within the inter-read distance from each other. The inter-read distance is calculated separately for each sample as ⌊(ETL –

RL) x 2 / 3⌋ where ETL is the extended template length (median template length + 3 STDs) and RL is the read length (100). We then found potential ME insertion points as a pair of such clusters from the same ME type and subtype such that the reference coordinates of a plus strand cluster were 5' of a minus strand cluster within twice the inter-read distance, and had at most 20bp of overlap. In addition, the cluster pairs had to have at least 6 combined supporting reads from the two clusters. These pairs were then filtered to exclude those with at least 25% of their length overlapping an ME of the same type and subtype annotated in the reference genome as defined by the UCSC repeat masker tract ME (bedtools intersect –f 0.25).

Confirming split-read mappings for remaining pairs were found as follows. All unmapped alignments, and any clipped alignments overlapping a pair region were aligned to the reference genome with YAHA using the same parameters as above; an 11/1 index on the mm9 reference genome, and these alignment parameters: (–H 9000 -M 15 –G 20). We then counted as a confirming split-read mapping one in which the type and subtype of the ME matched the one from the cluster pair, the portion of the read aligned to the reference fell within exactly one of the pair clusters, and the two split-read alignments together cover almost the entire read length with at most a few unmapped bps (the alignment mapped to the reference and the alignment mapped to the ME library ended within 3bp of opposite ends of the read, and there was a maximum of 4bp of unaligned sequence between them). We added the count of such split mappings to the total read count of the associated cluster, and kept a list of all of the reference loci for their reference aligned portion nearest to the implied insertion breakpoint to more precisely define where the breakpoint occurred (**Figure 5.7**).

We then filtered the cluster pairs formed above to find putative *de novo* somatic MEIs in a single MCNT-ES cell line using a similar strategy we used to identify *de novo* somatic SV events. We first eliminated cluster pairs that had fewer than 10 confirming reads. We then eliminated cluster pairs with evidence in other samples as follows. We separately intersected the genomic region of each cluster of a pair with clusters from all other samples that had the same ME type (disregarding subtype) and were on the same strand. We then eliminated all cluster pairs that had any confirming reads from such a matching cluster. Finally, we further filtered the remaining pairs to eliminate any pair that had any overlap with any MGP MEI call from any mouse strain regardless of ME type or subtype.

A1.11 Copy number variation detection by read-depth analysis

To detect copy number variation (CNV), we used a read-depth strategy very similar to the one described by Malhotra et al. (Malhotra et al., 2013) Assuming that Illumina genome sequencing uniformly samples the source DNA, the DNA copy number within a given genomic region should be directly proportional to the number of sequence reads mapped to the region relative to other regions. However, local read-depth is subject to two major sources of bias that must be overcome to make these calculations more accurate. First, Illumina sequencing exhibits significant GC bias such that local coverage depth falls off at GC content extremes, especially in regions with high GC percentages (Aird et al., 2011). To counteract this bias, we normalize the coverage data within small genomic regions by their percent GC content. The strategy used to do this normalization is based on the observation that the read-depth in regions of similar GC content approximates a normal distribution. Second, repetitive sequences are known to pose difficulties in sequence alignment and assembly, causing potentially large fluctuations in local read-depth mapping to the

reference genome. To counteract this bias, we base all of our calculations on read-depth in unique genomic regions.

We therefore start by breaking the reference genome up into regions ("windows") containing 5Kbp of unique sequence as defined by a mappability value equal to 1 in the UCSC 100mer mappability track (crgMappabilityAlign100mer). This results in 458,040 windows with a mean and median size of 5796 and 5030 bp, respectively.

We then process each of our cell samples separately as follows. We first count the number of reads mapped to the unique portion of each such 5Kbp window, then consider as a group those regions with the same percent GC content in 1-3% increments, e.g. (45.0-47.0%] GC. We then use the autosomal windows in each group to calculate the median and median absolute deviation (MAD) of read-depth for the group, and estimate its normal distribution using the MATLAB "normfit" function using all windows in each group that are within ±4 MADs from the median read-depth for the group. This yields a mean and standard deviation (STD) for each group as a whole. For each window within the group we then calculate the normalized read-depth as the raw read-depth for the window divided by the median read-depth for the group and multiply by two (assuming a diploid genome). Similarly we calculate a Z-score for each window as the raw read-depth for the window minus the mean read-depth for the group divided by the STD.

We next combine consecutive windows with similar Z-scores into copy number segments as described in (Malhotra et al., 2013) using the circular binary segmentation function in the DNAcopy package in R (http://cran.r-project.org/) with the following parameters: (undo.splits="sdundo", undo.SD=1 and alpha=0.001). For each segment we keep track of the count, mean, STD, median, and MAD of the read-depth values for windows it contains. In addition, for each SCNT-ESC cell line, we performed the same segmentation as above based on the $\log_2$ of the ratio of the normalized cell line read-depth divided by the corresponding thymus/spleen control sample read-depth. Such a division is useful for determining somatic CNVs as described below. We also calculated the total dataset median and MAD for each cell line and $\log_2$ ratio dataset separately for autosomes and the X chromosome to account for the expected difference in copy number on the X/Y chromosomes in males.

Finally, we called the somatic CNVs as follows. As CNV calling is fairly error-prone, we chose to use conservative filters that result in a low false positive rate, but potentially lower sensitivity. We find all segments in the $\log_2$ ratio datasets for the MCNT-ES cell lines that are formed from at least 3 windows and have a normalized segment median read-depth that is plus/minus at least 6 MADs above/below the full dataset median normalized read-depth for the corresponding chromosome set (autosomes or chrX as appropriate). From these, we remove any segment(s) that overlap with a segment in any of the 4 control samples with a normalized segment median read-depth that is plus/minus at least 6 MADs above/below the full dataset normalized read median read-depth. Together, these filters require a strong signal in one or more of the MCNT-ES cell lines in a genomic region that has no such signal in any control line, indicating a *de novo* somatic variant. Interestingly, this filter criteria results in putative CNV duplication calls in T-cell receptor alpha and/or gamma sites for all MCNT-ES cell lines using thymus as the control sample (B2, B3, B4, and E1). These are actually an artifact of the deletions in these regions in the thymus samples due to V(D)J recombination, and act as a positive control for the calling pipeline. Removing these spurious calls leaves us with four CNV calls all of which are also LUMPY SV breakpoint calls as shown in **Table S3**. Note that the above calling strategy requires segments of at least

three adjacent 5Kbp windows and is insensitive to any CNV below ~15Kbp in size. We have only five validated LUMPY breakpoint calls that are unbalanced variants of this length. Four of them are found as CNVs by read-depth analysis, and the fifth duplication call falls just below our detection thresholds in a segment three windows in length with a normalized copy number that is 4.7 MADs above the median.

A1.12 Somatic variant false negative rate estimations

To gauge the sensitivity of our somatic variant calling strategies in the absence of a known set of true positives, we estimate the false negative rate (FNR), and calculate the sensitivity as 1-FNR. The general strategy is to find a set of high confidence germline variants of the variant category of interest, called the gold standard set (GSS), and then count how many of these were detected in our analysis and would pass all relevant MCNT somatic call filters. To eliminate issues regarding sex chromosome differences across datasets from both male and female mice, all of our FNR estimates are based solely on autosomal variants. The detailed calculations of FNR estimates are shown in **Table S2**.

A1.13 Single nucleotide variant and indel false negative rate estimation

The GSS set for SNV calls was found on a per mouse basis as follows. We started with the set of all GATK autosomal SNP calls for a given donor mouse, and selected the subset of such calls that were also found as high confidence calls by the Mouse Genomes Project (MGP) in any inbred mouse strain. From this set, we further selected those that were called heterozygous in our data by GATK in at least one sample from the mouse in question. This is an important step, as we expect that, barring some rare event that causes loss of heterozygosity, all *de novo* somatic autosomal variants should be heterozygous. We therefore use solely heterozygous calls in our GSS as they should display similar patterns of variant allele frequencies and associated genotype phred likelihood scores as our sought-after somatic variants.

