Characterization and analysis of repetitive centromeres

Sivakanthan Kasinathan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Steven Henikoff, Chair

Robert K. Bradley

Toshio Tsukiyama

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Characterization and analysis of repetitive centromeres

Sivakanthan Kasinathan

Chair of the Supervisory Committee:
Affiliate Professor Steven Henikoff
Department of Genome Sciences, School of Medicine

Centromeres are specialized regions of eukaryotic chromosomes that ensure faithful transmission of genetic information at each cell division. The molecular architecture of centromeres is defined by evolutionarily dynamic protein and DNA components, which have been proposed to contribute to the origin of new species, while defects in centromeres have been linked to human disease. Centromeres are embedded in regions composed of large arrays of head-to-tail 'satellite' DNA elements, which are not amenable to many conventional genomic analyses.

Here, I describe the development of methods for the analysis of repetitive genomic regions and apply these tools to study primate centromeres, which are composed of ~170-bp α-satellite units. Although centromeric DNA is known to be polymorphic in humans, comprehensive cataloguing of variants at centromeres has not been possible. To gain insight into centromeric genetic variation, I developed a method that uses single-molecule sequencing for analyzing characteristic sequence periodicities called higher-order repeats that arise in human centromeres. The application of this approach to catalogue inter-individual, population-scale, and disease-associated structural variation identified extensive polymorphism in centromeres associated with binding sites for CENP-B, a sequence-specific DNA binding protein. This work also defined a set of functionally important α-satellite dimeric units that are underrepresented in current centromere models and demonstrated aberrations in centromeric sequence in breast cancer. I suggest a role for CENP-B in the evolution and maintenance of higher-order periodicities in centromeric arrays.

Although α-satellite is present at the centromeres of most primates, the precise mechanisms

of evolution of centromeric DNA and the contribution of genetic sequence to the specification of centromere identity remain unresolved. I examined centromere evolution in primates using a combination of data from different whole-genome sequencing methods. This approach demonstrated the presence of higher-order periodicities in all primates and identified an important role for CENP-B in shaping centromeric repeat organization. Further analysis of α-satellite uncovered interspecific variation in the presence of short inverted repeats, which may form hairpin and stem-loop structures. Based on these data, I propose a genetic mechanism for centromere specification that depends on the formation of cruciform or other non-B-form nucleic acid structures.

Taken together, this work enables the cataloguing of variation in satellite DNA, defines important evolutionary transitions in primate centromeres, and advances a model for primate centromere evolution and a theory for centromere specification.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables iv

Chapter 1: Introduction 1
1.1 Satellite DNA and primate centromeres 1
1.2 Aims 2
1.3 Dissertation overview 2

Chapter 2: The centromere 4
2.1 Satellite DNA 4
2.2 Centromere structure and organization 7
2.2.1 Centromeric DNA 8
2.3 Evolution of centromeric DNA 15
2.3.1 Centromeric proteins 18
2.4 Specification of centromere identity 24
2.5 Centromere drive 24

Chapter 3: Genetic variation in human centromeres 26
3.1 Introduction 26
3.2 Results 28
3.2.1 Diversity of α-satellite repeat structures at human centromeres 28
3.2.2 A method for characterization of higher order repeats using single-molecule sequencing 29
3.2.3 Functional alphoid dimers dominate human centromeres 31
3.2.4 A catalogue of inter-individual variation in human centromeres 33
3.2.5 Disease-associated variation in alphoid arrays 35
3.2.6 Variant HORs are enriched for CENP-B boxes 38
3.3 Discussion 38

Chapter 4: Evolutionary dynamics of primate centromeres 43
4.1 Introduction 43
4.2 Results 45
4.2.1 Identification of candidate centromeric satellites in primates 45
4.2.2 HORs are present in all simian primates and in the sifaka lemur 45
4.2.3 Extensive variation in primate α-satellite inferred from short-read sequencing 47
4.2.4 Presence of CENP-B boxes is inversely correlated with hairpin-forming dyad symmetries 50
4.3 Discussion 54

Chapter 5: Conclusions & Perspectives 57

Appendix A: Analysis of genetic variation in human centromeres 83
A.1 Datasets 83
A.2 Identification of SM reads containing repeat monomers 83
A.3 ASTRL: Characterization of array periodicities 84
A.4 ASTRL: Read segmentation 84
A.5 False discovery rate (FDR) estimation and validation of approach 85

Appendix B: Analysis of centromeric satellite in primates 86
B.1 Datasets 86
B.2 Definition of putative centromeric satellites 86
B.3 Centromeric satellite HMMs 88
B.4 Data processing and identification of satellite monomers in Illumina and PacBio datasets 88
B.5 Alignment-free comparison of AS monomers 88
B.6 Detection of dyad symmetries 88
B.7 DNA secondary structure prediction 89

LIST OF FIGURES

Figure Number		Page
2.1	Overview of eukaryotic centromere structure.	9
2.2	An abbreviated phylogeny of relevant primates.	12
2.3	Predominant modes of α-satellite repeat organization in hominoids.	14
2.4	Repeat evolution by unequal crossover.	17
3.1	Visualization of repeat organization in alphoid arrays.	29
3.2	Diversity in alphoid repeat structure at human centromeres.	30
3.3	Robust detection of repeat periodicities in error-prone single-molecule sequencing reads.	32
3.4	Comprehensive characterization of higher-order repeat periodicities in haploid human cell lines.	34
3.5	Higher order repeat structure in human individuals.	36
3.6	Aberrant centromeric repeat structure in a breast cancer cell line.	37
3.7	Centromeric structural variants occur proximal to CENP-B boxes.	39
3.8	A model for sequence organization at human centromeres.	41
4.1	Putative centromeric repeats dominate the tandem repeat spectrum of primates.	46
4.2	Higher order repeats are a general feature of Old and New World Monkey centromeres.	48
4.3	Characterization of the structure of putative centromeric repeats from two prosimians.	49
4.4	Characterization of α-satellite in primates using short-read sequencing.	51
4.5	Differential enrichment of dyad symmetries with the potential to form secondary structure in primate alphoid satellites.	52
4.6	Examples of neocentromeres with dyad symmetries.	53
4.7	A model for centromere specification.	55

