The Genetics and Epigenetics of Centromeres in Cancer

by

Anjan Kumar Saha

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Cancer Biology)
in the University of Michigan
2019

Doctoral Committee:

Professor David Markovitz, Chair
Professor Arul Chinnaiyan
Professor Uhn-Soo Cho
Professor Elizabeth Lawlor
Professor Muneesh Tewari

Anjan K. Saha

aksaha@umich.edu

ORCID iD: 0000-0001-6721-1599

Dedication

To my mother, Namita Saha, who sacrificed pursuing a career in academia to raise her children and support our success in whichever endeavor we choose to pursue.

Acknowledgements

First and foremost, to David Markovitz. You have watched me grow across all aspects of life during the close to 11 years that we have known each other and worked together. In fact, it was joining your laboratory in 2008 that precipitated the transition from pursuing a career in the physical sciences to one in the biological sciences. Many thanks for the guidance, the support, the space to craft ideas, and the doors that you have opened over the years to bring me to where I am presently. Many thanks also to my committee members, Arul Chinnaiyan, Uhn-Soo Cho, Elizabeth Lawlor, and Muneesh Tewari for the intellectual guidance during committee meetings and individual meetings and for the ongoing career guidance in my quest towards becoming a successful physician scientist.

To past and current members of the Markovitz laboratory who have watched me grow as a scientist, from not knowing how to pipette to now earning a doctorate. Special thanks to Scott Gitlin who forwarded my initial inquiries about gaining exposure to biomedical research to David, setting me on the path to the present. Additional thanks to Mark Kaplan, who has been a constant source of inspiration to keep producing and executing upon ideas, as an idea is only as good as the execution that follows to turn the idea into a practical real-world solution. Thank you also to Rafael Contreras-Galindo, who helped propel our adventure into the land of centromere genetics and epigenetics. Wishing you the best as an independent investigator at the Hormel Institute.

Special thanks to Arul Chinnaiyan and the Michigan Center for Translational Pathology for our collaborative efforts over the course of my dissertation training. Thank you to Yashar Niknafs and Steve Kregel who helped make this thesis possible. Special thanks also to Gil Omenn and the Department of Computational Medicine and Bioinformatics, for encouraging exploration into a line of thinking that is positioned to transform the future of science and medicine. Thank you to Maureen Sartor and Tingting Qin who helped make this thesis possible.

To the Medical Scientist Training Program (MSTP), Cancer Biology Program, faculty, and students, who have given me the opportunity to present my work, interact with incredible scientists, and who have guided my development as a scientist. Thank you to Ron Koenig, Justine Hein, Laurie Koivupalo, Hilkka Ketola, Zarinah Aquil, Dawn Storball and Ellen Elkin. I would also like to thank my funding sources: the Medical Scientist Training Program T32 Training Grant (T32GM007863-36), the Cancer Biology Program Fellowship, and the Ruth L. Kirschstein Individual Predoctoral NRSA F30 Award (F30CA210379-01)

To the friends who have been supportive of my pursuits of becoming a physician scientist. It has been fun to grow together. Thanks to Karina Anam, Apurba Chakrabarti, Katie Chakrabarti, Monica Choo, Chris Colonna, Evie Coves-Datson, Jessica Dominic, Daysha Ferrer-Torres, Chris Frost, Sumit Gupta, Akshay Hingwe, Nikhil Iyer, Sarina Khan, Mitra Nair, Dushyanth Srinivasan, Ramzi Tabbara, Benjamin White, Jerry Yan, and Alfred Yoon.

To my significant other and future life partner Kelsey Trotta. Thank you for helping me achieve this major milestone and thank you for all of the support in getting me here. I am looking forward to many more moments that we will be able to share together as we continue moving forward.

Finally, to my mother and sister. I doubt completing the requirements of any degree will improve your opinions of me, but I would have it no other way. Tough love is the motto in the Saha family. So onward.

Table of Contents

List of Tables

List of Figures

Abstract

The centromere is an essential structure required for the faithful segregation of chromosomes during mitosis, a process that is significantly dysregulated in cancer. The nucleic acid sequences that dominate the centromeric landscape are α-satellites, arrays of 171 base-pair monomer units arranged into higher-order arrays throughout the centromere. Nucleosomes that replace the canonical histone H3 with the centromeric H3 variant CENPA epigenetically mark the centromeric locus. Modern genomics and epigenomics are, however, unable to navigate the highly repetitive structure of this region, a technical shortcoming that greatly hinders the ability to study their role in human malignancy, which is the second leading cause of death globally. Methodologies aimed at characterizing centromeric sequence and function are thus needed to effectively study the genetics and molecular biology of the centromere in cancer. Our group has developed quantitative PCR assays capable of detecting chromosome-specific α-satellite sequences. This novel methodology has enabled us to study centromeres in the contexts of normal biology and human disease.

We demonstrate in this thesis that centromeres undergo both genetic and epigenetic alterations in the setting of cancer. Specifically, centromeric α-satellite sequences and pericentromeric HERV-K111 retroviral sequences experience copy number reductions and sequence homogenization in neoplastic cells and tissue. Furthermore, CENPA, the H3 variant that defines centromeric chromatin, was observed to be overexpressed across multiple cancer types, with functional significance to prostate cancer phenotypes. Intriguingly, we show that overexpressed CENPA possesses a previously uncharacterized function as a putative regulator of

gene transcription of important cell cycle, centromere, and kinetochore genes, through ectopic localization to their respective transcriptional start sites.

Our findings collectively underscore the necessity of studying diseases of cell division (i.e. cancer) from the perspectives of centromere genetics and epigenetics. Further insight into the mechanistic underpinnings of centromere derangements in cancer will provide additional molecular context to our understanding of this fundamental structure, thus providing opportunities to therapeutically reconfigure centromeres to structurally emulate their normal conformation. Repairing the form and function of this structure that is ubiquitously important for cancer proliferation may prove to be a viable and efficacious therapeutic strategy that has pan-cancer clinical potential.

Chapter 1 – Introduction: Centromeres in Biology and Disease

Cell Division

Over three centuries have transpired since Robert Hooke first coined the term “cell” in his 1665 publication *Micrographia* to describe what we now classify as the fundamental unit of life[1]. Within this historic work, Hooke detailed his use of microscopy to visualize empty spaces contained by walls within a thin slice of cork. He called these contained spaces “cells.” This discovery galvanized a period of unprecedented study into these building blocks that make up all living things[2]. The variation in structure and function of these microscopic units is now thought to underlie the diversity of life-forms on Earth[3]. This diversity ranges from simple unicellular prokaryotes to complex multicellular eukaryotes. Indeed, the complex tissues, organs, organ systems and symbiotic unicellular microbiota that govern human biology reflect the importance of cellular variation to the maintenance of human life and health[4].

Equally important to human life and health is cellular proliferation, the process by which cellular life propagates to facilitate growth and reproduction[5]. This process is highly conserved across all domains of life, where pre-existing cells give rise to new cells through cell division. The unifying principle that defines cell division is the segregation of hereditary material from the parent cell into each resulting daughter cell. Cell division takes on vastly different forms in prokaryotes and eukaryotes, owing to structural differences in their respective hereditary material[6]. While both rely upon chromosomes containing DNA to encode hereditary information, prokaryotes typically possess circular chromosomes while eukaryotes possess linear sets of chromosomes. These structural differences necessitate distinct mechanisms to partition chromosomes during cell division. Circular chromosomes typically undergo DNA replication and segregation simultaneously in a process known as binary fission[7]. Linear sets of chromosomes, on the contrary, require physical and temporal compartmentalization of DNA replication and segregation in order to preserve the fidelity of both processes[8]. The physical and temporal compartmentalization of replication and segregation is conferred through the cell cycle, whereby each linear set of DNA is replicated first within a membrane bound nucleus and subsequently partitioned by way of either meiosis or mitosis[8,9]. Meiosis is the process by which haploid gametes are produced whereas mitosis is the process by which somatic cells produce genetically identical progeny. While their products differ genetically and functionally, both result in the partitioning of linear sets of chromosomes, thus requiring similar molecular machinery.

Molecular Determinants of Cell Division

Microscopic evidence of cell division was first published in 1882, when Walther Flemming used dyes that selectively bound what were later called chromosomes to depict the segregation of threadlike material during various stages of cell division in salamander embryos[10]. Reconciling Flemming's molecular work with Gregor Mendel's laws of inheritance ultimately established chromosomes as the hereditary material that is transferred from parent to offspring[11,12]. Flemming's methodology represents one of the first instances by which molecular compatibility between a synthetic substance and naturally occurring biological structures (i.e. chromosomes) was repurposed for the study of cell biology. Advances in technology have since facilitated molecularly driven studies of cell division that demonstrate remarkable specificity. Modern renditions of Flemming's work rely upon compounds that selectively bind molecular targets of interest in order to precisely ascertain spatial and temporal distributions of the molecular components essential to cell division[13,14]. Aided by these techniques, we have uncovered enormous complexity within the molecular composition of the cell division machinery. Various structures, their subunits, and their derivatives work in a highly coordinated fashion in order to ensure that each cell's set of chromosomes is divided with high fidelity[8].

First, the cell cycle, a process that spatially and temporally separates DNA replication and segregation, is orchestrated by molecules termed cyclins and their effector kinases called cyclin-dependent kinases (CDKs)[15]. First characterized in 1971 through experimentation with frog oocytes, cyclin-CDK complexes (then called the maturation promoting factor) are now known to phosphorylate enzymes and structural proteins important for DNA replication, chromosome segregation, and cell cycle progression[16,17]. Cyclins have many isoforms, each of which exhibit cell cycle stage dependent abundance. For example, cyclin B and its effector kinase CDK1 together initiate and preside over mitosis during M-phase of the cell cycle. Similarly, the G1, S, and G2 phases of the cell cycle have their own dominant cyclin-CDK complexes[5,15,18].

Second, once the mitotic cyclin-CDK complex has initiated mitosis, a number of structural factors are required to physically separate replicated chromosomes (**Figure 1.1**). Going from distal to proximal relative to the chromosomes, centrosomes assemble at opposites poles of the cell during prophase, establishing the axis along which cell division will occur[19]. The centrosomes subsequently give rise to the mitotic spindle fibers, composed of the cytoskeletal protein tubulin[20]. Tubulin multimers polymerize towards the chromosomes until they engage multimeric structures on each chromosome termed kinetochores[21]. Kinetochores serve as the molecular interface between each chromosome and the spindle fibers tasked with separating replicated strands of DNA. Finally, the structural hallmark of chromosomes that specifies the assembly point for the overlying kinetochore is known as the centromere, a genomic locus that is typically situated centric to the ends of each chromosome[21–23].

The importance of each of these components to cell division thus suggests that derangements in their structure and function are observed in diseases of cell division, i.e. cancer. Indeed, numerous components of the molecular machinery for cell division demonstrate altered structure and activity in the setting of malignancy[24,25]. In particular, given the broad consensus surrounding genomic instability and epigenetic anomalies in cancer, the centromere presents an intriguing focus of study[26]. Indeed, these genetic loci and their overlying histone profiles serve as the nexus for all components of the cell division machinery, given their central positions within each chromosome[21]. Here, I review our current understanding of centromeres and their contributions to cancer pathogenesis.

Centromeres

When Walther Flemming outlined his observations about cell division in 1882, he paid particular attention to an area of each chromosome that he called the "primary constriction[10]." This region, now called the centromere, has since been the subject of much investigation. The centromere is an area of specialized chromatin that serves as the foundation upon which the kinetochore assembles[21–23]. The reliance upon centromeres for proper cell division is highly conserved across eukaryotic biology[27,28]. Centromeres have been observed in numerous species, with structural differences that guide their respective molecular biology (**Figure 1.2**). Point centromeres are compact loci whose DNA sequences are both necessary and sufficient to specify centromere identity and function, as these sequences serve as binding sites for essential kinetochore proteins. Regional centromeres are composed of large arrays of repetitive satellite

DNA that is often packaged into condensed and largely inaccessible heterochromatin. The DNA in regional centromeres contributes to but, does not necessarily define, function[28]. This variation in structure coincides with species variation in genome size and function; point centromeres are found in smaller and simpler organisms like budding yeast whereas regional centromeres are found in higher order organisms, including humans.

Point Centromeres

Point centromeres were first identified in 1980 by Louise Clarke and John Carbon[29,30]. Their finding represents the first time that centromere DNA was isolated for molecular analyses. Point centromeres are discreet loci that contain consensus sequences with binding affinity to essential kinetochore proteins. In budding yeast (*Saccharomyces cerevisiae*), where point centromeres are best characterized, a single ~125 base pair monomeric sequence is occupied by a single nucleosome composed of a histone H3 variant termed Cse4 (**Figure 1.2a**). Each of the 16 chromosomes in budding yeast contains the ~125 base pair consensus sequence bound nucleosome that subsequently directs the assembly of the kinetochore[30]. Since this discovery, the field has uncovered astounding structural complexity within centromeres across numerous species, complexity that serves as the bedrock upon which cell division finds its genesis. The discovery of point centromeres led Clarke to subsequently identify regional centromeres in fission yeast (*Schizosaccharomyces pombe)*, paving the way for understanding centromeres of higher order species that almost exclusively possess regional centromeres[31].

Regional Centromeres

Regional centromeres are large regions of DNA that serve as the assembly point for the kinetochore and the cell division machinery to facilitate separation of chromosomes. Regional centromeres are found across eukaryotic species and take on similar structure[22,23]. The positioning of these centromeres relative to the telomeres, or the ends of the chromosomes, produces varied chromosomal structures (acrocentric, submetacentric, or metacentric) based on the ratio between the length of the q (or long) arm and the p (or short) arm (**Table 1.1**). The genetic component of regional centromeres is defined by highly repetitive DNA sequences arranged in head to tail fashion across the entire centromere, where each repeat unit is largely similar but not identical. In humans, regional centromeres make up close to 5% of the human genome[22]. The repetitive elements that comprise human centromeres are called α-satellites, which are 171 base pair units that are rich in adenine and thymine (**Figure 1.2b**). Unlike point centromeres, the α-satellite rich backbone of regional centromeres is sufficient but not necessary for proper cell division to occur[23,28]. There are a handful of reports that outline the existence of structures termed neocentromeres, or centromeres that form in areas of the genome outside of the canonical centromere that do not contain α-satellites but that can still recruit the epigenetic factors required for cell division[32–35]. Neocentromeres have been identified in neoplastic settings and are thought to be a purely pathologic finding. In the absence of pathology, the epigenetic factors important for cell division decorate the α-satellite rich regional centromeres. The factors are components that are critical to proper localization and assembly of the kinetochore[21]. The distinction between the genetic and epigenetic components of regional centromeres is thus central to this thesis, as both contribute to centromere function in humans and both genomic instability and epigenetic anomalies are observed in human cancer.

Centromere Genomics: A Wild Frontier

While it is widely acknowledged that the α-satellite rich human centromere plays a significant role in cell division, large gaps remain in our collective understanding of the precise mechanisms by which the human centromere carries out its function[22,23]. A full end-to-end assembly of each human centromere is to this day unavailable, with the exception of the centromere from chromosome Y[36]. These regions thus remain as one of the final uncharted frontiers within the human genome. Despite major advancements in next generation sequencing (NGS) technology, low-complexity regions characterized by repetitive sequences, including centromere α-satellite DNA, have generally been considered refractory to sequencing due to non-unique alignments that arise during computational assembly of these regions. Bioinformatics approaches that require a reference genome for alignment thus traditionally exclude these low-complexity regions during analysis.

Recent efforts at overcoming the technical shortcomings of NGS approaches have focused on more conventional molecular biology techniques, including extended chromatin fiber analysis, fluorescent in-situ hybridization (FISH), and Southern blotting-based approaches[22,37–43]. Chromatin fiber analysis, FISH, and Southern blotting, while effective for qualitatively and quantitatively characterizing localization and size of given centromeric proteins and sequences, are labor, resource, and time intensive. Polymerase chain-reaction (PCR) based approaches, such as those recently developed by Dr. Rafael Contreras-Galindo in our group, offer expedited evaluation of the centromeric content within any given sample, making it more scalable than chromatin fiber analysis and hybridization-based approaches when evaluating samples derived from human cell lines and tissue[44]. Corroboration of the specificity and sensitivity of PCR-approaches by a number of orthogonal methodologies, including FISH, Southern blotting, and

Sanger sequencing, suggests that using rapid centromere targeted PCR methodologies is a viable strategy for studying centromere genetics[39,44–46]. In this thesis, all genetic interrogation of centromeres relies upon PCR-based methodologies, unless otherwise denoted in the data.

Centromere Epigenetics

Given the difficulties associated with studying the genetics of human centromeres, efforts to study them have focused on the epigenetics that drive centromere assembly (**Fig 1.3**)[47,48]. The α-satellite sequences that define centromere DNA are primarily occupied by a centromere specific histone H3 variant known as CENPA[47,49]. CENPA is a highly conserved ~17 kDa molecule that forms a centromere-specific nucleosome with H2A, H2B and H4[50]. Proper CENPA localization is a ubiquitin E3 ligase dependent process, requiring ubiquitination of lysine 124 (K124) for engagement with the CENPA specific chaperone HJURP[51]. HJURP, along with the MIS18 complex subsequently facilitates the incorporation of newly synthesized CENPA into nucleosomes occupying replicated α-satellite DNA[52,53]. CENPA nucleosomes have a unique set of binding partners that facilitate proper genomic localization, including CENPB, CENPC, and the constitutive centromere associated network (CCAN) that comprises the inner kinetochore[54,55]. The CCAN serves as a multimeric interface between the DNA-enveloped CENPA nucleosomes and the KMN (KNL-1/Mis12 complex/Ndc80 complex) network that comprises the outer kinetochore and directly interacts with the microtubule spindle fibers[56]. CENPA and its associated proteins therefore represent structural components that are essential to the integrity of cell division, and appropriate genomic localization of centromeric proteins is consequently a critical event in the cell cycle.

Cancer: A Disease of Cell Division

The importance of cell division to human growth and development therefore establishes that derangements that occur during the process of cell division can give rise to diseases defined by abnormal cell growth and proliferation. Indeed, three of the hallmarks presented in Hanahan and Weinberg's landmark paper in 2000 (and their follow-up paper in 2011) implicate dysregulated cell division as a fundamental characteristic that defines cancer[26,57]. In 2018, there were 17 million new cases of cancer worldwide[58]. Cancer is the second leading cause of death globally, claiming 9.6 million lives in 2018 according to the World Health Organization (WHO). While the underlying reasons are complex, a steadily aging and growing population as well as changes in the prevalence and distribution of risk factors are implicated in the rising prevalence and mortality of cancer[59]. Multi-institutional endeavors that catalogue molecular and clinical irregularities within cancer tissue in the form of large publicly available databases have provided oncologists and cancer biologists a wealth of information to comprehensively interrogate this heterogeneous class of diseases[60,61]. Understanding cancer from the standpoints of cellular and molecular biology is imperative to identifying novel therapeutics that can curb the growing burden of disease.

Genomic Instability

One of the enabling characteristics of cancer outlined by Hanahan and Weinberg in 2011 is genomic instability and mutation. Of the ten hallmarks, many of them depend upon successive genomic alterations to ultimately produce neoplastic cells[26]. Indeed, many genes that are important for regulating the timing and structural definitions of cell division are subjected to mutational events that relieve checks on uncontrolled proliferation or produce additional mutations in the form of rearrangements or copy number amplification through aneuploidy. These genes include those that encode cyclins, cyclin dependent kinases, and DNA-replication/repair machinery[62–66]. Certain cancers also demonstrate hypermutability phenotypes, owing to mutational loads that accumulate due to dysfunctional DNA repair machinery or to chemical carcinogens that increase mutation rates in actively dividing cells[67–69]. Comparative genomic hybridization (CGH) of chromosomes from various cancer types has additionally unveiled gross copy number alterations in specific genomic loci and sometime of whole chromosomes themselves[70]. The advent of NGS technologies eventually paved the way for The Cancer Genome Atlas (TCGA) Program, a multi-institutional effort that has molecularly catalogued over 20,000 primary cancer tissues and matched normal samples from 33 cancer types[60]. Parsing through this massive database has provided high-resolution assessments of cancer genomes, corroborating much of what we had previously known about genomic instability in cancer while simultaneously providing novel insights into biological processes that maybe preferentially going awry in neoplastic cells relative to their normal counterparts.

Epigenetic Destabilization

Running parallel to the well-established notion that cancer genomes display widescale alterations that confer selective advantage are the discoveries of pathogenic alterations to the epigenetic landscape in cancer. Comprehensive evaluation of gene expression, transcription factor binding, histone modifications, DNA methylation, and histone variant occupancies have uncovered gross alterations in gene regulation and epigenetic function across several cancer types[71–74]. Transcription factors such as c-Myc, FOXM1, and the E2F family, which are important for proper timing and initiation of cell division, are almost universally implicated in oncogenic gene expression patterns[75–81]. Polycomb repressor complex 2 (PRC2) component abundance and the resultant increase in the H3K27me3 repressive mark are established markers of disease across cancer types[82,83]. Global DNA hypomethylation is considered a ubiquitous feature of carcinogenesis thought to alter gene expression profiles through destabilizing transcriptional repression in cancer[72,84]. Gene regulatory elements that are uncharacteristically occupied by promiscuous histone variants, such as H3.3 or macroH2A, have been shown to alter gene expression profiles towards those that are reflections of malignant phenotypes[85–88].

Derangements in Centromeres

Genetics

Diseases of cell division, particularly cancer, remain largely unexplored within the realm of centromere genetics[33,89–92]. Gaining deeper insight into the contribution of centromere genetics to tumorigenesis and cancer progression thus has the potential to inform novel therapeutic strategies capable of improving long-term outcomes. Unfortunately, the oncogenic potential of centromeric sequences remains undetermined, due to the shortcomings of sequencing methodologies. TCGA has largely overlooked these sequences when evaluating loci that are frequently altered in cancer, due to the issue of non-unique alignments that arise during bioinformatic analyses of sequencing data. However, one can postulate that repetitive loci, which are known to undergo recombination events that produce amplifications and deletions, may experience expansions and contractions due to the genomic instability that is an enabling feature in cancer[92–94]. Indeed, recombination in centromeres has been observed in maize, resulting in sequence evolution with every generation[95–97]. Chapter 2 of this thesis presents a foray into the wild frontier of centromeres in cancer as an effort to gain a better understanding of potentially pathological processes that occur within this largely overlooked, but important locus. Using the centromere specific PCR assay developed in-house, we report wide-scale alterations in centromeric content in both cancer cell lines and tissue.

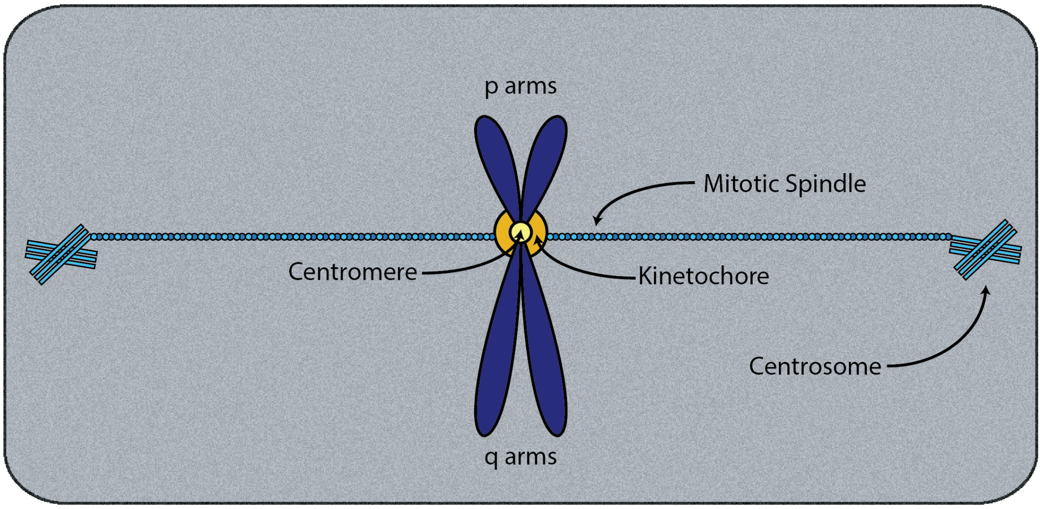
Epigenetics

Diseases of uncontrolled cell proliferation such as cancer, are compelling to examine from the epigenetic perspective of centromere biology. A number of studies have identified aberrant expression of centromeric/kinetochore proteins in cancers, where overexpression is predictive of survival and response to therapy, though their mechanistic contribution to cancer pathogenesis remains elusive[24,25,98,99]. In the setting of ectopic constitutive overexpression, CENPA, the histone H3 variant that epigenetically defines centromeric chromatin, has been shown to bind non-centromeric DNA. While CENPA localization normally requires engagement with HJRUP through a ubiquitin E3 ligase dependent process, CENPA promiscuity for non-centromeric DNA in the setting of ectopic overexpression is independent of aberrant E3 ligase activity and reliant upon the histone chaperone DAXX[100]. Ectopic localization of endogenously overexpressed CENPA has also been shown in colon cancer cell lines[101]. CENPA promiscuity thus presents another example of altered histone variant occupancy in cancer. The phenotypic consequences of such mislocalization in malignancy have yet to be elucidated, though ectopic binding to sites marked by DNase hypersensitivity and CTCF transcription factor affinity hint at a potential role in regulating gene transcription[100,101]. In Chapter 3 of this thesis, we aim to expand upon these intriguing findings to provide novel insight into the process of ectopic CENPA localization. Through the use of cell biology and integrated genomics approaches, we demonstrate a possible regulatory function in gene transcription for ectopically localized CENPA.

Conclusion

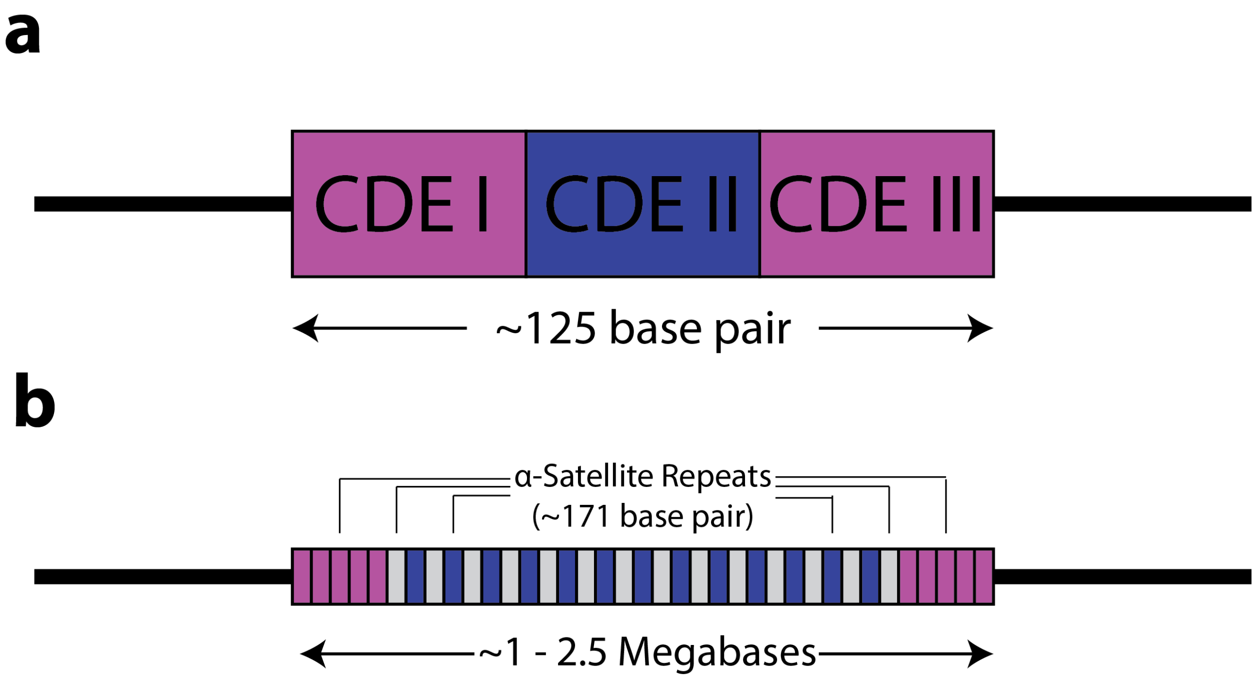
Centromeres represent essential structural components to cell division that ensure the faithful segregation of chromosomes during mitosis. Through interactions with the overlying kinetochore and mitotic spindle fibers, centromeres coordinate the cell division machinery to equally partition hereditary material from parent cells to each resulting daughter cell. Centromeres are thus compelling to study from both genetic and epigenetic perspectives in diseases of cell division such as cancer. Previous work has demonstrated that the epigenetic components of the centromere demonstrate uncharacteristic behavior in neoplastic settings, though the functional consequences of this behavior have yet to be ascertained. Unfortunately, technologic shortcomings limit a more comprehensive characterization of centromere genetics in cancer, and whether genetic aberrations may contribute to the promiscuous behavior of the epigenetic factors that define centromere chromatin. Novel and integrative methodologies are thus required to gain a more complete understanding of how centromeres behave during cancer pathogenesis and progression. We here demonstrate large scale genetic alterations in the centromeric locus in cancer cell lines and tissue. We further demonstrate that the epigenetic anomalies in centromeric components that arise in cancer, such as CENPA overexpression, have functional roles that affect gene regulation and transcription, using prostate cancer as our primary disease model. Collectively, this work aims to provide clarity to the roles that both the genetics and epigenetics of centromeres play in the setting of cancer, the most widely studied disease of cell division.

Figures



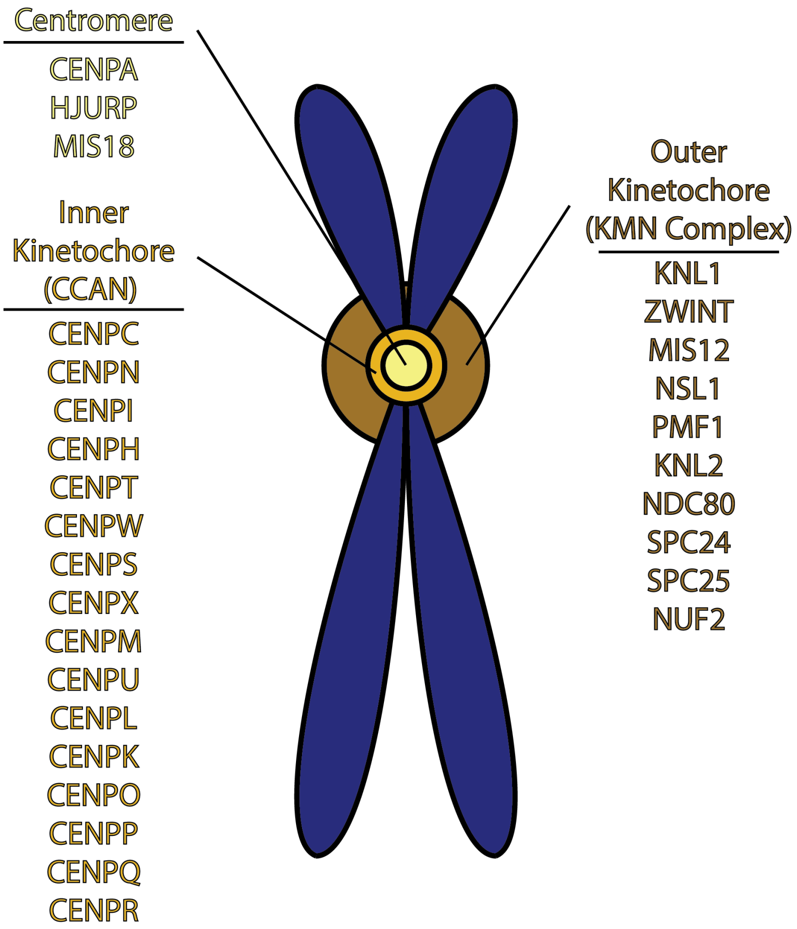
Authors: Anjan K. Saha, David M. Markovitz

Figure 1.1: Schematic depiction of the structural factors necessary for chromosomal segregation. Centrosomes produce mitotic spindle fibers that polymerize towards the kinetochore, a multimeric structure that assembles over the centromere.



Authors: Anjan K. Saha, David M. Markovitz

Figure 1.2: Schematic representation of point and regional centromeres. a) Point centromeres in budding yeast contain discreet consensus sequences within centromere DNA elements (CDE I – III) that recruit essential kinetochore proteins. The locus is occupied by a single CenH3 histone variant called Cse4 (the yeast equivalent of what is called CENPA in humans). b) Regional centromeres in humans are large loci composed of 171 base pair α-satellite DNA arranged in a head to tail fashion. Regional centromeres make up close to 5% of the human genome.



Authors: Anjan K. Saha, David M. Markovitz

Figure 1.3: Schematic illustration of the epigenetic compartmentalization of the centromeric locus in humans and the overlying kinetochore. Most proximal to the DNA on each chromosome is the centromeric H3 variant known as CENPA, along with its chaperone, HJURP, and recruitment machinery known as the MIS18 complex. The inner kinetochore, also known as the constitutive centromere associated network (CCAN), serves as the molecular interface between the centromere and the outer kinetochore (KMN complex) that directly interacts with the mitotic spindle fibers.