We then applied all of our MCNT-ES cell line filtering criteria except the "control" and "other" sample filters (filters 4 and 6 from above), and counted the percentage of the GSS calls that are eliminated in each MCNT-ES cell line. We take this as our estimate of overall FNR for that cell line. We then calculated the overall FNR rate for each mouse as the average of the FNRs of the (one or more) MCNT-ES cell line(s) from that mouse. The resulting per-mouse overall FNR estimates for all SNV calls range from 6.7% to 11.1%, and for our high confidence SNV calls from 19.0% to 22.8%.

We estimate the FNR rates for the indels in a similar fashion with one difference. The MGP does not report confidence levels for indels. Therefore, we intersected our per-mouse GATK heterozygous autosomal indels calls with all MGP inbred indel calls to find the per-mouse GSS set. The resulting per-mouse overall FNR estimates range from 22.5% to 27.0%, and for our high confidence calls from 24.3% to 28.6%. Note that the FNR estimates for our high confidence SNV and indel calls are quite similar, while those for all indel calls are significantly higher than for all SNV calls. This is not surprising given the increased difficulty in calling indels vs. SNP and the lower quality "truth" set we had available as input to the GATK tranche calculations, which resulted in GATK placing almost all of the indel calls in the high confidence tranches. See **Table S5** for details.

A1.14 Structural variant and mobile element insertion false negative rate estimation

For our SV and MEI FNR estimates, we also calculate a gold standard set (GSS) on a per mouse basis. To find our GSS set, we started with MGP calls from the 129S1 mouse strain,

and found the subset of these that are located in genomic regions that we predict to be in a haplotype block inherited from the 129 strain lineage in the mouse of interest. This is necessary because the different donor mice are mixed 129/Black6 genetic background, but due to their breeding history have inherited distinct 129 haplotype blocks. To find these haplotype regions, we first determined the set of germline SNPs called by GATK in each mouse that are also called by the MGP in the 129S1 mouse strain. We call these the 129S1-SNPs for that mouse.

Deletions are the most numerous and easiest to detect structural variants. We therefore have highest confidence in the deletion call annotations in the MGP. To estimate the FNR of our LUMPY SV breakpoint calls, we restricted our GSS to MGP deletions found in the 129S1 mouse strain. We further restricted the GSS to those calls that have two 129S1-SNPs within 250bp of both sides of the outer span of the call region. This results in per-mouse GSSs with ~2200 calls each. For initial FNR estimates, we counted the percentage of the GSS calls that do not have 50% reciprocal overlap with a LUMPY deletion call in the cell line of interest. The resulting per-mouse FNR estimates range from 38% to 42%. However, this dramatically overestimates the true FNR. Approximately half of the calls in each GSS are small (less than 500 bp). For these deletions, the uncertainty in the breakpoint location calculated by LUMPY is large relative to the size of the call. As a result, approximately 75% of these small calls failed the above test compared to only 6% to 9% for larger deletions. Therefore, for more accurate FNR estimates, we required 25% and 50% reciprocal overlap for the small and large calls respectively, then formed a weighted average of the resulting FNR estimates leading to final per-mouse FNR estimates ranging from 10.2% to 13.5%.

For our MEI calls, we formed an initial GSS in a similar fashion to the deletion calls. We chose those MGP MEI calls from the 129S1 mouse strain that have two 129S1-SNPs within 250bp on both sides of the insertion point, estimated as the midpoint of the insertion call region. We then counted the percentage of these calls that do not intersect the insertion region of any of our MEI calls of the same ME type in the cell line of interest. This results in initial per-mouse FNR estimates ranging from 45.8% to 48.2%. However, it is likely that this is an overestimate of FNR due to false positive MEI calls in the MGP. Therefore, we formed a stricter GSS for each mouse by adding the requirement that we have at least weak evidence for the insertion in our data. Specifically, we required there be at least two reads from any of our clusters from the same ME type that overlap the insertion region of the GSS call. We then again count the percentage of these restricted call set that do not intersect the insertion region of any of our MEI calls of the same ME type. The resulting per-mouse FNR estimates range from 19.4% to 15.5%. These probably underestimate the true FNR rate because we have pre-selected MGP calls that we are likely to find. The real FNR rate probably lies between these two extremes. See **Table S2** for details.

A1.15 Somatic variant validation strategy

To validate putative *de novo* somatic mutations, we performed PCR amplification of the genomic region containing the putative mutation using DNA from both the MCNT-ES cell line of interest and its associated control thymus or spleen sample from the same donor animal. The products were Sanger sequenced to verify that the mutation was present in the MCNT-ES cell line but not in the control sample. For SNVs and indels, a random subset of calls was tested. For SVs and MEIs, all calls were tested and we also performed additional tests to eliminate the potential for the mutation to have arisen during clonal expansion or reprogramming using subclone analyses. The number of mutations tested for each mutation category and the resulting false discovery rates are given in **Table S3**.

Bulk DNA for validation PCR reactions was prepared by standard methods. PCR reactions were composed of 7.55μL 16 of 2x Phusion High-Fidelity Master Mix (New England Biolabs), 0.5μL each the left and right primers at 10μM concentration, ~20 ng of DNA, and filled to 15μL with distilled water. PCR reactions were run in a Bio-Rad DNA Engine Dyad Peltier Thermal Cycler with annealing temperatures of both 60°C or 65°C with the following program: 1) 98°C for 30 seconds; 2) 98°C for 10 seconds; 3) 60°C or 65°C for 30 seconds; 4) 72°C for 1 minute; 5) go to step 2 25-30 times (depending on the locus); 6) 72°C for 5 minutes. Reactions were electrophoresed on 1.5-2% agarose gels made with 1X TAE buffer and containing 0.2μg/ml ethidium bromide, and visualized under UV light.

A1.16 Validation of putative somatic SNVs - Scripps

To test SNV calls, we designed PCR primers to amplify the region of genome containing the predicted SNV. PCR was performed on genomic DNA from MCNT-ES cells and from thymus or spleen of the original donor animal. The resulting PCR product was sequenced by Sanger sequencing, either directly, or after gel extraction if greater than one PCR product was amplified. Most PCR products were sequenced using either the forward or reverse primer from amplification. Any additional internal sequencing primers required are listed in **Table S4**. Sequencing results were aligned to the mouse genome to confirm the intended region was amplified before specifically looking for the presence or absence of the predicted SNV. If the predicted mutation was present in the predicted MCNT-ES cell sample and not in control donor tissue, the SNV was judged to be validated.

A1.17 Validation of putative somatic indels - Scripps

Indel validation was essentially identical to SNV validation. However, in SNV detection, single base polymorphisms are visible directly in the sequencing data. In indel validation, longer heterozygous sequences result in a decay of the quality of the sequencing data starting with the first base that differs between the reference and mutant alleles. So, the presence, bounds, and in most cases the actual sequence of the indel were confirmed by Sanger sequencing from both upstream and downstream of the predicted indel. PCR primers are listed in **Table S5**.

A1.18 Validation of putative somatic structural variants and MEIs

To validate putative *de novo* somatic SV and MEI breakpoints, PCR was performed on genomic DNA from MCNT-ES cells and donor animal thymus or spleen as control. Primers were designed to flank a putative SV breakpoint to produce a 200-800 bp product for the variant allele, and to produce either no product or a product of significantly different size for the reference allele. Primers were designed to be 18-25 bp in length, with a 57°C-63°C Tm, and 40%-60% GC content. All validating primers are listed with their corresponding variant call descriptions in **Tables S6-7**. CNV calls were not separately validated, as all somatic CNV calls were redundant with a validated SV call.

If a unique amplified product was present in the predicted MCNT-ES cell line(s) but not the control, the breakpoint was considered validated. If the same product(s) were present in both the predicted MCNT-ES cell line(s) and control DNA, the breakpoint was judged to be a germline variant. If amplified products were absent in all lines, or if the primers were non-specific (i.e. yielded multiple products) a second pair of primers were made. If the second pair of primers also failed to yield specific product(s) then the variant was judged to be a false positive. We note that this could result in a small number of false negatives due to off target amplification at loci that are difficult to amplify cleanly. The unique band produced

by validating primers was cut from the gel and sent to GENEWIZ (http://www.genewiz.com) for capillary sequencing of both strands.