LIST OF TABLES

Table Number Page

B.1 Primate short read whole-genome sequencing datasets. 87

ACKNOWLEDGMENTS

> [A]t school, I learned language and history, some geography and sums; but science was a black hole. My eager schoolteacher abandoned science to nature... Language, he used to say, was what made us different from the apes... I [have] learned the reverse: language is what you pick up naturally – everyone speaks, no problem – but science has to be learned methodically, by study, if one is ever to emerge out of the swamp of our psychotic superstitions. It is what transforms our lives.
>
> –Romesh Gunesekara, ***Reef***

This work would not have been possible without Steve Henikoff, who has been a wonderful mentor. The Henikoff lab would only be a theoretical ideal without Terri Bryson, Aaron Hernandez, and Karen Brighton. I am indebted to my lab mates, whose intellectual contributions were critical to helping me learn how science is done. I would especially like to thank Paul Talbert, Srinivas Ramachandran, and Pete Skene. I was also fortunate to work closely with and learn from Patrick Carroll and Brian Freie in the Eisenman lab.

Rob Bradley, Toshi Tsukiyama, Julie Overbaugh, Mark Groudine, and, in particular, Bob Eisenman have been excellent supervisory committee members and have provided sage advice. I would like to thank Rob Bradley and Toshi Tsukiyama who read this dissertation and provided valuable suggestions for improvement. I am also grateful to Dr. Jill Watanabe, my med school advisor, who was always quick to offer encouragement and to remind me to walk my own path. I am, of course, indebted to my friends and classmates. Katharine Willey, Vijay Ramani, and Jon Wang deserve special mention.

Many of the datasets analyzed in this work were generated through the efforts of others around the world; this work would not have been possible without the willingness of these groups to make their data publicly available. This work was was funded by the National Cancer Institute, the Howard Hughes Medical Institute, and the Micki and Bob Flowers Endowment of the Seattle Chapter of the ARCS Foundation. In addition to taking an active interest in and supporting my research, Micki and Bob have been role models as thoughtful, engaged citizens. An early stage

of this work was nurtured by the UW eScience Institute, where I was fortunate to work closely with Bryna Hazelton and Andrew Fiore-Gartland, who taught me a lot about programming and spectral analysis.

None of this would have been possible without my father, mother, and sister, who chose to join me on this journey through medical and graduate schools. And finally, I am indebted to Kelsey Lynch, who has been there from the beginning of my time in graduate school.

Chapter 1

INTRODUCTION

Assembled genomes have singularly underwritten advances in modern biology and promise to fundamentally revolutionize the practice of medicine. Yet almost two decades after the conclusion of foundational assembly efforts, including the Human Genome Project, many eukaryotic genome assemblies remain incomplete (Eichler et al., 2004; Miga, 2015). Gaps in assemblies occur in loci made up of highly repetitive 'satellite' DNAs, which can account for substantial fractions of complex genomes (Britten and Kohne, 1968) and are thought to play important roles in organismal evolution and the pathogenesis of disease (Csink and Henikoff, 1998; Henikoff et al., 2001; Marshall et al., 2008). Although these regions were historically prominent in genome biology, the experimental and computational intractability of satellite DNAs has led to the *de facto* exclusion of these genomic regions from what is arguably the major project of the post-genomic era: linking information encoded in the basepair sequence of the genome − genotype − to the actualization of that information in living cells and organisms − phenotype.

1.1 Satellite DNA and primate centromeres

In many eukaryotes large arrays of tandemly repeated satellites make up centromeres − the *cis*-acting loci that ensure faithful disjunction of chromosomes at each mitotic and meiotic cell division. Primates centromeres are made up of megabases of ~170-bp α-satellite repeats, which undergo rapid, concerted evolution. The conservation of alphoid DNA as the fundamental centromeric unit in simian primates suggests an important role for these genetic elements in centromere specification; however, centromere identity is thought to be determined independently of sequence via inheritance of the histone H3 variant CENP-A through poorly defined mechanisms.

The difficulty of studying satellite DNAs, particularly using genomic approaches, has precluded a clear understanding of intra- and inter-specific genetic variation in centromeres. Although alphoid DNA was first characterized in the early 1970s (Maio, 1971), inter-individual poly-

morphism in in centromeres remains understudied (Miga, 2015). In humans, α-satellite repeats are organized into dynamic, chromosome-specific multimeric units called higher-order repeats which are themselves repeated in head-to-tail fashion to make up homogenous, megabase-scale centromeric domains. Variation in centromeric DNA between individuals has been documented and has recently been shown to impact centromere function (Aldrup-MacDonald et al., 2016); however, it has not been possible to comprehensively characterize the scale of this variation in human populations or in diseases such as cancer, which may be associated with centromere dysfunction.

Relatedly, a clear picture of centromere evolution and repeat organization in primates has remained elusive. For example, while the mechanisms that direct tandem repeat evolution are thought to be universal (Dover, 1982), the architecture of centromeric repeats appears to be highly species-specific as higher-order repeats have not been described in Old World Monkeys or prosimians (Alexandrov et al., 2001). Further, although the foundational kinetochore structure in primates is thought to be the same (Schueler et al., 2010), binding sites for the deeply conserved sequence-specific DNA binding protein CENP-B are absent in the centromeres of many primates (Goldberg et al., 1996).

1.2 Aims

Given the challenges inherent in studying satellite DNA and important open questions surrounding variation in centromeric DNA, I endeavored to develop new approaches for characterizing repetitive centromeres. This work can be broadly categorized into two phases: First, I sought to establish an assembly-independent tool based on single-molecule sequencing for analyzing primate centromeres and apply this method to characterize inter-individual and disease-associated variations in human centromeres. Second, I aimed to use a sequence analysis approach to contribute new insights into mechanisms underlying the evolution and specification of centromeres in primates.

