UNIVERSITY OF CALIFORNIA SAN DIEGO

Assessing the Extent and Impact of Mutations in Human Induced Pluripotent Stem Cells and Young vs Aged Single Mouse Neurons

A dissertation submitted in satisfaction of the requirements for the degree Doctor of Philosophy

in

Biology

by

Michael A. Duran

Committee in charge:

Professor Kristin Baldwin, Chair
Professor Yimin Zou, Co-Chair
Professor Brenda Bloodgood
Professor Joseph Gleeson
Professor Yang Xu

2019

The Dissertation of Michael A. Duran. is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-chair

_____
Chair

University of California San Diego
2019

DEDICATION

This dissertation is dedicated to my parents, Mike and Kim, my sister Jordan, and the Lord God, all of whom have filled my life with love and laughter

EPIGRAPH

It shouldn't be the aim of education to make the pupil a perfect learner in all the sciences, or indeed in any one of them, but to give his mind the freedom, disposition, and habits that can enable him to acquire any knowledge that he wants or needs in the future course of his life

John Locke

The impartiality which, in contemplation, is the unalloyed desire for truth, is the very same quality of mind which, in action, is justice, and in emotion is that universal love which can be given to all, and not only to those who are judged useful or admirable. Thus contemplation enlarges not only the objects of our thoughts, but also the objects of our actions and our affections: it makes us citizens of the universe, not only of one walled city at war with all the rest. In this citizenship of the universe consists man's true freedom, and his liberation from the thralldom of narrow hopes and fears.

Bertrand Russell

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

C	Cytosine
A	Adenine
T	Thymine
G	Guanine
SNV	Single Nucleotide Variant
CNV	Copy Number Variant
SV	Structural Variant
MEI	Mobile Element Insertions
UTR	Untranslated Region
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
AID	Activity Induced Cytidine Deaminase
BER	Base Excision Repair
MMR	Mismatch Repair
CPD	Cyclobutane Pyrmidine Dimer
DSB	Double Strand Break
WGS	Whole Genome Sequencing
PCR	Polymerase Chain Reaction
VAF	Variant Allele Frequency
QS	Quality Score
iPSC	Induced Pluripotent Stem Cell
hiPSC	Human Induced Pluripotent Stem Cell
OSKM	Oct4, Sox2, Klf4, and C-myc

SCNT	Somatic Cell Nuclear Transfer
RC-seq	Retrotransposon Capture Sequencing
Xrcc2	X-Ray Repair Cross Complementing 2
hAPP	Human Amyloid Precursor Protein
OB	Olfactory Bulb
MT	Mouse Mitral and Tufted Neurons
MDA	Multiple Displacement Amplification
ESC	Embryonic Stem Cell
LiRA	Single Cell Linked-Read Analysis
SNP	Single Nucleotide Polymorphism
RPE	Retinal Pigment Epithelial
HDF	Human Dermal Fibroblast
GFP	Green Fluorescent Protein
CFP	Cyan Fluorescent Protein
YFP	Yellow Fluorescent Protein
VPA	Valproic Acid
DP	Read Depth
tcNER	Transcription-coupled Nucleotide Excision Repair
ENCODE	Encyclopedia of DNA Elements
NT-ESC	Nuclear Transfer-derived Embryonic Stem Cell
lncRNA	Long non-coding RNA
Hsp90ab1	Heat Shock Protein 90 alpha class B number 1
MEF	Mouse Embryonic Fibroblasts

PBS	Phosphate Buffered Saline
NRL	Neural retina-specific leucine zipper protein

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to Dr. Kristin Baldwin for serving as my thesis advisor; without her wonderful guidance and support this thesis would not have been possible. I would also like to thank the members of the Baldwin lab who have supported me and provided insights and critiques of my work. Specifically, I want to thank Dr. Jen Hazen and Dr. Valentina Lo Sardo for their rigorous mentorship and critical assistance with both projects discussed in the following work. I would also like to thank my collaborators in the Hall lab, Dr. Ira Hall, Krishna Kanchi, Amy Ly, and Ryan Smith, for their work on both Chapters and many productive conference calls. Thank you to Dr. Ramesh Nair, who performed alignment and initial variant calling for our sister lines. Finally I would like to thank the Scripps Mouse Genetics Core, Dr. Sergey Kupriyanov and Alberto Rodriguez, for performing the nuclear transfer experiments as outlined in Chapter 3 and providing me with embryos on which to practice micromanipulation.

This work was funded in part by the National Institute of Health and by the California Institute of Regenerative Medicine

Data from Chapter 2 have been prepared for submission. The material as it may appear in print is: Duran M.A., Lo Sardo V., Hazen J.L., Nair R.V., Kanchi K., Lala S., Tu N., Hall I.M., Baldwin K.K. “The Impact of Different Reprogramming Methods on Human Induced Pluripotent Stem Cell Genomes.” Cell Stem Cell. The dissertation author was the primary author of this paper. Dr. Valentina Lo Sardo and the dissertation author contributed equally as the primary investigators of this project.

Data from Chapter 3 include unpublished material that was coauthored with Rodriguez R.A., Kanchi K., Hall I.M., and Baldwin K.K. The dissertation author was the primary investigator for the work shown in Chapter 3

VITA

Education

2012-2019 Ph.D. Biology, University of California San Diego

2008-2012 B.S. Biology, University of California Irvine

Research Experience

2012-2018 Ph.D. student, University of California San Diego
Advisor: Kristin Baldwin, Ph.D.

2012 Visiting Post-Baccalaureate Research Fellow, King's College, London
Advisor: Maddy Parsons, Ph.D.

2010-2012 Undergraduate Research Associate, University of California Irvine
Advisor: Dave Gardiner, Ph.D.

2010 Summer Research Intern, Centre Superior D'Investigacio En Salut Publica, Spain
Advisor: Alex Mira, Ph.D.

Publications

Duran M.A., Lo Sardo V., Hazen J.L., Nair R.V., Kanchi K., Lala S., Tu N., Hall I.M., Baldwin K.K. "The Impact of Different Reprogramming Methods on Human Induced Pluripotent Stem Cell Genomes." *Manuscript in Prep*

Lo Sardo, V., Chubukov P., Ferguson W., Kumar A., Teng E.L., Duran M.A., Zhang L., Cost G., Engler A., Urnov F., Topol E.J, Torkamani A., Baldwin K.K.,. (2018). "Unveiling the Role of the Most Impactful Cardiovascular Risk Locus through Haplotype Editing." Cell 175(7): 1796-1810.e1720.

Hazen J.L., Duran M.A., Smith R.P., Rodriguez A.R., Martin G.S., Kupriyanov S., Hall I.M., Baldwin K.K., (2017), "Using Cloning to Amplify Neuronal Genomes for Whole-Genome Sequencing and Comprehensive Mutation Detection and Validation." Springer Neuromethods, vol. 131, Pp. 163-185

ABSTRACT OF THE DISSERTATION

Assessing the Extent and Impact of Mutations in Human Induced Pluripotent Stem Cells and Young vs Aged Single Mouse Neurons

By

Michael A. Duran

Doctor of Philosophy in Biology

University of California San Diego, 2019

Professor Kristin Baldwin, Chair

Genomic mutations pose a serious risk to the health of individuals both in terms of somatic cells and in stem cells used for clinical and research applications. Here we outline novel approaches for studying genome mutations in the context of human induced pluripotent stem cells and neurons. To definitively establish the extent of reprogramming-associated mutations, we utilize the fact that a single fibroblast can give rise to two iPSC colonies separated by at most

two divisions. Comparison of the two colonies allows us to distinguish mutations present in the original cell (those found in both colonies) from mutations arising during the reprogramming process. We find on average 150-450 single nucleotide variants per iPSC line, with iPSCs derived by episomal method being significantly more mutated than iPSCs derived with lentivirus. Further, we find that the mutations from episomal reprogramming show unique signatures compared to lentiviral and somatic reprogramming methods. We also find that reprogramming does not contribute significant number of structural variant or mobile element insertion classes of mutation. In the context of neurons, we utilize somatic cell nuclear transfer to reprogram rod photoreceptors from young and aged mice , allowing us to amplify single neuronal genomes without error-prone PCR methods. We show that these neurons accumulate 20-40 SNVs per year, and that these mutations are enriched in nucleotiode contexts that implicate APOBEC deaminase, as well as potential regions of somatic hypermutation

Chapter 1:
Introduction and Background

1.1 Introduction

Though it is often said that cells are the building blocks of life, this is at best a half truth, for whether one is composed of a single cell or a billion, we are all circumscribed by a living code written in DNA. This code is unique to each organism and defines the capacities of the cell, providing the blueprints which enable complex organisms to flourish. Evolutionary forces converged on DNA due to its remarkable balance of stability and flexibility; it can persist for eons yet is mutable enough to allow for the breathtaking diversity that epitomizes life on Earth. Despite these extraordinary properties, DNA is not without its faults; its mutability, so critical for diverse mechanisms of survival, is a double-edged sword that can hinder as easily as it can help. Changes to DNA can occur through a wide variety of mechanisms, which we broadly call mutagens, and the spectra of mutations arising from these mutagenic forces can have a devastating impact on the cell or, more rarely, on the organism as a whole.

1.1.1 Mutation types and sources

DNA is comprised of 4 nucleotide bases, cytosine (C), adenine (A), thymine (T), and guanine (G). Guanine and adenine are referred to as purines, while cytosine and thymine are pyrimidines. The double helix structure of DNA pairs purines (G,A) with pyrimidines (C,T), such that G is always linked to C, and A is always linked to T. Any single base pair can be erroneously altered to another base by a variety of processes, and this error can be made permanent by replication or by faulty repair machinery. A mutation of this type is called a single nucleotide variant (SNV), and is by far the most common type of mutation observed in cells. SNVs are categorized according to the nature of the bases that were changed. A purine to

pyrimidine swap is called a transversion, while a change from one purine to another purine (or pyrimidine to pyrimidine) is called a transition. SNVs arise from a wide variety of processes, as will be seen below, and are often benign, though they can occasionally pose serious problems for the functionality of the cell.

Insertions or deletions of one to fifty base pairs are called indels, and are generally caused by errors in DNA repair. Indels are difficult to call bioinformatically, which makes the prevalence of this class of mutation difficult to assess (1). The best available data indicates these events are an order of magnitude more rare than SNVs, but are still relatively common in comparison to larger genome rearrangements. A thorough study of indels in 179 human genomes by Montgomery et. al. (2013) found indel events at a frequency of 4.2×10^{-5} per base in non-repetitive regions compared to an SNV frequency of 6.9×10^{-4} (2). Large insertions or deletions are called copy number variants (CNVs) or structural variants (SVs). These types of mutations have been extensively studied and can have a dramatic effect on cell function. A meta analysis of 23 studies including 2,647 subjects found that CNVs were unevenly distributed in the genome, with subtelomeric and perichromatic regions being the most susceptible to CNVs (3). This same study found that roughly 5-10% of the human genome is responsible for most observed CNVs, and that these regions were enriched for paralogous genes (3).

A final class of mutation involves the activation of transposable elements to generate mobile element insertions (MEIs). Roughly 44% of the human genome is comprised of transposable elements, however 99.95% of these bases are remnants of ancient retroviral events and are incapable of being transposed in modern humans (4). The potentially active 0.05% of

transposable elements are divided into 4 classes based on their sequence, with additional variations seen within each class. L1 elements are roughly 6 kb long and are transcribed into RNA via an internal L1 promoter. The RNA is translated into an RNA binding protein and a reverse transcriptase/endonuclease, which form a complex with L1 transcripts that catalyzes the reverse transcription and integration of L1 back into the genome (5). Alu elements are roughly 300 bp long and are usually found in 3' UTRs, promoters, and intergenic regions, and are the most common MEI seen in the human genome (6), with a preference for gene-rich regions (7). It is transcribed by RNA polymerase III along with whatever is downstream (Alu elements lack a termination signal), and requires an L1-derived endonuclease to integrate into the genome (8). SVA elements are 300-fold rarer than Alu elements and, like Alu elements, require L1 activity for integration, though they do possess an internal reverse transcriptase. They are capable of truncation and inversion during integration, resulting in variable insertion lengths (9, 10). Finally, HERV-K elements code a glycosylated envelope protein as well as a homolog of HIV-1 Rev, though this does not yield infectious particles, and its disease relevance is poorly understood (11, 12).

These classes of mutation, SNV, indel, CNV, and MEI, are caused by a diverse array of cell processes and external stimuli. One common source of SNVs in humans is oxidative stress, which can cause G → T transversions by creating 8-oxoguanine, a guanine derivative that pairs with A (13). Additionally, the deamination of methyl-cytosine (5mC) by AID/APOBEC enzymes can convert cytosine to uracil (U), which is then converted to thymine by base excision repair (BER) and mismatch repair (MMR) machinery (14-16). UV radiation can cause damage via the creation of 6-4 photoproducts (which are pyrimidine adducts) or cyclobutane pyrimidine

dimers (CPDs) (17). 6-4 Photoproducts are enriched at transcription factor binding sites where DNA is naturally bent, such as the TATA-box, while CPDs act to inhibit the progress of DNA polymerase, which can lead to widespread genome damage (18). Finally, alkylating agents such as endogenous methylation machinery or bifunctional alkylating agents and chloroethylating agents found in medical treatments are all sources of DNA damage. Endogenous methylation machinery such as S-adenosyl methionine can alkylate oxygen and nitrogen atoms of DNA (19), resulting in adducts which have varying degrees of cytotoxicity and mutagenicity. If BER and MMR fail to remove these adduct, it can result in an SNV, most commonly a G → A transition as O^6MeG mispairs with T during replication (20, 21). Alkyl adducts remaining in template can also lead to double stranded breaks (DSBs), which can cause indels and CNVs (20).

1.1.2 Sequencing and calling mutations

The prevalence and impact of mutations is determined by combining whole genome sequencing (WGS), a bioinformatics pipeline, and validation by PCR. Modern approaches to WGS typically proceed as follows (with variations depending on the approach): DNA is sheared into short (100bp-1kb) fragments, after which adapters are ligated to both ends of the DNA. These adapters anneal to oligos on the surface of a flow cell, and the DNA is amplified in local clusters, with each cluster possessing a unique index identifier based on the original piece of DNA that bound the flow cell. Fluorescent nucleotides are then added and each base is recorded in a manner similar to Sanger Sequencing. Each cluster produces a single read, which is a short sequence of DNA representing the sequence of the original fragment. See (22) for review.

Each base pair in the genome will be present in a certain number of reads in the subsequent data set. Because DNA is subject to mutation during the library preparation and amplification steps

(23), multiple reads per base pair are necessary to ensure accurate data. Genomic DNA is also unevenly amplified and sequenced, with certain regions prone to being over or underrepresented (24). Because of this, more total reads are necessary to "see" underrepresented regions. In sequencing, researchers often speak of "targeted average read depth," which refers to the number of desired reads, on average, per base in the genome. So a sample sequenced to 30x read depth has an average of 30 reads per base pair with a bell curve distribution. Typical read depths range from 30x to 60x; higher read depths are used to find subclonal mutations within the starting population, and lower read depths are used when only commonly occurring mutations are required (budget limitations can also play a role, as sequencing becomes more expensive at higher depths). Targeted deep sequencing can achieve a read depth in the thousands by using specific DNA, such as a PCR product of a specific genomic region, as a template instead of the whole genome. Once the reads have been obtained from the sequencer, they must be aligned to a reference genome before mutations can be called (25).

There are a wide variety of tools available for alignment and variant calling, depending on the needs of the specific study (26, 27). The vast majority of initial variants called, however, will be false positives that must be filtered out before arriving at a dataset of high confidence mutations. Filtering can be done initially by metrics of read depth; variant calls with very few (<2) supporting reads can generally be discarded, and further filtering by read depth can be accomplished by considering variant allele frequency (VAF). VAF measures how frequently a variant call is observed compared to a wild type (# variant reads at base pair/ # total reads at base pair). A heterozygous mutation found in every cell of the sequenced population would be expected to show up in 50% of all reads, giving it a VAF of 0.5. A heterozygous mutation in half

the starting population of cells would show up in 25% of reads, with a VAF 0.25, and so on… VAF filters can be set based on what type of mutations are of interest in the study. Sequencing data also provides quality metrics which indicate the degree of confidence in a variant call, most commonly a phred score where quality (Q) is directly related to the probability (P) of an erroneous call according to the following equation (28):

$$P = 10^{(-Q/10)}$$

Most pipelines will apply additional criterion to arrive at a final quality score based on phred and various other metrics, therefore it is not advisable to compare quality scores across different variant calling pipelines.

1.1.3 Mutations in human iPSCs

One of the goals of modern biology is the regeneration of tissue damaged by age or disease. To that end, Takahashi et. al. (2007) reported on a protocol to create induced pluripotent stem cells (iPSCs) using a combination of four transcription factors, sox2, oct4, C-myc, and klf4 (OSKM) (29). iPSCs can be derived from commonly available cell types and have the capacity to differentiate into valuable post-mitotic populations of cells such as neurons and cardiomyocytes, opening up a wide range of possibilities in the clinic and academic research. Subsequent work in the field has described many variations on the original protocol for establishing iPSCs using different factors or small molecule compounds to enhance reprogramming of somatic cells (30).

Efforts to establish the mutational burden of iPSCs began when Kun Zhang's and Lawrence Goldstein's groups at UCSD (Gore and Li et. al. 2011) examined several

reprogramming methods and found 3 exome mutations on average that they associated with reprogramming (31). Ji et. al. (2012) found an average of 9 coding mutations in iPSCs derived by OSKM, though the authors made no attempt to assess false positive rates, likely overstating their results (32). Cheng et. al. (2012) found between 1058-1808 SNVs in three iPSC lines derived by episome, but this study did not distinguish reprogramming-associated mutations from mutations in the original donor cell (33). A study comparing iPSCs to ESCs derived by SCNT in mouse found about 9 exome SNVs per line (34). More recently Bhutani et. al. (2016) examined nine iPSC lines derived by three different methods (sendai, integrating retrovirus, and mRNA), all using Pou5f1, Sox2, Klf4, and C-Myc. Bhutani found on average 605 SNVs per iPSC, but crucially this study ignored any SNVs falling below a VAF of 0.4, ruling out mutations arising after the first division during reprogramming which could still contribute to phenotype. This study also did not distinguish mutations arising as a result of reprogramming from rare mutations present in the donor population (which was only sequenced to 40x average depth) (35). Lo Sardo et. al. (2017) did not examine reprogramming-associated mutations but did find that exome mutations in iPSCs increase linearly with the age of the donor (roughly 5-30 exome SNVs per line), to a plateau at around 90 years of age (36). In perhaps the most thorough study to date, Kwon et. al. (2017) found on average four exome mutations associated with reprogramming by OSKM; here all putative mutations were validated by targeted deep sequencing in the donor population, and they convincingly show that most of their initial calls are actually rare mutations in the parent cell population (37).

Studies have investigated CNVs and MEIs in iPSCs as well; Wissing et. al. (2012) used northern blots to show that an abundant type of MEI, L1 mRNA, was overexpressed in iPSCs

relative to human dermal fibroblasts (HDF), and further showed that L1 promoters were hypomethylated (38). A follow up study on 8 iPSC lines used retrotransposon capture sequencing (RC-seq) and found 11 MEIs, which were subsequently validated by PCR. This study also sequenced the donor population, but can't rule out the possibility that the MEIs existed at a very low frequency in the starting population of cells (39). Studies on CNVs have indicated as many as 57 per iPSC line (40), and that CNVs appear more often in iPSCs than fibroblasts, though the study in question made no effort to validate false positives (41). The same study indicated that regions including pluripotency genes such as Nanog were more likely to be duplicated in iPSCs, ostensibly due to selective pressure (41). A rigorous study on mouse iPSCs by Quinlan and Boland et. al. (2011) found little evidence of MEIs and SVs from reprogramming (42). Similarly Bhutani et. al. (2016) found almost no evidence of SVs arising from reprogramming (35).

The difficulty in assaying reprogramming-associated mutations stems from the fact that there is no way to recover the original genome once a somatic cell has been reprogrammed, and targeted deep sequencing of all putative SNVs is prohibitively expensive for any large study. The current literature on mutations in iPSCs thus leaves several key questions unanswered; no study to date has attempted whole genome sequencing of iPSCs with a mechanism to distinguish reprogramming-associated mutations from somatic ones, and no study has examined indels in non-exome sequencing. Furthermore, studies on MEIs and SVs have been contradictory, necessitating a more rigorous approach to determine the contribution of reprogramming to these classes of mutation. A more thorough understanding of how different reprogramming methods

contribute to the full spectrum of mutations in iPSCs will help clinicians and researchers choose reprogramming methods that are minimally mutagenic.

1.1.4 Mutations in neurons

Neurons pose a unique set of challenges for studying their genomic diversity at a single cell level. As post-mitotic cells, neurons do not divide, and efforts to force a mitotic program in neurons triggers an apoptotic response (43, 44). Despite these challenges, understanding how neuronal genomes change over time could shed light on age-associated cognitive decline and neurodegenerative disorders. The importance of genome integrity in neural function can be seen in the variety of disorders arising from mutations in DNA repair pathways. One of the earliest indicators of this relationship was a study by Barnes et. al. (1998), which showed that mutations in Ligase IV (crucial for DNA strand break repair) lead to embryonic lethality in mice that was associated with widespread neuronal death (45). Mutations in Xrcc2, involved in DNA repair by homologous recombination, were shown to be necessary for post-mitotic neuron development in mice (46). In humans, diseases such as Cockayne Syndrome and Xenoderma Pigmentosum are associated with neurological phenotypes and are caused by mutations that disrupt genome stability (47). Genome stability is also important for the daily functioning of neurons; a report by Suberbielle et. al. (2013) found evidence of widespread double strand breaks (DSBs) resulting from neuronal activity, a finding which was exacerbated in human amyloid precursor protein (hAPP) mouse models (48). A later study by Madabhushi et. al. (2015) found that neurons use DSBs to regulate rapid expression of neuronal early-response genes by eliminating topological barriers to their expression (49).

Given the importance of genome stability to neuronal function, researchers have endeavored to study neuronal genomes even absent a reliable means to capture single cell data. Early studies of aneuploidy showed that as many as 3% of mature neurons were missing a chromosome, indicating aneuploidy as a relatively common condition for neurons (50). Furthermore, neurons in mouse olfactory bulb (OB) and motor cortex stained positive for markers of neuronal activity even when aneuploidy, indicating they remain alive and may continue to contribute to brain function in some capacity (51). Studies on human cortex found rates of aneuploidy around 10%, with different chromosomes displaying different propensities for aneuploidy. The same study found an increase in aneuploidy among patients afflicted with Alzheimer's and Ataxia Telangectasia (AT) (52). McConnell et. al. (2013) analyzed 40 single neurons from human cortex and found 13 with CNVs not seen in bulk donor DNA, including a subset of hypermutated neurons (53). Early studies of MEIs in neurons indicated that L1 retroelement activity might play an important role in neural activity and diversity; a host of studies claimed rates as high as 80 MEIs per neuron (54-57). Recently however, several rigorous studies have called these results into question, showing insertions in the range of 0-1 per neuron (58-61). A particularly well-done study by Evrony et. al. (2016) showed that earlier studies incorrectly analyzed MEI calls, leading them to overstate the amount of MEIs present by a factor of 50 (60). Most recently Bedrosian et. al. (2018) show increased L1 activity in mice that receive less maternal care, possibly due to depletion of methylation in the L1 promoter in low care mice (62). Though intriguing, the paper does not find functional significance; it remains possible that the L1 activity is a mere byproduct of lower global methylation under different conditions of development. Furthermore, in arguing for a functional role of L1 MEIs it cites the earlier literature while ignoring the more recent studies arguing against it (62).

Initial attempts to elucidate the SNV burden in neurons relied on multiple displacement amplification (MDA) based single cell sequencing. Work by Lodato et. al. (2015) sequenced 36 cortical human neurons from three individuals to 40x average depth. They estimate each neuron possesses between 1458-1580 SNVs, with an enrichment in C → T transitions, however this study suffers from the typical challenges of SNV calling from single cell genomes; they cannot rule out mutations resulting from amplification of the initial genome, and amplification by MDA is known to introduce a C → T bias (63). A study by Hazen and Faust et. al. (2016) used somatic cell nuclear transfer (SCNT) to reprogram mature mouse mitral and tufted (MT) neurons of the olfactory bulb, generating embryonic stem cell lines (ESCs) and allowing them to amplify the genomic DNA without error-prone PCR methods. Sequencing 6 MT neurons (3 weeks to 6 months), they found roughly 69 SNVs, 17 indels, 1.5 SVs, and 0.7 MEIs per genome on average (59). Further, this study found significant enrichment of C → T transitions in agreement with the work by Lodato et. al. (2015), and found that the trinucleotide context was consistent with mutation by the cytosine deaminase Apobec 1 (59). This study also found that these mutations were enriched in expressed genes, indicating a potential functional significance. Most recently, a large study from the Walsh lab (Lodato, Rodin, Bohrson, Coulter, Barton, and Kwon et. al., 2017) used a novel method, single-cell linked-read analysis (LiRA), to overcome the issues of uneven amplification and mutation associated with single cell amplification by MDA. LiRA utilizes the fact that false positive from amplification or sequencing SNVs will be specific to reads associated with a single strand, while true SNVs will appear on both strands of a single chromosome. Thus, true SNVs will have reads supporting SNV calls on reads for both strands, while a false positive will have both alternate and reference allele support on reads associated

with a particular allele, using nearby germline SNPs to confirm which allele the mutation should have come from (64). Using this approach, the Walsh lab performed single cell sequencing of 93 cortical and 20 dentate gyrus (DG) neurons aged 4 months to 82 years. They found an increase in SNVs associated with age, with DG neurons accumulating roughly 40 SNVs/year and cortex accumulating roughly 23 SNVs/year (65). Further, they found that the fraction of C → T mutations decreased with age. Finally, they found the SNVs to be enriched in exons and neuronal genes, and they displayed a transcriptional strand bias (65).

Chapter 2:
Impact of Different Reprogramming Methods on iPSC Genomes

2.1 Introduction

The use of patient-derived induced pluripotent stem cells (iPSCs) is a promising approach for cell replacement therapies owing to the ease with which somatic cells can be collected and the efficiency of reprogramming methods. Patient-derived iPSCs can be differentiated into an array of somatic cell types while avoiding the immune system complications observed in classic stem cell replacement therapies. In 2014, Masayo Takahashi's team in Japan launched a clinical trial to treat macular degeneration using iPSC-derived retinal pigment epithelial cells (RPE), however this trial was halted after deleterious mutations were discovered in the iPSCs of the second patient, raising questions about the safety of iPSCs in a clinical context (66, 67).

To address these questions, previous studies have assessed mutations in iPSC colonies using several approaches. Bhutani et. al. (2016) reported that reprogramming adds to mutational burden. However this and earlier studies lack the means to distinguish reprogramming associated mutations from somatic mutations present at low frequencies in the donor fibroblast population (35). More recently Kwon et. al. (2017) used targeted deep sequencing to address this shortcoming and found that indeed many of their called mutants were somatic rather than reprogramming-associated. However their study was restricted to the exome of three iPSC lines derived with a single method (37). Thus the following questions remain: precisely how many mutations arise as a result of reprogramming, and do the number and character of these mutations vary based on the reprogramming method used (Table 2.1). To definitively establish the complete spectrum of mutations associated with different reprogramming processes we devised a novel approach, establishing iPSC colonies derived from the same progenitor

fibroblast separated by at most two divisions, which we called sister pairs. We performed WGS and comprehensive mutation discovery and validation on these sister iPSC colonies and compared the genomes with the assumption that any shared mutations must have existed in the original parent fibroblast, while any unique mutations must have arisen during the reprogramming process (Figure 2.1). We show for the first time that reprogramming contributes hundreds of SNVs to the mutational burden of the iPSC, and that the choice of reprogramming method can significantly impact SNV burden, but not the burden of other classes of mutation. These results represent the first whole genome analysis with the ability to distinguish somatic mutations from those arising during reprogramming.

Table 2.1: A list of studies which have examined mutations in human iPSCs. Our study aims to provide definitive answers to these classes of mutation for the entire genome, in addition to assessing MEIs and SVs.