To further determine that validated SV calls were present in the original donor neuron and did not arise either in culture or during reprogramming, PCR was performed with the validating primers on subclones from the relevant MCNT-ES cell line. DNA from MCNT-ES cell subclones was purified in 96-well format using the following protocol. MCNT-ES cell subclones were grown to confluency on MEF feeder cells. They were then washed with PBS and incubated in 50 uls lysis buffer (100 mM Tris pH8.0, 5mM EDTA, 0.2% SDS, 200mM NaCl, 100 µg/ml proteinase K) for 16 hours at 55˚C. To precipitate DNA, lysed cells were incubated in 100 uls of cold 100% ethanol for 30 minutes on an orbital shaker. Supernatant was removed, and precipitated DNA was washed twice with 70% ethanol and air dried for 20 minutes. The resulting DNA was resuspended in 35 µl of TE by incubating overnight at 37 ˚C. PCR was then performed on 1ul of DNA from subclones using the same primers and procedures described previously.

A1.19 Structural variant and MEI breakpoint determination

SV and MEI call breakpoints were determined to single base pair resolution primarily by split-read mapping of the capillary sequence data of the unique PCR product validating the call. Split-read mapping was done using YAHA with sensitive parameters and a breakpoint penalty neutral to variant length (a 11/1 index, -M 12 -BP 20 -MGDP 1 and –H 2000 for SVs and –H 65525 for MEIs). However, all of the four validated MEI LINE insertions had PCR validation of only 5' breakpoint due to the difficulty in finding usable primers in poly-a tails. Therefore, MEI breakpoints were determined by visual inspection of clipped alignments using the Integrative Genomics Viewer (https://www.broadinstitute.org/igv/home). Once the breakpoint locations were determined, we calculated additional breakpoint features by looking for additional features of the split-read mappings. Microhomology for SV breakpoints manifests as overlap of the two split-read alignments on the query, and target-site duplication for MEIs as the distance between the insertion breakpoints on the reference (**Figure 5.7**). The details of the breakpoint architectures of SVs and MEI are provided for each validated call in **Tables S6-7**. In addition, about half of the SV breakpoints were caused by complex genomic rearrangements as shown in **Figure 5.11**.

A1.20 Detection and validation of shared mutations.

We sought to identify somatic mutations that are shared among multiple MCNT-ES cell lines. Somatic variants that are shared among cell lines derived from a single donor mouse could exist due to clonal mutations that arose early in development, whereas variants that are shared among lines from different donor mice could exist due to recurrent mutation at hotspots, or conceivably due to programmed rearrangement (as in the immune system). Since it has long been hypothesized that recurrent structural mutations might be involved in generating neuronal diversity, we focused our search for mutations shared across different donor mice to SVs and MEIs. We restricted our search for shared SNVs to within-mouse mutations.

Within-mouse shared SNVs are naturally detected by our primary SNV calling procedures outlined above. We identified 13 such SNV calls in the B mouse and 8 in the C mouse (bottom of **Table S4)**. Shared SVs were identified from the primary LUMPY run as before, except that we modified criteria 1 to require at least 5 supporting reads in each of two or more MCNT-ES cell lines. These criteria identified 73 shared SV calls of which 13 well-supported candidates were tested; 10 were shared between two mice, and 3 were shared

among MCNT-ES cell lines from the same mouse. Shared MEIs were identified as before except that we selected pairs that had at least 6 overlapping cluster reads in at least one other MCNT-ES cell line. These criteria identified only three shared MEI calls; two within the B mouse, and one shared between two mice. No CNV calls made by read-depth analysis were shared among multiple cell lines.

We attempted to validate putative shared mutation calls using the same methods described above, except that we included all relevant MCNT-ES and control samples during PCR validation and subsequent Sanger sequencing. We were able to successfully make primers that yielded a product that could be sequenced for 10 of the shared SNV calls, all of which showed that the putative mutation was also in the control sample, and thus a germline SNP. It is also worth noting that all of the 21 putative shared SNVs were low confidence SNV calls that are known to have a low validation rate. Thus, our detection of zero high confidence SNVs that are shared among multiple neurons from the same mouse is by itself strong evidence that early-arising clonal mutations are extremely rare. For the 13 shared SV calls and 3 shared MEI calls, all failed validation either because the mutation was discovered in one or more of the control samples, or because we failed twice to successfully make usable primers (**Table S3**). Overall, we identified no bona-fide shared mutations either among MCNT-ES cell lines within a single mouse, or within different mice.

A1.21 Analysis of predicted functional consequences of somatic mutations

We first determined how many of the mutations have gene-coding effects. For SNVs and Indels, we used SnpEff (Cingolani et al., 2012) version 3.1m and filter for effects in codons. For SVs and MEIs we determined the coding effects using a combination of feature intersection (bedtools intersect) with RefSeq exome, as well as visual inspection. We identified five SNVs, one indel, and four SVs that disrupt exons in 11 different genes with various levels of predicted severity. Four of the genes involved are highly expressed in MT neurons as determined by our RNA-Seq data (**Table 5.5**). Many of the remaining tests focus on our high confidence SNV calls, as they are the most numerous and have been identified with high accuracy.

We compared our SNV base conversion profiles to those reported in other studies (Behjati et al., 2014; Kong et al., 2012; Welch et al., 2012; Young et al., 2012) by strand normalizing the base conversion and counting the number of mutations in each of the 6 possible categories (**Figure 5.10a**). As is common, we have more C→T conversions than any other base conversions. We compared the 3-base context in which these occur to our germline SNPs as a possible indicator of mutational process. The germline SNPs were determined for each mouse separately by using the same criteria used to identify somatic SNVs except that all MCNT-ES cell lines and the control sample from the same mouse were all used as sample lines, and no parent or other lines were used. As almost all of the calls occur in all mice, the final germline call set was determined by taking the union of the calls in each mouse. The strand corrected 3-base contexts were identified using bedtools getfasta. We find that the somatic SNVs in MT neurons are enriched in C→T conversions taking place in TpCpN contexts, as compared to germline SNPs using Fisher's Exact Test ($P<0.0001$) (**Figure 5.10b**).

We next sought to determine if our somatic SNV calls occur randomly throughout the genome, or instead co-locate more or less frequently than chance in certain genomic features. We restricted this study to autosomes to eliminate any issues with the fact that we have both male and female mice in this study. We chose to use nine genomic features that

were broadly diverse, of potential functional interest, and that were common enough to have more than 5 somatic SNVs fall within them. The chosen features were 100mer-uniquely-mappable regions, 100mer-unmappable regions, segmental duplications, elements conserved across placental mammals, simple repeats, LINEs, RefSeq exons, RefSeq transcripts, and RefSeq transcripts that are highly expressed in MT neurons. Almost all of these feature tracks were either directly downloaded from the UCSC table browser (/hgdownload.cse.ucsc.edu/goldenPath/mm9/database/) or readily derived from such a download. See **Table S9** for details.

We then calculated SNV enrichment relative to chance using two different strategies. First, we separately ran a 10,000 trial Monte Carlo simulation for each genomic feature. The 395 autosomal high confidence SNVs were distributed randomly throughout the autosomes using bedtools shuffle while excluding all reference genome assembly gaps and regions in which the read-depth in any sample was less than 10 or greater than 250. These latter regions were excluded from the simulations as they were also excluded by our SNV calling strategy. From the simulations, we captured the mean and standard deviation of the number of SNVs that fell in the feature, and the number of trials in which the number of SNVs in the feature was greater than or less than the actual count of the number of somatic SNV calls that fell in the feature. This latter provides an estimate of the p-value of an enrichment or depletion of our SNVs vs. random chance. Second, we calculated the expected value of the number of SNVs that should randomly fall in each feature based on the length of the feature vs. the accessible genome. We then also derived a p-value for the likelihood that our SNV count in the feature falls within the 95% confidence interval given the feature length using the Poisson Test. The simulation means and the expected values based on feature length are in very close agreement. All results are summarized in **Table S9.**

The mappable and unmappable regions were chosen as controls, as it is easier to correctly align reads and make mutation calls within unique regions of the genome. As expected, our SNV calls are significantly enriched in the mappable regions and depleted in unmappable regions. Also, our SNV calls are significantly depleted in segmental duplications. We estimate that at least half of this effect is due to the fact that only 38% of the segmental duplicates track falls within mappable regions, and 7 of our 9 SNV calls fall within that 38%. Similarly, our SNV calls are mildly depleted in simple repeats (not statistically significant). Interestingly, our MT neuron SNVs are significantly enriched in elements conserved across placental mammals where one might expect the opposite due to cellular selection pressure. Similarly, our SNVs are enriched in the RefSeq exome, transcripts, and transcripts highly expressed in MT neurons, but these enrichments do not quite reach statistical significance. However, we find this suggestive, and note that our statistical power to make such discriminations is limited by the relatively small number of SNVs found in this study.