1.3 Dissertation overview

Chapter 2 provides an overview of centromere biology and satellite repeats with a particular emphasis on centromeric DNA in humans and other primates.

Chapter 3 describes a method for assembly-independent characterization of the organization of centromeric DNA. This approach takes advantage of single-molecule sequencing and provides the means to comprehensively catalogue the sequence architecture of repetitive centromeres. The application of this approach revealed extensive inter-individual centromeric structural variation and changes to centromeric DNA in cancer. The methodological details are relegated to **Appendix A**.

Chapter 4 summarizes efforts to understand the evolution of centromeric DNA in the primate lineage. The application of the approach described in **Chapter 3** identified differences in the organization of centromeric sequence in a sampling of primates that corresponded with the presence of the recognition site for CENP-B, the only sequence-specific DNA-binding protein known to localize to centromeres. Analysis of centromeric satellite DNA variation in primates supports a general model for centromere specification based on the formation of cruciform structures in satellite DNA. A detailed of description of the methods is included in **Appendix B**.

Chapter 6 places this work in the broader context of centromere genomics and looks forward, outlining future directions with an emphasis on evolutionary biology and human disease.

Chapter 2

THE CENTROMERE

This chapter provides a DNA sequence-centric overview of centromere biology, beginning with foundational studies of eukaryotic genome organization and concluding with a survey of recent genomic analyses of centromeric repeats.

2.1 Satellite DNA

Eukaryotic genomes can be broadly classified on the basis of sequence composition into unique and repetitive fractions (Britten and Kohne, 1968). The unique component of genomes generally contains protein-coding genes and regulatory regions. The repetitive fraction can, in turn, be further subdivided based on the relative abundance of its constituent elements: Moderately repeated regions are comprised of interspersed genetic elements such as transposons, whereas the highly repeated fraction contains head-to-tail tandem repeats of elements known as satellite DNAs (satDNAs) (Britten and Kohne, 1968; Charlesworth et al., 1994; Jurka et al., 2007).

The meaning of the term 'satellite' in this context has evolved with the innovation and application of technologies for analyzing genomes. Equilibrium ultracentrifugation was the first of these foundational tools (Beridze, 1986). Investigation of the behavior of solutions of fragmented genomic DNA in sedimentation equilibrium density gradients (Meselson et al., 1957) led to the coining of the term 'satellite DNA' (Kit, 1961). Upon centrifugation, DNA fragments migrate to regions of zero net force where their tendency to diffuse is counterbalanced by sedimentation, forming isopycnic bands (*i.e.*, zones of molecules with similar densities). In comparison to DNA from bacteriophages λ and T4, which formed Gaussian bands, DNA from calf thymus DNA produced an apparently skewed unimodal band (Meselson et al., 1957), suggesting compositional heterogeneity. Subsequent studies established that this skewing was due to the presence of discrete subpopulations of peripheral or *satellite* bands and further demonstrated that eukaryotic DNA generally behaves in this manner in cesium chloride (CsCl) density gradients (Beridze, 1986). The use of

agents that increase the buoyant density of DNA (such as heavy metals) facilitated fine-scale analysis of minor fractions of satDNAs (Jones, 1973). Classical satDNAs are therefore defined based on their predilection for forming peripheral bands in CsCl equilibrium gradient ultracentrifugation.

The next major methodological advance was the development of C_0t analysis, in which the reassociation kinetics of denatured DNA is measured at varying temperatures with C_0t indicating the product of initial DNA concentration and incubation time. In these experiments, the presence of high-copy DNA was inferred from the observation of a rapidly reannealing fraction. A pioneering study of mouse DNA identified a fraction with rapid reassociation kinetics, which corresponded to a previously observed density gradient satellite band (Waring and Britten, 1966). C_0t analysis also helped estimate the length of the repeat unit at ~150-300 bp in rodents (Hennig and Walker, 1970). Measurement of DNA reassociation kinetics thereby expanded the conception of satDNAs by providing evidence of their repetitive nature and led to the definition of 'kinetic' satellites (Beridze, 1986). Further, the demonstration of an inverse proportionality between DNA content and reassociation rate in a variety of organisms established the universality of highly repeated DNA in eukaryotic genomes (Britten and Kohne, 1968).

Early molecular methods were critical in pinning down the sequence and localization of satDNAs. For example, complete nuclease digestion of DNA fractions coupled with polyacrylamide gel column analysis (Corneo et al., 1968) and partial enzymatic degradation-based fingerprinting (Southern, 1970) revealed profound compositional variation in satDNA compared to main band DNA consistent with their differing buoyant densities (Jones, 1973). The placement of satDNAs at particular genomic loci followed from three key observations: Schildkraut and Maio (1968) recovered satDNAs from nucleoli of mouse cells, Maio and Schildkraut (1969) then went on to demonstrate that mouse chromosomal DNA soluble in 2M NaCl was enriched for main band DNA while the satDNA remained in the insoluble pellet, and Yasmineh and Yunis (1969) isolated satDNAs directly from mouse heterochromatin. Shortly thereafter, some of the first *in situ* hybridization experiments employed radiolabelled satDNAs to localize these elements to centromere-proximal domains on mouse chromosomes (Pardue and Gall, 1970; Jones, 1970). Hints of the potential functions of satDNAs were beginning to emerge.

The third transition in the definition of satDNAs accompanied the advent of restriction analysis and DNA sequencing. Mowbray and Landy (1974) reported the liberation of specific repeated

fragments from calf thymus DNA upon restriction digestion and showed these ~1 kb fragments to have CsCl gradient buoyant densities similar to classical bovine satellites. Concomitantly, Botchan (1974) used a similar approach to identify a 1.4 kb tandemly repeated bovine satDNA and further uncovered internal repeats within a single satellite unit through analysis of renaturation kinetics. Both of these studies demonstrated the now classic laddering pattern produced by partial restriction digestion of satDNAs, heralding the conception of 'restriction' satellites (Beridze, 1986). Restriction digestion, especially in the context of Southern blotting (Southern, 1975a), continues to remain the gold standard for the characterization of the structure and organization of tandem repeats in eukaryotes. Sequence analysis of restriction-defined satDNAs from the African Green Monkey led Rosenberg et al. (1978) to distinguish between 'simple' and 'complex' satellites, composed of short oligonucleotide segments and long period repeats (170 bp in the African Green Monkey), respectively.