	Reprogramming SNV (avg)	Somatic SNV (avg)	Reprogramming Indel	Somatic Indel (avg)
Bhutani *et. al.* 2016	N/A	605*	N/A	173*
Kwon *et. al.* 2017	4 (exome)	76 (exome)	N/A	8 (exome)*
Lo Sardo *et al.* 2017	N/A	18 (exome)*	N/A	N/A
Ji *et. al.* 2012	N/A	12 (exome)*	N/A	N/A
Gore and Li *et. al.* 2011	3 (exome)	3 (exome)	N/A	0 (exome)*
Cheng *et. al.* 2012	N/A	1325*	N/A	1*

**Does not distinguish somatic from reprogramming-associated*

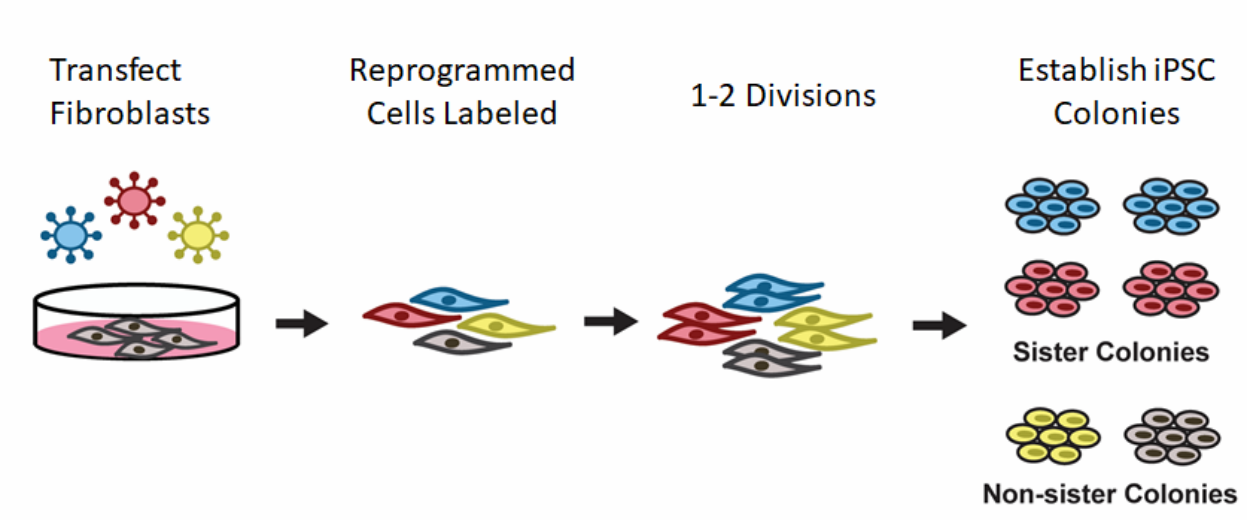


Figure 2.1: An overview of our sister iPSC paradigm. Fibroblasts were transfected with reprogramming factors using lentiviral or episomal delivery. Cells were simultaneously infected with fluorescent lentivirus. A small subset of cells would divide 1-2 times before forming proximal iPSC colonies from two fibroblasts that derived from the same progenitor cell.

2.3 Results

2.3.1 Developing a paradigm to assess reprogramming and somatic mutations in hiPSCs

We reprogrammed early post-natal human dermal fibroblasts (HDF) with one of two methods. (a) HDFs were infected with four dox inducible lentiviral constructs containing Oct4, Sox2, Klf4, and C-Myc, along with fluorescently labeled lentivirus with GFP, CFP, and YFP. (b) Alternatively, HDFs were transfected with episomal plasmids containing Oct4, Sox2, Klf4, Lin28, L-myc, and p53 shRNA, along with fluorescent lentivirus as mentioned above. Optimal conditions for obtaining sister iPSC colonies were determined by Dr. Valentina Lo Sardo (Fig. 2.2). Applying this approach we obtained 6 sister pairs derived from lentivirus and 2 pairs derived from episome, as well as 4 pairs of colonies which were found near one another but did not derive from the same original fibroblast (Fig. 2.3). Sister status was validated by Dr. Valentina Lo Sardo using Southern blot to check for the integration pattern of fluorescent lentivirus in iPSC colonies that shared a color and were found in close proximity. Pluripotency was validated by staining for pluripotency markers and confirming that the iPSCs could give rise to different cell types (Fig. 2.4).

Samples were sequenced to an average of 35x depth (Fig. 2.5). SNVs and Indels were filtered by depth and sequence quality. We defined candidate reprogramming mutations as those found in a single sister of an iPSC pair and in no other sample, while somatic mutations were those found in both sisters of a sister pair but in no other sample. In this way we are looking for (a) the mutations associated with reprogramming, and (b) rare mutations present in the original fibroblast that are missed by bulk sequencing the donor cell population. The rationale for screening mutants of one sister against all 23 other iPSC lines was that there are genomic regions which are known to be error-prone in sequencing, and this would help eliminate such mutations,

which are extremely unlikely to have arisen independently in two iPSC colonies by natural processes.

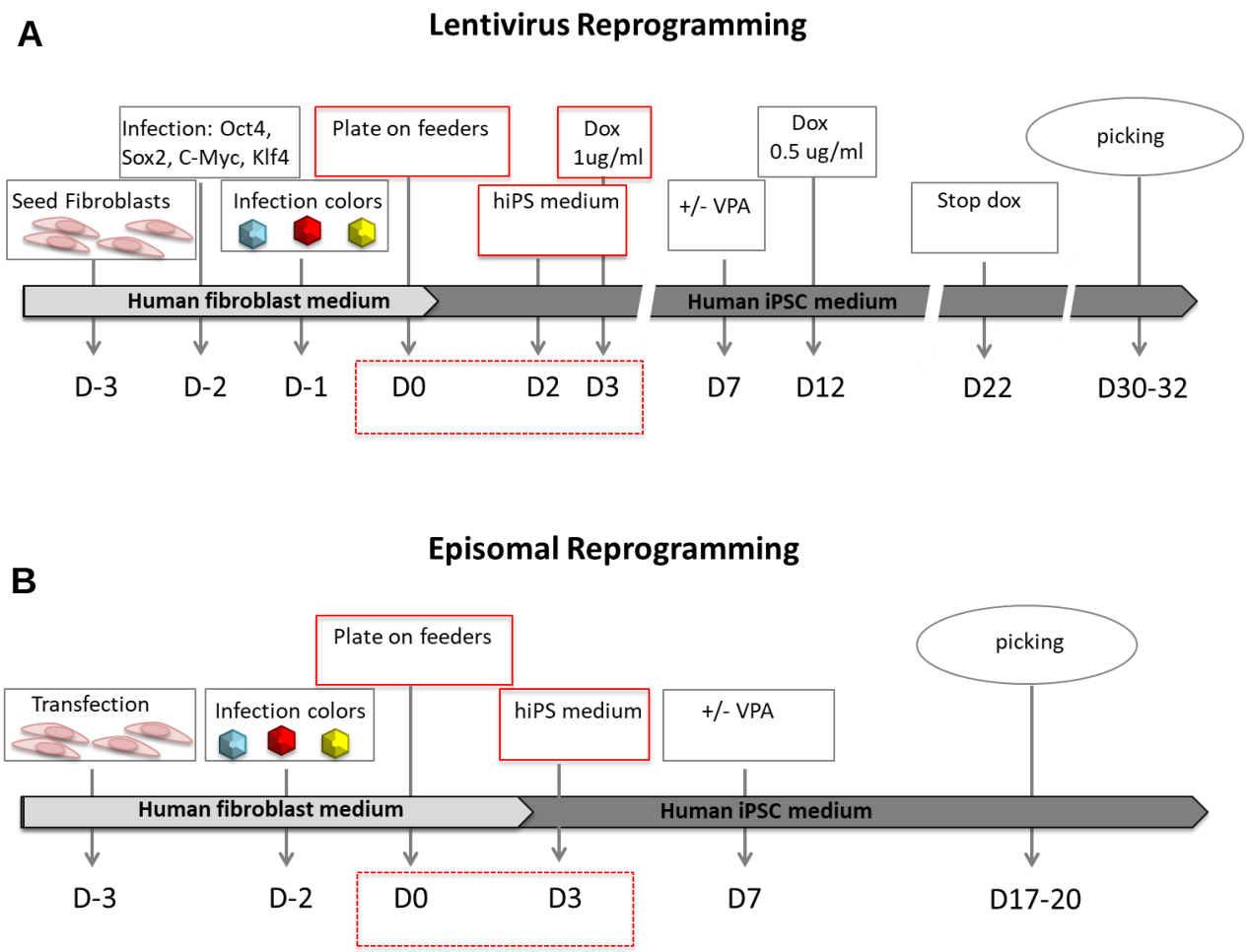


Figure 2.2: (a) Dox inducible OSKM casettes were integrated into the genome with lentivirus and were induced 3 days after plating on feeders and one day after switching to iPSC media. Colonies were picked approximately 28-30 days after induction (b) Fibroblasts were transfected with constitutively active episomal vectors and plated on feeders 3 days after transfection. iPSC colonies were picked approximately 20 days after initial transfection.

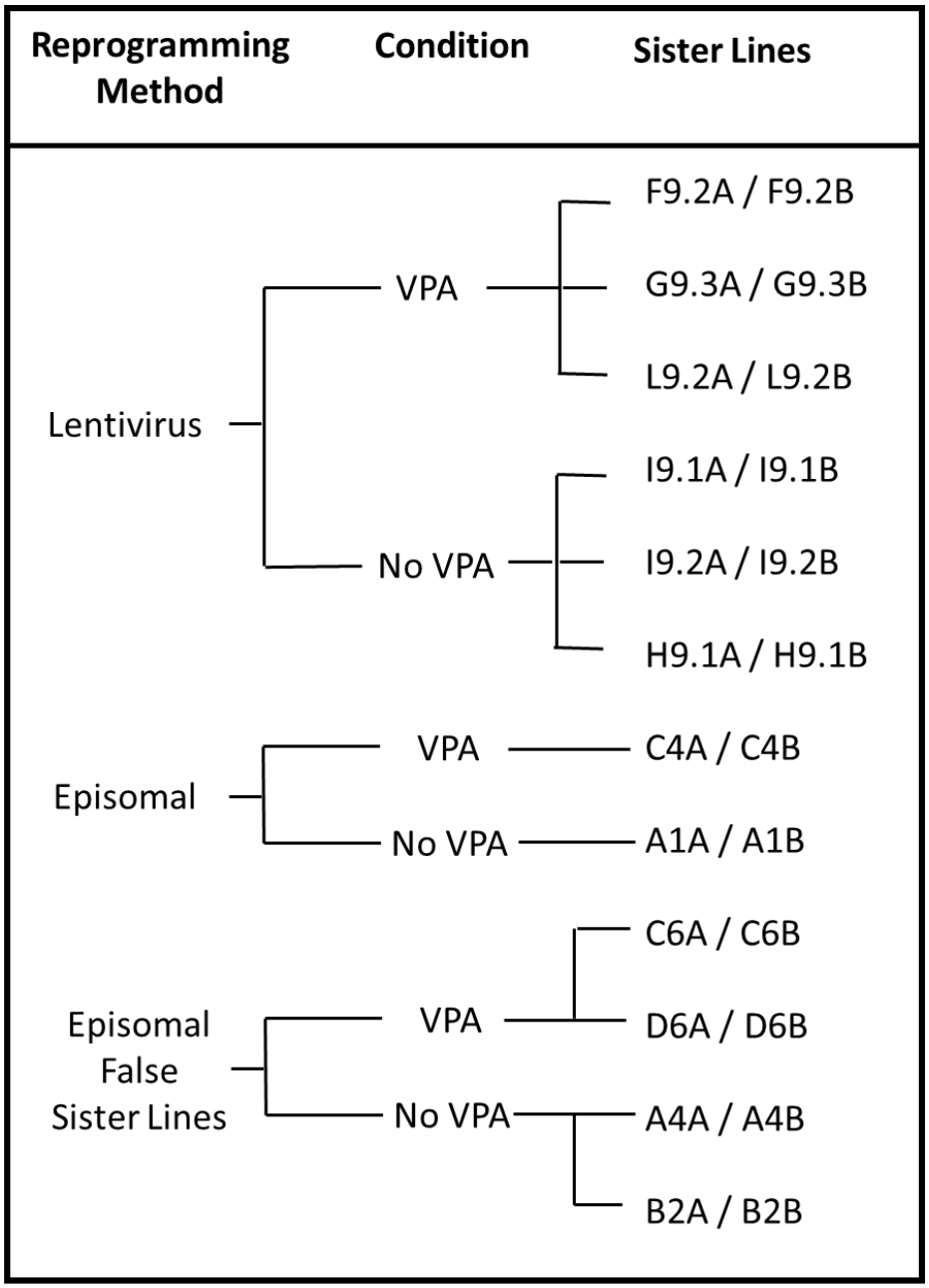


Figure 2.3: Summary of all iPSC lines collected and sequenced for further analysis. We derived 12 iPSCs via our lentiviral method, 3 pairs each with or without VPA. Two sister pairs were derived with the episomal method and, and 8 false sister pairs (proximal iPSCs that did not come from the same donor fibroblast) were collected as controls.

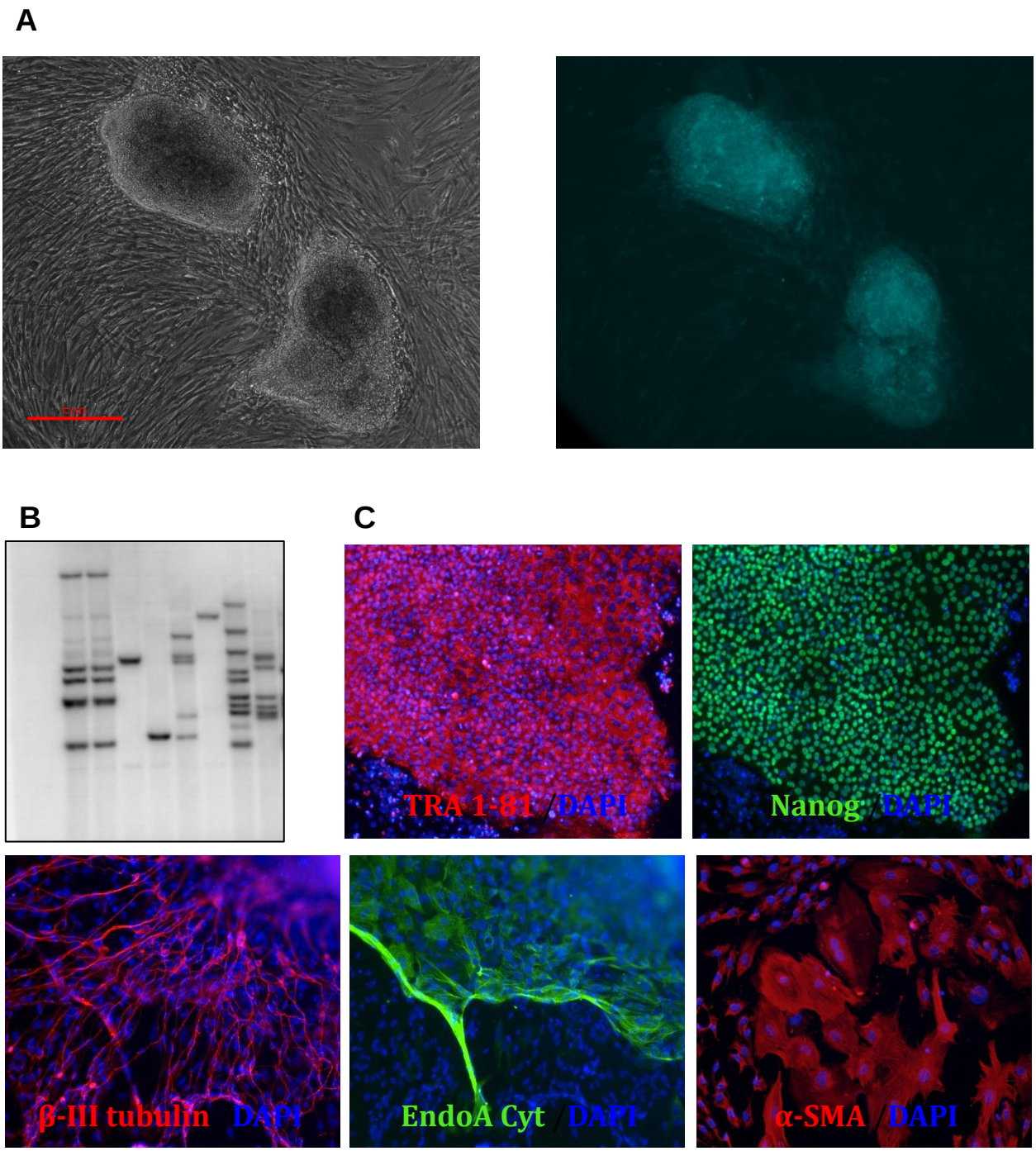


Figure 2.4: (a) Putative sister colonies were picked by proximity and shared fluorescent color. (b) Sister status was confirmed by southern blot examining the integration pattern of fluorescent lentiviral label; colonies with a shared pattern were deemed sister pairs. (c) Pluripotency was assessed by staining for pluripotency markers Tra 1-81 and Nanog, and by differentiating the iPSC colonies and staining for differentiation markers β-III tubulin, EndoA, and α-SMA. *Data in this fig. generated by Dr. Valentina Lo Sardo*

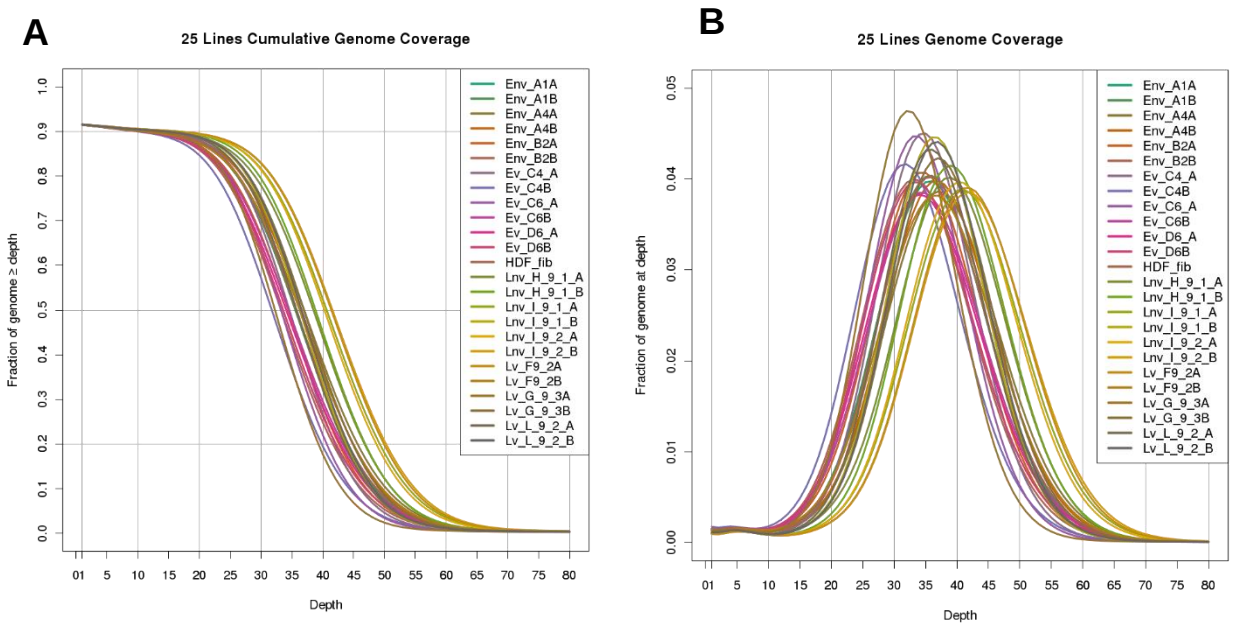


Figure 2.5: (a) The % of the genome covered at each read depth with targeted 35x average coverage. (b) The fraction of the genome covered at each read depth. Most of the genome was covered on average to between 30-40x read depth.

2.3.2 iPSCs contain hundreds of mutations from the parent cell not captured in bulk sequencing

To determine how many rare pre-existing mutations were present in our original fibroblasts we looked at mutations shared between two sister lines and missed in the bulk HDF sequencing and identified 779 predicted somatic SNVs and 39 indels on average per iPSC line (Fig. 2.6a,b). We see very few mutations in our fake sister controls, which is expected if the two colonies came from different starting fibroblasts. We found that the somatic calls have a VAF distribution centered at 0.5, as expected of heterozygous mutations present in every cell of the iPSC colony (Fig. 2.6c). Because we found a greater degree of noise in somatic indels vs SNVs in our fake sister controls, we assessed the quality score of each class of mutation and found that our somatic indel calls were, on average, called with lower confidence than our SNV calls (fig. 2.6d). These data indicate that simply sequencing the donor cells will miss a significant number of SNVs and indels, arguing for the need to also sequence the iPSC colony before proceeding in the lab or the clinic.

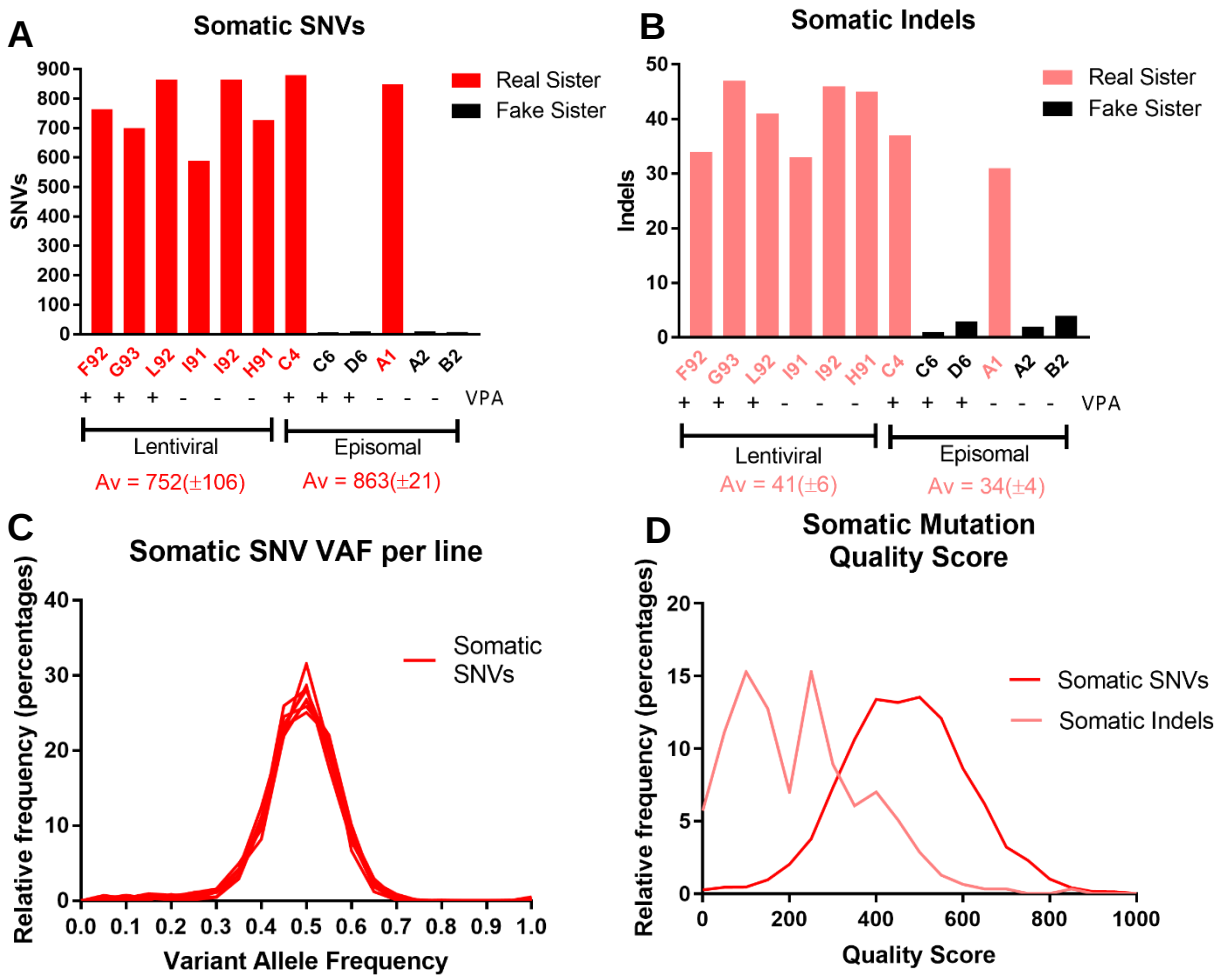


Figure 2.6 (a) All SNVs present in both iPSCs of a sister pair were called somatic mutants. The average is given below the graph for each condition with the standard deviation. (b) Somatic indels were using the same criteria. (c) VAF of somatic SNVs among each individual line. (d) The relative frequency of indels and SNVs at different quality score thresholds, totaled with all real sister lines.

2.3.3 Episomal reprogramming results in more SNVs than lentiviral reprogramming

Examining mutations present in only a single sister colony we initially found 250-1500 reprogramming-associated SNVs and around 250 indel candidate mutations per iPSC line (Fig. 2.7a,b). To refine our candidate mutation list, we assessed false positive rate by separating calls into 4-5 bins based on the quality score given by our pipeline. We then used PCR and sanger sequencing to validate a subset of our SNV calls at each quality threshold (Fig. 2.8a). To ensure that our approach was sufficiently sensitive to validate calls with a low variant allele frequency (VAF), we performed targeted deep sequencing on an additional subset of SNV calls to an average depth of 8000 reads/base. Indels were validated entirely through targeted deep sequencing. We attempted to validate 100 SNV calls and found that calls above quality score 200 were validated at 100% frequency, while calls between 100-200 quality score validated at 61%, calls between 10-100 validated at 18% frequency, and calls with a score below 10 failed to validate at all (Fig. 2.8b,c). We found that indels validated at much lower frequencies, even at high quality scores, reflecting the difficulties in calling this type of mutation (Fig. 2.8d).

We examined the percent of calls in each line at each quality bin to assess whether any samples showed uniformly low quality sequencing reads and found that 2 samples, G9.3B and I9.1A, do show a low percent of high quality calls (Fig. 2.7c). We also assessed whether differences in average read depth might result in some lines having more high quality calls than others. We did see a significant correlation between lines with lower average depth having more low quality calls, but did not see the concomitant correlation of lines with higher depth having more high quality calls (Fig. 2.7d).

For each line we multiplied the number of reprogramming-associated calls in each quality bin by the corresponding false positive rate to arrive at a corrected value for reprogramming-associated SNVs and indels. We predict 158 reprogramming-associated SNVs and 9 indels on average among iPSCs derived from lentivirus. For iPSCs from episomal vectors, we find 484 reprogramming-associated SNVs and 9 indels on average per line (Fig. 2.9a,b). We found that the variant allele frequency (VAF) for high confidence reprogramming-associated mutations showed a peak around 0.25 and 0.5, consistent with mutations arising in the first couple of divisions (Fig. 2.9c,d). These data imply an early-burst of reprogramming associated mutations within the first two divisions of the initial cell. It is important to note, however, that our sequencing depth precludes us from reliably calling low VAF SNVs, and we cannot say how many mutations arise in later divisions.