Tables

Table 1.1: Centromere position-dependent nomenclature of human chromosomes

Description	Definition	Examples
Telocentric	Centromere located near the telomere. The p-arm is barely visible if visible at all.	No examples in humans. Mouse chromosomes are characteristically telocentric.
Acrocentric	The p-arm of the chromosome is much shorter than the q-arm but longer than those in telocentric chromosomes.	Chromosomes 13, 14, 15, 21, 22, and Y
Submetacentric	The p arm and q arm are similar in length but not equal.	Chromosomes 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 17, 18, X
Metacentric	The p arm and q arm are equal in length.	Chromosomes 1, 3, 16, 19, 20

References

1. Hooke, R. Micrographia: Or, Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses. With Observations and Inquiries Thereupon. (J. Allestry, printer to the Royal Society, 1667).

2. Sharp, L. W. An Introduction To Cytology. (McGraw Hill Book Company Inc., 1921).

3. Kaliontzopoulou, A., Pinho, C. & Martínez-Freiría, F. Where does diversity come from? Linking geographical patterns of morphological, genetic, and environmental variation in wall lizards. BMC Evol. Biol. **18**, 124 (2018).

4. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. Biol. Rev. Camb. Philos. Soc. **81**, 425–455 (2006).

5. Golias, C. H., Charalabopoulos, A. & Charalabopoulos, K. Cell proliferation and cell cycle control: a mini review. Int. J. Clin. Pract. **58**, 1134–1141 (2004).

6. Moseley, J. B. & Nurse, P. Cell division intersects with cell geometry. Cell **142**, 184–188 (2010).

7. Graumann, P. L. Chromosome architecture and segregation in prokaryotic cells. J. Mol. Microbiol. Biotechnol. **24**, 291–300 (2014).

8. Pickett-Heaps, J. D., Tippit, D. H. & Porter, K. R. Rethinking mitosis. Cell **29**, 729–744 (1982).

9. Ohkura, H. Meiosis: an overview of key differences from mitosis. Cold Spring Harb. Perspect. Biol. **7**, (2015).

10. Flemming, W. Zellsubstanz, kern und zelltheilung.. (Leipzig, F. C. W. Vogel, 1882).

11. Abbott, S. & Fairbanks, D. J. Experiments on Plant Hybrids by Gregor Mendel. Genetics **204**, 407–422 (2016).

12. Mendel, G. Experiments in Plant Hybridisation. (Cosimo, Inc., 2008).

13. Donaldson, J. G. Immunofluorescence staining. Curr. Protoc. Cell Biol. **Chapter 4**, Unit 4.3 (2001).

14. Coons, A. H. Immunofluorescence. Public Health Rep. Wash. DC 1896 **75**, 937–943 (1960).

15. Schafer, K. A. The cell cycle: a review. Vet. Pathol. **35**, 461–478 (1998).

16. Maller, J. l et al. Maturation-promoting factor and the regulation of the cell cycle. J. Cell Sci. Suppl. **12**, 53–63 (1989).

17. Gautier, J. et al. Cyclin is a component of maturation-promoting factor from Xenopus. Cell **60**, 487–494 (1990).

18. Malumbres, M. & Barbacid, M. Cell cycle, CDKs and cancer: a changing paradigm. Nat. Rev. Cancer **9**, 153–166 (2009).

19. Conduit, P. T., Wainman, A. & Raff, J. W. Centrosome function and assembly in animal cells. Nat. Rev. Mol. Cell Biol. **16**, 611–624 (2015).

20. Prosser, S. L. & Pelletier, L. Mitotic spindle assembly in animal cells: a fine balancing act. Nat. Rev. Mol. Cell Biol. **18**, 187–201 (2017).

21. Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. Cell **112**, 407–421 (2003).

22. Aldrup-Macdonald, M. E. & Sullivan, B. A. The past, present, and future of human centromere genomics. Genes **5**, 33–50 (2014).

23. Hayden, K. E. Human centromere genomics: now it's personal. Chromosome Res. **20**, 621–633 (2012).
24. Zhang, W. et al. Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. Nat. Commun. **7**, 12619 (2016).
25. Valdivia, M., Hamdouch, K., Ortiz, M. & Astola, A. CENPA a Genomic Marker for Centromere Activity and Human Diseases. Curr. Genomics **10**, 326–335 (2009).
26. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. Cell **144**, 646–674 (2011).
27. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. Nat. Rev. Mol. Cell Biol. **17**, 16–29 (2016).
28. Rosin, L. F. & Mellone, B. G. Centromeres Drive a Hard Bargain. Trends Genet. TIG **33**, 101–117 (2017).
29. Clarke, L. & Carbon, J. Isolation of a yeast centromere and construction of functional small circular chromosomes. Nature **287**, 504–509 (1980).
30. Bloom, K. Anniversary of the discovery/isolation of the yeast centromere by Clarke and Carbon. Mol. Biol. Cell **26**, 1575–1577 (2015).
31. Clarke, L., Amstutz, H., Fishel, B. & Carbon, J. Analysis of centromeric DNA in the fission yeast Schizosaccharomyces pombe. Proc. Natl. Acad. Sci. U. S. A. **83**, 8253–8257 (1986).
32. Burrack, L. S. & Berman, J. Neocentromeres and epigenetically inherited features of centromeres. Chromosome Res. **20**, 607–619 (2012).
33. Vig, B. K., Sternes, K. L. & Paweletz, N. Centromere structure and function in neoplasia. Cancer Genet. Cytogenet. **43**, 151–178 (1989).
34. Chueh, A. C., Northrop, E. L., Brettingham-Moore, K. H., Choo, K. H. A. & Wong, L. H. LINE Retrotransposon RNA Is an Essential Structural and Functional Epigenetic Component of a Core Neocentromeric Chromatin. PLoS Genet. **5**, e1000354 (2009).
35. Marshall, O. J., Chueh, A. C., Wong, L. H. & Choo, K. H. A. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am. J. Hum. Genet. **82**, 261–282 (2008).
36. Jain, M. et al. Linear assembly of a human centromere on the Y chromosome. Nat. Biotechnol. **36**, 321–323 (2018).
37. Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K. & Sullivan, B. A. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. Genome Res. **26**, 1301–1311 (2016).
38. Li, X. et al. A fluorescence in situ hybridization (FISH) analysis with centromere-specific DNA probes of chromosomes 3 and 17 in pleomorphic adenomas and adenoid cystic carcinomas. J. Oral Pathol. Med. Off. Publ. Int. Assoc. Oral Pathol. Am. Acad. Oral Pathol. **24**, 398–401 (1995).
39. Liehr, T. Benign and Pathological Chromosomal Imbalances: Microscopic and Submicroscopic Copy Number Variations (CNVs) in Genetics and Counseling. (Academic Press, 2013).
40. Quénet, D. & Dalal, Y. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. eLife **3**, (2014).
41. Jabs, E. W., Goble, C. A. & Cutting, G. R. Macromolecular organization of human centromeric regions reveals high-frequency, polymorphic macro DNA repeats. Proc. Natl. Acad. Sci. U. S. A. **86**, 202–206 (1989).

42. Zahn, J. et al. Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans. Genome Biol. **16**, 74 (2015).

43. Du, Y., Topp, C. N. & Dawe, R. K. DNA Binding of Centromere Protein C (CENPC) Is Stabilized by Single-Stranded RNA. PLoS Genet. **6**, e1000835 (2010).

44. Contreras-Galindo, R. et al. Rapid molecular assays to study human centromere genomics. Genome Res. **27**, 2040–2049 (2017).

45. Shepelev, V. A. et al. Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. Genomics Data **5**, 139–146 (2015).

46. Miga, K. H. et al. Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res. **24**, 697–707 (2014).

47. Hayashi, T. et al. Mis16 and Mis18 Are Required for CENP-A Loading and Histone Deacetylation at Centromeres. Cell **118**, 715–729 (2004).

48. Kim, I. S. et al. Roles of Mis18α in Epigenetic Regulation of Centromeric Chromatin and CENP-A Loading. Mol. Cell **46**, 260–273 (2012).

49. Foltz, D. R. et al. The human CENP-A centromeric nucleosome-associated complex. Nat. Cell Biol. **8**, 458–469 (2006).

50. Hasson, D. et al. The octamer is the major form of CENP-A nucleosomes at human centromeres. Nat. Struct. Mol. Biol. **20**, 687–695 (2013).

51. Niikura, Y. et al. CENP-A K124 Ubiquitylation Is Required for CENP-A Deposition at the Centromere. Dev. Cell **32**, 589–603 (2015).

52. Barnhart, M. C. et al. HJURP is a CENP-A chromatin assembly factor sufficient to form a functional de novo kinetochore. J. Cell Biol. **194**, 229–243 (2011).

53. Foltz, D. R. et al. Centromere-Specific Assembly of CENP-A Nucleosomes Is Mediated by HJURP. Cell **137**, 472–484 (2009).

54. Falk, S. J. et al. CENP-C directs a structural transition of CENP-A nucleosomes mainly through sliding of DNA gyres. Nat. Struct. Mol. Biol. **23**, 204–208 (2016).

55. McKinley, K. L. et al. The CENP-L-N Complex Forms a Critical Node in an Integrated Meshwork of Interactions at the Centromere-Kinetochore Interface. Mol. Cell **60**, 886–898 (2015).

56. Schleiffer, A. et al. CENP-T proteins are conserved centromere receptors of the Ndc80 complex. Nat. Cell Biol. **14**, 604–613 (2012).

57. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. Cell **100**, 57–70 (2000).

58. Global Cancer Facts & Figures | American Cancer Society. Available at: https://www.cancer.org/research/cancer-facts-statistics/global.html. (Accessed: 21st May 2019)

59. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA. Cancer J. Clin. **68**, 394–424 (2018).

60. The Cancer Genome Atlas. National Cancer Institute (2018). Available at: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. (Accessed: 21st May 2019)

61. International Cancer Genome Consortium. Available at: https://icgc.org/. (Accessed: 21st May 2019)

62. Deshpande, A., Sicinski, P. & Hinds, P. W. Cyclins and cdks in development and cancer: a perspective. Oncogene **24**, 2909–2915 (2005).

63. Casimiro, M. C., Crosariol, M., Loro, E., Li, Z. & Pestell, R. G. Cyclins and cell cycle control in cancer and disease. Genes Cancer **3**, 649–657 (2012).

64. Curtin, N. J. DNA repair dysregulation from cancer driver to therapeutic target. Nat. Rev. Cancer **12**, 801–817 (2012).
65. Gavande, N. S. et al. DNA repair targeted therapy: The past or future of cancer treatment? Pharmacol. Ther. **160**, 65–83 (2016).
66. Torgovnick, A. & Schumacher, B. DNA repair mechanisms in cancer development and therapy. Front. Genet. **6**, 157 (2015).
67. Parsons, R. et al. Hypermutability and mismatch repair deficiency in RER+ tumor cells. Cell **75**, 1227–1236 (1993).
68. Strauss, B. S. Hypermutability in carcinogenesis. Genetics **148**, 1619–1626 (1998).
69. Chae, Y. K. et al. Genomic landscape of DNA repair genes in cancer. Oncotarget **7**, 23312–23321 (2016).
70. Houldsworth, J. & Chaganti, R. S. Comparative genomic hybridization: an overview. Am. J. Pathol. **145**, 1253–1260 (1994).
71. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. Science **362**, (2018).
72. Koch, A. et al. Analysis of DNA methylation in cancer: location revisited. Nat. Rev. Clin. Oncol. **15**, 459–466 (2018).
73. Dawson, M. A. The cancer epigenome: Concepts, challenges, and therapeutic opportunities. Science **355**, 1147–1152 (2017).
74. Jones, P. A., Issa, J.-P. J. & Baylin, S. Targeting the cancer epigenome for therapy. Nat. Rev. Genet. **17**, 630–641 (2016).
75. Chen, H., Liu, H. & Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. Signal Transduct. Target. Ther. **3**, 5 (2018).
76. Dang, C. V. MYC on the Path to Cancer. Cell **149**, 22–35 (2012).
77. Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. Nat. Rev. Cancer **8**, 976–990 (2008).
78. Kent, L. N. & Leone, G. The broken cycle: E2F dysfunction in cancer. Nat. Rev. Cancer (2019). doi:10.1038/s41568-019-0143-7
79. Polager, S. & Ginsberg, D. p53 and E2f: partners in life and death. Nat. Rev. Cancer **9**, 738–748 (2009).
80. Raychaudhuri, P. & Park, H. J. FoxM1: a master regulator of tumor metastasis. Cancer Res. **71**, 4329–4333 (2011).
81. Myatt, S. S. & Lam, E. W.-F. The emerging roles of forkhead box (Fox) proteins in cancer. Nat. Rev. Cancer **7**, 847–859 (2007).
82. Comet, I., Riising, E. M., Leblanc, B. & Helin, K. Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. Nat. Rev. Cancer **16**, 803–810 (2016).
83. Kim, K. H. & Roberts, C. W. M. Targeting EZH2 in cancer. Nat. Med. **22**, 128–134 (2016).
84. Klutstein, M., Nejman, D., Greenfield, R. & Cedar, H. DNA Methylation in Cancer and Aging. Cancer Res. **76**, 3446–3450 (2016).
85. Buschbeck, M. & Hake, S. B. Variants of core histones and their roles in cell fate decisions, development and cancer. Nat. Rev. Mol. Cell Biol. **18**, 299–314 (2017).
86. Vardabasso, C. et al. Histone variants: emerging players in cancer biology. Cell. Mol. Life Sci. CMLS **71**, 379–404 (2014).
87. Vardabasso, C. et al. Histone Variant H2A.Z.2 Mediates Proliferation and Drug Sensitivity of Malignant Melanoma. Mol. Cell **59**, 75–88 (2015).

88. Yuen, B. T. K. & Knoepfler, P. S. Histone H3.3 Mutations: A Variant Path to Cancer. Cancer Cell **24**, 567–574 (2013).

89. Black, E. M. & Giunta, S. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. Genes **9**, (2018).

90. Bersani, F. et al. Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. Proc. Natl. Acad. Sci. U. S. A. **112**, 15148–15153 (2015).

91. Natisvili, T. et al. Transcriptional Activation of Pericentromeric Satellite Repeats and Disruption of Centromeric Clustering upon Proteasome Inhibition. PloS One **11**, e0165873 (2016).

92. Giunta, S. & Funabiki, H. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. Proc. Natl. Acad. Sci. U. S. A. **114**, 1928–1933 (2017).

93. Klein, H. L. Genetic control of intrachromosomal recombination. BioEssays News Rev. Mol. Cell. Dev. Biol. **17**, 147–159 (1995).

94. Blanco, P. et al. Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J. Med. Genet. **37**, 752–758 (2000).

95. Schneider, K. L., Xie, Z., Wolfgruber, T. K. & Presting, G. G. Inbreeding drives maize centromere evolution. Proc. Natl. Acad. Sci. U. S. A. **113**, E987-996 (2016).

96. Shi, J. et al. Widespread gene conversion in centromere cores. PLoS Biol. **8**, e1000327 (2010).

97. Wolfgruber, T. K. et al. High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events. Front. Plant Sci. **7**, 308 (2016).

98. Thiru, P. et al. Kinetochore genes are coordinately up-regulated in human tumors as part of a FoxM1-related cell division program. Mol. Biol. Cell **25**, 1983–1994 (2014).

99. Tomonaga, T. et al. Overexpression and mistargeting of centromere protein-A in human primary colorectal cancer. Cancer Res. **63**, 3511–3516 (2003).

100. Lacoste, N. et al. Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. Mol. Cell **53**, 631–644 (2014).

101. Athwal, R. K. et al. CENP-A nucleosomes localize to transcription factor hotspots and subtelomeric sites in human cancer cells. Epigenetics Chromatin **8**, 2 (2015).

Chapter 2 – The Genomic Landscape of Centromeres in Cancers

Summary

Centromere genomics remain poorly characterized in cancer, due to technologic limitations in sequencing and bioinformatics methodologies that make high-resolution delineation of centromeric loci difficult to achieve. We here leverage a highly specific and targeted rapid PCR methodology to quantitatively assess the genomic landscape of centromeres in cancer cell lines and primary tissue. PCR-based profiling of centromeres revealed widespread heterogeneity of centromeric and pericentromeric sequences in cancer cells and tissues as compared to healthy counterparts. Quantitative reductions in centromeric core and pericentromeric markers (α-satellite units and HERV-K copies) were observed in neoplastic samples as compared to healthy counterparts. Subsequent phylogenetic analysis of a pericentromeric endogenous retrovirus amplified by PCR revealed possible gene conversion events occurring at numerous pericentromeric loci in the setting of malignancy. Our findings collectively represent a more comprehensive evaluation of centromere genetics in the setting of malignancy, providing valuable insight into the evolution and reshuffling of centromeric sequences in cancer development and progression.

Introduction

The centromere is essential to eukaryotic biology due to its critical role in genome inheritance[1,2]. The nucleic acid sequences that dominate the human centromeric landscape are α-satellites, arrays of ~171 base-pair monomeric units arranged into higher-order arrays throughout the centromere of each chromosome[1–3]. These α-satellites underlie a hierarchical network of proteins that collectively make up the kinetochore, a large multimeric structure that serves as a molecular bridge between chromosomes and microtubule polymers from the mitotic spindle during cell division. The interaction between centromeres, kinetochores and microtubule polymers lies at the nexus of metaphase and anaphase, ensuring faithful separation of the sister chromatids during mitosis.

Centromeres are thus critical to maintaining the fidelity of chromosomal segregation in proliferating tissues. While much is known about the hierarchical network of proteins that epigenetically compartmentalizes centromeres, the genomic foundation of the centromere remains largely uncharted. Centromeres remain a genetic black box that encompasses 2-5% of the human genome[4]. Despite advancements in next-generation sequencing (NGS) technologies, full assemblies of centromeric loci are still unavailable within the latest builds of the human genome, with the exception of a linear assembly of the centromere of chromosome Y[5]. Low complexity genomic regions, characterized by the contiguous arrangement of repetitive sequences, present computational challenges owing to nonunique alignments that are impractical for current informatics pipelines to navigate. Low complexity regions like centromeric loci are consequently excluded from most downstream bioinformatics analyses.

Methodologies that can add resolution to the genomic landscape of the centromere will thus play an integral role in developing a more nuanced understanding of its contribution to health and disease. Efforts to overcome the technical shortcomings of NGS approaches have focused on more conventional molecular biology techniques, including extended chromatin fiber analysis, fluorescent *in-situ* hybridization (FISH), Southern blotting, and polymerase chain-reaction (PCR) based approaches[4,6–13]. While effective for qualitatively and quantitatively characterizing localization and size of centromeric proteins and sequences, these methods are labor, resource, and time intensive. Diseases of cell division, particularly cancer, remain largely unexplored within the realm of centromere genetics, due to the lack of scalable alternatives to sequencing technology[14–18]. PCR-based approaches, corroborated for specificity and sensitivity, offer expedited evaluation of the centromeric content within any given sample, making it more scalable than chromatin fiber analysis and hybridization-based approaches when evaluating samples derived from human cell lines and tissue[8,9,19,20]. Applying scalable PCR-based approaches to the assessment of centromere size and structure in biological settings like cancer is therefore critical to contextualizing our knowledgebase on centromere genetics.

Here we report substantial heterogeneity in the centromeric landscape in cancer cell lines and tissues, in terms of copy number differences between tissues as well as differences between cancer cells/tissues and healthy cells. Both solid and hematologic tumors demonstrated marked copy number alterations in centromeric and pericentromeric repeats, as measured by a previously described quantitative centromere-specific PCR assay that targets core centromeric α-satellite DNA as well as pericentromeric human endogenous retrovirus (HERV) DNA[9]. Phylogenetic analysis of HERV sequences in several cancer cell lines suggests that pericentromeric sequences undergo aberrant recombination during tumorigenesis and/or disease progression, consistent with derangements that have been previous reported[12,18,21,22]. Strikingly, centromeric variation is a feature present across cancer tissue types, including primary tissue samples, providing further substantiation to the notion that genomic instability in centromeres is a ubiquitous occurrence in cancer. Evaluation of the centromeric landscape in the setting of malignancy thus reveals marked genetic alterations that may reflect novel pathophysiologic contributions to the development and progression of cancer.

Results

Cancer Cell Lines Demonstrate Heterogeneous Alterations in Centromeric and Pericentromeric DNA

NGS approaches to interrogating genetic alterations in cancer have repeatedly demonstrated ubiquitous genomic instability that is a hallmark of malignancy. However, the lack of an end-to-end assembly of centromeric loci prevents mapping of representative centromeric reads to a standardized reference. We have thus employed a rapid PCR-based approach that we previously described to evaluate the genomic landscape of centromeres and pericentromeres in several human cancer cell lines (**Figure 2.1**). The method was previously validated by comparison to meta-analyses of data from studies using NGS and southern blot, as well as through FISH analysis[9]. The cell lines studied here are representative of a variety of different tissue types, originating from both solid and hematologic malignancies. Our PCR-based methodology unveils significant heterogeneity in the centromeric and pericentromeric content in all 24 chromosomes across tissue types and as compared to healthy cells. This heterogeneity extends to HERVs, such as HERV K111, that we have previously shown to reside in pericentromeric regions. Unsupervised hierarchical clustering of the chromosome specific repeats demonstrates a striking organization to the patterns in centromere heterogeneity, differentiated by the region of the centromere (core or pericentromere) to which each repeat localizes. Similar clustering analysis applied across the different cell lines revealed that heterogeneity in centromeric and pericentromeric content is tissue type agnostic, with the exception of healthy peripheral blood lymphocytes (PBLs) that demonstrate higher relative concordance. The heterogeneity observed reflects a preference for contractions in centromeric and pericentromeric content, consistent across numerous tissue types (**Figure 2.2 – 2.6**). More specifically, D13Z1, D10Z1, D2Z1, D3Z1, D8Z2, D16Z2, and K111 demonstrated the most appreciable losses when collectively assessing all tested cancer cell lines. The nomenclature of these α-satellites begins with the letter D, followed by their resident chromosome number (1–22, X or Y), followed by a Z, and a number indicating the order in which the sequence was discovered. Consistent with the global loss of whole chromosomes previously reported in teratocarcinoma cells, we noted widescale loss of centromere arrays in teratocarinoma cell lines derived from male patients in this study (**Figure 2.6**)[23–26]. Of note, K111 deletion stood out as ubiquitous across all evaluated cell lines. Collectively comparing normal peripheral blood mononuclear cells (PBMCs) to cancer cell lines, grouped by tissue type, revealed marked reductions in pericentromeric material, using K111 copy number as a surrogate for pericentromeric content (**Figure 2.7**)[12,27].

A more focused analysis on breast cancer cell lines allowed us to cross-reference the observed heterogeneity in centromeric DNA against known molecular classifications and karyotypes for each cell line to ascertain whether centromeric and pericentromeric deletions were the result of previously described genetic derangements, such as recurrent molecular alterations or whole chromosome copy number loss, as seen in teratocarcinoma cell lines (**Figure 2.8**)[28–33]. Strikingly, the centromeric content demonstrated heterogeneity across the four molecular subtypes for breast cancer (Basal, HER2, Luminal A, and Luminal B); unsurprisingly, healthy PBLs clustered together. Similar to other tissue types tested, breast cancer cell lines also demonstrated a predilection for contracted centromeres and pericentromeres compared to healthy PBLs (**Figure 2.9**). While contraction of D13Z1 in Hs578T, BT474, and MDA-MB-361 can be attributed to loss of whole chromosome 13, contraction of D8Z2 in T47D, D3Z1 and D8Z2 in BT549, and D8Z2 and D10Z1 in SKBr3 were observed despite well characterized copy number amplifications of the respective chromosomes. K111 again demonstrated robust contractions relative to other markers. The strong reduction in DYZ3 (α-satellite on chromosome Y) to nearly undetectable levels provided validation for the specificity of the rapid PCR-based approach to evaluating centromeric content, given the absence of Y-chromosomes in breast cancer cell lines derived from females. Taken together, marked heterogeneity in centromeric and pericentromeric DNA is observed in cancer cell lines, with a predilection towards contraction when comparing cancer cell lines to healthy PBLs.

Gene Conversion of Pericentromeric HERV Sequences in Cancer Cell Lines

The genomic landscape of the centromere is characterized by thousands of copies of repetitive elements arranged in tandem to form higher order arrays[1]. Repetitive genomic regions are known to be subject to recombination due to sequence homology[18,34,35]. Intrachromosomal recombination is one example of repeat-associated recombination that can lead to either deletions that reduce the number of repeat units or gene conversion events that genetically homogenize the sequences of repeat units[36–38]. Interestingly, in contrast to healthy PBLs, we identified drastic reductions in pericentromeric K111 sequences across all evaluated cancer cell lines (**Figure 2.1 and 2.8**). While real-time PCR demonstrates deletion of centromeric and pericentromeric material in cancer cell lines, purely quantitative assessments do not provide insight into other recombination events, such as gene conversion. Furthermore, sequence analysis of α-satellites is unreliable for identifying gene conversion events. We thus conducted phylogenetic analysis on the sequences of real-time PCR amplicons from breast cancer cell lines to identify gene conversion events within

K111 loci, given ubiquitous loss of K111 across all cancer cell lines (**Figure 2.10a**). Our previous work has shown that divergence in K111 sequence similarity is dependent on chromosomal location of K111 loci[12,27]. We now show that K111 copies identified in breast cancer cell lines demonstrate cell line dependent sequence convergence towards K111 subtypes that organize into distinct clades (**Figure 2.10b**). The K151 cell line (pink) remarkably produced distinct clades that emerged in close proximity relative to each other from the same ancestral sequence. Sequences amplified from the K151 cell line were notably not distributed heterogeneously throughout the tree. Three additional breast cancer cell lines (MDA-MB-435, DT-13, and HCC1599) formed two exclusive subtypes that were also separated by phylogenetic analysis.

Phylogenetic analysis was also conducted in adult T-cell leukemia (ATL) cell lines and revealed similar patterns as in breast cancer (**Figure 2.11**). ATL26 alone formed three exclusive subtypes that diverge in homology from normal K111 clades. Of note, K111 clades arising from ATL43 and ATL16 demonstrated strong homology to K111 Solo LTRs, suggesting intrachromosomal recombination that has deleted K111, i.e. pericentromeric material. ATL43 and ATL16 indeed demonstrate the strongest reductions in K111 copy number relative to other ATL cell lines (**Figure 2.7**). As Solo LTRs are the result of homologous recombination between the LTRs flanking endogenous retroviral sequences[39–41], ATL cell lines having *de novo* K111 sequences with higher relative homology to Solo LTR sequences suggested that pericentromeric K111 sequences served as templates for gene conversion. Taken together, cell line dependent sequence convergence of HERV-K111 in cancer cell lines suggests that gene conversion events are driving sequence evolution within the pericentromeres of cancer cell lines.

Heterogeneous Loss of Centromere DNA in Cancer Tissue

Human cancer cell lines are useful models for evaluating cancer biology and genetics in an *in vitro* setting. Indefinite cellular propagation, however, results in clonal selection for cells that have a fitness advantage for growing *ex vivo*. Such a fitness advantage is sometimes conferred by abnormal karyotypes (aneuploidy), a cytogenetic feature that can influence the results of PCR based analyses. Cancer tissue itself thus presents the most accurate representation of malignancy-associated genomic instability that results from microenvironmental pressures that cannot be reproduced *ex vivo*. We thus applied our rapid PCR-based approach to DNA isolated from primary cancer tissue. Profiling the centromeric landscape in 9 different ovarian cancer samples against matched PBMCs revealed similarly significant loss of α-satellites across multiple chromosomes as observed in cell lines (**Figure 2.12**). Indeed, quantitative assessment of this heterogeneity again revealed copy number reductions in the cancer tissue, similar to findings noted in cell lines (**Figure 2.13**). Strikingly, a drastic reduction in the centromere of chromosome 17 (D17Z1) was seen in ovarian cancer tissue when compared to healthy tissue (**Figure 2.13**), corroborating previous reports of chromosome 17 anomalies in ovarian cancer[42]. No changes were seen in the single copy gene *GAPDH* found in the arm of chromosome 12. A significant loss in GAPDH is, however, noted in Sample 285, raising the possibility that this sample's karyotype displayed derangements that are reflected in the PCR data. Tumor karyotypes for tested samples were, however, unavailable for corroboration.

While matched blood samples provide reliable non-malignant references to their malignant counterparts, comparisons between primary ovarian cancer tissue and matched blood does not sufficiently deconvolute tissue specific genetic heterogeneity that may be present in normal biologic settings. To expand upon our findings, and to specifically address this latter issue, we profiled the centromeres of B-cells and T-cells that were separated by cell-surface marker selection from chronic lymphocytic leukemia (CLL) primary samples. CLL is a malignancy that arises in B-cells, as opposed to T-cells, within the bone marrow. Applying our methodology to compare patient matched B-cells and T-cells from CLL samples, both cells of lymphocytic lineage, thus largely eliminates the confounding contributions of normal development and tissue specificity to genetic heterogeneity in the centromere. Fewer repeats per sample were evaluated than in the experiments described above due to the limited availability of tumor DNA from each patient. Intriguingly, unsupervised hierarchical cluster analysis across patient samples cleanly separates healthy cells from diseased cells based on chromosome specific α-satellite abundance (**Figure 2.14**). We show contraction of numerous centromeres in malignant CD19+ B-cells as compared to their normal CD3+ T-cell counterparts, whereas no changes were seen in the housekeeping gene *GAPDH* found in the arm of chromosome 12 (**Figure 2.15**). Strikingly, we see no such centromeric differences between B-cells and T-cells separated from blood samples derived from healthy individuals. Taken together, centromeric contraction is a characteristic that is present in primary cancer samples, consistent with our data in cancer cell lines.

Discussion

The importance of centromeres to cell division provides a strong rationale for interrogating the genetics of the centromere in cancers. The challenges associated with studying the genomic landscape of centromeres, owing to the informatics impracticalities of evaluating low complexity regions, have however hindered meaningful progress in understanding the contributions of centromere genetics to tumorigenesis and cancer progression. Only one previous study reported the loss of centromere DNA in leukemia cells using fluorescent in situ hybridization (FISH)[43]. We demonstrate, for the first time, that centromeres and pericentromeres display heterogeneous alterations in the setting of malignancy, both in cancer cell lines and primary samples. We show that these heterogeneous alterations reflect marked reductions and gene conversions of repetitive elements and HERVs in multiple centromeres and pericentromeres, suggesting that oncogenic genomic instability selects against the presence of most centromeric sequences and perhaps for certain pericentromeric sequences. While mechanistically uncharacterized, these findings have direct implications for our understanding of global genomic instability in cancer, given the importance of centromeres to faithful segregation of chromosomes. The loss of centromeric material in chromosome 17 described above is an example of the concordance between centromere instability and ovarian cancer pathogenesis, given the recurrent alterations in chromosome 17 that have been previously described in ovarian cancer[42]. While in some cases loss of centromeric DNA could be attributed to a loss of that entire chromosome, there is also a substantial loss of centromeric DNA in specific chromosomes that are known to be euploid or even polyploid in a given cancer cell line. Further, we have shown previously that DNA from patients with trisomy 13 and trisomy 21 exhibit loss of pericentromeric K111 and that DNA from patients with trisomy 21 exhibit loss of D21Z1, suggesting that pericentromeric and centromeric contraction may drive mis-segregation of chromosomes 13 and 21[9]. It is thus conceivable that alterations in centromeres and pericentromeres may underlie chromosome segregation defects that are routinely observed in the context of abnormal cell proliferation. Gaining deeper insight into the mechanism driving gene conversion and centromere contraction may facilitate the identification of novel molecular drivers that can be targeted to prevent potentially oncogenic mis-segregation events.

While the genetics of centromeres in cancer continue to be elucidated, there is a body of work that has uncovered dysregulation of centromere epigenetics and transcriptional activity in malignancy. Overexpression of CENPA is observed ubiquitously across various cancers, with evidence of ectopic CENPA deposition at extra-centromeric loci across the human genome[44–47]. Satellite RNA abundance is an additional feature that been identified in cell lines and tissue[48–50]. Our findings of genomic contraction of centromeres provides a topographic rationale for the redirection of unbound CENPA to readily accessible ectopic loci in the setting of CENPA overexpression, though additional work is required to distinguish the role of cancer specific post-translational modifications in ectopic deposition of CENPA[51,52]. Moreover, while not mechanistically validated, regions that repress transcriptional homeostasis within centromeric loci may be lost (but beyond the sensitivity of PCR interrogation) during genomic contraction of centromeres and pericentromeres in cancer, thus driving transcriptional activity and overexpression of satellite RNAs in malignancy. Indeed, DNA methylation, an epigenetic mark of transcriptional repression, is prevalent within centromeric loci[53,54]. Selective deletion of methylated regions in centromeres during cancer pathogenesis may relieve transcriptional repression, resulting in overexpression of satellite RNAs. Cancer specific examination of DNA-methylation at the centromeric region that leverages our PCR methodology will be essential to validating this line of reasoning.

