Prevalence and Prevention of Large-Scale Somatic Copy Number Alterations

by

Kristin Ann Knouse
B.S. Biology
Duke University, 2010

Submitted to the Department of Biology
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

February 2017

© 2017 Kristin Ann Knouse. All rights reserved.

The author hereby grants to MIT permission to reproduce
and to distribute publicly paper and electronic
copies of this thesis document in whole or in part in any medium now known or
hereafter created.

Signature of Author: _

Signature redacted

Department of Biology
September 30, 2016

Certified by: ___

Signature redacted

Angelika Amon
Professor of Biology
Thesis Supervisor

Accepted by: ___

Signature redacted

Michael Hemann
Associate Professor of Biology
Co-Chair, Committee for Graduate Students
Prevalence and Prevention of Large-Scale Somatic Copy Number Alterations

by

Kristin Ann Knouse

Submitted to the Department of Biology on September 30, 2016 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the Massachusetts Institute of Technology

Organismal viability is contingent upon the transmission of a balanced genome to the next generation. Genomic imbalance on the order of megabases, as occurs with sub-chromosome copy number variants (CNVs) or whole-chromosome aneuploidy, is almost always associated with embryonic lethality or severe disease. However, while large-scale genomic imbalance has adverse consequences at the organismal level, both CNVs and aneuploidy are widespread in cancer and also reported to be prevalent in select untransformed tissues. In order to reconcile these discrepant observations and understand how genomic imbalance influences disease, it is necessary to perform directed analyses of genomic stability in vivo. In this thesis, we first employ single cell sequencing to examine megabase-scale somatic copy number alterations in mammalian tissues. We find that the somatic copy alterations are indeed more prevalent than constitutional copy number alterations, but not nearly as prevalent as previously reported for some tissues. We then analyze chromosome segregation of somatic cells both in vivo and in vitro and discover that chromosome segregation in epithelia is not cell autonomous. We find that tissue architecture facilitates chromosome segregation in vivo and its disruption leads to chromosome instability. Finally, we turn to the increasingly available cancer genomics data to assess karyotypic changes both within and across cancer types and within individual tumors. We find strong evidence for karyotype optimization that is highly cell-type specific. Together, our results emphasize the importance of context, whether it be organism-wide or somatic, in vivo or in vitro, or a specific cell type, in the generation and consequences of genomic imbalance and highlight the importance of in vivo experimentation in which context is considered.

Thesis supervisor: Angelika Amon
Title: Professor of Biology
ACKNOWLEDGEMENTS

Conducting my graduate work under Angelika’s supervision has been a true privilege and a life-changing experience. Angelika has the enthusiasm of someone who just picked up a pipette for the first time but the wisdom of someone who has been doing science for years—a rare but wonderful combination under which to train. I appreciate her granting me the freedom to exercise my scientific free spirit all while remaining readily available for guidance and direction. Her support has been extensive and steadfast, regardless of the state of my projects. She has been instrumental in my growth as a scientist and individual over these past several years and is a major source of inspiration for my future career.

I have also benefited from several other wonderful mentors throughout my scientific training. I am grateful to David Sabatini and Tyler Jacks for serving on my thesis committee. I appreciate them always providing interesting perspectives and ideas for my projects as well as their sincere advice and support for my future career. I thank Iain Cheeseman for his constant willingness to share his mitotic expertise and his dedication to assembling the local mitosis community. I thank Frank Solomon for wonderful conversations about all aspects of science. I appreciate David Pellman for serving as the outside member of my defense committee and for providing both scientific and career advice from the start of my time in Boston. I am also grateful to my undergraduate mentors, Bryan Tsou and Terry Lechler, for their patience and encouragement as they let an inexperienced undergraduate into their labs to fall in love with science.

The members of the Amon lab have provided a wonderful scientific environment. I am especially grateful for the long-term co-residents of the lab’s "Back Bay". Stacie, thank you for being my partner in crime and my sounding board for all things science and life. Stefano (Steve), I could always count on your company and advice in lab any hour of the day, any day of the week. Thank you for teaching me everything I know about the spindle assembly checkpoint and introducing me to Italian “culture”. Ian, thank you for enhancing any moment with insight or entertainment, whichever was most warranted at the time.

I have had the privilege of working with an incredibly talented undergraduate, Kristina (Tina) Lopez. She has amazed me with the pace at which she masters new techniques and the attention to detail invested into every single experiment. I would not have been able to make the progress that I have were it not for her hard work and creativity. Tina, it has been so fun to work alongside you these past two years and so rewarding to see you develop into an independent scientist. I look forward to seeing what you accomplish in your own scientific career.
I have benefited from work and collaborations with a number of individuals from different labs and institutions. Thank you to Jie Wu and Charlie Whittaker of the Koch Institute Bioinformatics Facility, Wendy Salmon of the Whitehead’s Keck Microscopy Facility, Matthew Frosch, Daniel Mordes, and Inma Cobos Sillero of the Neuropathology Department at Massachusetts General Hospital, Amitabh Srivastava and Melanie Johncilla of the Pathology Department at Brigham and Women’s Hospital, Kelly Stevens from Sangeeta Bhatia’s lab, and Hao Yin and Roman Bogorad from Dan Anderson’s lab. Your ideas, insight, and assistance have been major assets to various projects and our interactions have been a continuous source of exciting ideas.

Apart from the science, I will always remember graduate school for leading me to such a wonderful group of friends and their endless source of fun, inspiration, and support. Belinda, Alba, Sasha, Aurelia, Lisa, and Marianna, I am grateful there was never a moment without a trip, concert, or themed party to look forward to. Shari, our walks and talks have theoretically made us wiser than when we started here six years ago, thank you for always listening and getting it. Sahin, thank you for your endless advice and your commitment to keeping me cultured. Xenos, Daniel, and Alex, our long runs along the Charles were a great source of perspective and the perfect start to any day.

I am so grateful for the lifelong support of my family. To my siblings, Bryan and Melissa, you have both been a continuous source of bonding and I am so proud of the individuals you have become. To my parents, thank you for always being there—whether it was (literally) picking me up from marathon finish lines or bringing me takeout when you were visiting and I was inevitably still in lab. Most of all, thank you for raising me with (mostly) one requirement—to find what it is that I loved such that I’d never work a day in my life. It is because of you that these past four years have been the best years of my life.

And to Nick and Ken at Gordon’s, your Whiskey Wednesdays never let any week become too stressful.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................... 4

CHAPTER ONE: INTRODUCTION .......................................................................................... 8
INTRODUCTION .................................................................................................................... 9
CHROMOSOME SEGREGATION .......................................................................................... 10
History................................................................................................................................ 10
Mechanism.......................................................................................................................... 11
Regulation.............................................................................................................................. 14
Meroely................................................................................................................................. 17
CHROMOSOME MISSEGREGATION .................................................................................... 18
Causes................................................................................................................................ 18
Consequences....................................................................................................................... 20
ANEUPOLOIDY .................................................................................................................... 21
The aneuploid state .............................................................................................................. 21
Aneuploidy in healthy organisms ....................................................................................... 22
Aneuploidy in cancer ........................................................................................................... 25
COPY NUMBER VARIANTS .................................................................................................... 27
SUMMARY ............................................................................................................................. 28
REFERENCES ......................................................................................................................... 29

CHAPTER TWO: ASSESSMENT OF MEGABASE-SCALE SOMATIC COPY NUMBER VARIATION USING SINGLE CELL SEQUENCING ............................................................ 38
ABSTRACT ............................................................................................................................. 39
INTRODUCTION .................................................................................................................... 40
RESULTS ............................................................................................................................... 42
Characterizing sequencing data from single somatic cells ............................................... 42
Optimizing sensitivity and specificity of CNV detection in single cell sequencing data .... 45
Prevalence of somatic CNVs............................................................................................... 51
Validation of somatic CNVs ............................................................................................... 57
Characteristics of somatic CNVs ....................................................................................... 63
DISCUSSION ......................................................................................................................... 67
METHODS .............................................................................................................................. 72
ACKNOWLEDGMENTS ......................................................................................................... 76
REFERENCES ......................................................................................................................... 77

CHAPTER THREE: SINGLE CELL SEQUENCING REVEALS LOW LEVELS OF ANEUPOLOIDY ACROSS MAMMALIAN TISSUES ................................................................. 80
ABSTRACT ............................................................................................................................. 81
INTRODUCTION .................................................................................................................... 82
RESULTS ............................................................................................................................... 85
Detecting somatic aneuploidy by single cell sequencing .................................................... 85
Prevalence of aneuploidy in skin ......................................................................................... 90
Prevalence of aneuploidy in brain ...................................................................................... 91
Prevalence of aneuploidy in liver ....................................................................................... 96
DISCUSSION ......................................................................................................................... 100
The implications of low aneuploidy for hepatocyte proliferation ..................................... 101
The implications of low levels of aneuploidy for neural diversity .................................... 102
### CHAPTER FOUR: CELL NON-AUTONOMOUS REGULATION OF CHROMOSOME SEGREGATION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>113</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>114</td>
</tr>
<tr>
<td>RESULTS</td>
<td>115</td>
</tr>
<tr>
<td>Tissue architecture facilitates chromosome stability in epithelia</td>
<td>116</td>
</tr>
<tr>
<td>Tissue architecture enhances correction of aberrant microtubule-kinetochore attachments</td>
<td>120</td>
</tr>
<tr>
<td>Tissue architecture ensures chromosome stability in the setting of supernumerary centrosomes</td>
<td>123</td>
</tr>
<tr>
<td>Disruption of tissue architecture could contribute to chromosome instability in cancer</td>
<td>125</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>126</td>
</tr>
<tr>
<td>METHODS</td>
<td>127</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>128</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>137</td>
</tr>
</tbody>
</table>

### CHAPTER FIVE: ANEUPLOIDY IN CANCER: SEQ-ING ANSWERS TO OLD QUESTIONS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>140</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>141</td>
</tr>
<tr>
<td>ORIGIN OF ANEUPLOIDY IN CANCER</td>
<td>142</td>
</tr>
<tr>
<td>STATE OF ANEUPLOIDY IN CANCER</td>
<td>143</td>
</tr>
<tr>
<td>Pan-Cancer Patterns of Aneuploidy</td>
<td>147</td>
</tr>
<tr>
<td>Cancer-Specific Patterns of Aneuploidy</td>
<td>148</td>
</tr>
<tr>
<td>EVOLUTION OF ANEUPLOIDY IN CANCER</td>
<td>149</td>
</tr>
<tr>
<td>Aneuploidy in Time</td>
<td>156</td>
</tr>
<tr>
<td>Aneuploidy in Space</td>
<td>159</td>
</tr>
<tr>
<td>ROLE OF ANEUPLOIDY IN CANCER</td>
<td>162</td>
</tr>
<tr>
<td>Aneuploidy in Untransformed Cells and Organisms</td>
<td>162</td>
</tr>
<tr>
<td>Aneuploidy as Cancer's Arm</td>
<td>165</td>
</tr>
<tr>
<td>Aneuploidy as Cancer's Achilles Heel</td>
<td>169</td>
</tr>
<tr>
<td>CONCLUSIONS</td>
<td>171</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>172</td>
</tr>
</tbody>
</table>

### CHAPTER SIX: DISCUSSION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DETECTING SOMATIC COPY NUMBER ALTERATIONS USING SINGLE CELL SEQUENCING</td>
<td>178</td>
</tr>
<tr>
<td>PREVALENCE OF SOMATIC COPY NUMBER ALTERATIONS</td>
<td>179</td>
</tr>
<tr>
<td>ROLE OF TISSUE ARCHITECTURE IN CHROMOSOME STABILITY</td>
<td>181</td>
</tr>
<tr>
<td>ANEUPLOIDY IN CANCER</td>
<td>187</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>190</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>192</td>
</tr>
</tbody>
</table>

REFERENCES: 103, 109, 110

ACKNOWLEDGMENTS: 103

ANOLOGIES: 103

SEQUENCING: 109

REFERENCES: 110
CHAPTER ONE

Introduction
INTRODUCTION

The ultimate goal of life is to transmit an exact replica of the genome to another cell. This is achieved through the sequential processes of DNA replication and chromosome segregation, along with the continuous surveillance and repair of DNA damage. For sexually reproducing, multicellular organisms, transmission of DNA occurs during both the generation of diploid somatic cells and the production of haploid gametes. Diploid somatic cells are generated through mitosis, with DNA replication preceding the segregation of duplicated sister chromatids. Haploid gametes are generated through meiosis, with a single round of DNA replication preceding two independent chromosome segregation events. The first meiotic division segregates homologous chromosomes while the second segregates duplicated sister chromatids. Errors in the processes of DNA replication, repair, or chromosome segregation lead to mutations or structural alterations in the genome. Typically, these changes compromise cellular fitness. In some cases, however, these alterations can enhance cellular fitness and, if arising in a somatic cell, potentially lead to cancer. Defects in chromosome segregation are especially significant given that any single chromosome missegregation event will alter a substantial proportion of the genome. Over the past century, studies using model organisms and cell culture have identified several of the factors and mechanisms involved in executing and safeguarding chromosome segregation. While most of these discoveries are presumably generalizable across contexts, it is still unknown exactly how and to what extent these processes occur faithfully in mammals in vivo. In order to fully
understand how chromosome segregation influences both health and disease in mammals, it is critical to understand the prevalence and prevention of chromosome instability in vivo.

CHROMOSOME SEGREGATION

History
The idea that cells are generated by the partitioning of chromosomes into two daughter cells was first promulgated by Walther Flemming in 1882 (Flemming 1882). By observing developing animal embryos and regenerating adult tissues, he observed nuclei transforming into threads (mitos in ancient Greek) and the threads then partitioning into daughter cells. For nearly a century, investigation of chromosome segregation remained largely descriptive, with various imaging approaches revealing macromolecular structures such as the mitotic spindle and kinetochore which were assumed to facilitate this process (Inoue and Dan 1951; Metzner 1894). It wasn’t until nearly a century later that genetic screens in both budding and fission yeast began to uncover the specific players required for driving and regulating the cell cycle and chromosome segregation (Hartwell et al. 1974; Nurse 1975; Hoyt et al. 1991; Li and Murray 1991). Since then, further studies in yeast as well as other model organisms such as Drosophila and Xenopus and particularly animal cell culture have helped reveal the specific interactions among these players with respect to accurate chromosome segregation (Sunkel and Glover 1988; Glover et al. 1995; Lohka et al. 1988; Chen et al. 1996; Liu et al. 2009). Critical events in this process are the
capture of chromosomes by the mitotic spindle and the surveillance of such attachments by the spindle assembly checkpoint.

**Mechanism**

The forces required to move chromosomes into daughter cells are provided by the dynamic instability of the microtubules that comprise the mitotic spindle along with the motor proteins that associate with microtubules. Chromosome attachment to microtubules is mediated by the kinetochore, a multiprotein complex that assembles at the centromere of each chromosome (Cheeseman 2014). Proteins that comprise the inner kinetochore associate with the centromeric histone H3 variant CENP-A throughout the cell cycle (Figure 1A, left panel). At the onset of mitosis, additional kinetochore proteins assemble on the inner kinetochore to form the outer kinetochore (Figure 1A, right panel). A key component of the outer kinetochore is the Ndc80 complex (Figure 1B). This heterotetrameric complex of Ndc80 (Hec1 in humans), Nuf2, Spc24, and Spc25 forms a rod-like structure with globular head domains at either end of an α-helical coiled coil. The N-termini of Ndc80 and Nuf2 comprise the globular head domain of one end of this complex. Their N-terminal calponin-homology domains and nearby unstructured positively charged regions form electrostatic interactions with the negatively-charged C-terminal tails of tubulin and are thus responsible for microtubule binding by the kinetochore (Wei et al. 2007; Ciferri et al. 2008).

In mammals, but not all organisms, multiple kinetochore complexes are established in a tandem array along the centromere and thus a given
chromosome can interact with several (15-20) microtubules (Zinkowski et al. 1991; Cheeseman 2014).

**Figure 1**

A) Depiction of the inner kinetochore present throughout the cell cycle (left panel) and the outer kinetochore assembled during mitosis (right panel). Adapted from Cheeseman 2014. B) Depiction of the Ndc80 complex. Adapted from Tanaka 2010.

Upon nuclear envelope breakdown, the microtubule plus-ends emanating from the separating centrosomes begin to “search and capture” kinetochores...
The initial interaction between a microtubule and kinetochore typically involves the kinetochore making lateral attachments with the microtubule (Figure 2A, top panel). Eventually, through the combined action of motor proteins and microtubule depolymerization, these lateral interactions are converted into end-on attachments, with the kinetochore directly associating with the plus-end of the microtubule (Figure 2A, bottom panel). With time, the chromosomes, each a set of paired sister chromatids, congress to form a metaphase plate in preparation for anaphase onset and segregation of sisters into separate daughter cells.

In order for sister chromatids to segregate into separate daughter cells, the sister kinetochores must attach to microtubules emanating from opposite spindle poles, a condition known as amphitelic attachment (Figure 2B) (Tanaka 2010). However, amphitelic attachment is not the only possible outcome of microtubule search and capture. Sister kinetochores can also form a syntelic attachment, with both kinetochores attached to microtubules emanating from the same spindle pole, or a monotelic attachment, in which only one kinetochore is attached to microtubules (Figure 2B). Additionally, because a given kinetochore binds multiple microtubules, a kinetochore can also form a merotelic attachment, in which the kinetochore is attached to microtubules emanating from both spindle poles (Figure 2B). If syntelic, monotelic, or merotelic attachments persist into anaphase, the sister chromatids are unlikely to segregate properly into two daughter cells. To ensure that anaphase only ensues when all sister chromatids
have achieved the proper amphitelic attachment, the cell relies on the coupling of error correction and the spindle assembly checkpoint.

Figure 2

Figure 2. Kinetochore-microtubule attachment. A) Depiction of the initial lateral attachment of a kinetochore to a microtubule (top panel) and the ultimate end-on attachment of a kinetochore to a microtubule (bottom panel). B) Depiction of monotelic, syntelic, amphitelic, and merotelic kinetochore-microtubule attachments. Entire figure adapted from Tanaka 2010.

Regulation

Error correction, the elimination of monotelic, syntelic, and merotelic attachments in favor of amphitelic attachments, occurs through the stabilization of amphitelic attachments and destabilization of all other attachments. The kinase Aurora B is central to this process. Aurora B can phosphorylate various
components of the kinetochore, including Ndc80, to disrupt their interaction with microtubules (Cheeseman et al. 2002; 2006; DeLuca et al. 2006). Aurora B localizes to the inner centromere, in between sister kinetochores. Phosphorylation of substrates by Aurora B is dependent on their distance from the inner kinetochore (Liu et al. 2009). When kinetochore components are close to the inner centromere, they are phosphorylated by aurora B and kinetochore-microtubule attachments are destabilized (Figure 3A, left panel). However, when kinetochore components are pulled away from the inner centromere, phosphorylation by Aurora B is reduced and kinetochore-microtubule attachments are stabilized (Figure 3A, right panel). Importantly, amphitelic attachment, but not syntelic or monotelic attachment, stretches the kinetochore, thereby reducing Aurora B phosphorylation and stabilizing the attachment (Maresca and Salmon 2009).

While error correction can enrich for amphitelic attachments, it does not prevent anaphase onset in the presence of other attachments. The latter is the responsibility of the spindle assembly checkpoint. The spindle assembly checkpoint is activated by the presence of unattached kinetochores. When Ndc80 is not associated with microtubules, the kinase Mps1 can bind Ndc80 and initiate the pathway (Figure 3B) (Musacchio 2015). Mps1 phosphorylates kinetochore proteins to recruit other checkpoint components, including Bub3, Bub1, BubR1, Mad1, and Mad2. Together, the checkpoint complex inhibits Cdc20, a requisite activator of the anaphase-promoting complex (APC). Thus, in the presence of unattached kinetochores, checkpoint proteins bind the
kinetochore, APC is inhibited, and the cell does not progress into anaphase. Error correction and the checkpoint thus work synergistically. When Aurora B phosphorylates kinetochores without tension, it gives the kinetochore an opportunity to form an amphitelic attachment by simultaneously freeing the microtubule-binding site and delaying mitotic progression.
Figure 3. Error correction and the spindle assembly checkpoint. A) Depiction of kinetochore components relative to the zone of Aurora B activity (yellow circle) when the kinetochore is not (left panel) and is (right panel) under tension. B) Assembly of the mitotic checkpoint complex at an unattached kinetochore to activate the spindle assembly checkpoint (top panel). Entire figure adapted from Cheeseman 2014.
Given the current understanding of how error correction and the checkpoint function, merotelic attachments pose a problem for the cell. Because a kinetochore with merotelic attachment is bound by microtubules emanating from opposite spindle poles, the kinetochore may be stretched enough that it is protected from Aurora B phosphorylation and thus undetected by the checkpoint. However, if such an attachment persists into anaphase, the forces from opposite spindle poles can cause the chromosome to lag behind the remaining chromosomes and potentially segregate into the wrong daughter cell.

CHROMOSOME MISSEGREGATION

Causes

Analysis of animal cells in vitro has indicated that the most common, naturally arising chromosome segregation defect is a merotelic attachment that gives rise to a lagging chromosome in anaphase (Cimini et al. 2001; 2002; Thompson and Compton 2011). Because lagging chromosomes are spatially segregated from the two main masses of chromosomes, they tend to be excluded from the main nucleus and instead form a separate micronucleus when the chromosomes decondense and the nuclear envelope reforms in telophase. A study of a chromosomally stable human colorectal cancer cell line found micronuclei forming after 0.13% of mitoses (Thompson and Compton 2011). Importantly, in the majority (90%) of these cases, the micronucleus had formed in the proper daughter cell, presumably because the majority of microtubules attached to the chromosome emanated from the correct spindle pole (Cimini et al. 2004). In the remaining 10% of cases, the lagging chromosome
misseggregated and the micronucleus formed in the incorrect daughter cell, producing two aneuploid cells. From this study, the background rate of chromosome missegregation in mammalian cells was estimated to be once every $10^3-10^4$ divisions.

There are many mechanisms by which the rate of chromosome missegregation can be elevated above baseline—a condition known as chromosome instability. Certainly, direct disruption of the spindle assembly checkpoint will cause a dramatic increase in chromosome missegregation. Indeed, mutations in $Bub1b$, the gene encoding the checkpoint component BubR1, lead to chromosome missegregation and aneuploidy in up to one-third of cells in both mice and humans (Hanks et al. 2004; Baker et al. 2004; Knouse et al. 2014). Alternatively, certain situations that impaire kinetochore-microtubule attachments can go undetected by the spindle assembly checkpoint and lead to chromosome missegregation. For example, DNA damage and chromosome rearrangement can produce chromosomes lacking centromeres (acentric chromosomes) or chromosomes with two centromeres (dicentric chromosomes) which can fail to attach properly to the mitotic spindle without activating the spindle assembly checkpoint (Hande et al. 1999; Burrell et al. 2013). Additionally, defects in sister chromatid cohesion can lead to aberrant attachments that are unchecked by the error correction pathway (Barber et al. 2008). Finally, given that merotelic attachments can naturally lead to chromosome missegregation, any situations that increase merotely will increase chromosome missegregation. Impaired spindle formation, particularly the
delayed separation of centrosomes, can increase the prevalence of merotelic attachments (Silkworth et al. 2012). Additionally, in cells with multiple centrosomes it has been shown that centrosomes cluster into two poles prior to anaphase onset, a process that increases the prevalence of merotelic attachments (Figure 4) (Ganem et al. 2009; Silkworth et al. 2009).

**Figure 4**

**Figure 4. Supernumerary centrosomes as a source of increased merotelic attachments.** Depiction of mitotic progression in a cell harboring supernumerary centrosomes. Most cells will cluster centrosomes into two poles during mitosis and ultimately divide in a bipolar fashion, however this act of centrosome clustering increases the prevalence of merotelic attachments.

**Consequences**

Chromosome segregation defects can have immediate consequences for the chromosomes involved, particularly those that lag in anaphase. If lagging chromosomes remain in the spindle midzone at the onset of cytokinesis, they can become entrapped and damaged by the cleavage furrow (Janssen et al. 2011). Moreover, because micronuclei have impaired import of DNA replication and repair factors, the chromosomes within them often experience extensive damage and rearrangement (Crasta et al. 2012; Zhang et al. 2015). Thus, regardless of whether or not a lagging chromosome ultimately segregates into the proper
daughter cell, it can experience severe damage that would likely impact cellular fitness.

ANEUPLOIDY

The aneuploid state

If the chromosome segregates into the incorrect daughter cell, both daughter cells then experience the consequences of the aneuploid state. Given that changes in DNA copy number lead to commensurate changes in RNA and protein levels, these cells now experience a state of genome, transcriptome, and proteome imbalance (Torres et al. 2007; 2016). Within the first cell cycle following chromosome missegregation, aneuploid cells exhibit a prolonged G1 phase, impaired DNA replication, and proteotoxic stress (Santaguida et al. 2015).

Cells do not adapt to the immediate stresses caused by the aneuploid state. Instead, chronic aneuploidy is associated with a variety of fitness defects. Aneuploid cells exhibit transcriptional changes that mimic the environmental stress response (ESR) of yeast, a transcriptional signature that is induced by a variety of adverse conditions and associated with impaired growth and proliferation (Gasch 2007; Sheltzer et al. 2012). Indeed, aneuploid cells exhibit slowed growth (Torres et al. 2007; Williams et al. 2008; Gogendeau et al. 2015). Aneuploidy is also associated with inefficient energy utilization and sensitivity to inducers of energy stress (Williams et al. 2008; Tang et al. 2011). Additionally, aneuploid cells exhibit proteotoxic stress, including increased protein aggregation and increased autophagy, and are sensitive to inhibitors of protein quality control (Tang et al. 2011; Oromendia et al. 2012; Torres et al. 2007). These universal
characteristics of aneuploidy are believed to result from the cumulative effect of hundreds of proteins being expressed outside of their physiologic range. These changes in protein expression impair cell viability when the proteins are dosage sensitive or exist in complexes with other proteins for which expression is unchanged (Tang and Amon 2013). In addition to the universal effects of aneuploidy, there also exist phenotypes specific to a given karyotypic change (Dodgson et al. 2016).