We filtered our VAF data by total number of reads and found that mutations at bases with a higher read depth showed a more pronounced peak at lower VAFs (Fig. 2.9e,f), supporting the conclusion that higher depth sequencing would reveal more early mutations. Importantly, the episomal condition shows significantly more reprogramming SNVs than the lentiviral condition ($p<0.001$ by t test), and the fact that this difference is not observed among indels argues that this is a biological phenomenon and not an experimental artifact (Fig. 2.10a,b). To our knowledge this represents the first evidence that the choice of reprogramming method can have a significant impact on the mutational burden of iPSCs across the whole genome. Taken together, these data show that iPSCs have a substantial mutational burden arising both from somatic mutations in the donor cell and from reprogramming-associated processes. (Fig. 2.10c,d)

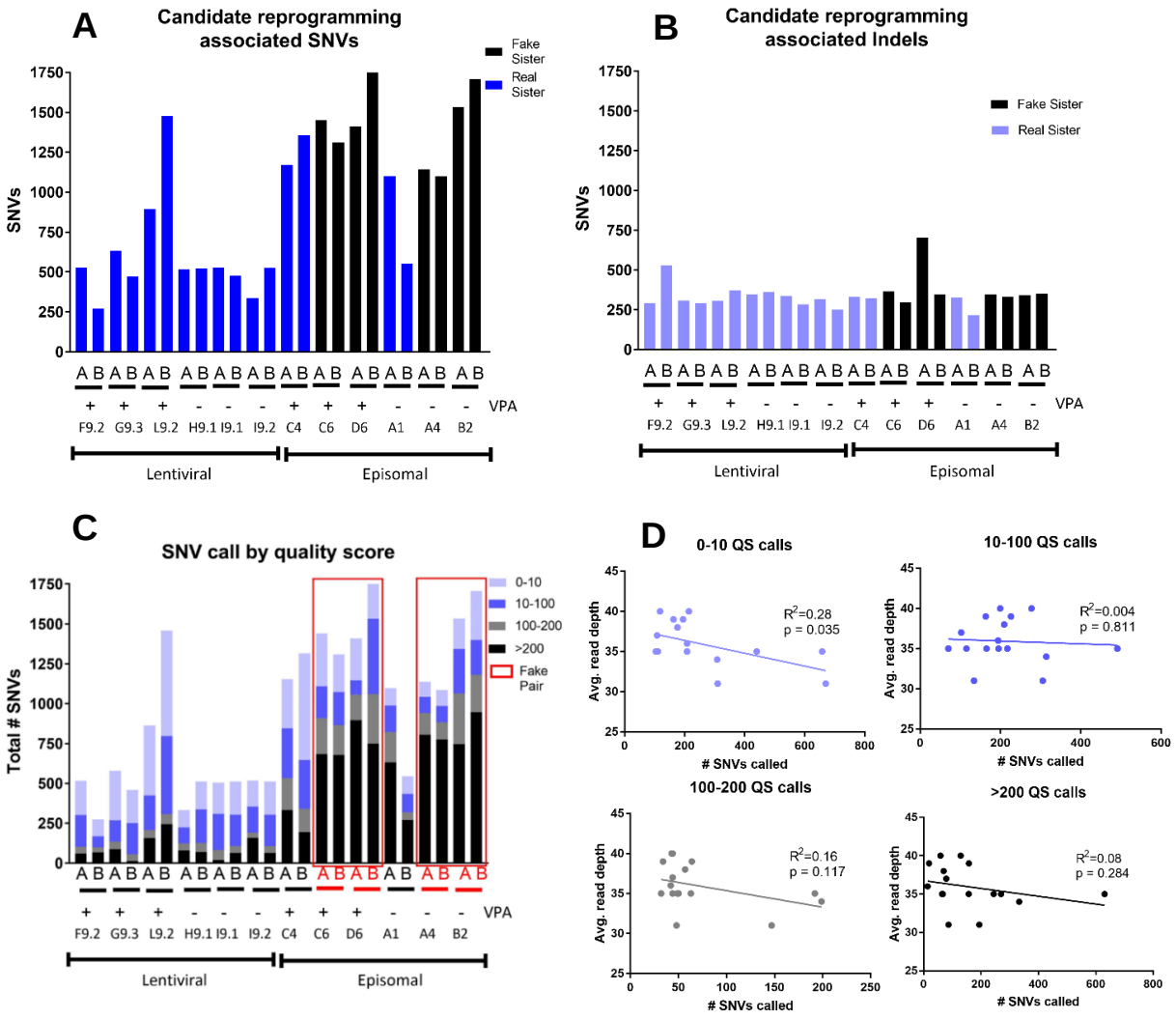


Figure 2.7: (a) Our initial assessment of candidate reprogramming-associated SNVs per line, without taking our false positive rate into account. (b) Candidate indels calculated without false positive rate for each line. (c) We examined the % of total SNV calls in each of our quality score bins per line. Our fake sisters appear to have higher average quality because they include somatic mutations. (d) Average depth vs the number of SNVs called for each line in each of our 4 quality bins.

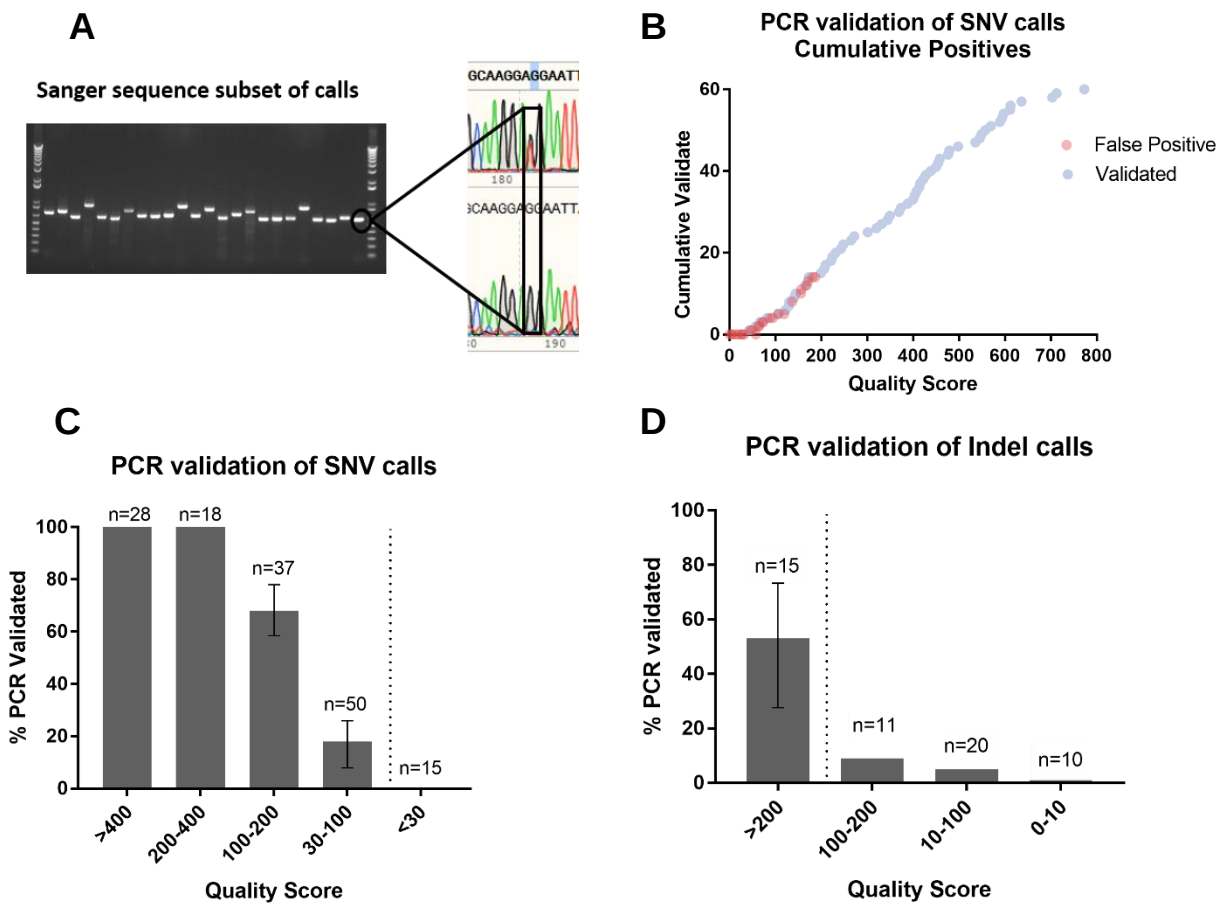
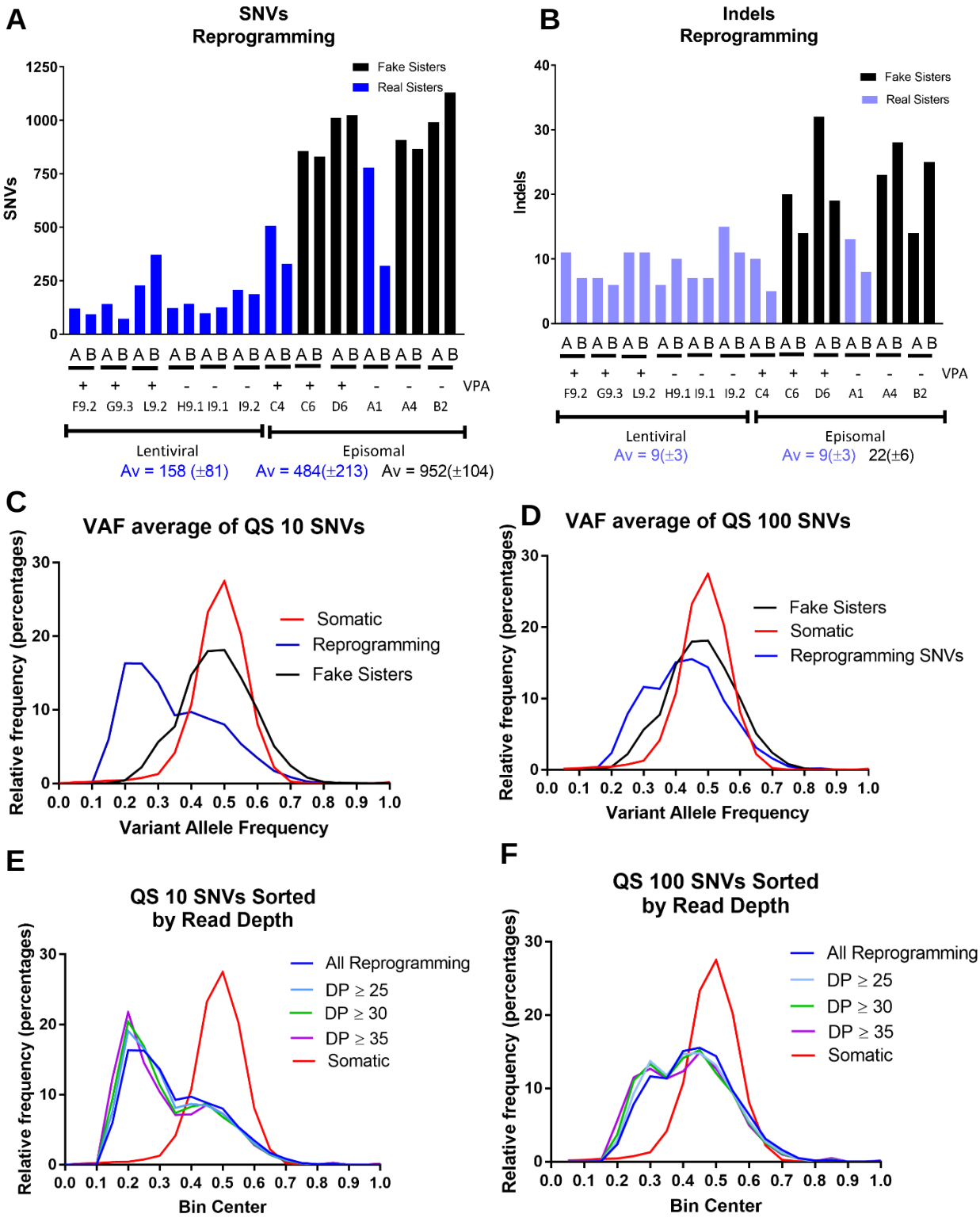


Figure 2.8: (a) We performed PCR on a subset of SNV calls and used Sanger sequencing to validate heterozygous mutants from our dataset. (b) Cumulative positive graph of false positives shows that the SNV call quality improves at the thresholds of QS 100 and then at 200. (c) The false positive rate for each of several quality score bins was calculated from validation experiments, and this rate was multiplied by the number of mutants in each bin per line to arrive at our final SNV estimates. (d) The same approach was applied to assess indel false positive rate, using targeted deep sequencing and Sanger sequencing.

Figure 2.9: (a) Predicted reprogramming mutations per line, adjusted for false positive rate. (b) Reprogramming-associated indels corrected for false positive rate. (c) The average of all reprogramming, somatic, and fake sister SNVs with QS > 10, plotted by the relative frequency of VAFs. (d) Same analysis as in e using a QS cutoff of 100. (e) Assessment of VAF of reprogramming mutations above QS 10 filtering for different minimum read depths (DP) to determine whether calls were being biased by the total number of reads at the locus. (f) Assessment of VAF frequency distribution using a QS 100 cutoff at different DP thresholds.


A
SNVs
Reprogramming
Fake Sisters
Real Sisters
SNVs
1250
1000
750
500
250
0
A B A B A B A B A B A B A B A B A B A B A B A B
+ + + - - - + + + - - - VPA
F9.2 G9.3 L9.2 H9.1 I9.1 I9.2 C4 C6 D6 A1 A4 B2
Lentiviral
Episomal
Av = 158 (±81)
Av = 484(±213)
Av = 952(±104)
B
Indels
Reprogramming
Fake Sisters
Real Sisters
Indels
40
30
20
10
0
A B A B A B A B A B A B A B A B A B A B A B A B
+ + + - - - + + + - - - VPA
F9.2 G9.3 L9.2 H9.1 I9.1 I9.2 C4 C6 D6 A1 A4 B2
Lentiviral
Episomal
Av = 9(±3)
Av = 9(±3)
22(±6)
C
VAF average of QS 10 SNVs
Relative frequency (percentages)
30
20
10
0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Variant Allele Frequency
Somatic
Reprogramming
Fake Sisters
D
VAF average of QS 100 SNVs
Relative frequency (percentages)
30
20
10
0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Variant Allele Frequency
Fake Sisters
Somatic
Reprogramming SNVs
E
QS 10 SNVs Sorted
by Read Depth
Relative frequency (percentages)
30
20
10
0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Bin Center
All Reprogramming
DP ≥ 25
DP ≥ 30
DP ≥ 35
Somatic
F
QS 100 SNVs Sorted
by Read Depth
Relative frequency (percentages)
30
20
10
0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Bin Center
All Reprogramming
DP ≥ 25
DP ≥ 30
DP ≥ 35
Somatic

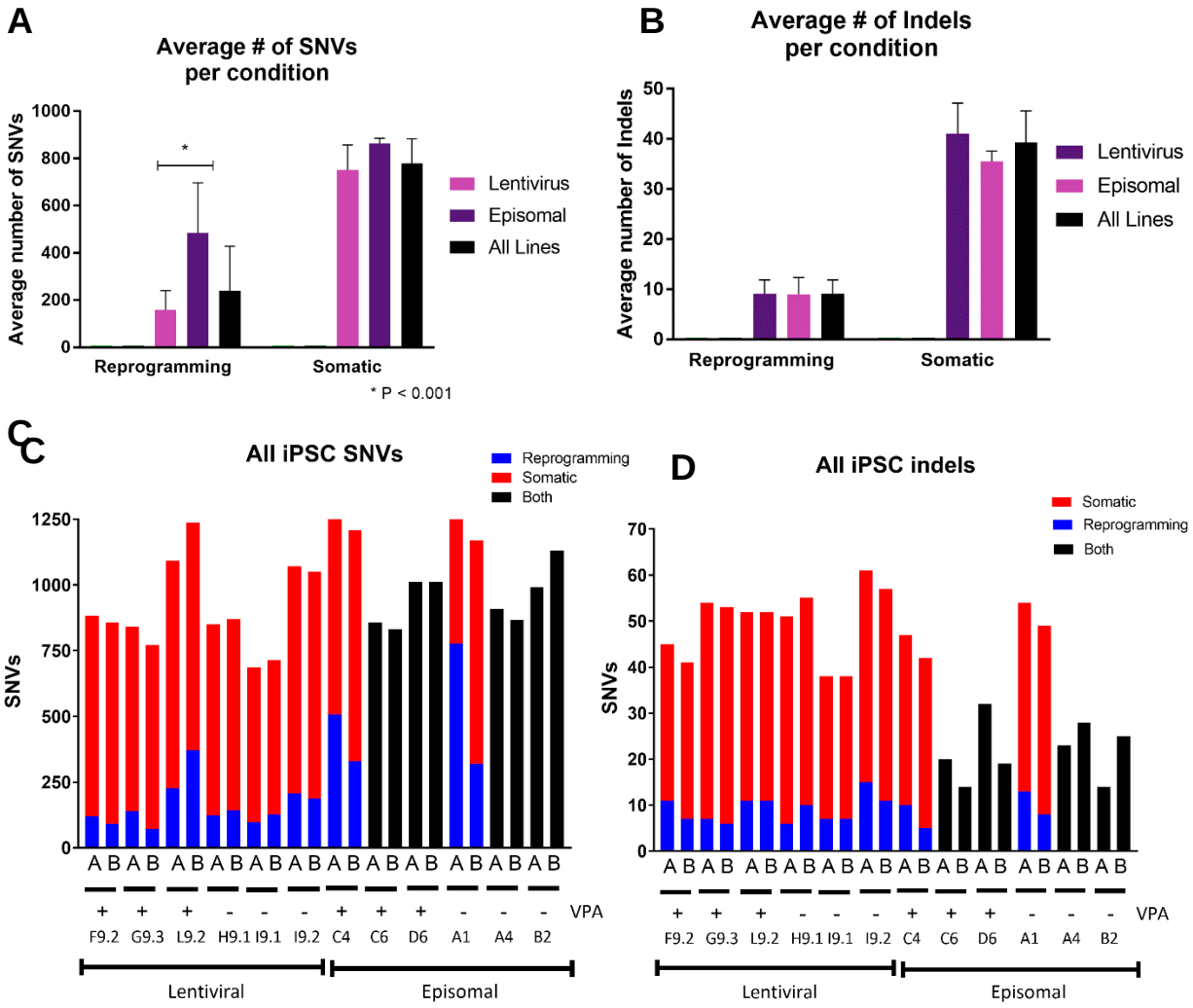


Figure 2.10: We summarized the averages for each condition and compared lentivirus to episomal using a T test (significance calculated by Holm-Sidak method). (b) Analyzing the average indels in each condition did not reveal any statistically significant differences. (c) Stacked graph assessing the total SNV burden, from reprogramming and somatic sources, in each iPSC line. (d) Stacked graph examining the average indel burden for each line. NB fake sisters appear to have fewer mutations because their somatic calls fell into the reprogramming bin and many were thus removed by our false positive filters despite being true mutations.

2.3.4 The nuclear context of SNVs in iPSC lines

To determine whether reprogramming-associated SNVs differed from somatic SNVs, we examined the nucleotide context of each mutation and found no significant differences between the nucleotide context of reprogramming-associated vs somatic SNVs (Fig. 2.11a). However, when we separated reprogramming-associated SNVs into episomal vs lentiviral-derived lines, we found that episomal reprogramming SNVs were significantly enriched in C → A mutations compared to somatic and lentiviral reprogramming mutations (Fig. 2.11b). To further explore this observation, we examined the trinucleotide contexts of lentiviral vs episomal SNVs and found episomal reprogramming SNVs were particularly enriched at TCN and GCN sites, and were relatively depleted in C → T SNVs (Fig. 2.11c,d).

One concern is that we only have four episomal iPSC lines based on two reprogramming events. We noted that our fake sister controls (proximal iPSC colonies that derived from separate fibroblasts) were derived by episomal method, and decided to assess nucleotide context in these cells despite not having the sister iPSC paradigm for these samples. To remove somatic SNVs from our analysis, we looked at the VAF of our real sister iPSCs and noted that most somatic mutations were found above VAF 0.35. We therefore analyzed all high confidence SNVs below VAF 0.35 for all lentiviral and episomal derived iPSC lines, ignorant of sister status, giving us an n of 12 for each condition. We found that we could still clearly see an enrichment of C → A mutations at specific trinucleotide contexts in episomal vs lentiviral derived iPSCs (Fig. 2.12a-c). This strongly indicates that reprogramming with episomal factors triggers different mechanisms of mutation than reprogramming with lentiviral OSKM.

To gain insight into the mechanisms of mutation operating during reprogramming, we utilized the mutational signatures first reported by Alexandrov et. al. (2013). They analyzed SNVs from over 7,000 cancer cell lines and found recurrent patterns of trinucleotide mutations termed mutational signatures. They were able to map many of these mutational signatures to specific cell processes in an effort to elucidate mechanisms responsible for specific types of cancer (68, 69). We used DeconstructSigs to map our SNVs to the mutational signature database maintained by COSMIC (70) and made several interesting observations. Processes underlying signature 18 and 24 made much larger contributions to episomal SNVs than lentiviral or somatic SNVs (Fig. 2.12d). While the mechanisms underlying signature 18 are unknown, this signature is most commonly observed in neuroblastoma. Signature 24 is associated with a transcriptional-strand bias for C → A mutations that are targeted by transcription-coupled nucleotide excision repair (tcNER). Interestingly, although the nucleotide contexts of lentiviral SNVs appear broadly similar to somatic SNVs, they show a unique enrichment for signature 5, associated with transcription-strand biased T → C mutations at ATN contexts, and signature 29, associated with CC → AA dinucleotide substitutions (Fig. 2.12e) (68). The mutational signatures indicate that transcription coupled nucleotide excision repair (tcNER) may play an outsized role in mutagenesis of episomal, but not lentiviral, reprogramming SNVs. Ultimately these data tell us that different methods of reprogramming cause mutations via mechanisms of action that are distinct from somatic processes and from each other.

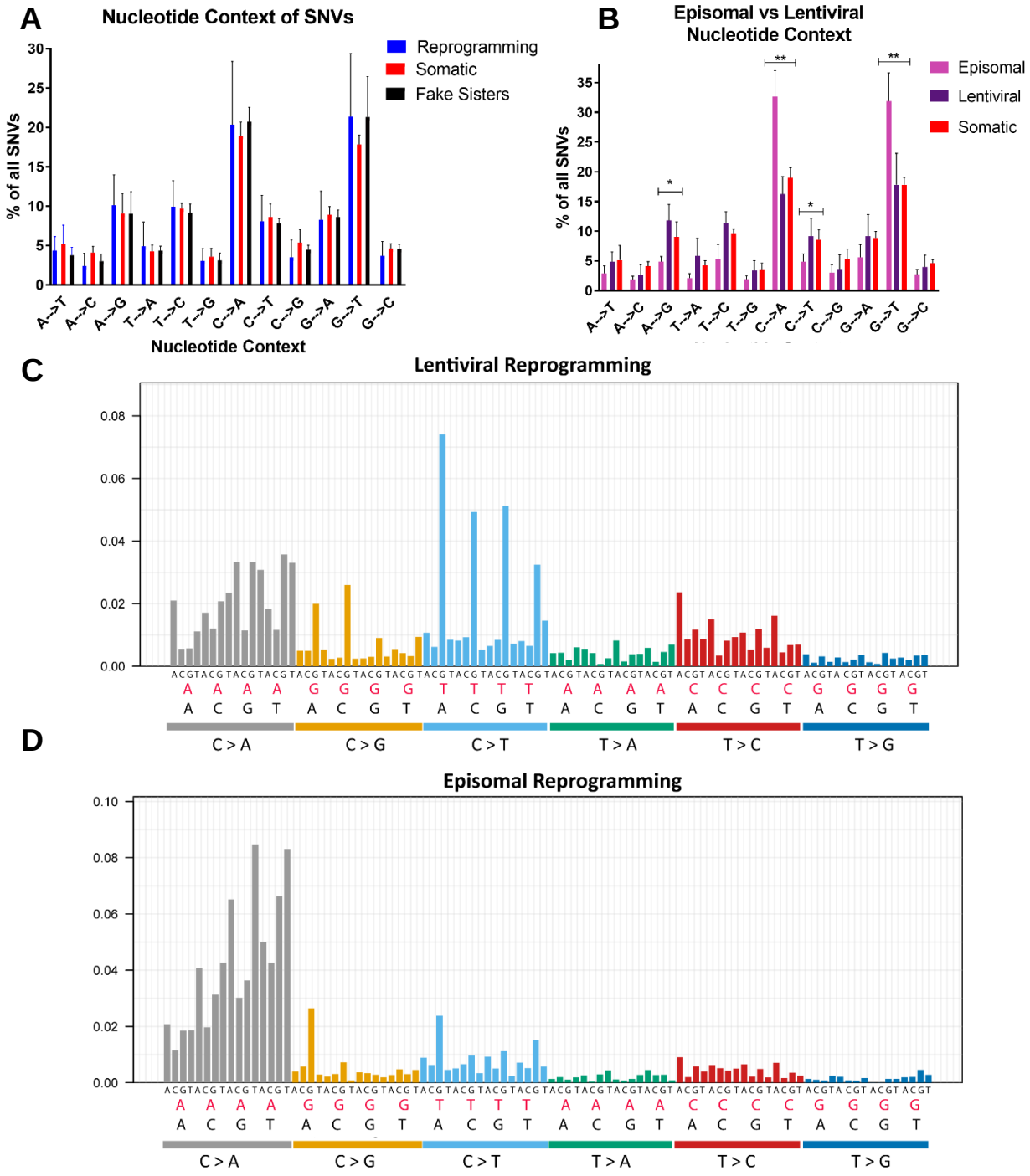
Because we found a unique nuclear signature of episomal reprogramming SNVs, we reasoned that these SNVs might be found in different regions of the genome. We looked for enrichment of SNVs in different genomic features and saw that reprogramming SNVs appear

depleted in 3'UTR and enriched in 5'UTR relative to somatic SNVs, but these results were not statistically significant, possibly due to insufficient sample size. (Fig. 2.13a). We next looked for enrichment in various histone contexts using ENCODE data on human embryonic stem cells (hESCs). However we did not observe any differences between reprogramming-associated and somatic SNVs (Fig. 2.13b). Because the epigenome of very early reprogramming might more closely mirror their original cell type, we gathered histone data from human dermal fibroblasts (HDFs), but we were again unable to discern any statistically significant differences in enrichment between somatic and reprogramming mutations (Fig. 2.13c). Because episomal SNVs are enriched in C → A SNVs, which are associated with oxidative stress, we wondered if mutations might be enriched in lamin-associated domains (LADS) which are known to be susceptible to oxidative damage. We looked for enrichment in LADs using a dataset from human embryonic stem cells (hESCs) as well as from fibrosarcoma (HT1080) but in both cases episomal reprogramming SNVs were not significantly more enriched than lentiviral or somatic SNVs (Fig. 2.13d)

Finally, we looked for regions of localized hypermutation, called katageis, which indicate particularly mutable regions of the genome. We found several katageis associated with lentiviral reprogramming but not with episomal reprogramming. Upon closer inspection, however, we realized that these katageis were largely due to a single line, L92B, which had several hypermutated regions, some falling in genes and some in intergenic regions (Fig. 2.14a). The mutations in genes were all silent, which at first made it appear as though lentiviral reprogramming was enriched in silent mutations relative to episomal reprogramming (Fig. 2.15a). We were concerned that L92B was an outlier that may have impacted our earlier results.

To investigate this potential source of error, we re-ran our previous analysis without L92B and confirmed that our prior observations on the total number of mutations and nucleotide context were still valid (Fig. 2.15c,d,e). Removing L92B eliminated the statistically significant enrichment in silent mutations as well (Fig. 2.15b). Our data indicates that some iPSC colonies can become hyper-mutated, though the reason for this is unknown.

Figure 2.11: (a) We considered all somatic and high confidence (QS > 100) reprogramming-associated SNVs and plotted the relative distribution of each nucleotide context. (b) The same analysis separating episomal derived and lentiviral derived sister iPSC lines. Statistically significant differences between the episomal sisters and other conditions calculated by T test. (c) The trinucleotide context of the episomal reprogramming SNVs, where the middle nucleotide is the SNV. (d) The same analysis as c on our lentiviral SNVs.