Instability in centromeric and pericentromeric loci in the setting of malignancy is consistent with the global genome instability that is a well characterized hallmark of cancer[55]. Subsets of breast and ovarian cancer have well studied DNA repair aberrations in homologous recombination proteins BRCA1/2[56]. Recent genomic profiling of several other malignancies has identified new disease subsets classified by molecular alterations in DNA repair genes and pathways[57,58]. It is conceivable that subsets of cancer that are dysfunctional in DNA repair may exhibit pronounced heterogeneity in centromeric content. Thus, it must be acknowledged that hypermutability in DNA that results from DNA repair dysfunction in cancer may alter centromere and pericentromere sequences enough to prevent detection by PCR, appearing like copy number loss or gene conversion in phylogenetic analysis instead of mismatches or single nucleotide polymorphisms (SNPs). Stratifying samples by DNA repair signatures prior to profiling the genomic landscape of centromeres may provide a strategy for identifying mechanistic contributors to centromere contraction in the setting of malignancy. Moreover, genomic profiling of centromeres in cancer tissue may produce signatures that are predictive of responders to therapies that target the DNA repair machinery, such as poly-ADP ribose polymerase (PARP) inhibitors.

In conclusion, we here provide quantitative resolution of the largely uncharacterized human centromere in the setting of cancer. We notably shed light on a region that has been widely considered a black box and impervious to rapid and comprehensive inquiry at the genomic level. The wide-spread alterations observed in cancer cell lines and primary tissue provide a sound rationale to mechanistically interrogate the molecular machinery that is likely driving the selection against centromeric material. Mechanistic characterization of genomic instability at centromeric loci has the potential to inform therapeutic approaches aimed at improving disease outcomes across several cancer types.

Materials and Methods

Cell Lines and Cell Culture. Cell lines were cultured according to American Type Culture Collection (ATCC) recommendations. Cell lines were grown at 37 °C in a 5% CO2 cell culture incubator and authenticated by short tandem repeat (STR) profiling for genotype validation at the University of Michigan Sequencing Core. ATL cell lines were cultured and authenticated as previously described[59].

DNA Isolation. DNA extraction was performed on cell lines and tissue with the DNeasy Blood and Tissue Kit (QIAGEN) according to manufacturer's instructions. DNA was preserved at -20° C.

Blood and Tumor Cell Separation. Between January 2005 and September 2016 patients with chronic lymphocytic leukemia (CLL) evaluated at the University of Michigan Comprehensive Cancer Center were enrolled onto this study. The trial was approved by the University of Michigan Institutional Review Board (IRB no. HUM00045507). Patients consented for tissue donation in accordance with a protocol approved by the University of Michigan's IRB (IRB no. HUM0009149). Written informed consent was obtained from all patients before enrollment in accordance with the Declaration of Helsinki. CLL diagnostic criteria were based on the National Cancer Institute Working Group Guidelines for CLL. Eligible patients needed to have an absolute lymphocytosis (> 5000 mature lymphocytes/µL), and lymphocytes needed to express CD19, CD23, sIg (weak), and CD5 in the absence of other pan-T-cell markers. Peripheral blood mononuclear cells (PBMCs) were isolated by venipuncture and separated using Histopaque-1077 (Sigma). Cryopreserved PBMCs (frozen after Ficoll-gradient purification) from CLL blood specimens were prepared for FACS and sorted into CD19+ (B-cells) and CD3+ (T-cells) cells as previously described[60]. Ovarian cancer DNA were isolated from Stage IIIc or Stage IV ovarian carcinomas. Tumor samples were obtained from the operating room and immediately taken to the laboratory for processing. Tissue was maintained in RPMI/10% FBS throughout processing. Fresh 4 × 4 × 2–mm tumor slices were rinsed several times to remove all loosely attached cells. The tissue was then placed in a tissue culture dish and DNA was extracted as described above.

Rapid Centromere Target PCR Assay. PCR was conducted on DNA samples from cell lines and primary cancer samples according the previously described conditions[9]. Briefly, copy numbers for each centromeric array, proviruses K111/K222, and single-copy genes were measured by qPCR using specific primers and PCR conditions as described. PCR amplification products were confirmed by sequencing. The qPCR was carried out using the Radiant Green Low-Rox qPCR master mix (Alkali Scientific) with an initial enzyme activation step for 10 min at 95°C and 16–25 cycles consisting of 15 sec of denaturation at 95°C and 30 sec of annealing/extension.

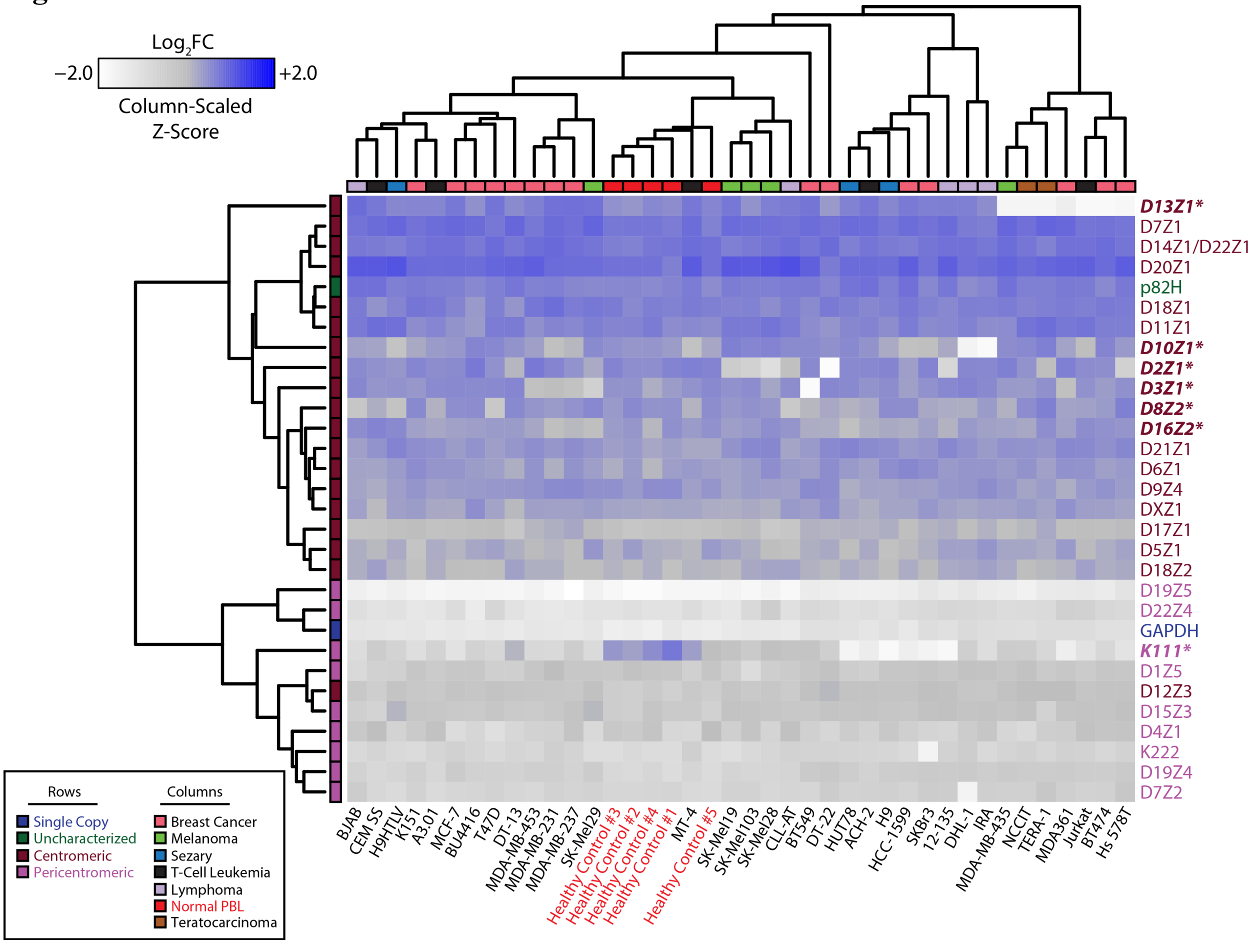
PCR for 5' and 3' K111 LTR Insertions. K111 insertions were amplified by PCR using the Expand Long Range dNTPack PCR kit (Roche Applied Science, Indianapolis, IN) as described. K111 5' and 3' LTRs and accompanying flanking regions were amplified. PCR was performed using an initial step of 94 °C for 2 min followed by 35 cycles consisting of denaturation at 94 °C for 30 sec, annealing at 55 °C for 30 sec, and extension at 68 °C for 5 min. The amplification products were cloned into the topo TA vector (Invitrogen, Carlsbad, CA) and sequenced.

Phylogenetic Analysis. Analysis was conducted as outlined previously[61]. The K111-related LTR sequences obtained from the DNA of cell lines, and DNA from human/rodent chromosomal cell hybrids were subjected to BLAST analysis against the NCBI nucleotide database. Sequences were aligned in BioEdit and exported to the MEGA5 matrix. LTR trees were generated using Bayesian inference (MrBayes v 3.2[62]) with four independent chains run for at least 1,000,000 generations until sufficient trees were sampled to generate more than 99% credibility.

Statistics and Data Analysis. All heatmaps were generated using the gplots, RColorBrewer, and plotrix packages within the RStudio integrated development environment for the R statistical programming language. Data were $\log_2$ normalized to the median values of healthy samples. Tests of statistical significance employed two-sided student t-tests, with level of significance denoted on appropriate plots.

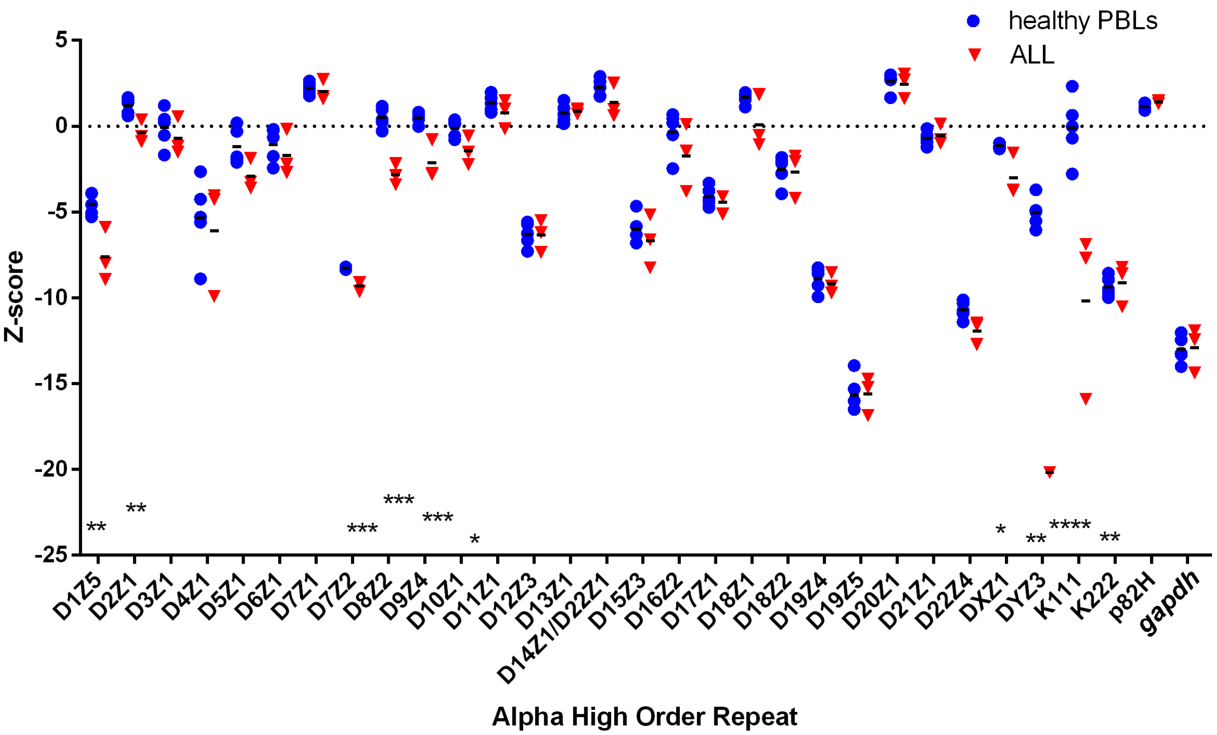
Data Availability. Sequences of K111-related insertions amplified from human DNA and human/rodent somatic chromosomal cell hybrids are deposited in the NCBI database with Accession Numbers (JQ790790 - JQ790967). All other data generated or analyzed during this study are included in this published article (and its Supplementary Information files).

Figures



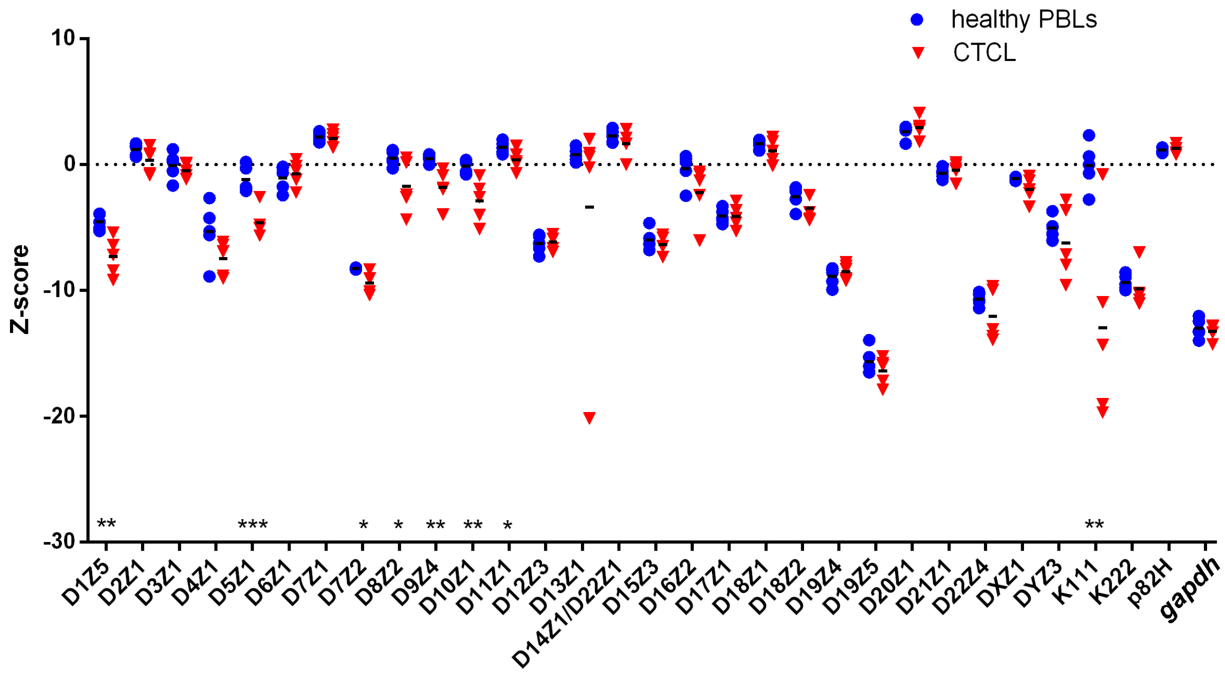
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.1: Heterogeneous alterations of centromere DNA in cancer cell lines. Heatmap representing the abundance of α-satellites specific for each centromere array (rows) obtained by qPCR in 50 ng of DNA from healthy cells and from cancer lines (columns). Relative abundance is denoted by the gradient legend (top left). Cancer type and α-satellite localization is depicted as indicated by the legend (bottom left). Repeats marked with an asterisk (also bolded and italicized) represent α-satellites with appreciable alterations across various cell lines relative to healthy controls. Data depicting α-satellite abundance are log$_2$ normalized to healthy PBL median values (asterisks, red). The nomenclature of these α-satellites begins with the letter D, followed by their resident chromosome number (1–22, X or Y), followed by a Z, and a number indicating the order in which the sequence was discovered. The DYZ3 repeat was excluded from the analysis to reduce confounding from gender.



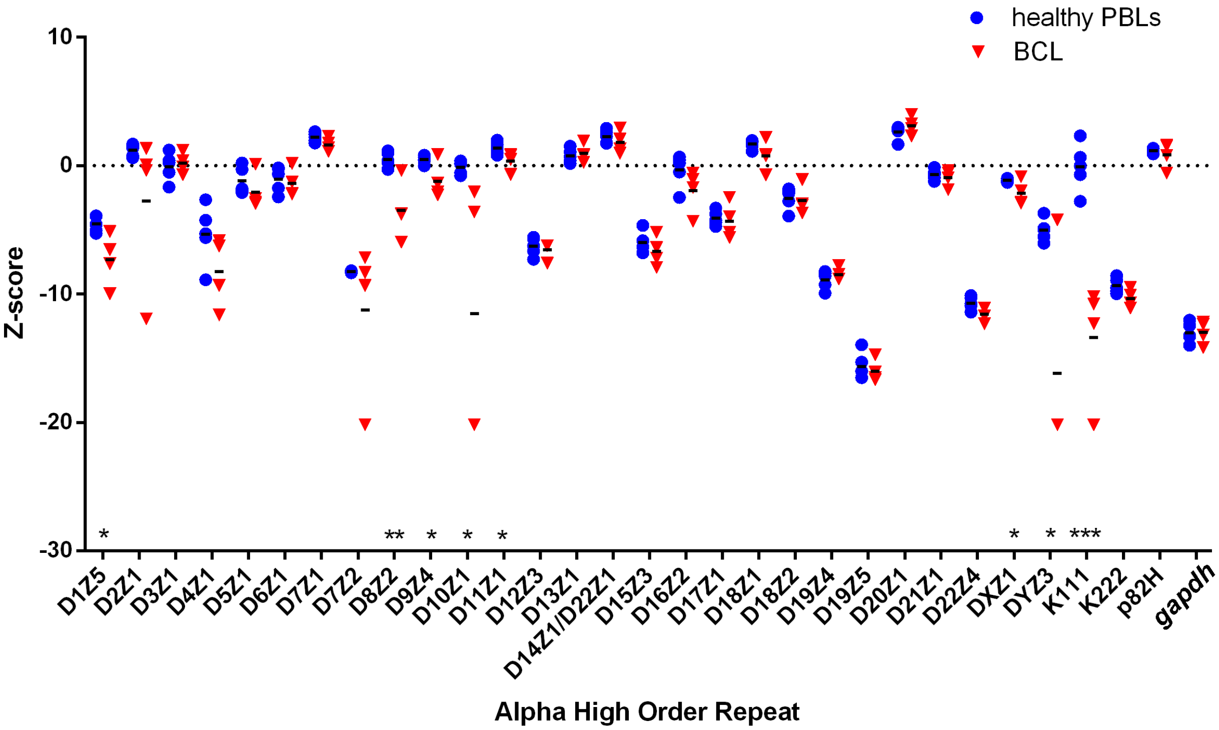
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.2: Heterogeneous loss of centromeres in acute lymphoblastic leukemia (ALL) cell lines. Abundance of centromere specific α-satellites is depicted by the Z-score (Y-axis) of each α-satellite. The nomenclature of these α-satellites begins with the letter D, followed by their resident chromosome number (1–22, X or Y), followed by a Z, and a number indicating the order in which the sequence was discovered. The $\log_2$ normalized numbers for each α-satellite were normalized to the average copy number of a given repeat in DNA from healthy cells (blue circles). Statistical significance was calculated using a t-test. * = *p*< 0.05, ** = *p*< 0.01, *** = *p*<0.001, **** = *p*< 0.0001.



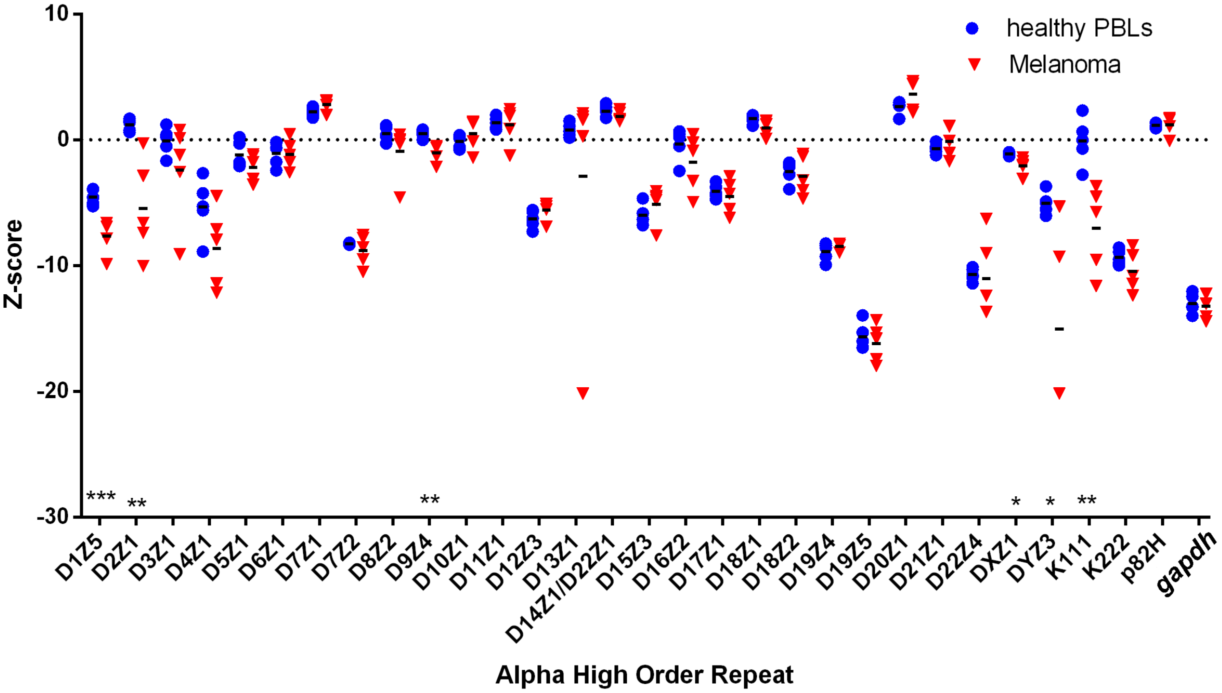
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.3: Heterogeneous loss of centromeres in cutaneous T-cell lymphoma (CTCL) cell lines. Abundance of centromere specific α-satellites is depicted by the Z-score (Y-axis) of each α-satellite.



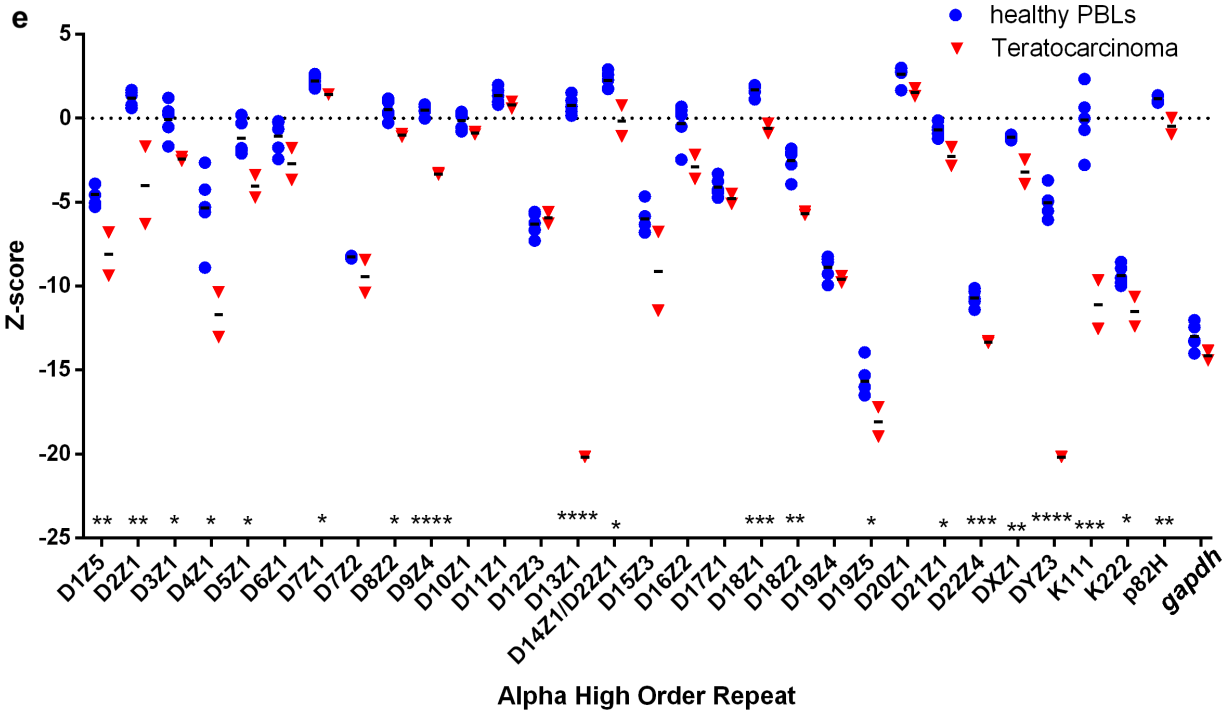
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.4: Heterogeneous loss of centromeres in B-cell lymphoma (BCL) cell lines. Abundance of centromere specific α-satellites is depicted by the Z-score (Y-axis) of each α-satellite.



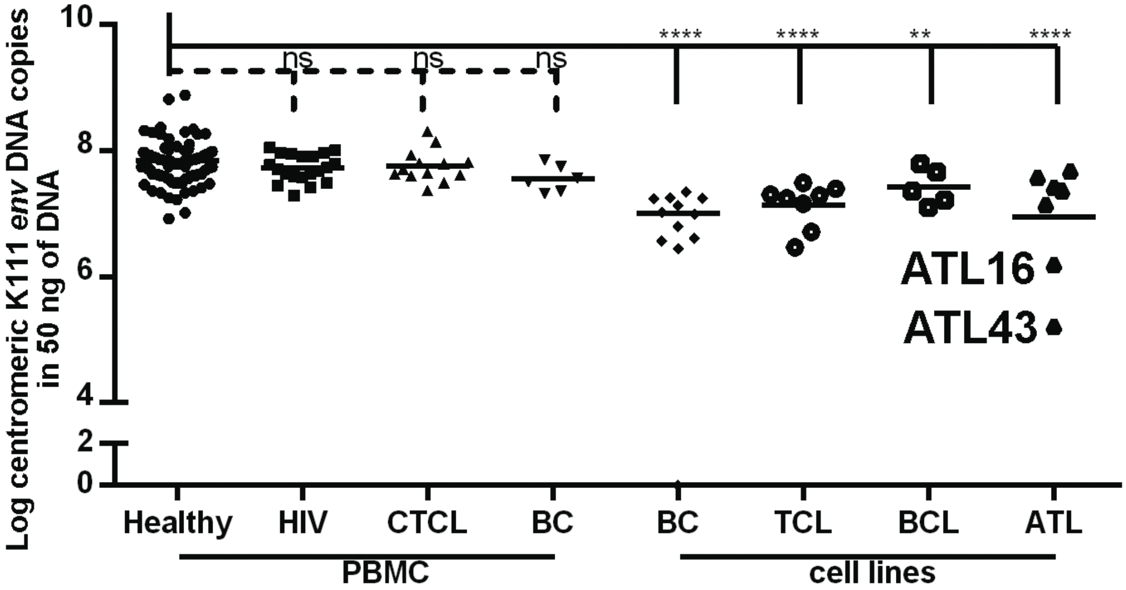
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.5: Heterogeneous loss of centromeres in melanoma cell lines. Abundance of centromere specific α-satellites is depicted by the Z-score (Y-axis) of each α-satellite.



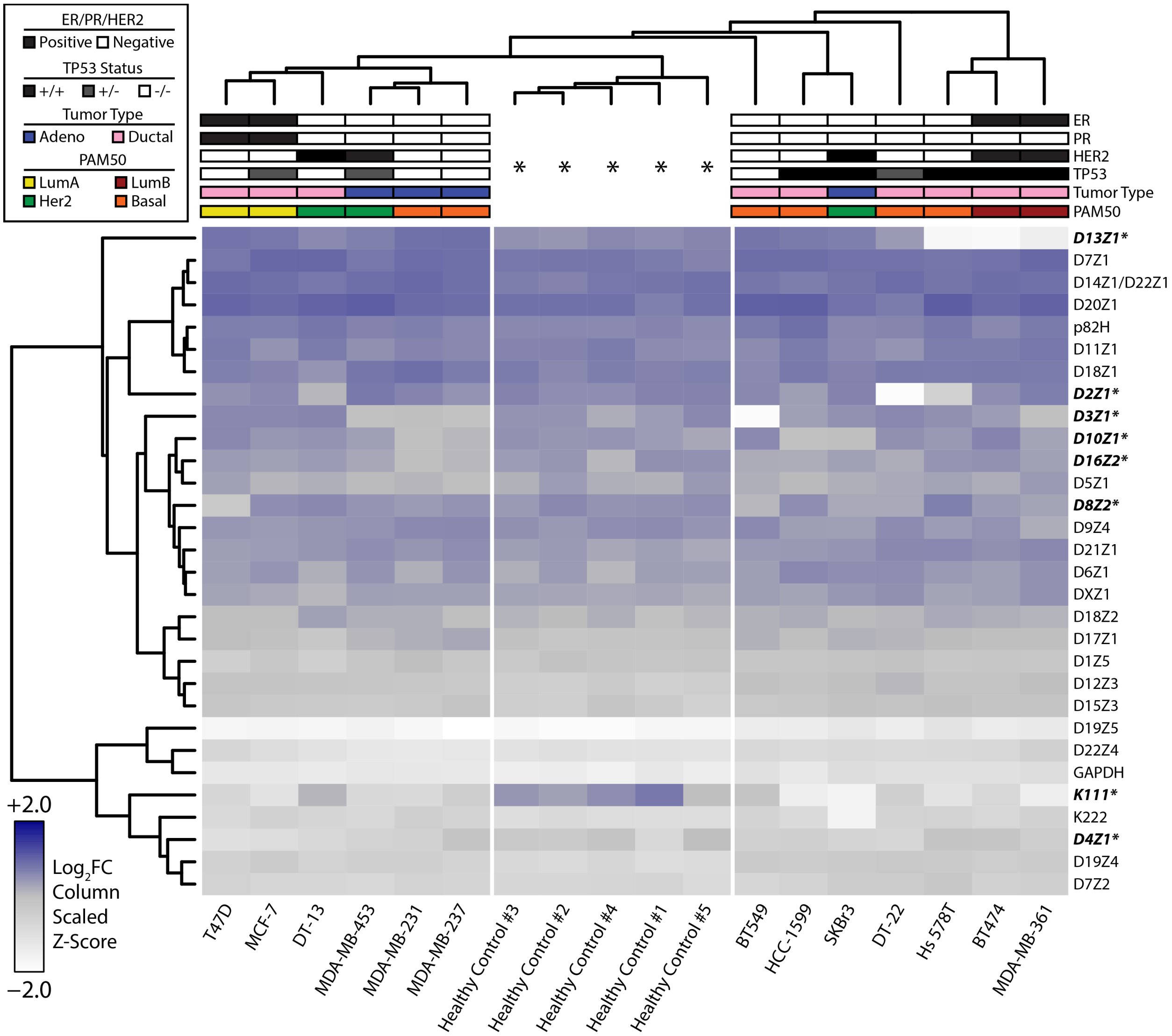
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.6: Heterogeneous loss of centromeres in cancer cell lines. Abundance of centromere specific α-satellites is depicted by the Z-score (Y-axis) of each α-satellite.



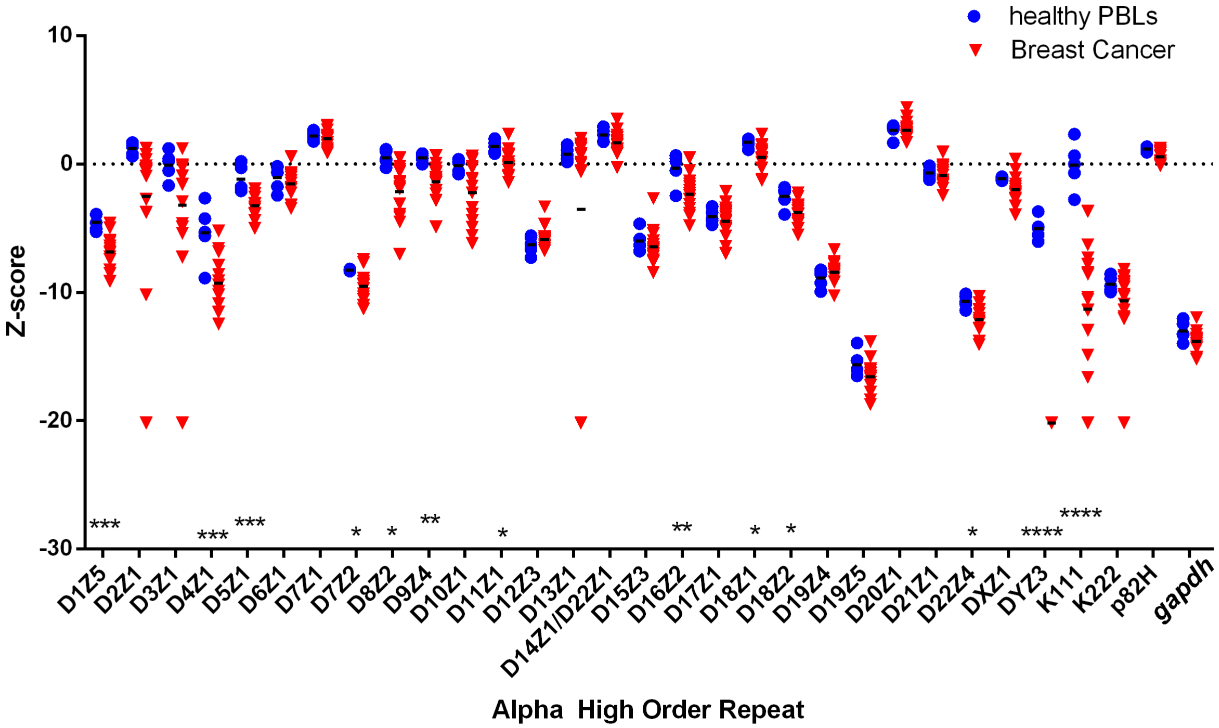
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.7: Abundance of pericentromeric K111 across cell lines and PBMCs. Dot plots representing the abundance of centromeric HERV *env* sequences from either PBMCs or cell lines with disease type designations listed along the X axis. The *gag* region of K111 was not assessed in our analysis, as select populations and CTCL patients are homozygous null for K111 *gag*. HIV = HIV patient sample, CTCL = Cutaneous T-Cell Lymphoma, BC = Breast Cancer, TCL = T-Cell Lymphoma, BCL = B-Cell Lymphoma, ATL = Adult T-Cell Lymphoma. Abundance of K111 is depicted by the $\log_2$ Z-score (Y-axis). Statistical significance was calculated using a t-test. * = $p<$ 0.05, ** = $p<$ 0.01, *** = $p<$0.001, **** = $p<$ 0.0001.