**Aneuploidy in healthy organisms**

Consistent with the fitness defects associated with aneuploidy at the cellular level, chromosome copy number change is also adverse for the organism. Constitutional aneuploidy—aneuploidy that is present in all cells of an organism—arises from chromosome missegregation events in meiosis that produce aneuploid gametes and ultimately an aneuploid zygote. Meiotic missegregation events are not uncommon, with over 15% of adult female oocytes harboring whole-chromosome aneuploidy (Hou et al. 2013). In both mice and humans, constitutional aneuploidy is almost always associated with embryonic lethality (Hassold and Jacobs 1984; Gropp et al. 1983). In humans, only three single autosomal aneuploidies—trisomy 13, 18, and 21—can sometimes survive to term, but all of these cases are associated with severe developmental defects and reduced lifespan.

While it is well established that chromosome missegregation during meiosis is quite common and its consequences—constitutinal aneuploidy—quite severe, it is unknown how frequently chromosome missegregation occurs
during mitosis of somatic cells—producing somatic aneuploidy—and whether these events can be tolerated by the organism. Indeed, it has been suggested that two tissues—the brain and liver—naturally harbor high levels of aneuploidy and that this karyotypic variation contributes to organ function. Through spectral karyotyping (SKY) of neural progenitor cells in culture and fluorescence in situ hybridization (FISH) of neurons in tissue sections, several studies have suggested that over 20% of mouse and human neural progenitor cells and neurons are aneuploid (Rehen et al. 2001; 2005; Pack et al. 2005; Yurov et al. 2007; Westra et al. 2010). These aneuploid cells allegedly arise from chromosome missegregation events in neural progenitor cells during embryonic development (Yang et al. 2003). It is suggested that aneuploidy in the brain contributes to phenotypic diversity among neurons while also predisposing this organ to neurodegeneration (Muotri and Gage 2006; Arendt et al. 2010; Kennedy et al. 2012).

The mammalian liver is reported to harbor multiple types of genomic alterations. It has been appreciated for several decades that adult mammalian hepatocytes are polyploid (Beams and King 1942). Although hepatocytes are diploid in the neonatal liver, they become polyploid during early postnatal development by initiating mitosis but failing cytokinesis (Guidotti et al. 2003; Margall-Ducos et al. 2007). This produces hepatocytes with multiple copies of their genome as well as additional centrosomes. In adult mice, hepatocyte ploidy ranges from diploid to hexadecaploid, with over 50% of hepatocytes being tetraploid (Saeter et al. 1989). In adult humans, polyploidy is less extensive, with
between 50 and 75% of hepatocytes being diploid and most of the remainder being tetraploid (Leuchtenberger et al. 1954; Anti et al. 1994).

More recently, studies employing FISH have suggested that hepatocytes are also highly aneuploid, with over 50% of mouse and human hepatocytes harboring gains and losses of individual chromosomes (Duncan et al. 2010; 2012b). Hepatocyte aneuploidy is believed to arise secondary to hepatocytes being polyploid and harboring supernumerary centrosomes. Analysis of hepatocyte mitoses in vitro has shown that the supernumerary centrosomes can either remain separate, leading to a multipolar mitosis with severe chromosome missegregation, or cluster into two poles, yielding a bipolar division but still generating appreciable levels of merotely and chromosome missegregation (Duncan et al. 2010). Aneuploidy is proposed to endow the liver with phenotypic variability and adaptability in the face of the various metabolic insults imposed on this organ (Duncan et al. 2012a).

Both SKY and FISH have limitations—SKY has a high rate of spreading artifacts and low resolution while FISH has a high risk of hybridization artifacts and can only query a few chromosomes at once. As such, a detailed understanding of the degree and types of aneuploidy in these tissues is lacking. Moreover, there as been little investigation of aneuploidy other tissues as a means for comparison. A complete assessment of aneuploidy in somatic tissues is necessary in order to understand how aneuploidy might facilitate the function of specific organs. This thesis will present an alternate approach to investigate aneuploidy in somatic tissues.
Aneuploidy in cancer

Reports of elevated aneuploidy in certain tissues call into question the overwhelmingly adverse consequences of aneuploidy and suggest that, in certain cases, aneuploidy can enhance cellular fitness. The most striking example of such an exception is cancer. Cancer, a state in which somatic cells acquire genomic alterations that dramatically enhance their fitness, is associated with high levels of aneuploidy. It has long been appreciated that the majority of solid tumors harbor gains and losses of one if not many chromosomes (Albertson et al. 2003). As with every other type of genomic alteration in cancer, the ultimate question is whether aneuploidy is a driver or passenger in tumorigenesis. For many years, scientists have investigated how aneuploidy arises, when it arises, and what patterns of aneuploidy exist in cancer in order to understand its role in tumorigenesis.

The high levels of aneuploidy in some cancers have been attributed to cancer cells harboring elevated rates of chromosome missegregation—chromosome instability. There is no simple explanation for the source of chromosome instability in cancer. Indeed, most chromosomally unstable cancer cell lines have functional spindle assembly checkpoints (Gascoigne and Taylor 2008; Haruki et al. 2001; Tighe et al. 2001). Instead, chromosome instability has been attributed to different defects in different cancer cell lines—including DNA damage, impaired sister chromatid cohesion, altered microtubule dynamics, and supernumerary centrosomes (Burrell et al. 2013; Barber et al. 2008; Bakhoum et al. 2009; Ertych et al. 2014; Ganem et al. 2009; Silkworth et al. 2009). However, it is unclear whether these mechanisms can explain chromosome instability and
aneuploidy in tumors \textit{in vivo}. This thesis will explore an alternate explanation for chromosome instability and aneuploidy in cancer.

Aneuploidy arises early in tumorigenesis, as aneuploidy can be identified in the premalignant lesions that precede certain tumors (Meijer et al. 1998; Ried et al. 1996; Chin et al. 2004). This suggests that aneuploidy might be an early, driving lesion in tumorigenesis. Indeed, some, though not all, mouse models of chromosome instability have elevated spontaneous tumorigenesis (Jeganathan et al. 2007; Li et al. 2009; Sotillo et al. 2007; Weaver et al. 2007). Moreover, analysis of karyotypic changes across tumors have revealed specific chromosome gains and losses that frequently occur in tumors (Albertson et al. 2003; Beroukhim et al. 2010; Zack et al. 2013). Analysis of these recurrent lesions has shown that frequently gained chromosomes tend to have a high density of oncogenes, while frequently lost chromosomes tend to be enriched for tumor suppressor genes (Davoli et al. 2013). These studies suggest that aneuploidy could drive tumorigenesis by changing the dosage of oncogenes and tumor suppressor genes. However, this hypothesis is based solely on the most common lesions in tumors and therefore it is unclear whether this can explain all of the karyotypic changes observed in tumors. Moreover, it is unknown how cancer can tolerate the adverse consequences of aneuploidy and instead exploit karyotypic change for enhanced fitness. This thesis will reevaluate karyotypic changes within and across cancer types in order to provide a thorough assessment of aneuploidy and a foundation for assessment of its role in tumorigenesis.
COPY NUMBER VARIANTS

Copy number variants (CNVs) refer to duplications or deletions of regions of the genome ranging from several base pairs (bp) to several megabases (Mb). Unlike whole-chromosome copy number changes, which arise during mitosis, sub-chromosome CNVs can arise at any point in the cell cycle by various mechanisms. These include erroneous recombination, particularly that which occurs between highly similar sequences in a process known as non-allelic homologous recombination (NAHR) (Sharp et al. 2005; Redon et al. 2006; Conrad et al. 2010). Other forms of recombination, such as non-homologous end-joining (NHEJ), can also lead to CNVs (Conrad et al. 2010). Finally, errors during DNA replication can lead to fork stalling and DNA breaks, ultimately leading to the generation of CNVs through various mechanisms (Lee et al. 2007; Hastings et al. 2009).

Large-scale CNVs, those exceeding one megabase in length, can alter the dosage and thus expression of up to hundreds of genes, leading to genome, transcriptome, and proteome changes that are almost as drastic as whole-chromosome aneuploidy (Henrichsen et al. 2009). Not surprisingly, large-scale CNVs that are present organism-wide, or constitutionally, are associated with adverse consequences. While constitutional CNVs exceeding 100 kilobases are present in at least 65% of individuals, constitutional CNVs exceeding one megabase are present in only about 1% of individuals (Itsara et al. 2009). Moreover, large-scale CNVs are often associated with disease and experience negative selection in populations (Itsara et al. 2010; Girirajan et al. 2011).
It is unclear whether large-scale CNVs could be tolerated at the somatic level. Sequencing-based approaches have been applied to tissues such as blood and skin and indeed have identified megabase-scale copy number alterations in up to 4% of individuals. However, because these approaches are applied to populations of cells they are only sensitive to changes present in greater than 5% of cells in the population (Forsberg et al. 2012; Jacobs et al. 2012; Laurie et al. 2012). Like aneuploidy, there is reason to believe that large-scale CNVs could be tolerated, and perhaps benefit, individual cells as they are extremely common in cancer (Beroukhim et al. 2010; Zack et al. 2013). This thesis will provide a high-resolution assessment of large-scale CNVs in somatic tissues.

SUMMARY

Much of our understanding of mammalian mitosis—the mechanisms for its fidelity and the causes and extent of infidelity—has been based on the expansion of cell lines in vitro, many of which have been transformed. However, mammalian mitosis normally takes place in the context of tissues, which have specialized architecture and function. It is possible that the surrounding tissue influences aspects of mitosis. Indeed, preliminary evidence from the brain and liver suggests that mitosis is modified in some situations for tissue-specific functions, although the methods of investigation suffer from poor sensitivity and high artifacts. In order to elucidate the finer details of mammalian mitosis and understand how defects in this process might lead to or be coopted in cancer, it is critical to investigate mitosis in the context of the tissue.
A fundamental question is whether aneuploidy and large-scale copy number alterations exist in normal tissues and, if so, to what extent. From this, it will be important to understand how tissues either prevent or generate and tolerate chromosome missegregation and copy number alteration. These questions are particularly relevant for the liver, where polyploidy is expected to produce high levels of aneuploidy. These investigations could ultimately explain how chromosome segregation in normal tissues might be disrupted during tumorigenesis.

This thesis will address these questions. First, we present the application, validation, and use of single cell sequencing to determine the prevalence of megabase-scale copy number variants and whole-chromosome aneuploidy. We then describe the unexpected finding that the architecture of a tissue is important for facilitating chromosome segregation in epithelial cells. Finally, we present an analysis of the growing cancer genome sequencing datasets to assess the landscape and evolution of aneuploidy in cancer in order to better understand the role of karyotypic changes in tumorigenesis.

REFERENCES


Meijer GA, Hermsen MAJA, Baak JPA, van Diest PJ, Meuwissen SGM, Belien JAM, Hoovers JMN, Joenje H, Snijders PJF, Walboomers JMM. 1998. Progression from colorectal adenoma to carcinoma is associated with non-


Santaguida S, Richardson A, Iyer DR, MSaad O, Zasadil LM, Knouse KA, Wong Y, Rhind N, Desai A, Amon A. Aneuploidy induction generates cell cycle-arrested cells with complex karyotypes that are eliminated by the immune system.


37
CHAPTER TWO

Assessment of megabase-scale somatic copy number variation using single-cell sequencing

Reprinted with permission from Cold Spring Harbor Laboratory Press


KAK led the study and performed all experiments with the exception of Figures 1A,D and 6 which were performed by JW. JW processed all sequencing data and KAK and JW analyzed the data.
ABSTRACT

Megabase-scale copy number variants (CNVs) can have profound phenotypic consequences. Germline CNVs of this magnitude are associated with disease and experience negative selection. However, it is unknown whether organismal function requires that every cell maintain a balanced genome. It is possible that large somatic CNVs are tolerated or even positively selected. Single cell sequencing is a useful tool for assessing somatic genomic heterogeneity but its performance in CNV detection has not been rigorously tested. Here we develop an approach that allows for reliable detection of megabase-scale CNVs in single somatic cells. We discover large CNVs in 8-9% of cells across tissues and identify two recurrent CNVs. We conclude that large CNVs can be tolerated in subpopulations of cells and that particular CNVs are relatively prevalent within and across individuals.
INTRODUCTION

Copy number variants (CNVs) can range in size from hundreds to millions of base pairs. Copy number changes affect approximately seven times as many base pairs as single nucleotide variants and are major contributors to inter-individual differences (Sudmant et al. 2015). Over 65% of individuals harbor a germline CNV of at least 100 kilobases (kb) and at least 1% of individuals have a CNV exceeding one megabase (Mb) (Itsara et al. 2009). Although megabase-scale CNVs could be considered collectively common, the specific CNVs themselves are rare and often associated with disease (Girirajan et al. 2011). Not surprisingly, large CNVs experience negative selection and their existence in a population is largely due to de novo events (Itsara et al. 2010).

While germline, megabase-scale CNVs are found in 1% of individuals, the prevalence of somatic CNVs is only beginning to be investigated. Array-based analyses of populations of cells from many individuals provided initial insight into this question. These studies identified megabase-scale somatic aberrations in up to 4% of individuals, however the sensitivity was limited to CNVs present in more than 5% of cells (Forsberg et al. 2012; Jacobs et al. 2012; Laurie et al. 2012). These studies are thus blind to alterations that arise late in development or adversely affect fitness, as this would limit their propagation in a cell population. With the emergence of methods to amplify the genome of a single cell, single cell sequencing now provides an alternate means of assessing the prevalence of somatic CNVs and offers the advantage of detecting variants that exist in as few as one cell. Recently, two groups performed low-coverage
sequencing of single human neurons and reported at least one megabase-scale CNV in over 40% of neurons (McConnell et al. 2013; Cai et al. 2014). These findings suggest much greater tolerance of large somatic CNVs compared to germline CNVs and raise the interesting possibility that somatic genomic heterogeneity contributes to phenotypic diversity within a tissue. However, it is still unclear how CNV detection methods perform when applied to individual cells, as single cell sequencing poses unique problems for CNV detection. For one, single cells are usually sequenced at very low coverage. Second, genome representation in the sequencing library can vary independently of copy number due to inefficient and uneven genome fragmentation and amplification. Moreover, any alterations identified in a single cell cannot be verified by an independent method. Therefore, it is imperative that appropriate quality control and analytic methods are employed such that the sensitivity—the likelihood that a real CNV of defined size is detected—and specificity—the likelihood that a detected CNV represents a real change in copy number—of an approach are known and optimized in the context of single cell sequencing data.

Here, we use a variety of methods to quantify the sensitivity and specificity of different approaches for megabase-scale CNV detection in single cell sequencing data. We develop an approach with higher specificity than those used previously and use this approach to analyze single cell sequencing data from normal human brain and skin. From this analysis, we infer the prevalence of megabase-scale CNVs across somatic tissues.
RESULTS

Characterizing sequencing data from single somatic cells

We previously isolated single cells from fresh postmortem brain and skin samples from four adults without neurologic or dermatologic disease (Knouse et al. 2014). Genomic DNA from a total of 105 brain cells (~75% of which are neurons) from all four individuals and a total of 55 keratinocytes from two of these individuals was amplified by linker adapter PCR and sequenced at low coverage (0.1X) (Table 1).

Table 1. Description of Dataset

<table>
<thead>
<tr>
<th>Individual</th>
<th>Age</th>
<th>Sex</th>
<th>Tissue</th>
<th>Total Cells Sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>130412</td>
<td>70</td>
<td>M</td>
<td>Brain</td>
<td>20</td>
</tr>
<tr>
<td>130604</td>
<td>52</td>
<td>M</td>
<td>Brain</td>
<td>22</td>
</tr>
<tr>
<td>140131</td>
<td>48</td>
<td>F</td>
<td>Brain</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Skin</td>
<td>30</td>
</tr>
<tr>
<td>140213</td>
<td>68</td>
<td>M</td>
<td>Brain</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Skin</td>
<td>25</td>
</tr>
</tbody>
</table>

To quantify variation in read depth across the genome and identify cells suitable for analysis, we previously calculated a variability score (VS) for each cell (Knouse et al. 2014). The variability score is generated by averaging the standard deviation in read depth in sliding windows across each chromosome and averaging the average standard deviation for the three autosomes with highest variability. While this is suitable for whole-chromosome copy number analysis, it could bias sub-chromosome copy number assessment as copy number changes within each chromosome could increase the VS. To assess the impact of CNVs on VS, we recalculated the VS of each cell by excluding
windows with read depths above or below threshold for diploid copy number. The VS of only three of 160 cells changed when we excluded non-diploid regions of the genome. In these three cells, the VS changed by less than 0.02 (Fig. 1A). This analysis indicates that copy number changes are not responsible for the majority of read depth variation. Regardless, we used the recalculated VS for all subsequent analyses.

In our dataset, the VSs ranged from 0.14 to 1.03, with a median of 0.19 and a long upper tail (Fig. 1B). The majority of cells with high VS were brain cells and the median VS in brain was significantly higher than in skin (0.2 vs. 0.18, Mann-Whitney U test, \( p < 0.02 \)). Cells with high VSs were characterized by a wide spread of read depths across adjacent genomic windows (Fig. 1C, middle panel) and/or an abundance of segments with copy number changes, many of which were non-integer and hence incompatible with gains and losses in a single cell (Fig. 1C, bottom panel, blue boxes). The VS is therefore an unbiased measure of sequencing data quality and a valid criterion for inclusion or exclusion of a cell from copy number analysis.
Figure 1

A

- All
- CNVs Excluded

B

VS

All Brain Skin

0.00

0.35

C

Hidden Markov Model

E = 0.9999999

E = 0.995

E = 0.99

E = 0.9

D

Minimum CNV Detected (Mb)

Non-Simulated CNVs

VS

0.17 0.2 0.21 0.24 0.28 0.35 0.5 0.85

E = 0.0001

E = 0.001

E = 0.01

E = 0.05

Minimum CNV Detected (Mb)

Non-Simulated CNVs

VS

0.17 0.2 0.21 0.24 0.28 0.35 0.5 0.85

Circular Binary Segmentation

VS

0.17 0.2 0.21 0.24 0.28 0.35 0.5 0.85

α = 0.0001

α = 0.001

α = 0.01

α = 0.05
Figure 1. Variability scores of dataset and testing sensitivity of CNV detection
A. The VSs of all brain and skin cells calculated from read depth across all genomic windows (all) or only genomic windows with diploid copy number (CNVs excluded).
B. The recalculated VSs of all cells sequenced. Lines and whiskers indicate median and interquartile range, respectively. The median VS in brain is significantly higher than in skin (0.2 vs. 0.18, Mann-Whitney $U$ test, $p < 0.02$).
C. Segmentation plots of cells with VSs of 0.20, 0.39, and 0.34 as processed by HMM at $E = 0.9999999$. Segments with log$_2$ ratios above the threshold for gain are colored in red and segments with log$_2$ ratios below the threshold for loss are colored in green. Cells with high VS have a wide spread in read depth across adjacent genomic windows (middle panel) and/or many segments with non-integer copy number states (bottom panel, blue boxes).
D. The minimum CNV size detectable in single cells with VSs ranging from 0.17 to 0.85 using a hidden Markov model (HMM) (first row) or circular binary segmentation (CBS) (third row) at different values of $E$ or $\alpha$, respectively. Gains and losses are represented by red and blue bars, respectively. The number of non-simulated CNVs identified in each cell by each approach is shown in the second and fourth rows. Asterisks indicate cases in which more than 20 non-simulated CNVs were identified.

Optimizing sensitivity and specificity of CNV detection in single cell sequencing data

To determine the prevalence of CNVs in brain and skin we tested two methods for copy number analysis. Both approaches quantify sequencing reads in genomic windows of approximately 500 kb and adjust read counts for mappability and GC bias within each window. The two approaches differ in how copy number segments are inferred from read depth in each window. One uses a hidden Markov model (HMM) (Ha et al. 2012) while the other uses circular binary segmentation (CBS) (Olshen et al. 2004). Both HMM and CBS feature an adjustable parameter that determines the likelihood that a change in copy number is identified. For HMM, the adjustable parameter $E$ dictates the
probability of extending a segment of defined copy number. Thus, as $E$ is lowered, more copy number changes are called. For CBS, the adjustable parameter $\alpha$ dictates the significance level required to accept a change in copy number. Therefore, higher levels of $\alpha$ allow for more copy number changes. Changing the values of these two parameters therefore adjusts the sensitivity and specificity of the two approaches.

To broadly test the performance of these two algorithms we simulated gains and losses ranging from 2.5 to 20 Mb in sequencing data from single cells with different variability scores. For each cell, we determined the number of simulated and non-simulated CNVs that were detected by HMM and CBS at different values of $E$ and $\alpha$. Several trends emerged. First, HMM had better sensitivity for gains while CBS had better sensitivity for losses (Fig. 1D, first and third panels). Second, overall sensitivity could be increased by reducing $E$ or increasing $\alpha$, improving the detection of 2.5 Mb and 5 Mb CNVs. Notably, doing so also resulted in the detection of many non-simulated CNVs especially in cells with high VS (Fig. 1D, second and fourth panels). Finally, for any algorithm at any parameter value, in cells with VS greater than 0.3 there was a dramatic reduction in the detection of simulated CNVs, especially gains, while as many as 100 non-simulated CNVs were identified (Fig. 1D).

To test the sensitivity of these approaches to real CNVs, we sequenced single fibroblasts from individuals known to harbor germline CNVs of defined size. All methods except HMM at $E = 0.9999999$ reliably detected 10 and 20 Mb gains and losses while only HMM at $E = 0.9$ detected CNVs less than 5 Mb (Fig.
We conclude that CNVs 5 Mb or larger can be detected by multiple approaches in single cells sequenced at 0.1X coverage.

Our simulations showed that algorithm sensitivity could be adjusted to identify CNVs of less than 5 Mb. However, this led to increased detection of non-simulated CNVs, suggesting that specificity was compromised (Fig. 1D, second and fourth panels). In these simulation experiments, as with all single cell sequencing experiments, it is impossible to determine whether the non-simulated CNVs represent true CNVs undetectable at less sensitive parameters or false CNVs caused by random fluctuations in read depth that are inappropriately identified as CNVs when sensitivity is increased. The next best way to verify CNVs is to sequence cells that are closely related to each other, ideally the two products of a cell division. Barring errors during DNA replication, two daughter cells should have identical or complimentary CNVs. A CNV present in only one of the cells, henceforth called a private CNV, therefore likely represents a false positive CNV or a CNV that failed to be detected in the other cell(s).

We expanded a single fibroblast in culture for approximately seven divisions to yield ~100 cells and sequenced five cells from this population. As any two cells in this population have been genetically distinct for at most seven generations, most CNVs should be shared by multiple cells. As a complimentary approach, we isolated single cells and split the cell contents in half after cell lysis and DNA fragmentation but prior to whole genome amplification. As the lysis and fragmentation steps generate DNA fragments ~1,000-fold smaller than the windows at which we bin reads, the CNVs should be identical between the two
samples. For both of these experiments, we then calculated the proportion of CNVs that were private in each cell or sample.

We were surprised to find that for most parameters of HMM and CBS, over half of the identified CNVs were private (Fig. 2B). Most of these private CNVs likely represent false CNVs, rather than real CNVs that failed to be detected in related cells or samples, because only 17% of private CNVs identified for a given algorithm at a less sensitive parameter were subsequently identified in other cells by the same algorithm using a more sensitive parameter. Across all algorithms and parameters, the private CNVs ranged in length from 0.5 to 34 Mb with the majority of private CNVs (63%) less than 5 Mb (Fig. 2D). Not surprisingly, the least sensitive algorithm and parameter, HMM at $E = 0.9999999$, was the most specific, with seven of nine cells lacking private CNVs when analyzed by this method. We also noted, especially in the split reactions, that increasing HMM and CBS sensitivity led to reduced specificity (Fig. 2B).

Importantly, while in a given sample HMM or CBS tended to identify the same private CNV at varying levels of sensitivity, only 15% of private CNVs were identified by both HMM and CBS (Fig. 2E). All of the private CNVs identified by both HMM and CBS were unique to a given cell with the exception of a 20 Mb gain on chromosome 19, which was identified in four of the nine samples. This CNV was not detected when cells were sequenced at higher (1X) coverage, indicating that it is an artifact of low-coverage sequencing.

From these experiments, we conclude that private CNVs, the majority of which we believe to be false CNVs, can occur at the scale of megabases and
can account for around half of all putative CNVs when algorithm parameters are
adjusted to increase the sensitivity to detect CNVs 5 Mb and smaller.
Furthermore, we find that HMM and CBS identify different private CNVs,
suggesting that employing both algorithms simultaneously could enhance
specificity. Indeed, when we only considered CNVs detected by both HMM at E
= 0.995 and CBS at α = 0.0001, and ignored the common gain on chromosome
19, the prevalence of private CNVs fell to zero in two of the five cells from a small
population and all of the split samples (Fig. 2B). Importantly, doing so did not
compromise the sensitivity for CNVs exceeding 5 Mb (Fig. 2A). We decided that
the overlap of HMM and CBS at E = 0.995 and α = 0.0001, respectively, afforded
the best combination of sensitivity and specificity for the purpose of detecting
megabase-scale CNVs in somatic cells. These parameters allow for the detection
of all gains and losses 10 Mb and larger. Approximately half of 5 Mb CNVs can
also be detected, with losses more easily detected than gains (Figure 2C).
Figure 2

A

B

Hidden Markov Model | Circular Binary Segmentation | Both (Chr 19 Gain Excluded)

<table>
<thead>
<tr>
<th>0.9999999</th>
<th>0.995</th>
<th>0.99</th>
<th>0.995</th>
<th>0.99</th>
<th>0.995</th>
<th>0.99</th>
<th>0.995</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Discovery (5 Mb Gain)</td>
<td>40%</td>
<td>40%</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
<td>60%</td>
<td>63%</td>
<td>60%</td>
</tr>
<tr>
<td>True Discovery (5 Mb Loss)</td>
<td>0%</td>
<td>75%</td>
<td>75%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>False Discovery (Small Population)</td>
<td>23%</td>
<td>53%</td>
<td>52%</td>
<td>49%</td>
<td>51%</td>
<td>45%</td>
<td>46%</td>
<td>59%</td>
</tr>
<tr>
<td>False Discovery (Split Amplification)</td>
<td>0%</td>
<td>25%</td>
<td>59%</td>
<td>61%</td>
<td>25%</td>
<td>46%</td>
<td>42%</td>
<td>43%</td>
</tr>
</tbody>
</table>

D

E

Hidden Markov Model | Circular Binary Segmentation | Both (Chr 19 Gain Excluded)

| E = 0.9999999 | α = 0.0001 | E = 0.995, α = 0.0001 | E = 0.995 | α = 0.0001 | E = 0.99 | α = 0.01 | E = 0.99, α = 0.01 | E = 0.9 | α = 0.05 |

| 50 |
Figure 2. Testing sensitivity and specificity of CNV detection in single cell sequencing data

A. The discovery rate of known CNVs ranging from 3 to 20 Mb in single cells using HMM, CBS, or both at different values of $E$ or $\alpha$. Four or five cells were sequenced for each CNV.