A
Nucleotide Context of SNVs
Reprogramming
Somatic
Fake Sisters
% of all SNVs
Nucleotide Context
B
Episomal vs Lentiviral
Nucleotide Context
Episomal
Lentiviral
Somatic
C
Lentiviral Reprogramming
D
Episomal Reprogramming
C > A
C > G
C > T
T > A
T > C
T > G

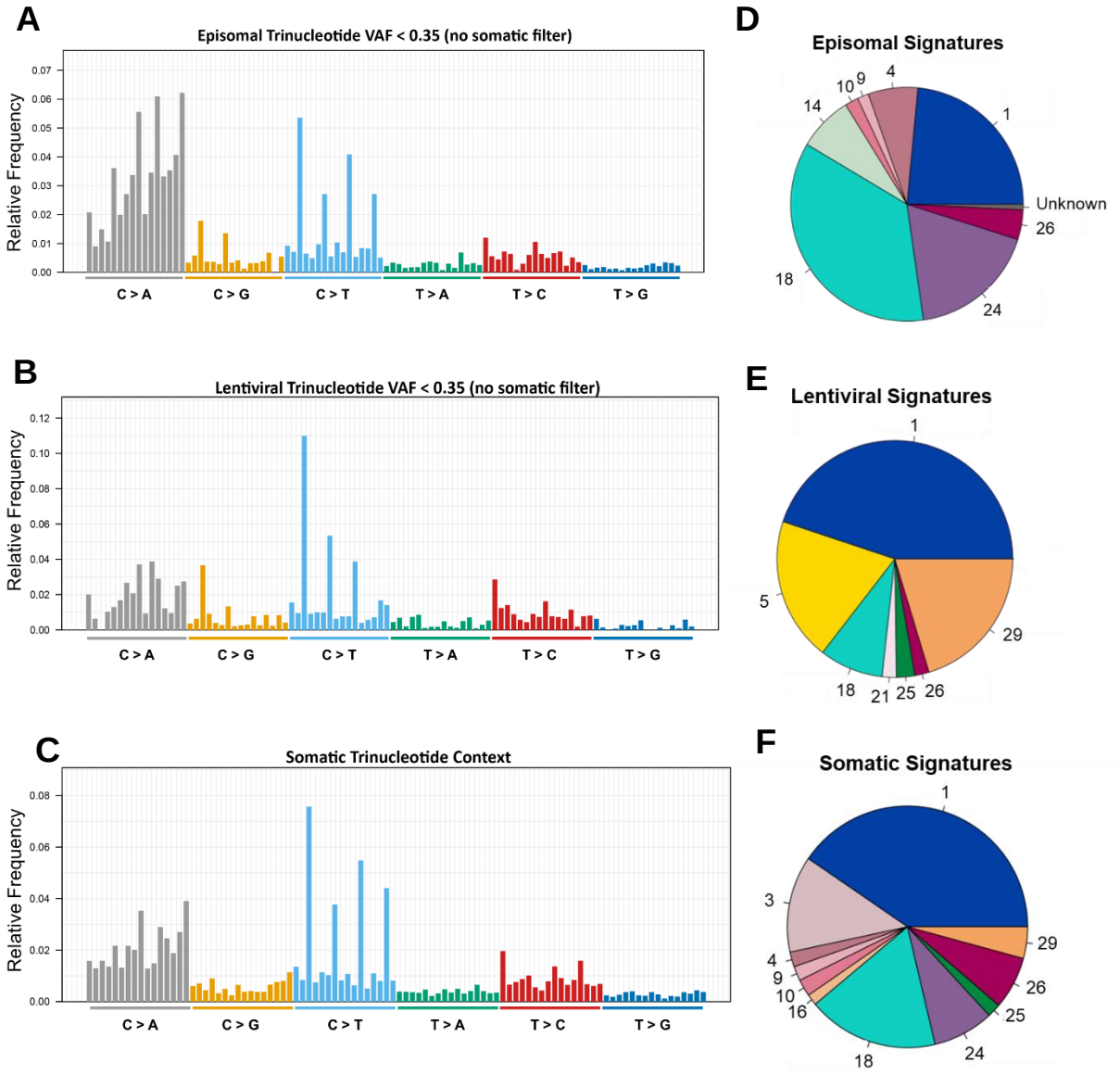
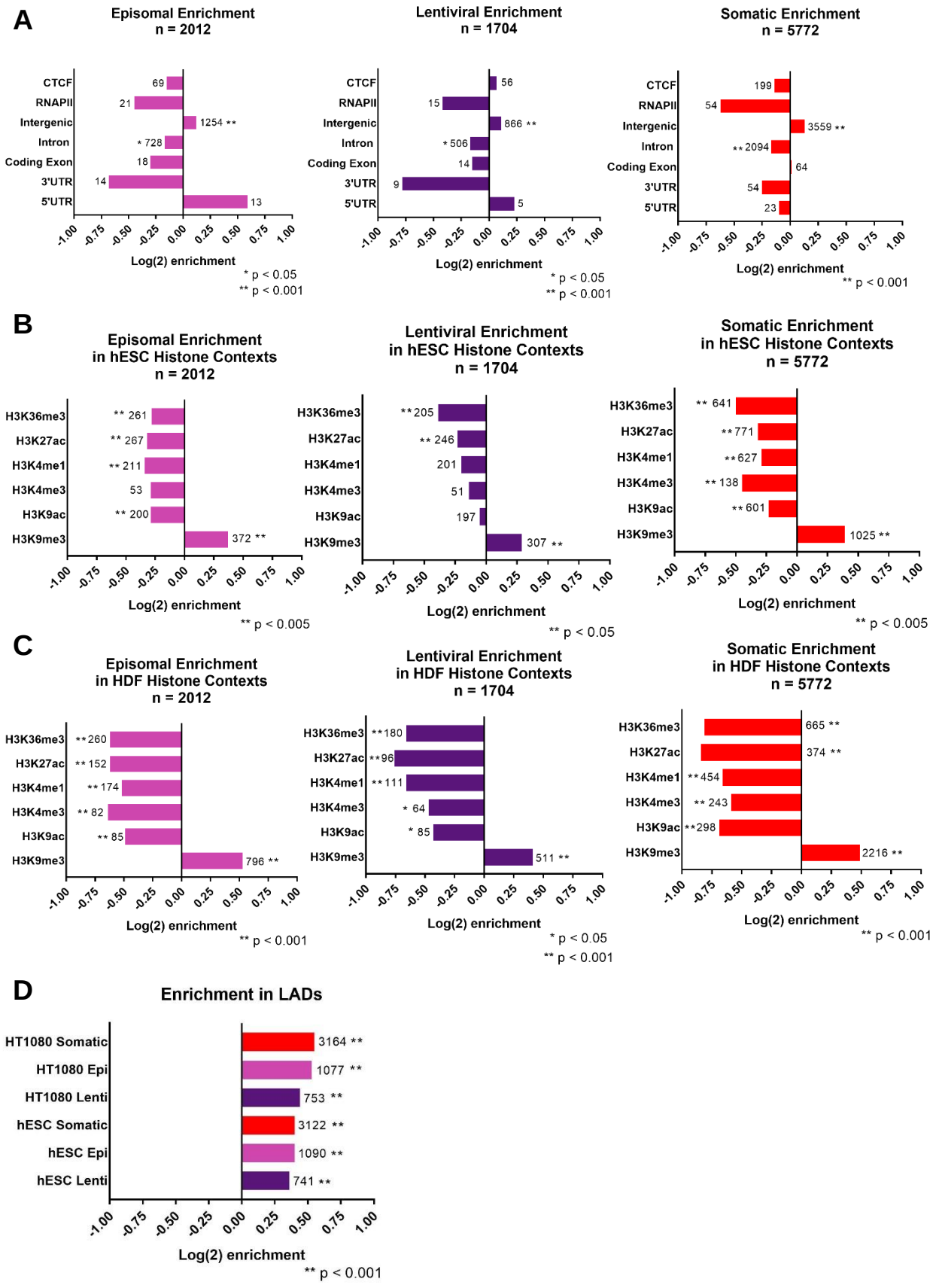


Figure 2.12: (a) Trinucleotide context for VAF < 0.35 SNVs of all episomal derived iPSCs ignorant of somatic vs reprogramming origin. It should be read from the bottom nucleotide to the top, with the nucleotide in red representing the mutated base. (b) Same analysis as above looking at all lentiviral SNVs of VAF < 0.35, ignorant of reprogramming vs somatic origin. (c) Trinucleotide analysis of all somatic SNVs. (d) A pie chart representing the % contribution of each mutational signature to the total mutational burden for the same SNVs assessed in 3a. (e) Mutational signatures associated with the data in 3b. (f) Mutational signatures associated with the data in 3c.

Figure 2.13: (a) Enrichment of SNVs in various genomic features of interest. The log(2) fold enrichment, where the n is the total number of SNVs considered, and the number next to each bar represents the total number of SNVs which overlapped with a particular feature. Significance was calculated by Fisher's exact (two-tailed) in bedtools to test the odds of the observed overlap given a random distribution of mutations. This analysis was run without L92B, an outlier which was strongly biasing the results. (b) The same analysis was run to find overlaps with an hESC histone ChIP assay. (c) Enrichment analysis using histone data from HDFs. (d) The same analysis for lamin-associated domains for hESCs as well as a fibrosarcoma line (HT1080)


A
Episomal Enrichment
n = 2012
CTCF
RNAPII
Intergenic
Intron
Coding Exon
3'UTR
5'UTR
69
21
1254 **
* 728
18
14
13
-1.00 -0.75 -0.50 -0.25 0.00 0.25 0.50 0.75 1.00
Log(2) enrichment
* p < 0.05
** p < 0.001
Lentiviral Enrichment
n = 1704
56
15
866 **
* 506
14
9
5
Somatic Enrichment
n = 5772
199
54
3559 **
** 2094
64
54
23
B
Episomal Enrichment
in hESC Histone Contexts
n = 2012
H3K36me3
H3K27ac
H3K4me1
H3K4me3
H3K9ac
H3K9me3
** 261
** 267
** 211
53
** 200
372 **
** p < 0.005
Lentiviral Enrichment
in hESC Histone Contexts
n = 1704
** 205
** 246
201
51
197
307 **
** p < 0.05
Somatic Enrichment
in hESC Histone Contexts
n = 5772
** 641
** 771
** 627
** 138
** 601
1025 **
** p < 0.005
C
Episomal Enrichment
in HDF Histone Contexts
n = 2012
** 260
** 152
** 174
** 82
** 85
796 **
** p < 0.001
Lentiviral Enrichment
in HDF Histone Contexts
n = 1704
** 180
** 96
** 111
* 64
* 85
511 **
* p < 0.05
** p < 0.001
Somatic Enrichment
in HDF Histone Contexts
n = 5772
665 **
374 **
** 454
** 243
** 298
2216 **
** p < 0.001
D
Enrichment in LADs
HT1080 Somatic
HT1080 Epi
HT1080 Lenti
hESC Somatic
hESC Epi
hESC Lenti
3164 **
1077 **
753 **
3122 **
1090 **
741 **
-1.00 -0.75 -0.50 -0.25 0.00 0.25 0.50 0.75 1.00
Log(2) enrichment
** p < 0.001

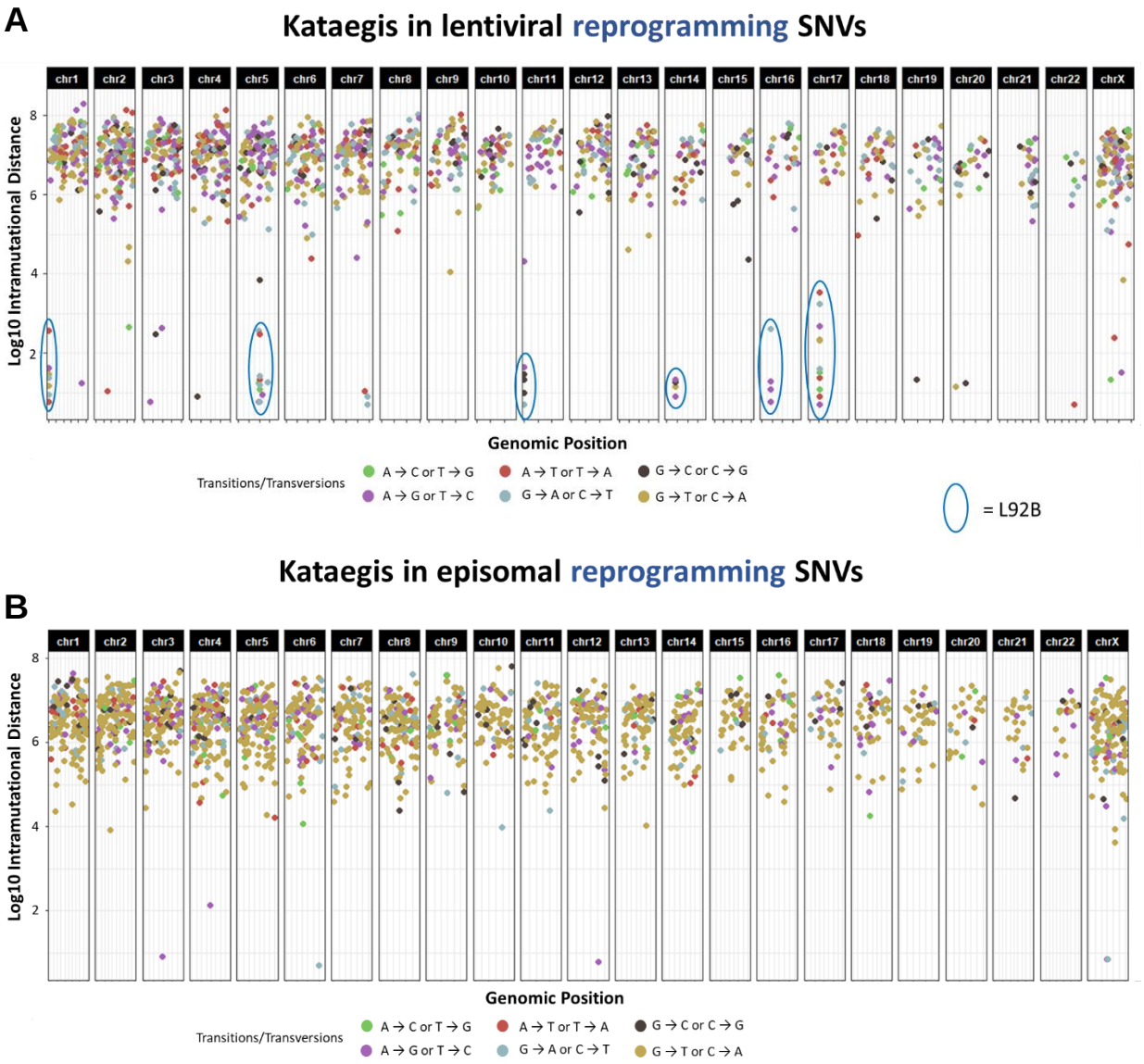


Figure 2.14: (a) Katagei (regions of localized hypermutation) were examined with MutationSigs. Each high confidence lentiviral SNV was plotted according to its location in the genome (x axis) and the distance between it and the previous mutation (y axis). Circled regions are katagei found in sample L92B (b) Locations of SNVs were plotted for high confidence episomal SNVs.

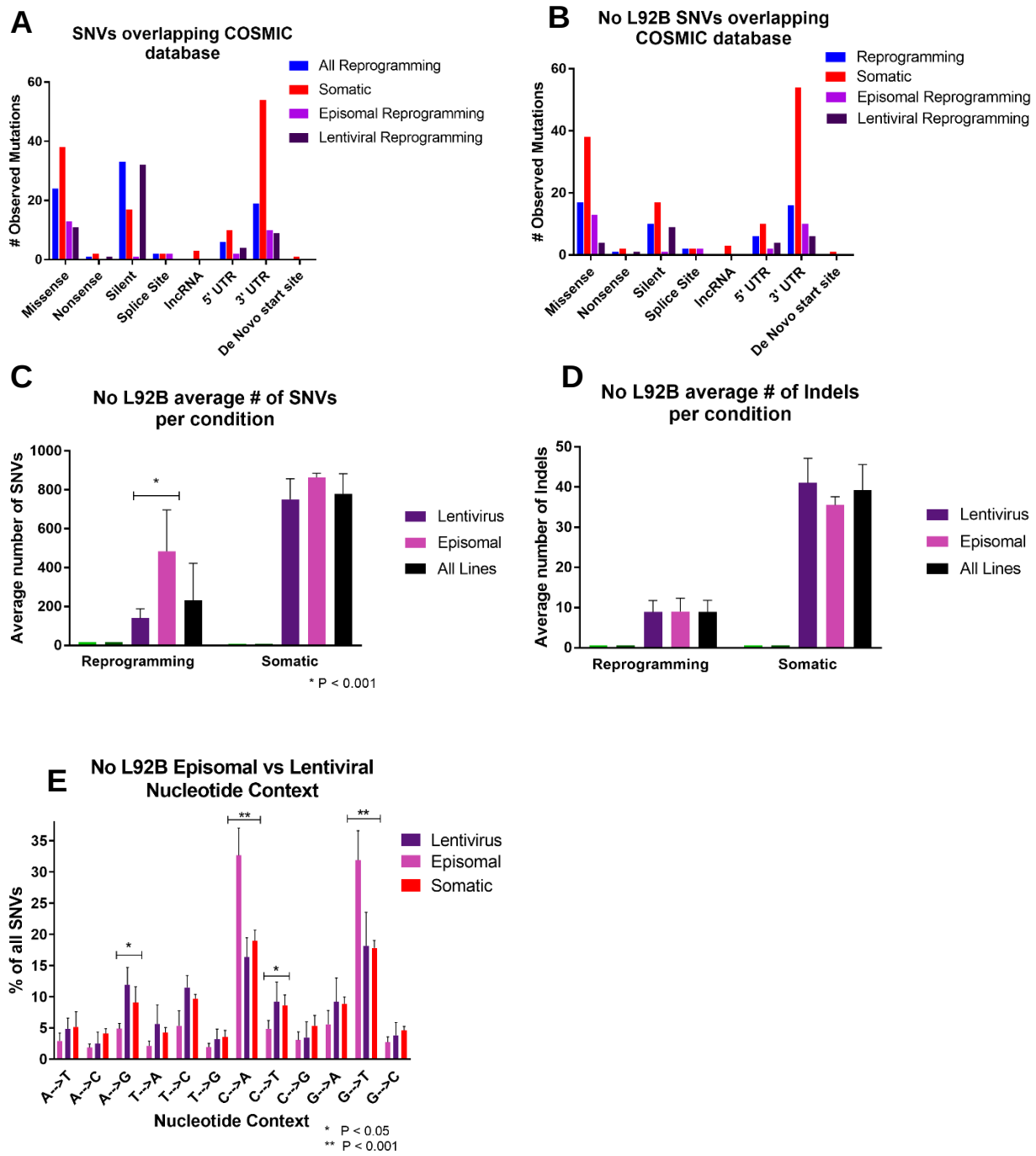


Figure 2.15: (a) High confidence SNVs were run against the COSMIC cancer database and plotted those that overlap an oncogene along with the predicted impact. (b) A rerun of the COSMIC analysis without sample L92B. (c) Assessment of the average number of SNVs per line without L92B, significance calculated by T test. (d) Assessment of the average number of indels per line without L92B. (e) Assessment of nucleotide context as previously described while omitting L92B from the analysis.

2.3.5 Physiological impact of mutations in iPSCs

Although we showed that reprogramming contributes several hundred SNVs and around 9 indels to our iPSC lines, we wanted to know whether these and somatic mutations would impact expressed genes. To assess the impact of mutational burden on physiology we used a panel of tools to predict functional impact of mutation. We first used two algorithms, SIFT and PolyPhen, which predict deleterious impact based on how evolutionarily conserved a base pair is (71-73). Looking at somatic SNVs, we found that 7/8 original fibroblasts contained at least one somatic SNV predicted to have a deleterious impact on a gene by one or both algorithms (Table 2.2). We ran this set of variants against the Online Mendelian Inheritance in Man (OMIM) database, which links genes to human disorders (74), and found 8 somatic mutations associated with a disease. We also ran these variants through ClinVar, which collates data on individual nucleotides of note (75). Of the somatic SNVs with an entry registered in NCBI (rsXXXX) none were directly linked with disease by ClinVar (Table 2.2). We performed the same analysis on our reprogramming-associated SNVs and found that 7/16 lines had at least one reprogramming-associated SNV predicted to have a deleterious impact on gene function. Several of these genes were associated with disease, though none of the specific base pairs were implicated by ClinVar (Table 2.3).

Table 2.2: Variant Effect Predictor (VEP) was used to assess somatic predicted deleterious SNVs by SIFT or PolyPhen in somatic SNVs. OMIM column represents where mutations overlapped genes with known diseases in the OMIM database or with a variant ID registered with NCBI. Rows in bold are SNV sites predicted to be deleterious by both algorithms. We also noted SNVs which result in a stop codon gained, which are not called by SIFT/PolyPhen but are generally deleterious.

Table S2: Somatic SNVs impacting gene function

iPSC Line	Gene	Mutant Type	Amino Acid Shift	OMIM Disease	Algorithm
A1	**Amer1**	**Missense**	**K/E**	**Osteopathia Striata**	**Both**
A1	**XRCC1**	**Missense**	**S/C**	**None**	**Both**
A1	**ZNF560**	**Missense**	**G/R**	**None**	**Both**
A1	**LTF**	**Missense**	**C/R**	**None**	**Both**
A1	ATP12A	Missense	C/F	None	SIFT
A1	BAZ1A	Missense	K/N	None	PolyPhen
C4	**Pdzrn4**	**Missense**	**R/H**	**None**	**Both**
C4	**Diras2**	**Missense**	**M/I**	**None**	**Both**
C4	**ASXL2**	**Missense**	**D/Y**	**Shashi-Pena Syndrome**	**Both**
C4	NRL	Missense	A/S	Retinitis Pigmentosa	SIFT
C4	Popdc2	Missense	Q/K	None	PolyPhen
C4	CDC3A	Splice Site	S/P	None	PolyPhen
I92	**Stx18**	**Missense**	**L/F**	**None**	**Both**
I92	**NELFB**	**Missense**	**L/S**	**None**	**Both**
I92	**Shroom2**	**Missense**	**Y/C**	**None**	**Both**
I92	OFD1	Missense	Q/K	Donnai-Barrow Syndrome	SIFT
I92	STAM1	Missense	L/F	Orofaciodigital Syndrome I	SIFT
F92	**HUWE1**	Missense	**L/S**	**Mental Retardation**	**Both**
F92	CLEC5A	Splice Site	V/I	None	SIFT
L92	CD163L1	Missense	H/Q	None	SIFT
L92	DCDC2C	Missense	V/L	Deafness	SIFT
L92	TDRD15	Missense	T/I	None	SIFT
L92	ZNF34	Missense	K/E	None	PolyPhen
L92	ZSCAN22	Missense	K/R	None	PolyPhen
H91	KIAA1024	Missense	S/I	None	SIFT
H91	PAX7	Missense	R/H	Rhabdomyosarcoma 2	SIFT
I91	LRP2	Missense	R/L	None	SIFT

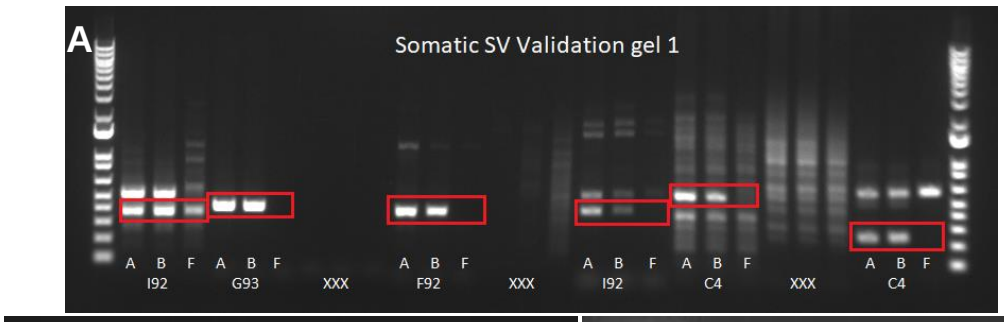
Table 2.3: An assessment of reprogramming-associated SNVs for potential impact using the same approach as Table 2.2. A little under half of all sister iPSC lines showed at least one deleterious SNV.

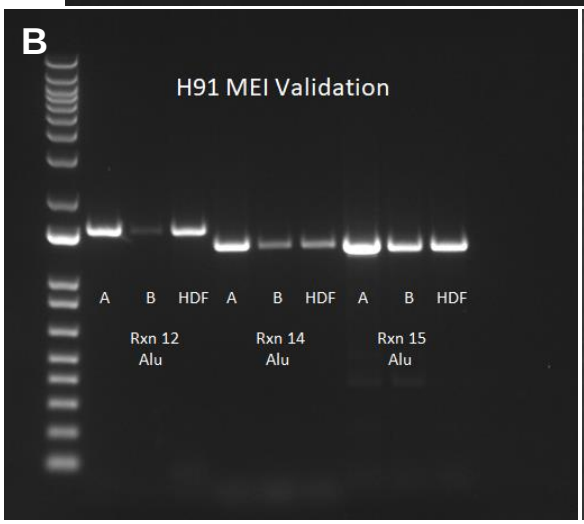
Table S1: Reprogramming associated SNVs impacting gene function

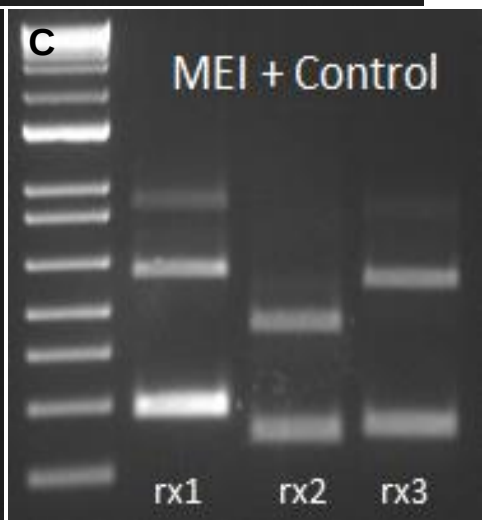
iPSC Line	Gene	Mutant Type	Amino Acid Shift	OMIM Disease	Algorithm
A1A	**Ano1**	**Missense**	**D/E**	**None**	**Both**
A1A	**Greb1L**	**Missense**	**F/L**	**None**	**Both**
A1A	**OS9**	**Missense**	**Q/H**	**None**	**Both**
A1A	**BHLH9**	**Missense**	**E/K**	**None**	**Both**
A1A	Golt1a	Missense	S/I	None	SIFT
A1A	DDI1	Missense	L/I	None	SIFT
A1A	RPL24	Missense	R/G	None	SIFT
A1B	Orc4	Missense	L/F	Meier-Gorlin Syndrome 2	SIFT
C4A	**IFT43**	**Missense**	**R/S**	**Cranioectodermal Dysplasia 3**	**Both**
C4A	EBF4	Missense	P/H	None	PolyPhen
C4A	Dock2	Splice Site	N/S	Immunodeficiency 40	SIFT
L92A	**ZC3H4**	**Missense**	**S/P**	**None**	**Both**
L92A	**ROR2**	**Stop Gained**	**--**	**None**	**--**
L92B	Taldo1	Missense	K/E	Transaldoase Deficiency	SIFT
I91B	Hoxd10	Missense	Y/S	Charcot-Marie tooth disease	PolyPhen
I91B	DGK2	Missense	A/T	None	PolyPhen
F92A	RECQL5	Missense	G/D	None	PolyPhen

2.3.6 Structural Variants and Mobile Element Insertions in iPSCs

SVs were called with LUMPY, a best in class SV caller which incorporates multiple methods of calling SVs from read pairs (76). High confidence somatic SVs were validated by PCR using primers that would only yield product in the presence of the SV (Fig. 2.16a). We called somatic SVs in 5 of 8 donor fibroblasts, all but one of which were validated by PCR (Fig. 2.16d). We did not call any high confidence reprogramming-associated SVs, but we attempted to validate low confidence reprogramming SVs and found that all were false positives (Fig. 2.16d). MEIs were called by MELT (77), and we performed PCR validation on a subset of MEIs from two lines, H92A and L92A, by designing primers that spanned the insertion. This should give two products, a wild type band and a higher weight MEI band. We ran PCRs for 20 putative MEIs but found a 100% false positive rate (Fig. 2.16b). To show that our approach was capable of validating MEIs we validated 6 germline MEI calls and found that these gave the expected product using our PCR approach (Fig. 2.16c). Because we didn't call any valid reprogramming-associated MEIs, we assessed our false negative rate by crossing our dataset with MEIs from the 1000 genomes project and determining how many of these MEIs were found in our fibroblast donors but were missed in one or more sister iPSC colonies (see methods A.1.4). From this we calculated a false negative rate of 1.6%. Together these data show that reprogramming does not significantly contribute to SV and MEI class mutations in iPSCs.