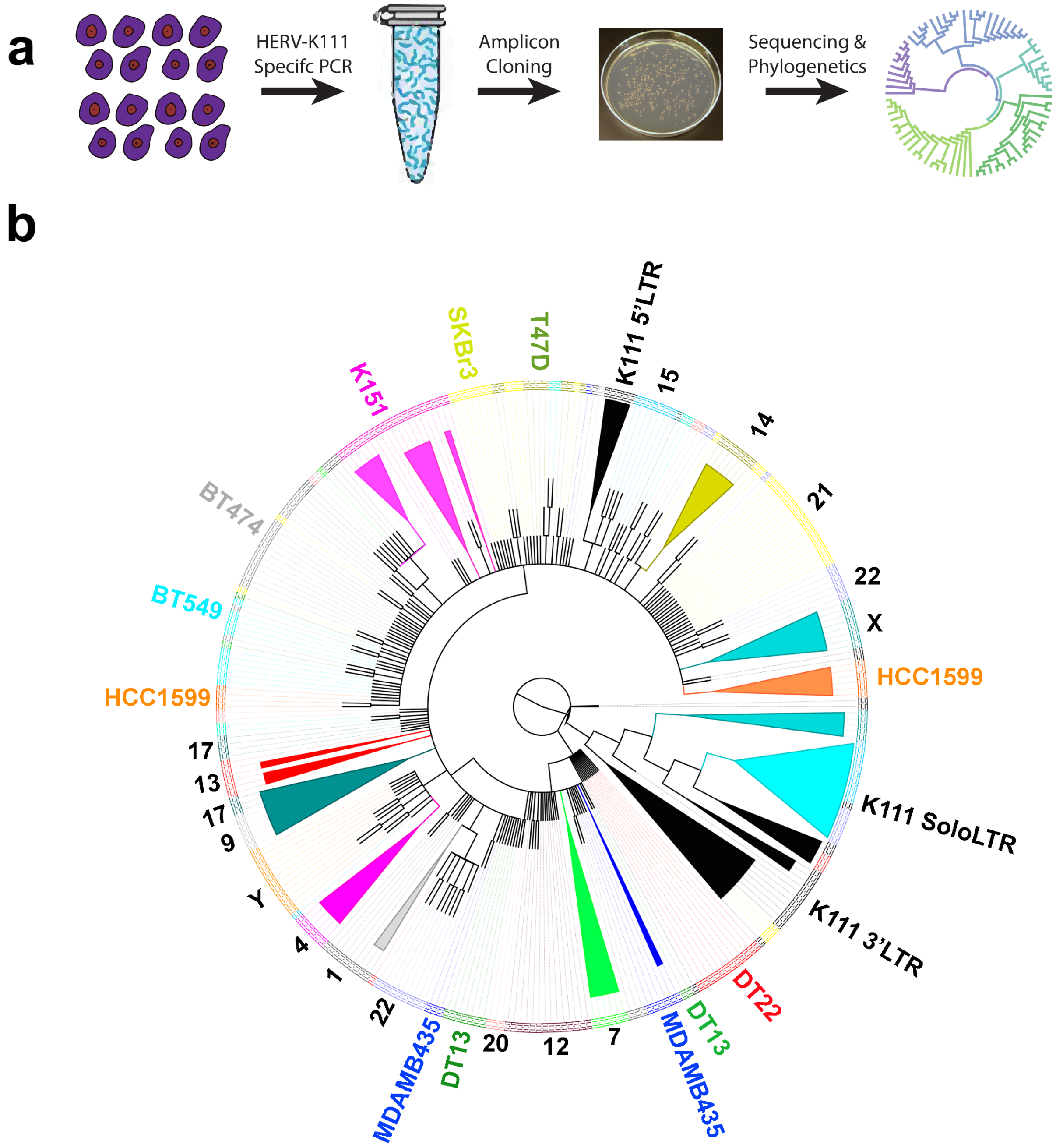
Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.8: Genomic profiling of centromeres in breast cancer cell lines. Heatmap representing the abundance of α-satellites specific for each centromere array (rows) obtained by qPCR in 50 ng of DNA from healthy cells and from breast cancer lines (columns). Relative abundance is denoted by the gradient legend (bottom left). Data depicting α-satellite abundance are $\log_2$ normalized to healthy PBL median values (asterisks). Repeats marked with an asterisk (also bolded and italicized) represent α-satellites with appreciable alterations across various cell lines relative to healthy controls. Hormone receptor, *TP53* status, histologic, and molecular classifications are depicted as indicated by the legend (top left). The DYZ3 repeat was excluded from the analysis to reduce confounding due to gender.



Authors: Anjan K. Saha, Mohamad Mourad, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

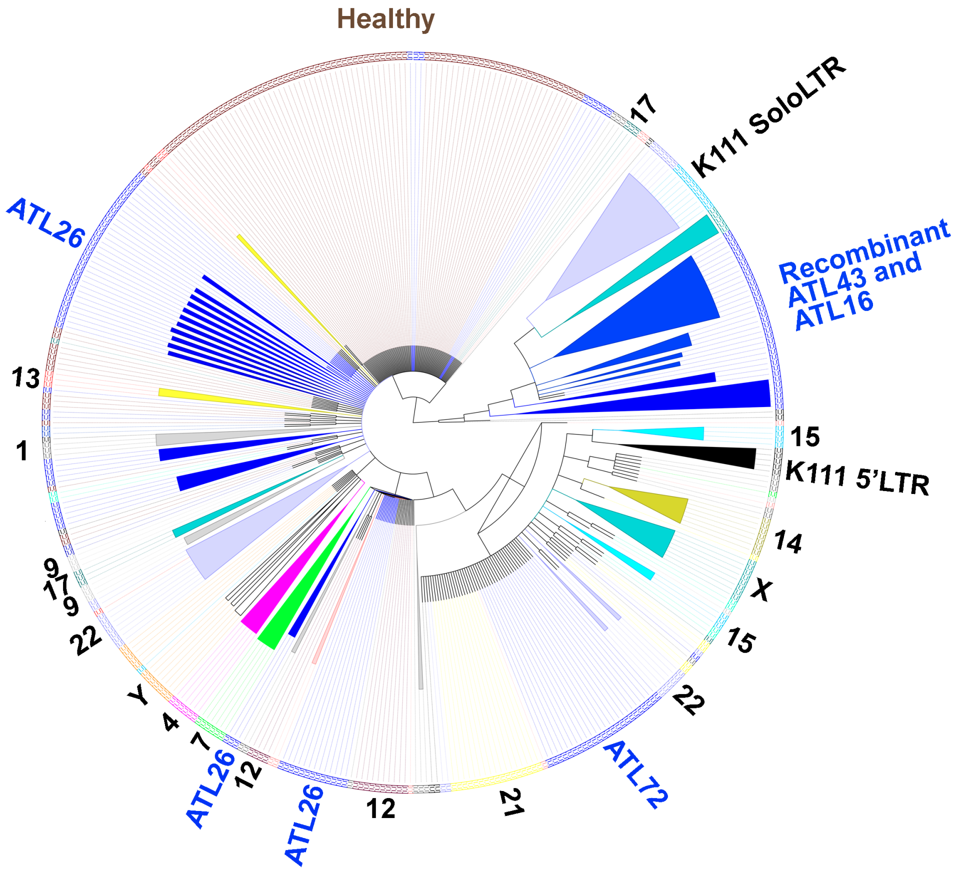
Figure 2.9: Heterogeneous loss of centromeres in breast cancer cell lines. Dot plots representing the abundance of α-satellites specific for each centromere array (X axis) in breast cancer cell lines. Abundance of centromere specific α-satellites is depicted by the Z-score (Y-axis) of each α-satellite. The log$_2$ normalized numbers for each α-satellite were normalized to the average copy number of a given repeat in DNA from healthy cells (blue circles). Statistical significance was calculated using a t-test. * = *p*< 0.05, ** = *p*< 0.01, *** = *p*<0.001, **** = *p*< 0.0001.



Authors: Anjan K. Saha, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

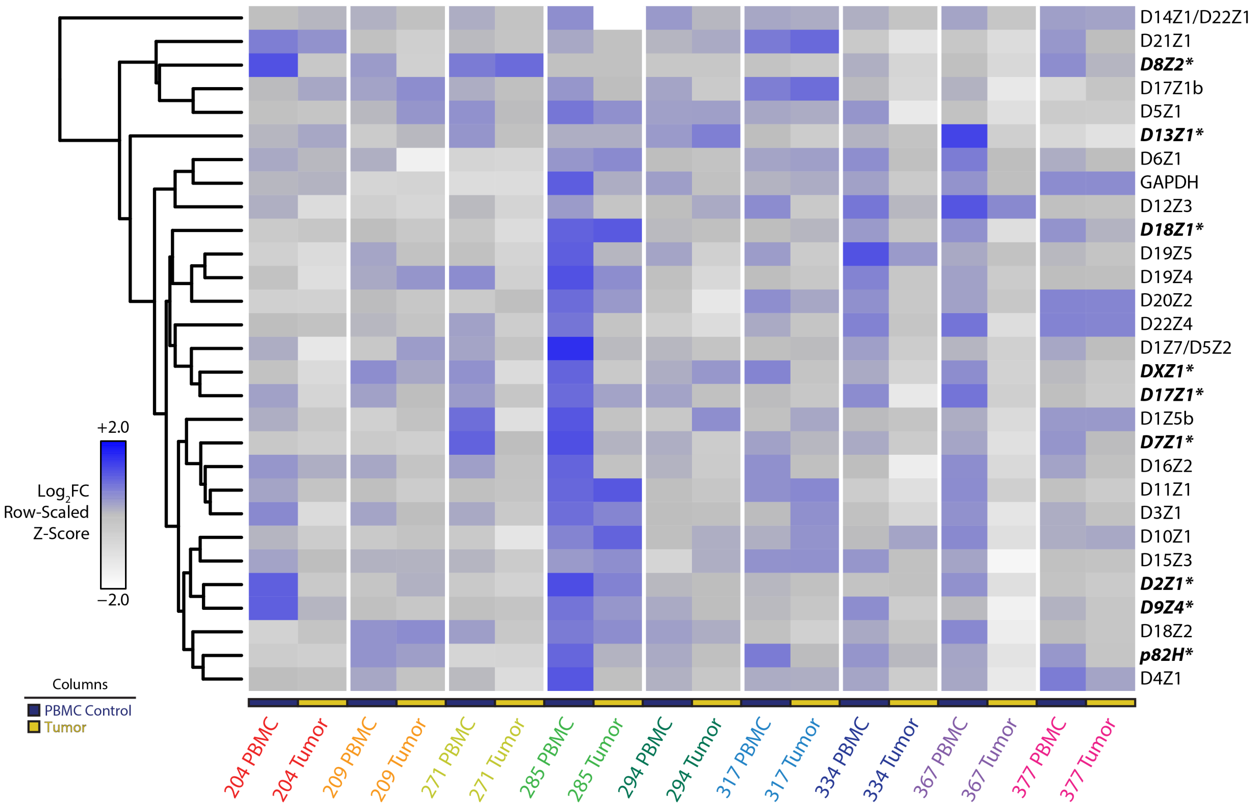
Figure 2.10: Gene conversion of HERV-K111 in breast cancer cell lines. a) Schematic outline of the experimental methodology employed to identify gene conversion events. b) Phylogenetic analysis conducted on K111 sequences amplified by PCR on breast cancer cell lines (T47D, BT549, HCC-1599, MD-MB-435, DT13, DT22, K151, and SKBr3) and human-hamster hybrid

cell lines (each containing a single human chromosome) as a reference. Amplicons are labeled and color-coded along the edge of the phylogenetic tree according to the cell line that produced the amplicon. Amplicons from human-hamster hybrid cell lines are denoted numerically by the human chromosome present in each hybrid cell line. Amplicons from K111 5'LTR, 3'LTR, and Solo LTR are additionally denoted. An example of gene conversion is shown in the cell line K151, possessing clades (pink) that localize in close proximity relative to each other but are not found heterogeneously throughout the tree. Convergence on two distinct K111 subtypes can additionally be identified within the MDA-MB-435, DT-13, and HCC1599 cell lines.



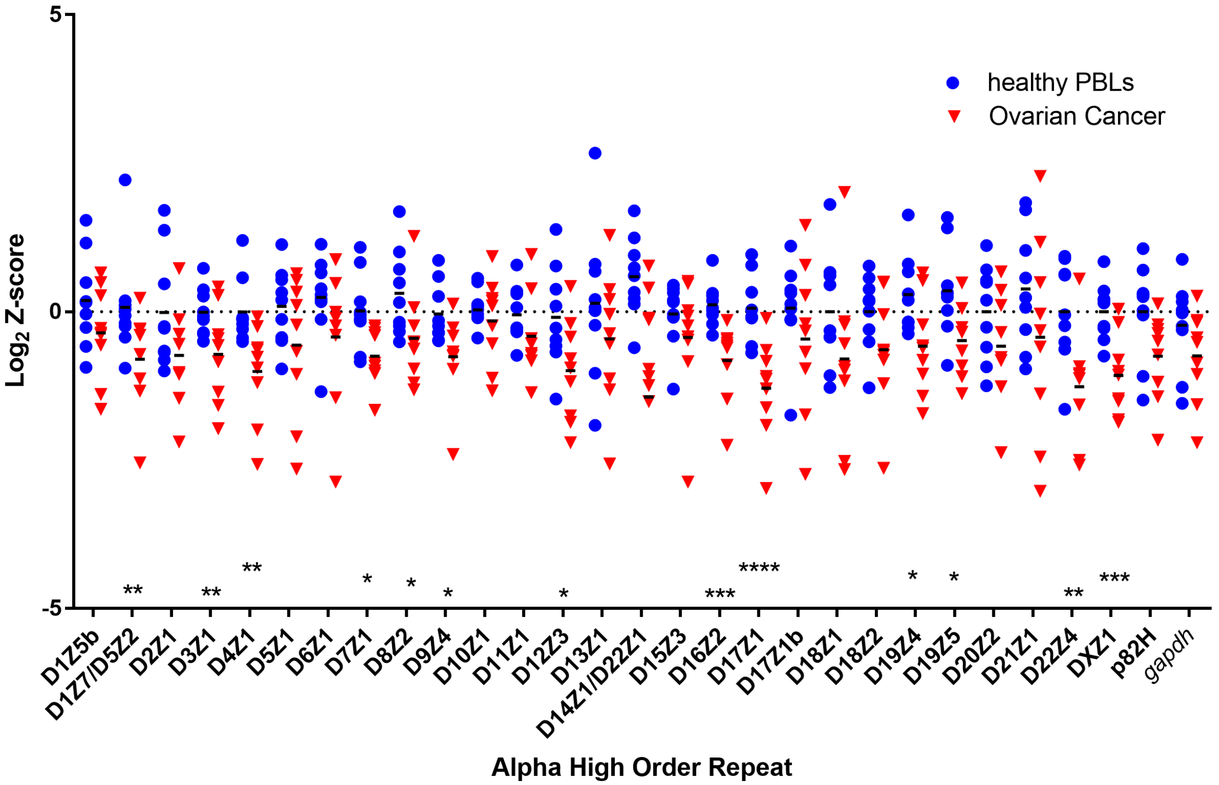
Authors: Anjan K. Saha, Mark H. Kaplan, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.11: Gene conversion of HERV K111 in adult T-cell leukemia (ATL) cell lines. Phylogenetic analysis conducted on K111 sequences amplified by PCR in ATL cell lines (ATL26, ATL72, ATL16, and ATL43) and heathy cells (brown labels). Amplicons are labeled and color-coded along the edge of the phylogenetic tree according to the cell line that produced the amplicon. Amplicons from human-hamster hybrid cell lines are denoted numerically by the human chromosome present in each hybrid cell line. Amplicons from K111 5'LTR and Solo LTR are additionally denoted. Recombinant K111 sequences resembling Solo LTRs are seen in cell lines ATL43 and ATL16, shown in blue arising from the same ancestral sequence to K111 Solo LTR sequences. K111 sequences from heathy PBLs show heterogeneous distribution along the tree and did not cluster in novel clades.



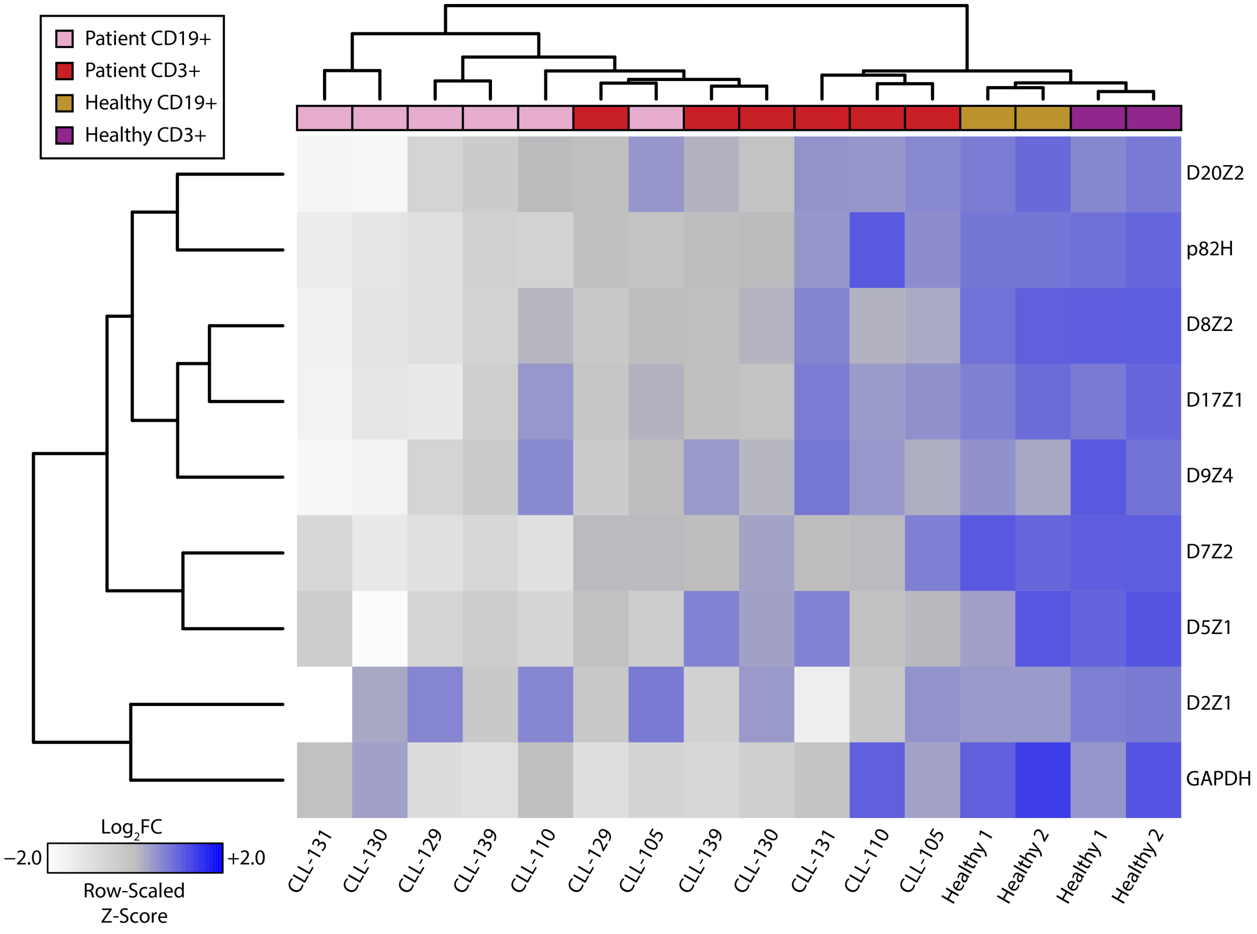
Authors: Anjan K. Saha, Ilana Chefetz, Ronald Buckanovich, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.12: Genomic profiling of centromeres in primary ovarian cancer tissue. Heatmap representation of rapid PCR data from nine primary ovarian cancer tissue samples with matched PBMC DNA. Matched sets from the same patient are grouped by color. PBMC control samples and tumor samples are labeled according to the legend (bottom left). Data depicting α-satellite abundance are $\log_2$ normalized to PBMC median values. Relative abundance is denoted by the gradient legend (bottom left). Repeats marked with an asterisk (also bolded and italicized) represent α-satellites with appreciable alterations across tissue samples relative to PBMC controls.



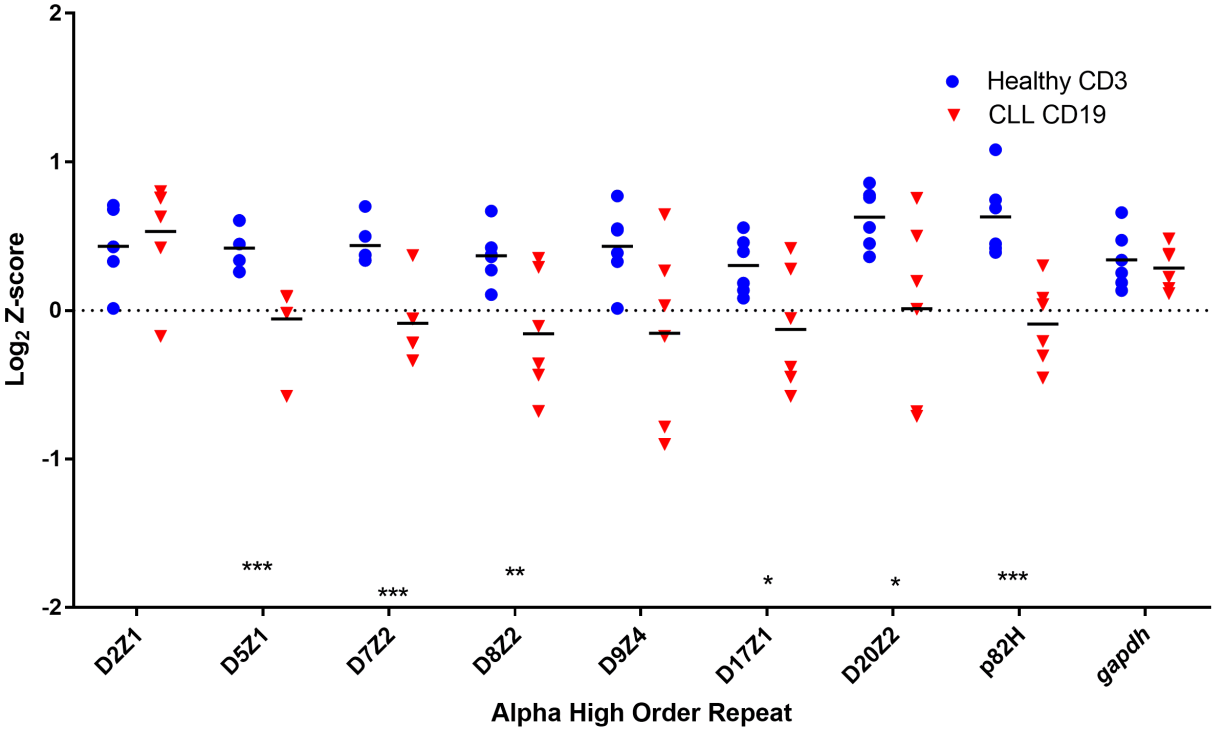
Authors: Anjan K. Saha, Ilana Chefetz, Ronald Buckanovich, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.13: Heterogeneous loss of centromeres in ovarian cancer patients. Dot plots representing the abundance of α-satellites specific for each centromere array (X axis) in ovarian cancer tumors. Abundance of centromere specific α-satellites is depicted by the log$_2$ Z-score (Y-axis) of each α-satellite. The log$_2$ normalized numbers for each α-satellite were normalized to the average copy number of a given repeat in DNA from healthy cells (blue circles). Statistical significance was calculated using a t-test. * = p< 0.05, ** = p< 0.01, *** = p<0.001, **** = p< 0.0001.



Authors: Anjan K. Saha, Sami N. Malek, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.14: CLL (malignant B-cells) and patient matched T-cells assessed for select centromeric α-satellite markers. Heatmap representation of rapid PCR data from six primary CLL and two healthy samples post-separation by indicated cell surface markers into B-cell (CD19+) and T-cell (CD3+) populations. Data depicting α-satellite abundance are log$_2$ normalized to T-cell median values. Relative abundance is denoted by the gradient legend (bottom left). Lymphocyte characterization and disease status is depicted as indicated by the legend (top left).



Authors: Anjan K. Saha, Sami N. Malek, David M. Markovitz, Rafael Contreras-Galindo

Figure 2.15: Heterogeneous loss of centromeres in chronic lymphoctyic leukemia (CLL) patients. Dot plot representing the abundance of α-satellites specific for each centromere array (X axis) in CLL tumors. Abundance of centromere specific α-satellites is depicted by the log$_2$ Z-score (Y-axis) of each α-satellite. The log$_2$ normalized numbers for each α-satellite were normalized to the average copy number of a given repeat in DNA from healthy cells (blue circles). Statistical significance was calculated using a t-test. * = p< 0.05, ** = p< 0.01, *** = p<0.001, **** = p< 0.0001.

References

1. Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**, 407–421 (2003).

2. Hayden, K. E. Human centromere genomics: now it's personal. *Chromosome Research* **20**, 621–633 (2012).

3. Henikoff, J. G., Thakur, J., Kasinathan, S. & Henikoff, S. A unique chromatin complex occupies young -satellite arrays of human centromeres. *Science Advances* **1**, e1400234–e1400234 (2015).

4. Aldrup-Macdonald, M. E. & Sullivan, B. A. The past, present, and future of human centromere genomics. *Genes (Basel)* **5**, 33–50 (2014).

5. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).

6. Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K. & Sullivan, B. A. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* **26**, 1301–1311 (2016).

7. Li, X. *et al.* A fluorescence in situ hybridization (FISH) analysis with centromere-specific DNA probes of chromosomes 3 and 17 in pleomorphic adenomas and adenoid cystic carcinomas. *J. Oral Pathol. Med.* **24**, 398–401 (1995).

8. Liehr, T. *Benign and Pathological Chromosomal Imbalances: Microscopic and Submicroscopic Copy Number Variations (CNVs) in Genetics and Counseling*. (Academic Press, 2013).

9. Contreras-Galindo, R. *et al.* Rapid molecular assays to study human centromere genomics. *Genome Res.* **27**, 2040–2049 (2017).

10. Quénet, D. & Dalal, Y. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. *eLife* **3**, (2014).

11. Jabs, E. W., Goble, C. A. & Cutting, G. R. Macromolecular organization of human centromeric regions reveals high-frequency, polymorphic macro DNA repeats. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 202–206 (1989).

12. Zahn, J. *et al.* Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans. *Genome Biol.* **16**, 74 (2015).

13. Du, Y., Topp, C. N. & Dawe, R. K. DNA Binding of Centromere Protein C (CENPC) Is Stabilized by Single-Stranded RNA. *PLoS Genetics* **6**, e1000835 (2010).

14. Vig, B. K., Sternes, K. L. & Paweletz, N. Centromere structure and function in neoplasia. *Cancer Genet. Cytogenet.* **43**, 151–178 (1989).

15. Black, E. M. & Giunta, S. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes (Basel)* **9**, (2018).

16. Bersani, F. *et al.* Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15148–15153 (2015).

17. Natisvili, T. *et al.* Transcriptional Activation of Pericentromeric Satellite Repeats and Disruption of Centromeric Clustering upon Proteasome Inhibition. *PLoS ONE* **11**, e0165873 (2016).

18. Giunta, S. & Funabiki, H. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 1928–1933 (2017).

19. Shepelev, V. A. *et al.* Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genom Data* **5**, 139–146 (2015).

20. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).

21. Yi, J.-M. & Kim, H.-S. Expression and phylogenetic analyses of human endogenous retrovirus HC2 belonging to the HERV-T family in human tissues and cancer cells. *J. Hum. Genet.* **52**, 285–296 (2007).

22. Hughes, J. F. & Coffin, J. M. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics* **171**, 1183–1194 (2005).

23. Nathanson, K. L. *et al.* The Y deletion gr/gr and susceptibility to testicular germ cell tumor. *Am. J. Hum. Genet.* **77**, 1034–1043 (2005).

24. Machiela, M. J. *et al.* Mosaic chromosome Y loss and testicular germ cell tumor risk. *J. Hum. Genet.* **62**, 637–640 (2017).

25. Mostert, M. M. *et al.* Fluorescence in situ hybridization-based approaches for detection of 12p overrepresentation, in particular i(12p), in cell lines of human testicular germ cell tumors of adults. *Cancer Genet. Cytogenet.* **87**, 95–102 (1996).

26. Summersgill, B. M. *et al.* Definition of chromosome aberrations in testicular germ cell tumor cell lines by 24-color karyotyping and complementary molecular cytogenetic analyses. *Cancer Genet. Cytogenet.* **128**, 120–129 (2001).

27. Contreras-Galindo, R. *et al.* HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Research* **23**, 1505–1513 (2013).

28. Taherian-Fard, A., Srihari, S. & Ragan, M. A. Breast cancer classification: linking molecular mechanisms to disease prognosis. *Brief. Bioinformatics* **16**, 461–474 (2015).

29. Davidson, J. M. *et al.* Molecular cytogenetic analysis of breast cancer cell lines. *Br. J. Cancer* **83**, 1309–1317 (2000).

30. Lagos, S. M. R. & Jiménez, N. E. R. Cytogenetic Analysis of Primary Cultures and Cell Lines: Generalities, Applications and Protocols. *Recent Trends in Cytogenetic Studies - Methodologies and Applications* (2012). doi:10.5772/34200

31. Morris, J. S., Carter, N. P., Ferguson-Smith, M. A. & Edwards, P. A. Cytogenetic analysis of three breast carcinoma cell lines using reverse chromosome painting. *Genes Chromosomes Cancer* **20**, 120–139 (1997).

32. Rummukainen, J. *et al.* Aberrations of chromosome 8 in 16 breast cancer cell lines by comparative genomic hybridization, fluorescence in situ hybridization, and spectral karyotyping. *Cancer Genet. Cytogenet.* **126**, 1–7 (2001).

33. Letessier, A. *et al.* Multicolour-banding fluorescence in situ hybridisation (mbanding-FISH) to identify recurrent chromosomal alterations in breast tumour cell lines. *Br. J. Cancer* **92**, 382–388 (2005).

34. Klein, H. L. Genetic control of intrachromosomal recombination. *Bioessays* **17**, 147–159 (1995).

35. Blanco, P. *et al.* Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *J. Med. Genet.* **37**, 752–758 (2000).

36. Schneider, K. L., Xie, Z., Wolfgruber, T. K. & Presting, G. G. Inbreeding drives maize centromere evolution. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E987-996 (2016).