B. The average proportion of CNVs that are private in individual cells sharing a recent ancestor or DNA split after fragmentation from a single cell when analyzed by HMM, CBS, or both at various parameters.

C. The true discovery rates (for 5Mb gains and losses) and the false discovery rates (averaged across samples from small population or split amplification) for CNV detection using HMM, CBS, or both at different values of $E$ or $\alpha$, regardless of VS.

D. The distribution of private CNV type and size in individual cells sharing a recent ancestor or DNA split after fragmentation from the same cell when analyzed by HMM, CBS, or both at various parameters.

E. The proportion of private CNVs called by more than one parameter of HMM or CBS and the proportion of private CNVs called by both HMM and CBS.

Prevalence of somatic CNVs

We applied the combination of HMM and CBS at $E = 0.995$ and $\alpha = 0.0001$, respectively, to all brain and skin cells. We ignored the gain on the proximal portion of chromosome 19 that we found to be an artifact of low sequencing coverage and only considered CNVs present on autosomes. In brain, but not skin, we observed a positive correlation between the VS of a cell and the number of CNVs identified (Pearson $r = 0.53$, $p < 0.0001$, Fig. 3). Brain cells in the top 10th percentile for VS contained 72% of all CNVs identified in brain (Fig. 4A). Thus, the inferred CNV prevalence is strongly influenced by the cutoff used to select cells for analysis (Fig. 4B,C).
Figure 3. Relationship between variability score and CNV prevalence
The total number of CNVs and the VS of individual brain and skin cells. The number of CNVs and VS for brain cells are positively correlated (Pearson $r = 0.5308$, $p < 0.0001$). Arrow points to a brain cell with 6 CNVs and a VS of 0.22.
Figure 4.

A. The variability scores for all brain and skin cells sequenced. The number adjacent to a given point or group of points indicates the total number of CNVs identified in that cell or group of cells. The dashed line indicates a VS cutoff of 0.26.

B. The percent of brain and skin cells harboring at least one megabase-scale CNV using different VS cutoffs to exclude cells from analysis.

C. The distribution of the total number of megabase-scale CNVs per cell in brain and skin using different VS cutoffs to exclude cells from analysis.
Understanding the causality between VS and CNVs is crucial given the positive correlation between a cell's VS and the number of CNVs in brain. High variability in read depth could lead to identification of CNVs that are not real. However, it is also possible that many CNVs increase read depth variability. Several observations argue against the latter. For one, when we recalculated VSs by excluding CNVs, the VSs hardly changed, indicating that the many CNVs in cells with high VS are not solely responsible for the high VS. Indeed, much of the variation in read depth in cells with high VS came in the form of fluctuations within the range of diploid copy number (Fig. 1C, blue boxes). Furthermore, our dataset includes a brain cell with six CNVs and a VS of only 0.22 (Fig. 3, arrow), indicating that high VS is not an obligatory consequence of harboring many CNVs.

The above observations suggest that high VS is secondary to reasons other than variation in copy number and that the many CNVs identified in such cells are likely to be false. Our private CNV analysis supports this hypothesis. The number of private CNVs was much higher in cells with high VS (Fig. 2B). Moreover, our simulations indicated that the sensitivity of CNV detection drops as VS exceeds 0.3 (Fig. 1D, first and third panels). Because high VS not only complicates detection of real CNVs but also increases the discovery of false CNVs, we only considered cells with VS of less than 0.26 for our analysis (Fig. 4A, dashed line). This VS value renders the brain and skin cells equivalent in VS distribution (mean = 0.19 for both tissues). When applying HMM and CBS at $E = 0.995$ and $\alpha = 0.0001$, respectively, to cells with VS < 0.26, we expect to detect
20% of 5 Mb gains, 75% of 5 Mb losses, and all CNVs 10 Mb and larger with a false discovery rate of < 17% (Figure 2). Out of 132 brain and skin cells with VS < 0.26, we identified a total of 23 CNVs distributed across 7 brain and 4 skin cells (Fig. 5, Table 2).

**Figure 5**

![Karyogram showing the 23 gains and losses identified across 80 brain cells and 52 skin cells with VS < 0.26. Gains and losses are represented by red and blue bars, respectively.](image)

**Figure 5. Megabase-scale CNVs in somatic cells**

Karyogram showing the 23 gains and losses identified across 80 brain cells and 52 skin cells with VS < 0.26. Gains and losses are represented by red and blue bars, respectively.
<table>
<thead>
<tr>
<th>Individual</th>
<th>Tissue</th>
<th>Cell (VS)</th>
<th>Chr</th>
<th>Start (Mb)</th>
<th>End (Mb)</th>
<th>Length (Mb)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>130412</td>
<td>Brain</td>
<td>1780 (0.19)</td>
<td>22</td>
<td>0.00</td>
<td>51.18</td>
<td>51.18</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1783 (0.22)</td>
<td>1</td>
<td>245.72</td>
<td>248.32</td>
<td>2.60</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>13.36</td>
<td>20.92</td>
<td>7.56</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>0.00</td>
<td>3.20</td>
<td>3.20</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1785 (0.16)</td>
<td>6</td>
<td>0.00</td>
<td>3.65</td>
<td>3.65</td>
<td>L</td>
</tr>
<tr>
<td>130604</td>
<td>Brain</td>
<td>2595 (0.19)</td>
<td>8</td>
<td>0.00</td>
<td>4.41</td>
<td>4.41</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2607 (0.17)</td>
<td>8</td>
<td>0.00</td>
<td>4.96</td>
<td>4.96</td>
<td>L</td>
</tr>
<tr>
<td>140131</td>
<td>Skin</td>
<td>692 (0.24)</td>
<td>2</td>
<td>143.27</td>
<td>148.36</td>
<td>5.09</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>696 (0.19)</td>
<td>5</td>
<td>46.32</td>
<td>180.72</td>
<td>134.40</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>698 (0.25)</td>
<td>2</td>
<td>0.00</td>
<td>8.27</td>
<td>8.27</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>0.00</td>
<td>6.11</td>
<td>6.11</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>136.00</td>
<td>141.00</td>
<td>5.00</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>120.00</td>
<td>122.50</td>
<td>2.50</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td>28.42</td>
<td>32.50</td>
<td>4.08</td>
<td>L</td>
</tr>
<tr>
<td>140213</td>
<td>Brain</td>
<td>962 (0.24)</td>
<td>4</td>
<td>156.00</td>
<td>163.00</td>
<td>7.00</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>988 (0.21)</td>
<td>9</td>
<td>117.50</td>
<td>120.50</td>
<td>3.00</td>
<td>G</td>
</tr>
</tbody>
</table>
Validation of somatic CNVs

While the nature of single cell sequencing renders it impossible to validate the CNVs identified in a single cell by an orthogonal method, we employed multiple analyses to increase our confidence that the CNVs we identified represent true gains and losses of genomic material. First, we resequenced 4 cells identified to have 6, 5, 3 and 1 CNVs and 2 cells with no CNVs at 10-fold higher coverage (1X). This analysis revealed that our CNV identification parameters are robust. The two cells without CNVs did not have any CNVs upon increased coverage. Furthermore, 87% of CNVs identified in cells at 0.1X coverage were also identified in cells at 1X coverage (Table 3). Two small (less than 5.5 Mb) CNVs identified at 0.1X coverage were no longer detected at 1X coverage, and 4 small (less than 3 Mb) CNVs were identified at 1X coverage that were not detected at 0.1X coverage (Table 3).
### Table 3. CNVs Identified in Somatic Cells at Low (~0.1X) and High (~1X) Coverage

<table>
<thead>
<tr>
<th>Individual</th>
<th>Tissue</th>
<th>Cell (VS)</th>
<th>Low (~0.1X) Coverage</th>
<th>Higher (~1X) Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>130412</td>
<td>Brain</td>
<td>1783 (0.22)</td>
<td>Chr</td>
<td>Start (Mb)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>245.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1786 (0.22)</td>
<td>8</td>
<td>13.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>161.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>99.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>127.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>66.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>62.51</td>
</tr>
<tr>
<td>140131</td>
<td>Skin</td>
<td>696 (0.19)</td>
<td>5</td>
<td>46.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>698 (0.25)</td>
<td>2</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>136.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>129.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td>28.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>7.50</td>
</tr>
</tbody>
</table>

*Appeared at higher coverage* | *Disappeared at higher coverage*
Resequencing cells at higher coverage allowed us to look for loss of heterozygosity (LOH) in putative deletions. We pooled all cells sequenced from each individual to identify single nucleotide polymorphisms (SNPs) for which the individual was heterozygous. We then determined the state of these heterozygous SNPs in the cells sequenced at higher coverage. We first measured the distance between two successive heterozygous SNPs outside the regions of putative deletion in the single cells. These distances ranged from 2 base pairs to 58 Mb with a median of 2.8 Mb (Figure 6A). We then calculated the distance between heterozygous SNPs flanking (and, if present, within) putative deletions. The distance between heterozygous SNPs flanking/inside putative deletions was significantly higher than the distance between heterozygous SNPs outside putative deletions for two of the three cells (Mann-Whitney U test, \( p = 0.014, p < 0.0001 \), Figure 6A). Encouraged by this result, we analyzed the state of SNPs within the putative deletions. At 1X coverage the status of only a handful of SNPs could be assessed in each putative deletion. However, when we combined all SNPs for the 13 putative deletions that could be analyzed, there was a significant depletion of heterozygous SNPs within the putative deletions compared to outside the alleged deletions (Fisher’s exact test, \( p = 0.0001 \)). Nine of the 13 putative deletions did not harbor a single heterozygous SNP (Fig. 6B). The remaining four deletions had a single heterozygous SNP. For one of these deletions, a loss on the proximal portion of chromosome 8, the heterozygous SNP occurred within the first 500 kb of the putative deletion and thus could be a consequence of binning reads in ~500 kb windows. For two other deletions, the
heterozygous SNP was within at least 1.5 Mb of one boundary. This suggests that the CNV either was identified incorrectly or that there are two deletions joined by a small region of normal copy number. Finally, one heterozygous SNP was identified in a deletion that was not detected upon sequencing at higher depth, further supporting the conclusion that this deletion in not real.

We also analyzed 13 putative deletions in two cells with VS > 0.26. As before, we observed a significant increase in the distance between heterozygous SNPs within deletions as compared to outside deletions for both cells (Mann-Whitney U test, $p = 0.0013$, $p < 0.0001$, Figure 6C). However, of the 13 deletions, only three lacked heterozygous SNPs and there was no significant depletion of heterozygous SNPs within the putative deletions compared to outside these deletions (Fisher’s exact test, $p = 0.892$, Figure 6D). Together, our SNP analysis revealed that in low-coverage, single cell sequencing data, continuous homozygosity can span megabases even in allegedly euploid regions. Thus, loss of heterozygous SNPs cannot rigorously confirm megabase-scale deletions, but the identification of heterozygous SNPs can identify false deletions. The significant increase in the distance between heterozygous SNPs flanking/inside deletions combined with the significant decrease in heterozygous SNPs within deletions in our dataset supports but does not prove their existence. However, the identification of heterozygous SNPs within many of the deletions from cells excluded from analysis underscores the association between high VS and false CNVs and justifies the exclusion of such cells from analysis.
Figure 6

A

B

C

D

Percent Het SNPs
Het in Single Cell
Outside Inside

Outside Inside

1 Mb

61
Figure 6. Analysis of SNPs in putative deletions in single cells
A. The distance between successive heterozygous SNPs outside or flanking/inside putative deletions in cells resequenced at higher coverage. For deletions, the distance between the two heterozygous SNPs located closest to either breakpoint was calculated. If heterozygous SNPs were present within the deletion, multiple smaller distances were calculated using the heterozygous SNPs flanking the breakpoints and the heterozygous SNPs within the deletion. Lines and whiskers indicate median and interquartile range, respectively. The distance between heterozygous SNPs flanking/inside deletions is significantly higher than the distance outside deletions for cells 1783 and 1786 but not for cell 696 (Mann-Whitney U test, p = 0.014, p < 0.0001, p = 0.14).
B. The status of heterozygous SNPs in putative losses from cells resequenced at higher coverage. Heterozygous SNPs which were heterozygous in the single cell are represented with E while heterozygous SNPs which were homozygous in the single cell are represented with O. The percent of heterozygous SNPs which are heterozygous outside (n = 50 per CNV) and inside (n = variable) the putative deletion are indicated.
C. The distance between successive heterozygous SNPs outside or flanking/inside putative deletions in cells excluded from analysis (VS > 0.26) which were resequenced at higher coverage. The distance between heterozygous SNPs flanking/inside deletions is significantly higher than the distance outside deletions for both cells (Mann-Whitney U test, p = 0.0013, p < 0.0001).
D. The status of heterozygous SNPs in putative losses from cells excluded from analysis (VS > 0.26) which were resequenced at higher coverage.

While loss of heterozygosity is consistent with a deletion, it cannot distinguish a true genomic deletion from loss of DNA during cell lysis or whole genome amplification. However, in contrast to loss of DNA during whole genome amplification, most mechanisms that would generate true genomic deletions would join two previously separated regions of the genome. This juxtaposition could be identified by paired-end sequencing, assuming the fragment containing the junction did not drop out during whole genome amplification. To determine whether we could identify such chimeric DNA fragments we resequenced one cell harboring multiple deletions to 2X coverage with paired-end reads. We identified multiple (3-5) discordant reads flanking the breakpoints for half of the
deletions. However, we also identified thousands of discordant reads mapping to distant positions within and across chromosomes in allegedly euploid regions, and in some cases a handful mapped to similar positions as they did for half of the putative deletions. We suspect this is due to chimera formation during whole genome amplification. In light of the not insignificant probability that junction fragments dropout during whole genome amplification (low signal), the high level of discordant reads in euploid regions (high background), and the financial costs associated with resequencing at higher coverage, we concluded that paired-end sequencing is not a viable approach for validating CNVs identified across many single cells. In summary, although SNP analysis can provide supporting evidence for the existence of deletions, it remains difficult to validate putative CNVs identified across many cells. This emphasizes the importance of using a CNV detection algorithm that minimizes the false discovery rate.

Characteristics of somatic CNVs

Using the CNVs detected by our empirically validated detection algorithm and exclusion criteria (Fig. 5, Table 2), we went on to characterize the CNVs identified in brain and skin. To avoid biasing the data, we kept the two deletions that disappeared at higher sequencing depth in our dataset and did not incorporate the four additional CNVs identified at higher depth. Our analysis revealed that 9% of brain cells and 8% of skin cells harbor at least one megabase-scale CNV (Fig. 4B). Across the individuals we analyzed, the prevalence of cells harboring megabase-scale CNVs ranged from 0% to 20% in brain and 4% to 11% in skin (Fig. 7). Interestingly, CNV occurrence was not
independent, as the 23 CNVs were distributed among only 11 of 132 cells (Poisson $p < 10^{-5}$; Fig. 4C). Losses were much more common than gains, but given that one of the gains comprised an entire chromosome arm, losses affected only slightly more of the genome (186.23 Mb of losses versus 137.4 Mb of gains, Fig. 8A). Telomeric CNVs were highly enriched (Fisher’s exact test, $p < 0.0001$), with 11 of the 23 CNVs extending to the telomere.

**Figure 7**

![Figure 7](image)

**Figure 7. Prevalence of CNVs across individuals and tissues**

Percent of cells harboring megabase-scale CNVs in brain and skin across four individuals. Skin cells were not sequenced from individuals 130412 and 130604.

To determine whether CNVs or their boundaries shared certain characteristics, we independently tested whether each CNV and a 1 Mb region centered at each CNV boundary (or the first 0.5 Mb of a telomeric boundary) were enriched for various genomic features compared to a random genomic region of equivalent size. Thirty-five percent of CNVs were enriched for at least one type of repetitive sequence, such as SINEs, LINEs, DNA repeat elements (i.e. DNA transposons) and noncoding RNAs (i.e. tRNA, rRNA, snRNA, scRNA, and srpRNA) (Fig. 8B). CNV boundaries were even further enriched for
repetitive sequence, with 54% of boundaries enriched for at least one of these elements. Although CNV boundaries often had repetitive elements, only 2 CNVs showed enrichment for the same type of repetitive sequence at both boundaries (noncoding RNA in both cases). Additionally, 8 boundaries (17% of all boundaries) were enriched for segmental duplications, with 7 of these boundaries occurring at telomeres. However, no CNVs were enriched for segmental duplications at both boundaries.
Figure 8. Characteristics of megabase-scale CNVs in somatic cells
A. The distribution of CNV length among CNVs identified in brain and skin cells with VS < 0.26.
B. The percent of CNVs and CNV boundaries showing significant enrichment or depletion of various genomic features.
In light of the low CNV burden—we identified 23 CNVs that in total comprise only 324 Mb (~10%) of the genome—we were surprised to find two CNVs that recurred with nearly identical coordinates in at least two cells. The first was a ~7 Mb loss on the proximal portion of chromosome 5 in two brain cells from two different individuals (Fig. 5, Table 2). The second was a ~5 Mb loss on the proximal portion of chromosome 8 in two brain cells from the same individual and one skin cell from a different individual. This region of chromosome 8 has previously been identified as a rearrangement hotspot because of an abundance of segmental duplications at chromosome coordinates 0, 1, 7, and 8 Mb that predispose this region to nonallelic homologous recombination (Bailey et al. 2002; Yu et al. 2010). A study of 1,000 individuals with developmental defects revealed megabase-scale CNVs on the proximal portion of chromosome 8 in 1% of patients (Yu et al. 2010). This region of chromosome 8 has also been identified as a peak region of deletion across multiple tumor types and contains the tumor suppressor CSMD1 (Zack et al. 2013; Ma et al. 2009; Midorikawa et al. 2009). In light of these preexisting data, it seems likely that CNVs in this region occur more frequently, and perhaps provide a selective advantage, at the somatic level.

DISCUSSION

Low coverage sequencing of single cells is emerging as a popular and powerful tool to assess genomic heterogeneity in health and disease. However, it has been unclear what types of variants can be detected and, more importantly, what the likelihood is that the detected variants are real. Through a
combination of *in silico* and *in vivo* approaches, we assessed the sensitivity and specificity of a variety of analytical approaches for CNV detection in low coverage, single cell sequencing data. We developed an approach that allows for the robust yet specific detection of CNVs exceeding 5 Mb. We applied this approach to single cell sequencing data from brain and skin to provide, to our knowledge, the first assessment of somatic CNVs in multiple tissues at genome-wide, single cell resolution. We find that around 10% of somatic cells harbor at least one megabase-scale CNV regardless of tissue of origin.

Our analysis shows that the specificity of CNV detection is extremely compromised when algorithm parameters are adjusted to detect CNVs less than 5 Mb and when cells exhibiting high variability in read depth are analyzed. In these cases, false positive CNVs can account for over half of CNVs detected. This can explain the much higher prevalence of CNVs reported for neurons in two recent single cell sequencing studies. McConnell et al. sequenced 110 neurons at \(~0.1\)X coverage and identified megabase-scale CNVs in 45 (41%) of these cells (McConnell et al. 2013). Their dataset included 14 cells with \(\text{VS} > 0.26\) and CNVs were detected using CBS with \(\alpha = 0.001\). When we reanalyzed their data using our VS cutoffs and overlapping algorithms, we identified megabase-scale CNVs in 17% of cells. Cai et al. sequenced 26 neurons at \(~0.08\)X coverage, 19 of which they analyzed, and 13 (68%) of these cells were reported to harbor megabase-scale CNVs (Cai et al. 2014). In their dataset, 16 cells had \(\text{VS} > 0.26\) and CNVs were detected using CBS with \(\alpha = 0.02\). When we analyzed the 10 cells with \(\text{VS} < 0.26\) by our approach, we identified CNVs in
only 1 neuron. Thus, the differences between our results and those reported by McConnell et al. and Cai et al. can be attributed to differences in analytic methods rather than differences in the cell populations.

We acknowledge that the majority of cells we excluded with VS > 0.26 were from brain. Thus, it is possible that the brain has a subset of cells with high CNV burden that was excluded from our dataset. However, we believe that including cells with high VS is more likely to generate artifacts than to report on true CNVs for several reasons. For one, our specificity experiments show that cells with higher VS have a greater proportion of false CNVs. Moreover, our analysis of SNPs in cells with high VS revealed that the majority of their putative deletions were false. Finally, we note that CNVs do not appear to be responsible for high VS and that much of the variability in these cells is in the form of non-integer copy number changes. It is unclear why high VSs are more common in brain than skin but could reflect biological differences in chromatin structure that affect the efficiency of whole genome amplification. Until we better understand the biological and/or technical origin of high VS we must assume that the high CNV burden in cells with high VS is an artifact.

Our approach generated a dataset of high confidence, megabase-scale somatic CNVs, leading to several conclusions about somatic copy number variation. First, we observe 10-fold more losses than gains. However, because one gain was much larger than all other CNVs, losses and gains affected a similar amount of the genome. We note the increased frequency of losses compared to gains could be secondary to our approach having better sensitivity
to losses for CNVs less than 10 Mb. Second, somatic CNVs are not distributed uniformly throughout the genome but instead tend to occur at telomeres. Many of the CNVs were enriched for repetitive sequence such as SINEs, LINEs, DNA repeat elements, and noncoding RNAs. Presumably CNVs affecting repetitive sequence are better tolerated than CNVs enriched for coding sequence.

Over 20% of germline CNVs are associated with segmental duplications, and it is believed that non-allelic homologous recombination among these segmental duplications or other repetitive sequences is a common source of germline duplications and deletions (Redon et al. 2006; Sharp et al. 2005). On the other hand, high-resolution analyses of cancer genomes point to nonhomologous end joining and alternative end joining as the primary source of somatic deletions in tumors (Yang et al. 2013). We observed enrichment for repetitive sequences and segmental duplications at the boundaries of somatic CNVs, suggesting that non-allelic homologous recombination might also underlie somatic structural variation in non-transformed cells. However, only two CNVs were enriched for the same class of repetitive sequence at both breakpoints and no CNVs were flanked by segmental duplication on both sides. With limited resolution at breakpoints and without obvious repetitive sequence flanking both boundaries of most CNVs, we are unable to implicate specific molecular mechanisms in generating somatic CNVs in healthy tissues.

We were surprised to find two recurrent CNVs, deletions on the proximal portions of chromosomes 5 and 8, in 2 and 3 cells, respectively. The deletion on chromosome 8 thus occurs in 2% of all cells analyzed and accounts for 13% of
all the CNVs we detected. That this particular CNV has been previously identified in multiple population-based copy number analyses and was never identified as a private CNV in our study further increases our confidence that this loss represents a real, recurrent CNV as opposed to a single cell sequencing artifact. Moreover, its presence at the individual level and in cancer suggests that it not only may be prone to arising because of local genome structure but may also provide a selective advantage for cells that harbor it.

We find that megabase-scale CNVs are ten times more prevalent at the somatic level compared to the organismal level. This suggests that these aberrations arise more frequently in mitotic cells and/or, more likely, that these changes are better tolerated when they occur sporadically in tissues. However, like germline CNVs, somatic CNVs show an inverse relationship between CNV size and prevalence. In this current analysis, we observe whole chromosome copy number changes in 0.8% of cells. Sub-chromosome CNVs exceeding 10 Mb are present in 1.5% of cells, and CNVs between 5 and 10 Mb are found in 3.8% of cells. Extrapolating to events beyond our detection limit of 5 Mb, we expect that more than 15% of cells harbor CNVs smaller than 5 Mb.

We present and validate an approach that allows for detection of megabase-scale CNVs with high specificity in low coverage single cell sequencing data. With this approach, we now have the power to address various questions of genomic heterogeneity in health and disease. It remains to be determined whether somatic CNVs accumulate with age and whether cells harboring these changes will undergo further transformation or contribute to
tissue dysfunction. Sequencing cells at much higher coverage, which will necessitate collaboration across multiple groups and financial sources, will enable the characterization of even smaller variants. Further technological development catered toward single cell sequencing will also help to enhance the sensitivity and specificity at any coverage level.

METHODS

Tissues, Cell Lines, and Sequencing Data

The tissue sources were described previously (Knouse et al. 2014). The following cell lines harboring known CNVs were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM10374, GM05401, GM08263, GM05875, GM06801, GM13330, GM08696, and GM03918. Data from McConnell et al. 2013, Cai et al. 2014, and Knouse et al. 2014 were downloaded from the National Center for Biotechnology Information Sequence Read Archive using accession numbers SRP030642, SRP051114, and SRP041670, respectively.

Single Cell Whole Genome Amplification and Sequencing

Single cell isolation, whole genome amplification, and sequencing were performed as previously described (Knouse et al. 2014) with the following modifications. To sequence cells from a small population, single fibroblasts from GM08696 were transferred to individual wells of a 96-well plate using a homemade microaspirator. The cells were cultured until ~100 cells were present in a single well, at which point the contents of the well were harvested by
trypsinization and individual cells were prepared for sequencing. To sequence two separate amplifications of a single cell, the solution containing a lysed and fragmented cell was split into two separate tubes of equal volume. Both tubes were then subjected to whole genome amplification using a half volume of each subsequent reagent.

**Quality Control**

Variability scores (VS) for all cells were calculated previously (Knouse et al. 2014). To recalculate VS with CNVs excluded, windows with read depth at log$_2$ ratios above the threshold for gain or loss were eliminated from the calculation.

**CNV Simulation**

Sequencing data from twelve samples with different VS were used as input. To simulate copy number loss, we randomly downsampled half of the reads in the desired CNV interval. To simulate copy number gain, all the reads in the desired CNV region were retained while one-third of the reads outside of the window were removed. Functions implemented in SAMtools (version 0.1.19) (Li et al. 2009) and BEDTools (version 2.17.0) (Quinlan and Hall 2010) were used in the region sampling. The lengths of these simulated CNVs varied from 2.5 Mb to 20 Mb and, for each length, 5 CNVs were simulated throughout the genome. This modified sequencing data was then analyzed for CNVs as described below. A CNV of defined size was considered detectable if it was identified in at least three of five cases in a single cell.
CNV Detection Using HMM

Sequence reads were trimmed to 40 nucleotides and aligned to the major chromosomes of human (hg19) using BWA (version 0.6.1) (Li and Durbin 2009) with default options. HMMcopy (version 0.1.1) (Ha et al. 2012) was used to detect CNVs by estimating copy number in 500 kb bins controlling for mappability (downloaded from UCSC Genome Browser http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/) and GC content (calculated by HMMcopy gcCounter). The parameter E was varied from default (E = 0.9999999) to E = 0.995, 0.99, 0.95 and 0.9 for testing. Log₂ cutoffs of 0.4 and -0.35 were used for gains and losses, respectively. These cutoffs were set just below the minimum log₂ ratio at which known CNVs were observed.