D Line	Somatic SV Validated	Reprogramming SV Validated (either sister)
I92 (lenti)	2/3	0/0
G93 (lenti)	1/1	0/1
F92 (lenti)	2/2	0/0
L92 (lenti)	1/1	0/0
H91 (lenti)	0/0	0/1
C4 (epi)	2/2	0/1
A1 (epi)	0/0	0/1

Figure 2.16: (a) High confidence structural variants were validated by PCR. Valid somatic SVs were found in both sisters (A,B) but not the fibroblast (F). Reprogramming SVs should be found only in one sister. (b) Mobile element insertions (MEI) called by MELT were validated with PCR primers spanning the insert. A valid MEI will show two bands; a wildtype and an insert product shifted up 250-350bp. This is a representative gel of a few validation reactions. (c) We validated several germline MEIs and saw the expected two products. (d) All but one somatic SV was validated by PCR, however none of the 4 called reprogramming SVs were validated.

2.3.7 Application of sister cell paradigm data to other approaches

Although the sister iPSC paradigm presents a powerful approach for assessing somatic and reprogramming-associated mutations, it is technically challenging and requires considerable labor to obtain and validate sister iPSC pairs. We thus sought to use our data to create a general framework to quickly determine the contribution of reprogramming vs somatic mutations to an iPSC line. To accomplish this, we combined all high confidence SNVs for all real sister lines (somatic and reprogramming-associated), and sorted this dataset into VAF bins of 0.02 (VAF 0.200 - 0.219, 0.220 - 0.239, etc…). We then assessed the relative contribution of somatic and reprogramming-associated mutations to the total number of called mutations in each of these bins (Fig. 2.17a). This gave us a "reprogramming coefficient" and a "somatic coefficient" for each VAF bin (Fig. 2.17b). Any dataset can be divided into VAF bins, and the number of somatic or reprogramming-associated mutations can be estimated by multiplying the total number of SNVs in each bin by the associated somatic or the reprogramming coefficient for that bin. To test the accuracy of this approach, we ran our own data through this pipeline ignorant of sister status. We found that our pipeline was able to predict somatic and reprogramming-associated SNVs to a reasonable degree of accuracy (Fig. 2.17c). Importantly, our pipeline showed that episomal reprogramming resulted in significantly more reprogramming-associated mutations, even without a priori knowledge of which mutations come from reprogramming and which from the original fibroblast (Fig. 2.17c). This dataset will allow researchers to quickly assess the mutagenicity of novel reprogramming methods by separating reprogramming-associated mutations from somatic ones, allowing them to rule out somatic background noise.

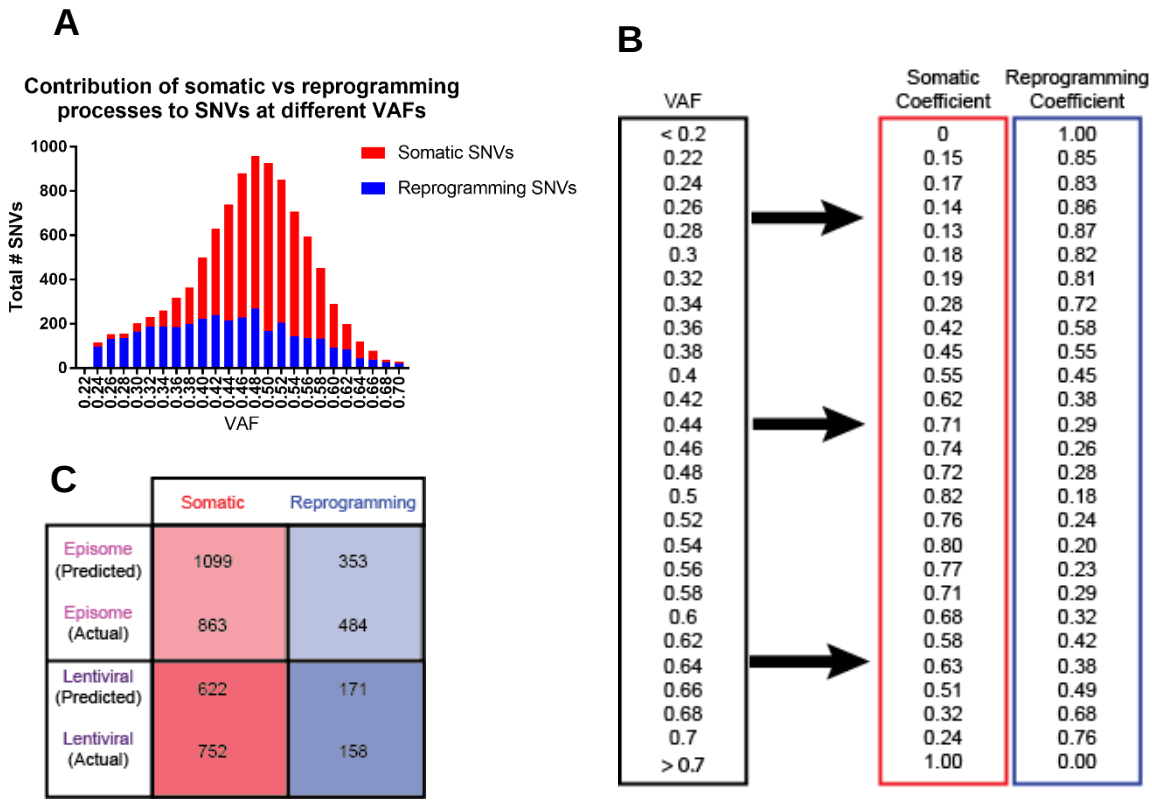


Figure 2.17: (a) The total number of high confidence somatic and reprogramming SNVs for all sister lines were plotted along VAF bins of 0.2. (b) At each VAF bin, we calculated the % contribution for somatic and reprogramming processes to generate a constant coefficient for each process at any given VAF. (c) We analyzed our data using these coefficients to predict average reprogramming and somatic mutations per line ignorant of sister status, then compared these data to the averages observed using our sister paradigm.

2.4 Discussion

Our sister iPSC paradigm provides a rigorous system for definitively establishing the mutagenicity of reprogramming, and for determining the mutational burden of the original somatic cell. Understanding the number of mutations caused by different reprogramming methods can inform clinicians and researchers as they pursue iPSCs as a tool for disease modeling or regenerative medicine. In the disease modeling field, it isn't uncommon for studies to be published with results based on one or two iPSC lines, which are often characterized by sequencing the donor fibroblast population, if whole genome sequencing is conducted at all (see Ebert, Liang, and Wu, 2013, Table 1 for a collection of many such studies)(78). Our data show that this approach is insufficient; reprogramming can induce hundreds of mutations, and the initial cell can itself possess hundreds of mutations which are not captured in the donor sequence.

In considering our observation that episomal reprogramming resulted in more SNVs than lentiviral reprogramming, there are several points that should be emphasized. Most importantly, our episomal condition used a different set of transcription factors (Oct4, Klf4, Sox2, L-Myc, lin28, p53 shRNA) than our lentiviral condition (Oct4, Sox2, Klf4, and C-Myc). We cannot distinguish between the effects of different delivery methods and different transcription factors. Bhutani et. al. (2016) compared retrovirus, Sendai virus, and mRNA all using Pou5F1, Sox2, Klf4, and C-Myc, and found no differences in total number of mutations in the iPSCs (35), though it is possible that the differences were obscured by somatic noise as this study didn't distinguish reprogramming mutations from more numerous somatic variants. Bhoutani's study indicates that our results might be due to different transcription factors rather than different methods of delivery. However, Kwon et. al. (2017) reprogrammed fibroblasts with episomal vectors using the same combination of factors we use in our lentiviral condition and found an

enrichment of C → A in exome (37). This is in agreement with our episomal data, and indicates that it is the episomal delivery method itself which is the source of additional mutations over the lentiviral method. Because neither of the above studies (and no study to date) performed whole genome sequencing with a means to distinguish reprogramming from somatic mutations, it is at present impossible to resolve these conflicting possibilities. Adding to the complexity is the fact that the lentiviral factors were inducible by Dox while the episomal vectors were constitutively expressed. Because cells were not switched to iPSC culture media until 3 days post transfection, this means the episomal lines were initially reprogrammed under conditions of fibroblast culture, while the lentiviral lines were reprogrammed under iPSC culture conditions (see Fig. 2.2). Although this design was necessary given the difficulties we encountered in establishing sister iPSC colonies, it precludes us from knowing which aspect of reprogramming contributed to the differences observed between episomal and lentiviral conditions. What we can say with some confidence is that different reprogramming conditions can lead to different degrees of mutational burden in the final iPSC line, and novel methods of reprogramming should be assayed for mutagenicity before being adopted for widespread use.

Not only does the reprogramming condition impact mutation in iPSCs, but out second key observation is that the choice of donor cell also impacts mutational burden. Early studies on iPSCs made assumptions that donor cells had no or very few mutations that were not accounted for by sequencing the donor individual. While Kwon et. al. (2017) showed by targeted deep sequencing that many iPSC exome mutations derived from the donor cells (37), we show here for the first time the full contribution of somatic mutations to the whole genome of iPSCs. Our findings of 250-1500 somatic SNVs are broadly in-line with a study assessing SNVs in single

human fibroblasts (79), and the high variance emphasizes the degree of genetic mosaicism found in cell populations, as well as the difficulty in assessing reprogramming mutagenicity with somatic background noise. Researchers and clinicians should be cognizant of how many sources of mutation their initial samples may have been exposed to. Lo Sardo et. al. (2016) showed that exome mutations in iPSCs increase with the age of the donor (36), while a study of iPSCs from human cord blood showed roughly 2 exomic SNVs per line, far less than reported in most iPSCs derived from fibroblasts (80). Selecting younger and less mutagenized samples, wherever possible, will limit the impact of somatic mutations on downstream applications of iPSCs.

The third key observation of our study is that the increased mutations evident in episomal reprogramming are not randomly distributed, but are biased toward C → A and specific trinucleotide contexts which implicate specific cellular processes. There are several possible explanations for this, though we stress that further research is required to definitively assess mechanisms of reprogramming-associated mutation. C → A transversions are a hallmark of oxidative stress, associated with high metabolic function (81). It is noteworthy that the reprogramming process is known to induce a switch in metabolic profile from oxidative phosphorylation via the Krebs cycle to oxidative glycolysis (82). This is thought to be a genome protective measure, synthesizing the large number of metabolic intermediates required in pluripotent cells while minimizing oxidative stress (83). It is also noteworthy that C-Myc is a key promoter of glycolysis (84, 85), and one of the differences in our episomal condition was the absence of c-Myc. We instead used L-Myc, which is known to promote proliferation similar to the other Myc members, but has not been shown to play any role in promoting glycolysis. Thus it is possible that different kinetics in metabolic switching between our reprogramming methods

could result in an increase in DNA damage enriched in C → A, though this is speculative. Our trinucleotide and mutational signatures analysis gave additional insight into potential mechanisms of mutagenesis. Episomal reprogramming SNVs are enriched for signatures 18 and 24, while lentiviral reprogramming SNVs are enriched for signatures 5 and 29. Though little is known about signature 18 beyond its prevalence in neuroblastoma, signature 24 is listed in the COSMIC database as: "Exhibits a very strong transcriptional strand bias for C>A mutations indicating guanine damage that is being repaired by transcription-coupled nucleotide excision repair (tcNER)." Despite the proposed role for transcription-coupled NER in episomal mutations, we found no evidence of enrichment in introns or coding exons, and found that mutations were significantly depleted in regions with histone marks associated with active transcription. One possibility is that our histone mark database does not accurately reflect early reprogramming iPSCs. Performing RNAseq on our iPSCs for both episome and lentiviral conditions, and then testing for enrichment of SNVs in actively transcribed genes, could resolve this discrepancy. Ultimately, it seems that different methods of reprogramming introduce mutations via mechanisms that are distinct from one another, though what these mechanisms are remains an unresolved question.

We found that 7/8 sister pairs had at least one SNV predicted to be deleterious by SIFT or PolyPhen, stemming from the original fibroblast but not caught by bulk sequencing of the donor cell population. The impact of these mutations depends on the downstream application of the iPSC line; C4A/B have a deleterious mutation in NRL, a transcription factor involved in differentiation of rod photoreceptors (86). This makes C4A and C4B poor choices for applications involving the rods, however these lines would be acceptable for use in cardiac

research, as they seem to have no mutations in genes associated with cardiomyocytes. Reprogramming contributes to potentially deleterious mutations in 7/16 lines, several of which are also associated with human disease. These data show the importance of thoroughly characterizing an iPSC line before using it in the clinic or the lab, rather than relying on available sequencing data for the donor or the initial cell culture.

A careful assessment of SVs and MEIs found no evidence that reprogramming contributed to these classes of mutation, providing a valuable data point amidst a sea of contradictory literature (see section 1.1.3). These conflicting studies are a result of the fact that both SVs and MEIs pose unique challenges for variant callers. Because SVs are by definition greater than the read length of the sequence, SV mapping algorithms are unable to accurately map reads within an inversion, as they cannot distinguish wild type from inverted sequence without a breakpoint. They also struggle to determine the origin of duplicated sequence reads (did the read come from the original or duplicated sequence?), and can become confused by multiple types of reads supporting an SV. Our approach utilizes a highly sensitive algorithm that incorporates multiple sources of SV detection, as opposed to earlier studies which generally relied on a single detection method (76). Similarly, while MEIs are traditionally difficult to call and validate for a variety of reasons (60), we have made every effort to observe best practices. The recently reported MELT tool, developed for the final phase of the 1000 genomes project, has been shown to be a best in class MEI caller (77), and indeed we find a false negative rate of only 1.6% using MELT's pipeline. We hand validated a subset of MEI calls with PCRs that span both junctions of the putative insertion, and found a 100% false positive rate. Importantly, as a positive control we were able to validate all tested germline MEIs with this method. Taken

together, we are confident in concluding that our reprogramming methods do not appreciably contribute MEIs or SVs to the iPSC genome.

The total mutational burden of an iPSC colony can have major implications in both clinical and research contexts. We show that the early stages of reprogramming is significantly more mutagenic than would be expected by traditional mitotic processes, and that these mutations arise from mechanisms that differ from those present in the parent cell. While these mutations are depleted in exons, several of them are predicted to have a deleterious impact on gene function, and mutations in non-coding regulatory regions can significantly impact physiology. Even so, we find little evidence of large deleterious variants arising from reprogramming, nor do we find predicted deleterious mutations in notable oncogenes. Taken together, our study provides several points of best practice which should be considered when working with iPSCs. Namely, care should be taken in selecting the reprogramming method and the donor cell population, and resulting colonies should be thoroughly characterized prior to use in downstream applications. Accounting for genomic integrity will allow researchers and clinicians to safely use iPSCs to their full potential.

Data from Chapter 2 have been prepared for submission. The material as it may appear in print is: Duran M.A., Lo Sardo V., Hazen J.L., Nair R.V., Kanchi K., Lala S., Tu N., Hall I.M., Baldwin K.K. “The Impact of Different Reprogramming Methods on Human Induced Pluripotent Stem Cell Genomes.” Cell Stem Cell. The dissertation author was the primary author of this paper. Dr. Valentina Lo Sardo and the dissertation author contributed equally as the primary investigators of this project.

Chapter 3:
Single Cell Mutational Burden in Young and Old Rod Photoreceptors

3.1 Introduction

Of all the cells in the body, perhaps none are as valuable and unique as neurons. Intimately linked with the attributes we most clearly associated with sentient life, most neurons are derived early in an organism's life and persist for many years. A neuron's ability to survive a century or more is made all the more remarkable when one considers the relentless onslaught of mutagens that assail a neuron's genome even in the absence of DNA replication. As was discussed in Chapter 1, neurons suffer double stranded breaks (DSBs) in response to neural activity, probably as a means of allowing quick activation of early response genes (48, 49). DSBs are particularly harmful for postmitotic cells because they lack the ability to repair them through homologous recombination (HR), which requires a duplicated genome in S phase (87). This forces neurons and other postmitotic cells to rely on nonhomologous end-joining (NHEJ), which is versatile but error-prone (87, 88). In addition, it has long been known that neurons have a highly active metabolism, producing a steady stream of oxidative stress (89). The impact of these factors is evident in aged human neurons, which have well over a thousand somatic SNVs, more SNVs than are found in our dividing fibroblasts from Chapter 2 or in most other normal single cell datasets (65). The most thorough human study to-date, however, found a two-fold difference in mutation rates between cortical neurons and neurons from the dentate gyrus ((65) see Chapter 1.1.4). Several questions remain unanswered in this important field, among them are; (1) do other populations of neurons possess significantly different rates of mutation, and if so why? (2) To what extent do mutations impact cognition in aged individuals? (3) Are aged mouse neurons comparable to aged human neurons in terms of mutational burden, and how might this impact mouse models of cognitive age-related disease? To begin to address these and other questions,

we expanded on an innovative approach pioneered by our lab and first published by Hazen and Faust et. al. (2016)(59). Here, we utilize somatic cell nuclear transfer (SCNT) to study the complete spectra of mutations in young, middle aged, and old rod photoreceptors from mouse retina.

3.2 Results

3.2.1 Establishing the extent of SNVs in rod neurons of different ages

To obtain a pure population of rod photoreceptors, we bred mice with GFP tagged neural retina-specific leucine zipper protein (NRL), which is a specific marker for rods (90). To amplify these post-mitotic genomes without PCR, we initially utilized somatic cell nuclear transfer (SCNT) with the aim of collecting two cell embryos or establishing an embryonic stem cell line (NT-ESC). This approach was previously used to establish NT-ESC lines from mitral and tufted neurons of young mice (Hazen and Faust et. al., 2016, (59)). Retinae were dissected from mice of various ages and a single neuron suspension was prepared for injection (Fig. 3.1a,b). We performed over a thousand neuron nucleus injections via SCNT (Fig. 3.1c), and initially established 3 NT-ESC lines from p6 mice. As we began experiments on older mice, however, we observed a decrease in reprogramming efficiency of aged neurons, with fewer of these embryos developing morula/blastocysts (Fig. 3.1d). We further noted that even in the case where aged neurons generated blastocysts they generally appeared less healthy than their young counterparts. It has been reported that excessive epigenetic marks can impair reprogramming efficiency (91), and that aged brains show hypermethylation (92). Therefore we attempted to enhance our efficiency by cloning Kdm4d, a histone demethylase that was reported to enhance reprogramming of SCNT in fibroblasts and cumulus cells by direct mRNA injection (93). Despite numerous attempts at optimization, however, we did not observe enhanced reprogramming efficiency via this method. We attempted to collect the embryo at the 2 cell stage, reasoning that we could use single cell amplification on each cell of a 2 cell pair, calling only mutations that are found in both samples (and thus came from the original neuron), an approach similar to our sister iPSC strategy outlined in Chapter 2. We developed a method for dissociating the 2 cell embryo which involved removing the zona pellucida via Tyrode's solution

followed by vigorous trituration in trypsin by mouth pipette, and we collected a number of sister pairs with this approach (Fig. 3.2a). However the single cell amplification technology at the time was not optimized for SNV detection, and the uneven genome coverage made variant calling extremely difficult (Fig. 3.2b). To examine whether our collection method was leaving residual reagents that might interfere with amplification, we tested our collection method on a control blastocyst and saw even coverage after amplification (Fig. 3.2c).

In reconsidering our approach, we noticed that we were able to develop blastocysts even when reprogramming aged neurons (albeit at a lower efficiency), and we reasoned that the number of cells in a blastocyst (150-250) would be sufficient for high quality whole genome sequencing, although unlike with an NT-ESC line we would not have an unlimited source of DNA. We collected blastocysts and split them prior to amplification by multiple displacement amplification (MDA), allowing us to rule out any mutations arising from amplification and library prep by only calling mutations found in both halves of the blastocyst (Fig. 3.3). We sequenced an initial set of blastocysts in this way and noticed that the data varied greatly in quality (Fig. 3.4a), likely due to the differences in health of the original blastocysts. To screen out poor quality blastocysts prior to WGS, we developed a PCR based quality control panel which tested the capacity of the amplified DNA to PCR amplify various loci across the genome (Fig. 3.4b). We assigned each blastocyst a quality score based on this PCR, with samples receiving 0, 1, or 2 points for no band, a faint band, or a strong band at each QC loci. We sequenced only those blastocysts which showed a quality comparable to that of our high quality blastocysts from our initial sequencing. Altogether we sequenced 3 NT-ESC lines from p6 rods and 10 blastocyst pairs from rods aged p20 to 2.5 years (Fig. 3.4c).

This dataset was processed as described in chapter 2. We included only SNVs called by both halves of a blastocyst but not found in the bulk control for that mouse. To reduce common sequencing artifacts, we further filtered out any SNVs found in more than one sample. We included a VAF cutoff of 0.3 based on our prior findings that most somatic mutations are found above that VAF. The three ES lines required a slightly different approach because of their high depth compared to our blastocysts samples (60x vs 30x average depth), which resulted in false positives even at high quality scores (which are in part based on the number of reads supporting a call). These samples were analyzed by our collaborators in the Hall lab using the approach reported by Hazen and Faust et. al. (59). We estimate roughly 109 (78-116) SNVs in p6 rods, 151 (113-224) SNVs in p21, and on average 226 (164-273) SNVs amongst older rods (Fig. 3.4d). Very old rod photoreceptors thus seem to have significantly more SNVs than found in p6 rods as assessed by T test and one-way ANOVA (Fig. 3.4e).

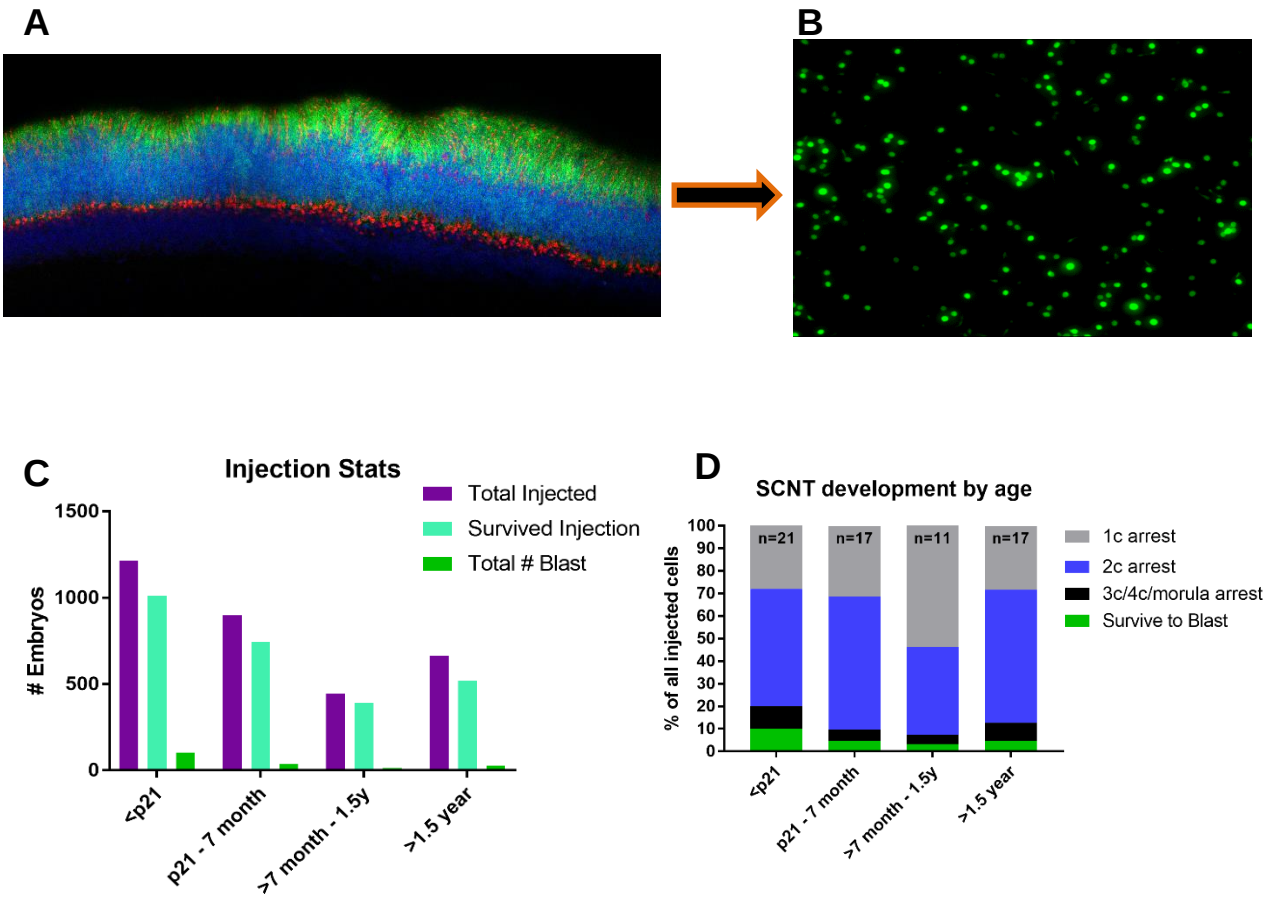
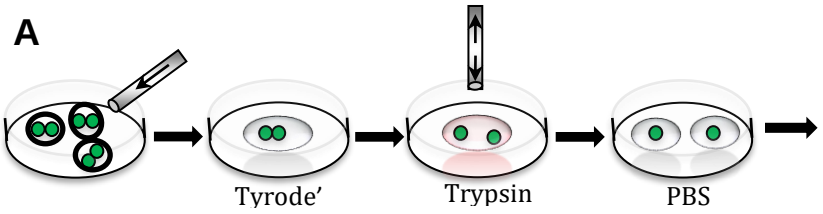


Figure 3.1: (a) To confirm accurate labeling, retinae were sectioned and stained for markers of cones, glia, and proliferative cells. No overlap was observed with GFP marking rod photoreceptors. Shown above is a zoomed out image of retina from an NRL-GFP mouse co-stained with cone arrestin (labels cones) and DAPI. (b) Neuronal quality was examined under fluorescent dissecting scope prior to injection to confirm cells were not lysing (such cells appear swollen, while the healthiest cells sometimes have processes attached, though this was less commonly observed in rods than other neuron types.) (c) Observation of the fraction of embryos surviving the injection process for all NT experiments, separated into different age groups. (d) Observation of efficiency of development as a function of age. N is the number of individual mice used per condition.