Since the 1980s, the deluge of DNA sequence information and progress in comparative genomics has lead to the usage of the term "satellite" as a catch-all for repetitive genetic elements, with tandemly repeated elements defined based on length of the repeated unit: microsatellites are 2-5 bp in length, minisatellites are ~16-60 bp in length, while satDNAs tend to be >100 bp in length (Charlesworth et al., 1994). Minisatellites and microsatellites, which are also called variable number tandem repeats (VNTRs) and have important clinical (Gatchel and Zoghbi, 2005) and forensic (Chambers et al., 2014) implications due to instability/polymorphism, correspond to simple oligonucleotide repeats, whereas 'satellite' in today's parlance refers to the complex elements described by Rosenberg et al. (1978). This expansive definition of satDNAs as complex, long-period repeats will be used throughout.

The function of satDNAs with regard to their heterochromatic and centromeric localization generated great interest and was the subject of intense speculation as early as 1970 – in part driven by the lack of evidence for transcription and protein coding potential of satDNAs (Walker, 1971). A "News & Views" piece published in *Nature* (1970) provides a glimpse into the thinking at the time:

> Localization of satellite DNA in the centromere regions of mitotic chromosomes, however, is most interesting... Their restricted and highly reiterated base sequences might

> be the molecular basis of chromosome pairing between... A host of experiments and speculations leap to mind. Perhaps satellite DNA plays some part in the assembly of the mitotic spindle, for example, by influencing... the attachment of chromosomes to the spindle... The localization of satellite DNA in centromere regions may suggest explanations of some of its bizarre properties, but many questions remain.

Major tenets of the contemporary theory of satDNAs, especially relating to their centromeric function, were already being considered and debated in 1970; however, a coherent picture of the functional role of satDNAs remained elusive. Evaluating the field at the end of the 1970s, Miklos and John (1979) wrote (emphasis theirs):

> These studies in man, like comparable ones in numerous other organisms, have failed to yield positive results on what should, after all, be their primary objective, namely, the *functional* aspects of human heterochromatin and satellite DNA. Indeed, much of the current research on satellite DNA appears to be directionless from the point of view of function.

Below, I attempt to chronicle the progress made towards understanding the centromeric function of satDNAs.

2.2 Centromere structure and organization

Nearly a century before the characterization of the first satDNAs, centromeres were located by microscopy at the primary constrictions of metaphase chromosomes (Flemming, 1882), where they interact with the spindle apparatus to effect disjunction of chromosomes at each cell division. These cell biological studies were also important in cataloguing the diversity of strategies for centromere organization (Figure 2.1): The simplest centromere is the point centromere of the budding yeast *Saccharomyces cerevisiae*, which interacts with a single spindle microtubule; whereas 'regional' centromeres found in the fission yeast *Schizosaccharomyces pombe* and 'satellite' centromeres in many metazoans are discrete loci composed of kilobases to megabases of repetitive DNA and interact with multiple microtubules (Steiner and Henikoff, 2015; Drinnenberg et al., 2016). In contrast to regional/satellite centromeres (Figure 2.1), holocentromeres, found in a variety of organisms such

as some nematodes and insects, span the length of chromosomes (Steiner and Henikoff, 2015; Drinnenberg et al., 2016). This work is primarily concerned with regional/satellite centromeres.

The molecular architecture of the primary constriction is complex. Electron microscopy defined a deeply conserved trilaminar organization of the kinetochore − the large proteinaceous complex assembled at the centromere to functionally link chromosomes to the spindle apparatus (Cheeseman and Desai, 2008; Cleveland et al., 2003; Santaguida and Musacchio, 2009). The outer regions of the kinetochore contain structural and signal integration components that play critical roles in maintaining genome stability through their interaction with microtubules during cell division, involvement in the resolution of sister chromatid cohesion, and participation in the metaphase and anaphase cell cycle checkpoints (Cleveland et al., 2003; Santaguida and Musacchio, 2009). The inner kinetochore directly interfaces with centromeric chromatin − a highly specialized nucleoprotein complex that can be distinguished from chromatin elsewhere in the genome by the presence of unique DNA and protein components (Figure 2.1).

2.2.1 Centromeric DNA

Although the basepair sequence of centromeric satDNA can be extremely divergent even among members of closely related species, the general pattern of sequence organization is conserved. Excepting point- and holo-centromeres, centromeric DNA is composed of repetitive elements. Functional centromeric sequences, which are competent for kinetochore formation, tend to be AT-rich and embedded within homogenous arrays of head-to-tail tandem repeats. These repeat arrays are flanked by regions that harbor interspersed elements such as transposons, which constitute pericentric heterochromatin. The length of centromeric satDNA units can vary substantially from 5 bp to >1.4 kb; however, they tend to evenly divide or be integer multiples of mononucleosome length (~150 bp), which likely facilitates chromatinization (Melters et al., 2013; Heslop-Harrison and Schwarzacher, 2013).