CNV Detection Using CBS

The mappability track (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeCrgMapabilityAlign40mer.bigWig) downloaded from UCSC Genome Browser was processed to define dynamic windows containing 500 kb uniquely mapped locations in each window as previously described (McConnell et al. 2013). The GC percentages were computed for these windows. The reads were first mapped by BWA (version 0.6.1) and uniquely mapped reads were kept. PCR duplicates were then removed using MarkDuplicates from Picard
Read-counts in the dynamic windows were summarized by BEDTools (version 2.17.0) coverageBed (Quinlan and Hall 2010). Read count in each window was normalized by the genome-wide median read count of all windows with similar GC percentage, as measured in 1% intervals, then multiplied by 2 (McConnell et al. 2013). Log₂ ratios were then used as input for DNAcopy in R (Olshen et al. 2004). The parameter $\alpha$ in the DNAcopy package was varied (0.0001, 0.001, 0.01, and 0.05) for testing. Cutoffs of 1.32 and 0.6 were used for gains and losses, respectively.

**CNV Detection Using Both Algorithms**

Sequencing data from each cell was independently processed by the two methods described above at the specified values of $E$ and $\alpha$. A CNV was called only in regions in which both algorithms identified a CNV of the same type (loss or gain). If the two algorithms identified the same CNV but the boundaries differed, the coordinates of overlap were set as the boundaries of the CNV.

**SNP Analysis**

SAMtools (version 0.1.19) mpileup and bcftools (Li et al. 2009) were used to identify variants in the BAM alignment output. VCFtools (version 0.1.8a) vcf-annotate (Danecek et al. 2011) was then used to match these variants to dbSNP build 138 (Sherry et al. 2001). For pooled samples, a mapping quality score of 30 and a read depth of 4 were required to identify heterozygous SNPs. The DP4 tags in the VCF files were used to characterize the status of known SNPs that
are not located in repetitive regions defined by RepeatMasker track downloaded from UCSC genome browser.

**Enrichment Analysis**

To test for telomere enrichment, BEDTools shuffle was used to identify 10 random regions of the genome that were the same length as each CNV in our dataset. The location of these 230 coordinates was compared to the location of the 23 CNVs. To test for enrichment of other genomic features, the following BED format annotation files were downloaded from UCSC Genome Browser: segmental duplications (Segmental Dups), repeats (RepeatMasker), CpG islands (cpgIslandExt), and genes (refGene). Fragile sites identified by a previous study (Fungtammasan et al. 2012) were extracted and lifted over to hg19. BEDTools was used to overlap the CNVs and their boundaries to these annotation files and count the number of features per million base pairs. These densities were compared with the background feature densities throughout the whole genome and z-scores for each were calculated. A CNV or boundary was considered significantly enriched or depleted for a given feature if the z-score was greater than 1.96 or less than -1.96, respectively.

**ACKNOWLEDGMENTS**

We thank the MIT BioMicro Center for sequencing. We are grateful to Kareem Carr of the Research Technology Consulting team at the Institute for Quantitative Social Science at Harvard University for statistical support. We thank Mark Daly,
Peter Sudmant, Cheng-Zhong Zhang, and Gavin Ha for their comments on this manuscript. This work was supported by the National Institutes of Health Grant GM056800 and the Kathy and Curt Marble Cancer Research Fund to A.A. and in part by the Koch Institute Support Grant P30-CA14051. A.A is also an investigator of the Howard Hughes Medical Institute and the Glenn Foundation for Biomedical Research. K.A.K. is supported by the NIGMS Training Grant T32GM007753.

REFERENCES


CHAPTER THREE

Single cell sequencing reveals low levels of aneuploidy across mammalian tissues

Reprinted with permission from the National Academy of Sciences of the United States of America


KAK led the study and performed all experiments. JW and CAW processed sequencing data and KAK, JW, and CAW analyzed the data.
Whole chromosome copy number alterations, also known as aneuploidy, are associated with adverse consequences in most cells and organisms. However, high frequencies of aneuploidy have been reported to occur naturally in the mammalian liver and brain, fuelling speculation that aneuploidy provides a selective advantage in these organs. To explore this paradox, we used single cell sequencing to obtain a genome-wide, high-resolution assessment of chromosome copy number alterations in mouse and human tissues. We find that aneuploidy occurs much less frequently in the liver and brain than previously reported and is no more prevalent in these tissues than in skin. Our results highlight the rarity of chromosome copy number alterations across mammalian tissues and argue against a positive role for aneuploidy in organ function. Cancer is therefore the only known example, in mammals, of altering karyotype for functional adaptation.
INTRODUCTION

DNA copy number alterations, deviations from the diploid DNA content, can vary in length from a few base pairs to entire genomes. When the entire genome is duplicated, cells are referred to as polyploid. Aneuploidy refers to a change in the copy number of individual chromosomes. The changes in gene copy number caused by these karyotype alterations typically change expression of the affected genes (1-3). Therefore, copy number alterations can have profound effects on cellular and organismal physiology. This is especially true for aneuploidy, where the relative dosage of many genes is altered.

In multicellular organisms, aneuploidy can be present either in all cells, termed constitutional aneuploidy, or only in select cells, called somatic aneuploidy. Constitutional aneuploidy typically has adverse consequences for the organism. In humans a third copy of chromosome 21 causes Down syndrome, and a third copy of any other autosome almost always results in embryonic or childhood lethality (4). Supporting the adverse effects of aneuploidy at the organismal level, chromosome copy number changes are also associated with impaired fitness at the cellular level. Aneuploid yeast and mammalian cells exhibit slower proliferation and proteotoxic stress (2, 3, 5, 6).

Somatic aneuploidy also has adverse consequences for the organism. In humans, mutations in \textit{BUB1B} cause the disease mosaic variegated aneuploidy (MVA) (7). \textit{BUB1B} encodes BubR1, a component of the spindle assembly checkpoint required for accurate chromosome segregation (8). Cells of MVA patients are frequently aneuploid for one or more chromosomes, and MVA
patients exhibit growth retardation, developmental defects, and childhood cancers. Mice homozygous for a hypomorphic mutation of \( BUB1B \) \( \text{BUB1B}^{H/H} \) show similar phenotypes (9). Cancer is another example of a disease characterized by aneuploidy. In this case, however, changes in gene copy number are thought to increase the fitness of tumor cells relative to untransformed cells because of selection for oncogene gains and tumor suppressor gene losses (10).

In light of the many reports describing significant adverse effects of aneuploidy on cellular and organismal fitness, it is surprising that two organs, the liver and brain, are reported to harbor high levels of aneuploidy. Hepatocytes, the primary cell type of the liver, are diploid in neonates. However, hepatocytes become polyploid during early postnatal development by initiating mitosis but failing cytokinesis (11, 12). Importantly, a recent reexamination of hepatocyte ploidy by fluorescence in situ hybridization (FISH) reported aneuploidy in over 50% of hepatocytes (13, 14). Duncan et al. suggested that polyploid hepatocytes underwent aberrant mitoses with high levels of chromosome mis-segregation, thereby generating aneuploidy among polyploid cells (13). The authors further speculated that aneuploidy endows hepatocytes with phenotypic variability and adaptability upon exposure to various noxious agents and metabolic stresses (15).

Rehen et al. were first to report elevated aneuploidy in the mammalian brain (16). Using spectral karyotyping (SKY), they reported that 33% of embryonic mouse neural progenitor cells harbored loss or gain of one or more
chromosomes. Subsequent studies using FISH indicated around 20% of adult mouse and human brain cells were aneuploid (16-20). These aneuploid cells were believed to arise from chromosome mis-segregation events in neural progenitor cells and were shown to integrate into brain circuitry (21, 22). These high levels of aneuploidy were proposed to provide the brain with its notable phenotypic diversity while simultaneously predisposing the organ to neurodegeneration (23-26).

The reports of aneuploidy in the liver and brain suggest that, in these tissues, aneuploidy may not compromise organ function but instead provide beneficial phenotypic plasticity. Indeed a positive role for aneuploidy is well documented in experimental evolution studies in microorganisms (27, 28). The reports of high levels of aneuploidy in the brain and liver thus raised the exciting possibility that these organs somehow avoid the adverse consequences of aneuploidy and instead employ these changes to their benefit, all while avoiding oncogenic transformation. Here, we use single cell sequencing to provide a genome-wide, high-resolution assessment of aneuploidy in mammalian tissues. Unlike previous reports, we find that the prevalence of aneuploidy in liver and brain is low and comparable to the frequency in skin. Our findings argue against a positive role for somatic chromosome copy number alterations in tissue function and instead reinforce their adverse consequences for cellular and organismal fitness.
RESULTS

Detecting somatic aneuploidy by single cell sequencing

The studies reporting aneuploidy in the brain and liver employed fluorescence in situ hybridization (FISH) and spectral karyotyping (SKY) to quantify aneuploidy. Both methods are prone to artifacts. In FISH, probe stretching, probe clustering, failed hybridization, or off-target hybridization can cause signals to be lost or gained inappropriately. Even a low frequency of artifacts for a single chromosome can lead to a gross overestimation of aneuploidy when extrapolated across all chromosomes. SKY can also overestimate aneuploidy as the spreading procedure can move chromosomes between nuclei. We therefore utilized single cell sequencing as an alternative method to assess somatic aneuploidy in brain and liver. We dissociated tissues from mice and humans, amplified the genomic DNA of single cells, sequenced the amplified DNA, and inferred copy number from sequencing read depth (Fig. 1A). When possible, we isolated tissues from males as monosomy of the X chromosome provides an internal control for identifying copy number alterations. We avoided permeabilization and fixation of target cells in order to minimize interference with whole genome amplification. In the liver, we isolated nuclei instead of whole cells in order to facilitate detection of copy number changes in a polyploid setting.
Figure 1. Validating detection of aneuploidy by single cell sequencing
(A) Overview of the method used to detect copy number alterations in single cells by whole genome sequencing.
(B) Segmentation plots of a euploid brain cell (left) and a trisomy 16 brain cell (right), isolated from male mouse embryos euploid and trisomic for chromosome 16, respectively. Segmentation plots show copy number of single cells relative to a euploid reference on a log2 scale. Segments above threshold for gain are colored in red, segments below threshold for loss are colored in green.
(C) Heat map of chromosome losses and gains in aneuploid brain cells and liver nuclei from BUB1B?/? mice.
(D) Segmentation plots of five aneuploid cells from a \(BUB1B^{H/H}\) mouse brain. (E) Segmentation plots of three aneuploid nuclei from a \(BUB1B^{H/H}\) mouse liver.

For each cell, on average, \(8.5 \times 10^6\) reads aligned to the reference genome, corresponding to 0.1X coverage. We quantified reads in 500 kilobase (kb) windows and used a hidden Markov model to generate segments of defined copy number relative to a euploid reference (29). We did not analyze the Y chromosome for copy number alterations because of its low amount of unique sequence.

Approximately 10% of sequenced cells exhibited aberrantly high intrachromosomal sequence depth variability (Fig. 2A, top two panels) as might arise from incomplete fragmentation of genomic DNA or biased amplification of fragments during whole genome amplification. To identify these cells and exclude them from the analysis, we measured the standard deviations of corrected read copies (log2 based) within sliding windows (30 adjacent 500 kb bins) across a single chromosome, averaged the standard deviations for each chromosome, computed the mean average standard deviation across the three most variable autosomes, and excluded cells for which this variability score (VS) exceeded 0.34 (see Methods, Fig. 2B-C). In total, we excluded 10.1% of sequenced cells (Fig. 2D). This metric does not bias against aneuploidy, because variability is measured within each chromosome. Consistent with this, the average VS of all aneuploid cells we identified was equivalent to the average VS of all euploid cells (Fig. 2A, bottom two panels, 2C).
Figure 2

A

Variability Score | Excluded?
---|---
0.45 | Yes
0.36 | Yes
0.25 | No
0.24 | No

B

All Sequenced Cells

C

Analyzed Cells Only

D

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number Sequenced</th>
<th>Number Excluded</th>
<th>Percent Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse Keratinocytes</td>
<td>38</td>
<td>1</td>
<td>2.6%</td>
</tr>
<tr>
<td>Human Keratinocytes</td>
<td>55</td>
<td>2</td>
<td>3.6%</td>
</tr>
<tr>
<td>Mouse BubR1&lt;sup&gt;+/−&lt;/sup&gt; Hepatocyte Nuclei</td>
<td>16</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Mouse Hepatocyte Nuclei</td>
<td>68</td>
<td>2</td>
<td>2.9%</td>
</tr>
<tr>
<td>Human Hepatocyte Nuclei</td>
<td>103</td>
<td>3</td>
<td>2.9%</td>
</tr>
<tr>
<td>Mouse Neural Progenitor Cells</td>
<td>47</td>
<td>11</td>
<td>23.4%</td>
</tr>
<tr>
<td>Mouse Euploid+Ts16 Brain Cells</td>
<td>11</td>
<td>2</td>
<td>18.1%</td>
</tr>
<tr>
<td>Mouse BubR1&lt;sup&gt;+/−&lt;/sup&gt; Brain Cells</td>
<td>22</td>
<td>1</td>
<td>4.5%</td>
</tr>
<tr>
<td>Mouse Brain Cells</td>
<td>51</td>
<td>8</td>
<td>15.7%</td>
</tr>
<tr>
<td>Mouse NeuN Neurons</td>
<td>18</td>
<td>9</td>
<td>50%</td>
</tr>
<tr>
<td>Mouse DRD2 Neurons</td>
<td>10</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Human Brain Cells</td>
<td>105</td>
<td>16</td>
<td>15.2%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>544</strong></td>
<td><strong>55</strong></td>
<td><strong>10.1%</strong></td>
</tr>
</tbody>
</table>
We validated several aspects of our approach to ensure that this method reliably detects somatic aneuploidy. First, we confirmed that loss or gain of a chromosome caused a change in relative read depth that exceeded background fluctuations in read depth. We sequenced single brain cells from trisomy 16 mouse embryos (3) and were indeed able to detect an additional copy of chromosome 16 in all cells from the trisomy 16 embryo (Fig. 1B, right panel). We next determined whether the method was able to detect aneuploidies of unknown composition. We sequenced single brain cells from an adult male $BUB1B^{H/H}$ mouse, as the $BUB1B$ mutation is reported to cause aneuploidy in one-third of cells (9). Indeed, 8 of the 21 (38.1%, 95% CI 18.1-61.6%) brain cells we analyzed were aneuploid for one or more chromosomes (Fig. 1C-D).

To determine whether we were able to detect aneuploidy in polyploid cells, we sequenced hepatocyte nuclei from an adult male $BUB1B^{H/H}$ mouse. Here, 3 of the 16 (18.8%, 95% CI 4-45.6%) hepatocyte nuclei we analyzed harbored one or more chromosomes with read depths indicating loss or gain in a tetraploid
nucleus (Fig. 1C, E). We conclude that our method can detect single as well as multiple chromosome gains or losses in diploid and polyploid cells.

Lastly, we tested whether our approach selected against aneuploid cells because of differences in viability during tissue dissociation or operator bias during microaspiration. We mixed equal volumes of trisomy 16 and euploid brain, dissociated the tissue together, and sequenced single cells from this mixture. We found that five of seven cells (71.4%, 95% CI 29-96.3%) were trisomic for chromosome 16 indicating that our procedure does not discriminate against aneuploid cells.

Together, these experiments show that single cell sequencing can detect simple and complex aneuploidies in an unbiased manner. Furthermore, by compiling the relative copy numbers (log₂ ratios) for all aneuploid chromosomes and X chromosome monosomies, we were able to establish cutoffs for detecting copy number variants in all subsequently sequenced diploid and polyploid cells (see Methods).

**Prevalence of aneuploidy in skin**

There are no reports of elevated aneuploidy in skin. Therefore, we sequenced single cells from skin to establish a baseline level of somatic aneuploidy. We isolated keratinocytes by dissociating epidermis and selecting uncornified cells. In the mouse we found one aneuploid cell (trisomy 12) among 37 cells analyzed, indicating that 2.7% of mouse keratinocytes are aneuploid (95% CI 0.1-14.2%, Table 1). To examine aneuploidy in human skin, we isolated keratinocytes from epidermis obtained upon autopsy of a 48 year-old female and
a 68 year-old male. Of the 53 keratinocytes analyzed, none were aneuploid (Table 1). We conclude that aneuploidy in human skin is low (95% CI 0-6.7%, Table 1).

**Table 1. Prevalence of Aneuploidy in Somatic Cells**

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Total Cells Analyzed</th>
<th>Total Aneuploid Cells</th>
<th>Description of Aneuploidy</th>
<th>Prevalence of Aneuploidy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse Keratinocytes</td>
<td>37</td>
<td>1</td>
<td>Trisomy 12</td>
<td>2.7% (0.1-14.2%)</td>
</tr>
<tr>
<td>Human Keratinocytes (48F, 68M)*</td>
<td>53 (29, 24)</td>
<td>0</td>
<td></td>
<td>0% (0-6.7%)</td>
</tr>
<tr>
<td>Mouse Neural Progenitor Cells</td>
<td>36</td>
<td>0</td>
<td></td>
<td>0% (0-9.7%)</td>
</tr>
<tr>
<td>Mouse Brain Cells</td>
<td>43</td>
<td>1</td>
<td>Trisomy 15</td>
<td>2.3% (0.1-12.3%)</td>
</tr>
<tr>
<td>Mouse Neurons</td>
<td>19</td>
<td>0</td>
<td></td>
<td>0% (0-17.6%)</td>
</tr>
<tr>
<td>Human Brain Cells (70M, 52M, 48F, 68M)</td>
<td>89 (20, 22, 22, 25)</td>
<td>2 (2, 0, 0, 0, 0)</td>
<td>Monosomy 22 Trisomy 18 in tetraploid</td>
<td>2.2% (0.3-7.9%)</td>
</tr>
<tr>
<td>Mouse Hepatocyte Nuclei</td>
<td>66</td>
<td>0</td>
<td></td>
<td>0% (0-5.4%)</td>
</tr>
<tr>
<td>Human Hepatocyte Nuclei (46M, 51M)</td>
<td>100 (39, 61)</td>
<td>4 (1, 3)</td>
<td>Uninterpretable cell Pentasomy 5, 7 in tetraploid Pentasomy 7 in tetraploid Pentasomy 15 in tetraploid</td>
<td>4% (1.1-9.9%)</td>
</tr>
</tbody>
</table>

**Prevalence of aneuploidy in brain**

In the mammalian brain, aneuploidy is reported to be highest in neural progenitor cells with a prevalence of 33% (16). To investigate aneuploidy in
embryonic neural progenitor cells, we isolated brains from mouse embryos expressing the neural progenitor marker nestin-GFP. Of the 36 cells analyzed, none were aneuploid (Table 1). To assess the degree of aneuploidy in the adult mouse brain, we dissociated grey matter from cerebral cortex and enriched for neurons and glia by density gradient sedimentation. Single cell gene expression analysis revealed that the cell suspension obtained by this method contained 75% neurons, 12.5% glia, and 12.5% unknown cell types (Fig. 3). Of the 43 cells analyzed, we identified one aneuploid cell (trisomy 15). Thus, only 2.3% of mouse brain cells are aneuploid (95% CI 0.1-12.3%, Table 1).

Figure 3

![Expression of NeuN, GFAP, and MBP](image)

**Figure 3. Identity of brain cell preparations and segmentation plots of aneuploid brain cells**

Expression of the neuronal marker NeuN, the astrocyte marker GFAP, and the oligodendrocyte marker MBP in single mouse brain cells determined by single cell qRT-PCR.

To specifically examine neurons, we immunostained adult mouse brain cells with an antibody against the neuronal nuclear marker NeuN and enriched for NeuN-positive cells by FACS. Of 18 cells sequenced, we had to exclude 9
cells because of high variability in read depth between adjacent genomic windows (Fig. 2B-D). Of the 9 cells analyzed, none were aneuploid. We presume that the high VS in many of the neurons isolated by this method was due to permeabilization, fixation, and/or immunostaining interfering with whole genome amplification. As an alternate approach, we purified DRD2-expressing medium spiny neurons of the basal ganglion from an adult mouse carrying a DRD2-GFP transgene. We sequenced 10 cells, all of which could be analyzed, and none of which were aneuploid. We conclude that the sequence variability observed in the neurons purified by NeuN immunostaining was due to the immunostaining procedure interfering with sample preparation. Importantly, our analysis of 19 neurons did not reveal a single aneuploid cell (95% CI 0-17.6%, Table 1). This result is not significantly different from that for the mixed brain cell population. In total, the prevalence of aneuploidy in the mouse brain is 1% (95% CI 0-5.6%, Fig. 4A).

To extend these studies to humans, we dissociated grey matter of frontal lobe obtained during autopsy of four individuals: a 52 year-old male, 68 year-old male, 70 year-old male, and a 48 year-old female. None of these individuals had a history of neurologic disease or evidence of brain pathology upon autopsy. Of the 89 cells analyzed, all but two cells were euploid (Fig. 4C). Of the aneuploid cells, one harbored a monosomy for chromosome 22 (Fig. 4D, top panel). The other cell had a read depth on chromosome 18 consistent with trisomy in a tetraploid cell (Fig. 4D, bottom panel). Because cell suspensions were extremely dilute, we believe that this indeed represents a tetraploid cell with chromosome
loss rather than two cells isolated together, one of which was aneuploid. We conclude that the prevalence of aneuploidy in the adult human brain is 2.2% (95% CI 0.3-7.9%, Fig. 2B, Table S1), significantly less than the over 20% aneuploidy reported by prior studies (16-20).
Figure 4. Prevalence of aneuploidy in brain
(A, B) The prevalence of euploid and aneuploid cells in mouse (A) and human (B) brain.
(C) Segmentation plots of euploid human brain cells.
(D) Segmentation plot of an aneuploid human brain cell harboring a monosomy for chromosome 22 (top panel) and a presumably tetraploid human brain cell harboring a loss of chromosome 18 (bottom panel).
Prevalence of aneuploidy in liver

To determine the degree of aneuploidy in the liver, we first assessed the ploidy distribution in this organ using nuclear size measurements coupled with dual-color, single-chromosome FISH. In liver, but not other tissues, we observed many binucleate cells and many cells with more than two copies of a single chromosome. The elevated copy number states were associated with increases in nuclear diameter (Fig. 5A). Because polyploidy increases nuclear diameter, the increases in chromosome copy number in hepatocytes likely reflect polyploidy. We thus classified cells having two, four, eight, or sixteen copies of a single chromosome as being diploid, tetraploid, octaploid, or hexadecaploid, respectively. In mouse liver, the majority of hepatocytes were polyploid, with tetraploid cells occurring most frequently (Fig. 5B). Approximately half of tetraploid cells were binucleate, whereas the majority of octaploid and hexadecaploid cells were binucleate. In human liver, ploidy ranged from diploid to octaploid, with diploid and tetraploid cells occurring most often and with similar frequency (Fig. 5B).

Aneuploidy is more difficult to detect in polyploid cells because the adjusted read depth on the aneuploid chromosome is smaller. As described above, we were able to detect aneuploidy in tetraploid hepatocyte nuclei from $BUB1B^{H/H}$ mice (Fig. 1E). Importantly, in all presumably diploid cells we analyzed from other tissues, only one brain cell harbored a chromosome with an adjusted read depth consistent with a single chromosome gain or loss in a tetraploid cell. Thus, the chromosome changes we observed in $BUB1B^{H/H}$ hepatocyte nuclei likely represent actual aneuploidy rather than background fluctuations in read
depth. Detection of aneuploidy in octaploid nuclei is more challenging because the adjusted read depth from a single chromosome loss or gain is further reduced. Indeed, 14% of all diploid cells analyzed harbored one or more chromosomes with an average read depth that would be consistent with a chromosome gain or loss in an octaploid cell. Thus, chromosome gains or losses in an octaploid cell cannot be distinguished from background fluctuations in read depth. We therefore did not consider fluctuations of this magnitude as aneuploidy. This did not significantly affect the outcome of our analysis because octaploid hepatocyte nuclei are rare (Fig. 5B, 6).

Having established cut-offs for detecting aneuploidy in a polyploidy setting, we isolated nuclei from mouse liver. To enrich for hepatocytes, we preferentially picked nuclei with diameters consistent with tetraploidy (Fig. 6). Of the 66 cells analyzed, we did not identify a single aneuploid nucleus (Fig. 5C). We next sequenced commercially available human hepatocytes prepared from a 46 year-old male and a 51 year-old male. Neither of these individuals had liver disease, however the second individual had an extensive drug and alcohol history. The purity of these hepatocyte preparations exceeds 70%. Of the 39 hepatocytes analyzed from the first donor, all cells except one were euploid (Fig. 5E). Though we could not assign a karyotype to the other cell due to extensive fluctuation of segments, it met our quality control criteria and was considered aneuploid (Fig. 5F, first panel). Of the 61 nuclei analyzed from the second donor, all but three cells were euploid (Fig. 5E). All three aneuploid nuclei were tetraploid with gains of one or two chromosomes (Fig. 5F, second through fourth
panels). In total, the prevalence of aneuploidy in human hepatocytes is 4% (95% CI 1.1-9.9%, Fig. 5D). Thus, single cell sequencing indicates that less than 5% of hepatocytes are aneuploid (Table 1).

Figure 5

A

B

C

D

E

F
Figure 5. Prevalence of aneuploidy in liver
(A) Increase in chromosome copy number correlates with an increase in nuclear diameter in mouse and human hepatocytes. Note that only a single octaploid human hepatocyte nucleus was identified.
(B) Prevalence of cells of different ploidy for a single chromosome in mouse and human skin (keratinocytes), liver (hepatocytes), and brain (neurons and glia). Polyploid hepatocytes are indicated as being mononucleate (mono) or binucleate (bi). For example, tetraploid (bi) describes a cell with two nuclei, each of which contains two copies of a single chromosome. n = 200 cells per tissue.
(C, D) The prevalence of euploid and aneuploid nuclei in mouse (C) and human (D) liver.
(E) Segmentation plots of euploid human hepatocyte nuclei.
(F) Segmentation plots of a human hepatocyte nucleus for which a karyotype could not be determined (first panel), and tetraploid human hepatocyte nuclei harboring pentasomy for chromosomes 5 and 7 (second panel), pentasomy 7 (third panel) and pentasomy 15 (fourth panel).