Therefore, to further explore the enrichment of MT neuron SNVs in genes, we compared our SNVs to those found in a recent study of clonal organoids formed from mouse prostate, stomach, small intestine and bowel (Behjati et al., 2014). We find that the MT neuron SNVs are enriched in genes ($p=0.0039$, Fisher's Exact, **Figure 5.10d**) and genes highly expressed in MT neurons ($p=0.025$, Fisher's Exact, **Figure 5.10f**) when compared to those found in the mouse organoid study. We further found that the SNVs from small intestine organoids are depleted in genes highly expressed in Lgr5 positive small intestine stem cells relative to MT neuron SNVs ($p=7.06 \times 10^{-4}$, Fisher's Exact, **Figure 5.10g**).

We assessed whether MT neuron derived SNVs were enriched in genes that could impact neuronal function as defined by GO terms. We composed a list of gene names for the 164 autosomal high confidence MT neuron SNVs that fall in refSeq transcripts, and submitted it to DAVID (Huang da et al., 2009a, b) (http://david.abcc.ncifcrf.gov/). We then specified the *Mus musculus* background, de-selected all annotations except for the three default GO annotation categories (GOTERM_BP_FAT, GOTERM_CC_FAT, GOTERM_MF_FAT) and created a Functional Annotation Chart that was then downloaded and inserted into **Table S10**. We followed the same procedure for the 2250 mouse organoid SNVs and inserted the resulting GO terms into **Table S11**. For the MT neuron SNVs, the top 6 GO terms, and 11 of 59 or 19% of all GO terms are neuronal. In contrast, the 2,250 mouse organoid SNVs fall into genes in which only 51 of 488 or 11% of all associated GO terms are neuronal, with the highest ranking neuronal GO term 14th on the list (**Figure 5.10e**).

Finally, we examined whether our high confidence SNVs were occurring in clusters to determine hotspots, or evidence for other mutational processes that might create multiple SNVs at once such as kataegis or base excision repair. Clusters were identified using bedtools closest (-N –d –t first). There is one cluster of two adjoining SNVs, one with 1bp between them, and one with three SNVs separated by 49bp and 199bp. The next closest pair of SNVs are more than 40Kbp apart. To assess whether these SNVs were likely created simultaneously by a single process, we determined if the SNVs in each cluster appear on the same parental chromosome, (i.e., haplotype). For the first two clusters, as well as the two SNVs in the third cluster that are 49bp apart, we were able to establish that they occur on the same parental chromosome by visual inspection of the reads containing them. In all cases, a read that spanned the region of the two SNVs contained either both SNVs or neither, indicating that they are indeed on the same parental chromosome.

A1.22 Haplotype Determination by Amplicon Cloning - Scripps

The two SNVs separated by 199bp were too far apart to use this technique. Therefore, we designed PCR primers to generate a single amplicon containing all three mutations, (forward primer: AGAAACAAATGCTTAGGGTTGGGTTC, reverse primer: GACTGTGTTCTGGGAGTTCATCTACAAAC) and cloned the resulting PCR product into pCR™-Blunt II-TOPO® vector using the Zero Blunt® TOPO® PCR Cloning Kit. We performed Sanger sequencing on 30 independent TOPO clones and found that 16 clones contained all three mutations and 13 clones contained no mutations. A final clone contained two of three mutations that likely resulted from a rare template-switching event which are observed during PCR amplification (Odelberg et al., 1995).

A1.23 MT Neuron RNA-Seq Sample Preparation - Scripps

Mitral and tufted cells were dissociated from Pcdh21/Cre-Ai9 mice as for nuclear transfer (see above) and flow sorted using the MoFlo® Astrios™ (Beckman Coulter). Ten minutes prior to sorting, DAPI (1 μM) and DRAQ5 (BioStatus DR50050, 1 μM) were added to the cell suspension. Dead cells and debris were first gated out using side and forward scatter. Objects were identified as cells by positive staining for DRAQ5, and as live cells by the absence of DAPI staining. From this population, MT neurons were identified by TdTomato expression, and sorted directly into TRIzol® LS Reagent (Life Technologies). The following lasers were used: DRAQ5 (642nm laser), DAPI (405nm laser), TdTomato (561nm laser). Three biological replicates were collected on independent days using this method.

Prior to RNA extraction, all samples were adjusted to 1.75 mL TRIzol® LS and 1 ug of linear acrylamide (Ambion AM9520) was added. RNA was extracted using Direct-zol™ RNA

MiniPrep (Zymo Research) using their Zymo-Spin™ IC columns for low amounts of RNA. The optional in-column DNase treatment was included. RNA was eluted in 10ul of water and RNA quality was assessed using Agilent RNA 6000 Pico Kit. All RNA samples had RIN scores >7.5.

Prior to sequencing, 10ng of RNA from each biological replicate was amplified using SMARTer® Ultra™ Low Input RNA for Illumina® Sequencing – HV (Clontech Laboratories, Inc.). Amplified cDNA was checked for quality using High Sensitivity DNA Kit (Agilent Technologies) and acoustically sheared using the Covaris system. Sequencing libraries were prepped from sheared cDNA using NEBNext® Ultra™ DNA Library Prep Kit for Illumina® and sequenced on an Illumina HiSeq.

A1.24 RNA-Seq Analyses

We analyzed the RNA-Seq data using TopHat (Kim et al., 2013) v2.0.10 and Cufflinks (Trapnell et al., 2010) v2.0.2 from the Tuxedo suite. We first created the genome and annotation indexes by downloading the mm9 annotation data (mm9/Mus_musculus_UCSC_mm9.tar.gz) from (ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com) and using gtf_to_fasta. Each of the three MT neuron samples were then aligned separately using bowtie through the tophat interface (-r 160 –libarary-type fr-unstranded –coverage-search –b2-sensitive). The BAM files for the reads from the three MT neuron samples were then merged using samtools merge. We then assembled the reads and determined expression levels for the combined MT neuron samples using cufflinks (--library-type fr-unstranded --multi-read-correct –max-intron-length 500000). The resulting "genes.fpkm_tracking" file was converted to bed format for further processing. Finally, we considered those genes with greater than the median expression level of ~0.78 to be "highly expressed".

RNA-Seq datasets from 3 Lgr5+ small intestine stem cells with accession ids ERX421326, ERX421327 and ERX421329 were downloaded from (http://www.ncbi.nlm.nih.gov/sra/) in SRA format. From these files, fastq files of the RNA-Seq reads were extracted using fastq-dump v2.1.18 (--gzip) from the SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/). The reads were processed as above, with a resulting median expression level of ~0.69.

Bibliography

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M.S., Tomasini, L., Ferrandino, A.F., Rosenberg Belmaker, L.A., Szekely, A., Wilson, M., *et al.* (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature *492*, 438-442.

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res *21*, 974-984.

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P., and Kawashima, E. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic acids research *28*, E87.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology *12*, R18.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., *et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415-421.

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

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Alzualde, A., Moreno, F., Martinez-Lage, P., Ferrer, I., Gorostidi, A., Otaegui, D., Blazquez, L., Atares, B., Cardoso, S., Martinez de Pancorbo, M., Juste, R., Rodriguez-Martinez, A.B., Indakoetxea, B., and Lopez de Munain, A. (2010). Somatic mosaicism in a case of apparently sporadic Creutzfeldt-Jakob disease carrying a de novo D178N mutation in the PRNP gene. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics *153B*, 1283-1291.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003-1007.