Fungi

Centromeres in the budding yeast *Saccharomyces cerevisiae* are completely specified by ~120-125 bp DNA sequences (Clarke and Carbon, 1985) with tripartite structure of characteristic centromere

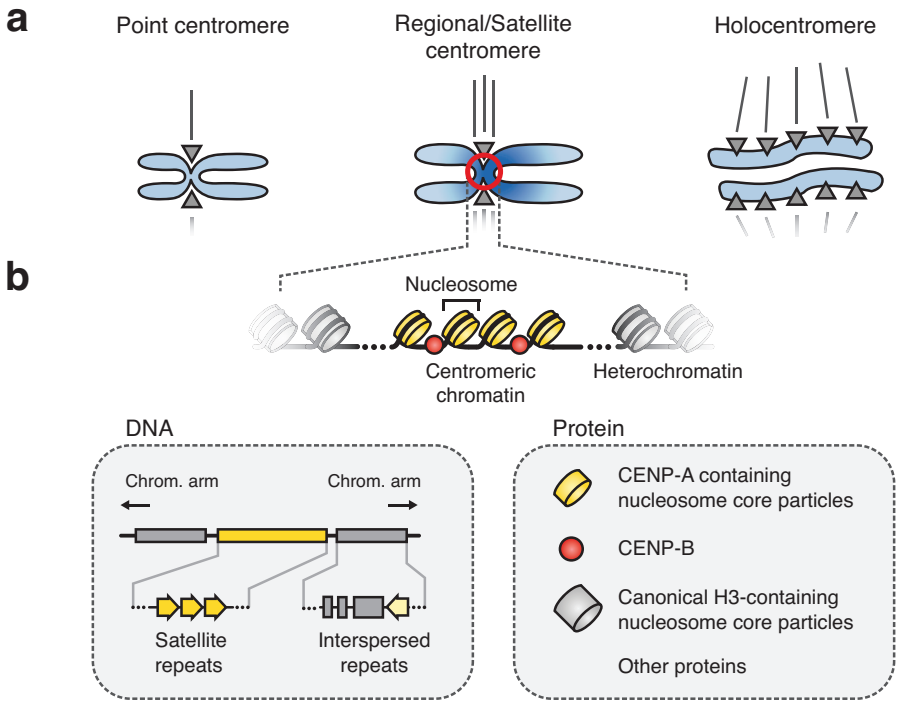

Figure 2.1 | Overview of eukaryotic centromere structure. (**a**) Chromosomes (blue) interact with spindle microtubules (vertical lines) via the kinetochore (grey triangles), which forms at centromeres. There are three ways in which centromeres are organized in different organisms: point centromeres, which interact with a single microtubule, regional/satellite centromeres with multiple attachments, and holocentromeres in which the centromere is distributed over the length of the chromosome. (**b**) Stereotypical chromatin organization at repetitive/satellite centromeres with the functional centromere composed of head-to-tail satellite repeats (gold arrows) and the pericentric heterochromatin made up of interspersed repeats. In addition to specialized DNA, centromeres are marked by the presence of nucleosomes containing the histone H3 variant CENP-A (in gold) and other proteins, *e.g.*, the sequence-specific DNA binding protein CENP-B.

DNA elements (CDEs). CDEI contains an 8-bp recognition site for the basic helix-loop-helix transcription factor Cbf1; CDEII is ~90% AT-rich and wraps a specialized nucleosome; and CDEIII contains a 26-bp motif that is bound by the CBF3 complex. Unlike budding yeast, the fission yeast *Schizosaccharomyces pombe* has large, regional centromeres that are approximately 35-110 kb in size and are characterized by a non-repetitive central core (~4-5 kb) flanked by a set of inner (~6 kb on each arm) and outer (~5 kb on each arm) inverted repeats. Only the central core and the inner repeats are required for centromere function, while the outer repeats constitute pericentric heterochromatin. Inter-repeat homologous recombination resulting in the formation of a covalently closed loop has been proposed to play a role in *S. pombe* centromere function (McFarlane and Humphrey, 2010). Interestingly, fission yeast centromeres also contain tandemly arranged clusters of tRNA genes (Kuhn et al., 1991).

Insects

Centromeric satellite sequences in many insects have not been definitively identified (Palomeque and Lorite, 2008). In *Drosophila*, centromeric satDNAs are thought to be derived from species-specific expansions of short micro- or mini-satellite sequences (Lohe et al., 1993). Although native centromeric sequences remain poorly characterized due to the challenges of analyzing regions dominated by short period repeats (Hoskins et al., 2002; Krassovsky and Henikoff, 2014), Murphy and Karpen (1995), Sun et al. (1997), and Sun et al. (2003) characterized a 420 kb region required for minichromosome transmission and uncovered blocks of AT-rich pentameric minisatellites with intervening transposon insertions consistent with the proposed architecture of *Drosophila* centromeres.

Nematodes

Caenorhabditis elegans has holocentric chromosomes with centromere formation occurring along the length of the chromosome, suggesting that holocentromeres may be assembled agnostic of DNA sequence (Steiner and Henikoff, 2015). This model was supported by analysis of centromeric histone-associated sequences using chromatin immunoprecipitation and microarray hybridization (Gassmann et al., 2012). More recently, a high resolution mapping study identified ~700 discrete

centromeric loci distributed along each chromosome, with each site containing a single centromeric nucleosome (Steiner and Henikoff, 2014). These sites were enriched for GA-rich motifs previously identified as 'high occupancy target' regions on the basis of promiscuous transcription factor binding (Niu et al., 2011). Interestingly, these sites have a chromatin profile similar to budding yeast point centromeres, suggesting that holocentromeres represent dispersed organization of point centromeres (Steiner and Henikoff, 2014).

Plants

The centromeres of *Arabidopsis thaliana* are composed of AT-rich ~180 bp restriction satellite units organized into large (0.4 - 1.4 Mb) arrays (Richards et al., 1991; Copenhaver et al., 1998, 1999). Unlike the centromeres of other eukaryotes, centromeres of *Arabidopsis* and other plants are enriched for particular classes of retrotransposons, which are proposed to play a role in homogenization and centromere evolution (Slotkin, 2010; Neumann et al., 2011; Birchler and Presting, 2012; Gao et al., 2015). Centromeric regions in rice (*Oryza sativa*) have also been characterized and are relatively unusual in that they are composed of both repetitive and non-repetitive sequences reminiscent of *de novo* centromere formation in humans (see below) Nagaki et al. (2004); Yan et al. (2008). Surprisingly, a repetitive rice centromere was shown to contain actively transcribed genes, possibly representing a transition state in the centromerization of a genic region (Nagaki et al., 2004).