37. Shi, J. *et al.* Widespread gene conversion in centromere cores. *PLoS Biol.* **8**, e1000327 (2010).
38. Wolfgruber, T. K. *et al.* High Quality Maize Centromere 10 Sequence Reveals Evidence of Frequent Recombination Events. *Front Plant Sci* **7**, 308 (2016).
39. Dangel, A. W., Baker, B. J., Mendoza, A. R. & Yu, C. Y. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics* **42**, 41–52 (1995).
40. Vitte, C. & Panaud, O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Mol. Biol. Evol.* **20**, 528–540 (2003).
41. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1668–1672 (2004).
42. Tavassoli, M. *et al.* Whole chromosome 17 loss in ovarian cancer. *Genes Chromosomes Cancer* **8**, 195–198 (1993).
43. MacKinnon, R. N. & Campbell, L. J. The Role of Dicentric Chromosome Formation and Secondary Centromere Deletion in the Evolution of Myeloid Malignancy. *Genetics Research International* **2011**, 1–11 (2011).
44. Sun, X. *et al.* Elevated expression of the centromere protein-A(CENP-A)-encoding gene as a prognostic and predictive biomarker in human cancers: Elevated Expression of the CENP-A-Encoding Gene in Cancer. *International Journal of Cancer* **139**, 899–907 (2016).
45. Zhang, W. *et al.* Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nature Communications* **7**, 12619 (2016).
46. Athwal, R. K. *et al.* CENP-A nucleosomes localize to transcription factor hotspots and subtelomeric sites in human cancer cells. *Epigenetics Chromatin* **8**, 2 (2015).
47. Lacoste, N. *et al.* Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. *Mol. Cell* **53**, 631–644 (2014).
48. Ting, D. T. *et al.* Aberrant Overexpression of Satellite Repeats in Pancreatic and Other Epithelial Cancers. *Science* **331**, 593–596 (2011).
49. Kishikawa, T. *et al.* Satellite RNA Increases DNA Damage and Accelerates Tumor Formation in Mouse Models of Pancreatic Cancer. *Mol. Cancer Res.* **16**, 1255–1262 (2018).
50. Kishikawa, T. *et al.* Satellite RNAs promote pancreatic oncogenic processes via the dysfunction of YBX1. *Nat Commun* **7**, 13006 (2016).
51. Niikura, Y. *et al.* CENP-A K124 Ubiquitylation Is Required for CENP-A Deposition at the Centromere. *Developmental Cell* **32**, 589–603 (2015).
52. Deyter, G. M. R. & Biggins, S. The FACT complex interacts with the E3 ubiquitin ligase Psh1 to prevent ectopic localization of CENP-A. *Genes & Development* **28**, 1815–1826 (2014).
53. Gopalakrishnan, S., Sullivan, B. A., Trazzi, S., Della Valle, G. & Robertson, K. D. DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. *Hum. Mol. Genet.* **18**, 3178–3193 (2009).
54. Kim, I. S. *et al.* Roles of Mis18α in Epigenetic Regulation of Centromeric Chromatin and CENP-A Loading. *Molecular Cell* **46**, 260–273 (2012).
55. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

56. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12**, 68–78 (2011).

57. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **161**, 1215–1228 (2015).

58. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).

59. Maeda, M. *et al.* Origin of human T-lymphotrophic virus I-positive T cell lines in adult T cell leukemia. Analysis of T cell receptor gene rearrangement. *Journal of Experimental Medicine* **162**, 2169–2174 (1985).

60. Kujawski, L. *et al.* Genomic complexity identifies patients with aggressive chronic lymphocytic leukemia. *Blood* **112**, 1993–2003 (2008).

61. Contreras-Galindo, R. *et al.* Characterization of human endogenous retroviral elements in the blood of HIV-1-infected individuals. *J. Virol.* **86**, 262–276 (2012).

62. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).

Chapter 3 – Non-Canonical Function of CENPA as a Regulator of Gene Expression in Prostate Cancer

Summary

Overexpression of centromeric proteins has been identified in a number of human malignancies, though their functional and mechanistic contributions to disease progression have not been characterized. CENPA, the centromeric histone H3 variant, is the epigenetic mark that determines centromere identity. Here, we demonstrate that CENPA is highly overexpressed in prostate cancer in both tissue and cell lines, and the level of CENPA expression correlates with the stage of disease. Gain-of- and loss-of-function experimentation confirms that CENPA promotes prostate cancer cell line growth. Integrated sequencing studies further reveal a previously unidentified function of CENPA as a transcriptional regulator that modulates expression of critical proliferation, cell-cycle, and centromere/kinetochore genes. Our findings, therefore, suggest a previously undescribed biological function for CENPA, a protein normally thought to be solely and importantly involved in centromere identity. Identifying a novel function for CENPA as a regulator of gene expression represents a major shift in our understanding of the role it plays in biology and disease. While CENPA is indeed a crucial factor that epigenetically compartmentalizes centromere function during cell division, our findings shed light on a parallel mechanism for CENPA that might be involved in perpetuating malignant phenotypes through gene regulation. Though this study is focused on interrogating prostate cancer, the scope of these

findings might be generalizable to other malignancies and can ultimately serve as a foundation for designing therapeutics that selectively target CENPA's gene-regulatory activity.

Introduction

Centromeres are cellular structures that are necessary for the propagation of hereditary information[1,2]. Located centric to the ends of each chromosome, centromeres provide the structural foundation for kinetochores, multimeric complexes that serve as molecular interfaces between microtubule spindle fibers and individual chromatids during mitosis[1]. The centromere-kinetochore-microtubule interaction facilitates separation of the sister chromatids as mitosis proceeds from metaphase to anaphase. Centromeres are thus essential to ensuring faithful segregation of chromosomes in actively dividing cells.

Efforts to study human centromeres have focused on the epigenetics that drive centromere assembly[3,4]. Alpha satellite sequences that define centromere DNA are primarily occupied by the centromere-specific histone H3 variant known as CENPA, a highly conserved ~17 kDa molecule that forms a centromere-specific nucleosome with H2A, H2B and H4[3,5,6]. The CENPA-specific chaperone HJURP facilitates proper localization and incorporation of newly synthesized CENPA into nucleosomes occupying replicated alpha satellite DNA through a ubiquitin E3 ligase dependent process[7–9]. CENPA nucleosomes subsequently engage a unique set of binding partners that ensure proper genomic localization. These binding partners including CENPB, CENPC, and the constitutive centromere associated network (CCAN) that comprises the inner kinetochore[10,11]. The CCAN further serves as a multimeric interface between CENPA nucleosomes and the KMN (KNL-1/Mis12 complex/Ndc80 complex) network that comprises the outer kinetochore and directly interacts with the microtubule spindle fibers[12].

CENPA and its associated proteins therefore represent structural components that are essential to the integrity of cell division, and appropriate genomic localization of centromeric proteins is consequently a critical event in the cell cycle. Diseases of uncontrolled cell division, particularly cancer, are thus compelling to examine from the epigenetic perspective of centromere biology, primarily as it pertains to the key epigenetic mark CENPA. A number of studies have identified aberrant expression of centromeric/kinetochore proteins in cancers, where overexpression is predictive of survival and response to therapy, though their mechanistic contribution to cancer pathogenesis remains elusive[13–16]. In the setting of ectopic constitutive overexpression, CENPA mislocalization in HeLa cells is independent of aberrant E3 ligase activity, but rather demonstrates a reliance on the histone chaperone DAXX[17]. Endogenously overexpressed CENPA has also been shown to ectopically localize in colon cancer cell lines[18]. The effects of such mislocalization on phenotypes in malignancy have yet to be elucidated, though enrichments in ectopic binding to DNase hypersensitivity sites and CTCF transcription factor consensus sequences hint at a potential role in regulating gene transcription[17,18].

Here we report that CENPA is highly overexpressed in prostate cancer and that disease progression correlates with CENPA expression within a large patient cohort. CENPA knockdown markedly decreases proliferation of prostate cancer cells but not that of benign prostate cells, and increased expression of CENPA causes benign prostate epithelial cells to proliferate more rapidly. Most strikingly, CENPA appears to affect proliferation of prostate cancer cells by acting as a transcriptional regulator that modulates expression of genes critical to proliferation, cell cycle progression, and centromere/kinetochore integrity. Thus, overexpression of CENPA, a histone variant studied in great detail in its role as a centromere epigenetic mark, unexpectedly also contributes to prostate cancer pathogenesis by an alternative and previously uncharacterized mechanism.

Results

Overexpression of Centromeric Factors in Prostate Cancer

The significance of centromeres to cell division suggests that centromeric components may play important roles in development and in diseases of cell division gone awry, particularly in cancer. Previous work identified a centromere-kinetochore (CEN-KT) signature that was associated with aggressive, treatment-refractory malignancy[16]. We therefore profiled the transcriptomes of different types of malignancies across a compiled catalogue of publicly available RNA-sequencing (RNA-seq) databases[19]. We found that *CENPA* is ubiquitously overexpressed in malignant tissue relative to respective normal counterparts (**Figure 3.1A, Table 3.1**). These observations, combined with the well-characterized contributions of centromeric components like CENPA to cell division, suggested conducting a more focused interrogation of these components in cancers that display poor prognosis in the context of high proliferation indices. Prostate cancer is one such disease, where a high proliferation index is predictive of poor outcomes[20,21]. New treatment strategies are much needed for prostate cancer, which remains the most diagnosed malignancy in men and the second leading cause of cancer-related death in men[22]. While hormonal therapy and chemotherapeutic options are available, resistant metastatic disease and life-altering side effects, such as urinary incontinence and erectile dysfunction, are everlasting concerns[23]. In view of the above considerations, we performed Sample Set Enrichment Analysis in the prostate tissue type cohort containing RNA-seq data from 685 tissue samples[19]. Gene expression of numerous centromeric components exhibited strong enrichments in prostate cancer tissue relative to their normal counterparts (**Figure 3.2A, Table 3.2**).

Our analysis corroborates previous reports that characterize some of these components as part of the CEN-KT signature that is strongly associated with poor disease outcomes[16]. We selected CENPA from this panel of genes for further assessment, given its central role in centromere biology, importance for development, and highly conserved structure, and found a significant increase in expression with disease progression (**Figure 3.2B**)[24]. This *in silico* finding was validated at the protein level through prostate tissue microarrays stained for CENPA, notably demonstrating marked overexpression of CENPA that increased with disease severity (**Figure 3.2C**). Importantly, receiver operator characteristic (ROC) analysis of the CENPA-stained prostate tissue microarray produced an area under the curve (AUC) of 0.89, orthogonally demonstrating a strong association between elevated CENPA expression and metastatic prostate cancer (**Figure 3.1B**). Assessment of CENPA expression was also examined in cancer cell line models to determine feasibility for more focused molecular inquiry. We verified robust overexpression of CENPA in prostate cancer cell lines, as compared to benign prostatic epithelial lines (**Figure 3.2D**). The PNT2 benign cell line was a notable exception, likely due to its rapid proliferation rate relative to other cell lines we tested (**Figure 3.1C**). Taken together, CENPA is a well-conserved, developmentally important factor abundant in prostate cancer tissue, as seen in a large number of patients, and in prostate cancer cell lines, and an increase in its expression at the RNA and protein levels is highly correlated with more aggressive disease.

CENPA is Associated with Cell Division in Prostate Cancer

The abundance of CENPA in prostate cancer raised the question as to whether overexpression plays a functional role in disease pathogenesis and progression. We thus first conducted a comparative analysis of *CENPA* expression relative to the remaining transcriptome in prostate cancer to identify associations with biologic concepts that could computationally guide functional assessments. Our efforts to profile transcriptomes in human cancer and normal tissue facilitates performing transcriptome-wide correlations against nominated genes of interest in a tissue specific manner within a large catalogue of samples (n = 10,848). We thus correlated *CENPA* mRNA levels to the expression levels of all other protein coding elements (**Table 3.3**) to deconvolute its relative contribution to prostate cancer progression. *CENPA* expression tracks tightly with a number of previously identified prostate cancer pathogenesis factors including *CENPF*, *UBE2C* and *EZH2* (**Figure 3.3A, 3.4C, and 3.4D, Table 3.3**). *MKI67* (gene encoding proliferation marker Ki67) also performed well in our analysis, further suggesting a role for *CENPA* in cellular proliferation (**Figure 3.3B**). Of note, *CENPA* does not tightly correlate with *ACTB* (housekeeping gene), *AMACR* (prostate cancer biomarker), or *AR* (**Figures 3.4A and 3.4B, Table 3.3**), suggesting a pathogenic process that is independent of androgen signaling, a pharmacologically relevant molecular pathway that is frequently targeted in prostate cancer treatment.

Strong associations with cellular proliferation genes and select pathogenesis factors independent of *AR* implicates CENPA as a contributor to a biologic process that is involved in androgen refractory prostate cancer progression. In fact, we found that AR signaling actually represses CENPA expression in cell culture (**Figure 3.5A**). We additionally used the Database for Annotation, Visualization, and Integrated Discovery (DAVID) to conduct ontology assessments on the highest performing genes from our transcriptome-wide correlation against CENPA expression in prostate cancer ($r > 0.8$)[25]. Our analysis revealed a correlation between CENPA gene expression and biologic concept clusters that highlight centromeres, kinetochores, mitosis, and cell division (**Figure 3.3C and 3.5B**). Concepts that include genes encoding components of the CCAN were additionally captured by our analysis. Of note, *CDC25C*, *CDCA5*, *TOP2A,* and *CENPU*, genes known to play roles in cellular proliferation, cell cycle progression and centromere/kinetochore integrity, were included in these biological concepts. Pre-ranked Gene Set Enrichment Analysis (GSEA) independently confirmed enrichments in gene signatures important for cell cycle, cell division, and mitosis (**Figure 3.3D and 3.5C**). Taken together, CENPA expression is strongly linked to gene signatures that underlie processes that govern proliferation, cell cycle progression, and centromere/kinetochore integrity in prostate cancer.

CENPA Dependent Proliferation in Prostate Cancer

Significant association between CENPA and proliferation signatures is expected given the role CENPA plays in the structural integrity of the centromere. There is limited evidence, however, concerning CENPA function in human malignancy. We therefore performed loss-of- and gain-of-function experiments in cell lines stably expressing either doxycycline-inducible short hairpin RNAs against CENPA or EF1A-promoter driven full-length CENPA. Doxycycline administration at 2 µg/mL was sufficient to produce robust knockdown of CENPA after 72 hours (**Figure 3.6A, 3.7A, and 3.7B**). CENPA depletion led to a profound growth-inhibitory effect on 22Rv1, LnCaP, and DU145 prostate cancer cells (**Figure 3.6B, 3.6C, 3.6E, 3.7C, and 3.7D**). CENPA depletion in prostate cancer cells results in an accumulation of cells in G1 that seem to be unable to progress through the cell cycle (**Figure 3.6D, 3.7E, and 3.7F**). Conversely, overexpression of CENPA in the 957E-hTERT benign prostate epithelial cell line leads to a profound growth-promoting effect (**Figure 3.6F and 3.6G**). Interestingly, benign 957E-hTERT cells depleted of CENPA do not demonstrate significant proliferative changes, consistent with previous reports that cells can proliferate with low levels of CENPA (**Figure 3.6G**)[26]. Taken together, our data show that CENPA is an essential factor for progression through the cell cycle and that overexpression drives proliferation of prostate cancer cells.

Non-Canonical Genomic Localization of CENPA

Given prior reports of ectopic deposition in the setting of CENPA overexpression and the marked overexpression of CENPA in prostate cancer, we performed native chromatin immunoprecipitation followed by sequencing (NChIP-seq) to identify non-centromeric and potentially regulatory binding sites for CENPA in prostate cancer. As expected, four α-satellites were enriched relative to the IgG control antibodies using a PCR assay we previously devised that can distinguish chromosome specific α-satellite DNA from any given centromere, verifying the validity of the CENPA ChIP (**Figure 3.8A**)[27]. CENPA directed ChIP-seq identified 569 non-centromeric binding sites in the VCaP prostate cancer cell line within three experimental replicates (**Figure 3.9A, 3.8C, and 3.8D**). One example of such a CENPA binding site is present in the promoter region of *CDC25C* (**Figure 3.9B**), a cell cycle phosphatase that is critical for progression through anaphase that was also identified in our comparative gene expression analysis described above (**Figure 3.3A**). Intriguingly, the promoter of *CENPA* itself was also bound by CENPA, consistent with previously reported results[18]. CENPA directed ChIP was additionally conducted in the benign prostatic epithelial cell line 957E-hTERT to determine whether ectopic CENPA binding is a cancer specific observation (**Figure 3.8B**). CENPA enrichment over the four previously assessed α-satellites was significantly lower than that observed in the VCaP cell line, consistent with each cell line's respective CENPA abundance observed above (**Figure 3.2D**). CENPA directed ChIP-seq for the 957E-hTERT cell line was thus deferred.

We next conducted a global assessment of CENPA binding sites to obtain a functional taxonomy of CENPA-bound genes. Ontologic assessment of genes whose transcriptional start sites were in close proximity to CENPA binding sites revealed enrichments in biologic concepts that are involved with maintenance of nuclear architecture and organization, such as protein-DNA complex assembly ($p = 6.44 \times 10^{-18}$) and chromosome organization ($p = 8.70 \times 10^{-13}$) (**Figure 3.9C**). Furthermore, binning CENPA binding sites into categories corresponding to discreet locations within the human genome demonstrates a predilection towards binding regulatory elements such as promoters and CpG islands (**Figure 3.9D**). Comparing the number of peaks present within any two genomic regions reveals significant overlap between loci considered to be regulatory areas (**Figure 3.8E**). Taken together, we show that CENPA localizes to non-canonical genomic loci, with a predilection towards the regulatory elements of genes that control cellular proliferation.

Gene Regulation by CENPA

Histone variants have been well characterized as modulators of aberrant gene expression in cancer. H2A.Z.2, macroH2A, and H3.3 are well documented as key contributors to malignant phenotypes in a number of cancer types[28–31]. CENPA is a centromere specific histone H3 variant overexpressed in cancer but whose functional contributions to malignancy have remained elusive. CENPA localizing to regulatory elements outside of the centromere near genes involved in maintaining nuclear architecture and chromosome organization, however, presents the intriguing possibility that CENPA plays a direct and previously unsuspected role in gene regulation. We thus conducted RNA-seq on CENPA-depleted cell lines to evaluate whether removing CENPA drastically alters the gene expression profiles of genes bound by CENPA. Doxycycline administration at 2 μg/mL was sufficient to produce ~71% (shCENPA1) and ~85% (shCENPA2) knockdown of CENPA mRNA relative to the non-targeting control (shNT) after 72 hours (**Figure 3.10A**). RNAi off-target effects were excluded by filtering genes that were either differentially expressed or dimensionally inconsistent between each independent CENPA-targeted shRNA (**Figure 3.11A and 3.11B**). The remaining 427 differentially expressed genes (DEGs) illustrated global transcriptional downregulation in the setting of CENPA depletion (**Figure 3.10B and 3.11C**). Indeed, when conducting ontologic assessments on the RNA-seq dataset, we identified overlap between concepts bound by CENPA and concepts that are transcriptionally perturbed, specifically nuclear architecture and organization (**Figure 3.10C**). Formal integrative analysis between the ChIP-seq and RNA-seq data (obtained from two different prostate cancer cell lines for technical reasons) additionally identified a number of genes essential to cellular proliferation and centromere/kinetochore integrity that were both bound by CENPA and differentially expressed in the setting of CENPA depletion (**Figure 3.10D**). *CDC25C*, *CDCA5*, *TOP2A*, and *CENPU*, genes whose expression levels were strongly correlated to CENPA expression in prostate cancer tissue, were all notably downregulated with CENPA depletion as well as bound by CENPA. Of note, cell division was also found to be an additional enriched biologic concept within our RNA-seq data, again exhibiting similarity with earlier correlative findings in tissue. These findings collectively suggest that CENPA is a regulator of transcription for genes important for proliferation and cell cycle progression in prostate cancer.

Discussion

The centromeric histone H3 variant CENPA is overexpressed in cancer and has the propensity to localize to genomic loci that lie outside of the canonical centromere in the setting of overexpression. Yet to date, the functional significance of the ectopic localization of CENPA in the biological setting of malignancy has been largely unexplored. We demonstrate for the first time that CENPA mislocalization has functional consequences through an unexpected role as a regulator of gene expression in prostate cancer. CENPA is infrequently mutated and amplified in metastatic prostate cancer[32,33]. Given the relative genomic stability of the CENPA locus, the observed phenotypic aberrations that are the result of modulating CENPA expression are likely to be epigenetically driven. While histone variants exhibiting aberrant biologic properties in malignancy have been studied extensively, previous thought has been that CENPA's importance to cell division and proliferation is purely a function of its role as a structural node for the CCAN and KMN network. While its role in the centromere is certainly vital, we now show that ectopic deposition of CENPA in prostate cancer likely plays a role in modulating cell division, functioning as a regulator of critical proliferation, cell cycle, and centromere/kinetochore genes. These observations thus suggest an additional critical way in which this much-studied protein can affect cellular proliferation when overexpressed in the setting of cancer.

Given the ubiquitous nature of CENPA overexpression in cancer, it is conceivable that ectopically bound CENPA driving proliferation through transcriptional regulation is a generalizable feature exhibited across numerous cancer types. Previous work shows that while centromeric/kinetochore proteins are infrequently mutated or amplified in cancer, coordinate overexpression of centromeric/kinetochore factors is a common characteristic identified in malignancy[13]. Our findings suggest a potential epigenetic feedforward mechanism by which progressively increasing levels of CENPA drive gene expression of critical proliferation, cell cycle progression, and centromere/kinetochore factors that complement shifting mutational landscapes previously identified through cancer genomics approaches[34]. Intriguingly, AR signaling, the most commonly targeted and mutated molecular signature in metastatic-castration resistant prostate cancer, was not captured by a CENPA focused analysis, a finding that we further confirmed in tissue culture experiments. This phenomenon mirrors EZH2 expression changes in response to androgen stimulation, implicating CENPA as an additional epigenetic factor that may contribute to androgen-refractory progression[35]. Of note, we observed EZH2 expression indeed tracked tightly with CENPA expression in prostate cancer tissue (*SI Appendix*, Fig. S2D).

While we show here that prostate cancer tissue and cell lines overexpress CENPA, previous work suggests that occupation of only ~4% of the alpha-satellite rich centromere is sufficient for producing functional centromeres[36]. Indeed, benign prostatic epithelial cells do proliferate in spite of expressing low levels of CENPA and exhibiting reduced CENPA binding to α-satellite DNA. Loss-of-function experimentation yields pronounced growth inhibition in prostate cancer cell lines while producing no observable phenotype in a benign prostatic epithelial cell line. Moreover, overexpression of CENPA drives proliferation in benign prostatic epithelial cells. These findings are consistent with the presence of an ancillary gene regulatory function for CENPA that further dictates its control of proliferation in the setting of overexpression and malignancy. Work in HeLa cells suggests that overexpressed CENPA is directed to gene regulatory elements through interaction with DAXX, a protein that has been previously been shown to be overexpressed in prostate cancer[37]. It is thus conceivable that DAXX carries out a chaperone like function for CENPA when they are both overexpressed in prostate cancer, though Re-ChIP assays involving

CENPA and DAXX together would be necessary to formally prove the presence of such a mechanism.

In conclusion, we show here that the much-studied centromeric histone H3 variant CENPA likely also has a previously uncharacterized function as an epigenetic regulator of transcriptional activity involving genes important for proliferation, cell cycle progression, and centromere and kinetochore integrity in prostate cancer. CENPA overexpression, driven by as yet uncharacterized oncogenic events, thus potentiates a feedforward loop designed to maintain uncontrolled proliferation in cancer. The ubiquitous nature of CENPA overexpression in other malignancies in addition to prostate cancer suggests that CENPA-driven gene expression will be present across different cancer types, providing a generalizable rationale to exploit CENPA and its downstream targets for therapeutic purposes.

Materials and Methods

Cell Lines and Cell Culture. LnCaP, 22Rv1 and DU145 prostate cancer cell lines were cultured in Roswell Park Memorial Institute (RPMI) medium supplemented with 10% fetal bovine serum (FBS) (Atlanta Biologics) and 1% penicillin/streptomycin (P/S), as was the PNT2 prostatic epithelial cell line. VCaP and PC3 prostate cancer cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS and 1% P/S. The RWPE-1 and 957E-hTERT immortalized prostatic epithelial cell lines were cultured in keratinocyte serum free medium (K-SFM) supplemented with 0.05 mg/mL Bovine Pituitary Extract (BPE) and 5 ng/mL epidermal growth factor (EGF). All cell lines were grown at 37 °C in a 5% CO2 cell culture incubator, authenticated by short tandem repeat (STR) profiling for genotype validation at the University of Michigan Sequencing Core and tested for Mycoplasma contamination.

Tissue RNA-seq Differential Expression Analysis. Analysis was performed on a compendium of 10,848 poly(A)+ RNA-sequencing (RNA-seq) libraries containing primary cancer tissue, normal tissue, and cancer cell lines from the TCGA, Michigan Center for Translational Pathology, and other public sources. Sample Set Enrichment Analysis (SSEA) was performed as previously described across all libraries to query CENPA expression in a cancer vs. normal fashion[19]. Further analysis restricted to the prostate tissue type cohort was conducted to determine CENPA expression levels at different stages of malignancy.

Ontologic Assessments. CENPA mRNA levels within the cancer cohort were subjected to transcriptome-wide correlation studies against all protein-coding genes. All genes satisfying the r > 0.8 criteria were included in a custom list for ontologic assessments. This list of genes was ranked by the Spearman rho coefficient and subject to pathway analysis. The Database for Annotation, Visualization, and Integrated Discovery (DAVID) tool performed ontologic assessments against Gene Ontology (GO) terms as well as UniProt concepts. Weighted, pre-ranked Gene Set Enrichment Analysis (GSEA) was performed against MSigDB datasets.

Tissue Microarray. CENPA expression in prostatic epithelium was assessed by immunohistochemistry (IHC) on a Tissue Microarray (TMA) using a mouse anti-CENPA antibody. Benign prostate tissue, high grade prostatic intraepithelial neoplasia (HGPIN), localized prostate cancer and metastatic castration resistant prostate cancer (CRPC) tissues were spotted in triplicate on the core (n=58 total tissues, n=174 cores). Staining was evaluated by assessing the most frequent pattern of intensity at 20x in addition to the percentage of cells showing that pattern.

A product score was subsequently calculated (intensity x percentage of cells with the pattern) for each core. Receiver operator characteristic (ROC) curve was generated based on average product scores.

Quantitative Reverse Transcriptase–Polymerase Chain Reaction (qRT-PCR) Assay. The All Prep DNA/RNA Mini Kit (Qiagen) was utilized to isolate RNA from cell lysates. RNase-Free DNase (Qiagen) was used to eliminate contaminating genomic DNA. RNA was quantified by the NanoDrop 2000 (ThermoFisher Scientific) and diluted to 25 ng/µL. The Step One Plus Real-Time PCR System (Applied Biosystems) was utilized for One-Step qRT-PCR reactions and Moloney Murine Leukemia Virus Reverse Transcriptase (Promega) for reverse transcription. Gene-specific primers were designed and subsequently synthesized by IDT Technologies. A relative quantification method was used to analyze qRT-PCR data and subsequently presented as average fold change over an internal reference (as internal reference, GAPDH was utilized). All primers used for qPCR are detailed in **Table 3.4**. Three technical replicates were used in each assay, and all data shown were from three biological replicates.

Lysates, Antibodies and Immunoblotting. Cells were pelleted, dissolved in 4x Laemmli Buffer, sonicated for 30 seconds and immediately placed in ice. Whole cell extracts were then heated for an additional two minutes at 95° C and returned to ice. Samples were subsequently separated on 4-20% SDS-polyacrylamide gels (BioRad) and transferred to polyvinylidene fluoride (PVDF) membranes via wet transfer at 80 V for 90 minutes. Membranes were then incubated in blocking buffer (phosphate buffered saline, 0.1% Tween, (PBS-T) 5% non-fat dry milk) for an hour at room temperature. Primary antibody incubations were conducted with indicated antibodies in blocking buffer at 4° C overnight. Secondary antibody incubations were conducted the following day with species appropriate horseradish peroxidase (HRP) conjugated secondary antibodies. Blots were developed using enhanced chemiluminescence substrate according to manufacturer's protocol (Millipore). Antibodies against CENPA (Abcam ab13939) and GAPDH (ab181602) were used as primary antibodies.

Knockdown and Overexpression Studies. Stable knockdown of CENPA was achieved using the pTRIPZ Tet-On system (Dharmacon). Commercial glycerol stocks of bacteria propagating plasmids containing two different CENPA-directed shRNAs and a non-targeting shRNA were inoculated and cultured for 24 hours. Plasmid DNA was subsequently isolated using the Plasmid Maxi Kit (Qiagen) and sent to the University of Michigan Vector Core for lentiviral production. 22Rv1, DU145, and LnCaP cells were transduced with lentivirus in the presence of 8 μg/mL polybrene. After 24 hours, cells were cultured in the presence of 2 μg/mL puromycin. Knockdown was achieved through culturing cells with doxycycline at a final concentration 2 μg/mL. Doxycycline response was assessed by microscopy, immunoblot, and qRT-PCR. Stable CENPA overexpression was achieved using the pLV system driven by an EF1A promoter (VectorBuilder). Plasmids and lentivirus were prepared as described for the pTRIPZ system. 957E-hTERT cells were incubated with lentivirus in the presence of 8 μg/mL polybrene. After 24 hours, cells were cultured in standard K-SFM and subjected to fluorescence-activated cell sorting (FACS) to select mCherry positive cells by the University of Michigan Flow Cytometry Core. Overexpression was verified by qRT-PCR and immunoblot. Maps of all vectors are provided in Supplemental Information.

R1881 Treatment. To evaluate the effect of androgen signaling, cells were cultured in medium containing 10% charcoal-treated FBS and treated with DMSO vehicle or with 1 nM or 10 nM R1881. After 24 and 48 hrs, RNA was isolated and qRT-PCR was performed as described above using FastStart SYBR Green Mastermix (Roche).

Cell Proliferation Assays. Cells were seeded in T-25 flasks at $1x10^6$ cells per flask. Flasks were evaluated by microscopy 24 hrs following seeding to assess initial confluence. Growth curves were constructed by imaging flasks by microscopy, where the growth curves are generated from confluence measurements acquired from 6 fields per condition, using the NIS Elements microscope imaging software. Proliferation assay was performed for shRNA-mediated knockdown, overexpression, and basal growth experiments.

Crystal Violet Assays. Cells were seeded in triplicate in 6-well plates at $1x10^4$ cells per well. Cells were given 24 hours to adhere and were subsequently subjected to doxycycline treatments at 2 µg/mL. Doxycycline was replenished every 2 days to maintain continuity of the pTRIPZ system. Treatment was discontinued 7 days after induction. Cells were washed with ice-cold PBS and subsequently fixed with methanol. Cells were then stained with 0.5% crystal violet for 10 minutes, washed with water, and air-dried.

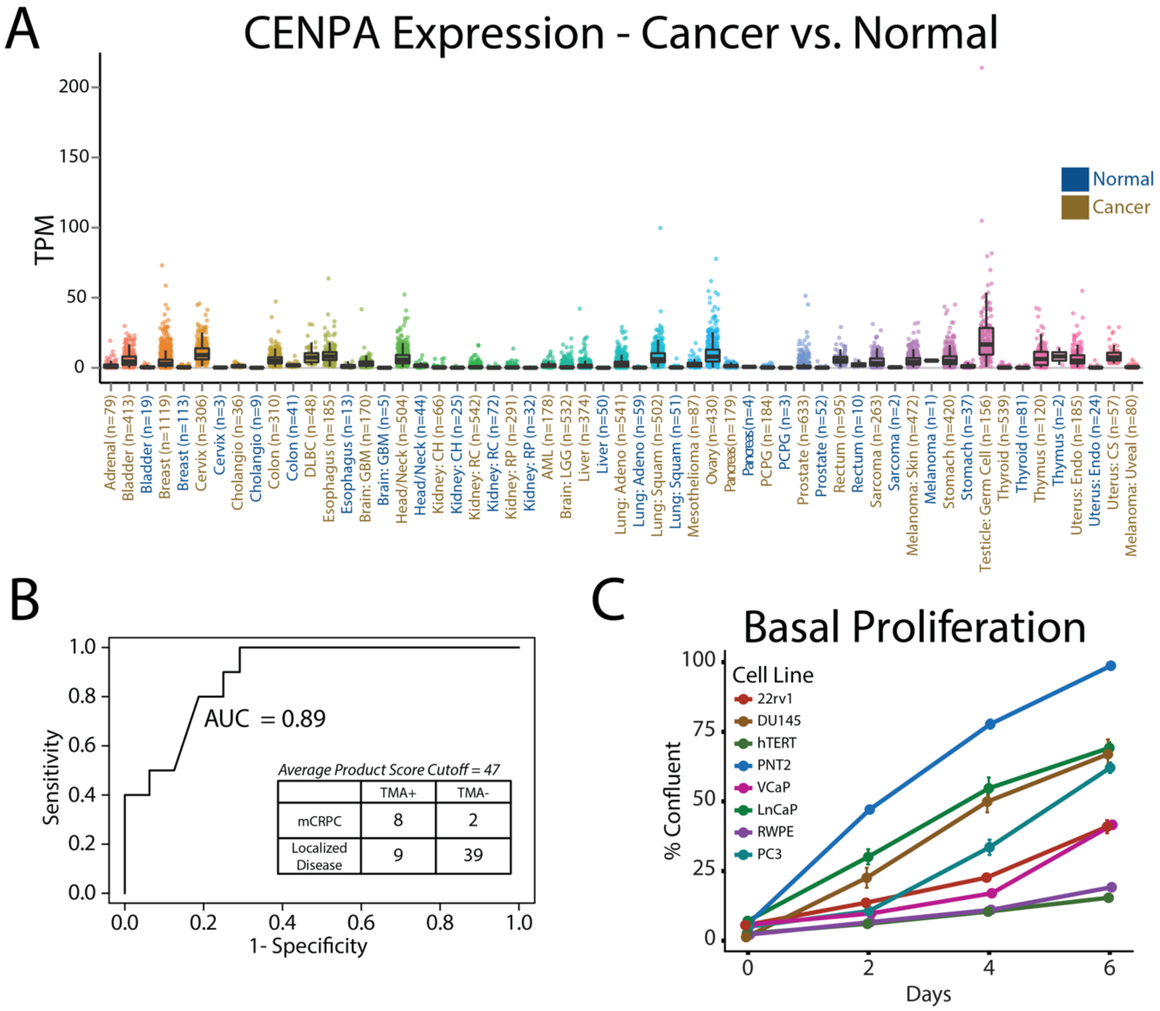
Cell Cycle Analysis. Cells were subjected to 15 minutes of ethanol fixation at -20° C and subsequently collected by centrifugation. Cells were rehydrated at room temperature in PBS, pelleted and re-suspended in 3 µM DAPI diluted in staining buffer (100 mM Tris pH 7.4, 150 mM NaCl, 1 mM $CaCl_2$, 0.5 mM $MgCl_2$, 0.1% Nonidet P-40). Cells were incubated for 15 minutes prior to flow cytometry by University of Michigan Flow Cytometry Core. Cell cycle distribution was evaluated using FCS Express.