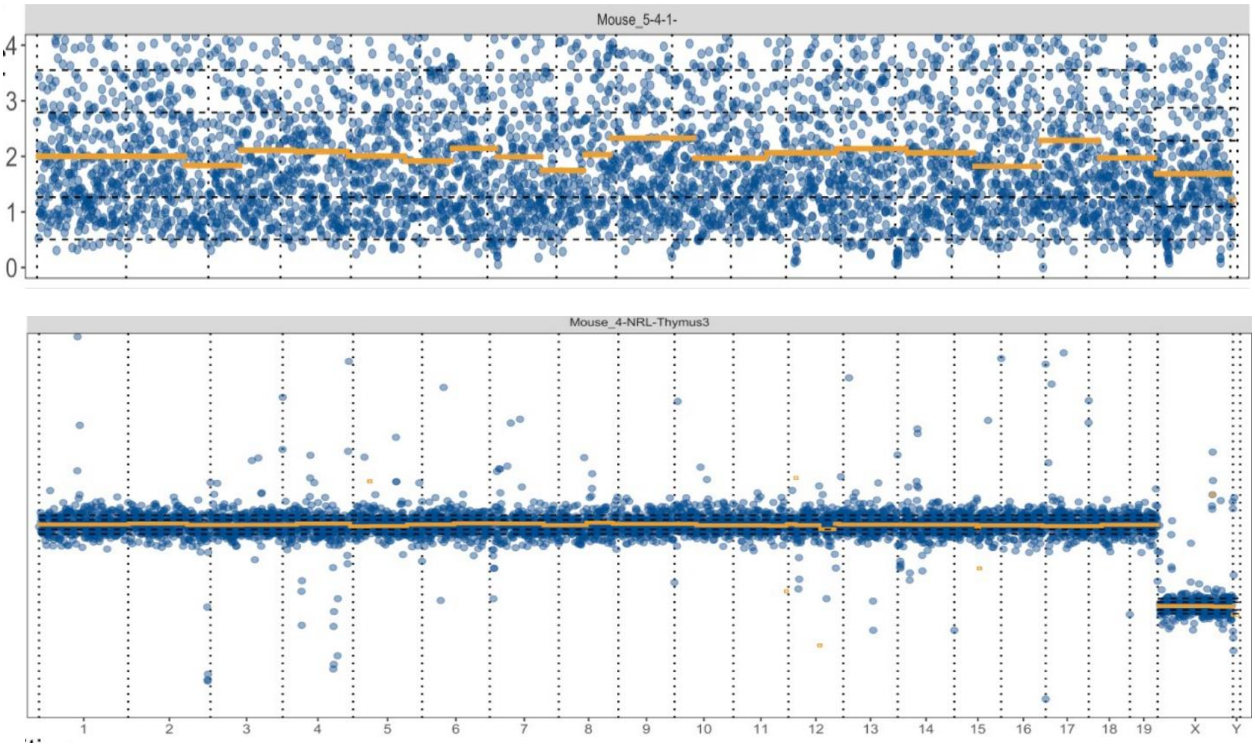
Figure 3.2: (a) An outline of our approach for dissociating 2c embryos into single cells. Cells were transferred between droplets by mouth pipette, with a 3x wash in M2 media between each step. The # collected refers to the number of embryo pairs dissociated and collected. (b) The diploid copy number is shown on the y axis for each region of the genome along the x axis, where a standard diploid genome would center around 2. This analysis is traditionally used to assess copy number variants, which will appear as 3x copy number (duplication) or 1x (deletion) on the y axis. Here it is used as a rough assessment for the evenness of amplification. The upper graph is the data from single cell amplification, compared to the lower graph of bulk control DNA. The yellow line denotes the average along the genome. An evenly covered sequencing dataset will show a straight yellow line except at regions of mono or polyploidy, as shown in the thymus female control. (c) Our control blastocyst halves showed an evenness that was sufficient for thorough analysis.

A

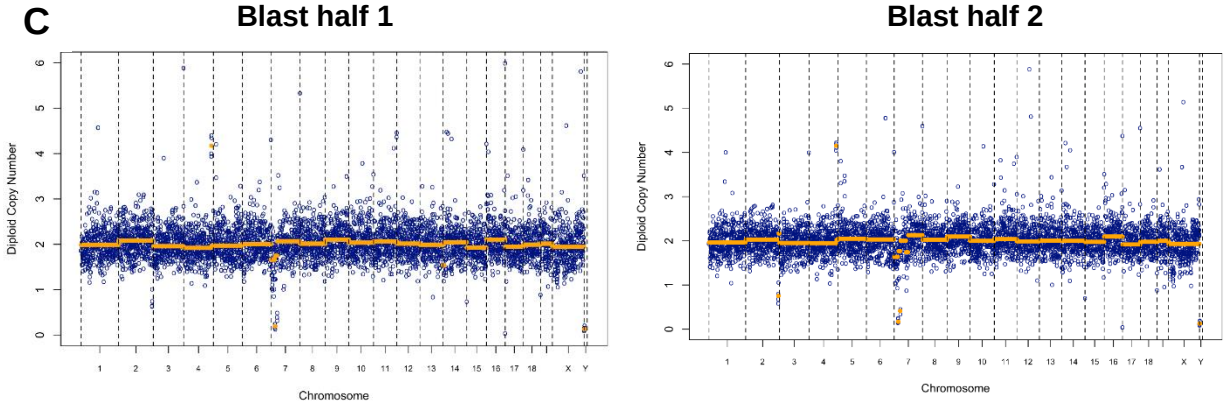


Neuron	Age Range	# Collected
NRL (Rod)	<p21	7
NRL (Rod)	>p21-1m	61
NRL (Rod)	2 year	31

B



C



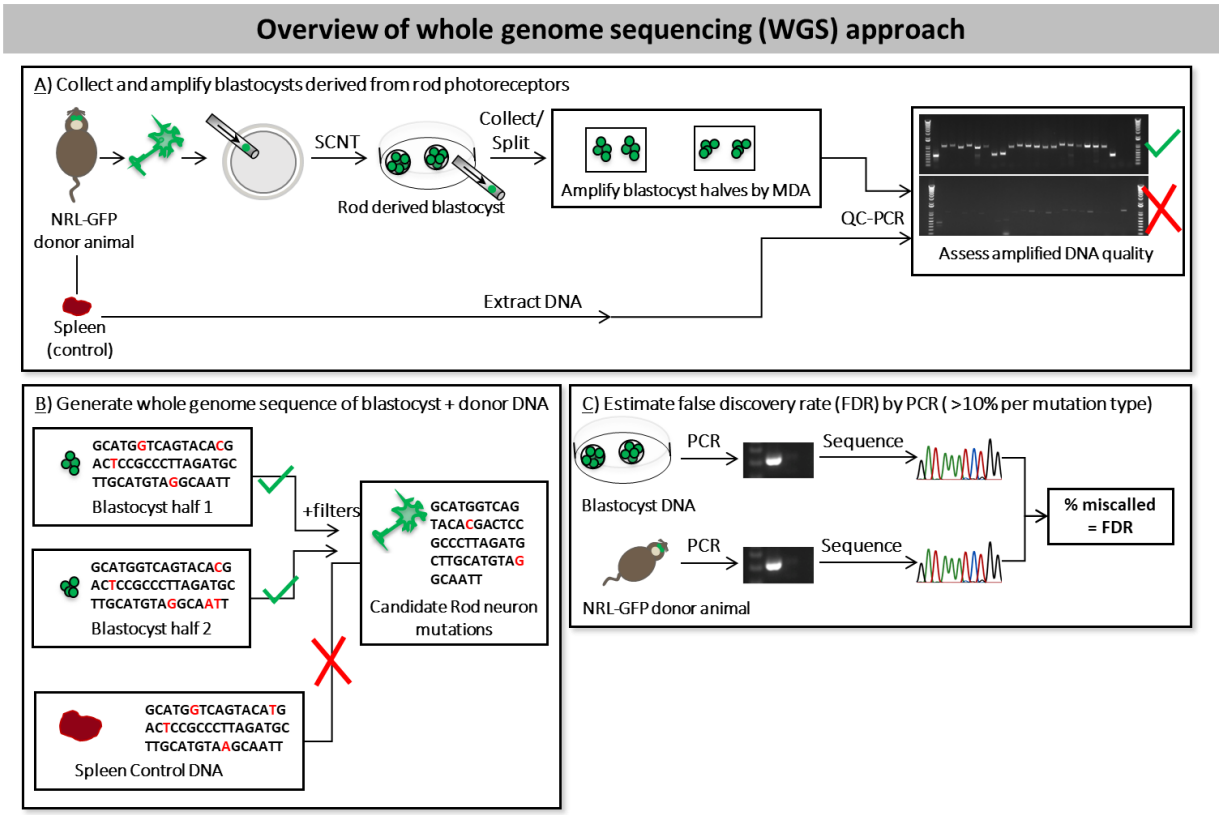
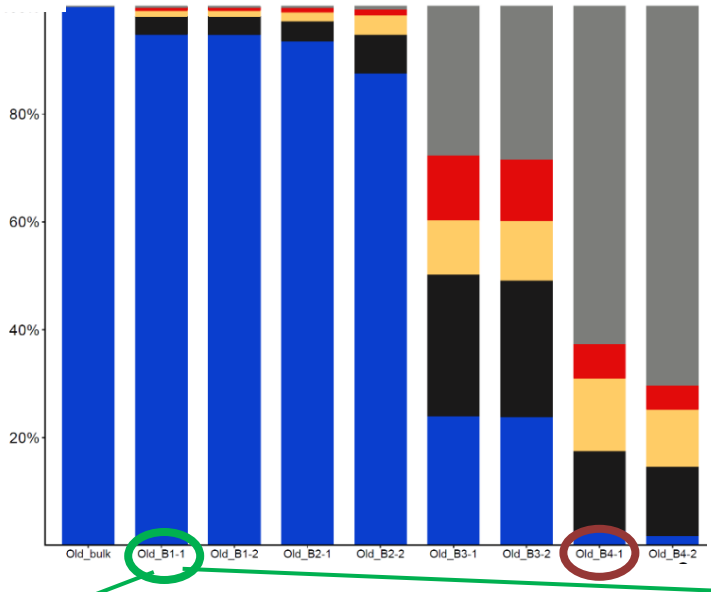


Figure 3.3: (a) Rod photoreceptors were collected from mice of various ages and their nuclei were injected into enucleated oocytes. Some of the fertilized eggs developed to blastocysts and were collected and carefully split into two tubes before being subjected to MDA based whole genome amplification. Amplified DNA was then tested for genomic integrity by PCR. (b) High quality DNA from the previous step was whole genome sequenced to 35x average depth, and mutations present in both halves of the original blastocyst, but not present in spleen control, were called candidate neural mutations. (c) A subset of candidate mutations were hand validated by PCR to assess the false positive rate of our variant calling pipeline.

Figure 3.4: (a) Genotype concordance, called for a preliminary set of amplified blastocysts. A set of germline mutations in the bulk control DNA (0/1 for heterozygous SNV) were called, and the percent those control SNVs found in each of our amplified blastocysts was plotted. A high quality dataset will make a heterozygous SNV call at > 80% of the control loci. 0/0 means no mutant SNVs were called, while 1/1 means a homozygous mutant call, and DP < 10 means the read depth threshold was below our filter cutoff. (b) We designed PCRs for 24 genomic loci and assayed how effectively our DNA samples were able to amplify them by PCR. We found a strong correlation between this quality control assay and our sequencing results for our preliminary blastocysts, so we used this assay on all future blastocysts prior to sequencing to select for samples which were likely to give good sequencing results. (c) A summary of the ages and sample names of our sequenced blastocysts. (d) The number of predicted SNVs for each sample by age, providing the average per age group at the bottom. (e) Average predicted SNV burden across three age ranges, significant differences calculated by ANOVA.

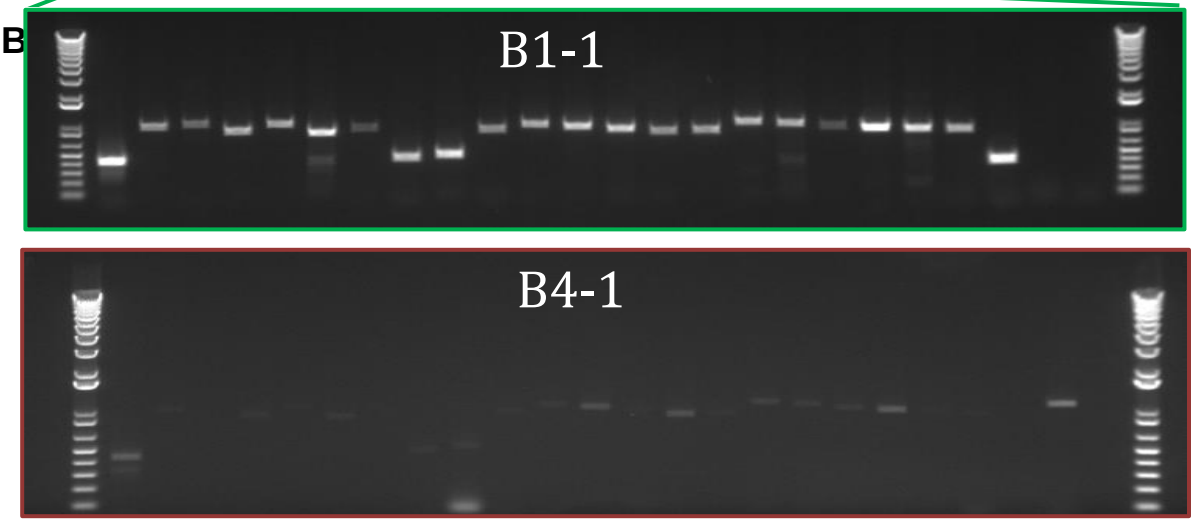
A



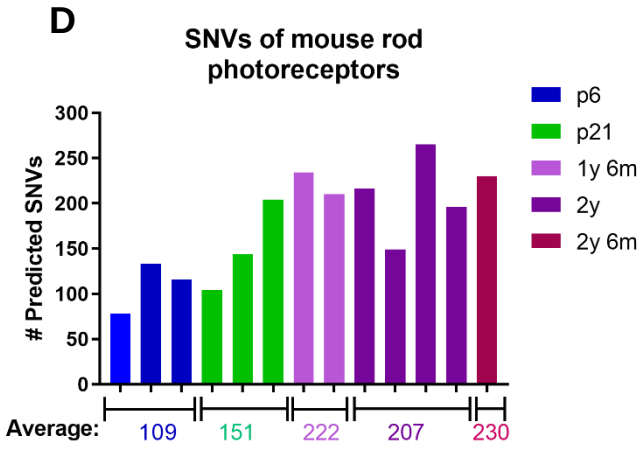
C

Sample ID	Mouse Age
ES1	p6
ES2	p6
ES3	p6
NY1-1	p21
NY1-2	p21
NY6-1	p21
NO11-21	1y 6m
NO11-22	1y 6m
NO1-1	2y
NO1-2	2y
NO4-8	2y
NO4-9	2y
NO8-17	2y 6m

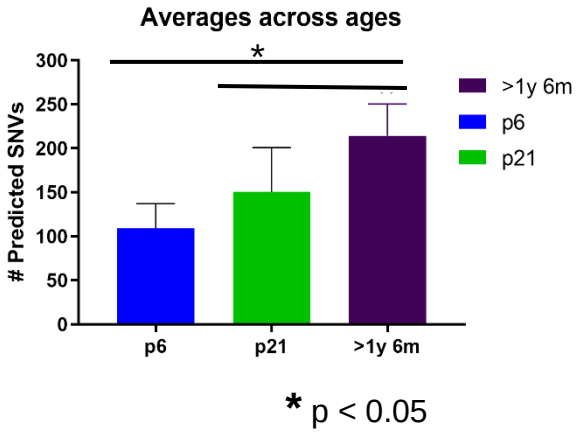
B



D



E



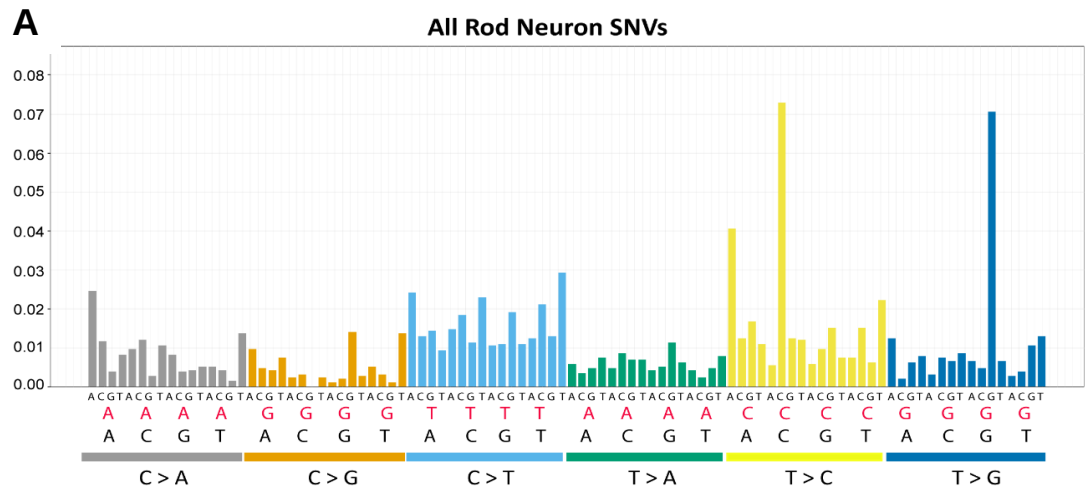
* $p < 0.05$

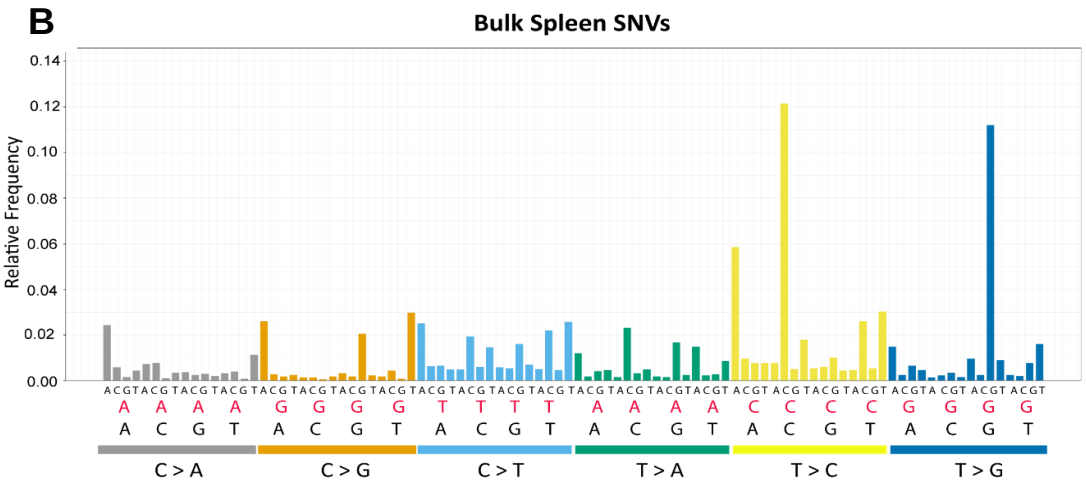
3.2.2 Nucleotide Context of SNVs in rod neurons

To determine possible sources of mutation in these neurons, we assessed the trinucleotide context of our highest confidence SNVs using DeconstructSigs and plotted the signatures of all neuronal SNVs vs the combined SNVs of our bulk spleen samples (Fig. 3.5a,b). We looked at the mutational spectra in each individual neuron and found that p6 SNVs had a greater proportion of A > G and A > C mutations compared to SNVs from older neurons, indicating mechanisms which could be active in development but not in mature, post-mitotic cells (Fig. 3.5c).

We assessed each context for differences between our neuronal and germline SNVs (Fig. 3.6). Upon close inspection, we found a handful of contexts which significantly differed; in C > T, neuronal SNVs were enriched in GTG mutations and TTG, while in T > C, neuronal SNVs were enriched in ACA and CGC contexts. Importantly, we previously reported enrichment in TTG contexts in mouse mitral and tufted cells (59), putatively linked to the activity of the deaminase APOBEC. Our findings here indicate that this might be a general mechanism of mutagenesis in neurons, though we can find significance in only one of the 4 common APOBEC contexts and further studies are required to definitively link APOBEC as the source of these mutations.

Figure 3.5: (a) High confidence SNVs from our rod sequences (all ages) were combined and their trinucleotide sequence context analyzed by DeconstructSigs. The nucleotide sequence is 5' > 3' from bottom to top, with the mutated nucleotide in red in the middle. (b) The same analysis on high confidence germline mutations using bulk spleen DNA. (c) The proportion of transitions and transversions for each individual sample.


A
All Rod Neuron SNVs
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
C > A
C > G
C > T
T > A
T > C
T > G


B
Bulk Spleen SNVs
Relative Frequency
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
C > A
C > G
C > T
T > A
T > C
T > G

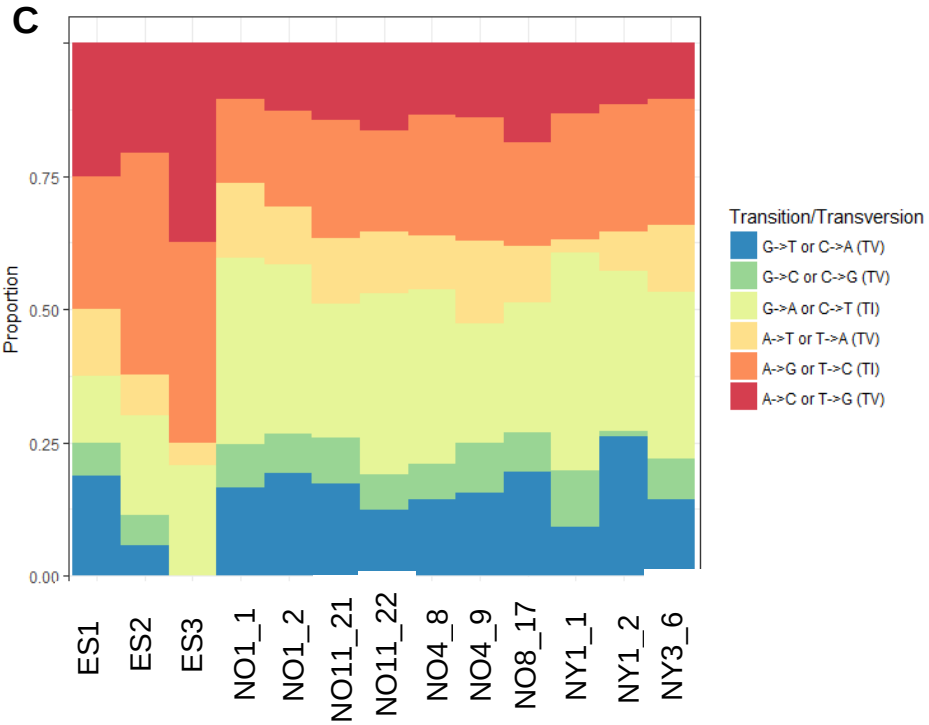
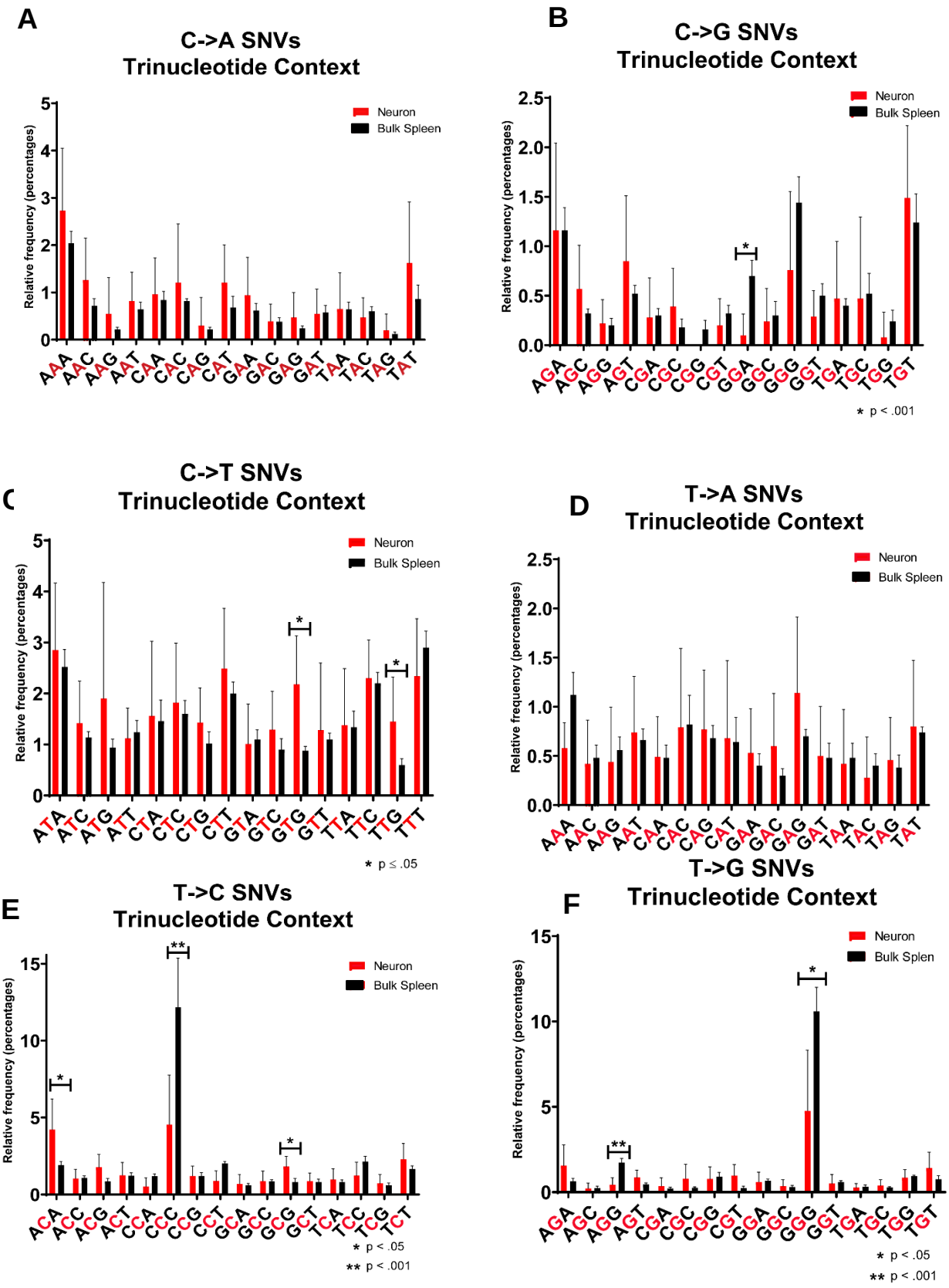

C
Proportion
0.75
0.50
0.25
0.00
ES1
ES2
ES3
NO1_1
NO1_2
NO11_21
NO11_22
NO4_8
NO4_9
NO8_17
NY1_1
NY1_2
NY3_6
Transition/Transversion
G->T or C->A (TV)
G->C or C->G (TV)
G->A or C->T (TI)
A->T or T->A (TV)
A->G or T->C (TI)
A->C or T->G (TV)

Figure 3.6: (a-f) Analysis of each class of trinucleotide context mutation for all samples as well as bulk spleen control (control spleen from all samples were analyzed together). Statistical significance was assessed by T test to find putative differences between neuron and bulk spleen contexts for each mutation category.