Mouse

Centromere-proximal satDNAs in mouse include major and minor satellite units, which form distinct domains (Guenatri et al., 2004; Komissarov et al., 2011). Major satellite units are 234 bp classical satellites, account for 6 Mb of total sequence, and map to pericentric heterochromatin (Hörz and Altenburger, 1981). In contrast, the 120-bp minor satellite units associate with centromeric proteins and constitute a relatively small fraction of the genome (~120 kb of total sequence) (Joseph et al., 1989; Broccoli et al., 1990; Kipling et al., 1991). Both major and minor satellites are AT-rich and are ~80% homologous in certain regions, suggesting that up to two-thirds of the minor satellite sequence was derived from major satellite (Wong and Rattner, 1988). Minor satellite is associated with a 17-nt sequence recognized by the sequence-specific centromere binding protein CENP-B

(reviewed below); however, major satellite does not appear to bind CENP-B (Kipling et al., 1995; Ohzeki et al., 2002). Evolutionary analysis of these satellite sequences supports their recent amplification in the mouse genome (Cazaux et al., 2013).

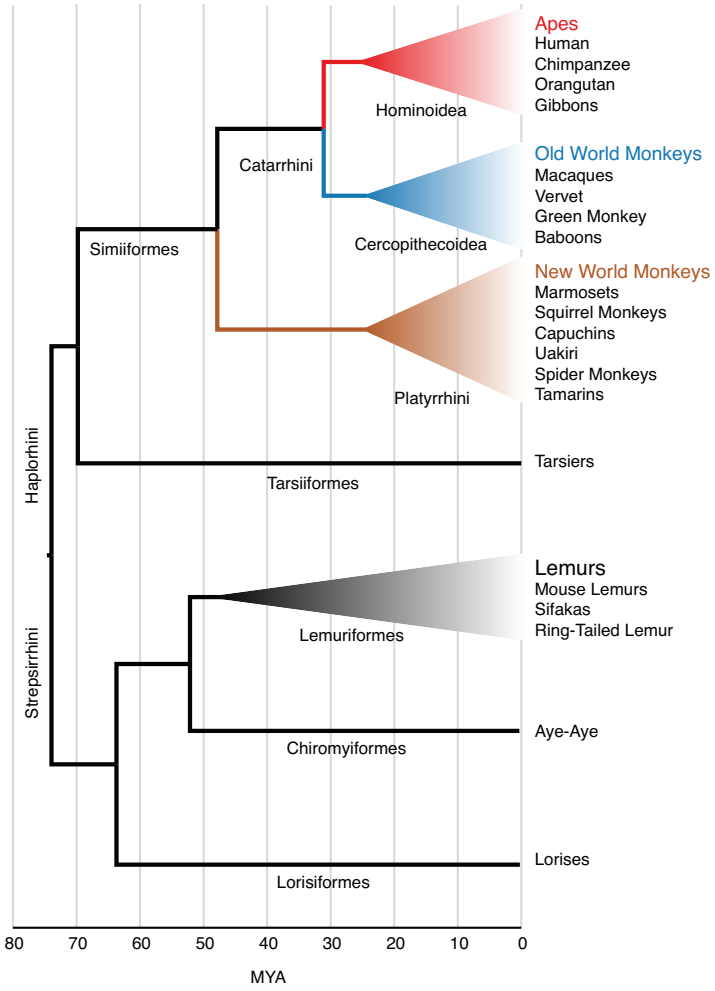

Figure 2.2 | An abbreviated phylogeny of relevant primates. Chronogram with divergence times (in millions of years) estimated from mitochondrial genomes; based on Pozzi et al. (2014) and the 10kTrees Project.

Primates

Primates have regional centromeres composed of megabase-scale arrays of tandem repeats. Excepting species such as lemurs (Lee et al., 2011; Shepelev et al., 2009; Alexandrov et al., 2001), most

primates have centromeres made up of ~170 bp α-satellite (AS) repeats (Musich et al., 1980; Maio et al., 1981; Willard, 1991; Alves et al., 1994; Alexandrov et al., 2001). AS was first identified as a classical/kinetic satellite in *Cercopithecus aethiops*, the African Green Monkey (AGM) (Maio, 1971). Alphoid sequence was subsequently found in humans by restriction digestion and homology to AGM AS and was further localized to primary constrictions by *in situ* hybridization (Manuelidis, 1978b; Manuelidis and Wu, 1978; Manuelidis, 1978a). Similarly, AS was shown to be present in tandemly repeated arrays in a broad sampling of the primate lineage that included Old and New World Monkeys by Maio et al. (1981) by homology to AGM and human alphoid DNAs. Combined with cross-hybridization of AS, sequence analysis of AS repeat units from a diversity of primates suggested that alphoid units arose in an ancestor of primates and underwent subsequent expansion and mutation (Rosenberg et al., 1978; Thayer et al., 1981; Waye and Willard, 1989; Durfy and Willard, 1990; Alkan et al., 2007).

AS sequence, repeat structure, and organization in primates is species- or clade-specific. Almost a quarter of the AGM genome is composed of 172 bp AS units with very low (1-5%) inter- and intra-chromosomal divergence (Musich et al., 1980; Thayer et al., 1981). In contrast, AS from other OWMs such as macaques and baboons accounts for 8-10% of genomic sequence and is more diverse at the monomer level. However, in these species, AS is organized into dimeric units, which are $<$10% divergent (Singer and Donehower, 1979; Donehower et al., 1980; Musich et al., 1980; Pike et al., 1986). Similarly, most NWMs have dimeric units composed of diverged AS monomers; in some NWM genera (*Chiropotes* and *Pithecia*) the major repeat organization is reported to be trimeric (Fanning et al., 1989, 1993; Alves et al., 1994, 1998).