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., *et al.* (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Ball, M.P., Thakuria, J.V., Zaranek, A.W., Clegg, T., Rosenbaum, A.M., Wu, X., Angrist, M., Bhak, J., Bobe, J., Callow, M.J., *et al.* (2012). A public resource facilitating clinical use of genomes. Proceedings of the National Academy of Sciences of the United States of America *109*, 11920-11927.

Ballif, B.C., Rorem, E.A., Sundin, K., Lincicum, M., Gaskin, S., Coppinger, J., Kashork, C.D., Shaffer, L.G., and Bejjani, B.A. (2006). Detection of low-level mosaicism by array CGH in routine diagnostic specimens. American journal of medical genetics Part A *140*, 2757-2767.

Bartenhagen, C., and Dugas, M. (2013). RSVSim: an R/Bioconductor package for the simulation of structural variations. Bioinformatics *29*, 1679-1681.

Baserga, R. (1985). The Biology of Cell Reproduction (Harvard University Press).

Beale, R.C., Petersen-Mahrt, S.K., Watt, I.N., Harris, R.S., Rada, C., and Neuberger, M.S. (2004). Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. Journal of molecular biology *337*, 585-596.

Beck, J.A., Poulter, M., Campbell, T.A., Uphill, J.B., Adamson, G., Geddes, J.F., Revesz, T., Davis, M.B., Wood, N.W., Collinge, J., and Tabrizi, S.J. (2004). Somatic and germline mosaicism in sporadic early-onset Alzheimer's disease. Human molecular genetics *13*, 1219-1224.

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G., *et al.* (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature *advance online publication*.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., and Canaider, S. (2013). An estimation of the number of cells in the human body. Annals of human biology *40*, 463-471.

Bielanska, M., Tan, S.L., and Ao, A. (2002). Chromosomal mosaicism throughout human preimplantation development in vitro: incidence, type, and relevance to embryo outcome. Human reproduction (Oxford, England) *17*, 413-419.

Blanchart, A., De Carlos, J.A., and Lopez-Mascaraque, L. (2006). Time frame of mitral cell development in the mice olfactory bulb. The Journal of comparative neurology *496*, 529-543.

Boland, M.J., Hazen, J.L., Nazor, K.L., Rodriguez, A.R., Gifford, W., Martin, G., Kupriyanov, S., and Baldwin, K.K. (2009). Adult mice generated from induced pluripotent stem cells. Nature *461*, 91-94.

Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. Journal of cell science *121 Suppl 1*, 1-84.

Brewer, G.J., and Torricelli, J.R. (2007). Isolation and culture of adult neurons and neurospheres. Nature protocols *2*, 1490-1498.

Broad Institute. Picard Tools.

Bruder, C.E., Piotrowski, A., Gijsbers, A.A., Andersson, R., Erickson, S., Diaz de Stahl, T., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A.*, et al.* (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. American journal of human genetics *82*, 763-771.

Brush, S.G. (2002). How theories became knowledge: Morgan's chromosome theory of heredity in America and Britain. Journal of the history of biology *35*, 471-535.

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

Burrell, R.A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. Nature *501*, 338-345.

Bushman, D.M., and Chun, J. (2013). The genomically mosaic brain: aneuploidy and more in neural diversity and disease. Seminars in cell & developmental biology *24*, 357-369.

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

Campbell, P.J., Pleasance, E.D., Stephens, P.J., Dicks, E., Rance, R., Goodhead, I., Follows, G.A., Green, A.R., Futreal, P.A., and Stratton, M.R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proceedings of the National Academy of Sciences of the United States of America *105*, 13081-13086.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061-1068.

Carlson, E.A. (2004). Mendel's Legacy: The Origin of Classical Genetics (Cold Spring Harbor Laboratory Press).

Celniker, S.E., and Rubin, G.M. (2003). The Drosophila melanogaster genome. Annual review of genomics and human genetics *4*, 89-117.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P.*, et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods *6*, 677-681.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R.*, et al.* (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research *40*, D700-705.

Chiang, C., Jacobsen, J.C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills, R.E., Kirby, A., Lindgren, A.M., Rudiger, S.R.*, et al.* (2012). Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nature genetics *44*, 390-397, S391.

Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2014). SpeedSeq: Ultra-fast personal genome analysis and interpretation. bioRxiv.

Chun, J., and Schatz, D.G. (1999). Rearranging views on neurogenesis: neuronal death in the absence of DNA end-joining proteins. Neuron *22*, 7-10.

Chun, J.J., Schatz, D.G., Oettinger, M.A., Jaenisch, R., and Baltimore, D. (1991). The recombination activating gene-1 (RAG-1) transcript is present in the murine central nervous system. Cell *64*, 189-200.

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

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80-92.

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nature nanotechnology *4*, 265-270.

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

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

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nature reviews Genetics *10*, 691-703.

Cotterman, C.W. (1956). Somatic mosaicism for antigen A2. Acta genetica et statistica medica *6*, 520-521.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Marchetto, M.C., Muotri, A.R., Mu, Y., Carson, C.T., Macia, A., Moran, J.V., and Gage, F.H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proceedings of the National Academy of Sciences of the United States of America *108*, 20382-20387.

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

De, S. (2011). Somatic mosaicism in healthy human tissues. Trends in genetics : TIG *27*, 217-223.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J.*, 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.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics *43*, 491-498.

Di Noia, J.M., and Neuberger, M.S. (2007). Molecular mechanisms of antibody somatic hypermutation. Annual review of biochemistry *76*, 1-22.

Dizdaroglu, M. (2012). Oxidatively induced DNA damage: mechanisms, repair and disease. Cancer letters *327*, 26-47.

Douville, R., Liu, J., Rothstein, J., and Nath, A. (2011). Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Annals of neurology *69*, 141-151.

Eggan, K., Baldwin, K., Tackett, M., Osborne, J., Gogos, J., Chess, A., Axel, R., and Jaenisch, R. (2004). Mice cloned from olfactory sensory neurons. Nature *428*, 44-49.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Erickson, R.P. (2010). Somatic gene mutation and human disease other than cancer: an update. Mutation research *705*, 96-106.

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

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J., and Walsh, C.A. (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49-59.

Farrar, M. (2007). Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics *23*, 156-161.

Faust, G.G., and Hall, I.M. (2012). YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinformatics *28*, 2417-2424.

Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics *30*, 2503-2505.

Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. Genomics, proteomics & bioinformatics.

Ferragina, P., and Manzini, G. (2000). Opportunistic data structures with applications. In Proceedings of the 41st Annual Symposium on Foundations of Computer Science (IEEE Computer Society), pp. 390.

Ferreira, T., Wilson, S.R., Choi, Y.G., Risso, D., Dudoit, S., Speed, T.P., and Ngai, J. (2014). Silencing of odorant receptor genes by G protein betagamma signaling ensures the expression of one odorant receptor per olfactory sensory neuron. Neuron *81*, 847-859.

Flores, M., Morales, L., Gonzaga-Jauregui, C., Dominguez-Vidana, R., Zepeda, C., Yanez, O., Gutierrez, M., Lemus, T., Valle, D., Avila, M.C.*, et al.* (2007). Recurrent DNA inversion rearrangements in the human genome. Proceedings of the National Academy of Sciences of the United States of America *104*, 6099-6106.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J.W., Campbell, P.J., Stratton, M.R., and Futreal, P.A. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research *39*, D945-950.

Forsberg, L.A., Rasi, C., Razzaghian, H.R., Pakalapati, G., Waite, L., Thilbeault, K.S., Ronowicz, A., Wineinger, N.E., Tiwari, H.K., Boomsma, D.*, et al.* (2012). Age-related somatic structural changes in the nuclear genome of human blood cells. American journal of human genetics *90*, 217-228.

Frank, K.M., Sekiguchi, J.M., Seidl, K.J., Swat, W., Rathbun, G.A., Cheng, H.L., Davidson, L., Kangaloo, L., and Alt, F.W. (1998). Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. Nature *396*, 173-177.

Freed, D., Stevens, E.L., and Pevsner, J. (2014). Somatic mosaicism in the human genome. Genes *5*, 1064-1094.

Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U., and Shapiro, E. (2005). Genomic variability within an organism exposes its cell lineage tree. PLoS computational biology *1*, e50.