A
C->A SNVs
Trinucleotide Context
Neuron
Bulk Spleen
Relative frequency (percentages)
AAA AAC AAG AAT CAA CAC CAG CAT GAA GAC GAG GAT TAA TAC TAG TAT
B
C->G SNVs
Trinucleotide Context
Neuron
Bulk Spleen
Relative frequency (percentages)
AGA AGC AGG AGT CGA CGC CGG CGT GGA GGC GGG GGT TGA TGC TGG TGT
* p < .001
C
C->T SNVs
Trinucleotide Context
Neuron
Bulk Spleen
Relative frequency (percentages)
ATA ATC ATG ATT CTA CTC CTG CTT GTA GTC GTG GTT TTA TTC TTG TTT
* p ≤ .05
D
T->A SNVs
Trinucleotide Context
Neuron
Bulk Spleen
Relative frequency (percentages)
AAA AAC AAG AAT CAA CAC CAG CAT GAA GAC GAG GAT TAA TAC TAG TAT
E
T->C SNVs
Trinucleotide Context
Neuron
Bulk Spleen
Relative frequency (percentages)
ACA ACC ACG ACT CCA CCC CCG CCT GCA GCC GCG GCT TCA TCC TCG TCT
* p < .05
** p < .001
F
T->G SNVs
Trinucleotide Context
Neuron
Bulk Splen
Relative frequency (percentages)
AGA AGC AGG AGT CGA CGC CGG CGT GGA GGC GGG GGT TGA TGC TGG TGT
* p < .05
** p < .001

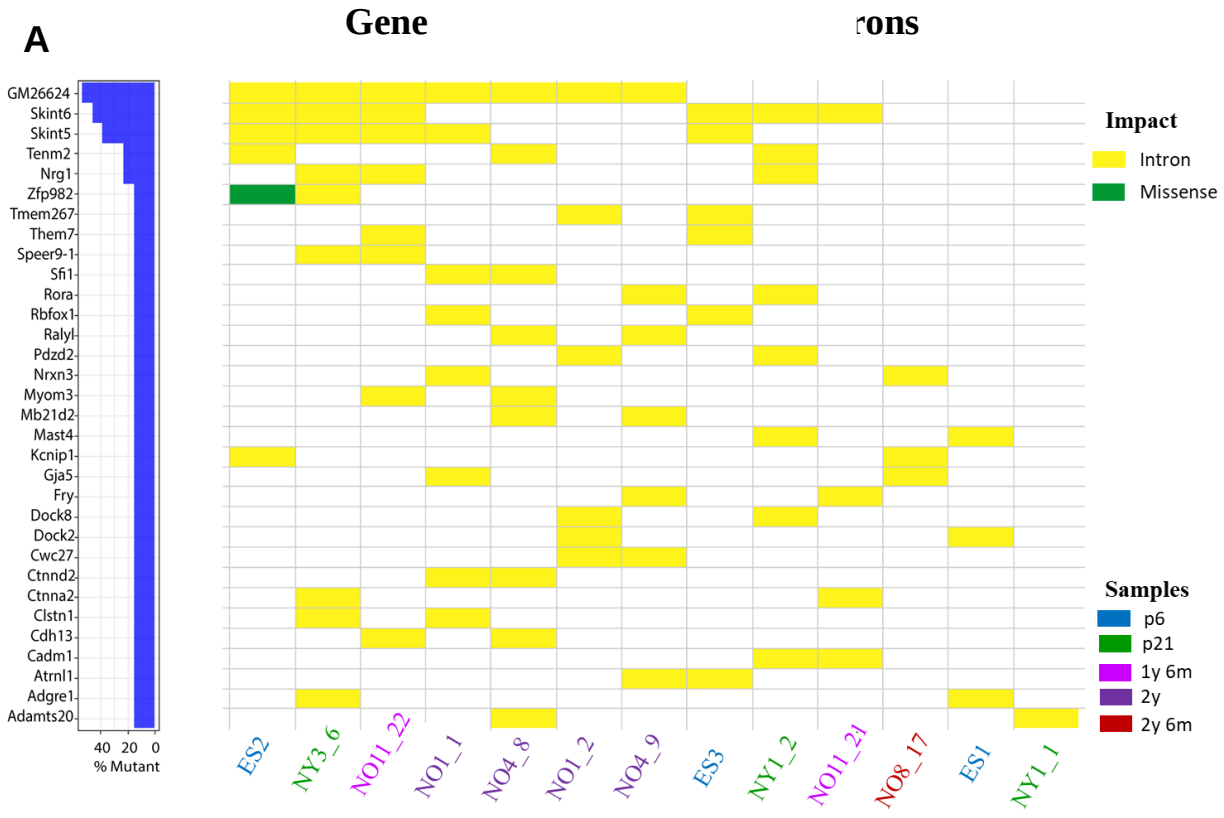
3.2.3 Functional assessment of rod SNVs

We wanted to know whether certain genes or regions were particularly susceptible to SNV mutation in our neurons, so we generated a waterfall plot to examine all genes with mutations in multiple neuronal samples (Fig. 3.6a). We found a predicted long non-coding RNA (lncRNA), GM26624, which was mutated in over half of our neuronal samples. Although GM26624 is long (320kb), these SNVs are not always randomly distributed, and accounted for 22 total SNVs across all neuronal samples (Table 3.1). In contrast we find only 12 randomly distributed SNVs in this region among all germline mutations, despite this dataset having an order of magnitude more SNVs than our neuron calls. Mutations in Skint6, Skint5, Tenm2, and Nrg1 did not differ significantly in total number from the bulk control.

Intrigued by the mutational hotspot observed in GM26624, we asked whether any samples showed other localized hypermutations and found that sample NO4_8 possessed 10 SNVs in Hsp90ab1, a heat shock protein highly expressed in rod photoreceptors. We plotted these mutations in the context of the protein and found that they all clustered in the C-terminal domain and were mostly missense mutations, though one introduces a stop codon (Fig. 3.6b).

To determine how many of our observed SNVs might functionally impair our neurons, we utilized an RNAseq dataset generated by Kim. *et. al.* (2016) to find genes with high expression in rods ((94), GSE74660). We took the top 50% expressed genes by normalized fpkm across three replicate RNAseq experiments of p6 rods and crossed it with our SNV list. We annotated the resulting dataset using variant effect predictor (VEP) and noted that 8/12 neurons have at least one potentially damaging mutation in a highly expressed gene (for the sake of simplicity we plotted only the most damaging mutation for Hsp90ab1) (Table 3.2).

Figure 3.7: (a) Waterfall plot showing genes with mutations in multiple samples. Each filled box represents a mutation of the labeled type corresponding to the gene on the y axis and the sample on the x axis. The translational effect is the number of mutations per MB of the sample genome for both synonymous and non-synonymous SNVs. (b) Lolliplot of Hsp90ab1 SNVs in sample NO4_8, which showed a large number of mutations in this one gene. The mutations cluster in the C terminal domain.


A
Gene
rons
GM26624
Skint6
Skint5
Tenm2
Nrg1
Zfp982
Tmem267
Them7
Speer9-1
Sfi1
Rora
Rbfox1
Ralyl
Pdzd2
Nrxn3
Myom3
Mb21d2
Mast4
Kcnip1
Gja5
Fry
Dock8
Dock2
Cwc27
Ctnnd2
Ctnna2
Clstn1
Cdh13
Cadm1
Atrnl1
Adgre1
Adamts20
40 20 0
% Mutant
ES2
NY3_6
NO11_22
NO1_1
NO4_8
NO1_2
NO4_9
ES3
NY1_2
NO11_21
NO8_17
ES1
NY1_1
Impact
Intron
Missense
Samples
p6
p21
1y 6m
2y
2y 6m

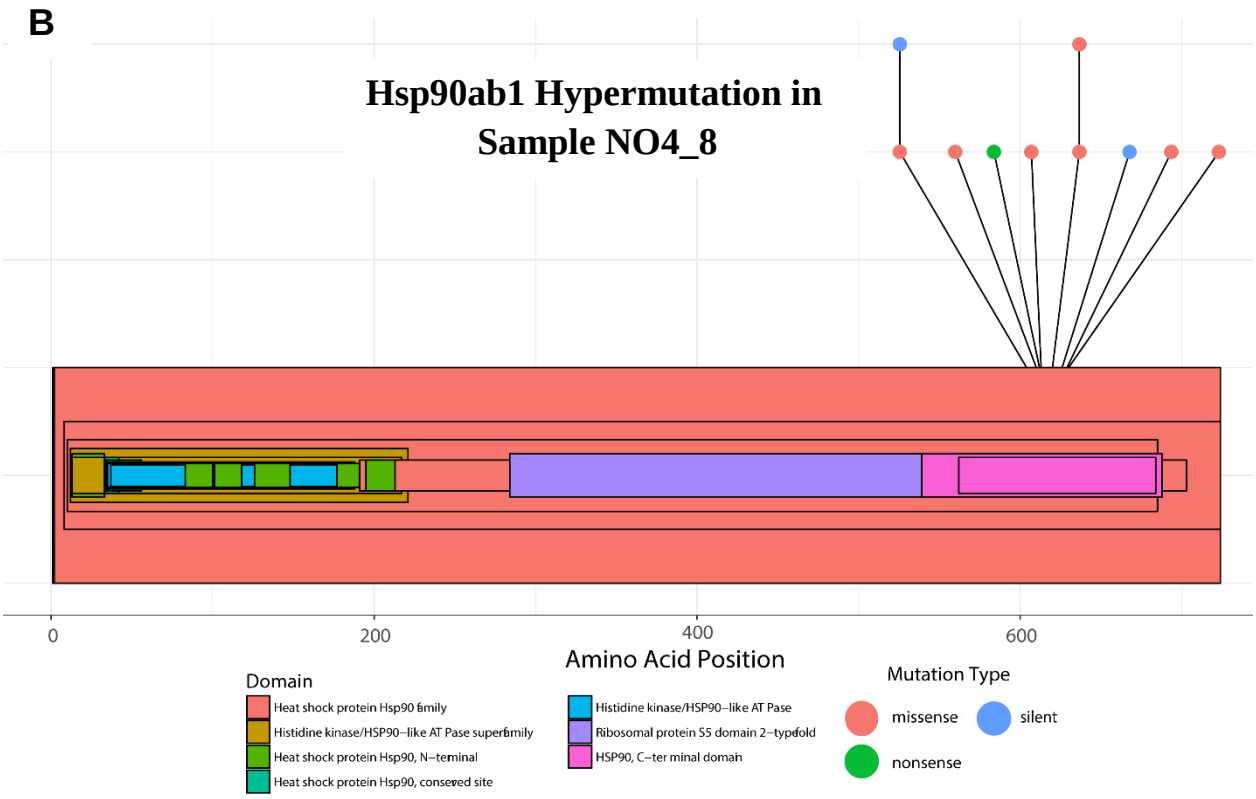

B
Hsp90ab1 Hypermutation in
Sample NO4_8
0
200
400
600
Amino Acid Position
Domain
Heat shock protein Hsp90 family
Histidine kinase/HSP90−like ATPase superfamily
Heat shock protein Hsp90, N−terminal
Heat shock protein Hsp90, conserved site
Histidine kinase/HSP90−like ATPase
Ribosomal protein S5 domain 2−type fold
HSP90, C−terminal domain
Mutation Type
missense
silent
nonsense

Table 3.1: (a) We manually inspected each neuronal SNV falling in GM26624 and sorted them by genomic coordinate, taking the distance between SNVs to assess whether they were evenly distributed or not.Context was taken for 5bp upstream and downstream of each mutation.

Sample	Coordinate	Ref	Alt	Distance from Prior	Context
NO1_1	147198706	G	A		GGGGGGAAAAA
NO4_8	147267020	T	G	68314	AATTGTAGGCC
NO4_8	147267022	G	A	2	TTGTAGGCCAC
NO4_9	147267037	G	A	15	AGTTAGTATCT
NO1_1	147272236	G	A	5199	TTGCCGCAGAA
NO1_2	147331917	A	G	59681	TACAAATGCAG
NO1_1	147336351	T	C	4434	TGTTGTCTGGA
NO1_2	147339535	G	A	3184	GAATGGTGGCT
NO1_2	147339539	C	T	4	GGTGGCTCACA
NO1_2	147339767	A	G	228	GCTGGAGGGTT
NO1_2	147339825	T	A	58	AACTCTCTGCA
NO1_2	147343184	T	G	3359	TTGTGTACAGT
NO1_2	147343204	T	G	20	GGCCTTCGTAC
NY3_6	147403855	G	C	60651	TGTCTGTCTCT
NO4_9	147405630	T	A	1775	TTTTTTAAGAT
NO1_1	147417529	C	T	11899	AATTACAATTT
NO1_2	147420400	C	T	2871	AGATGCTGTAA
NO1_1	147429091	C	T	8691	TCTCACGGTTG
NO11_22	147434719	G	T	5628	TGTCTGTATGG
NO11_22	147434744	C	T	25	ACATGTGTGTC
ES2	147448062	C	G	13318	TTTCTCTCTCT
ES2	147448068	C	T	6	TCTCTCTCTCT

Table 3.2: (a) Ensemble's variant effect predictor was used to assess SNVs in highly expressed genes. All missense and nonsense mutations were recorded, as well as mutations of unknown impact.

Table 3.2: Predicted impactful mutations in highly expressed genes

Sample	Gene	Mutation Type	Amino Acid Shift	Codon	Position (exon #/total)	Predicted Impact
NO4_8	Hsp90ab1	Stop Gained	R/*	CGA/TGA	10/11	Deleterious
NO4_8	Ip6k2	Missense	P/S	CCC/TCC	3/3	Deleterious
NO4_8	Gm10721	Missense	L/V	CTG/GTG	1/4	Deleterious
NO4_9	Mir149	Non-coding exon var.	-/-	--	-/-	Deleterious
NO4_9	Gm10719	Missense	L/W	TTG/TGG	3/4	Deleterious
NO4_9	Gm10720	Missense	S/R	AGT/CGT	3/5	Deleterious
NO4_9	Zfp933	Splice Site	-/-	--	-/-	Low/Med
NO8_17	Dynll1	LncRNA exon var.	-/-	--	-/-	Unknown
NO8_17	GM33676	LncRNA exon var.	-/-	--	-/-	Unknown
NY1_1	Ubr3	Missense	G/S	GCG/ACG	18/39	Deleterious
NY1_1	Gm10801	Missense	I/M	ATA/ATG	4/4	Deleterious
NY3_6	Map7	Non-coding exon var.	-/-	--	-/-	Unknown
ES1	Gm11168	Missense	R/T	AGG/ACG	2/6	Deleterious
ES3	Gm11168	Missense	F/L	TTC/CTC	2/4	Deleterious
ES2	Gm13152	Missense	K/Q	AAA/CAA	4/4	Tolerated
ES2	Zfp982	Missense	L/P	CTC/CCG	3/3	Deleterious
ES2	Gm13152	Missense	L/P	CTC/CCG	4/4	Deleterious

3.3 Discussion

In this study we looked at SNVs in young and old mouse rod photoreceptors at a single cell level. We first showed that we could derive blastocysts from single rod neurons as old as 2.5 years by SCNT, something not attempted in any neuronal subtype previously. We were able to derive ES lines from p6 rods with an efficiency of 0.2%, below the efficiency of mitral and tufted neurons (1-2%) and comparable to attempts on p6 cortical neurons (0.5%). Neurons in general exhibit lower reprogramming efficiency in our hands than for control cumulus (8%) and sertoli (7%) cells. Unfortunately, attempting to derive ESCs precludes the collection of blastocysts, and we were unable to derive any ES lines from mature rod neurons. This was possibly due to their inverted nuclear architecture, which places open chromatin on the periphery and condensed chromatin in the nuclear center, an evolutionary adaptation to allow light to more easily pass the retina in nocturnal mammals (95). This inversion takes place around p12, and indeed we observe a sharp decrease in development to blastocyst between p6 (10%) and p21 (4.7%), concomitant with a decrease in ESC development from 0.2% to 0. However the low efficiency even in p6 rods, before the inversion occurs, indicates that these neurons are particularly resistant to forming ES lines, despite the fact that they develop to blastocysts at a higher rate than MT neurons (10% vs 7%). Ultimately we collected data from 12 single rod neurons spanning ages p6 to 2.5 years, developing a set of QC metrics to limit amplification artifacts and ensure quality sequencing data.

We find between 78-265 SNVs per neuron, a wider range than reported in our previous work on MT neurons (50-112 SNVs). However our MT neuron dataset didn't look at any neurons older than 7 months, and our data indicates that these neurons are accumulating mutations as they age. Taking our data on p6 vs 2 year old neurons, we estimate these neurons

accumulate mutations at a rate of 30-50 SNVs/year, broadly consistent with a study of human cortical and hippocampal neurons which found annual rates of 23-40 SNVs/year (65). It is not surprising that our neurons should have a marginally higher rate of mutation; a recent study comparing mouse and human mutation rates found a higher somatic mutation rate in mice (79), though the study was not examining post-mitotic mutations. Because we examined a neuronal subtype never before sequenced at the single cell level in any species, however, we cannot distinguish subtype specific processes from species-level differences in genome maintenance.

Our data allowed us to assess putative sources of mutation in rod neurons. It is possible that APOBEC plays some role in neuronal mutations, as we found an enrichment of SNVs at TCG contexts compared to germline, consistent with APOBEC deaminase activity. APOBEC family enzymes normally converts C → U on mRNAs, but can erroneously act on DNA at TpCpN trinucleotides, resulting in a C → T transition (see chapter 1.1.1). Importantly, we found this signature in our previously published study on MT neurons (59). We also find C → T mutations in GCG contexts, though we could find no clear mechanistic explanation for this. While C → T mutations are a known artifact of MDA (96), the fact that we observe these transitions only in two specific trinucleotide contexts argues that it is a biological phenomena and not an artifact of amplification. In addition, we observe two signatures of T → C mutations at GTG and ATA trinucleotide contexts, which implicates fatty acid metabolism as its most likely source (97), though as with APOBEC we cannot definitively conclude that it is a causative mechanism. Similar to our previous mouse study, we did not observe an enrichment of C → A transversions associated with oxidative phosphorylation. A C → A signature was observed in aged human neurons however (65), which indicates that oxidative stress can play a role in

neuronal mutations, but the rate may low enough not to be evident after 2 years, or may not be operative in mouse neurons.

Our functional assessment indicated that our neurons acquire deleterious mutations in highly expressed genes, however we did not observe a significant difference in the number of functional mutations between young and old neurons. Intriguingly, we observed that a long non-coding RNA, Gm26624, was heavily mutated across multiple neurons compared to somatic controls. We also observed that these mutations tended to cluster non-randomly. Gm26624 is predicted to have 3 exons, generating a 3,769bp transcript of unknown function. Several whole-brain RNAseq studies have shown Gm26624 expression in neurons, however while an RNAseq study of p6 rods finds evidence for transcripts from this region, these do not appear to be highly expressed (86), indicating it may play a role at a different point in time. There is virtually no literature focusing on this genomic region, in neurons or otherwise, making it difficult to attribute a cause or effect to our observation. Further work in this region could reveal some functional role in the development or maintenance of photoreceptors, and reinforces the idea of using mutations as a tool to discover genomic regions of interest.

Less enigmatic is our observation of localized hypermutation in the heat shock protein Hsp90ab1. The phenomenon of multi-nucleotide hotspots has been previously reported to comprise up to 3% of de novo human SNVs (98, 99). These are regions of higher mutagenic activity than would be predicted by chance, (mean distance = 538bp, many within 20bp). These studies, however, have focused on germline mutations; to our knowledge our current data represents the first evidence that similar mechanisms are at work in a subset of neurons.

Hsp90ab1 is highly expressed in rod neurons at least up to p12 (86), and plays an important role in protein folding and signal transduction. Because the mutations observed are predominately deleterious, it is highly likely that Hsp90ab1 was nonfunctional in this aged neuron, which likely had a negative impact on the cell's ability to carry out its functions. Analysis of larger datasets could reveal the prevalence of these functional mutation hotspots in other neuronal populations.

Our observations of Hsp90ab1 dysfunction bring us to a broader question; what precisely is the impact of these mutations in the context of whole brain function? It is tempting to speculate that one aspect of age-associated cognitive decline involves the gradual abeyance of neural function as deleterious mutations accumulate in individual neurons, indeed recent work in humans has found that the SNV burden in 82 year old neurons of the prefrontal cortex is in the thousands, and though they did not characterize functional impact, simple chance dictates that at least some will be deleterious (65). On the other hand, a recent study showed that even early neural precursors possess an unexpectedly high mutational burden, in line with the Walsh lab's findings that even young neurons harbor hundreds of developmental mutations (100). No study to-date has definitively linked mutational burden of single neurons to age-associated cognitive decline, likely because of the immense challenges associated with such a study. Our data indicate that similar mechanisms of post-mitotic mutation are at work in mice, though given the lifespan of even the oldest mouse neurons, it is unsurprising that the total mutational burden is an order of magnitude lower than is observed in humans.

In favor of a functional impact of SNVs, young progeroid neurons have been shown to have a similar level of mutational burden as aged normal neurons (65), but testing causation is

challenging. There is evidence that mouse visual acuity declines with age (101, 102), and our data provide potential mechanisms for intervention in addition to bolstering global DNA repair pathways. It would be interesting to see whether shielding rods from mutational burden rescued some degree of visual function in old age, though even this would not be conclusive given the large differences we find in total mutational burden between aged mouse and human neurons. Ultimately our data show that post-mitotic mutations are not unique to humans, are caused by unique mechanisms of action, and can functionally impact neurons. Future studies should seek to address the cumulative impact of these mutations on organism function, possibly by using ours and others data to design intervention strategies aimed at limiting the most common sources of DNA damage in neurons.

Data from Chapter 3 include unpublished material that was coauthored with Rodriguez R.A., Kanchi K., Hall I.M., and Baldwin K.K. The dissertation author was the primary investigator for the work shown in Chapter 3

Chapter 4: Conclusions

Though the importance of mutations in cancer and inherited disease have long been understood, it is only recently that the scope of genetic mosaicism among single somatic cells has been grasped. Studies of disease modeling using iPSCs have often relied on available sequences of donor individuals, ignoring mutations accumulated by the individual fibroblast that gave rise to their projects. These same studies often rely on a single iPSC line, promulgating results which could be influenced by undetected mutations lurking in the genome. Even if a group were cognizant of the potential impact of somatic mutations, choosing best practices to avoid them is challenging. Differences in mutagenicity between different reprogramming methods can be easily obscured by the more numerous somatic mutations. Determining the sources of mutation in iPSCs requires a method for disentangling mutations present in the original cell from mutations arising during the reprogramming process. Here we reported on a unique paradigm which allows us to directly test the mutagenicity of reprogramming and the contribution of the donor fibroblast to the total mutational load of the iPSC.

By isolating pairs of iPSC colonies derived from the same donor fibroblast, we definitively show that reprogramming by OSKM with lentivirus induces approximately 158 SNVs and 9 indels per iPSC line on average. Similarly, we show that reprogramming by episomal vectors using a different set of factors induces 484 SNVs and 9 indels per iPSC line on average. By analyzing the variant allele frequency of these mutations we predict that they arise within the first 2 divisions after reprogramming. These data indicate that reprogramming is responsible for far more mutations than would be expected by mitotic processes, and that the choice of reprogramming method can have a significant impact on the mutational burden of the final iPSC line. The presence of 800 SNVs from the original fibroblast, missed in bulk

sequencing, is unsurprising given the literature on single fibroblast SNVs. Even so, studies continue to be published with little regard for the source of donor cells contributing to the iPSC lines under study. Researchers often select samples and methods based on ease of acquisition and their efficiency of reprogramming. While these are important considerations, researchers should also weigh mutagenicity in the balance. In cases where primary cells from an elderly patient are necessary, every effort should be made to secure the least mutagenized cells possible. The fact that we find unique mutagenic mechanisms at work in different reprogramming methods is evidence that the method of reprogramming can impact both the number and the character of mutations in the iPSC line. Researchers and clinicians should utilize a method of reprogramming limits exposure to mutations.

Although recent studies of human cortical and hippocampal neurons indicate that neurons accumulate post-mitotic mutations with age, it remains unanswered whether this is a general property of other neurons or species. We assess rod photoreceptor neurons in young and old mice and find evidence of post-mitotic mutations in these neurons, indicating that mice undergo a similar age-dependent accumulation of mutations. We find that several of these mutations are predicted to be deleterious and in highly expressed genes, including one aged neuron exhibiting localized hypermutation in a heat shock protein with a variety of important roles. Our finding that neurons accumulate functional mutations poses an intriguing question about their role in the gradual cognitive decline that is associated with aging. Although our data are not sufficient to provide a definitive answer, the findings here suggest that mice might be viable, if imperfect, model to begin addressing these questions. Future studies will need to adopt large-scale approaches to examine differences in mutation rates of different neuronal populations, and to

map these mutational burdens to functional impact on the cell and ultimately the cognition of the animal.

Appendix A

A.1 Methods for Chapter 2

A.1.1 Deriving iPSC sister lines

NB: Section A.1.1 was written and performed by Dr. Valentina Lo Sardo. It is being included here with her permission.

iPSCs were generated from human dermal fibroblast (Cat. #2300 ScienCell Research Laboratories) via Yamanaka episomal-based and lentiviral-based reprogramming.

For episomal-derived iPSC: fibroblasts were cultured in DMEM supplemented with 10% FBS, glutamax, Pen/Strep, NEAA. $5x10^5$ cells were transfected (Amaxa Nucleofector Technology) with plasmids containing the reprogramming transcription factors (pCXLE-hOCT3/4-shp53; pCXLE-hSK; pCXLE-hUL), 1ug each plasmid. Human Embryonic Stem Cell kit 1 (Cat#VPH-5012) was used with program P-022 on an Amaxa Nucleofector II device. 24h after transfection cells were infected with lentiviral vectors containing fluorescent proteins, media was changed the day after to remove viral particles. 24h after cells were plated on MEF feeders at density $3x10^4$ in 10 cm dishes in fibroblast media (day 0). At day3 media was switched to mTeSR1 (StemCell Technology) for iPSC cultures and fed everyday for about 17-20 more days. At day 7 media was supplemented with VPA 0.5mM for VPA condition.

For lentiviral-derived iPSC: fibroblast were seeded at $1x10^5$ cells per well of a 6 well plate. The day after, cells were infected with lentiviruses encoding reprogramming factors (hSOX2, hKLF4, hOCT4, hMYC). After 24h cells were infected with lentiviruses encoding fluorescent proteins. The following day cells were plated on MEF feeders at density $3x10^4$ in 10 cm dishes in fibroblast media (day 0). At day 2 media was switched to mTeSR1. At day 3 doxycycline 1ug/ml was added to the culture until day 12, when concentration was reduced to 0.5ug/ml until day 22. At day 7 VPA 0.5mM was added for VPA condition. Colonies were picked around day 30-32.
All iPSC colonies were picked and expanded in mTeSR medium in matrigel coated dishes.

A.1.2 Calling SNVs/indels/CVs/MEIs

Samples were sequenced to 35x on an Illumina hiSeq. Reads were aligned to CRh37 and intial variants were called using the SpeedSeq pipeline. From the initial variants list, SNVs and indels were separated using VCFtools. We then removed any calls with multiple variant alleles at the same site. We split HDF and sister iPSC calls into two separate files, then removed any calls

with a read depth < 10 or > 250 in either the HDF or sister lines (meaning a variant in a sister line required a read depth of at least 11 at that position in both sisters and HDF datasets). We used bedtools to produce datasets of (1) mutations present in only one sister iPSC and no other sample and (2) mutations present in both sisters of a sister iPSC pair, but not present in any other sample. VAF for each call was made by extracting the read depth and alternate observations (variant reads) in VCFtools, then dividing the # variant reads by the total # reads.

```
//vcftools --vcf/file/location/name.vcf --extract-FORMAT-info AO
//vcftools --vcf/file/location/name.vcf --extract-FORMAT-info DP
```

We removed any calls which had a variant allele count > 0 in the HDF callset (a position can have a few variant calls and not be called a variant by our earlier pipeline). And removed any calls with fewer than 2 reads supporting reference or alternate calls.

A.1.3 False positive assessment for variant calls

We selected several lines from our episomal and lentiviral conditions that showed average sequencing depth and genome coverage, and sorted the SNV or indel calls from highest to lowest quality score. We designed PCRs to be evenly spaced along the spectrum of quality scores, giving us an even coverage for validation. We further considered the VAF of each of these, to make sure we weren't biasing our validations in any way. We designed PCRs to span 300bp upstream and downstream of the SNVs using NCBI primerblast to select unique primers. We performed PCRs for each site using the HDF DNA and the DNA of both sisters of the associated sister pair. We selected for sequencing PCRs that gave a clear product, using gel extraction if multiple products were detected. These samples were sent for Sanger sequencing with the forward primer, and the traces of both sisters and HDF were compared. Validated SNVs were those that showed a heterozygous mutant peak at the location of the SNV in the called sister but not the HDF or other sister. We did not validate somatic calls, as it is extremely unlikely that the same mutant called in two sisters but in no other sample is an artifact of sequencing. We subsequently performed validation on

To ensure that we weren't missing low VAF mutants due to the sensitivity of our Sanger method, we submitted a set of SNVs and indels to be validated by targeted deep sequencing. We pulled a group of SNVs from C4A and a group of indels from L92A and I92A as described above. Because many indel calls are in repetitive regions we performed nested PCR to obtain 100bp products spanning the mutant site. We submitted our samples to The Scripps Research Institute Genetics Core for targeted deep sequencing, who sequenced our samples to an average depth of several thousand reads per mutant. Files were aligned to GRCh19 with bowtie and processed with samtools, and variants were called using the mpileup function

```
$bowtie2 -x /ref/hg19 -U /sample/name.fastq -S DSeq_A.sam
$samtools view -S -b -h /location/DSeq_A.sam > OutA.bam
$samtools sort /locationt/OutA.bam A_Sort.bam
$samtools mpileup -uf /location/human_g1k_v37.fasta /location/A_Sort.bam | bcftools view -vcg - > A_SNV.vcf
```

To assess false positives in MEIs, we took a subset of our calls from two lines, A1A and I92A, and designed PCRs to span the insertion site. We then ran a 2% gel at 100v to resolve the wild-type product from the product with the insertion (300-800bp depending on the MEI). We validated this approach by selecting several germline MEIs called in our HDF dataset and overlapping with the 1000 genomes list of population MEIs. We performed PCR on these as described above and confirmed the presence of a wild type and mutant allele, showing that our PCR is sufficiently able to amplify through MEI regions.