Figure 6

Figure 6. Ploidy of hepatocyte nuclei preparations and segmentation plots of aneuploid hepatocyte nuclei
Ploidy of mouse hepatocyte nuclei inferred by FISH (chromosome 16 copy number) in tissue sections and nuclear diameter in dissociated nuclei.
DISCUSSION

Here we provide the first assessment of chromosome copy number alterations in multiple tissues at genome-wide, single-cell resolution. Our sequencing results suggest that the prevalence of aneuploid cells in the mammalian brain and liver is less than 5% (Table 1). This observation is in agreement with a recent single neuron sequencing study which, in spite of identifying a high frequency of sub-chromosome copy number variants, did not observe high levels of whole-chromosome aneuploidy in neurons (30). Prior reports used SKY and FISH to assess the prevalence of aneuploidy and found it to exceed 50% and 20% in the liver and brain, respectively (13, 14, 16-20). We attribute this difference to drawbacks associated with using FISH to detect somatic aneuploidy. Indeed, using dual-color, single-chromosome FISH we too observed low levels of aneuploidy for a single chromosome that would, when extrapolated across all chromosomes, indicate that over 40% of hepatocytes and 10% of brain cells are aneuploid (Fig. 5B). However, we also observed similar rates of aneuploidy in skin. That we had to exclude up to one-third of cells from analysis because of discrepancies in signal number between two probes targeting the same chromosome illustrates the limitations of FISH. The rare cases of signal gain or loss for both probes could just as likely represent two simultaneous hybridization artifacts. FISH artifacts that are interpreted as aneuploidy even when occurring at low frequency can lead to gross overestimation of aneuploidy when used to infer the prevalence of aneuploidy
across the genome. We conclude that single cell sequencing is a superior method for characterizing somatic karyotypic changes.

Whether the rare cells harboring copy number alterations that we did identify in normal tissues represent incipient transformation remains to be determined. Furthermore, it is important to note that we only analyzed skin, brain, and liver tissue that had no evidence of disease. It is possible that aneuploidy is elevated in disease states other than cancer. Indeed, it is interesting to note that proteotoxic stress is a hallmark of both aneuploidy and neurodegenerative disease (6, 31).

**The implications of low aneuploidy for hepatocyte proliferation**

The low prevalence of aneuploidy in normal tissues means that either the events producing such aberrations occur rarely or the cells harboring such changes are eliminated. Both possibilities are likely to occur. The wide array of detrimental phenotypes associated with constitutional aneuploidy *in vitro* and *in vivo* suggests that aneuploid somatic cells may indeed be selected against when arising amid a euploid population. On the other hand, our sequencing of $BUB1B^{H/H}$ mice indicates that cells can survive in spite of multiple copy number alterations, though their survival in these mice may be due to the presence of fewer euploid cells and thus reduced selective pressure (9).

The low level of aneuploidy has particular implications for our understanding of liver biology. Hepatocytes are not the only example of polyploidy in mammals. Trophoblast giant cells, which comprise the placenta,
and megakaryocytes, the precursors of platelets, are also polyploid. However, in these two cases, and indeed in cases in other organisms, polyploidy is associated with a terminally differentiated state and cell cycle arrest (32). Hepatocytes are an exception to this rule. Upon liver damage, hepatocytes reenter the cell cycle and proliferate extensively (33). This is a nontrivial task as polyploid cells, in addition to having multiple copies of their genome, have multiple centrosomes. When a cell with multiple centrosomes enters mitosis, a multipolar spindle forms which, if uncorrected, randomly segregates chromosomes to produce highly aneuploid cells. Studies of cancer cells have shown that multiple centrosomes can cluster to form bipolar spindles (34). However, the process of centrosome clustering produces many merotelic kinetochore-microtubule attachments, where a kinetochore attaches to microtubules emanating from both spindle poles. Chromosomes with merotelic attachments are frequently mis-segregated (35). It is difficult to imagine that the liver culls the appreciable fraction of aneuploid hepatocytes expected to arise from these faulty mitoses. This implies that hepatocytes have either enhanced mechanisms for detecting and correcting merotelic attachments or employ a unique mechanism for segregating chromosomes in the presence of multiple centrosomes. Determining how hepatocytes prevent or correct merotelic attachments will be an important future question.

The implications of low levels of aneuploidy for neural diversity

The brain is notable for its remarkable diversity, both at the cellular and functional level. There are over 10,000 different types of neurons, and even
neurons of the same type can differ in their expression of synapse components and in their thresholds for excitation and firing (23). Diversity at the cellular level presumably translates into diversity at the functional level. There is remarkable inter-individual heterogeneity in cognition, intellect, and behavior. For example, even isogenic mice exhibit remarkable diversity in stress responses (36). The report of high levels of aneuploidy in the brain led to extensive speculation that random aneuploidies endowed the brain with cellular and thus functional diversity (23). However, our study shows that somatic aneuploidy is much less common than previously reported. Given the low prevalence of aneuploidy in normal tissues, it is difficult to imagine that such changes play a positive role in organ function. We infer that normal mammalian tissues do not employ aneuploidy as a means of phenotypic diversification or adaptation.

METHODS

Tissue sources

Frozen human liver was obtained from the Brigham and Women’s Hospital Tissue and Blood Repository. Hepatocytes isolated from human cadaveric donors without liver disease were purchased from Life Technologies. Frontal lobe and epidermis were isolated from humans without neurologic disease or brain hypoxia upon autopsy at Massachusetts General Hospital through the Neuropathology Core of the Massachusetts Alzheimer Disease Research Center. All human procedures were approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects.
Murine liver, cerebral cortex, and epidermis were isolated from C57BL/6J mice purchased from the Jackson Laboratory. \( BUB1B^{H/H} \) mice were obtained from Jan van Deursen, DRD2-GFP mice from David Housman, and nestin-GFP mice from David Scadden. Trisomy 16 embryos were generated from \( Rb(6.16)24Lub \) and \( Rb(16.17)7Bnr \) mice purchased from the Jackson Laboratory. All animal procedures were approved by the Massachusetts Institute of Technology Committee on Animal Care.

**Immunostaining and fluorescence in situ hybridization (FISH)**

Tissues were fixed in 4% paraformaldehyde in PBS for 16-24 hours. Fixed tissues were washed with PBS, cryoprotected with 30% sucrose, and frozen in O. C. T. Compound (Tissue-Tek). Slides with 10 or 30 \( \mu \)m thick sections were prepared using a cryostat and stored at -80°C until use. Slides were thawed, aged at room temperature for 16-24 hours, rehydrated in PBS for 15 minutes and boiled in sodium citrate buffer (10 mM tri-sodium citrate dihydrate, 0.05% Tween 20, pH 6.0) for 20 minutes. Slides were washed in PBS, dried briefly, and sections outlined with a hydrophobic pen. Sections were incubated with PBS containing 0.5% Triton X-100 (0.5% PBST) for 5 minutes followed by incubation in blocking solution (4% BSA, 10% goat serum, 10% donkey serum in 0.5% PBST) for 1 hour at room temperature. Sections were incubated with a rabbit pan-cadherin antibody (Abcam) diluted 1:500 in block at 4°C for 16-24 hours. Sections were washed three times with 0.5% PBST for 5 minutes each. Sections were incubated with an AlexaFluor 647 goat anti-rabbit antibody (Life
Technologies) diluted 1:1000 in blocking solution at room temperature for 1 hour. Sections were washed three times with 0.5% PBST for 5 minutes each. Sections were fixed with 2% paraformaldehyde in PBS for 10 minutes and washed three times with PBS for 5 minutes each. Hydrophobic ink was removed and slides were incubated in 2X SSC for 5 minutes followed by incubation in 50% formamide, 2X SSC for 2 hours. Fluorescently labeled probes targeting different regions of mouse chromosome 16 (RP23-354F11 and RP23-18M23) and human chromosome 7 (RP11-243E12 and RP11-377B19) (Empire Genomics) were diluted in hybridization buffer as per manufacturer's instructions and applied to sections. Sections were sealed with a coverslip and rubber cement and incubated in the dark at 45°C for 2 hours to allow probes to infiltrate section. This was followed by incubation at 85°C for 5 minutes to denature the DNA. Hybridization was performed in the dark at 37°C for 48 hours. Slides were then washed with 0.4X SSC containing 0.3% NP-40 for 2 minutes at 73°C followed by 2X SSC containing 0.1% NP-40 for 1 minute at room temperature. Slides were incubated in 0.05 μg/mL DAPI in 2X SSC for 30 minutes and mounted with ProLong Gold Antifade Reagent (Life Technologies). Images were acquired on a spinning disk confocal microscope (PerkinElmer) and analyzed using the Volocity software package (PerkinElmer).

**Tissue dissociation**

To isolate keratinocytes from human and mouse skin, skin was floated in 0.05% thermolysin in HEPES-buffered saline for 1-2 hours at 37°C. The epidermis was
liberated from dermis using forceps, dissociated with 0.05% trypsin in 0.05% EDTA in PBS at 37°C for 30 minutes, combined with an equal volume of DMEM with 10% FBS, passed through a 40μm filter, centrifuged at 100 g for 5 minutes, and resuspended in DMEM with 10% FBS.

To isolate neural progenitor cells from embryonic nestin-GFP mice and medium spiny neurons from adult DRD2-GFP mice, brains were dissociated as previously described (37) and GFP-positive cells isolated by flow cytometry. To isolate brain cells from adult human and mouse cortex, tissue pieces were processed as previously described (38). Fraction 3 was isolated to enrich for neurons.

To isolate nuclei from human hepatocytes, hepatocytes were washed with cold PBS, incubated in 0.2X PBS on ice for 10 minutes, lysed with a dounce homogenizer for 15 strokes, centrifuged at 1,000 g at 4°C for 10 minutes, resuspended in sucrose buffer (250 mM sucrose, 5 mM MgCl₂, 10 mM Tris-Cl pH 7.4), and centrifuged and resuspended again. To isolate nuclei from mouse liver, liver lobes were dissected in PBS on ice, pressed through a 40 μM filter into sucrose buffer, centrifuged at 600 g at 4°C for 10 minutes, resuspended in sucrose buffer, and centrifuged and resuspended again.

**Single cell qRT-PCR**

A total of 2.5 x 10³ dissociated cells were added to 20 mL of media or sucrose buffer containing 5% FBS in a 15 cm petri dish. Single cells were isolated using
a homemade microaspirator and prepared for qRT-PCR using the Single Cell-to-
C\textsubscript{T} Kit (Life Technologies). Specific transcripts were quantified using TaqMan
Gene Expression Assays (Life Technologies) targeting mouse Rbfox3 (NeuN)
(Mm01248771\_m1), mouse GFAP (Mm01253033\_m1), mouse MBP
(Mm01266402\_m1), and mouse GAPDH (Mm99999915\_g1)

**Single cell whole genome amplification**

A total of $2.5 \times 10^3$ dissociated cells or nuclei were added to 20mL of media or
sucrose buffer containing 5% FBS in a 15 cm petri dish. Single cells and nuclei
were isolated using a homemade microaspirator and transferred to 8 \( \mu \)L water in
a 96-well plate. The microaspirator needle was cleaned with 10% bleach
followed by water after transferring each cell. After up to 32 cells or nuclei had
been isolated, cells were lysed and genomic DNA amplified using the
GenomePlex Single Cell Whole Genome Amplification Kit (Sigma). All reagents
were added using aerosol-resistant pipette tips in a laminar flow hood.
Following amplification, 4\( \mu \)L of each sample were analyzed by agarose gel
electrophoresis. Samples producing a smear ranging from 100 to 1,000 base
pairs were sequenced.

**Single cell sequencing**

Products of whole genome amplification were purified using the MinElute
Cleanup kit (Qiagen). 50 ng of purified DNA were prepared and barcoded using
Nextera reagents (Illumina) and tagmented material was PCR amplified for 7
cycles. Libraries were quantified using an AATI Fragment Analyzer and qPCR before pooling. Between 15 and 24 cells were combined for sequencing in a single lane. Libraries were sequenced (40 or more nucleotide reads) on an Illumina HiSeq2000. Reads were demultiplexed using custom scripts allowing single mismatches within the reference barcode.

**Copy number analysis**

Sequence reads were trimmed to 40 nucleotides and aligned to genomes of human (hg19) or mouse (mm9) using BWA (0.6.1) with default options (39). HMMcopy (0.1.1) was used to detect CNVs by estimating copy number in 500 kb bins controlling for mappability (downloaded from UCSC Genome Bioinformatics (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/ or http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeMapability/) and GC content (calculated by HMMcopy gcCounter) (29).

Standard deviations of the corrected read copies (log2 based) from HMMcopy were computed within sliding windows (30 adjacent 500 kb bins) for all chromosomes and the average was calculated for each chromosome. The average standard deviations of the three autosomes with highest variability were averaged to generate a variability score (VS). The distribution of VS for all cells had a positive skew, with a range of 0.14 to 2.01, a mean of 0.24, and a standard deviation of 0.16. Cells with a VS exceeding 0.34 were excluded from analysis. This cutoff is 1.25 standard deviations from the minimum VS and resulted in
exclusion of 55 of 544 cells. For autosomes in brain and skin cells, a log₂ ratio exceeding 0.44 was considered a chromosome gain and a log₂ ratio below -0.57 was considered a chromosome loss. For autosomes in hepatocyte nuclei, a log₂ ratio exceeding 0.25 was considered a chromosome gain and a log₂ ratio below -0.3 a loss. For the X chromosome in males (normal log₂ ratio < -0.57), a log₂ ratio greater than -0.23 was considered gain.

ACKNOWLEDGMENTS

We thank the MIT BioMicro Center for sequencing, the Koch Institute Swanson Biotechnology Center for flow cytometry and histology, and the Whitehead Institute Keck Microscopy Facility for imaging. We are grateful to Jan van Deursen, David Scadden and David Housman for providing mouse strains, Matthew Frosch, Daniel Mordes, and Inma Cobos Sillero for providing autopsy samples, and Zachary Crook for technical assistance. We thank Mark Daly for helpful discussions and David Housman, Frank Solomon, Mark Daly and members of the Amon laboratory for their critical reading of this manuscript. This work was supported by an NIH Physical Sciences Oncology Center (PSOC) grant (5-U54-CA143874), a senior scholar award from the Ellison Medical Foundation, the Koch Institute Support Grant (P30-CA14051) from the National Cancer Institute, and by grants from the Howard Hughes Medical Institute and the Kathy and Curt Marble Cancer Research Fund. The Neuropathology Core of the Massachusetts Alzheimer Disease Research Center is supported by an NIH
grant (P50 AG005134). KK is supported by an NIGMS training grant (T32GM007753).

REFERENCES


CHAPTER FOUR

Cell non-autonomous regulation of chromosome segregation

This chapter is not published as of September 2016

Kristin A. Knouse, Kristina E. Lopez, and Angelika Amon

KAK led the study and performed all experiments with assistance from KEL.
KAK analyzed the data.
ABSTRACT

Chromosome segregation has been assumed to be a cell autonomous process. We tested this assumption by comparing the fidelity of chromosome segregation in epithelial cells proliferating in vivo and in vitro. We observed that epithelial cells segregate chromosomes more accurately in vivo compared to in vitro. This difference arises from the enhanced correction of aberrant microtubule-kinetochore attachments in the presence of tissue architecture. We conclude that chromosome segregation in epithelial cells is in fact cell non-autonomous. This could explain both chromosome instability in tumorigenesis as well as the tendency for epithelial cells to expand poorly ex vivo.
INTRODUCTION

Decades of study of chromosome segregation have led to the assumption that it is a cell autonomous process. However, most of these studies have been based on cells proliferating in isolation in vitro, hindering the accumulation of any evidence to the contrary. Indeed, some observations in vivo indicate that chromosome segregation might be influenced by external factors. While in vitro studies indicate that polyploid cells have high levels of chromosome missegregation, the liver, a naturally polyploid organ, has minimal aneuploidy (Ganem et al. 2009; Silkworth et al. 2009; Knouse et al. 2014). Additionally, cancer cells often have high levels of chromosome instability, yet they rarely have mutations in core regulators of chromosome segregation (Tighe et al. 2001; Haruki et al. 2001; Gascoigne and Taylor 2008). When epithelial cells divide in vivo they do so in the context of tissue architecture, with distinct forms of cell-cell adhesion, cell-matrix adhesion, and cell polarity. In many cases, these features are maintained during mitosis in order to preserve tissue structure and function. As chromosome segregation is a mechanical process, it seems plausible that it would be influenced by elements at the cell periphery.

Critical events in chromosome segregation include formation of a bipolar spindle, capture of kinetochores by microtubules, correction of aberrant microtubule-kinetochore attachments, and segregation of sister chromatids into separate daughter cells (Tanaka 2010; Musacchio 2015). Correction of improper attachments of chromosomes to the spindle is achieved by the complementary activities of Aurora B and the spindle assembly checkpoint. Errors in this
process lead to chromosome missegregation. Evidence of chromosome missegregation includes chromosomes that lag behind the two separating masses of chromosomes in anaphase. These lagging chromosomes can become trapped in ingressing cleavage furrow and often form a small, satellite nucleus termed a micronucleus, both of which can lead to extensive DNA damage and structural rearrangements (Janssen et al. 2011; Crasta et al. 2012; Zhang et al. 2015). Lagging chromosomes also have a tendency to end up in the incorrect daughter cell, leading to aneuploidy (Cimini et al. 2004; Thompson and Compton 2011).

RESULTS

Tissue architecture facilitates chromosome stability in epithelia

To test if chromosome segregation is indeed cell autonomous, we analyzed chromosome segregation in various epithelial cells both in their native tissue (*in vivo*) and after dissociation and culture as monolayers (*in vitro*) (Figure 1a). We specifically quantified lagging chromosomes in anaphase. We never identified lagging chromosomes *in vivo* (Figure 1b). However, when these same cells were isolated from tissues and expanded *in vitro*, we observed a significant increase in lagging chromosomes within the first 48 hours of culture (Figure 1b). These lagging chromosomes harbored kinetochores as judged by the presence of the kinetochore protein CENP-C, indicating that they are the product of chromosome segregation errors, as opposed to DNA bridges or acentric chromosomes which would also manifest during anaphase but secondary to DNA damage (Figure 1c). One established mechanism for increased lagging
chromosome is the presence of supernumerary centrosomes (Ganem et al. 2009; Silkworth et al. 2009). However, this is unlikely to explain the chromosome segregation defects we observe as we did not observe a dramatic increase in the presence of supernumerary centrosomes (Y-tubulin foci) in cells expanded in vitro (Figure 2a).

The in vitro environment differs from the in vivo environment in several ways. Upon expansion in vitro, cells lose their characteristic cell-cell and cell-matrix adhesions, are exposed to a hyperoxic atmosphere, and cultured in nonphysiological concentrations of growth factors. In theory, any of these changes could affect chromosome segregation. To test the possibility that loss of tissue architecture was responsible for the chromosome segregation defects, we seeded mammary epithelial cells in Matrigel as this leads to the formation of spheroids that closely resemble mammary gland acini (Jechlinger et al. 2009). After 48 hours in Matrigel, cells had formed small clusters that had neither a central lumen nor obvious apicobasal polarity (henceforth termed immature spheroids) (Figure 1d, left image, Figure 2c). However, by 96 hours in Matrigel the cells had developed into spheroids with a single cell layer surrounding a central lumen (henceforth termed mature spheroids). These cells harbored apicobasal polarity as evidenced by the basal localization of the cell-matrix adhesion protein α6 integrin and the apical ribbon of the tight junction protein ZO-1 (Figure 1d, right image, Figure 2c). Notably, while immature spheroids still harbored similar levels of chromosome segregation defects as monolayers, these defects were almost entirely eliminated in mature spheroids (Figure 1e).
Importantly, when we disrupted the architecture and cell polarity of mature spheroids by depleting β1 integrin, chromosome segregation defects increased to the levels observed in monolayers and immature spheroids (Akhtar and Streuli 2012) (Figure 1e, Figure 2b,c). These defects in immature spheroids and spheroids lacking β1 integrin were not secondary to supernumerary centrosomes (Figure 2d). Together, these data demonstrate that tissue architecture facilitates chromosome segregation in epithelial cells and that disruption of this architecture can lead to chromosome instability. Moreover, because immature and mature spheroids are identical in all respects except for the duration of culture in Matrigel and the presence of tissue architecture, they provide an ideal system for further mechanistic analysis of the relationship between tissue architecture and chromosome segregation.
Figure 1. Epithelial cells depend on tissue architecture for chromosome stability. 

**a.** Images of proliferative epithelial cells from mammary gland, skin, and neonatal liver *in vivo* (tissue sections) and *in vitro* (monolayers) immunostained for epithelial markers (keratin 8 for mammary epithelial cells, keratin 14 for keratinocytes, and pan-cadherin for hepatocytes), αtubulin, and Ytubulin. Scale bars, 10 μm. 

**b.** Prevalence of lagging chromosomes in anaphases of epithelial cells from mammary gland, skin, and neonatal liver *in vivo* (tissue sections) and *in vitro* (monolayers). Error bars indicate SD. p = 0.02 (mammary gland *in vivo* versus *in vitro*), 0.03 (skin *in vivo* versus *in vitro*), 0.03 (neonatal liver *in vivo* versus *in vitro*) by one-tailed Fisher’s exact test. n > 75 anaphases per condition. 

**c.** Images of lagging chromosomes in keratinocytes proliferating *in vitro* immunostained for CENP-C. Scale bars, 5 μm. 

**d.** Images of mammary epithelial cells after 48 and 96 hours of culture in Matrigel immunostained for ZO-1 and α6 integrin. 

**e.** Prevalence of lagging chromosomes in anaphases of wildtype mammary epithelial cells in immature and mature spheroids and control and β1 integrin knockout spheroids. Error bars indicate SD. p = 0.04 (immature versus mature spheroids), 0.047 (control versus β1 integrin knockout spheroids) by one-tailed Fisher’s exact test. n > 100 anaphases per condition. 

**f.** Images of control and β1 integrin knockout mammary epithelial cells after 96 hours of culture in Matrigel immunostained for ZO-1 and α6 integrin. Scale bars, 10 μm.
Figure 2. Acute disruption of tissue architecture leads to chromosome instability. 

a, Quantification of centrosome number (Y tubulin foci) in mitotic epithelial cells from mammary gland, skin, and neonatal liver in vitro and in vivo.
b, Western blot of whole cell lysates from control and β1 integrin knockout mammary epithelial cells cultured in Matrigel. Actin was used as a loading control.
c, Proportion of spheroids with no, partial, or complete lumens generated by wildtype mammary epithelial cells after 48 and 96 hours of culture in Matrigel or control and β1 integrin knockout mammary epithelial cells after 96 hours of culture in Matrigel.
d, Quantification of centrosome number (Y tubulin foci) in mitotic mammary epithelial cells in immature and mature spheroids and control and β1 integrin knockout spheroids.

Tissue architecture enhances correction of aberrant microtubule-kinetochore attachments

The increase in chromosome missegregation in the absence of tissue architecture could be a consequence of three non-mutually exclusive defects: 1) increased formation of aberrant microtubule-kinetochore attachments 2) impaired detection of aberrant microtubule-kinetochore attachments by the spindle assembly checkpoint, or 3) impaired correction of aberrant microtubule-kinetochore attachments by Aurora B and the mitotic spindle. To test which of
these processes underlie the defects we observed, we performed live imaging of spheroids prepared from Centrin 2-GFP;H2B-mCherry mice to simultaneously monitor spindle pole and chromosome movements in mitosis. Notably, the duration of chromosome congression (prometaphase and metaphase) was significantly higher in the immature spheroids compared to the mature spheroids (Figure 3a, first and third columns). This difference was mostly due to an increased duration of metaphase (Figure 3a, ninth and eleventh columns). This suggests that cells in immature spheroids were arresting prior to anaphase onset, perhaps as a consequence of spindle assembly checkpoint activation. Indeed, the differences between immature spheroids and mature spheroids were eliminated upon treatment with the checkpoint inhibitor reversine (Figure 3a). These results suggest that the spindle assembly checkpoint is activated in the absence of tissue architecture.

The prolonged spindle assembly checkpoint activation observed in the absence of tissue architecture could be the result of either increased formation of aberrant attachments or impaired correction of aberrant attachments. To distinguish between these possibilities, we examined the prevalence of lagging chromosomes in the presence of reversine. By inhibiting the spindle assembly checkpoint, error correction is largely eliminated such that the prevalence of lagging chromosomes directly reflects the formation of aberrant attachments. In the presence of reversine, the prevalence of lagging chromosomes was only slightly higher in immature spheroids compared to mature spheroids, and the fold increase was not nearly of the magnitude observed in the absence of reversine
From these experiments, we conclude that tissue architecture facilitates accurate chromosome segregation at least in part by enhancing the correction of aberrant microtubule-kinetochore attachments. In the absence of architecture, correction of aberrant attachments is impaired, the spindle assembly checkpoint is activated, and some of these incorrect attachments persist into anaphase leading to chromosome missegregation.