Gao, Y., Sun, Y., Frank, K.M., Dikkes, P., Fujiwara, Y., Seidl, K.J., Sekiguchi, J.M., Rathbun, G.A., Swat, W., Wang, J.*, et al.* (1998). A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. Cell *95*, 891-902.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. In ArXiv e-prints, pp. 3907.

Ghosh, S., Larson, S.D., Hefzi, H., Marnoy, Z., Cutforth, T., Dokka, K., and Baldwin, K.K. (2011). Sensory maps in the olfactory cortex defined by long-range viral tracing of single neurons. Nature *472*, 217-220.

Gole, J., Gore, A., Richards, A., Chiu, Y.J., Fung, H.L., Bushman, D., Chiang, H.I., Chun, J., Lo, Y.H., and Zhang, K. (2013). Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. Nature biotechnology *31*, 1126-1132.

Gollob, M.H., Jones, D.L., Krahn, A.D., Danis, L., Gong, X.Q., Shao, Q., Liu, X., Veinot, J.P., Tang, A.S., Stewart, A.F.*, et al.* (2006). Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. The New England journal of medicine *354*, 2677-2688.

Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E.*, et al.* (2011). Somatic coding mutations in human induced pluripotent stem cells. Nature *471*, 63-67.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. Journal of molecular biology *162*, 705-708.

Gottlieb, B., Beitel, L.K., and Trifiro, M.A. (2001). Somatic mosaicism and variable expressivity. Trends in genetics : TIG *17*, 79-82.

Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., *et al.* (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nature neuroscience *17*, 215-222.

Guo, J.U., Su, Y., Zhong, C., Ming, G.L., and Song, H. (2011). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell *145*, 423-434.

Han, J.S., and Shao, S. (2012). Circular retrotransposition products generated by a LINE retrotransposon. Nucleic acids research *40*, 10866-10877.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nature reviews Genetics *10*, 551-564.

Hazen, J.L., Faust, G.G., Rodriguez, A.R., Ferguson, W., Shumilina, S., Clark, R.A., Boland, M.J., Martin, G., Chubukov, P., Tsunemoto, R., Torkamani, A., Kupriyanov, S., Hall, I.M., and Baldwin, K.K. (2015). Cloning mice from neurons reveals extensive genome diversity in the adult brain. Unpublished.

Heid, C.A., Stevens, J., Livak, K.J., and Williams, P.M. (1996). Real time quantitative PCR. Genome Res *6*, 986-994.

Hercus, C. (2009). Novoalign (Unpublished).

Hinds, J.W. (1968a). Autoradiographic study of histogenesis in the mouse olfactory bulb. I. Time of origin of neurons and neuroglia. The Journal of comparative neurology *134*, 287-304.

Hinds, J.W. (1968b). Autoradiographic study of histogenesis in the mouse olfactory bulb. II. Cell proliferation and migration. The Journal of comparative neurology *134*, 305-322.

Hochedlinger, K., and Jaenisch, R. (2002). Monoclonal mice generated by nuclear transfer from mature B and T donor cells. Nature *415*, 1035-1038.

Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B., *et al.* (2014). Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res *24*, 733-742.

Hormozdiari, F., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res *19*, 1270-1278.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., *et al.* (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell *148*, 873-885.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research *37*, 1-13.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols *4*, 44-57.

Imamura, F., Ayoub, A.E., Rakic, P., and Greer, C.A. (2011). Timing of neurogenesis is a determinant of olfactory circuitry. Nature neuroscience *14*, 331-337.

International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789-796.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., *et al.* (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285-299.

Iourov, I.Y., Soloviev, I.V., Vorsanova, S.G., Monakhov, V.V., and Yurov, Y.B. (2005). An approach for quantitative assessment of fluorescence in situ hybridization (FISH) signals for applied human molecular cytogenetics. The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society *53*, 401-408.

Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., Hutchinson, A., Deng, X., Liu, C., Horner, M.J., *et al.* (2012). Detectable clonal mosaicism and its relationship to aging and cancer. Nature genetics *44*, 651-658.

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

Jeong, B.H., Lee, Y.J., Carp, R.I., and Kim, Y.S. (2010). The prevalence of human endogenous retroviruses in cerebrospinal fluids from patients with sporadic Creutzfeldt-Jakob disease. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology *47*, 136-142.

Ji, J., Ng, S.H., Sharma, V., Neculai, D., Hussein, S., Sam, M., Trinh, Q., Church, G.M., McPherson, J.D., Nagy, A., and Batada, N.N. (2012). Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. Stem cells (Dayton, Ohio) *30*, 435-440.

Johnson, L.A. (1995). Sex preselection by flow cytometric separation of X and Y chromosome-bearing sperm based on DNA difference: a review. Reproduction, fertility, and development *7*, 893-903.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research *110*, 462-467.

Kaneko, H., Dridi, S., Tarallo, V., Gelfand, B.D., Fowler, B.J., Cho, W.G., Kleinman, M.E., Ponicsan, S.L., Hauswirth, W.W., Chiodo, V.A.*, et al.* (2011). DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration. Nature *471*, 325-330.

Kang, H., and Roh, S. (2011). Extended exposure to trichostatin A after activation alters the expression of genes important for early development in nuclear transfer murine embryos. The Journal of veterinary medical science / the Japanese Society of Veterinary Science *73*, 623-631.

Kano, H., Godoy, I., Courtney, C., Vetter, M.R., Gerton, G.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes & development *23*, 1303-1312.

Kawase, E., Yamazaki, Y., Yagi, T., Yanagimachi, R., and Pedersen, R.A. (2000). Mouse embryonic stem (ES) cell lines established from neuronal cell-derived cloned blastocysts. Genesis (New York, NY : 2000) *28*, 156-163.

Kazazian, H.H., Jr. (2011). Mobile DNA transposition in somatic cells. BMC biology *9*, 62.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M.*, et al.* (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289-294.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

Kingsbury, M.A., Friedman, B., McConnell, M.J., Rehen, S.K., Yang, A.H., Kaushal, D., and Chun, J. (2005). Aneuploid neurons are functionally active and integrated into brain circuitry. Proceedings of the National Academy of Sciences of the United States of America *102*, 6143-6147.

Kishigami, S., Mizutani, E., Ohta, H., Hikichi, T., Thuan, N.V., Wakayama, S., Bui, H.T., and Wakayama, T. (2006a). Significant improvement of mouse cloning technique by treatment with trichostatin A after somatic nuclear transfer. Biochemical and biophysical research communications *340*, 183-189.

Kishigami, S., Wakayama, S., Thuan, N.V., Ohta, H., Mizutani, E., Hikichi, T., Bui, H.T., Balbach, S., Ogura, A., Boiani, M., and Wakayama, T. (2006b). Production of cloned mice by somatic cell nuclear transfer. Nature protocols *1*, 125-138.

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

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

Kuwabara, T., Hsieh, J., Muotri, A., Yeo, G., Warashina, M., Lie, D.C., Moore, L., Nakashima, K., Asashima, M., and Gage, F.H. (2009). Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nature neuroscience *12*, 1097-1105.

Laerum, O.D., and Farsund, T. (1981). Clinical application of flow cytometry: a review. Cytometry *2*, 1-13.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*, 311-317.

Lathe, R., and Harris, A. (2009). Differential display detects host nucleic acid motifs altered in scrapie-infected brain. Journal of molecular biology *392*, 813-822.

Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., *et al.* (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nature genetics *44*, 642-650.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., *et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

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

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., *et al.* (2012a). Landscape of somatic retrotransposition in human cancers. Science *337*, 967-971.

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

Lee, W.P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., and Marth, G.T. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PloS one *9*, e90581.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In ArXiv e-prints, pp. 3997.

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

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589-595.

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

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Briefings in bioinformatics *11*, 473-483.

Li, J., Ishii, T., Feinstein, P., and Mombaerts, P. (2004). Odorant receptor gene choice is reset by nuclear transfer from mouse olfactory sensory neurons. Nature *428*, 393-399.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009b). SNP detection for massively parallel whole-genome resequencing. Genome Res *19*, 1124-1132.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713-714.