ChIP-sequencing Library Preparation and Analysis. Native ChIP was conducted as previously described[38] using a monoclonal antibody against CENPA (Abcam ab13939). CENPA targeted ChIP DNA and input control DNA were prepared for parallel sequencing using the TruSeq ChIP Library Preparation kit (Illumina) according to the manufacturer's protocol. Library preparations were done in conjunction with the University of Michigan Sequencing Core. Paired-end libraries were sequenced with the Illumina HiSeq 4000 (2X150 nucleotide read length) with sequence coverage to >50M total reads per sample. FASTQC was employed to assess overall quality of each sample followed by TrimGalore processing to trim low-quality bases and adapter sequences. Reads were aligned to *hg38* using Bowtie2 (version 2.2.1) with default parameters. Principal component analysis was conducted to determine the degree of variation between samples. PePr (version 1.1.14) was employed to identify CENPA bound regions. P-values were adjusted for multiple testing using the false-discovery rate (FDR) approach and CENPA bound regions were considered significant peaks when p-values $< 1.0 \times 10^{-5}$. Peaks were annotated with annotatr (version 1.0.3). Ontology assessments were conducted using ChIP-Enrich[39].

RNA-sequencing Library Preparation and Analysis. RNA was isolated as described and RNA integrity was evaluated using an Agilent TapeStation. Strand-specific libraries were prepared using the TruSeq Stranded mRNA kit (Illumina) according to the manufacturer's protocol. Library preparations were done in conjunction with the University of Michigan Sequencing Core. Paired-end, strand-specific libraries were sequenced with the Illumina HiSeq 4000 (2X50 nucleotide read length) with sequence coverage to >50 M paired-end reads and >100 M total reads per sample. Sequencing results were run through a computational pipeline to trim low-quality bases and adapters (TrimGalore), align reads (STAR) to a reference genome (UCSC hg38 from iGenomes), quantify gene expression levels (HTseq), and call differentially expressed genes (edgeR). Ontology assessments were conducted using RNA-Enrich[40].

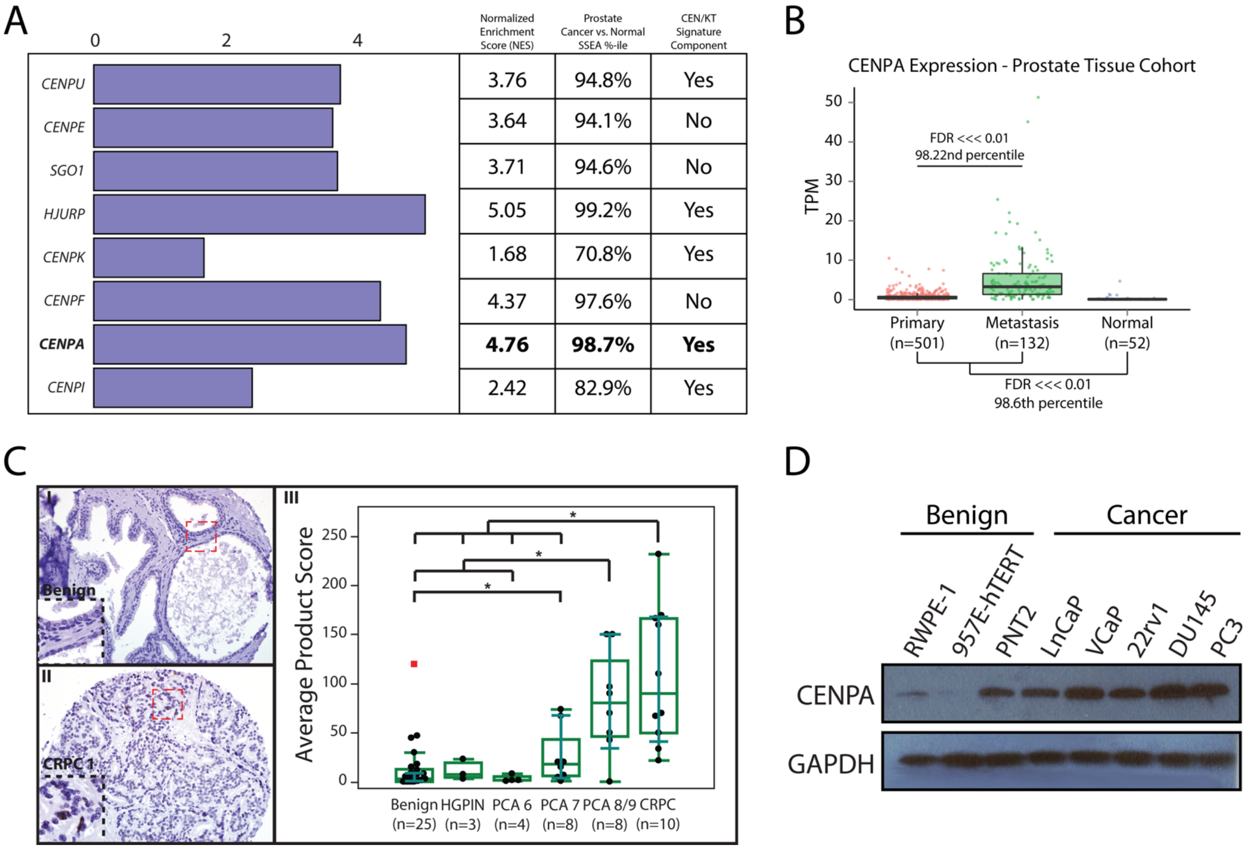
Data Availability. RNA-sequencing and ChIP-sequencing data are deposited in the MIAME-compliant GEO repository (NCBI) with Accession Numbers GSM3639705 - GSM3639722.

Figures



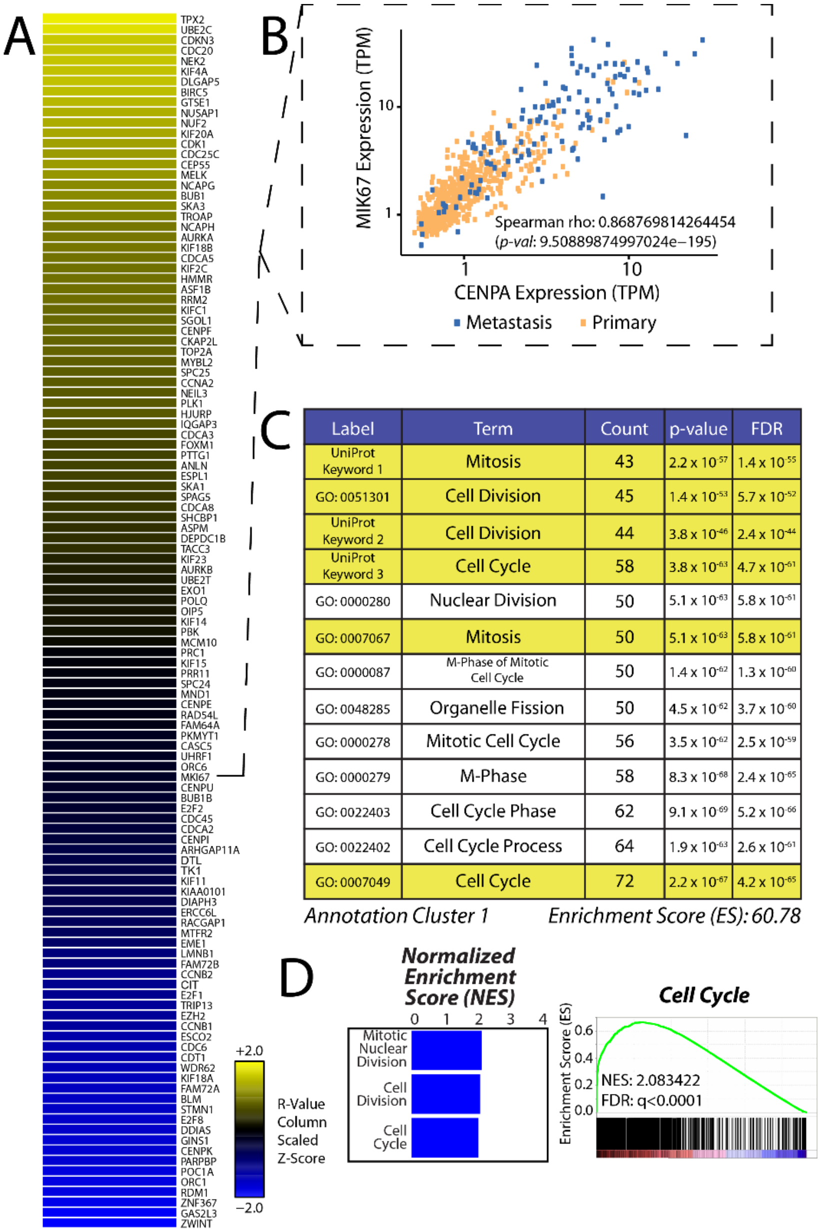
Authors: Anjan K. Saha, Yashar S. Niknafs, Matthew Iyer, Javed Siddiqui, Scott Tomlins, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

Figure 3.1: Characterization of CENPA in cellular proliferation, and cancer. A) Cancer vs. normal analysis conducted across a curated RNA-seq catalogue querying tissue CENPA levels by SSEA. B) Receiver operator characteristic (ROC) separating metastatic castration resistant prostate cancer (mCRPC) from localized disease by CENPA staining in a tissue microarray. C) Proliferation rates of a panel of prostate cancer (LnCaP, VCaP, 22rv1, DU145, and PC3) and benign prostatic epithelial (RWPE-1, 957E-hTERT, and PNT2) cell lines.



Authors: Anjan K. Saha, Yashar S. Niknafs, Matthew Iyer, Javed Siddiqui, Scott Tomlins, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

Figure 3.2: Overexpression of CENPA in prostate cancer. A) Sample Set Enrichment Analysis (SSEA) was used to query a catalogue of curated RNA-seq libraries (n=685) for differentially expressed centromeric genes in the prostate tissue type cohort. Genes were selected based on associations identified in prior studies with cancer progression and were characterized by their inclusion in the previously described CEN/KT signature that negatively impacts therapy response and survival. B) Focused SSEA on CENPA mRNA levels in normal prostate (n=52), primary prostate cancer (n=501), and metastatic prostate cancer (n=132) tissue. C) Tissue Microarray (TMA, n=58 total tissues, n=174 cores) of benign prostate (I), high grade prostatic intraepithelial neoplasia (HGPIN), Gleason grade 6-9 prostate cancer, and castration resistant prostate cancer (CRPC) (II) tissue stained for CENPA, *P<0.05. Staining was evaluated by assessing the most frequent pattern of intensity at 20x and percentage of cells exhibiting that pattern (III). D) Immunoblot for CENPA and GAPDH (loading control) in a panel of benign and malignant prostate cell lines. Note that PNT2, although benign, proliferates the most rapidly of all cell lines tested (**Figure 3.1C**).



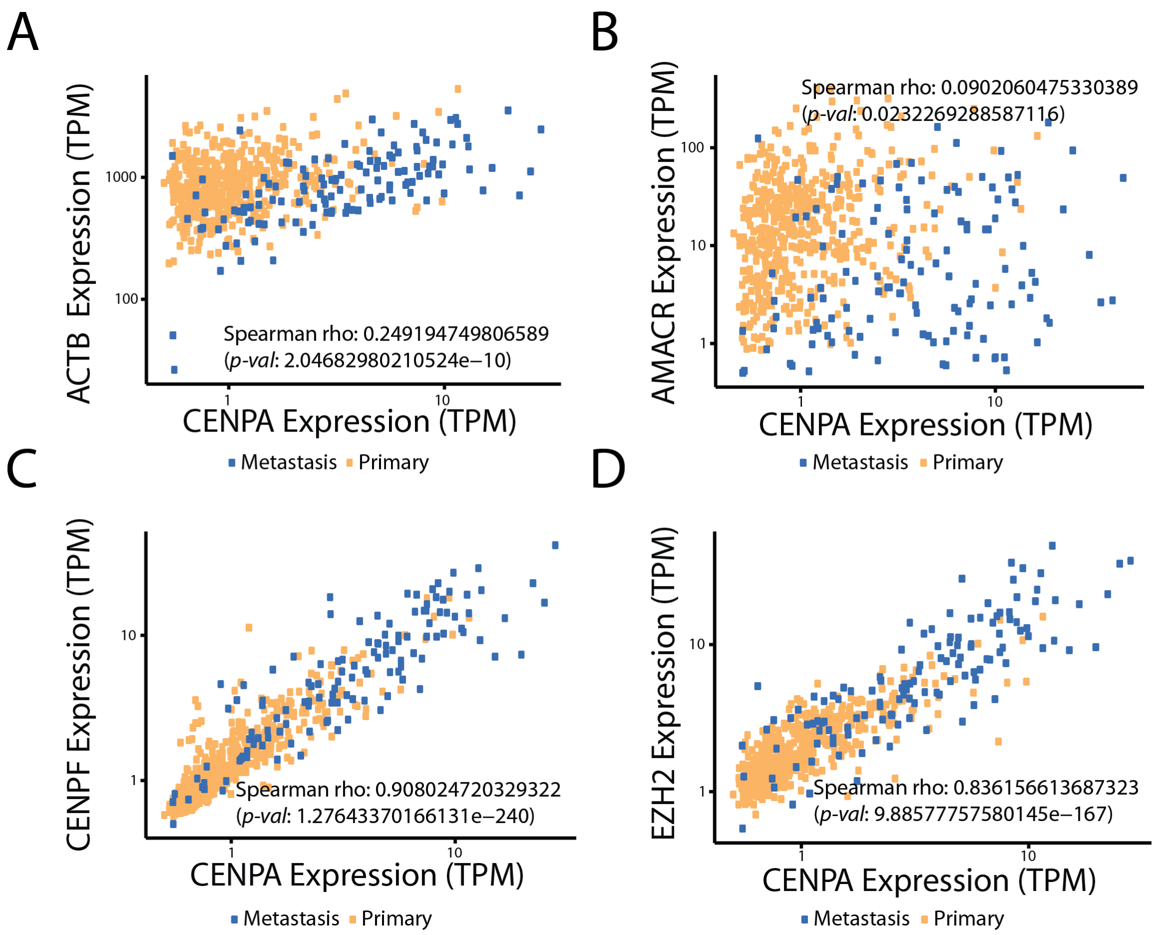
Label	Term	Count	p-value	FDR
UniProt Keyword 1	Mitosis	43	2.2×10^{-57}	1.4×10^{-55}
GO: 0051301	Cell Division	45	1.4×10^{-53}	5.7×10^{-52}
UniProt Keyword 2	Cell Division	44	3.8×10^{-46}	2.4×10^{-44}
UniProt Keyword 3	Cell Cycle	58	3.8×10^{-63}	4.7×10^{-61}
GO: 0000280	Nuclear Division	50	5.1×10^{-63}	5.8×10^{-61}
GO: 0007067	Mitosis	50	5.1×10^{-63}	5.8×10^{-61}
GO: 0000087	M-Phase of Mitotic Cell Cycle	50	1.4×10^{-62}	1.3×10^{-60}
GO: 0048285	Organelle Fission	50	4.5×10^{-62}	3.7×10^{-60}
GO: 0000278	Mitotic Cell Cycle	56	3.5×10^{-62}	2.5×10^{-59}
GO: 0000279	M-Phase	58	8.3×10^{-68}	2.4×10^{-65}
GO: 0022403	Cell Cycle Phase	62	9.1×10^{-69}	5.2×10^{-66}
GO: 0022402	Cell Cycle Process	64	1.9×10^{-63}	2.6×10^{-61}
GO: 0007049	Cell Cycle	72	2.2×10^{-67}	4.2×10^{-65}

Annotation Cluster 1 *Enrichment Score (ES): 60.78*

Authors: Anjan K. Saha, Yashar S. Niknafs, Matthew Iyer, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

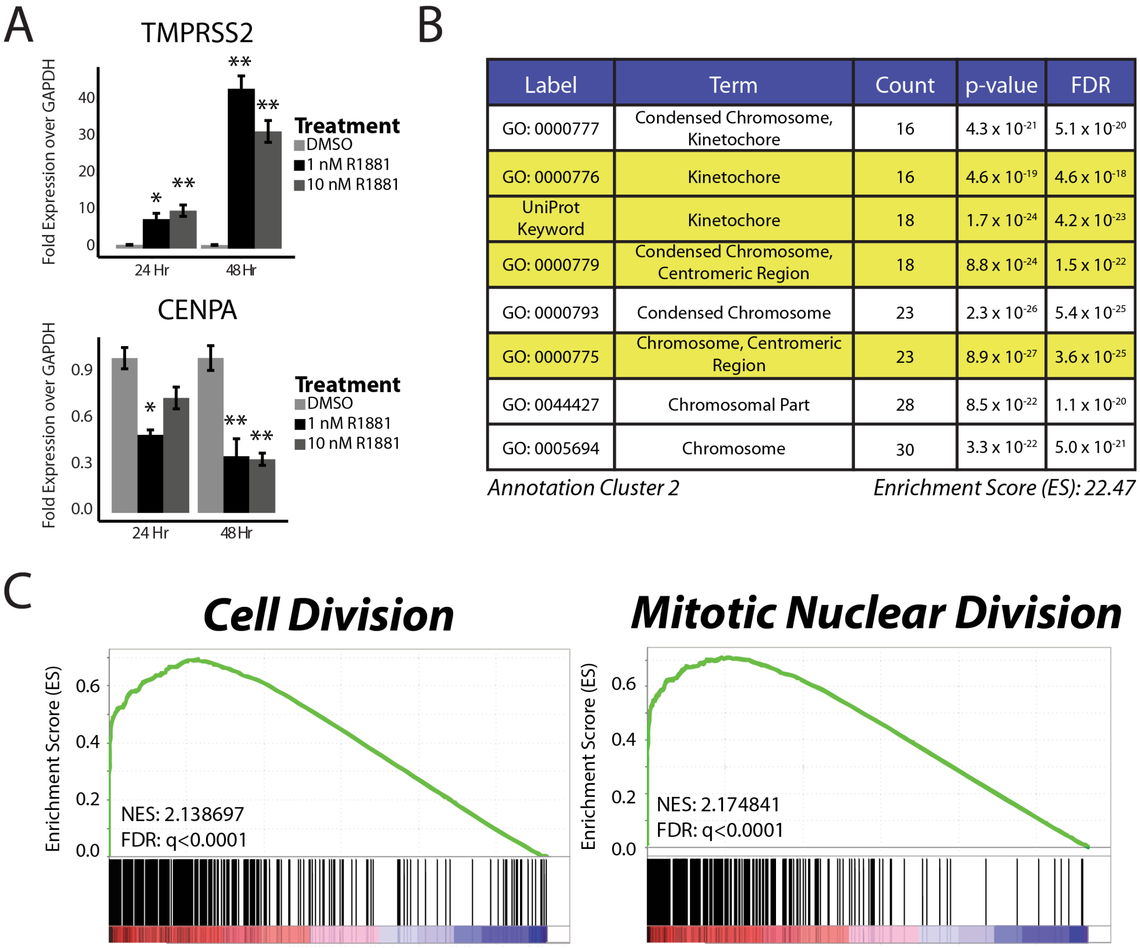
Figure 3.3: Proliferation signature associated with CENPA. A) CENPA mRNA levels from SSEA subjected to a transcriptome-wide correlation. Results were rank-ordered by the strength of correlation. Heatmap depicts genes that performed at r ≥ 0.8. B) Scatterplot depicting strong concordance between CENPA and the proliferation marker MKI67. C) Top 117 performers from transcriptome-wide correlation subjected to functional annotation analysis using the publicly available Database for Annotation, Visualization, and Integrated Discovery (DAVID). Enriched biological concepts are rank-ordered by their false discovery rate (FDR). D) Independent Gene Set Enrichment Analysis (GSEA) of mitotic nuclear division, cell division, and cell cycle gene signatures conducted on transcriptome-wide correlation values pre-ranked by the strength of

correlation. Barplot depicts enrichment scores from biologic concepts designated along vertical axis (left). Representative enrichment plot from “Cell Cycle” gene signature (right).



Authors: Anjan K. Saha, Yashar S. Niknafs, Matthew Iyer, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

Figure 3.4: Transcriptome-wide correlation against CENPA mRNA levels in prostate cancer. A-D) Individual scatterplots depicting correlation strength between CENPA expression, previously characterized prostate cancer pathogenesis factors (AMACR, CENPF, and EZH2) and a housekeeping control gene (ACTB).

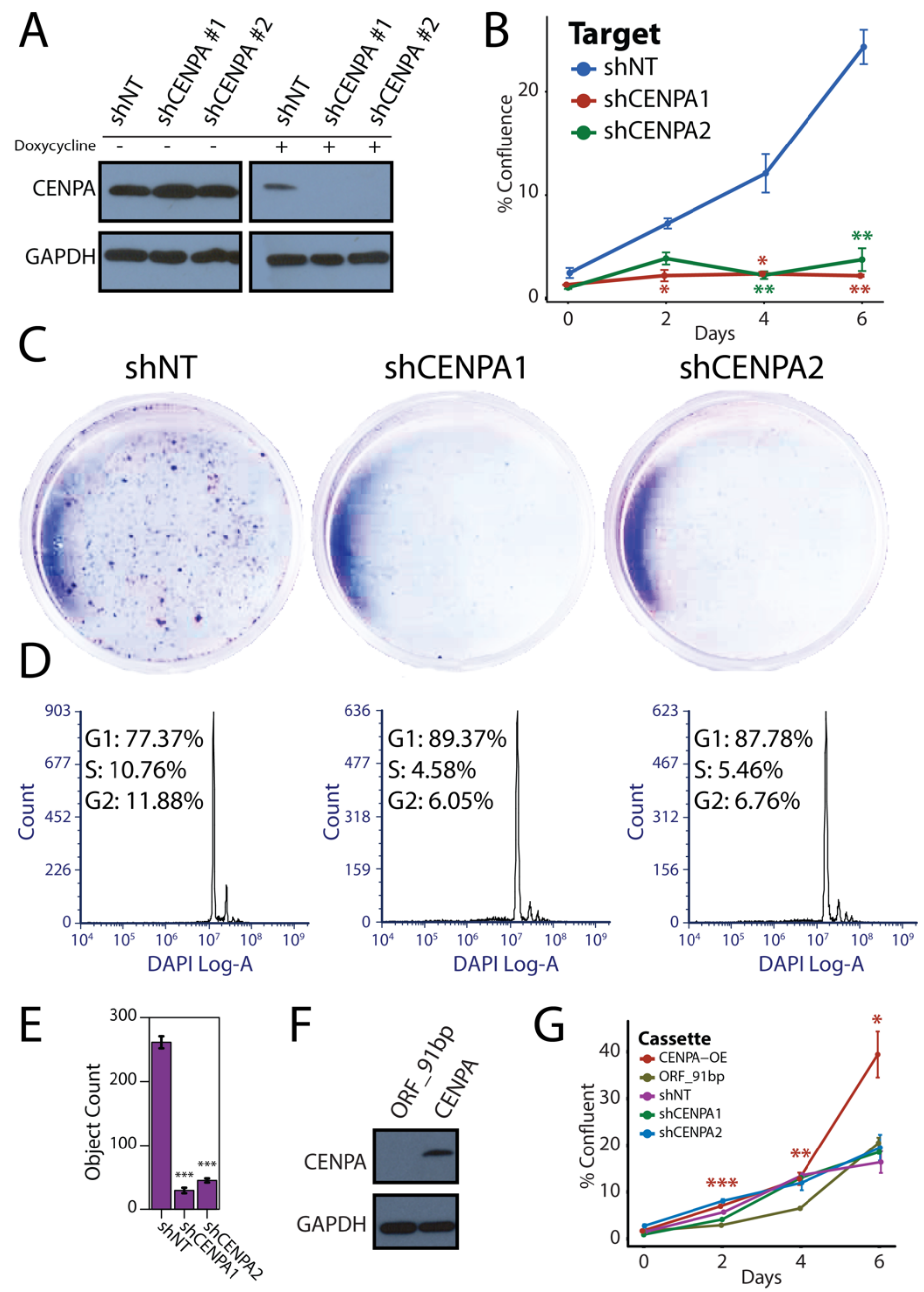


Label	Term	Count	p-value	FDR
GO: 0000777	Condensed Chromosome, Kinetochore	16	4.3×10^{-21}	5.1×10^{-20}
GO: 0000776	Kinetochore	16	4.6×10^{-19}	4.6×10^{-18}
UniProt Keyword	Kinetochore	18	1.7×10^{-24}	4.2×10^{-23}
GO: 0000779	Condensed Chromosome, Centromeric Region	18	8.8×10^{-24}	1.5×10^{-22}
GO: 0000793	Condensed Chromosome	23	2.3×10^{-26}	5.4×10^{-25}
GO: 0000775	Chromosome, Centromeric Region	23	8.9×10^{-27}	3.6×10^{-25}
GO: 0044427	Chromosomal Part	28	8.5×10^{-22}	1.1×10^{-20}
GO: 0005694	Chromosome	30	3.3×10^{-22}	5.0×10^{-21}

Annotation Cluster 2 *Enrichment Score (ES): 22.47*

Authors: Anjan K. Saha, Yashar S. Niknafs, Matthew Iyer, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

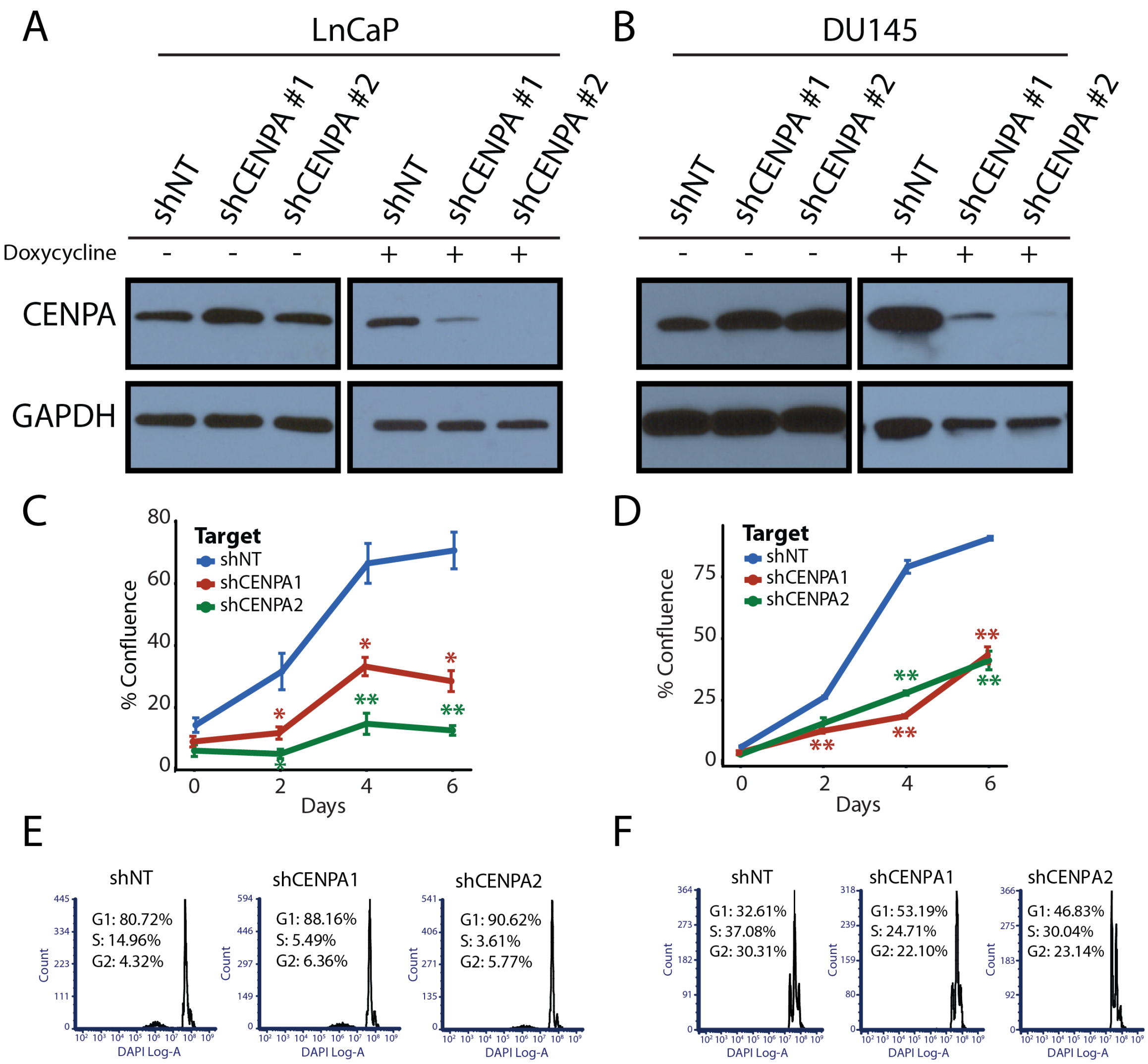
Figure 3.5: Association between CENPA mRNA levels and cell division in prostate cancer. A) VCaP prostate cancer cell lines treated with the androgen agonist R1881 and evaluated for CENPA and TMPRSS2 (androgen-responsive positive control) expression by qRT-PCR. *P<0.05, **P<0.01, comparing to DMSO for each condition and time point via Student's t-test. B) Additional biological concepts identified as associated with CENPA expression through functional annotation analysis using DAVID, highlighting enrichments in concepts important for centromeric and kinetochore integrity. C) Representative enrichment plots from GSEA of "Cell Division" (left) and "Mitotic Nuclear Division" (right) gene signatures conducted on transcriptome-wide correlation values pre-ranked by the strength of correlation.



Authors: Anjan K. Saha, Rafael Contreras-Galindo, Monica Palande, Claire Wang, Brian Qian, Elizabeth Ward, Tara Tang, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

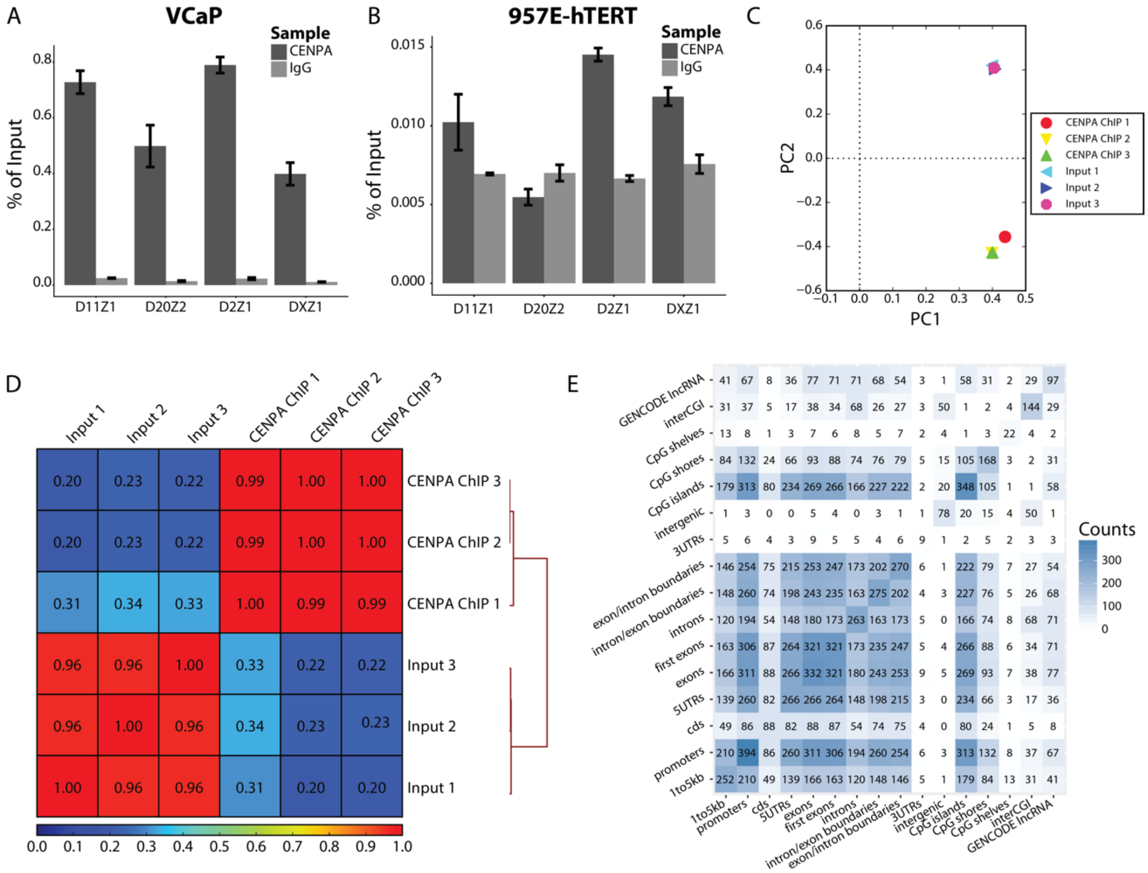
Figure 3.6: Functional importance of CENPA in prostate cancer cells. A) Immunoblot for CENPA and GAPDH in 22Rv1 cells expressing a doxycycline inducible vector encoding a non-targeted and two independent CENPA-targeted shRNAs. B) Growth curve depicting proliferation over 7 days following doxycycline induction in CENPA knockdown cell lines. Error bars represent the standard error of three biologic replicates. *P<0.05, **P<0.01, comparing to shNT for each condition via Student's t-test. C) Crystal violet cell proliferation assay conducted 7 days post-doxycycline induction. D) Cell cycle analysis with DAPI in CENPA shRNA-depleted cells

compared to shNT. E) Quantification of crystal violet colonies in panel C. Error bars represent the s.e.m. of three biologic replicates. F) Immunoblot for CENPA and GAPDH in 957E-hTERT benign prostatic epithelial cells expressing a vector encoding a constitutively active CENPA construct (CENPA-OE). G) Growth curve depicting proliferation over 7 days following CENPA overexpression or knockdown in 957E-hTERT cells. Error bars represent the standard error of three biologic replicates. *P<0.05, **P<0.01, ***P<0.001, comparing CENPA-OE to ORF_91bp (vector control) via Student's t-test.