**Figure 3.** Tissue architecture facilitates the correction of erroneous microtubule-kinetochore attachments. a, Duration of different aspects of mitosis in Centrin 2-GFP;H2B-mCherry mammary epithelial cells in immature and mature spheroids the absence (-) and presence (+) of 500 nm reversine. Chromosome congression (prometaphase and metaphase) is defined as the time from when individual chromosomes could be observed to when anaphase began. Prometaphase is defined as the time from when individual chromosomes could be observed to when the centrin foci became bipolar. Metaphase is defined as the time from when the centrin foci became bipolar to when anaphase began. b, Quantification of lagging chromosomes in anaphases of mammary epithelial cells in immature and mature spheroids in the presence of 500 nm reversine.
Tissue architecture ensures chromosome stability in the setting of supernumerary centrosomes

If tissue architecture does indeed facilitate the correction of erroneous microtubule-kinetochore attachments, it should enhance chromosome stability in situations known to increase aberrant attachments. The adult liver provides an ideal in vivo system to test this prediction because adult, but not neonatal, hepatocytes are polyploid and harbor supernumerary centrosomes and the presence of supernumerary centrosomes is known to increase aberrant microtubule-kinetochore attachments (Beams and King 1942; Margall-Ducos et al. 2007; Ganem et al. 2009; Silkworth et al. 2009). We performed partial hepatectomies on adult mice to stimulate liver regeneration and analyze hepatocyte mitoses in vivo. We confirmed that polyploid, multicentrosomic hepatocytes proliferate during regeneration (Figure 4a,b). As has been reported for polyploid cells in vitro, we found that polyploid hepatocytes entered mitosis with multipolar spindles but as mitosis progressed centrosomes clustered to yield bipolar anaphase spindles (Figure 4a). However, in contrast to what has been reported for polyploid hepatocytes and other polyploid cells in vitro, we observed very few lagging chromosomes during anaphase in adult hepatocytes in vivo. Importantly, this surprising chromosome segregation fidelity of polyploid hepatocytes was dependent on tissue architecture, as knockout of β1 integrin or expansion in vitro both led to significant increases in chromosome missegregation (Figure 4c,d).
Figure 4. Tissue architecture ensures chromosome stability in the setting of supernumerary centrosomes. 

a, Images of hepatocytes in prometaphase, metaphase, and anaphase during development (neonate) and regeneration (adult) in vivo immunostained for pan-cadherin, αtubulin, and γtubulin. 

b, Ploidy of hepatocytes in interphase, prometaphase, and anaphase during development (neonate) and regeneration (adult) as inferred by chromosome 16 FISH (interphase) and centrosome number (γtubulin foci, prometaphase and anaphase). n = 100 for interphase FISH, n > 50 for γtubulin foci in prometaphase, n > 25 for γtubulin foci in anaphase. 

c, Prevalence of lagging chromosomes in anaphases of hepatocytes during development (neonate), wildtype and β1 integrin knockout regeneration (adult), and in vitro expansion. p = 0.0184 (wildtype versus β1 integrin knockout regeneration), p = 0.0004 (wildtype regeneration versus in vitro) by one-tailed Fisher’s exact test. n > 50 anaphases per condition. 

d, Images of hepatocytes expanded in vitro with
lagging chromosomes in anaphase immunostained for CENP-C. Scale bars, 5 μm

Disruption of tissue architecture could contribute to chromosome instability in cancer

Our data indicate that disruption of tissue architecture by various mechanisms is sufficient to induce chromosome instability. This leads us to the hypothesis that the disruption of tissue architecture that occurs during tumor formation and metastasis could contribute either entirely or in part to the chromosome instability observed in many tumors. Testing this hypothesis during tumorigenesis is difficult as it remains possible that any of the pre-existing mutations could affect chromosome segregation. However, one of the most universal disruptions of tissue architecture during tumorigenesis, the epithelial-mesenchymal transition (EMT), can be induced by cytokines regardless of mutational background. We induced EMT in mammary epithelial cell spheroids by treating spheroids with TGFβ and TNFα. This caused many cells to lose apicobasal polarity, downregulate epithelial cell-cell adhesion, and fill the spheroid lumen (Figure 5a, Figure 6a,b). Notably, this was also associated with an increase in chromosome missegregation that was not secondary to supernumerary centrosomes (Figure 5b, Figure 6c). This supports the idea that disruption of tissue architecture during tumorigenesis could indeed lead to chromosome instability. We note that gene expression analyses of tumors have correlated chromosome instability with higher expression of genes involved in EMT, and suggest that this could reflect reverse causality (Roschke et al. 2008).
Figure 5. Disruption of tissue architecture during tumorigenesis can lead to chromosome instability. a, Images of mammary epithelial cells after 96 hours of culture in Matrigel in the absence (-) or presence (+) of 5 ng/mL TGFβ and TNFα for the final 24 hours immunostained for ZO-1 and α6 integrin. Scale bars, 10 µm. b, Prevalence of lagging chromosomes in anaphases of wildtype mammary epithelial cells after 96 hours of culture in Matrigel in the absence (-) or presence (+) of 5 ng/mL TGFβ and TNFα for the final 24 hours. Error bars indicate SD. p = 0.0512 by one-tailed Fisher’s exact test. n > 100 anaphases per condition.

Figure 6. Induction of epithelial-mesenchymal transition in mammary epithelial cell spheroids. a, Western blot of whole cell lysates from wildtype mammary epithelial cells after 96 hours of culture in Matrigel in the absence (-) or presence (+) of 5 ng/mL TGFβ and TNFα for the final 24 hours. Actin was used as a loading control. b, Proportion of spheroids with no, partial, or complete lumens generated by wildtype mammary epithelial cells after 96 hours of culture
in Matrigel in the absence (-) or presence (+) of 5 ng/mL TGFβ and TNFα for the final 24 hours. c, Quantification of centrosome number (Ytubulin foci) in mitotic cells in spheroids generated by wildtype mammary epithelial cells after 96 hours of culture in Matrigel in the absence (-) or presence (+) of 5 ng/mL TGFβ and TNFα for the final 24 hours.

**DISCUSSION**

Our results indicate that chromosome segregation in epithelia is cell non-autonomous. Epithelial cells rely on tissue architecture to correct erroneous microtubule-kinetochore attachments. Disruption of tissue architecture therefore leads to chromosome instability in epithelial cells. This could explain chromosome instability in cancer, with loss of tissue architecture during early tumor formation as well as during metastasis giving rise to chromosome missegregation regardless of the genetic background and in the presence of functional mitotic machinery. While the chromosome segregation defects that we observe upon loss of tissue architecture are mild, it has been shown that mild levels of chromosome missegregation are more conducive to tumorigenesis than high levels of chromosome instability (Silk et al. 2013). Moreover, these individual chromosome missegregation events can beget further genomic instability, as lagging chromosomes can undergo extensive DNA damage and aneuploid cells have elevated DNA damage and chromosome missegregation (Janssen et al. 2011; Crasta et al. 2012; Zhang et al. 2015; Sheltzer et al. 2011; Passerini et al. 2016). Our observation could also in part explain why epithelial cells expand poorly ex vivo. In spite of the tremendous regenerative capacity of hepatocytes in vivo, attempts to expand untransformed hepatocytes in vitro have not been successful (Overturf et al. 2007; Levy et al. 2015). It seems likely that
the high level of chromosome instability of polyploid hepatocytes upon disruption
of tissue architecture limits their proliferative capacity. Culture systems that
restore the in vivo architecture, and thus chromosome stability, are likely to
enhance ex vivo expansion.

We recognize that not all somatic cells divide in the context of defined
architecture. It seems plausible that other, potentially cell autonomous, features
can facilitate error correction in a manner similar that provided by tissue
architecture for epithelial cells. Indeed, the tendency of established tissue culture
cell lines to round up and increase cortical tension during mitosis has been
shown to be important for chromosome segregation fidelity (Lancaster et al.
2013).

METHODS

Animals. C57BL/6J mice, Centrin 2-GFP mice, Cre-ERT2 mice, and β1
integrin\textsuperscript{flox} mice were purchased from the Jackson Laboratory. H2B-mCherry
mice were obtained from the RIKEN Center for Life Science Technologies
(CLST) (Kobe, Japan) and Albumin-CreER mice were obtained from Eric Taylor
(University of Iowa, Iowa City, IA). All animal procedures were approved by the
Massachusetts Institute of Technology Committee on Animal Care.

Tissues. Proliferative mammary glands were harvested from pregnant females
at 4.5 days gestation. Proliferative skin was harvested from embryos at 14.5 to
16.5 days gestation. Proliferative livers were harvested from pups at 8 to 10
days of age.
**Partial hepatectomies.** Partial hepatectomies were performed on 8 week-old male mice as previously described with minor modifications (Mitchell and Willenbring 2008). Instead of removing the entire median lobe and the gallbladder, only the right median lobe was removed while the left median lobe and gallbladder were spared. Mice were sacrificed 48 hours after surgery to process livers for immunofluorescence.

**Tissue dissociation.** To isolate mammary epithelial cells, mammary glands from a single 6 to 8 week-old female mouse were harvested and digested in 10 mL DMEM-F12 with 2.5 mM L-glutamine, 25 mM HEPES, 100 U/mL and 100 μg/mL penicillin-streptomycin, 20 μg/mL Liberase TM (Roche), and 150 U/mL collagenase III at 37 °C with 5% CO₂ for 16 hours. Digested tissue was resuspended with gentle trituration, diluted with 25 mL PBS, and pelleted at 1,000 RPM at room temperature for 5 minutes. The supernatant was aspirated to remove all but 5 mL of PBS and an equal volume of 0.25% trypsin-EDTA was added. Digested tissue was trypsinized at 37 °C with 5% CO₂ for 45 minutes with occasional agitation. Trypsin was inactivated by adding 25 mL of DMEM-F12 with 2.5 mM L-glutamine, 25 mM HEPES, 100 U/mL and 100 μg/mL penicillin-streptomycin, and 10% FBS. Cells were treated with 100 μL of 1 mg/mL DNase and pelleted at 1,000 RPM at room temperature for 5 minutes. Cells were resuspended in phenol red-free Mammary Epithelial Basal Medium (MEBM) with growth factors (Lonza). To isolate keratinocytes, neonatal (one day-old) mice were decapitated and cleaned with betadine followed by isopropyl alcohol. Skin was removed and floated on 0.25% trypsin in PBS at 4 °C for 16
hours. Epidermis was peeled from dermis, transferred to DMEM with 10% FBS, minced using a razor blade, and triturated using a serological pipette. Cells were centrifuged at 150 g at 4 °C for 5 minutes, resuspended in DMEM with 10% FBS, and passaged through at 70 μm mesh strainer. Cells were again pelleted, resuspended in Complete Defined Keratinocyte Serum-Free Medium (Thermo Fisher Scientific) and passaged through another 70 μm mesh strainer. To isolate neonatal hepatocytes, livers were isolated from neonatal (8-10 day-old) mice, cut into small pieces, and incubated in disruption solution (HBSS without calcium and magnesium, 25 mM HEPES, 0.5 mM EDTA, and 0.9 mM MgCl₂, pH 7.4) for 10 minutes at 37 °C with shaking for 10 minutes. This step was repeated twice with fresh disruption solution. Liver pieces were then incubated in digestion solution (HBSS with calcium and magnesium, 25 mM HEPES, and 0.05% collagenase IV, pH 7.4) for at 37 °C with shaking for 5 minutes. Remaining liver pieces were transferred to fresh digestion solution for another 5 minutes and the two digestion solutions were combined. The suspension was triturated with a serological pipette, passed through a 70 μm filter, and diluted in hepatocyte medium (DMEM with 4.5 g/L glucose, 10% FBS, 1X MEM Non-Essential Amino Acids (Thermo Fisher Scientific)). Cells were pelleted at 30 g at 4 °C for 3 minutes, resuspended in 5 mL hepatocyte medium, and combined with 5 mL Percoll-HBSS (4.5 mL Percoll, 0.5 mL 10X HBSS). Cells were pelleted at 150 g at 4 °C for 3 minutes, washed with hepatocyte media, and resuspended in SUM3 media (75% DMEM with 4.5 g/L glucose, 25% Waymouth’s MB 752/1, 0.5% FBS, 2 mM L-glutamine, 10 mM HEPES, 50 ng/mL epidermal growth factor, 1
µg/mL insulin, 30 nM sodium selenite, 10 µg/mL transferrin, 50 ng/mL somatotropin, 670 ng/mL triiodo-L-thyronine). To isolate adult hepatocytes, mice were anesthetized with 2% isoflurane-oxygen delivered via nosecone. The liver was perfused with disruption solution for 10 minutes followed by digestion solution for 10 minutes by incising the portal vein with a 25G needle attached to a peristaltic pump. The inferior vena cava was incised immediately after perfusion began to allow fluid outflow. After perfusion was complete, the digested liver was transferred to 10 mL fresh disruption solution, shaken to liberate cells, triturated with a serological pipette, passaged through a 70 µm filter, and diluted with 10 mL hepatocyte media. Cells were pelleted at 30 g at 4 °C for 3 minutes, resuspended in 12 mL hepatocyte media, and combined with 12 mL Percoll-HBSS (10.8 mL Percoll, 0.5 mL 10X HBSS). Cells were pelleted at 150 g at 4 °C for 3 minutes, washed with hepatocyte media, and resuspended in hepatocyte media.

Cell culture. To culture mammary epithelial cells as monolayers, dissociated cells were cultured in MEBM with supplements on dishes coated with 30 µg/mL collagen I at 37 °C with 5% CO₂ for 24 hours. Cells were then washed with PBS three times and treated with 0.25% trypsin-EDTA at 37 °C for 45 seconds to remove contaminating fibroblasts. Remaining cells were then trypsinized with 0.25% trypsin-EDTA at 37 °C for 6 minutes and resuspended in DMEM-F12 with 2.5 mM L-glutamine, 25 mM HEPES, 100 U/mL and 100 µg/mL penicillin-streptomycin, and 10% FBS. Cells were pelleted at 1,000 RPM for 5 minutes, washed with PBS, resuspended in MEBM with supplements, 1 µM progesterone,
1 μg/mL prolactin, plated on coverslips coated with 30 μg/mL collagen I, and incubated at 37 °C with 5% CO₂ for 24 hours. To culture mammary epithelial cells as spheroids, cells were trypsinized and washed as above and resuspended in MEBM with supplements at 7 x 10⁵ cells per mL. Cells were then combined with four volumes of phenol red-free, growth factor-reduced Matrigel (Corning) and 100 μL gels were cast in individual wells of 24-well glass-bottom plates (Mat Tek). Gels were polymerized at 37 °C for 30 minutes before adding 500 μL of MEBM with supplements, 1 μM progesterone, 1 μg/mL prolactin. Gels were cultured at 37 °C with 5% CO₂ for 24 to 96 hours, with media changed after 48 hours of culture. To deplete β1-integrin in mammary epithelial cells from Cre-ER<T₂;β1 integrin<sup>lox/lox</sup> mice, 100 nM 4-hydroxytamoxifen was added to the media for the 24 hours of culture in a monolayer and the first 48 hours of culture in Matrigel. To inhibit the spindle assembly checkpoint, 500 nm reversine (Cayman) was added to media 2 hours prior to fixation or imaging. To induce epithelial-mesenchymal transition, 5 ng/mL TGFβ and TNFα (R&D Systems) were added to media 24 hours prior to fixation or imaging. To culture keratinocytes as monolayers, dissociated cells were cultured in Complete Defined Keratinocyte Serum-Free Medium (Thermo Fisher Scientific) on coverslips coated with 10 μg/mL fibronectin at 37 °C with 5% CO₂ for 24 hours. To culture neonatal hepatocytes as monolayers, dissociated cells were cultured in SUM3 medium on coverslips coated with 30 μg/mL collagen I at 37 °C with 5% CO₂ for 48 hours. To culture adult hepatocytes as monolayers, dissociate cells were first cultured in hepatocyte media on coverslips coated with 30 μg/mL.
collagen I at 37 °C with 5% CO₂ for 24 hours followed by SUM3 medium for an additional 48 hours.

**Live imaging.** Mammary epithelial cells from Centrin 2-GFP;H2B-mCherry mice were combined with Matrigel as described above and gels were cast in 35 mm glass-bottom dishes (Mat Tek). Dishes were imaged under an incubated stage set to 37 °C with 5% CO₂ (Pathology Devices). Images were acquired using a CSU-22 spinning disc confocal head (Yokogawa) with Borealis modification (Andor) mounted on an Axiovert 200M microscope (Zeiss) with a 40X water immersion objective, an Orca-ER CCD camera (Hamamatsu), and MetaMorph acquisition software (Molecular Devices). Only cells that entered mitosis after imaging began was considered for analysis. If the cell did not enter anaphase before imaging ceased, it was analyzed only if it had been in mitosis for at least 60 minutes. In these cases, the duration of chromosome congression was the time interval from the start of mitosis to the end of imaging.

**Immunostaining.** To immunostain tissue sections, tissues were fixed in 4% paraformaldehyde in PBS for 16-24 hours. Fixed tissues were washed with PBS, cryoprotected with 30% sucrose, and frozen in O. C. T. Compound (Tissue-Tek). Slides with 30 μm thick sections were prepared using a cryostat and stored at -80°C until use. Slides were dried at room temperature for 4-8 hours, rehydrated in PBS for 15 minutes and boiled in sodium citrate buffer (10 mM tri-sodium citrate dihydrate, 0.05% Tween 20, pH 6.0) for 20 minutes. Slides were washed in PBS, dried briefly, and sections outlined with a hydrophobic pen. Sections were incubated with extraction buffer (1% Triton X-100 in PBS) for 15 minutes
followed by incubation in block (3% BSA, 0.3% Triton X-100 in PBS) for 1 hour at room temperature. Sections were incubated primary antibodies diluted in block at room temperature for 16-24 hours. Sections were washed three times with blocking solution for 10 minutes each. Sections were incubated with secondary antibodies diluted in block with 5 µg/mL Hoechst 33342 (Thermo Fisher Scientific) at room temperature for 1-2 hours. Sections were washed with block for 5 minutes twice and once with PBS for 5 minutes. Sections were mounted with ProLong Gold Antifade Reagent (Life Technologies). To immunostain cultured cells, cells were washed with PBS and fixed with 4% PFA in PBS at room temperature for 5 minutes followed by ice-cold methanol at –20 °C for 5 minutes. Fixed cells were washed with PBS for 5 minutes, permeabilized with 0.1% Triton X-100 in PBS (0.1% PBST) for 10 minutes, and blocked with 4% BSA in 0.1% PBST for 20 minutes. Primary antibodies were diluted in block and applied for 30 minutes. Cells were washed with 0.1% PBST for 5 minutes thrice. Secondary antibodies were diluted in block with 5 µg/mL Hoechst 33342 dye (Thermo Fisher Scientific) and applied for 30 minutes. Cells were washed with 0.1% PBST for 5 minutes twice followed by PBS for 5 minutes. All incubations were performed at room temperature. Cells were mounted with ProLong Gold Antifade Mountant (Thermo Fisher Scientific). To immunostain spheroids, gels were washed with PBS and fixed with 4% PFA in PBS at room temperature for 10 minutes followed by ice-cold methanol at –20 °C for 10 minutes. Fixed gels were washed with PBS for 5 minutes, permeabilized with 0.5% Triton X-100 in PBS (0.5% PBST) for 20 minutes, and blocked with 10% goat serum, 0.2% BSA,
0.2% Triton X-100, 0.1% Tween 20 in PBS for 1 hour with rocking. Primary antibodies were diluted in block and applied for 16-24 hours with rocking. Gels were washed with 0.2% BSA, 0.2% Triton X-100, 0.1% Tween-20 for 10 minutes thrice. Secondary antibodies were diluted in block with 5 μg/mL Hoechst 33342 dye (Thermo Fisher Scientific) and applied for 1-2 hours. Gels were washed for 10 minutes twice followed by PBS for 10 minutes. All incubations were performed at room temperature in the original glass-bottom wells. Gels were coated with ProLong Gold Antifade Mountant (Thermo Fisher Scientific). The following primary antibodies were used: α-Tubulin-FITC (clone DM1A, 1:1,000, Sigma), Ytubulin (clone GTU88, 1:500, Sigma), keratin 14 (clone Poly19053, 1:1,000, BioLegend), cytokeratin 8 (clone EP1628Y,1:1,000, Abcam), pan-cadherin (ab6529, 1:500, Abcam), CENP-C (gift from Iain Cheeseman,1:10,000), ZO-1 (61-7300, 1:100, Thermo Fisher Scientific), and α6 integrin (clone MAB1378, 1:250, Millipore). All secondary antibodies were various AlexaFluor conjugates (1:1,000, Thermo Fisher Scientific). The Ytubulin antibody was directly conjugated using the AlexaFluor 568 Antibody Labeling Kit (Thermo Fisher Scientific). Images were acquired using a CSU-22 spinning disc confocal head (Yokogawa) with Borealis modification (Andor) mounted on an Axiovert 200M microscope (Zeiss) with a 63X oil immersion objective (Zeiss), an Orca-ER CCD camera (Hamamatsu), and MetaMorph acquisition software (Molecular Devices).

**Western blotting.** To prepare tissue lysates, flash-frozen tissues were resuspended in RIPA buffer (50 mM Tris pH 8.0, 150 mM sodium chloride, 1%
NP-40, 0.5% sodium deoxycholate, 0.1% SDS) containing Complete Mini Protease Inhibitor Cocktail (Roche). Tissues were homogenized using a dounce homogenizer, incubated on ice for 20 minutes, and centrifuged at 13,000 rpm at 4 °C for 20 minutes. Protein concentration of supernatants was measured using Bradford dye (Bio-Rad) on a spectrophotometer. To prepare spheroid lysates, gels were digested with Matrigel Dispase (Corning) for 1 hour at 37 °C, spheroids were pelleted at 1,000 RPM at room temperature for 5 minutes, resuspended in RIPA buffer containing Complete Mini Protease Inhibitor Cocktail (Roche), lysed using a cell scraper, and centrifuged at 13,000 rpm at 4 °C for 20 minutes. Protein concentration of supernatants was measured using Bradford dye (Bio-Rad) on a spectrophotometer. To prepare samples for western, lysates were diluted in 5X sample buffer (250 mM Tris pH 6.8, 50% glycerol, 5% β-mercaptoethanol, 0.025% bromophenol blue, 5% SDS). Samples were separated on homemade polyacrylamide gels and transferred to Immobilon-P membranes (Millipore) via wet transfer. Membranes were blocked with 5% milk in TBST (50 mM Tris pH 8.0, 150 mM NaCl, 0.1% Tween-20) for 1 hour at room temperature. Membranes were incubated in primary antibodies diluted in 5% milk in TBST at 4 °C with rocking for 16 hours and washed with TBST for 10 minutes thrice. Membranes were incubated in secondary antibodies diluted in TBST at room temperature with rocking for 1 hour and washed with TBST for 10 minutes thrice. Membranes were incubated in ECL Prime Western Blotting Detection Reagent (GE Healthcare) for 5 minutes and imaged on an ImageQuant LAS 4000 luminescent image analyzer (GE Healthcare). The following primary
antibodies were used: β1 integrin (clone MAB1997, 1:1,000, Millipore), E-cadherin (clone DECMA-1, 1:500, Abcam), and β-actin (clone AC-74, 1:20,000, Sigma). For β1 integrin, β-mercaptoethanol was eliminated from the 5X sample buffer. All secondary antibodies were ECL horseradish peroxidase linked (1:5,000, GE Healthcare).

ACKNOWLEDGMENTS

We thank Iain Cheeseman for helpful discussions, Konstantina Rowald and Rocio Sotillo for advice in establishing mammary epithelial cell culture, and Wendy Salmon for imaging advice. This work was supported by the National Institutes of Health Grant GM056800 and the Kathy and Curt Marble Cancer Research Fund to A.A. A.A is also an investigator of the Howard Hughes Medical Institute and the Glenn Foundation for Biomedical Research. K.A.K. is supported by the NIGMS Training Grant T32GM007753.

REFERENCES


Beams HW, King RL. 1942. The Origin of Binucleate and Large Mononucleate Cells in the Liver of the Rat. The Anatomical Record 83: 281–297.


CHAPTER FIVE

Aneuploidy in cancer: seq-ing answers to old questions

Reprinted with permission from Annual Reviews


KAK led the study. TD downloaded and processed TCGA data and KAK and TD analyzed the data.
ABSTRACT

Aneuploidy, the state of having gained and lost chromosomes, is a hallmark of cancer. Approximately 90% of tumors have gained or lost at least one chromosome. Aneuploidy is therefore as frequent, if not more so, than disruption of the p53 pathway. However, whether and how aneuploidy influences tumorigenesis is still poorly understood. Here we take advantage of large-scale tumor sequencing efforts to assess karyotypic alterations across many cancer types and review recent sequencing studies that show how karyotypes change in space and time. We further summarize findings that describe the effects of aneuploidy on untransformed cells and mechanisms by which aneuploidy could drive tumorigenesis while also serving as a unique therapeutic target.
Cancer is, in its purest sense, an isolated environment of mutation and selection occurring within a multicellular host. In this process, somatic cells acquire various genomic alterations, some of which endow the cells with a fitness advantage and enable the expansion of these cells within a tissue. These genomic alterations can take many forms. The linear sequence can change, through base substitutions, insertion of novel sequence, and structural rearrangements. Additionally, nucleotide and histone modifications can be altered. Lastly, the copy number of a region of the genome can change, through the gain or loss of segments ranging from a few to millions of nucleotides. All of these alterations have the potential to affect cellular fitness by changing the function or expression of RNAs and proteins. Most cancers harbor a combination of different alterations throughout the genome. However, across cancers, the extent of gene copy number changes far exceeds changes in nucleotide sequence. While the average tumor harbors approximately one point mutation every megabase, approximately one-third of its genome is affected by copy number changes (reviewed in (Garraway & Lander 2013)). Most of this copy number change occurs in the form of large-scale gains and losses that can be detected cytologically as changes in karyotype (Beroukhim et al. 2010). These alterations of entire chromosomes or large portions of chromosomes are known as whole-chromosome aneuploidy and segmental aneuploidy, respectively. Whole-chromosome and segmental aneuploidy arise by distinct
mechanisms but affect a similar amount of the cancer genome. In this review we focus on whole-chromosome aneuploidy.

Because any single chromosome gain or loss will alter the expression of up to thousands of genes, aneuploidy has pleiotropic effects on cells and unraveling its role in cancer is complex. To characterize this relationship, we must understand 1) the spectrum of karyotypic changes in tumors, 2) the evolution of these changes in time and space, and 3) the requirement of these changes for tumorigenesis. As karyotypic change in tumors has been documented for over a century, these questions are not novel and attempts to answer them abound. However, the increasing availability of cancer genome sequencing data, both of whole tumors and of individual tumor cells, provides us with new power to revisit these questions. Here, we combine insights from classical cell biology and novel genomic approaches to re-evaluate the relationship between aneuploidy and cancer.

ORIGIN OF ANEUPLOIDY IN CANCER

Whole-chromosome aneuploidy arises from chromosome missegregation. During mitosis, microtubules attach to kinetochores to move sister chromatids within the cell. For sister chromatids to partition into separate daughter cells, the paired kinetochores must attach to microtubules emanating from opposite spindle poles (amphitelic attachment) (Figure 1A). Erroneous attachments, in which only a single sister chromatid is attached to microtubules (monotelic attachments), in which both sister chromatids are attached to microtubules emanating from the same spindle pole (syntelic attachments), or in which a single sister chromatid is
attached to microtubules emanating from both spindle poles (merotelic attachments), can all lead to chromosome missegregation and aneuploidy.