Li, W., Jin, Y., Prazak, L., Hammell, M., and Dubnau, J. (2012). Transposable elements in TDP-43-mediated neurodegenerative disorders. PloS one *7*, e44099.

Li, W., Prazak, L., Chatterjee, N., Gruninger, S., Krug, L., Theodorou, D., and Dubnau, J. (2013). Activation of transposable elements during aging and neuronal decline in Drosophila. Nature neuroscience *16*, 529-531.

Li, Z., Lu, H., Yang, W., Yong, J., Zhang, Z.N., Zhang, K., Deng, H., and Xu, Y. (2014). Mouse SCNT ESCs Have Lower Somatic Mutation Load Than Syngeneic iPSCs. Stem cell reports *2*, 399-405.

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

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America *107*, 961-968.

Ma, H., Morey, R., O'Neil, R.C., He, Y., Daughtry, B., Schultz, M.D., Hariharan, M., Nery, J.R., Castanon, R., Sabatini, K., *et al.* (2014). Abnormalities in human pluripotent cells due to reprogramming mechanisms. Nature *511*, 177-183.

Madisen, L., Zwingman, T.A., Sunkin, S.M., Oh, S.W., Zariwala, H.A., Gu, H., Ng, L.L., Palmiter, R.D., Hawrylycz, M.J., Jones, A.R., Lein, E.S., and Zeng, H. (2010). A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. Nature neuroscience *13*, 133-140.

Maher, C.A., and Wilson, R.K. (2012). Chromothripsis and human disease: piecing together the shattering process. Cell *148*, 29-32.

Mair, R.G., and Gesteland, R.C. (1982). Response properties of mitral cells in the olfactory bulb of the neonatal rat. Neuroscience *7*, 3117-3125.

Makino, H., Yamazaki, Y., Hirabayashi, T., Kaneko, R., Hamada, S., Kawamura, Y., Osada, T., Yanagimachi, R., and Yagi, T. (2005). Mouse embryos and chimera cloned from neural cells in the postnatal cerebral cortex. Cloning and stem cells *7*, 45-61.

Malhotra, A., Lindberg, M., Faust, G.G., Leibowitz, M.L., Clark, R.A., Layer, R.M., Quinlan, A.R., and Hall, I.M. (2013). Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. Genome Res *23*, 762-776.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

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

Mardis, E.R. (2013). Next-generation sequencing platforms. Annual review of analytical chemistry (Palo Alto, Calif) *6*, 287-303.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.*, et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Marsh, M., Tu, O., Dolnik, V., Roach, D., Solomon, N., Bechtol, K., Smietana, P., Wang, L., Li, X., Cartwright, P., Marks, A., Barker, D., Harris, D., and Bashkin, J. (1997). High-throughput DNA sequencing on a capillary array electrophoresis system. Journal of capillary electrophoresis *4*, 83-89.

McClintock, B. (1951). Chromosome organization and genic expression. Cold Spring Harbor symposia on quantitative biology *16*, 13-47.

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

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A., and Fiston-Lavier, A.S. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PloS one *9*, e106689.

McCulloch, S.D., and Kunkel, T.A. (2008). The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell research *18*, 148-161.

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

Meissner, A., Eminli, A., and Jaenisch, R. (2011). Stem Cells in Regenerative Medicine (Methods in Molecular Biology) (Humana Press).

Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A.*, et al.* (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell *151*, 1431-1442.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K.*, et al.* (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59-65.

Misra, S., Agrawal, A., Liao, W.K., and Choudhary, A. (2011). Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing. Bioinformatics *27*, 189-195.

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

Miyamichi, K., Amat, F., Moussavi, F., Wang, C., Wickersham, I., Wall, N.R., Taniguchi, H., Tasic, B., Huang, Z.J., He, Z., Callaway, E.M., Horowitz, M.A., and Luo, L. (2011). Cortical representations of olfactory input by trans-synaptic tracing. Nature *472*, 191-196.

Mombaerts, P. (2004). Odorant receptor gene choice in olfactory sensory neurons: the one receptor-one neuron hypothesis revisited. Current opinion in neurobiology *14*, 31-36.

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

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

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

Muotri, A.R., Zhao, C., Marchetto, M.C., and Gage, F.H. (2009). Environmental influence on L1 retrotransposons in the adult hippocampus. Hippocampus *19*, 1002-1007.

Myers, E.W., and Miller, W. (1988). Optimal alignments in linear space. Computer applications in the biosciences : CABIOS *4*, 11-17.

Nag, A., Bochukova, E.G., Kremeyer, B., Campbell, D.D., Muller, H., Valencia-Duarte, A.V., Cardona, J., Rivas, I.C., Mesa, S.C., Cuartas, M.*, et al.* (2013). CNV analysis in Tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1. PloS one *8*, e59061.

Nagai, Y., Sano, H., and Yokoi, M. (2005). Transgenic expression of Cre recombinase in mitral/tufted cells of the olfactory bulb. Genesis (New York, NY : 2000) *43*, 12-16.

Nagy, A., Gocza, E., Diaz, E.M., Prideaux, V.R., Ivanyi, E., Markkula, M., and Rossant, J. (1990). Embryonic stem cells alone are able to support fetal development in the mouse. Development (Cambridge, England) *110*, 815-821.

Nagy, A., Rossant, J., Nagy, R., Abramow-Newerly, W., and Roder, J.C. (1993). Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America *90*, 8424-8428.

Nakayama, T., Fujiwara, H., Tastumi, K., Fujita, K., Higuchi, T., and Mori, T. (1998). A new assisted hatching technique using a piezo-micromanipulator. Fertility and sterility *69*, 784-788.

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

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V.*, et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242-245.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A.*, et al.* (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979-993.

Ning, L., Liu, G., Li, G., Hou, Y., Tong, Y., and He, J. (2014). Current challenges in the bioinformatics of single cell genomics. Frontiers in oncology *4*, 7.

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001). SSAHA: a fast search method for large DNA databases. Genome Res *11*, 1725-1729.

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.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D.*, et al.* (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246-250.

Odelberg, S.J., Weiss, R.B., Hata, A., and White, R. (1995). Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic acids research *23*, 2049-2057.

Ogura, A., Inoue, K., and Wakayama, T. (2013). Recent advancements in cloning by somatic cell nuclear transfer. Philosophical transactions of the Royal Society of London Series B, Biological sciences *368*, 20110329.

Osada, T., Kusakabe, H., Akutsu, H., Yagi, T., and Yanagimachi, R. (2002). Adult murine neurons: their chromatin and chromosome changes and failure to support embryonic development as revealed by nuclear transfer. Cytogenetic and genome research *97*, 7-12.

Osada, T., Tamamaki, N., Song, S.Y., Kakazu, N., Yamazaki, Y., Makino, H., Sasaki, A., Hirayama, T., Hamada, S., Nave, K.A., Yanagimachi, R., and Yagi, T. (2005). Developmental pluripotency of the nuclei of neurons in the cerebral cortex of juvenile mice. The Journal of neuroscience : the official journal of the Society for Neuroscience *25*, 8368-8374.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America *85*, 2444-2448.

Perrat, P.N., DasGupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., and Waddell, S. (2013). Transposition-driven genomic heterogeneity in the Drosophila brain. Science *340*, 91-95.

Peterson, S.E., Westra, J.W., Rehen, S.K., Young, H., Bushman, D.M., Paczkowski, C.M., Yung, Y.C., Lynch, C.L., Tran, H.T., Nickey, K.S., *et al.* (2011). Normal human pluripotent stem cell lines exhibit pervasive mosaic aneuploidy. PloS one *6*, e23018.

Pettersson, E., Lundeberg, J., and Ahmadian, A. (2009). Generations of sequencing technologies. Genomics *93*, 105-111.

Pfeifer, G.P., Denissenko, M.F., Olivier, M., Tretyakova, N., Hecht, S.S., and Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene *21*, 7435-7451.

Pinkel, D., and Albertson, D.G. (2005). Comparative genomic hybridization. Annual review of genomics and human genetics *6*, 331-354.

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

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R., *et al.* (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191-196.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C., *et al.* (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184-190.

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

Quick, J., Quinlan, A.R., and Loman, N.J. (2014). A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. GigaScience *3*, 22.