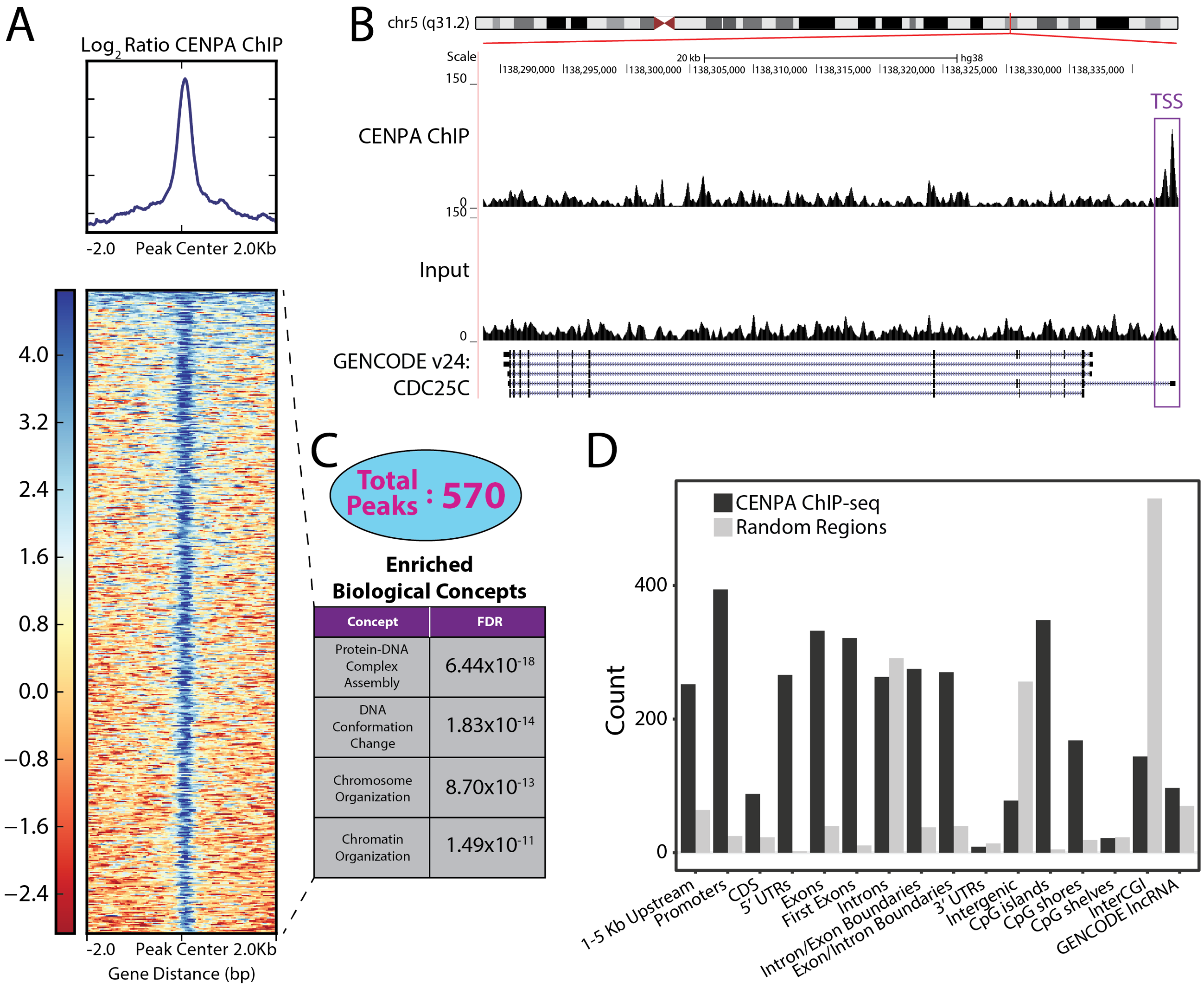
Authors: Anjan K. Saha, Rafael Contreras-Galindo, Monica Palande, Claire Wang, Brian Qian, Elizabeth Ward, Tara Tang, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

Figure 3.7: Panel of prostate cancer cell lines subjected to CENPA depletion. A, B) Immunoblot for CENPA and GAPDH in LnCaP and DU145 cells expressing a doxycycline inducible vector encoding a non-targeted and two independent CENPA-targeted shRNAs. C, D) Growth curve depicting proliferation over 7 days following doxycycline induction in CENPA knockdown cell lines (left – LnCaP, right – DU145). Error bars represent the standard error of three biologic replicates. *P<0.05, **P<0.01, comparing to shNT for each condition via Student's t-test. E, F) Cell cycle analysis with DAPI in CENPA shRNA-depleted LnCaP and DU145 cells compared to shNT.



Authors: Anjan K. Saha, Tingting Qin, Karthik Padmanabhan, Maureen Sartor, Gil S. Omenn, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

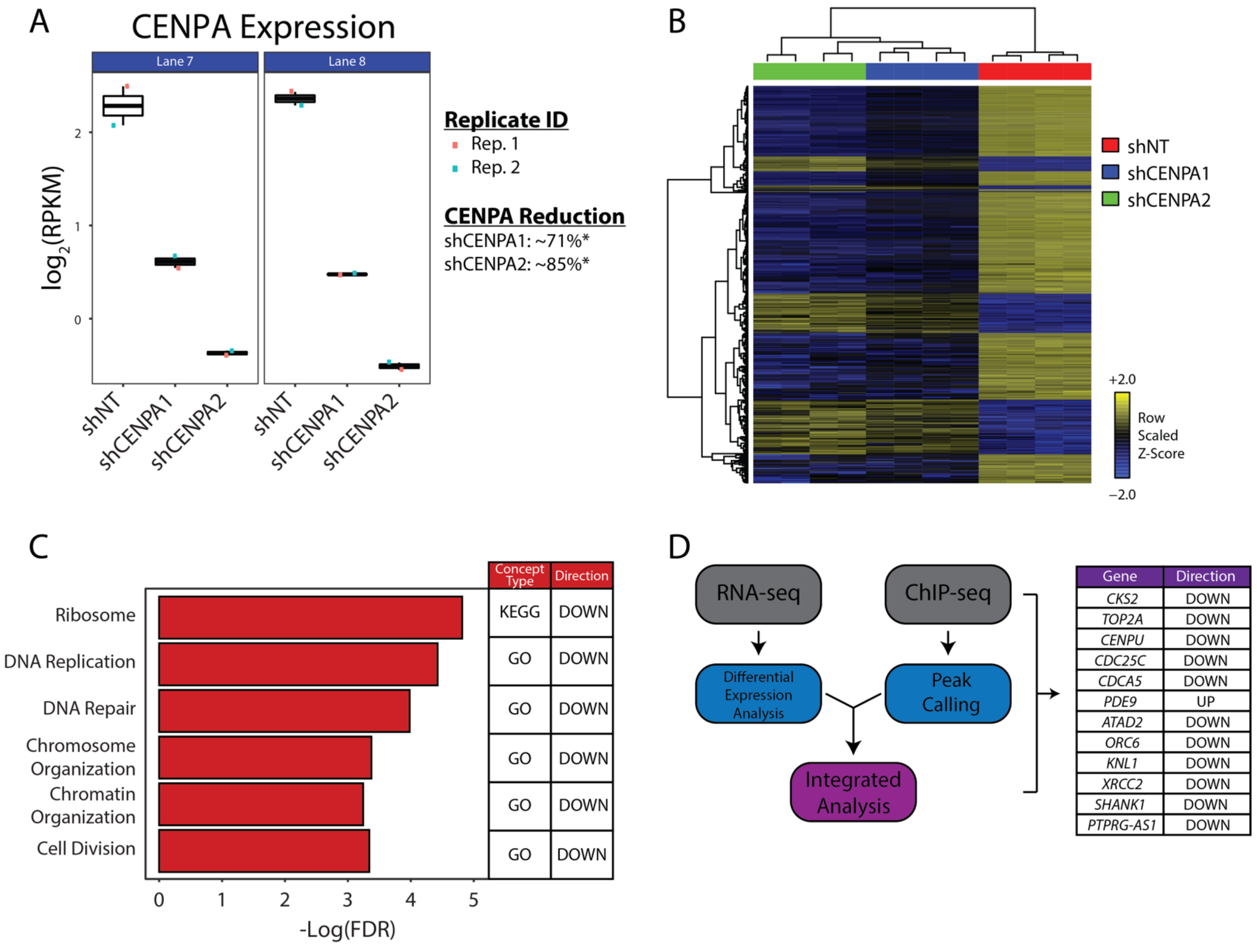
Figure 3.8: CENPA N-ChIP-seq validation. A) Validation of native chromatin immunoprecipitation (N-ChIP) efficiency of CENPA, through PCR targeting individual centromeric repeats. B) N-ChIP-PCR conducted in benign prostatic epithelial cell line 957E-hTERT. C) Principal Component Analysis (PCA) to determine variation in N-ChIP-seq replicates. D) Heatmap depicting correlation strength between each individual N-ChIP-seq sample. Bottom left and top right indicate near perfect consistency between ChIP and input replicates, further validated by the unsupervised hierarchical clustering (see dendrogram along the vertical component). E) Matrix depicting the degree of overlap of CENPA occupancy between different genomic regions. Abbreviations: CDS – coding sequence, UTR – untranslated region, CGI – CpG Island.



Concept	FDR
Protein-DNA Complex Assembly	$6.44\text{x}10^{-18}$
DNA Conformation Change	$1.83\text{x}10^{-14}$
Chromosome Organization	$8.70\text{x}10^{-13}$
Chromatin Organization	$1.49\text{x}10^{-11}$

Authors: Anjan K. Saha, Tingting Qin, Karthik Padmanabhan, Maureen Sartor, Gil S. Omenn, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

Figure 3.9: Deposition of CENPA at regulatory regions across the genome in prostate cancer cells. A) Heatmap across 4 kb windows of CENPA ChIP vs. input signals centered at the CENPA peaks. B) UCSC Genome Browser illustration of CENPA binding to transcriptional start site (TSS) of the *CDC25C* gene on chromosome 5. C) 569 CENPA peaks were subjected to Gene Ontology assessment. Representative concepts are rank ordered by their FDR. D) CENPA peaks were annotated against 16 genomic regions relative to known genes. Abbreviations: CDS – coding sequence, UTR – untranslated region, CGI – CpG Island. Peak abundance (black bars) was compared to abundance from random selection (grey bars) within each genomic region.

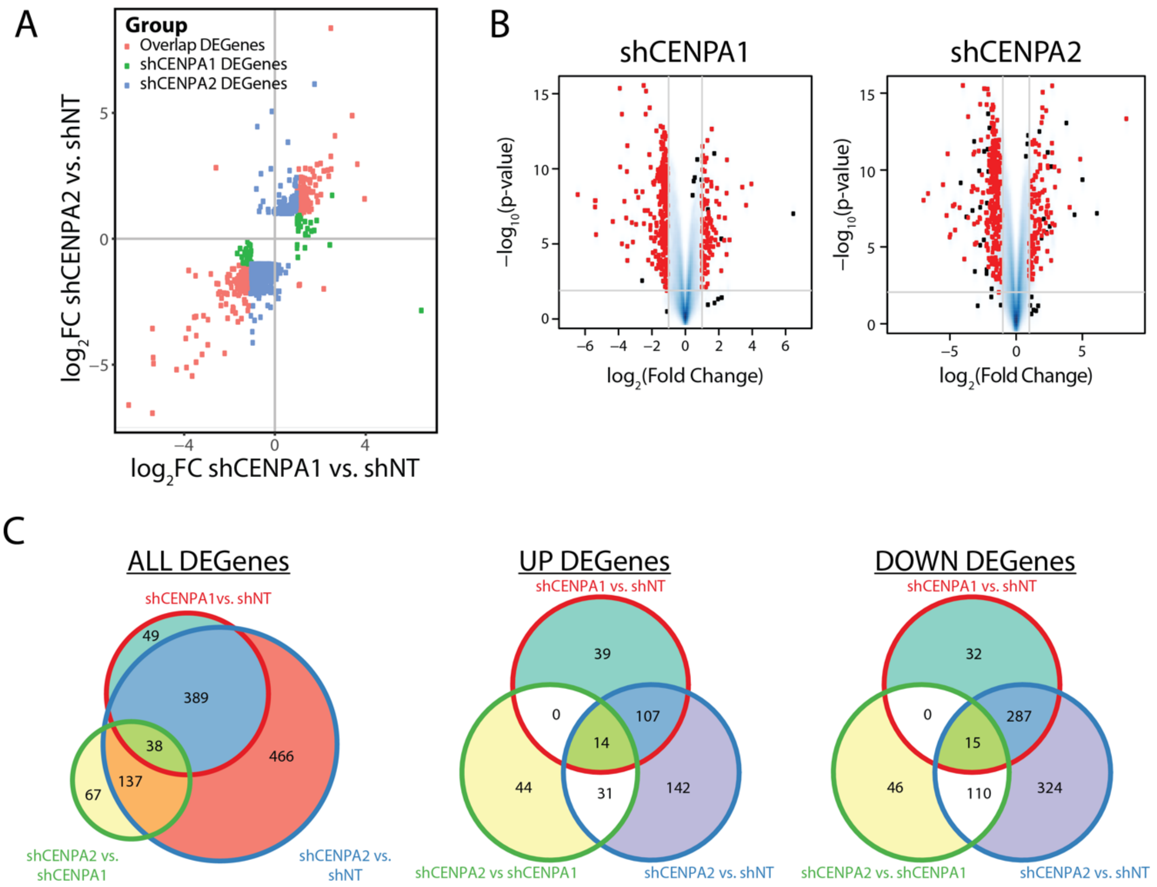


Concept Type	Direction
KEGG	DOWN
GO	DOWN
GO	DOWN
GO	DOWN
GO	DOWN
GO	DOWN

Gene	Direction
CKS2	DOWN
TOP2A	DOWN
CENPU	DOWN
CDC25C	DOWN
CDCA5	DOWN
PDE9	UP
ATAD2	DOWN
ORC6	DOWN
KNL1	DOWN
XRCC2	DOWN
SHANK1	DOWN
PTPRG-AS1	DOWN

Authors: Anjan K. Saha, Tingting Qin, Karthik Padmanabhan, Maureen Sartor, Gil S. Omenn, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

Figure 3.10: Transcriptional profile of CENPA depleted prostate cancer cells. A) Jitterplot reflecting CENPA knockdown efficacy across all replicates. B) Heatmap representation of the 427 differentially expressed genes (DEGs) when comparing a non-targeting shRNA to two independent CENPA targeted shRNAs. Unsupervised hierarchical clustering was performed to group samples (columns) and genes (rows) by similarities in data structure. C) Ontologic assessments conducted on the 427 DEGs using the RNAEnrich program. A subset of significant concepts from the analysis of CENPA depleted cells are depicted. KEGG and GO are databases that reflect ontologies representative of connected biologic processes. D) Transcriptional profile of CENPA depleted 22Rv1 cells merged with CENPA ChIP-seq data from VCaP. Genes listed demonstrate both differential expression with CENPA depletion as well as CENPA binding. Directionality of differential expression for each gene is depicted in the right column. Only genes that satisfy the absolute log fold change > 2 and FDR < 0.05 were considered for integrative analysis.



Authors: Anjan K. Saha, Tingting Qin, Karthik Padmanabhan, Maureen Sartor, Gil S. Omenn, Scott Gitlin, Arul M. Chinnaiyan, David M. Markovitz

Figure 3.11: CENPA-depletion RNA-seq quality control. A) Scatterplot comparing the directionality of differentially expressed genes from the two independent CENPA-targeted shRNAs. B) Volcano plot depicting significance against fold change (FC) between shNT and two independent CENPA-targeted shRNAs. Genes that satisfied the absolute FC > 1.5 (blue lines) and p-value < 0.01 (top right and top left) criteria were considered differentially expressed. C) Venn diagrams illustrating the overlap in differential gene expression by comparison of analysis pairs for all differentially expressed genes (DEGs), upregulated DEGs and downregulated DEGs.

Tables

Table 3.1: Cancer vs. normal sample set enrichment analysis performed across tissue types for tissue CENPA expression.

*CH: Chromophobe Renal Cell Carcinoma; GBM: Glioblastoma multiforme

Tissue Type	*Enrichment Score (ES)*	*Normalized ES (NES)*	*FDR*	*Percentile*
Uterus: Endometrial	0.82400674	4.45339332	<<<0.00001	0.993245958
Prostate	0.62010533	4.7605022	<<<0.00001	0.986821833
Kidney: CH*	0.5436092	2.61474901	0.001257495	0.882046881
Colon	0.82709628	5.88013794	<<<0.00001	0.982360294
Bladder	0.75358915	3.98232218	<<<0.00001	0.994756839
Thyroid	0.27684852	2.61333875	0.000382503	0.654878626
Lung: Squamous	0.98630798	7.85019046	<<<0.00001	0.999219786
Esophagus	0.86499405	3.59421046	<<<0.00001	0.997247915
Head/Neck	0.82944953	6.20450438	<<<0.00001	0.998761127
Stomach	0.76970363	5.34637779	<<<0.00001	0.991716617
Breast	0.79628032	9.41603282	<<<0.00001	0.996246823
Rectum	0.68856615	2.52076371	0.002767945	0.90579198
Kidney: Renal Cell	0.75847375	6.87649952	<<<0.00001	0.975200584
Cholangiocarcinoma	1	3.38185023	<<<0.00001	0.867112947
Kidney: Renal Papillary	0.75839376	4.50619268	<<<0.00001	0.983743856
Lung: Adenocarcinoma	0.86407524	7.14574446	<<<0.00001	0.992993828
Liver	0.90532684	6.72891874	<<<0.00001	0.994865907
Brain: GBM*	1	2.66158809	3.06E-06	0.966514945

Table 3.2: Sample set enrichment analyses comparing expression levels of centromere and kinetochore genes in normal prostate tissues, primary prostate cancers, and metastatic prostate cancers.

Transcript ID	*Comparison Name*	*Enrichment Score (ES)*	*Normalized Enrichment Score (NES)*	*FDR*	*Percentile*
ENSG00000115163.13	Metastasis vs. Primary	0.677312	7.644799	<<<0.00001	0.982245411
ENSG00000115163.13	Metastasis vs. Normal	0.831278	5.604697	<<<0.00001	0.983979735
ENSG00000115163.13	Cancer vs. Normal	0.620105	4.760502	<<<0.00001	0.986821833
ENSG00000129810.13	Metastasis vs. Primary	0.733815	8.111382	<<<0.00001	0.988213656
ENSG00000129810.13	Metastasis vs. Normal	0.854931	5.903665	<<<0.00001	0.989456749
ENSG00000129810.13	Cancer vs. Normal	0.474081	3.714663	4.43E-07	0.946241884
ENSG00000117724.11	Metastasis vs. Primary	0.674255	7.850147	<<<0.00001	0.984572651
ENSG00000117724.11	Metastasis vs. Normal	0.832376	5.956698	<<<0.00001	0.990244069
ENSG00000117724.11	Cancer vs. Normal	0.551631	4.368803	<<<0.00001	0.976147243
ENSG00000123485.10	Metastasis vs. Primary	0.652578	7.61156	<<<0.00001	0.98187005
ENSG00000123485.10	Metastasis vs. Normal	0.844991	5.728154	<<<0.00001	0.986033615
ENSG00000123485.10	Cancer vs. Normal	0.643455	5.051252	<<<0.00001	0.992104105
ENSG00000138778.10	Metastasis vs. Primary	0.7042	8.101394	<<<0.00001	0.988025975
ENSG00000138778.10	Metastasis vs. Normal	0.874649	6.131515	<<<0.00001	0.993256427
ENSG00000138778.10	Cancer vs. Normal	0.450099	3.641762	4.43E-07	0.941152195
ENSG00000151725.10	Metastasis vs. Primary	0.605186	7.05999	<<<0.00001	0.974775722
ENSG00000151725.10	Metastasis vs. Normal	0.80322	5.746591	<<<0.00001	0.986581317
ENSG00000151725.10	Cancer vs. Normal	0.468724	3.758717	4.43E-07	0.948470342
ENSG00000102384.12	Metastasis vs. Primary	0.682269	7.739223	<<<0.00001	0.983446567
ENSG00000102384.12	Metastasis vs. Normal	0.848675	5.837062	<<<0.00001	0.988053264
ENSG00000102384.12	Cancer vs. Normal	0.308641	2.41594	0.003485	0.828656322
ENSG00000123219.11	Metastasis vs. Primary	0.654545	7.487768	<<<0.00001	0.980593822
ENSG00000123219.11	Metastasis vs. Normal	0.758751	5.215585	<<<0.00001	0.974531886
ENSG00000123219.11	Cancer vs. Normal	0.214367	1.679269	0.121066	0.708237042

Table 3.3: Transcriptome-wide correlation results against CENPA expression in prostate cancer tissue samples (n=633) rank ordered by correlation strength. Table restricted to genes with r > 0.8.

Gene ID	*Gene Name*	*Gene Type*	*Coefficient*	*p-value*
ENSG00000115163.13	CENPA	protein_coding	1	0
ENSG00000088325.14	TPX2	protein_coding	0.953247713	0
ENSG00000175063.15	UBE2C	protein_coding	0.949233385	0
ENSG00000100526.18	CDKN3	protein_coding	0.942322286	3.90E-302
ENSG00000117399.12	CDC20	protein_coding	0.940619244	2.88E-298
ENSG00000117650.11	NEK2	protein_coding	0.940405145	8.65E-298
ENSG00000090889.11	KIF4A	protein_coding	0.939524776	7.66E-296
ENSG00000126787.11	DLGAP5	protein_coding	0.939319726	2.16E-295
ENSG00000089685.13	BIRC5	protein_coding	0.938308668	3.36E-293
ENSG00000075218.17	GTSE1	protein_coding	0.936094029	1.60E-288
ENSG00000137804.11	NUSAP1	protein_coding	0.933599848	1.88E-283
ENSG00000143228.11	NUF2	protein_coding	0.932731673	9.85E-282
ENSG00000112984.10	KIF20A	protein_coding	0.931277019	6.64E-279
ENSG00000170312.14	CDK1	protein_coding	0.929232336	4.96E-275
ENSG00000158402.17	CDC25C	protein_coding	0.928026105	8.43E-273
ENSG00000138180.14	CEP55	protein_coding	0.927882891	1.54E-272
ENSG00000165304.6	MELK	protein_coding	0.926369475	8.45E-270
ENSG00000109805.8	NCAPG	protein_coding	0.922418309	6.45E-263
ENSG00000169679.13	BUB1	protein_coding	0.92241255	6.60E-263
ENSG00000165480.14	SKA3	protein_coding	0.922179349	1.64E-262
ENSG00000135451.11	TROAP	protein_coding	0.919163683	1.62E-257
ENSG00000121152.8	NCAPH	protein_coding	0.916186482	9.00E-253
ENSG00000087586.16	AURKA	protein_coding	0.916174417	9.40E-253
ENSG00000186185.12	KIF18B	protein_coding	0.916080692	1.32E-252
ENSG00000146670.8	CDCA5	protein_coding	0.915046688	5.30E-251
ENSG00000142945.11	KIF2C	protein_coding	0.914678507	1.95E-250
ENSG00000072571.18	HMMR	protein_coding	0.914549332	3.08E-250
ENSG00000105011.7	ASF1B	protein_coding	0.91427943	7.97E-250
ENSG00000171848.12	RRM2	protein_coding	0.913730469	5.46E-249
ENSG00000237649.6	KIFC1	protein_coding	0.912124451	1.41E-246
ENSG00000129810.13	SGOL1	protein_coding	0.911795314	4.36E-246
ENSG00000117724.11	CENPF	protein_coding	0.911751692	5.06E-246
ENSG00000169607.11	CKAP2L	protein_coding	0.910515383	3.33E-244
ENSG00000131747.13	TOP2A	protein_coding	0.910413271	4.69E-244
ENSG00000101057.14	MYBL2	protein_coding	0.908588507	2.01E-241
ENSG00000152253.7	SPC25	protein_coding	0.908524197	2.49E-241
ENSG00000145386.8	CCNA2	protein_coding	0.907887864	2.00E-240
ENSG00000109674.3	NEIL3	protein_coding	0.90739416	9.94E-240

ENSG00000166851.13	PLK1	protein_coding	0.907201273	1.86E-239
ENSG00000123485.10	HJURP	protein_coding	0.90714193	2.25E-239
ENSG00000183856.9	IQGAP3	protein_coding	0.90557122	3.45E-237
ENSG00000111665.10	CDCA3	protein_coding	0.901495447	1.09E-231
ENSG00000111206.11	FOXM1	protein_coding	0.901026443	4.51E-231
ENSG00000164611.11	PTTG1	protein_coding	0.900555263	1.87E-230
ENSG00000011426.9	ANLN	protein_coding	0.89990179	1.32E-229
ENSG00000135476.10	ESPL1	protein_coding	0.89973951	2.15E-229
ENSG00000154839.8	SKA1	protein_coding	0.899334885	7.16E-229
ENSG00000076382.15	SPAG5	protein_coding	0.897243566	3.33E-226
ENSG00000134690.9	CDCA8	protein_coding	0.897162757	4.21E-226
ENSG00000171241.7	SHCBP1	protein_coding	0.895676201	3.05E-224
ENSG00000066279.15	ASPM	protein_coding	0.894498322	8.66E-223
ENSG00000035499.11	DEPDC1B	protein_coding	0.894092574	2.72E-222
ENSG00000013810.17	TACC3	protein_coding	0.893768311	6.76E-222
ENSG00000137807.12	KIF23	protein_coding	0.89050008	5.58E-218
ENSG00000178999.11	AURKB	protein_coding	0.889938345	2.56E-217
ENSG00000077152.8	UBE2T	protein_coding	0.887617124	1.26E-214
ENSG00000174371.15	EXO1	protein_coding	0.887581464	1.38E-214
ENSG00000051341.12	POLQ	protein_coding	0.887002447	6.36E-214
ENSG00000104147.7	OIP5	protein_coding	0.886376225	3.28E-213
ENSG00000118193.10	KIF14	protein_coding	0.88611361	6.50E-213
ENSG00000168078.8	PBK	protein_coding	0.885390953	4.24E-212
ENSG00000065328.15	MCM10	protein_coding	0.88287278	2.65E-209
ENSG00000198901.12	PRC1	protein_coding	0.876931157	5.93E-203
ENSG00000163808.15	KIF15	protein_coding	0.876690103	1.06E-202
ENSG00000068489.11	PRR11	protein_coding	0.875783463	9.16E-202
ENSG00000161888.10	SPC24	protein_coding	0.874162081	4.18E-200
ENSG00000121211.6	MND1	protein_coding	0.873879568	8.09E-200
ENSG00000138778.10	CENPE	protein_coding	0.873655133	1.37E-199
ENSG00000085999.10	RAD54L	protein_coding	0.87323259	3.65E-199
ENSG00000129195.14	FAM64A	protein_coding	0.873200215	3.93E-199
ENSG00000127564.15	PKMYT1	protein_coding	0.870136858	4.39E-196
ENSG00000137812.18	CASC5	protein_coding	0.869693748	1.19E-195
ENSG00000276043.3	UHRF1	protein_coding	0.8690451	5.13E-195
ENSG00000091651.7	ORC6	protein_coding	0.868969543	6.08E-195
ENSG00000148773.11	MKI67	protein_coding	0.867715775	9.94E-194
ENSG00000151725.10	CENPU	protein_coding	0.867167902	3.34E-193
ENSG00000156970.11	BUB1B	protein_coding	0.865106175	3.05E-191
ENSG00000007968.6	E2F2	protein_coding	0.864803155	5.88E-191
ENSG00000093009.8	CDC45	protein_coding	0.864549042	1.02E-190
ENSG00000184661.12	CDCA2	protein_coding	0.861725496	4.25E-188
ENSG00000102384.12	CENPI	protein_coding	0.8615679	5.93E-188

ENSG00000198826.9	ARHGAP11A	protein_coding	0.858654441	2.59E-185
ENSG00000143476.16	DTL	protein_coding	0.858074052	8.57E-185
ENSG00000167900.10	TK1	protein_coding	0.85737153	3.61E-184
ENSG00000138160.5	KIF11	protein_coding	0.85714429	5.75E-184
ENSG00000166803.9	KIAA0101	protein_coding	0.855859296	7.80E-183
ENSG00000139734.16	DIAPH3	protein_coding	0.855388765	2.02E-182
ENSG00000186871.6	ERCC6L	protein_coding	0.853544067	8.05E-181
ENSG00000161800.11	RACGAP1	protein_coding	0.851907563	2.03E-179
ENSG00000146410.10	MTFR2	protein_coding	0.85137979	5.71E-179
ENSG00000154920.13	EME1	protein_coding	0.848101678	3.20E-176
ENSG00000113368.10	LMNB1	protein_coding	0.841708517	4.81E-171
ENSG00000188610.11	FAM72B	protein_coding	0.841587693	6.00E-171
ENSG00000157456.6	CCNB2	protein_coding	0.836448844	5.92E-167
ENSG00000122966.12	CIT	protein_coding	0.835824278	1.77E-166
ENSG00000101412.12	E2F1	protein_coding	0.835231932	4.98E-166
ENSG00000071539.12	TRIP13	protein_coding	0.83289595	2.84E-164
ENSG00000106462.9	EZH2	protein_coding	0.832630908	4.47E-164
ENSG00000134057.13	CCNB1	protein_coding	0.83135339	3.96E-163
ENSG00000171320.13	ESCO2	protein_coding	0.826995672	5.88E-160
ENSG00000094804.8	CDC6	protein_coding	0.825740153	4.64E-159
ENSG00000167513.7	CDT1	protein_coding	0.824746176	2.35E-158
ENSG00000075702.15	WDR62	protein_coding	0.823395844	2.10E-157
ENSG00000121621.6	KIF18A	protein_coding	0.821762945	2.90E-156
ENSG00000196550.9	FAM72A	protein_coding	0.820033411	4.53E-155
ENSG00000197299.9	BLM	protein_coding	0.819740648	7.19E-155
ENSG00000117632.19	STMN1	protein_coding	0.819190897	1.71E-154
ENSG00000129173.11	E2F8	protein_coding	0.818021052	1.07E-153
ENSG00000165490.11	DDIAS	protein_coding	0.817674061	1.84E-153
ENSG00000101003.9	GINS1	protein_coding	0.815175278	8.75E-152
ENSG00000123219.11	CENPK	protein_coding	0.815104073	9.75E-152
ENSG00000185480.10	PARPBP	protein_coding	0.815102512	9.78E-152
ENSG00000164087.6	POC1A	protein_coding	0.811596424	2.00E-149
ENSG00000085840.11	ORC1	protein_coding	0.810709703	7.55E-149
ENSG00000278023.3	RDM1	protein_coding	0.80986939	2.64E-148
ENSG00000165244.6	ZNF367	protein_coding	0.807180403	1.39E-146
ENSG00000139354.9	GAS2L3	protein_coding	0.802631464	9.90E-144
ENSG00000122952.15	ZWINT	protein_coding	0.800761444	1.40E-142

Table 3.4: Primers used in this chapter.