Many potential chromosome missegregation events are prevented by the spindle assembly checkpoint, a surveillance mechanism that detects aberrant attachments and prevents anaphase onset until the faulty attachments are corrected (reviewed in (Musacchio & Salmon 2007)). However, while the spindle assembly checkpoint can easily detect monotelic and syntelic attachments, it is believed to be less sensitive to merotelic attachments. As a result, even when cells have a functional spindle assembly checkpoint, merotelic attachments can persist into anaphase and, in some cases, lead to chromosome missegregation (Cimini et al. 2001, Thompson & Compton 2011). It is estimated that a single chromosome is missegregated approximately once every $10^3$-$10^4$ cell divisions and indeed aneuploidy can be found at low levels in normal tissues (Cimini et al. 1999, Knouse et al. 2014, Thompson & Compton 2011).

The rare chromosome segregation errors that occur in untransformed cells could underlie aneuploidy in cancer, producing karyotypic changes that remain stable during tumorigensis. However, in light of the high levels of karyotypic change observed in a large fraction of tumors, many cancer cells are believed to have an increased frequency of chromosome missegregation, termed chromosome instability (CIN). There is no single mechanistic explanation for CIN in cancer. While disruption of the spindle assembly checkpoint would provide the most direct route to chromosome instability, most chromosomally unstable cancer cell lines have functional spindle assembly checkpoints (Gascoigne &
Taylor 2008, Haruki et al. 1999, 2001; Tighe et al. 2001). As such, many investigations have focused on alterations in cancer cells that would cause chromosome missegregation in spite of a functional spindle assembly checkpoint (Figure 1B). The presence of multiple centrosomes, a feature of many cancer cells, has been shown to increase the frequency of merotelic attachments and thus chromosome missegregation (Ganem et al. 2009, Silkworth et al. 2009). Additionally, some cancer cell lines have been shown to have increased microtubule stability and assembly, which also leads to increased merotely (Bakhoum et al. 2009, Ertych et al. 2014). Alternatively, some cancers exhibit defects in sister chromatid cohesion, which can prevent detection of all types of attachment errors (Barber et al. 2008). Finally, errors occurring prior to mitosis, such as DNA damage or telomere attrition, can produce acentric or dicentric chromosomes that can attach improperly to the mitotic spindle and missegregate (Burrell et al. 2013, Hande et al. 1999).
Figure 1

A. Outcomes of Potential Kinetochore-Microtubule Attachments

- Amphitelic: Checkpoint satisfied, Proper segregation
- Monotelic: Checkpoint activated
- Syntelic: Checkpoint activated
- Merotelic: Checkpoint possibly satisfied, Possible mis-segregation

B. Checkpoint-Independent Sources of Chromosome Misseggregation

- Supernumerary centrosomes
- Hyperstable microtubules
- Impaired sister chromatid cohesion
- Chromosome damage

Figure 1. Origin of aneuploidy in cancer

A. Description of the ways in which sister kinetochores can attach to microtubules and the effects of these attachments on activation of the spindle assembly checkpoint and subsequent chromosome segregation.

B. Depiction of possible defects in cancer cells that could lead to chromosome missegregation in spite of a functional spindle assembly checkpoint.

There are clearly multiple pathways to changing chromosome copy number. However, much of the characterization of chromosome instability has been performed in cancer cell lines in vitro. The mechanism and extent to which CIN contributes to cancer karyotypes in vivo remains to be determined. Are different mechanisms responsible for CIN in different tumors? Are complex karyotypes the product of continuous CIN, transient CIN, isolated aberrant mitoses, or a combination of these processes? Regardless of mechanism, it is clear that many tumors somehow acquire high levels of aneuploidy. In the next
section, we analyze the growing number of cancer genome sequences to assess the degrees and patterns of aneuploidy across cancers.

STATE OF ANEUPLOIDY IN CANCER

Traditionally, karyotypic aberrations in tumors were assessed by chromosome spreads and comparative genomic hybridization (CGH). These approaches revealed striking heterogeneity across cancers (reviewed in (Albertson et al. 2003)). Some tumors had very few alterations while others had dozens, and this heterogeneity was present both across and within different cancers. In the past decade, the genome sequences of thousands of tumors have become available thanks to the collaborative work of The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium. These sequencing data allow for the detection of various types of genomic alterations, including aneuploidy, so long as they exist in a substantial portion of cells comprising the tumor.

To better understand the degree of karyotypic change within and across different cancers, we processed sequencing data from TCGA to identify whole-chromosome aneuploidies. We analyzed a total of 10,249 individual tumor samples (henceforth referred to as tumors) comprising 24 different cancer types (henceforth referred to as cancers) using thresholds of copy number change that could detect single chromosome gains or losses even if the tumor had undergone a single whole genome doubling event. These analyses allowed us to reassess, in detail, the state of aneuploidy across and within cancers. We reaffirmed the striking heterogeneity of karyotypic change across cancers but find
that cancers can be classified in terms of the extent and uniformity of such change.

**Pan-Cancer Patterns of Aneuploidy**

Of the 10,249 tumors we analyzed, 84% had at least one aneuploid chromosome. There were notable biases in the types of aneuploidy. First, chromosome loss was more common than chromosome gain (Figure 2A). This is consistent with previous reports that a tumor’s absolute DNA content is usually somewhat less than the predicted ploidy (Carter et al. 2012, Zack et al. 2013). Second, certain chromosomes, such as 7 and 20, were frequently gained, while other chromosomes, such as 13 and 22, were frequently lost. Another subset of chromosomes, including 1, 2, and 3, were rarely altered. It is difficult to envision mechanisms that cause the gain or loss of specific chromosomes. Instead, these biases across chromosomes likely emerge from selective pressure. Notably, there is a significant inverse correlation between the number of genes on a chromosome and the frequency with which it is lost across all cancers, suggesting that large-scale changes in gene expression are poorly tolerated (Figure 2B). The minimal overlap between chromosomes that are frequently lost and those that are frequently gained further suggests that the combined effect of genes on a given chromosome can have either a positive or negative impact on tumorigenesis. We revisit this possibility later.
Figure 2. Pan-cancer patterns of aneuploidy
A. The proportion of all tumors analyzed that show gain or loss for a given chromosome.
B. The correlation between the number of genes on a given chromosome and the frequency with which that chromosome is gained or lost across all tumors.

Cancer-Specific Patterns of Aneuploidy
Although aneuploidy was identified in a substantial fraction (84%) of tumors analyzed, it manifested in unique ways across cancer types. First, the percent of aneuploid tumors for a given cancer type ranged from 5% to 99% (Figure 3). Aneuploidy was identified in only 5% of acute myeloid leukemias and thyroid carcinomas while 99% of ovarian carcinomas were aneuploid. Moreover, the degree of aneuploidy across cancers was highly variable. The average
number of aneuploid chromosomes per tumor ranged from 0 to 9 across cancers (Figure 3). Finally, there was variation in the extent to which the tumors of a given cancer had similar karyotypes. While many cancer types show high karyotype variability some cancer types exhibit highly characteristic karyotypes (Figure 3). For example, for cervical carcinoma no single chromosome was gained or lost in more than 25% of tumors. On the other hand, over half of kidney renal papillary cell carcinomas had gained chromosomes 16 and 17, while across all cancers the frequency of gain for chromosomes 16 and 17 was less than 8% (Figure 3). To quantify this intra-cancer homogeneity—the extent to which tumors of the same cancer shared specific aneuploidies—we computed the variance in copy number for each chromosome and averaged variance across all chromosomes for each cancer.

Although each cancer type presented a distinct pattern of aneuploidy and level of homogeneity, we were able to classify tumors into four groups based on whether their degrees of aneuploidy and homogeneity were above or below the median (Figure 3). The first group was characterized by both low levels of aneuploidy and low homogeneity. This class includes thyroid carcinoma and prostate adenocarcinoma (Figure 3, group 1 and Figure 4, first and second panels). The second group of cancers exhibits low but homogenous levels of aneuploidy. This class includes glioblastoma and kidney papillary carcinoma (Figure 3, group 2 and Figure 4, third and fourth panels). For example, the average glioblastoma had only 5 aneuploid chromosomes. Notably, 75% of glioblastomas had gained chromosome 7, 79% had lost chromosome 10, and
65% of tumors harbored both aneuploidies. The third group of cancers harbor high levels of aneuploidy with little homogeneity among tumors of the same type. This group includes breast carcinoma and lung adenocarcinoma (Figure 3, group 3 and Figure 4, fifth and sixth panels). For both cancers, no single chromosome gain or loss was present in more than 40% of tumors. The final group of cancers is characterized by a high level of stereotypic aneuploidies. For example, the average colorectal tumor had 6 aneuploid chromosomes, with gain of chromosome 13 and loss of chromosome 18 present in over 55% of samples (Figure 3, group 4 and Figure 4, seventh panel). Melanoma had a similar pattern of high levels of recurrent aneuploidies (Figure 4, eighth panel).
Figure 3. Cancer-specific patterns of aneuploidy

Characteristics of the 24 cancers analyzed individually (first 24 columns) and together (final column). The first row indicates the percent of tumors of each cancer that are aneuploid for at least one chromosome. The second row indicates the percent of tumors of each cancer that had undergone whole genome doubling (NA, information not available, pan-cancer value obtained from (Zack et al. 2013)). The subsequent two heatmaps show the frequency with which a given chromosome is gained or lost across cancers. The bottom two heatmaps describe the extent of karyotype homogeneity within a cancer and the average number of aneuploid chromosomes for that cancer.
Figure 4

Group 1
Low aneuploidy
Low homogeneity

Group 2
Low aneuploidy
High homogeneity

Group 3
High aneuploidy
Low homogeneity

Group 4
High aneuploidy
High homogeneity

Figure 4. Examples of cancer karyotype classes
Plots showing the copy number state of chromosomes (log₂ scale) for each tumor analyzed of representative cancers of each karyotype class.
Our analysis confirms that there is no single “cancer karyotype”, with different cancers varying in their levels of aneuploidy and homogeneity. It is not intuitive that these metrics would vary across cancers. Two broad explanations could underlie either or both of these features. First, these could be artificial differences imposed by differential timing of diagnosis across cancers. Alternatively, these could be genuine differences reflecting cell type-specific variability in the generation of and response to karyotypic change. With respect to the former, it has been shown that the degree of aneuploidy increases with increasing tumor grade and stage (Ried et al. 1999). As such, if different cancers tend to be diagnosed at different times this could lead to differences in the perceived degree of aneuploidy. However, our analyses suggest that this unlikely explains differences in the degrees of aneuploidy across cancers. For tumors in which staging information was available, the distribution of tumor stages was similar in cancers of the low aneuploidy groups (groups 1 and 2) and cancers in the high aneuploidy groups (groups 3 and 4).

Variability in how different cells types generate and respond to karyotypic change could lead to different degrees of aneuploidy and karyotypic homogeneity across cancers. There are several possible sources of variability across cell types. For example, genome doubling, a feature observed in some premalignant lesions, is associated with increased generation of and tolerance to aneuploidy and could therefore explain differences in the level of aneuploidy across cancers ((Galipeau et al. 1996), reviewed in (Davoli & de Lange 2011)). In our analyses, we were able to infer the proportion of tumors that had likely undergone whole
genome doubling for some of the cancers (Figure 3). Indeed, whole genome
duplication was more frequent in cancers with high aneuploidy (groups 3 and 4).
Thus, although increased genome doubling could explain the high aneuploidy
observed for some cancers, it is still unclear why genome doubling would affect
some cell types more than others.

Differences in the extent of karyotypic homogeneity across cancers could
reflect aneuploidy serving largely as a driver in some cancers while existing
mostly as a passenger in others. For cancers with highly homogenous
karyotypes, such as glioma, kidney papillary carcinoma, and colorectal
carcinoma, it seems plausible that the highly recurrent aneuploidies are selected
for and contribute to tumorigenesis. Consistent with this, the recurrent
aneuploidies observed in kidney papillary carcinoma and colorectal carcinomas
are already recurrent in the tumors of low stage. For example, the recurrent
gains on chromosomes 16 and 17 observed for kidney papillary carcinoma were
already present in 45% and 65%, respectively, of stage one tumors. On the other
hand, for cancers with high but variable aneuploidy it is possible that some of the
more frequent aneuploidies act as drivers and coexist with many other gains and
losses that minimally affect fitness. It is unclear why aneuploidy would
predominate as a driver alteration in some cancers while existing largely as a
passenger in others. If aneuploidy is indeed driving some cancers, different
aneuploidies are driving different cancers. While the biological basis is unclear, it
is consistent with different cancers showing enrichment for different sequence
mutations (Kandoth et al. 2013, Lawrence et al. 2014).
In summary, analysis of cancer genomes reveals that most cancers harbor whole chromosome gains and losses, and that specific but distinct events tend to recur within and across cancer types. We are now left with questions of evolution. How does the karyotype change during tumor development and dissemination? How does karyotype vary, if at all, across tumor space?

EVOLUTION OF ANEUPLOIDY IN CANCER

The analysis of tumor genomes provides only a snapshot of aneuploidy—the changes present in most of the tumor at a given point in time. We are unable to say when each change arose and whether each change is present in every single cell. We therefore now investigate the evolution of karyotypic change in time and space. By examining aneuploidy at distinct stages in tumor development and varying positions within a tumor we can gain insight into when and how chromosome missegregation shapes the tumor genome.

Understanding the evolution of the tumor karyotype can provide initial insights into how such changes may influence tumor development.

Aneuploidy in Time

Insight into the timing of karyotypic change during tumor development has been enabled by the analysis of defined premalignant lesions that precede certain cancers. Studies of unmatched colorectal adenomas (premalignant lesion) and colorectal carcinomas by comparative genomic hybridization (CGH) showed that aneuploidy was present in both the adenomas and the carcinomas, however the proportion of samples of each type having any aneuploidy and the amount of aneuploidy per sample was higher in carcinomas compared to
adenomas (Meijer et al. 1998, Ried et al. 1996). For example, of the lesions analyzed by Ried et al. 1996, 21% of low-grade adenomas, 58% of high-grade adenomas, and 88% of carcinomas harbored karyotypic changes. Moreover, in aneuploid low-grade adenomas only one chromosome was affected, whereas the average aneuploid carcinoma had six aneuploid chromosomes.

Several recent studies have addressed similar questions using Barrett’s esophagus, a precursor lesion for esophageal adenocarcinoma. Two groups performed whole-genome sequencing of patient-matched, synchronous biopsies of Barrett’s esophagus and esophageal adenocarcinoma (Ross-Innes et al. 2015, Stachler et al. 2015). In concordance with the studies in colon, copy number alterations were present at low levels in Barrett’s esophagus but widespread in esophageal carcinoma. For example, in the samples analyzed by Ross-Innes et al. 2015, on average 0.3% of the genome in Barrett’s esophagus samples had copy number alteration whereas the average esophageal carcinoma had copy number alterations affecting 37.6% of the genome. Similarly, in the evolution of breast cancer copy number changes have been identified at low levels in ductal hyperplasia (premalignant lesion) but much higher levels in ductal carcinoma in situ (Chin et al. 2004). Together, these analyses of premalignant lesions and their malignant counterparts demonstrate that karyotypic change is present early in tumor development but that copy number changes increase with malignant progression.

The analyses described above, which assess only two points in the tumor lifetime, provide little information regarding the rate of change. A recent
longitudinal case-cohort study of individuals with Barrett’s esophagus provides
greater insight into the pace of karyotypic change (Li et al. 2014). Li et al.
followed and serially biopsied patients for up to 22 years. For patients that did
not progress to esophageal adenocarcinoma, the biopsies showed little
karyotypic change, with fewer than 200 Mb of copy number alteration for the
duration of the study. However, the patients that did progress to esophageal
carcinoma showed a drastic increase in copy number alteration (between 200
and 3,000 Mb of alterations) appearing in the Barrett’s esophagus biopsies within
the four years preceding diagnosis of esophageal carcinoma. Moreover, many of
the lesions underwent whole genome doubling in the two years prior to cancer
diagnosis. This study shows that premalignant lesions harbor a low level of
karyotypic change that can remain stable for decades. However, in many cases
the transition from premalignant lesion to malignancy is accompanied by a large
increase in copy number changes in the few years preceding diagnosis.
Although this burst of karyotypic change seems rapid in comparison to many
years with which Barrett’s esophagus can exist with minimal copy number
change, a few years is a long time with respect to the doubling time of single
cells. Perhaps this delay between karyotypic change and malignancy represents
a time during which the nascent tumor is sampling, adapting, and optimizing a
combination of complex aneuploidies and other mutations until a highly fit
genome emerges. Consistent with the hypothesis of premalignant cells sampling
the space of karyotypic variation is the fact that the level of clonal diversity at the
genomic level in Barrett’s esophagus predicts the likelihood of progression to
esophageal adenocarcinoma (Maley et al. 2006). While more investigations of this nature are required before reaching general conclusions about karyotypic evolution, it appears that aneuploidy is a longstanding but changing feature of the cancer genome. As a tumor progresses so too does its karyotype, sometimes quite rapidly.

Aneuploidy in Space

The evolution of karyotypes over time is driven by individual cells acquiring new alterations that allow these cells to expand relative to other cells. Thus, karyotypic heterogeneity in space drives karyotypic changes over time. Indeed, population-based sequencing of different sectors of individual breast tumors has shown that the majority of tumors are comprised of multiple clones with unique genetic changes (Navin et al. 2010, Torres et al. 2007b, Yates et al. 2015). While population-based analyses of different regions of a tumor allow for insight into karyotypic variability across dominant clones, they are insensitive to rare changes within clones. Single cell sequencing is emerging as a powerful tool for assessing intratumoral heterogeneity. This approach has the resolution to detect events that occur in a single cell, and therefore is especially useful for documenting the extent of karyotypic heterogeneity within a tumor.

Single cell sequencing was first applied to a triple negative breast tumor (Navin et al. 2011). Flow cytometry analysis of cells isolated from different regions of the tumor revealed that it was comprised of three spatially distinct aneuploid populations: a hypodiploid population and two separate sub-tetraploid populations as well as a ubiquitous diploid population. Sequencing single cells
from each of the populations revealed very homogenous copy number changes among cells within each aneuploid population consistent with three separate clonal expansions. Interestingly, while most of the diploid cells had no copy number alterations and were assumed to be infiltrating lymphocytes, some had extensive copy number changes that were not shared by other cells. A subsequent single cell sequencing study revealed similar trends (Wang et al. 2014). Together, these two studies document high intratumoral heterogeneity in karyotype, confirming previous reports of intratumor heterogeneity in breast cancer (Navin et al. 2010, Torres et al. 2007b, Yates et al. 2015). However, instead of uniformly distributed variability, this heterogeneity exists in the form of distinct populations of defined karyotype. As these tumors evolved, novel karyotypes arose that expanded but remained relatively stable. This leads to specific conclusions about chromosome instability in these tumors. Perhaps chromosome instability is transient, generating a burst of novel karyotypes, one of which is highly fit and subsequently expands with few other changes. Alternatively, chromosome instability is constantly generating deviants of the clonal karyotype, but most of these cells are unfit and hardly contribute to tumor mass. The presence of pseudodiploid cells with unique copy number alterations is especially interesting. Are these cells the remnants of karyotypic sampling at distinct stages of tumorigenesis or are they the unfit products of continuous chromosome instability? In the above studies, these pseudodiploid cells were more closely related to the infiltrating cells than the aneuploid clones, suggesting that, if indeed cancerous, they arose early in tumorigenesis.
The most extreme form of spatial heterogeneity is that between a primary tumor and its metastases. Primary tumors constantly shed cancer cells into the circulation and, in rare cases, these cells colonize a new location. Given that a tumor karyotype can evolve in both time and space, the relationship between a primary tumor and its metastases is presumably highly dependent on the heterogeneity of the primary tumor at the time of dissemination and the amount of time elapsed since dissemination. Most evidence supports a “parallel progression” model of tumorigenesis (reviewed in (Klein 2009)), whereby metastases result from cells disseminating early in primary tumor development and evolving separately from the primary tumor. This would allow primary tumors and their metastases to follow distinct mutational and karyotypic trajectories. Indeed, most studies that have compared karyotypes of metastatic lesions with their primary tumors have supported the notion of branched evolution. Analysis of primary breast tumors and asynchronous metastases revealed similar degrees of copy number change across samples, however the extent to which the samples shared copy number alterations varied (Kuukasjärvi et al. 1997). Two-thirds of samples shared many copy number alterations allowing for inference of a clonal relationship while the remaining one-third of samples had such divergent copy number profiles that no clonal relationship could be inferred. Similar trends were observed in a recent study comparing endometrial tumors and abdominopelvic metastases (GIBSON ET AL. 201X). For approximately 80% of pairs, the percent of genome affected by copy number alteration was similar in the primary tumor and metastasis, while in the remaining
pairs the metastasis had a much higher level of copy number alteration. However, although most pairs harbored similar levels of copy number alteration, only around half of copy number alterations were shared between primary tumors and metastases. Similar divergence has been observed for the brain metastases of various primary tumors (Brastianos et al. 2015). Less divergence was observed across different metastatic lesions of primary prostate cancer, although the different metastases were not always compared to the primary tumor (Liu et al. 2009). While metastases typically have similar degrees of aneuploidy and share some of the aneuploidies with their primary tumor, these secondary tumors have undergone at least some karyotypic divergence. This suggests that the ideal karyotype of a tumor cell—the karyotype that provides the greatest fitness advantage—might vary depending on environment.

Certainly more studies of this scope are required before these concepts can be generalized across cancer. However, these initial results suggest that cancer karyotypes are finely tuned, implying an active role in tumorigenesis. However, it remains possible that these distinct karyotypes are passively but stably associated with a specific set of other genetic alterations that drive tumorigenesis. We now address the question of whether karyotypic change alone can drive tumorigenesis.

ROLE OF ANEUPLOIDY IN CANCER

Aneuploidy in untransformed cells and organisms

Before assessing the role of aneuploidy in cancer it is helpful to understand how aneuploidy affects untransformed cells and organisms. At the
At the cellular level stable aneuploidy is associated with impaired fitness. Across species, aneuploidy is associated with transcriptional changes that mimic the environmental stress response (ESR), a transcriptional signature first described in yeast which is induced by a variety of adverse conditions and associated with impaired cell growth and proliferation (Gasch 2007, Sheltzer et al. 2012). In *Drosophila*, induction of aneuploidy can lead to cell cycle arrest or apoptosis (Dekanty et al. 2012, Gogendeau et al. 2015). Aneuploid yeast and mammalian cells have impaired energy utilization and are sensitive to inducers of energy stress (Tang & Amon 2013, Williams et al. 2008). Aneuploid cells also exhibit proteotoxic stress and are sensitive to various inhibitors of protein quality control (Oromendia et al. 2012, Tang et al. 2011, Torres et al. 2007a). Lastly, aneuploid cells exhibit genomic instability in the form of increased mutation, increased structural rearrangements, and increased mutations (Nawata et al. 2011, Passerini et al. 2016, Sheltzer et al. 2011, Zhu et al. 2012). In addition to
these universal consequences of aneuploidy, experiments in yeast have revealed a variety of distinct cellular stresses resulting from specific karyotypic changes (Dodgson et al. 2016).

While karyotypic change is typically detrimental for cells and organisms in their native context, experimental evolution studies in microorganisms indicate that whole-chromosome copy number changes can increase fitness in stressful environments. Aneuploidy is frequently observed as an early event in adaptation to nutrient limitation, drug treatment, and other stresses (Dunham et al. 2002, Rancati et al. 2008, Selmecki et al. 2006, Yona et al. 2012). Consistent with these observations, it was recently shown that aneuploid mammalian cells grow better than euploid cells in adverse environments (Rutledge et al. 2016).

Effects of continuous, random karyotypic change—chromosome instability—have also been examined in humans and mouse models. In humans, mutation of the spindle assembly checkpoint gene \textit{Bub1b} causes mosaic variegated aneuploidy (MVA) and an appreciable level of somatic karyotypic variation. This disease is characterized by developmental defects and an increased incidence of various childhood malignancies (Hanks et al. 2004, Kajii et al. 2001). Mouse models of chromosome instability have been generated by mutating various components of the mitotic machinery. The phenotypes are diverse and at times contradictory. At one end, mutation of the spindle assembly checkpoint component \textit{Bub1b} in mice results in a progeroid phenotype and reduced incidence of spontaneous tumors (Baker et al. 2004). On the other hand, mutation of the spindle assembly checkpoint component \textit{Bub1},
overexpression of the spindle assembly checkpoint component Mad2, or hyperactivation of the spindle assembly checkpoint target Cdc20 results in increased spontaneous tumorigenesis (Jeganathan et al. 2007, Li et al. 2009, Sotillo et al. 2007). The pro- and anti-tumorigenic effects of increased chromosome missegregation are best captured by loss of function mutations in the gene encoding the mitotic motor CENP-E. Reducing levels of CENP-E leads to elevated aneuploidy and an increase in spontaneous lymphomas and lung tumors but reduction in spontaneous liver tumors, carcinogen-induced tumors, and genetically-induced tumors (Weaver et al. 2007). From these various mouse models, it appears that tumorigenesis is a possible but not obligatory outcome of increased chromosome missegregation and is likely influenced by the degree of instability and tissue context.

Aneuploidy as Cancer’s Arm

It is clear that neither aneuploidy nor chromosome instability are necessary or sufficient for tumorigenesis. First, our analyses of karyotypic changes across cancers revealed some tumors that had no karyotypic changes (Figure 3). Second, as discussed above, many cases of stable aneuploidy or chromosome instability are actually associated with reduced incidence of cancer. However, though neither necessary nor sufficient, in specific cases both stable aneuploidy and chromosome instability appear to facilitate tumorigenesis. This idea that aneuploidy can, in select circumstances, promote tumorigenesis is not unreasonable. Aneuploidy encompasses a variety of copy number alterations. It seems likely that only certain patterns of copy number and gene expression
changes would support tumorigenesis, perhaps only in the context of a specific cell type or other accompanying mutations. Studies of cancer predisposition in Down syndrome support the importance of context of the aneuploid state. In mouse models of Down syndrome, the reduced burden of solid tumors can be attributed in part to increased expression of tumor suppressor genes on the duplicated chromosome (Baek et al. 2009, Sussan et al. 2008). Likewise, a subset of genes on chromosome 21 that regulate lymphoid development are suspected to be involved in the increased risk of leukemia in Down syndrome patients (Lane et al. 2014).