In great apes, AS repeat structure and distribution are considerably more complex. Early studies identified a dimeric AS unit in humans released by EcoRI digestion (Manuelidis and Wu, 1978) and estimated to occur up to 22,000 times in the genome based on reassociation kinetics (Darling et al., 1982), supporting the idea that AS organization might be similar in humans and OWMs. However, it soon became clear that there was extensive divergence of up to ~40% between adjacent monomers and chromosome-specific distributions of AS monomers (Yang et al., 1982; Willard et al., 1983; Jabs et al., 1984a,b; Jørgensen et al., 1986; Willard et al., 1986; Alexandrov et al., 1988; Jørgensen et al., 1988). In addition to AS units that are unique to the centromeres of particular chromosomes, there are also AS arrays that are shared between non-homologous chromosomes

(Baldini et al., 1989; Greig et al., 1993; Jørgensen et al., 1988), suggesting efficient homogenization of repeats between chromosomes.

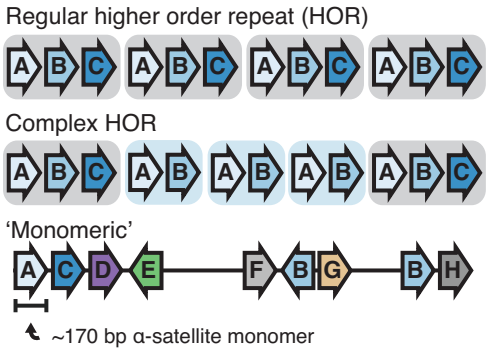

Figure 2.3 | Predominant modes of α-satellite repeat organization in hominoids. Alphoid sequence is apes (genera *Homo*, *Pan*, *Pongo*, and gibbons) is organized in two major types of arrays: Monomeric arrays are composed of highly divergent AS units and are typically located proximal to chromosome arms at the periphery of the centromere. In contrast, higher-order repeats (HORs) are found closer to the centromeric core and are composed of repeats of AS multimers, which tend to be >95% identical.

Chromosome-specific monomeric units were used to define restriction fragment length polymorphisms termed higher-order repeats (HORs). In contrast to monomeric AS units, HORs are multimers of 2-30 AS units in which the constituent monomers are easily distinguished (~70% sequence identity), but adjacent HORs are nearly identical (Willard, 1991). Similar patterns of AS organization were observed in other great apes (Jørgensen et al., 1987; Waye and Willard, 1989; Baldini et al., 1991; Haaf and Willard, 1997, 1998). HORs were thought to be specific to hominoids (Rudd et al., 2006; Alkan et al., 2007; Cellamare et al., 2009; Terada et al., 2013) and absent from the genomes of OWMs and NWMs; however, recent studies have identified HORs at the centromeres of NWMs (Prakhongcheep et al., 2013; Baicharoen et al., 2014; Sujiwattanarat et al., 2015; Suntronpong et al., 2016; Kugou et al., 2016). Interestingly, there do not appear to be reports of HORs in OWMs . In addition to homogenous HORs, which are thought to make up centromeric cores, AS is also present in so-called 'monomeric' arrays, which are diverged AS sequences that are present at the centromeric periphery of all chromosomes (Baldini et al., 1993) and are thought to correspond to ancestral centromeric sequences (Alexandrov et al., 2001; Shepelev et al., 2009). These monomeric arrays are targets of transposon insertion, which is a rare event in homogenous

centromeric cores (Laurent et al., 1999; Schueler et al., 2001, 2005).

Unlike mouse and bovine satDNAs, initial characterization of AS did not reveal any unusual features internal to monomeric units (Alexandrov et al., 2001) with the exception of a concentration of mutations in a restricted region of some AS units (Romanova et al., 1996) corresponding the recognition sequences for the sequence-specific DNA-binding proteins pJα (Gaff et al., 1994) and CENP-B (discussed in more detail below). Binding sites for CENP-B have generally not been detected in OWMs or NWMs, but are present at the centromeres of a host of other mammals (Haaf and Ward, 1995).

That alphoid "sequences appear as old as the primate Order itself" (Maio et al., 1981) coupled with the disruption of endogenous centromere function upon integration of human AS into AGM chromosomes suggested early on that AS "DNA provides the primary sequence information for centromere protein binding and for at least some functional aspect(s) of a mammalian centromere" (Haaf et al., 1992). Although this strong genetic view of centromere identity in primates and other species has gone out vogue (see below), the contribution of DNA sequence to centromere function is indisputable.

2.3 Evolution of centromeric DNA

What is the origin and mechanism of maintenance of arrays of satellite repeats? How are higher-order periodicities generated? Like other satDNAs, centromeric satellites carry the signature of concerted evolution – high intraspecific and low interspecific sequence identity (Brown et al., 1972; Coen et al., 1982; Dover, 1982). AS present in the common ancestor of primates presumably underwent unique, species-specific evolutionary trajectories to generate the extensive variation in sequence and organization of monomers and arrays observed today (Durfy and Willard, 1990; Warburton et al., 1993, 1996; Waye and Willard, 1986b; Alexandrov et al., 2001). Mechanisms of molecular drive, which include gene conversion, unequal crossing-over, and transposition (Coen et al., 1982; Strachan et al., 1982; Dover, 1982; Coen and Dover, 1983; Strachan et al., 1985), are therefore thought to direct the evolution of centromeric satDNAs.