Quick, J., Quinlan, A.R., and Loman, N.J. (2015). Erratum: A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. GigaScience *4*, 6.

Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C., and Hall, I.M. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Research *20*, 623-635.

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 *28*, 43-53.

Rausch, T., Jones, D.T., Zapatka, M., Stutz, A.M., Zichner, T., Weischenfeldt, J., Jager, N., Remke, M., Shih, D., Northcott, P.A., *et al.* (2012a). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell *148*, 59-71.

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

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

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

Richardson, S.R., Morell, S., and Faulkner, G.J. (2014). L1 retrotransposons and somatic mosaicism in the brain. Annual review of genetics *48*, 1-27.

Riviere, J.B., Mirzaa, G.M., O'Roak, B.J., Beddaoui, M., Alcantara, D., Conway, R.L., St-Onge, J., Schwartzentruber, J.A., Gripp, K.W., Nikkel, S.M., *et al.* (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nature genetics *44*, 934-940.

Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013). The advantages of SMRT sequencing. Genome biology *14*, 405.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M.*, et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348-352.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Biotechnology (Reading, Mass) *24*, 104-108.

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

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

Sebat, J., Levy, D.L., and McCarthy, S.E. (2009). Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. Trends in genetics : TIG *25*, 528-535.

Sheaffer, K.L., Kim, R., Aoki, R., Elliott, E.N., Schug, J., Burger, L., Schubeler, D., and Kaestner, K.H. (2014). DNA methylation is required for the control of stem cell differentiation in the small intestine. Genes & development *28*, 652-664.

Shibata, Y., Kumar, P., Layer, R., Willcox, S., Gagan, J.R., Griffith, J.D., and Dutta, A. (2012). Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. Science *336*, 82-86.

Shin, S.C., Ahn do, H., Kim, S.J., Lee, H., Oh, T.J., Lee, J.E., and Park, H. (2013). Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PloS one *8*, e68824.

Sindi, S.S., Onal, S., Peng, L.C., Wu, H.T., and Raphael, B.J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. Genome biology *13*, R22.

Smit, A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature *321*, 674-679.

Sosulski, D.L., Bloom, M.L., Cutforth, T., Axel, R., and Datta, S.R. (2011). Distinct representations of olfactory information in different cortical centres. Nature *472*, 213-216.

Specchia, V., Piacentini, L., Tritto, P., Fanti, L., D'Alessandro, R., Palumbo, G., Pimpinelli, S., and Bozzetti, M.P. (2010). Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. Nature *463*, 662-665.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A.*, et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27-40.

Suberbielle, E., Sanchez, P.E., Kravitz, A.V., Wang, X., Ho, K., Eilertson, K., Devidze, N., Kreitzer, A.C., and Mucke, L. (2013). Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-beta. Nature neuroscience *16*, 613-621.

Tan, H., Qurashi, A., Poidevin, M., Nelson, D.L., Li, H., and Jin, P. (2012). Retrotransposon activation contributes to fragile X premutation rCGG-mediated neurodegeneration. Human molecular genetics *21*, 57-65.

Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjold, M., Ponder, B.A., and Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics *13*, 718-725.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.

Upton, Kyle R., Gerhardt, Daniel J., Jesuadian, J.S., Richardson, Sandra R., Sánchez-Luque, Francisco J., Bodea, Gabriela O., Ewing, Adam D., Salvador-Palomeque, C., van der Knaap, Marjo S., Brennan, Paul M., Vanderver, A., and Faulkner, Geoffrey J. (2015). Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell *161*, 228-239.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J.*, et al.* (2002). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In Current Protocols in Bioinformatics (John Wiley & Sons, Inc.).

Vanneste, E., Voet, T., Le Caignec, C., Ampe, M., Konings, P., Melotte, C., Debrock, S., Amyere, M., Vikkula, M., Schuit, F.*, et al.* (2009). Chromosome instability is common in human cleavage-stage embryos. Nature medicine *15*, 577-583.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Vorsanova, S.G., Yurov, Y.B., and Iourov, I.Y. (2010). Human interphase chromosomes: a review of available molecular cytogenetic technologies. Molecular cytogenetics *3*, 1.

Waisfisz, Q., Morgan, N.V., Savino, M., de Winter, J.P., van Berkel, C.G., Hoatlin, M.E., Ianzano, L., Gibson, R.A., Arwert, F., Savoia, A., Mathew, C.G., Pronk, J.C., and Joenje, H. (1999). Spontaneous functional correction of homozygous fanconi anaemia alleles reveals novel mechanistic basis for reverse mosaicism. Nature genetics *22*, 379-383.

Wakayama, S., Ohta, H., Kishigami, S., Thuan, N.V., Hikichi, T., Mizutani, E., Miyake, M., and Wakayama, T. (2005). Establishment of male and female nuclear transfer embryonic stem cell lines from different mouse strains and tissues. Biology of reproduction *72*, 932-936.

Wakayama, T., Perry, A.C., Zuccotti, M., Johnson, K.R., and Yanagimachi, R. (1998). Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. Nature *394*, 369-374.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P.*, et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Watson, I.R., Takahashi, K., Futreal, P.A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. Nature reviews Genetics *14*, 703-718.

Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T.*, et al.* (2008). Association between microdeletion and microduplication at 16p11.2 and autism. The New England journal of medicine *358*, 667-675.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J.*, et al.* (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264-278.

Westra, J.W., Peterson, S.E., Yung, Y.C., Mutoh, T., Barral, S., and Chun, J. (2008). Aneuploid mosaicism in the developing and adult cerebellar cortex. The Journal of comparative neurology *507*, 1944-1951.

Westra, J.W., Rivera, R.R., Bushman, D.M., Yung, Y.C., Peterson, S.E., Barral, S., and Chun, J. (2010). Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. The Journal of comparative neurology *518*, 3981-4000.

Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A., and Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nature genetics *40*, 880-885.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H.*, et al.* (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, 886-895.

Yagi, T. (2003). Diversity of the cadherin-related neuronal receptor/protocadherin family and possible DNA rearrangement in the brain. Genes to cells : devoted to molecular & cellular mechanisms *8*, 1-8.

Yalcin, B., Wong, K., Bhomra, A., Goodson, M., Keane, T.M., Adams, D.J., and Flint, J. (2012). The fine-scale architecture of structural variants in 17 mouse genomes. Genome biology *13*, R18.

Yamazaki, Y., Makino, H., Hamaguchi-Hamada, K., Hamada, S., Sugino, H., Kawase, E., Miyata, T., Ogawa, M., Yanagimachi, R., and Yagi, T. (2001). Assessment of the developmental totipotency of neural cells in the cerebral cortex of mouse embryo by nuclear transfer. Proceedings of the National Academy of Sciences of the United States of America *98*, 14022-14026.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (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.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res *19*, 1586-1592.

Young, J.M., and Trask, B.J. (2002). The sense of smell: genomics of vertebrate odorant receptors. Human molecular genetics *11*, 1153-1160.

Young, M.A., Larson, D.E., Sun, C.W., George, D.R., Ding, L., Miller, C.A., Lin, L., Pawlik, K.M., Chen, K., Fan, X., *et al.* (2012). Background mutations in parental cells account for most of the genetic heterogeneity of induced pluripotent stem cells. Cell stem cell *10*, 570-582.

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

Yurov, Y.B., Iourov, I.Y., Vorsanova, S.G., Liehr, T., Kolotii, A.D., Kutsev, S.I., Pellestor, F., Beresheva, A.K., Demidova, I.A., Kravets, V.S., Monakhov, V.V., and Soloviev, I.V. (2007). Aneuploidy and confined chromosomal mosaicism in the developing human brain. PloS one *2*, e558.

Zhang, J., Wheeler, D.A., Yakub, I., Wei, S., Sood, R., Rowe, W., Liu, P.P., Gibbs, R.A., and Buetow, K.H. (2005). SNPdetector: a software tool for sensitive and accurate SNP detection. PLoS computational biology *1*, e53.

Zhang, Z., Berman, P., and Miller, W. (1998). Alignments without low-scoring regions. Journal of computational biology : a journal of computational molecular cell biology *5*, 197-210.

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