For large SV deletions we designed PCRs to span the insertion site, with the assumption that only samples with an SV would be able to give product. For duplications we designed PCRs on spanning the duplication break point, which gave product only in the presence of the duplication (for tandem duplications). We validated all high confidence calls in all lines using this method.

A.1.4 Assessing false negative rate

Because we did not find evidence of reprogramming-associated MEIs, we used bedtools to intersect our MEI calls with the 1000 genomes MEI data set. We then pulled all overlapping MEIs (which are known to be real) and asked how many of these were called in all but one sample, which would mean a false negative rate of 1 in 25 alleles (one allele for each sample plus the bulk HDF). We repeated this process for samples called in all but two samples, and so on for all germline MEIs. We estimated false negative rate by taking the total number of missed alleles over the total number of possible alleles.

A.1.5 Assessing nucleotide context and signatures

Nucleotide context was done by hand for each line by taking all SNVs with QS > 100 and sorting them into reference A, C, T, or G, and then further sorting them by mutant base. These were totaled and the standard deviation was assessed in PRISM for significance by t test. Trinucleotide context was assessed using DeconstructSigs (70). To call mutational signatures we used the following argument:

```
Sample = whichSignatures(tumor.ref = sigs.input, signatures.ref = signatures.cosmic, contexts.needed = TRUE, tri.counts.method = 'genome', signature.cutoff = .01)
```

This normalized the trinucleotide counts to the number of times that trinucleotide appears in the reference genome, and it draws mutation signatures from the COSMIC database. Data were plotted in R. Statistical significance for nucleotide contexts was calculated by taking the contexts for each line individually and examining the standard deviation for episomal, lentiviral, and somatic conditions. We used a t test to compare each condition to each other for each context (lentivirus to episomal, lentivirus to somatic, somatic to episomal).

A.1.6 Assessing enrichment in genomic regions

We sourced the literature for databases of genomic regions of interest; a collection of these can be found at https://data.mendeley.com/datasets/ghrt3ngzrm/1, based on the work of Yoshihara et. al. (2017)(REF HERE). We ran all variants of QS > 100 against several of these databases using bedtools, and we noted the total number of overlapping nucleotides for each analysis. Relative enrichment was calculated as follows:

$$\frac{\text{(Overlapping SNVs / length of database regions)}}{\text{(All called SNVs / length of the reference genome)}}$$

Where overlapping SNVs is the # of SNVs from the sample that overlap the reference database, length of database is the total number of base pairs in the reference database, and all called SNVs is the total number of SNVs in the sample. The numerator denotes how often an SNV is called in a genomic feature, normalized to how common that feature is, while the denominator denotes how many mutations there were in total, normalized to the number of bases in the genome. We took the Log(2) of this value and plotted it as relative enrichment. Significance was calculated using bedtools fisher tool.

A.1.7. Assessing genomic impact

High confidence (QS > 100) SNVs and indels were run through Ensemble's variant effect predictor (VEP)(https://uswest.ensembl.org/info/docs/tools/vep/index.html). For SNVs, we selected any calls predicted by SIFT of PolyPhen to be deleterious. We also ran our calls through the Broad Institute's Oncotator tool (https://portals.broadinstitute.org/oncotator/help/) to determine status in COSMIC and the OMIM databases. Information on ClinVar was obtained by manually checking each SNV predicted to be deleterious with NCBI's variation viewer (https://www.ncbi.nlm.nih.gov/variation/view/) and checking for entries.

A.1.8. Developing a general approach for other datasets

We took all calls of QS > 100 and sorted them by VAF. We established bins every 0.02 for VAF 0.2 to 0.7 and queried how many of the calls in each bin were somatic vs reprogramming. We assessed what % of the total calls in the bin this represented, and noted that number as a "reprogramming coefficient" or a "somatic coefficient." We then removed somatic or reprogramming identifiers from our data and sorted it into VAF bins, multiplying the total in each bin by the reprogramming or somatic coefficient to get an estimate of the contribution of somatic or reprogramming mutations in each bin. We then summed the SNVs in each bin to estimate the total burden of somatic vs reprogramming-associated mutations.

A.2 Methods for Chapter 3

A.2.1 Preparation of primary rods from mouse retina

We prepared 40ml solutions of hibernate-A from gibco with 800ul of B27 without vitamin A and 100ul glutamax, (HAGB), these solutions were kept on ice. A solution of papain was prepared by dissolving one vial of Pap2 from Worthington in 20ml of Hibernate-A with 50ml glutamax. This solution was heat activated at 37℃ for 30 minutes then placed on ice. Mice were anesthetized with isoflurane and killed by decapitation. Retina were extracted from both eyes and treated with 2ml of the papain solution for 10 minutes in a 37℃ water bath, shaking gently to prevent the retina from sinking to the bottom. We replaced the papain with 2ml HAGB and triturated the retina 10 times using fire-polished glass pipettes, using a moderate speed. We then waited 1 min for the chunks to settle, and pipetted the supernatant through a 40um nylon cell strainer. We added 2ml HAGB to the tube with tissue chunks and triturated as before. In total we repeated this step three times, triturating for a total of 30x. After, we added 2ul of 2U/ul DNAse and incubate on ice for 5 mins, to remove any DNA in solution which makes micromanipulation more challenging. The solution was centrifuged at 200x G for 5 mins and resuspended in 50ul HAGB.

A.2.2 Performing SCNT and collecting 2c embryos or blastocysts

SCNT was performed as reported in Hazen and Faust et. al. (2016) (59). Briefly; Females were superovulated and oocytes collected and enucleated by micromanipulation with an 8u injection pipette. Nuclei from green fluorescent neurons were extracted and injected into enucleated oocytes. We activated the resultant embryos with strontium chloride and 5nM Trichostatin A. Trichostatin A was used both during 6h activation and additionally for 10h overnight. To derive ESCs, the zona pellucida of blastocysts were removed via a piezo-actuated drill needle before being transferred to a MEF feeder plate in ESC derivation medium containing: 500 mls Knockout DMEM (Gibco 10829-018), 80 mlsKnockout 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).

To collect blastocysts and single cells from 2c embryo, we used a 37℃ heated microscope stage. Blastocysts were collected in PBS droplets by mouth pipette after a 3x PBS wash from their incubation media and were immediately spun down and placed on dry ice. To dissociate 2c embryos, zona pellucida was removed by a brief wash in a drop of Tyrode's Solution, watching carefully to remove the embryo as soon as the zona was eliminated. Embryos were then washed 3x in M2 media before being dropped in 0.05% trypsin. Embryos were vigorously triturated by mouth pipette until the two cells became dissociated. Each cell was then washed 3x in PBS before being collected in a PCR tube, spun down, and immediately placed on dry ice. In some cases cells were collected quickly with TE. Single cells were amplified by SigmaPlex amplification kit. Blastocysts were amplified by GenomiPhi (GE) low input amplification

(MDA-based). After lysis, PBS was added to make a total volume of 4ul, and 2ul was carefully transferred to a second tube before both halves were amplified separately.

A.2.3 Assessing quality of amplified DNA

After amplification, samples were purified by phenol chloroform extraction with RNAse and were quantified by nanodrop. We assayed the evenness of amplification by designing PCRs for 22 genes across various chromosomes and 2 intergenic regions. The genes and associated primers were:

Sfi1	GAAAGCAGCACTGGCGATTC	CCGCTACGAGGACAGCTATG
Cenbp	ACCAAAGACCTGGTTGGGTG	GCCCTCGGACATAGCAACTT
Spry1	ACTGCACCAAGACCCGAAAA	GTCCACGATCCCACAGTACC
Bag1	AGAAGTCACCTGCTGGATGC	TCAGGAAGAGTGTTGCCGTC
Steap1	AACACAGCCCTAGTGAACCG	TGAGGTGACTTGATTGGGGC
Exoc4	CCAGAGGTGTCGTCGTGAAA	GGTTTCGCTTTGGAGGGGTA
Mesp2	GGGCCTTCACTAGCTGGAAA	TTTATCTGCCCCAGACACCG
Vac14	CCACTCAGCGTCACAGAAGT	AGCTCCCGGAATAAAACGCA
Casp1	AAACATGCGCACACAGCAAT	CTGGAGCTGAAGGTGAGTGG
Stx11	AAGGGAAGGAGTTCACGCAG	ACCAGGCCGAGATGAAACAG
Drg1	GCATCCAGCAAACTGCAGAC	CCGTAGTCTAAACCTGCGCT
Dnmt3a	AAACGTCTGCTGAGCATCCA	CCAGTGATGGGTGCAGTTCT
Smad5	TGAGGGGGATATGCTGGTGT	GAGCTAAGGAGGCATCGCAA
Parp2	GAGCCTCGGGAAGAATCAGG	ACACTCATGTTCTGTGGCGT
Lynx1	TTTGTGTCCCGAGCTCTCAC	GTCTGCTAGAGGTGAGGGGA
Mkl2	TGAGTTGCCTATGGGAAGCG	CCAGGGCTGGCAAAGAAGTA
Park2	CACGGTGAAAGTAGCCGAGT	AAAGCTTCCTCCGGATTGGG
Bambi	CCTGTGACAAAATGCGGCAA	TCGTTGCAGAGAAAGCGGAT
Men1	CACTCGCACTAAGGGTTGGT	TTGATGGCGCTCGAGTTGAT
GAS5	GGAGGTTGGTTCTGCGTGTA	GCCCTGACTTCAGACTTCCC
HotAir	TGCATAGACCTGCCTCCTCT	AAGGCTGAAATGGAGGACCG
Nespas	AGGAACGCGCATTTTGCTTT	CACCGTTGTCTCTGCTCAGT

We used 15ng DNA per reaction and ran PCRs for 30 cycles using Q5 HF polymerase at 20s extension and 60℃ annealing in a 30ul reaction volume. 20ul of the results were run on a 2% agarose gel, 150v. We assigned 0 points for an absent band, 1 point for a faint band, and 2 points for a clear, strong band.

QC of ESCs and single cells was conducted by the Hall lab and involved examining the estimated copy number along the genome, looking for samples which showed the roughly expected 2n copy number along the length of the genome.

A.2.4 Calling variants in single rod photoreceptors

Samples were sequenced to an average depth of 30x on an Illumina HiSeq. Sequences were aligned to GRCm38 and variants were initially called with GATK. We removed all calls with more than one variant allele, as well as any calls with <10 or >250 read depth. We compared both halves of each blast and kept only mutations called in both. We then screened this dataset against the bulk control DNA and removed all calls found in bulk DNA. We removed all calls with VAF < 0.3 and with variant or reference read support <2. Based on our work in Chapter 2, we ruled out all mutations with a quality score less than 30, and assessed false positive rate as described previously.

A.2.5 Nucleotide context and transvesion/transition ratio

Nucleotide context was assessed as describe in A.1.5 with the following changes. Because DeconstructSigs was not designed to work with mice, we downloaded the source from github (https://github.com/raerose01/deconstructSigs) and found a bsg reference which we changed from NULL to BSgenome.Mmusculus.UCSC.mm10 (this is not standard install from Bioconductor but can be installed as described here: https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Mmusculus.UCSC.mm10.html). We confirmed that this approach worked by creating a test set of 20 trinucleotide contexts that we confirmed by hand with the UCSC genome browser. The specific argument made to generate the final data was as follows:

```
sigs.input <- mut.to.sigs.input(bsg = mouseGenome, mut.ref =
"~/R/InputFiles/QS100_Somatic.txt", sample.id = "Sample", chr = "chr", pos = "pos", ref
= "ref", alt = "alt"
```

To assess each individual set of trinucleotides as shown in fig. 3.6, we imported the raw data from DeconstructSigs into PRISM. Importantly, to generate a full trinucleotide plot of all data we removed sample ID and set the tag to simple “p6”, “p21,” etc… However to generate error bars in PRISM it is necessary to run the analysis for each individual sample, maintaining the sample ID and then gathering each individual output file into a single tab delimited txt document, which we did in excel.

The Tv/Ti ratio plots seen in fig. 3.5 was done with GenVisR after generating a tab delimited txt file with 3 columns; sample, reference, and variant, referring to the sample id, reference nucleotide, and variant nucleotide. The samples can be plotted by reading the txt file into R and using the TvTi function in the GenVisR package (we used supplementary arguments lab_txtAngle = 75, fileType = "MGI")

A.2.6 Functional assessment of Rods

We downloaded an RNAseq dataset for mouse rods (GSE74660) and sorted by normalized fpkm, taking the top 50% expressed genes as described in Hazen and Faust et. al. (2016). We took the genomic loci of these genes and crossed our SNVs against them with Bedtools. We processed the resulting hits with ensemble's variant effect predictor (VEP) tool, which can be cloned from Github here https://github.com/Ensembl/ensembl-vep and recorded all hits of interest. The specific arguments used were

> ./vep --appris --biotype --check_existing --distance 1 --domains --pick --plugin miRNA --regulatory --sift b --species mus_musculus --symbol --tsl --cache --input_file [input_data]

We assessed SNVs in the same gene across multiple samples with GenVisR's waterfall function (https://bioconductor.org/packages/release/bioc/html/GenVisR.html). Important: this requires the VEP output, but as is the VEP mutation type output labels do not match the expected arguments for the waterfall function. These must be changed by hand (trv_type) to match the annotation outlined in the documentation for GenVisR (so, missense_mutation must be changed to missense, for example).

Works Cited

1. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res. 2011;21(6):961-73.

2. Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. 2013;23(5):749-61.

3. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16(3):172-83.

4. Mills RE, Bennett EA, Iskow RC, Devine SE. Which transposable elements are active in the human genome? Trends Genet. 2007;23(4):183-91.

5. Rodic N, Burns KH. Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? PLoS Genet. 2013;9(3):e1003402.

6. International Human Genome Sequencing C, Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC. Initial sequencing and analysis of the human genome. Nature. 2001;409:860.

7. Chen C, Gentles AJ, Jurka J, Karlin S. Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. Proceedings of the National Academy of Sciences. 2002;99(5):2930.

8. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3(5):370-9.

9. Hancks DC, Kazazian HH, Jr. SVA retrotransposons: Evolution and genetic instability. Semin Cancer Biol. 2010;20(4):234-45.

10. Eric M Ostertag JLG, Yue Zhang, Haig H. Kazazian Jr. SVA Elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet. 2008;73:1444-51.

11. Tönjes RR1 LR, Boller K, Denner J, Hasenmaier B, Kirsch H, König H, Korbmacher C, Limbach C, Lugert R, Phelps RC, Scherer J, Thelen K, Löwer J, Kurth R. HERV-K: the biologically most active human endogenous retrovirus family. J Acquir Immune Defic Syndr Hum Retrovirol. 1996;13:261-7.

12. Geoffrey Turner* MB, Mei Su*,, Michael I. Jensen-Seaman†‡§ KKKaJL. Insertional polymorphisms of full-length endogenous retroviruses
in humans. Curr Biol. 2001;11.

13. Kino K, Sugiyama H. Possible cause of G·C→C·G transversion mutation by guanine oxidation product, imidazolone. Chem & Biol. 2001;8(4):369-78.

14. Nabel CS, Manning SA, Kohli RM. The Curious Chemical Biology of Cytosine: Deamination, Methylation,and Oxidation as Modulators of Genomic Potential. ACS Chemical Biology. 2012;7(1):20-30.

15. Bransteitter R, Pham P, Scharff MD, Goodman MF. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proceedings of the National Academy of Sciences. 2003;100(7):4102.

16. Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980;287:560.

17. D.L. M, D. K. The Induction and Repair of DNA Photodamage in the Environment. In: A.R. Y, J. M, L.O. B, W. N, editors. Environmental UV Photobiology. Boston: Springer; 1993.

18. Sinha RP, Häder D-P. UV-induced DNA damage and repair: a review. Photochemical & Photobiological Sciences. 2002;1(4):225-36.

19. De Bont R, van Larebeke N. Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis. 2004;19(3):169-85.

20. Margison GP, Santibanez-Koref MF. O6-alkylguanine-DNA alkyltransferase: role in carcinogenesis and chemotherapy. Bioessays. 2002;24(3):255-66.

21. Kondo N, Takahashi A, Ono K, Ohnishi T. DNA damage induced by alkylating agents and repair pathways. J Nucleic Acids. 2010;2010:543531.

22. Illumina. Illumina Sequencing Introduction 2016 [Available from: https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf.

23. Potapov V, Ong JL. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLoS One. 2017;12(1):e0169774.

24. Borgstrom E, Paterlini M, Mold JE, Frisen J, Lundeberg J. Comparison of whole genome amplification techniques for human single cell exome sequencing. PLoS One. 2017;12(2):e0171566.

25. Geurts-Giele WR, Dirkx-van der Velden AW, Bartalits NM, Verhoog LC, Hanselaar WE, Dinjens WN. Molecular diagnostics of a single multifocal non-small cell lung cancer case using targeted next generation sequencing. Virchows Arch. 2013;462(2):249-54.

26. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875.

27. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellstrom-Lindberg E, Jansen JH, Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci Rep. 2017;7:43169.

28. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8(12):e85024.

29. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007;131(5):861-72.

30. Malik N, Rao MS. A review of the methods for human iPSC derivation. Methods Mol Biol. 2013;997:23-33.

31. Gore A, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Bourget J, Daley GQ, Goldstein LSB, Zhang K. Somatic coding mutations in human induced pluripotent stem cells. Nature. 2011;471(7336):63-7.

32. Ji J, Ng SH, Sharma V, Neculai D, Hussein S, Sam M, Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. Stem Cells. 2012;30(3):435-40.

33. Cheng L, Hansen NF, Zhao L, Du Y, Zou C, Donovan FX, Low incidence of DNA sequence variation in human induced pluripotent stem cells generated by nonintegrating plasmid expression. Cell Stem Cell. 2012;10(3):337-44.

34. Li Z, Lu H, Yang W, Yong J, Zhang ZN, Zhang K, Mouse SCNT ESCs have lower somatic mutation load than syngeneic iPSCs. Stem Cell Reports. 2014;2(4):399-405.

35. Bhutani K, Nazor KL, Williams R, Tran H, Dai H, Dzakula Z, Whole-genome mutational burden analysis of three pluripotency induction methods. Nat Commun. 2016;7:10536.

36. Lo Sardo V, Ferguson W, Erikson GA, Topol EJ, Baldwin KK, Torkamani A. Influence of donor age on induced pluripotent stem cells. Nat Biotechnol. 2017;35(1):69-74.

37. Kwon EM, Connelly JP, Hansen NF, Donovan FX, Winkler T, Davis BW, iPSCs and fibroblast subclones from the same fibroblast population contain comparable levels of sequence variations. Proc Natl Acad Sci U S A. 2017;114(8):1964-9.

38. Wissing S, Munoz-Lopez M, Macia A, Yang Z, Montano M, Collins W, Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. Hum Mol Genet. 2012;21(1):208-18.

39. Klawitter S, Fuchs NV, Upton KR, Munoz-Lopez M, Shukla R, Wang J, Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. Nat Commun. 2016;7:10286.

40. Hussein SM, Batada NN, Vuoristo S, Ching RW, Autio R, Narva E, Copy number variation and selection during reprogramming to pluripotency. Nature. 2011;471(7336):58-62.

41. Laurent LC, Ulitsky I, Slavin I, Tran H, Schork A, Morey R, Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. Cell Stem Cell. 2011;8(1):106-18.

42. Quinlan Aaron R, Boland Michael J, Leibowitz Mitchell L, Shumilina S, Pehrson Sidney M, Baldwin Kristin K, Genome Sequencing of Mouse Induced Pluripotent Stem Cells Reveals Retroelement Stability and Infrequent DNA Rearrangement during Reprogramming. Cell Stem Cell. 2011;9(4):366-73.

43. Liu DX, Greene LA. Neuronal apoptosis at the G1/S cell cycle checkpoint. Cell Tissue Res. 2001;305:217-28.

44. Becker EB, Bonni A. Cell cycle regulation of neuronal apoptosis in development and disease. Prog Neurobiol. 2004;72(1):1-25.

45. Barnes DE, Stamp G, Rosewell I, Denzel A, Lindahl T. Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. Current Biology. 1998;8(25):1395-8.

46. Deans B, Griffin CS, Maconochie M, Thacker J. Xrcc2is required for genetic stability, embryonic neurogenesis and viability in mice. The EMBO Journal. 2000;19(24):6675.

47. Jeppesen DK, Bohr VA, Stevnsner T. DNA repair deficiency in neurodegeneration. Prog Neurobiol. 2011;94(2):166-200.

48. Suberbielle E, Sanchez PE, Kravitz AV, Wang X, Ho K, Eilertson K, Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-beta. Nat Neurosci. 2013;16(5):613-21.

49. Madabhushi R, Gao F, Pfenning AR, Pan L, Yamakawa S, Seo J, Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. Cell. 2015;161(7):1592-605.

50. Rehen SK, Yung YC, McCreight MP, Kaushal D, Yang AH, Almeida BS, Constitutional aneuploidy in the normal human brain. J Neurosci. 2005;25(9):2176-80.

51. Kingsbury MA, Friedman B, McConnell MJ, Rehen SK, Yang AH, Kaushal D,. Aneuploid neurons are functionally active and integrated into brain circuitry. Proceedings of the National Academy of Sciences. 2005;102(17):6143.

52. Iourov IY, Vorsanova SG, Liehr T, Yurov YB. Aneuploidy in the normal, Alzheimer's disease and ataxia-telangiectasia brain: differential expression and pathological meaning. Neurobiol Dis. 2009;34(2):212-20.

53. McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Mosaic copy number variation in human neurons. Science. 2013;342(6158):632-7.

54. Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K, L1 retrotransposition in neurons is modulated by MeCP2. Nature. 2010;468(7322):443-6.

55. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, L1 retrotransposition in human neural progenitor cells. Nature. 2009;460(7259):1127-31.

56. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sanchez-Luque FJ, Bodea GO, Ubiquitous L1 mosaicism in hippocampal neurons. Cell. 2015;161(2):228-39.

57. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005;435(7044):903-10.

58. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell. 2012;151(3):483-96.

59. Hazen JL, Faust GG, Rodriguez AR, Ferguson WC, Shumilina S, Clark RA, The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning. Neuron. 2016;89(6):1223-36.

60. Evrony GD, Lee E, Park PJ, Walsh CA. Resolving rates of mutation in the brain using single-neuron genomics. Elife. 2016;5.

61. Erwin JA, Paquola AC, Singer T, Gallina I, Novotny M, Quayle C, L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat Neurosci. 2016;19(12):1583-91.

62. Bedrosian TA, Quayle C, Novaresi N, Gage FH. Early life experience drives structural variation of neural genomes in mice. Science. 2018;359(6382):1395.

63. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Somatic mutation in single human neurons tracks developmental and transcriptional history. Science. 2015;350(6256):94.

64. Bohrson CL, Barton AR, Lodato MA, Rodin RE, Viswanadham V, Gulhan D,. Linked-read analysis identifies mutations in single-cell DNA sequencing data. bioRxiv. 2017.

65. Lodato MA, Rodin RE, Bohrson CL, Coulter ME, Barton AR, Kwon M, Aging and neurodegeneration are associated with increased mutations in single human neurons. Science. 2018;359(6375):555.

66. Garreta E, Sanchez S, Lajara J, Montserrat N, Belmonte JCI. Roadblocks in the Path of iPSC to the Clinic. Curr Transplant Rep. 2018;5(1):14-8.

67. Mandai M, Watanabe A, Kurimoto Y, Hirami Y, Morinaga C, Daimon T, Autologous Induced Stem-Cell-Derived Retinal Cells for Macular Degeneration. N Engl J Med. 2017;376(11):1038-46.

68. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV,. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21.

69. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014;24:52-60.

70. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology. 2016;17(1):31.

71. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003;31(13):3812-4.

72. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.

73. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17(1):122.

74. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research. 2005;33(Database issue):D514-D7.

75. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014;42(Database issue):D980-D5.

76. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biology. 2014;15(6):R84.

77. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome research. 2017;27(11):1916-29.

78. Ebert AD, Liang P, Wu JC. Induced pluripotent stem cells as a disease modeling and drug screening platform. J Cardiovasc Pharmacol. 2012;60(4):408-16.

79. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. Nat Commun. 2017;8:15183.

80. Su R-J, Yang Y, Neises A, Payne KJ, Wang J, Viswanathan K, Few Single Nucleotide Variations in Exomes of Human Cord Blood Induced Pluripotent Stem Cells. PLOS ONE. 2013;8(4):e59908.

81. Grollman AP, Moriya M. Mutagenesis by 8-oxoguanine: an enemy within. Trends in Genetics. 1993;9(7):246-9.

82. Panopoulos AD, Yanes O, Ruiz S, Kida YS, Diep D, Tautenhahn R, The metabolome of induced pluripotent stem cells reveals metabolic changes occurring in somatic cell reprogramming. Cell Res. 2012;22(1):168-77.

83. Zhang J, Khvorostov I, Hong JS, Oktay Y, Vergnes L, Nuebel E, UCP2 regulates energy metabolism and differentiation potential of human pluripotent stem cells. EMBO J. 2011;30(24):4860-73.

84. Dang CV, O'Donnell KA, Zeller KI, Nguyen T, Osthus RC, Li F. The c-Myc target gene network. Semin Cancer Biol. 2006;16(4):253-64.

85. Miller DM, Thomas SD, Islam A, Muench D, Sedoris K. c-Myc and cancer metabolism. Clin Cancer Res. 2012;18(20):5546-53.

86. Kim JW, Yang HJ, Brooks MJ, Zelinger L, Karakulah G, Gotoh N, NRL-Regulated Transcriptome Dynamics of Developing Rod Photoreceptors. Cell Rep. 2016;17(9):2460-73.

87. Chapman JR, Taylor MR, Boulton SJ. Playing the end game: DNA double-strand break repair pathway choice. Mol Cell. 2012;47(4):497-510.

88. Lieber MR. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu Rev Biochem. 2010;79:181-211.

89. Magistretti PJ, Allaman I. A cellular perspective on brain energy metabolism and functional imaging. Neuron. 2015;86(4):883-901.

90. Akimoto M, Cheng H, Zhu D, Brzezinski JA, Khanna R, Filippova E, Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(10):3890.

91. Long CR, Westhusin ME, Golding MC. Reshaping the transcriptional frontier: epigenetics and somatic cell nuclear transfer. Mol Reprod Dev. 2014;81(2):183-93.

92. Akbarian S, Beeri MS, Haroutunian V. Epigenetic determinants of healthy and diseased brain aging and cognition. JAMA Neurol. 2013;70(6):711-8.

93. Matoba S, Liu Y, Lu F, Iwabuchi KA, Shen L, Inoue A, Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. Cell. 2014;159(4):884-95.

94. Kim J-W, Yang H-J, Brooks MJ, Zelinger L, Karakülah G, Gotoh N, NRL-Regulated Transcriptome Dynamics of Developing Rod Photoreceptors. Cell reports. 2016;17(9):2460-73.

95. Solovei I, Kreysing M, Lanctôt C, Kösem S, Peichl L, Cremer T, Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. Cell. 2009;137(2):356-68.

96. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. Cell. 2012;148(5):873-85.

97. De Bont R, Van Larebeke N. Endogenous DNA damage in humans: A review of quantitative data2004. 169-85 p.

98. Schrider DR, Hourmozdi JN, Hahn MW. Pervasive multinucleotide mutational events in eukaryotes. Curr Biol. 2011;21(12):1051-4.

99. Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Multi-nucleotide de novo Mutations in Humans. PLoS Genet. 2016;12(11):e1006315.

100. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science. 2018;359(6375):550.

101. Lehmann K, Schmidt K-F, Löwel S. Vision and visual plasticity in ageing mice. Restorative Neurology and Neuroscience. 2012;30(2):161-78.

102. Kolesnikov AV, Fan J, Crouch RK, Kefalov VJ. Age-related deterioration of rod vision in mice. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2010;30(33):11222-31.