Direction	*Target*	*Method*	*Sequence*
GAPDH F	GAPDH	qRT-PCR	ACATCGCTCAGACACCATG
GAPDH R	GAPDH	qRT-PCR	TGTAGTTGAGGTCAATGAAGGG
CENPA F	CENPA	qRT-PCR	GTGTGGACTTCAATTGGCAAG
CENPA R	CENPA	qRT-PCR	TGCACATCCTTTGGGAAGAG
TMPRSS2 F	TMPRSS2	qRT-PCR	CCTGCAGGGACATGGGCTATA
TMPRSS2 R	TMPRSS2	qRT-PCR	CCGGCACTTGTGTTCAGTTTC
D2Z1 F	D2Z1	ChIP-PCR	TCGTTGGAAACGGGATTGT
D2Z1 R	D2Z1	ChIP-PCR	CTGCTCTATGAAAGGGACTGTT
D11Z1 F	D11Z1	ChIP-PCR	CTTCCTTCGAAACGGGTATATCT
D11Z1 R	D11Z1	ChIP-PCR	GCTCCATCAGCAGGATTGT
DXZ1 F	DXZ1	ChIP-PCR	CGGGATCACCTTCCCATAAC
DXZ1 R	DXZ1	ChIP-PCR	GGTGTTGCAAACCTGAACTATC
D20Z2 F	D20Z2	ChIP-PCR	TGCTTGGAAACGGGAATGT
D20Z2 R	D20Z2	ChIP-PCR	CCTGCTCTATGAAAGGGAATGT

References

1. Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**, 407–421 (2003).
2. Hayden, K. E. Human centromere genomics: now it's personal. *Chromosome Research* **20**, 621–633 (2012).
3. Hayashi, T. *et al.* Mis16 and Mis18 Are Required for CENP-A Loading and Histone Deacetylation at Centromeres. *Cell* **118**, 715–729 (2004).
4. Kim, I. S. *et al.* Roles of Mis18α in Epigenetic Regulation of Centromeric Chromatin and CENP-A Loading. *Molecular Cell* **46**, 260–273 (2012).
5. Foltz, D. R. *et al.* The human CENP-A centromeric nucleosome-associated complex. *Nature Cell Biology* **8**, 458–469 (2006).
6. Hasson, D. *et al.* The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nature Structural & Molecular Biology* **20**, 687–695 (2013).
7. Niikura, Y. *et al.* CENP-A K124 Ubiquitylation Is Required for CENP-A Deposition at the Centromere. *Developmental Cell* **32**, 589–603 (2015).
8. Barnhart, M. C. *et al.* HJURP is a CENP-A chromatin assembly factor sufficient to form a functional de novo kinetochore. *The Journal of Cell Biology* **194**, 229–243 (2011).
9. Foltz, D. R. *et al.* Centromere-Specific Assembly of CENP-A Nucleosomes Is Mediated by HJURP. *Cell* **137**, 472–484 (2009).
10. Falk, S. J. *et al.* CENP-C directs a structural transition of CENP-A nucleosomes mainly through sliding of DNA gyres. *Nature Structural & Molecular Biology* **23**, 204–208 (2016).
11. McKinley, K. L. *et al.* The CENP-L-N Complex Forms a Critical Node in an Integrated Meshwork of Interactions at the Centromere-Kinetochore Interface. *Molecular Cell* **60**, 886–898 (2015).
12. Schleiffer, A. *et al.* CENP-T proteins are conserved centromere receptors of the Ndc80 complex. *Nature Cell Biology* **14**, 604–613 (2012).
13. Thiru, P. *et al.* Kinetochore genes are coordinately up-regulated in human tumors as part of a FoxM1-related cell division program. *Molecular Biology of the Cell* **25**, 1983–1994 (2014).
14. Tomonaga, T. *et al.* Overexpression and mistargeting of centromere protein-A in human primary colorectal cancer. *Cancer Res.* **63**, 3511–3516 (2003).
15. Valdivia, M., Hamdouch, K., Ortiz, M. & Astola, A. CENPA a Genomic Marker for Centromere Activity and Human Diseases. *Current Genomics* **10**, 326–335 (2009).
16. Zhang, W. *et al.* Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nature Communications* **7**, 12619 (2016).
17. Lacoste, N. *et al.* Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. *Mol. Cell* **53**, 631–644 (2014).
18. Athwal, R. K. *et al.* CENP-A nucleosomes localize to transcription factor hotspots and subtelomeric sites in human cancer cells. *Epigenetics Chromatin* **8**, 2 (2015).
19. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
20. Fisher, G. *et al.* Prognostic value of Ki-67 for prostate cancer death in a conservatively managed cohort. *Br. J. Cancer* **108**, 271–277 (2013).

21. Li, R. *et al.* Ki-67 Staining Index Predicts Distant Metastasis and Survival in Locally Advanced Prostate Cancer Treated With Radiotherapy: An Analysis of Patients in Radiation Therapy Oncology Group Protocol 86-10. *Clin Cancer Res* **10**, 4118–4124 (2004).

22. American Cancer Society. Cancer Facts & Figures 2018. Available at: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2018.html. (Accessed: 4th June 2018)

23. Cooperberg, M. R. Prostate cancer: A new look at prostate cancer treatment complications. *Nature Reviews Clinical Oncology* **11**, 304–305 (2014).

24. Smith, C. L. *et al.* Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res.* **46**, D836–D842 (2018).

25. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).

26. Black, B. E. *et al.* Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. *Mol. Cell* **25**, 309–322 (2007).

27. Contreras-Galindo, R. *et al.* Rapid molecular assays to study human centromere genomics. *Genome Res.* **27**, 2040–2049 (2017).

28. Sporn, J. C. *et al.* Histone macroH2A isoforms predict the risk of lung cancer recurrence. *Oncogene* **28**, 3423–3428 (2009).

29. Vardabasso, C. *et al.* Histone variants: emerging players in cancer biology. *Cell. Mol. Life Sci.* **71**, 379–404 (2014).

30. Vardabasso, C. *et al.* Histone Variant H2A.Z.2 Mediates Proliferation and Drug Sensitivity of Malignant Melanoma. *Molecular Cell* **59**, 75–88 (2015).

31. Yuen, B. T. K. & Knoepfler, P. S. Histone H3.3 Mutations: A Variant Path to Cancer. *Cancer Cell* **24**, 567–574 (2013).

32. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **161**, 1215–1228 (2015).

33. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174**, 433-447.e19 (2018).

34. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).

35. Bohrer, L. R., Chen, S., Hallstrom, T. C. & Huang, H. Androgens suppress EZH2 expression via retinoblastoma (RB) and p130-dependent pathways: a potential mechanism of androgen-refractory progression of prostate cancer. *Endocrinology* **151**, 5136–5145 (2010).

36. Bodor, D. L. *et al.* The quantitative architecture of centromeric chromatin. *Elife* **3**, e02137 (2014).

37. Puto, L. A., Brognard, J. & Hunter, T. Transcriptional Repressor DAXX Promotes Prostate Cancer Tumorigenicity via Suppression of Autophagy. *J. Biol. Chem.* **290**, 15406–15420 (2015).

38. Brind'Amour, J. *et al.* An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nat Commun* **6**, 6033 (2015).

39. Welch, R. P. *et al.* ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Res.* **42**, e105 (2014).

40. Lee, C., Patil, S. & Sartor, M. A. RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power. *Bioinformatics* **32**, 1100–1102 (2016).

Chapter 4 – Conclusion: Future Directions for Studying Centromeres

Summary of Dissertation

The aim of the work encapsulated in this thesis was to ascertain the role centromeres play in cancer. The well-established contributions of centromeres to faithfully orchestrating chromosomal segregation, as well as some early data in the literature, bestow biologic rationale to further interrogate this essential subcellular structure in the context of malignancy. The findings presented here reflect the degree to which alterations in centromere genetics and epigenetics have been overlooked in our quest towards a more complete understanding of the molecular determinants of cancer. Two immediate future directions can be readily surmised from this thesis work: 1) instability in centromeric loci requires further elucidation to mechanistically delineate factors that drive the aberrations observed in the genomic centromere; and 2) epigenetic promiscuity exhibited by CENPA in prostate cancer may be generalizable across additional centromeric factors as well as across additional cancer tissue lineages.

Genomic Derangements

Global genome instability in cancer underscored our supposition that centromere DNA undergoes molecular alterations that have gone unnoticed due to technologic shortcomings in NGS methodologies that hinder interrogation of repetitive loci[1]. The repetitive landscape of the genomic centromere thus necessitated that we use a methodology that can successfully navigate low-complexity regions of the human genome. Previous work had characterized a novel PCR-based methodology that takes advantage of the rare divergences in homology between chromosome specific α-satellites to quantify copy numbers of chromosome specific repeat units[2]. The specificity of this assay lends itself to applications that seek to characterize centromeric genetics through copy number assessments of the α-satellites that reside within each chromosome. Prior reports in a number of model organisms described copy number alterations at centromeric loci due to recombination and retrotransposition of centromeric DNA, motivating us to apply this PCR methodology to study potential α-satellite copy number alterations in cancer[3–10]. We thus began by applying this methodology to DNA samples from cancer cell lines, tissues, and normal controls, hypothesizing that our assay would unveil cancer-associated copy number variation in centromeric DNA.

The ease of use of our centromere-specific PCR assay allowed us to broadly examine the genetic landscape of centromeres across numerous cancer types. We found heterogeneous losses in centromeric material in terms of copy number differences between tissues as well as differences between cancer cells/tissues and healthy cells in a manner independent of aneuploid karyotypes. The heterogeneous copy number losses identified in centromeric DNA from cancer cell lines was also identified in both bulk cancer tissue and flow sorted cancer tissue. Within a particular locus, we identified evolution in the sequences of the HERV-K111 virus in the form of gene conversion events that homogenized HERV-K111 sequence relative to K111 copies evaluated within normal cells. Taken together, we confirm our hypothesis that gross derangements in centromeric DNA in the form of copy number alterations and gene conversion is nearly ubiquitous within cancer. Our novel PCR assay provides insight into a genetic locus that has been widely-deemed as impervious to genomic inquiry. These findings thus shed light on a largely overlooked region of the human genome in the context of cancer (**Figure 4.1**).

Epigenetic Rewiring

Numerous reports in the literature have however deemed α-satellite DNA as sufficient but not necessary for centromere function[11–14]. The field widely accepts that human centromere function is defined not just by its genetic underpinnings, but also by its epigenetic composition. It thus follows that these postulates must inform our interrogation of centromeres in cancer. The advent of publicly available databases that have molecularly catalogued cancer tissues (TCGA) facilitated a more nuanced assessment of gene expression signatures that display an overrepresentation of centromeric epigenetic factors. Furthermore, RNA-sequencing and ChIP-sequencing are valuable tools by which centromere dysfunction can be epigenetically unraveled in the setting of malignancy. Underlying derangements in the centromeric locus compelled us to investigate the epigenetic components that traditionally occupy centromeric DNA. We thus leveraged a wide array of NGS-based molecular profiling techniques to investigate epigenetic destabilization of centromeres, using prostate cancer as our primary model system. We hypothesized that overexpression of centromeric factors such as CENPA has functional consequences to prostate cancer pathogenesis.

Mining transcriptomic data within a compendium of 10,848 RNA-sequencing libraries revealed widescale overexpression of CENPA in cancer tissue relative to lineage matched normal tissue. This finding corroborated previous reports of wide-scale overexpression of CENPA in cancer, though in much larger cohort sizes[15–17]. A more focused analysis within the prostate cancer tissue type cohort within this compendium demonstrated overexpression of a number of centromere/kinetochore proteins relative to normal prostate tissue, and further showed a successive increase of CENPA mRNA with disease progression, a finding that was validated at the protein level in cell lines and tissue. Ontologic assessments of transcriptional signatures that correlate tightly with CENPA expression in prostate cancer tissue further revealed significant associations with proliferation, cell division, centromere, and kinetochore concepts, corroborating the notion that CENPA overexpression in prostate cancer is linked to cellular programs that govern cell division. Loss-of- and gain-of-function experimentation in cell lines revealed that CENPA modulation indeed affects proliferative phenotypes, where CENPA depletion impairs prostate cancer proliferation by altering flux through the cell cycle while overexpression of CENPA propels proliferation of benign prostatic epithelial cells. As seen previously in colon cancer cell lines, we demonstrate that in a cell line that overexpresses CENPA, we identify non-canonical binding across the genome to regulatory regions in close proximity to genes involved with nuclear architecture and DNA-replication, including the *CENPA* locus itself. Intriguingly, subsequent RNA-seq analysis on CENPA depleted cells reveals drastic changes in the transcriptional profile of prostate cancer cells. Integrative analysis between ChIP-seq and RNA-seq analysis indeed revealed an intersection between genes that are bound by CENPA and genes that are differentially expressed with CENPA depletion. Of note, this list of genes includes those that encode factors crucial for cell cycle, centromere, and kinetochore integrity. Collectively, our findings implicate a previously uncharacterized function for CENPA in regulating transcription of genes important for cell cycle, centromere, and kinetochore function when overexpressed in the setting of prostate cancer (**Figure 4.1**).

Furthering our Understanding of Centromeres in Cancer

The work outlined in this thesis presents a lens into what remains to be accomplished to more completely delineate the contribution of centromere genetics and epigenetics to the pathogenesis of cancer. While we present a previously uncharacterized taxonomy of the molecular irregularities that arise within centromeres in neoplastic settings, much work is required to inform our understanding of the mechanisms that drive these changes in cancer.

Genetics

Revisiting publicly available genomics data can help us build new tools that can navigate repetitive regions within NGS datasets. A thorough dissection of the abundance of sequencing reads across TCGA samples that contain chromosome specific α-satellites would provide powerful corroboration for the PCR findings presented in this thesis. Cross-referencing these findings to known genetic anomalies within each sample has the potential to reveal mechanistic drivers that result in the heterogeneous losses observed in α-satellite copy number. Tools that can survey variation in α-satellites across large datasets like TCGA can be subsequently leveraged to evaluate transcription arising from centromeric loci from RNA-seq datasets. These tools can additionally be repurposed to prospectively and/or cross-sectionally evaluate additional non-neoplastic biologic and disease processes, including development, aging, and inflammation (**Figure 4.2**). Indeed, as we have shown previously, contractions in the centromere of chromosome 21 were associated with trisomy 21 in DNA samples from those afflicted by Down's Syndrome, a widely known disease due to aberrant development[2]. Cancer is widely known to be associated with increased age, and thus evaluating centromeric content in a prospective cohort of individuals as they age in conjunction with the diseases they develop during the aging process, may shed light into whether centromeric contraction observed in malignancy is in reality a signature that arises due to aging. Oxidative stress due to reactive-oxygen species (ROS) tension in an inflammatory setting may also have a propensity to alter centromeric DNA in a manner that facilitates copy number variation when comparing healthy cells to cells from an inflammatory milieu in tissue. The insights drawn from these sorts of inquiries can collectively further our understanding of how the genomics underlying this critical structure for cell division affect outcomes under any number of biological conditions.

From a functional standpoint, advances in gene-editing technologies have augmented our ability to genetically manipulate regions of interest to gain a better understanding of their contributions to biology. CRISPR-Cas9 gene-targeting strategies that employ guide RNAs with sequence complementarity to chromosome specific α-satellites and/or HERV-K111 have the potential to unlock deeper comprehension of the role centromere genetics play in both *in vitro* and *in vivo* settings. Cre-Lox recombination systems activated by tissue specific factors can additionally enhance these insights by offering an orthogonal methodology to validate CRISPR-Cas9 findings while simultaneously determining if tissue lineage affects phenotypes in cells/tissues devoid of chromosome specific α-satellites. Given our findings presented in this thesis, it is conceivable that selective deletion of α-satellites will have oncogenic effects, both *in vitro* and *in vivo*, phenotypes that can be readily assessed through examination of cell proliferation and cell cycle in cell culture and in tissue. To ascertain the mechanisms that underlie the losses in centromere material, one can conduct parallel loss-of-function experiments that selectively target effectors within one of the several DNA-repair pathways, to determine if dysfunctional DNA-repair machinery result in the heterogeneous losses in centromeric DNA in cancer. Finally, the use of human artificial chromosomes (HACs) may clarify the debate surround necessity and sufficiency of the genetic and the epigenetic components of the centromere to cell division, with insights into whether centromeric polyploidy can itself drive malignant phenotypes.

Epigenetics

There are two related but separate components of our epigenetic findings that would benefit from future cell biology and mechanistic assessments: 1) canonical centromere epigenetics and 2) non-canonical centromere epigenetics.

Canonical

The altered gene expression patterns linked to CENPA expression that we uncovered in prostate cancer are unlikely to be restricted to the prostate lineage. A handful of studies have investigated centromeric signatures within cancer datasets that, while not nearly as comprehensive as our findings, present trends that are indeed similar to patterns we present here[16,18]. Thus, a natural extension is to investigate the significance of each of the genes within these centromeric signatures (genes that encode members of the CCAN and KMN complex) to malignant phenotypes and identify potentially synergistic partners through combinatorial loss-of-function experimentation. Identifying the most promising genes to target through such a strategy can involve using biochemical approaches such as immunoprecipitation of CENPA from cancer tissue followed by mass spectrometry (IP-MS) to nominate strong interactors that maybe therapeutically actionable. As an example, in prostate cancer tissue, CENPA expression is tightly associated with expression of HJURP, the chaperone responsible for localizing CENPA to α-satellite DNA. The interaction between CENPA and HJURP that is dependent upon ubiquitination of lysine 124 on CENPA, can potentially be a druggable interface between two molecules important for the fidelity of cell division. Dual loss-of-function studies involving both CENPA and HJURP that demonstrate a synergistic reduction in cell proliferation would serve as an immediate prerequisite to demonstrate the potential utility targeting this essential interaction.

Non-Canonical

With regards to the localization defects of CENPA tied to its overexpression in cancer, additional work is necessary to separate the contributions of ectopic CENPA deposition from those of canonical CENPA function in prostate cancer. Complementation experiments involving genes that are both CENPA bound and differentially expressed with CENPA modulation will partly address the importance of non-canonical CENPA function to prostate cancer pathogenesis. Preliminary attempts at reversing the proliferation defects in CENPA-depleted cells with CDC25C, a cell cycle phosphatase both bound by CENPA and differentially expressed under CENPA-depleted conditions, did not show complementation, as the proliferation defect is likely polygenic. A combinatorial approach to complementation will be required to better distinguish the role ectopic CENPA binding plays in perpetuating malignant phenotypes.

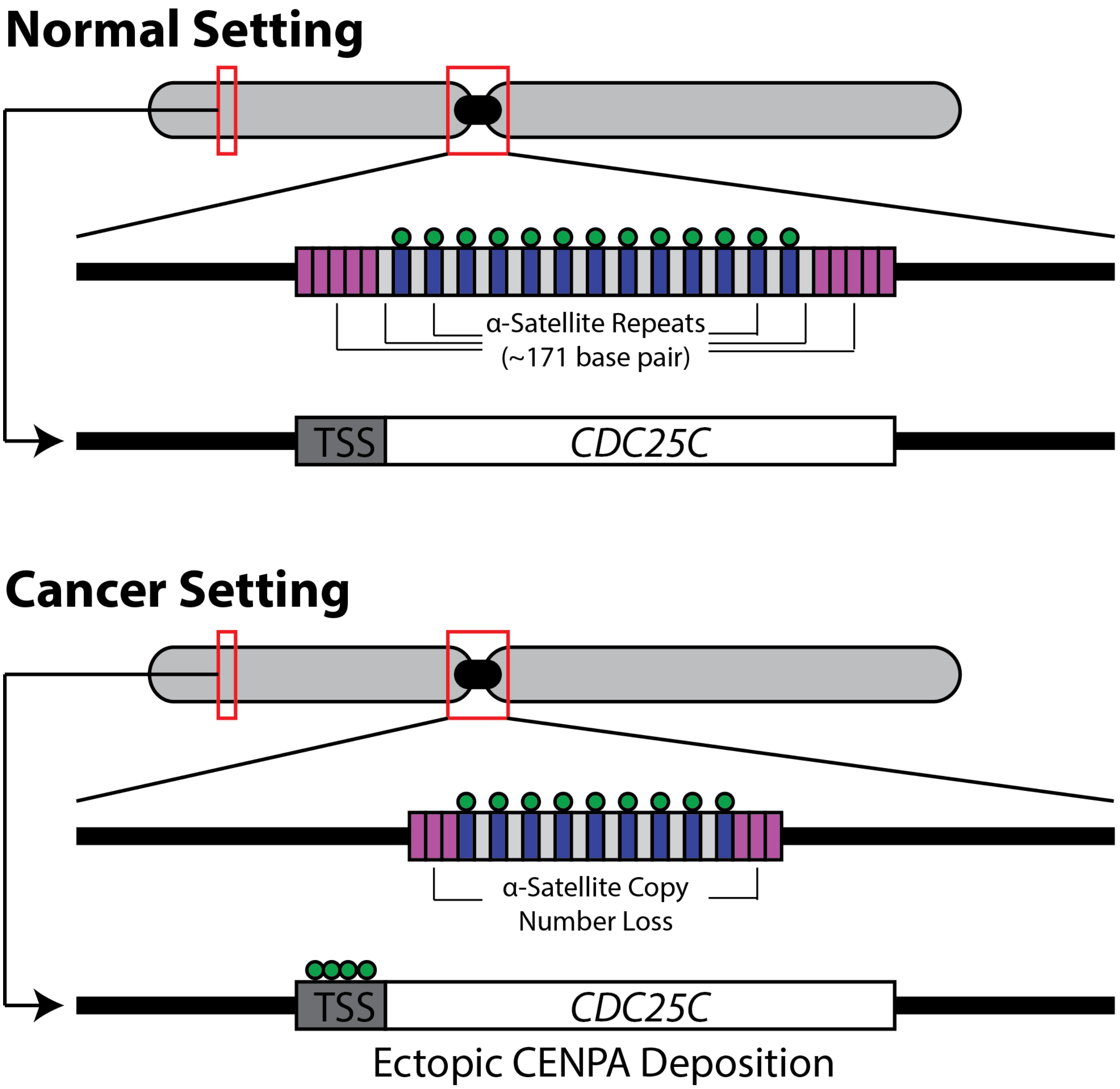
Reproducing the prostate cancer findings presented in this thesis through a similar integrative genomics approach in additional cancer types will provide necessary parallel validation of the importance of gene expression signatures linked to ectopic CENPA deposition. Furthermore, mechanistically interrogating the binding partners that drive this ectopic deposition through similar IP-MS approaches as described above may inform additional therapeutic development. Previous work suggests that distinct binding partners direct canonical and non-canonical CENPA localization, with canonical deposition requiring HJURP and non-canonical deposition requiring DAXX[19]. Carefully choosing cell line models that either express intermediate or high levels of CENPA will aid in determining whether CENPA levels determine binding capacity to these chaperone proteins, potentially through similar use of a doxycycline-inducible system. Finally, a wholistic understanding of centromeric derangements in cancer will require mechanistically linking the genetic findings of heterogenous α-satellite loss with the epigenetic findings involving ectopic CENPA localization. CRISPR-Cas9 mediated or Cre-Lox recombination directed deletion of specific centromeric markers followed by CENPA ChIP-seq may provide insight into whether ectopic localization of CENPA is related to deficiencies in α-satellites available for canonical binding.

Conclusion

The importance of centromeres to cell division underscores our motivation to study these structures in the context of malignancy. Centromeres are vital to coordinating the faithful segregation of chromosomes, providing the foundation for kinetochore assembly and spindle fiber recognition. Eukaryotic biology is thus dependent on this highly conserved process in order to maintain proper growth and development. Aberrant regulation and control of cellular growth and development provide the underpinnings for the development diseases of cell division, i.e. cancer. Highly repetitive sequences termed α-satellites arranged in a head-to-tail fashion across each human centromere however present methodologic challenges to evaluating their genomic structure and function. Moreover, epigenetic characterization of centromeric components in cancer has to date yielded an incomplete understanding of the role they play in driving cancer pathogenesis. We thus sought to comprehensively dissect the genetic and epigenetic components of centromeres in the context of malignancy through methodologies that provide insight into the changes that occur in cancer, while functionally characterizing a subset of these changes. We found anomalies in both the α-satellite and HERV-K111 rich genetic landscape as well as in CENPA expression and localization in the setting of cancer. Our data suggest that the findings reported in this thesis have broad implications to cancer genetics and cancer biology, as the anomalies in α-satellite copy number, sequence homogenization in HERV-K111, and the overexpression/ectopic localization of

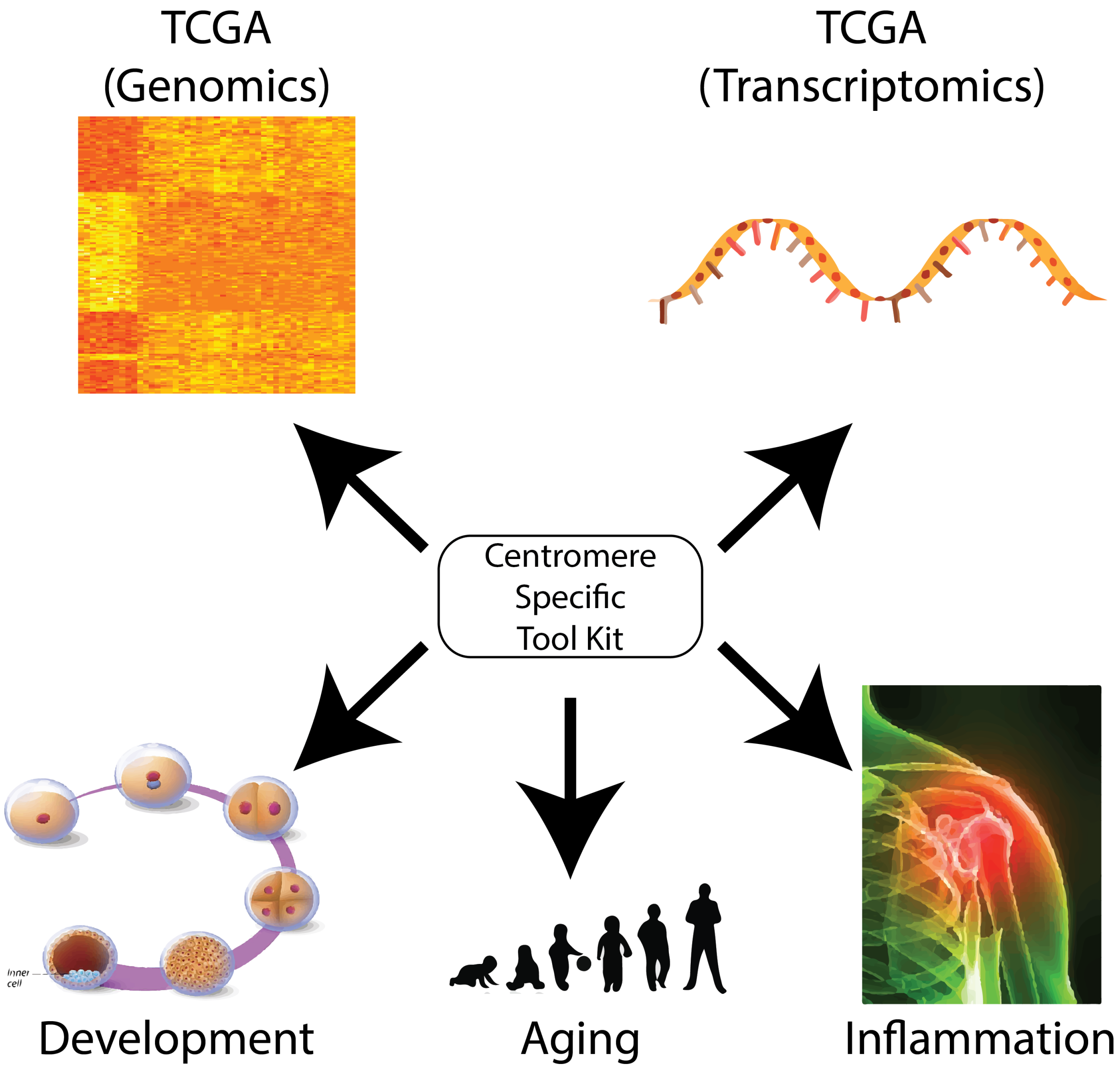
CENPA are present in a number of cancers, described above and in the literature. A deeper understanding of centromere structure and function in cancer will thus give rise to the future development of novel therapeutic strategies that may have clinical utility across several cancer types. Future work that is focused on mechanistically characterizing as well as generalizing the findings presented above will set the stage for high impact discoveries in both science and in medicine.

Figures



Authors: Anjan K. Saha, David M. Markovitz

Figure 4.1: Schematic depiction of centromeric molecular alterations in cancer. Copy number alterations in the form of α-satellite deletions are observed across cancer types in both cell lines and tissue. CENPA, the H3 variant that traditionally occupies α-satellite DNA, ectopically binds gene regulatory elements, such as transcriptional start sites (TSS), of genes important for cell cycle progression, such as CDC25C, when overexpressed in cancer. Future studies are necessary to functionally link the ectopic localization to the α-satellite deletions.



Authors: Anjan K. Saha, David M. Markovitz

Figure 4.2: Schematic outline of future directions with chromosome specific PCR assay. Widescale applications of rapid centromere-specific PCR approaches have the potential to markedly expand our understanding of numerous biologic and disease oriented processes. The collection of α-satellite consensus sequences we have produced through rapid PCR-approaches can be computationally compiled into a centromere-specific reference genome build. Tools that can parse NGS datasets such as TCGA genomic and transcriptomic libraries for these consensus sequences can corroborate our PCR findings in a high throughput fashion. These tools can be further leveraged to study additional non-neoplastic processes such as development, aging, and inflammation.

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Contreras-Galindo, R. *et al.* Rapid molecular assays to study human centromere genomics. *Genome Res.* **27**, 2040–2049 (2017).
3. Zahn, J. *et al.* Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans. *Genome Biol.* **16**, 74 (2015).
4. Vig, B. K., Sternes, K. L. & Paweletz, N. Centromere structure and function in neoplasia. *Cancer Genet. Cytogenet.* **43**, 151–178 (1989).
5. Black, E. M. & Giunta, S. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes (Basel)* **9**, (2018).
6. Bersani, F. *et al.* Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15148–15153 (2015).
7. Natisvili, T. *et al.* Transcriptional Activation of Pericentromeric Satellite Repeats and Disruption of Centromeric Clustering upon Proteasome Inhibition. *PLoS ONE* **11**, e0165873 (2016).
8. Giunta, S. & Funabiki, H. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 1928–1933 (2017).
9. Yi, J.-M. & Kim, H.-S. Expression and phylogenetic analyses of human endogenous retrovirus HC2 belonging to the HERV-T family in human tissues and cancer cells. *J. Hum. Genet.* **52**, 285–296 (2007).
10. Hughes, J. F. & Coffin, J. M. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics* **171**, 1183–1194 (2005).
11. Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**, 407–421 (2003).
12. Rosin, L. F. & Mellone, B. G. Centromeres Drive a Hard Bargain. *Trends Genet.* **33**, 101–117 (2017).
13. Aldrup-Macdonald, M. E. & Sullivan, B. A. The past, present, and future of human centromere genomics. *Genes (Basel)* **5**, 33–50 (2014).
14. Hayden, K. E. Human centromere genomics: now it's personal. *Chromosome Research* **20**, 621–633 (2012).
15. Valdivia, M., Hamdouch, K., Ortiz, M. & Astola, A. CENPA a Genomic Marker for Centromere Activity and Human Diseases. *Current Genomics* **10**, 326–335 (2009).
16. Zhang, W. *et al.* Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nature Communications* **7**, 12619 (2016).
17. McGovern, S. L., Qi, Y., Pusztai, L., Symmans, W. F. & Buchholz, T. A. Centromere protein-A, an essential centromere protein, is a prognostic marker for relapse in estrogen receptor-positive breast cancer. *Breast Cancer Res.* **14**, R72 (2012).
18. Thiru, P. *et al.* Kinetochore genes are coordinately up-regulated in human tumors as part of a FoxM1-related cell division program. *Molecular Biology of the Cell* **25**, 1983–1994 (2014).
19. Lacoste, N. *et al.* Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. *Mol. Cell* **53**, 631–644 (2014).

Appendix

Author Contributions

CHAPTER 1

This chapter was written and conceived by Anjan Saha and edited by David Markovitz.

CHAPTER 2

This chapter is currently under review at *Nature Scientific Reports*. Co-authors: Mohamad Mourad, Mark H. Kaplan, Ilana Chefetz, Sami N. Malek, Ronald Buckanovich, David M. Markovitz, and Rafael Contreras-Galindo. A pre-print is available within *bioRxiv:* https://doi.org/10.1101/505800. For this chapter, Anjan Kumar Saha and Rafael Contreras-Galindo conceived, designed and conducted the experiments, analyzed the results, and wrote the main text; Mohamad Mourad helped conduct experiments; Mark H. Kaplan helped conduct experiments and helped analyze the results; Ilana Chefetz, Sami N. Malek, and Ronald Buckanovich provided patient samples; David M. Markovitz helped analyze the results and revise the text.

CHAPTER 3

This chapter is currently under review at *Proceedings of the National Academy of Sciences*. Co-authors: Rafael Contreras-Galindo, Yashar S. Niknafs, Matthew Iyer, Tingting Qin, Karthik

Padmanabhan, Javed Siddiqui, Monica Palande, Claire Wang, Brian Qian, Elizabeth Ward, Tara Tang, Scott Tomlins, Scott Gitlin, Maureen Sartor, Gil S. Omenn, Arul M. Chinnaiyan, and David M. Markovitz. For this chapter, Anjan Kumar Saha and David Markovitz designed the project and directed the experimental studies. Anjan Saha, Rafael Contreras-Galindo, Monica Palande, Claire Wang, Brian Qian, Elizabeth Ward, and Tara Tang helped conduct *in vitro* experiments. Yashar S. Niknafs, Matthew Iyer, Tingting Qin, Karthik Padmanabhan, Maureen Sartor, and Gil S. Omenn helped with sequencing analysis, and statistical analyses. Javed Siddiqui and Scott Tomlins coordinated biospecimens and performed pathology assessments. Scott Gitlin and Arul M. Chinnaiyan helped interpret results. Anjan Kumar Saha and David Markovitz wrote the text.

CHAPTER 4

This chapter was written and conceived by Anjan Saha and edited by David Markovitz.