The case of Down syndrome indicates that aneuploidy could promote tumorigenesis if gains enrich for oncogenes and losses enrich for tumor suppressor genes. Indeed, analyses of the gene content of recurrent gains and losses support the idea that copy number changes can drive tumorigenesis by altering the expression of genes that regulate cell proliferation. For example, for focal copy number alterations (changes that are smaller than a chromosome arm), recurrent deletions are enriched in genes that restrain cellular proliferation and depleted in essential genes and oncogenes (Solimini et al. 2012). At the level of chromosome arm and whole chromosome gains and losses, the density and potency of tumor suppressor genes and oncogenes as inferred by analysis of point mutations in tumors also correlates with the frequency of deletion and amplification, respectively, across all tumor types (Davoli et al. 2013). Since the vast majority of these copy number events are hemizygous, this implies that the copy number changes are either cooperating with mutation of the other allele or
are exerting fitness effects in isolation through haploinsufficiency and
triplosensitivity.

In addition to the recurrent patterns of aneuploidy across all tumor types, there exist recurrent aneuploidies specific to individual tumor types. The latter could reflect cell type-specific sensitivities to different oncogenes and tumor suppressor genes as well interactions with other mutations unique to each tumor. Moreover, tumor cell fitness is influenced not only by cell autonomous processes but also by the cell’s ability to avoid external stresses such as hypoxia, immunosurveillance, and, notably, chemotherapy. Given the role of aneuploidy in adaptation to stress in microorganisms, it seems plausible that aneuploidy may help certain tumors adapt to such external stresses. These external stresses may vary with tumor size and location, and thus the karyotypic changes that provide adaptation would be expected to vary too. While the model based on the cumulative balance of cancer gene dosages provides support for the hypothesis that specific recurring aneuploidy events are a driving force during the evolution of human cancer, precise modeling of these genomic events in mice will be required to further test these hypotheses.

The above theories suggest that only specific aneuploidies will facilitate tumorigenesis, while others will actually inhibit the process. However, there is also evidence from mouse models that the aneuploid state can facilitate tumorigenesis regardless of karyotype (Jeganathan et al. 2007, Li et al. 2009, Sotillo et al. 2007, Weaver et al. 2007). Given that genomic instability is a feature of many models of stable aneuploidy, it is possible that aneuploidy
provides a level of mutability sufficient to adapt to novel environments while still compatible with cell viability.

The importance of context with respect to the relationship between aneuploidy and tumorigenesis also holds for the relationship between chromosome instability and tumor development. Analysis of the CENP-E model of chromosome instability revealed that in situations in which CENP-E reduction inhibited tumorigenesis, these situations had high basal levels of chromosome missegregation that were exacerbated by CENP-E reduction. The situations where the reduction of CENP-E function increased tumor burden were not found to have elevated baseline chromosome missegregation (Silk et al. 2013). Therefore, like the aneuploid state per se, moderate levels of chromosome instability could create karyotypes that provide cancer cells with a selective advantage. Moreover, the act of missegregating chromosomes can lead to other forms of genomic alterations and therefore potential for the evolution of favorable variants. Several studies have shown that missegregating chromosomes experience DNA damage, sometimes leading to alterations as severe as chromothripsis (Crasta et al. 2012, Janssen et al. 2011, Zhang et al. 2015). It is thus possible that chromosome missegregation and aneuploidy provide an appropriate level of mutability that increases the pace by which cancer cells activate oncogenes, disrupt tumor suppressor genes, and adapt to stress.

Given that aneuploidy increases the mutability and adaptability of untransformed systems, the effects of chromosome copy number changes on cancer cells could extend beyond the adjustment of oncogene and tumor
suppressor gene levels. If these additional effects do indeed enhance the mutability and adaptability of tumors, one might expect aneuploidy to be associated with poor prognosis in cancer. Although there is some evidence that aneuploidy is associated with improved prognosis in childhood tumors (Niggli et al. 1994, Taylor et al. 1988, Tomita et al. 1988), studies of all other cancers indeed associate aneuploidy with poor prognosis. Leukemias with single or multiple chromosome gains or losses have worse prognoses than leukemias without numerical cytogenetic abnormalities (Byrd et al. 2002, Grimwade et al. 1998, Wolman et al. 2002). Analyses of DNA ploidy and proliferation by flow cytometry have associated aneuploidy with increased proliferation and worse prognosis in a variety of solid tumors (Allison et al. 1991, Emdin et al. 1987, Ohyama et al. 1990, Wenger et al. 1993). It remains to be determined whether this relationship is causal or correlative. If causal, the contribution of aneuploidy to tumorigenicity might not be simply increased proliferation through adjustment of oncogenes and tumor suppressor genes but also through increased adaptability in the setting of environmental stresses and therapeutic interventions.

Aneuploidy as Cancer's Achilles Heel

The aneuploid state, regardless of specific karyotype, is associated with adverse fitness of untransformed cells in their native environment. This suggests that cancer cells have either acquired karyotypic change in the setting of novel internal or external conditions or, more likely, acquired additional mutations that help cancer cells tolerate the adverse consequence of aneuploidy while
exploiting the benefits of altered gene dosage. Indeed, there is evidence that cancer cells have adapted to proteotoxic stress that could in part be generated by the aneuploid state. Many tumors have increased expression of HSF1, an inducer of a variety of mediators of proteostasis, and many cancer cell lines are dependent upon HSF1 for survival (Dai et al. 2007, Santagata et al. 2011). Although cancer cells likely mask the adverse consequences of aneuploidy, karyotype changes could represent hidden vulnerabilities of cancer cells, and thus therapeutic targets (Tang et al. 2011). Therapeutic targeting of this and other karyotype-nonspecific consequences of aneuploidy has several advantages. First, aneuploidy is unique to cancer cells, with aneuploidy rarely occurring in untransformed tissues (Knouse et al. 2014). This is in contrast to rapid proliferation, which is a feature shared by both tumors and many normal tissues. Second, given that aneuploidy could serve as a means of adapting to selective pressures such as chemotherapy, targeting karyotypic change could prevent evolution of resistance.

That specific karyotypes can endow cells with unique susceptibilities is especially interesting in light of the many cancers with highly specific karyotypes. Indeed, analysis of glioblastoma showed a correlation between chromosome 7 gain and sensitivity to irinotecan (Chen et al. 2015). It is possible that these recurrent aneuploidies play a critical role in tumorigenesis, and applying selective pressure against recurrent aneuploidies could be effective therapy for such cancers.
CONCLUSIONS

Our analyses of bulk tumor sequencing data reveal that there is no single portrait of aneuploidy in cancer. There are some cancers with hardly any karyotypic changes while other cancers have extensive aneuploidy. For a given type of cancer, the aneuploidies may be random or recurrent. Studies of karyotype evolution and intratumoral heterogeneity suggest that complex karyotypes arise early in clonal expansion and remain relatively fixed across cells within a clone. The forces that shape these cancer karyotypes remain to be characterized in detail in vivo. Does chromosome instability continuously spawn new variants from which optimal karyotypes are selected or is chromosome instability a punctuated phenomenon that drives tumor evolution in discrete steps?

The presence of recurrent karyotypes in select cancers and fixed copy number alterations within clones suggest that, in at least some cases, karyotypic change is selected for and contributes to tumorigenesis. Indeed, in vivo and in silico experiments support the idea that aneuploidy can drive tumorigenesis. The complexity of karyotypic changes in cancer further emphasizes the role of context. Aneuploidy is but one means of altering the genome. The fitness consequences of a given copy number change are shaped by the other genomic alterations present at that time. This merits further analyses that assess, simultaneously, all forms of genome alteration in a single cancer genome and their effects in the setting of a particular cell type and tissue environment.
REFERENCES


Cimini D, Tanzarella C, Degrassi F. 1999. Differences in malsegregation rates obtained by scoring ana-telophases or binucleate cells. *Mutagenesis*. 172


175


CHAPTER SIX

Discussion
DETECTING SOMATIC COPY NUMBER ALTERATIONS USING SINGLE CELL SEQUENCING

The viability and reproduction of organisms is contingent upon transmission of a balanced genome. Large-scale copy number alterations that exist throughout an organism are typically associated with embryonic lethality or disease (Girirajan et al. 2011; Hassold and Jacobs 1984). However, there has been speculation that such changes might be better tolerated, and perhaps even selected for, at the somatic level (Rehen et al. 2001; 2005; Duncan et al. 2010; 2012). Confirming these speculations has been complicated by limitations in the methods for detecting copy number alterations in single cells. Chromosome spreads are prone to false positive aneuploidies due to chromosomes spreading between cells and can only be performed on mitotic cells. Fluorescence in situ hybridization (FISH), while applicable to interphase cells, is subject to hybridization artifacts and assesses only a few predetermined loci in any given experiment. The development of single cell sequencing provided an option for unbiased, genome-wide assessment of copy number in individual cells regardless of proliferative status. This technique was promptly employed for the assessment of copy number alterations in the human brain, yielding reports that over 40% of neurons harbored megabase-scale copy number alterations (McConnell et al. 2013; Cai et al. 2014). However, these early applications of single cell sequencing did not include thorough validation of the approach.

We performed a variety of in silico and in vitro experiments to test the detection performance of single cell sequencing with respect to detecting copy
number alterations. We found that false positive CNVs are extremely common, accounting for up to half of all CNV calls, especially when algorithms are adjusted to detect CNVs less than 5 Mb and when samples with high levels of noise are analyzed. However, we found that by establishing stringent criteria to exclude cells of poor quality and by overlapping two distinct algorithms for CNV detection, we were able to reliably detect CNVs exceeding 5 Mb with a low false discovery rate. Single cell sequencing is therefore an improved method for investigating copy number alterations in somatic cells, but it must be employed carefully.

Currently, the two main drawbacks of single cell sequencing, along with other methods for detecting somatic copy number alterations, are that their resolution is limited to megabase-scale alterations and that their use prevents the confirmation of putative alterations by an orthogonal approach. However, improvements in single cell sequencing technology could enhance the detection, validation, and characterization of alterations. Optimizing whole genome amplification to ensure more uniform amplification across the genome will lead to less noise in the sequencing data. This will first and foremost allow for detection of smaller alterations without high false discovery rates. Moreover, when combined with paired-end sequencing reads and high sequencing depth, this will allow for identification of breakpoints and analysis of single nucleotide polymorphisms (SNPs), both of which can validate putative deletions. Finally, the emergence of methods to simultaneously amplify the genome and transcriptome of single cells will not only help validate putative alterations but
may also lend insight into their consequences for cell physiology (Macaulay et al. 2015).

**PREVALENCE OF SOMATIC COPY NUMBER ALTERATIONS**

Using our validated single cell sequencing approach, we first determined the prevalence of megabase-scale copy number alterations (both sub-chromosome CNVs and whole-chromosome aneuploidy) in human brain and skin. We could not determine the prevalence of sub-chromosome CNVs in liver, as hepatocytes are polyploid and the reduced copy number ratios associated with single-copy gains and losses in a polyploid background cannot reliably be distinguished from noise at the sub-chromosome level. We identified megabase-scale copy number alterations in 9% of brain cells and 8% of skin cells. While deletions were more common than amplifications, they also tended to be smaller in size than amplifications. We validated the majority of these alterations through increased sequencing depth and analysis of SNPs. In spite of only identifying a handful of copy number alterations, we were surprised to find two recurrent alterations. The most common alteration was a loss on the proximal portion of chromosome 8, occurring in 3 of the cells analyzed across two individuals and accounting for 13% of all CNVs identified.

We then focused exclusively on whole-chromosome aneuploidy in mouse and human brain, skin, and liver. We identified aneuploidy in between 0 and 3% of brain and skin cells and between 0 and 4% of hepatocytes. Thus, the prevalence of somatic aneuploidy is less than 5%, regardless of tissue of origin. It appears that the prevalence of aneuploidy might be slightly higher in the liver.
than in the brain and skin, but establishing a significant difference would require analyzing more cells. This difference could be secondary to hepatocytes being polyploid and harboring supernumerary centrosomes, as supernumerary centrosomes can lead to increased chromosome missegregation in vitro (Ganem et al. 2009; Silkworth et al. 2009).

The frequencies of sub-chromosome and whole-chromosome copy number alterations inferred from our studies are substantially less than those previously reported. While we discovered megabase-scale CNVs in less than 10% of brain cells, other studies reported megabase-scale CNVs in over 40% of neurons (McConnell et al. 2013; Cai et al. 2014). Based on our analyses of the other datasets, we conclude that the higher frequencies reported by these studies are due to the inclusion of cells of poor sequencing quality and the use of algorithms with low stringency, both of which lead to identification of many false positive CNVs. Similarly, while we discovered whole-chromosome aneuploidy in less than 5% of cells across tissues, other studies using chromosome spreading and FISH reported levels of aneuploidy exceeding 20% and 50% in the brain and liver, respectively (Rehen et al. 2001; 2005; Duncan et al. 2010; 2012). We attribute the alleged high levels of aneuploidy to artifacts associated with the use of chromosome spreading and FISH. Indeed, we too observed higher levels of aneuploidy when using FISH, but use of dual-color labeling indicated most of these events were actually hybridization artifacts. While it could be that preparation of cells for sequencing selects against aneuploid cells, we excluded this possibility through control experiments with known mixtures of aneuploid and
euploid cells. Thus, we are confident that our single cell sequencing approach provides more reliable estimates of copy number alterations than previous studies.

Although the frequency of somatic copy number alterations that we determined is less than the frequencies reported by prior studies, this frequency of somatic variants is still much higher than that reported for constitutional (organism-wide) variants. While we identified megabase-scale CNVs in nearly 10% of somatic cells, constitutional CNVs of this size exist in only 1% of individuals (Itsara et al. 2009). Moreover, though we found whole-chromosome aneuploidy in up to 4% of somatic cells, constitutional aneuploidy is found in only 0.3% of live births (Hassold and Jacobs 1984). These differences could be the result of either or both increased generation of such variants or reduced selection against these changes at the somatic level. For whole chromosome aneuploidy, it is clear that differential selection is responsible for at least part of this difference. Chromosome missegregation is common in female meiosis, and aneuploidy has been observed in over 15% of human oocytes (Hou et al. 2013). The frequency of aneuploidy in oocytes (>15%) is much higher than the prevalence of aneuploidy in live births (0.3%), reflecting strong selection against constitutional aneuploidy (Hassold and Jacobs 1984). For CNVs, it seems that differential selection could again account for at least part of this difference. Constitutional megabase-scale CNVs typically exist as de novo events rather than heritable alterations, indicating that they experience strong negative selection (Itsara et al. 2010). Thus, it appears that large-scale copy number
alterations are better tolerated when they exist in a subset of somatic cells rather than organism-wide.

While the selective pressures against large-scale copy number alterations might be lower in somatic cells, there is still evidence that these events do experience negative selection. When somatic aneuploidy is induced in *Drosophila* via genetic manipulation, many of the cells undergo cell cycle arrest or apoptosis (Dekanty et al. 2012; Gogendeau et al. 2015). Moreover, for megabase-scale CNVs we observed an inverse correlation between the size of the CNV and its frequency in single cells (Figure 1). While we cannot exclude the possibility that larger CNVs are less likely to arise, we believe this result is in part due to the strength of selection increasing with the magnitude of genomic imbalance. Notably, the occurrence of copy number changes of chromosome arms and whole chromosomes is higher than one would expect given the relationship between CNV size and frequency. These large-scale events might be more frequent because the mechanisms that give rise to them, such as chromosome missegregation, occur more often than those that give rise to sub-chromosome CNVs. In addition, while whole-chromosome events can affect genome balance without sequence changes, sub-chromosome CNVs involve DNA breaks and thus sequence changes. Depending on the location of the break, sub-chromosome copy number alterations, though smaller, could have larger consequences for a cell than whole chromosome aneuploidy.
Figure 1. Relationship between the size and frequency of copy number alterations. Plot showing the frequency with which a copy number variant of a given size (in megabases) or a chromosome arm or whole chromosome event was identified in single cells.

We were surprised to find that both CNVs and aneuploidy, though uncommon, tended to cluster in individual cells. For example, the 23 CNVs we identified in 132 brain and skin cells were distributed among only 11 cells. Similarly, of the four aneuploid hepatocytes we identified out of 100 hepatocytes, one of these cells harbored two aneuploid chromosomes. We acknowledge that the cells with multiple CNVs could be an artifact of analyzing cells that satisfied quality control criteria but still had high levels of noise. However, assuming that these concurrent alterations represent real events, we consider two non-mutually exclusive conclusions about somatic copy number alterations. It is possible that the multiple aberrations all result from the same event, such as severe DNA
damage or a highly aberrant mitosis. Alternatively or in addition, it is possible that having a single alteration predisposes a cell to developing further alterations. Indeed, aneuploid cells have been shown to have higher frequencies of both DNA damage and chromosome missegregation (Nawata et al. 2011; Passerini et al. 2016; Sheltzer et al. 2011; Zhu et al. 2012).

Our assessment of copy number alterations in somatic cells leaves several questions for further inquiry. First, our study indicates the presence of CNVs and aneuploidy in somatic cells at a single point in time. This value is influenced by both the rate at which these events arise and the rate at which these events are eliminated. Our study cannot distinguish between the generation and selection of copy number alterations. Query of somatic cells immediately after cell division could address the former, while a means of tracking the fate of altered cells would address the latter. If these cells do persist, it will be important to determine how they impact organ function. Perhaps these cells senesce or perhaps they transform. We note that the most common CNV we identified, a deletion on the proximal portion of chromosome 8, is also an extremely frequent CNV in cancer (Zack et al. 2013). This region contains a fragile site and thus its prevalence might reflect its increased frequency of occurring. However, this region also contains the tumor suppressor CSMD1, making it plausible that this deletion provides cells with a proliferative advantage and ultimately, in the presence of additional mutations, facilitates transformation (Bailey et al. 2002; Yu et al. 2010; Ma et al. 2009; Midorikawa et al. 2009). Lastly, our study did not provide the power to address the relationship
between the age of an individual and the frequency of somatic copy number alterations, but it seems possible that these alterations would increase with age if negative selection did not fully compensate for the rate at which they arise or persist through cell division.

ROLE OF TISSUE ARCHITECTURE IN CHROMOSOME STABILITY

Studies of chromosome segregation have revealed a complex network of intracellular regulators and safeguards, leading to the assumption that it is a cell autonomous process. If this is true, the process and accuracy of chromosome segregation should be unaffected by the surrounding environment. However, much of our understanding of chromosome segregation is derived from cells expanded as monolayers in vitro, preventing the accumulation of evidence that chromosome segregation may not be cell autonomous. We tested this assumption by analyzing chromosome segregation in defined cell types in different environments. We found that chromosome missegregation is significantly more frequent when cells are expanded in vitro compared to in vivo. The chromosome instability observed in vitro is secondary to disruption of tissue architecture, and it can be rescued by culturing cells in an organotypic fashion. We found that tissue architecture facilitates chromosome segregation by enhancing the correction of aberrant kinetochore-microtubule attachments. This enhancement is sufficient to ensure relatively accurate chromosome segregation even in the setting of supernumerary centrosomes, as occurs in hepatocytes. This is consistent with the low levels of aneuploidy we observed in the mammalian liver in spite of what has been shown for polyploid cells in vitro.
We conclude that chromosome segregation is in fact non-cell autonomous.

While we know that tissue architecture facilitates chromosome segregation by enhancing the correction of erroneous kinetochore-microtubule attachments, we still do not understand the mechanistic details linking architecture and correction. Further experiments will aim to determine how elements at the cell periphery, including transmembrane proteins and polarity factors, influence the function of the error correction machinery.

The observation that disruption of tissue architecture leads to chromosome missegregation provides a potential explanation for chromosome instability in solid tumors. While many cancer cells have high levels of chromosome missegregation, there has been no single explanation for chromosome instability in cancer. As most cancer cell lines have fully functional spindle assembly checkpoints, chromosome instability has instead been attributed to different checkpoint-independent mechanisms in different cancer cell lines (Gascoigne and Taylor 2008; Haruki et al. 2001; Tighe et al. 2001; Ganem et al. 2009; Silkworth et al. 2009; Bakhoum et al. 2009; Ertych et al. 2014; Barber et al. 2008). However, a universal feature of solid tumors is the disruption of tissue architecture—both during the early stages of tumorigenesis as a clone expands outside the confines of the tissue and during metastasis when individual cells colonize a new area. Our results indicate that this alone, regardless of the genetic background and in the absence of mutations in the mitotic machinery,
could lead to chromosome instability. Importantly, we note that the chromosome missegregation that occurs upon disruption of tissue architecture is mild. However, studies suggest that mild levels of chromosome missegregation are indeed optimal for tumors—allowing for gradual karyotypic optimization without leading to sudden genomic imbalance (Silk et al. 2013).

Primary epithelial cells typically proliferate poorly ex vivo. Indeed, in spite of the remarkable regenerative capacity of hepatocytes in vivo, attempts to expand untransformed, primary hepatocytes in vitro has had limited success (Overturf et al. 2007; Levy et al. 2015). Our findings suggest that chromosome instability secondary to loss of tissue architecture could be responsible, at least in part, for this reduced proliferative capacity. In vitro systems that recapitulate the cell-cell adhesion and cell polarity patterns of the in vivo environment for a given cell type will restore chromosome stability and likely enhance proliferative capacity. Indeed, organotypic culture methods have been established for select cell types and further efforts should be invested in developing approaches for other cell types (Shamir and Ewald 2014).

We acknowledge that not all proliferative somatic cells have the defined tissue architecture—cell-cell adhesion, cell-matrix adhesion, and cell polarity—that is typical for epithelial cells. However, we speculate that tissue architecture may not be the only means of enhancing chromosome stability in somatic cells. Many epithelial cells must maintain their contacts and polarity regardless of their cell cycle state for the purpose of tissue function and thus might rely on their surrounding environment to facilitate chromosome segregation. However, cells
that exist in isolation, such as the blood cells, might employ alternate, cell
autonomous mechanisms for enhancing chromosome segregation. For example,
the tendency for cells in culture to round and increase cortical stiffness has been
shown to be important for accurate chromosome segregation (Lancaster et al.
2013). We note that blood cancers have minimal aneuploidy, suggesting that the
safeguards they employ for chromosome segregation are not disrupted during
tumorigenesis.

ANEUPLOIDY IN CANCER

Though aneuploidy has been recognized as a common occurrence in solid
tumors for decades, there is still no consensus as to how it influences
tumorigenesis. Given that aneuploidy is associated with fitness defects at the
cell and organismal level in untransformed systems, it is difficult to reconcile how
it would be a hallmark of a disease of enhanced proliferation. Indeed, both
human examples and mouse models of stable aneuploidy and chromosome
instability have discrepant findings when it comes to tumorigenesis. While a
detailed assessment of karyotypic changes within and across individual tumors
and within and across cancer types would provide a solid foundation for
addressing these discrepancies, traditional methods for assessing cancer
karyotypes were not amenable to amassing all of this information. The growing
field of cancer genomics, with large-scale sequencing efforts for both bulk tumors
and individual tumor cells, provides a wealth of data suitable for this purpose.

We analyzed bulk tumor sequencing data from many tumors of a variety of
cancer types as well as single cell sequencing data from individual tumors to
better understand karyotypic changes within and across cancers. We reaffirm assumptions from previous studies that there is no universal cancer karyotype. Some cancers have minimal aneuploidy, while others are highly aneuploid. For both cases, there are cancers in which given aneuploidies recur across tumors (homogenous tumors) and there are cancers in which the aneuploidies are random across tumors (heterogeneous tumors). Within a given tumor, there can be high karyotypic heterogeneity across clones within a tumor but very little karyotypic heterogeneity within a given clone.

Our analysis of bulk tumor karyotypes suggests that karyotypes are optimized in a cell type-specific fashion. As there is no evidence suggesting that different cell types would preferentially missegregate different chromosomes, we assume that all cancer cells generate the same spectrum of karyotypic changes. However, because the bulk karyotype differs across cancer types but is often similar across tumors of the same cancer, it must be true that a given karyotypic alteration has different fitness effects on different cell types. Presumably the epigenetic and transcriptional landscape of a given cell type, and the pre-existing mutations that arose during tumorigenesis, will influence whether a given karyotypic change has a positive, neutral, or negative effect on that cell’s fitness. This suggests that aneuploidy indeed influences tumorigenesis by influencing the fitness of tumor cells. To fully understand how aneuploidy influences tumorigenesis, karyotypic changes must be studied with consideration for the other genomic alterations and the cell type of the tumor.
Studies that have analyzed the karyotypes of single cells within a tumor further suggest that karyotypes are optimized and therefore influence tumorigenesis. While some tumors are composed of clones with different karyotypes, the karyotypes within a clone are rather homogeneous. If karyotype were irrelevant, one would expect to observe more heterogeneity within a clone. These studies also lead to inferences about the nature of chromosome instability in cancer. If chromosome instability is indeed continuous, then the deviants of the preferred karyotype that are constantly being generated must be of such reduced fitness that they hardly contribute to tumor mass. Alternatively, we suggest that chromosome instability might be transient. Perhaps transient insults to a nascent tumor, whether it be loss of epithelial architecture or hypoxic stress, acutely disrupt mitosis leading to a burst of altered karyotypes, one of which emerges as a fit but stable clone. More high-resolution studies of intratumoral heterogeneity as well as assessment of chromosome instability in vivo throughout tumorigenesis will help better understand how chromosome instability influences karyotype variation and tumorigenesis.

**SUMMARY**

Model organisms and systems are extremely useful for elucidating fundamental principles of biology. However, these tools may not always reveal processes optimized for a specific context and might not recapitulate how disruption of biological processes give rise to a particular disease. The work presented in this thesis emphasizes the importance of context and highlights the validity of direct, in vivo observations for addressing specific questions. I found
that somatic copy number alterations are more frequent, and presumably better tolerated, than constitutional copy number alterations. I subsequently found that chromosome segregation by primary cells is not perfectly recapitulated in vitro as disruption of tissue architecture interferes with chromosome segregation. Finally, I reaffirm that cancer karyotypes are highly specific to a given cancer, highlighting the importance of cell-type specificity in the disease process.

REFERENCES


Cancer Biology & Therapy 8: 907-916.


Silk AD, Zasadil LM, Holland AJ, Vitre B, Cleveland DW, Weaver BA. 2013. Chromosome missegregation rate predicts whether aneuploidy will promote

195