The first model that sought to explain the origins of satDNA was advanced by Britten and Kohne (1968). In this "saltatory replication" model, a DNA sequence is copied many times prior

to chromosomal integration and fixation by natural selection through a succession of relatively low-probability events. Short repeats could be amplified via polymerase slippage during replication, while longer repeats may undergo rolling circle amplification (Beridze, 1986). A second model, which is more widely accepted in the context of centromere evolution (Willard, 1991) but is not mutually exclusive with saltatory amplification, invokes unequal crossover (Smith, 1976). In this model (Figure 2.4), a variant arises in an array of tandemly repeated satDNAs through random mutation, encouraging out-of-register pairing and unequal crossover to yield tandem duplication and deletion products. Expansions and contractions of the variant repeat array arise as the duplicated and deleted products undergo subsequent rounds of unequal exchange and arrays with desired characteristics are selected and maintained. This model explains the divergence of AS arrays on different chromosomes, the homogenization of arrays in *cis*, and the observation of suprachromosomal families of AS, which are evolutionarily related arrays on different chromosomes that have similar higher-order organization (Alexandrov et al., 1988, 2001). Specifically, suprachromosomal families with chromosome-specific distributions are thought to arise through periodic unequal crossover between homologous chromosomes and homogenization through intrachromosomal exchange (Alexandrov et al., 1988; Durfy and Willard, 1989; Alexandrov et al., 1991; Schueler et al., 2001; Schindelhauer and Schwarz, 2002; Schueler et al., 2005).

Given the pattern of AS distribution, unequal exchange likely occurs both inter- and intrachromosomally. Centromeres of homologous chromosomes are notoriously resistant to crossover recombination during meiosis likely to prevent aneuploidy (Choo, 1998; Talbert and Henikoff, 2010b; Nambiar and Smith, 2016). A number of mechanisms for inhibition have been proposed including a role for the compacted state of pericentric heterochromatin (Choo, 1998; Nambiar and Smith, 2016). Recently, Vincenten et al. (2015) described a non-canonical role for kinetochore components in preventing initiation of meiotic recombination by inhibiting both break and crossover formation. However, there are some examples of meiotic events in centromere evolution. Meiotic exchanges have been observed in maize, where gene conversion events (in which segments of homologous chromosomes are copied) in centromeric cores are widespread (Shi et al., 2010; Talbert and Henikoff, 2010b). Schindelhauer and Schwarz (2002) have suggested that highly homogenous human DXZ1 AS array variants lacking *de novo* mutations implicate a gene conversion mechanism. However, this latter observation is not necessarily incompatible with unequal crossover (Schindel-

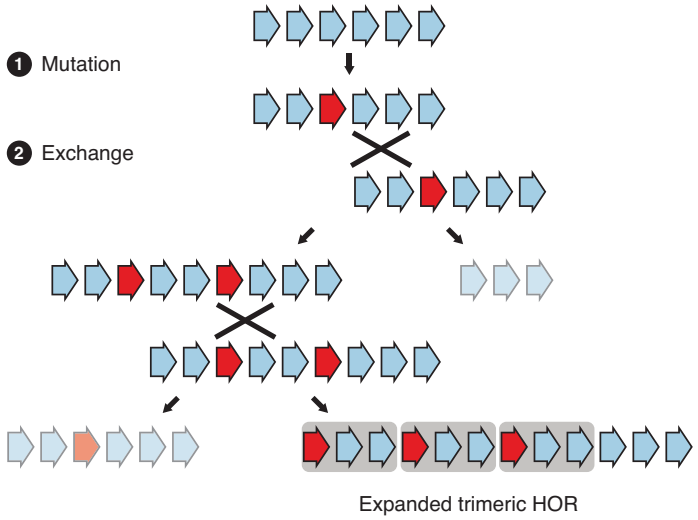

Figure 2.4 | Repeat evolution by unequal crossover. A model for evolution of periodicities in repeated sequences based on Smith (1976). A repeat array acquires a random mutation and undergoes out-of-register pairing and recombination. The products of this exchange contain expansion or deletion of monomeric units relative to the original array. Subsequent rounds of random mutation and unequal exchange can spontaneously give rise to periodicities.

hauer and Schwarz, 2002; Roizès, 2006).

Recombination likely occurs between centromeric arrays of sister chromatids during mitosis. For example, Wang et al. (2008) describe the microscopic characterization of alphoid bridges between sister chromatids that are resolved late in the cell cycle and Jaco et al. (2008) uncovered DNA methylation-regulated mitotic recombination at centromeres in excess of crossover in chromosome arms. Recently, Giunta and Funabiki (2017) observed centromeric exchanges between sister chromatids at a frequency of 5% in cultured human cells and suggested a role for centromeric proteins including CENP-A in maintaining the AS array integrity. Further, consistent with mitotic recombination, population-scale sequence analyses have identified polymorphic signatures suggestive of repeat evolution along haplotypic lineages (Marçais et al., 1991; Warburton and Willard, 1995; Roizès, 2006).

It is important to note that unequal crossover may not be the only mechanism operating on centromeric satDNAs. Computational analyses of the evolutionary relationships of AS monomers have raised the possibility of an unknown mechanism of transposition at play in AS evolution (Alkan et al., 2002, 2004). Using PCR primers spanning L1/Alu-AS junctions, Prades et al. (1996)

detected a surprisingly high number of L1/Alu polymorphisms at human centromeres possibly related to an elevated rearrangement rate associated with the presence of these interspersed repeats. L1 elements, which are known to be associated with monomeric AS (Laurent et al., 1999; Kazakov et al., 2003), have also been proposed to actively participate in centromere evolution by impairing centromere function through insertion and/or limitation of expansion of selfish centromeric DNA (Csink and Henikoff, 1998; Laurent et al., 1999). Interestingly, L1 polymorphisms in AS tend to occur in blocks such that a cluster of L1s is present or absent, possibly due to the insertion or excision of extrachromsomal circles containing L1s and alphoid sequence (Laurent et al., 1999). Consistent with this hypothesis, extrachromosomal circles containing alphoid sequences have been detected in human cells (Cohen et al., 2010).

Based on the preponderance of evidence for repeat evolution through mechanisms of molecular drive, one might expect HORs to be a general feature of centromeric DNA. However, as noted above, in primates, HORs are present in restricted clades (see above) possibly due to more efficient homogenization of AS in other primates through unknown mechanisms (Alexandrov et al., 2001). Outside of the primate lineage, HORs have been described in a number of species (Melters et al., 2013). Notably, HORs have not been described in mouse minor satellite (excepting the Y-chromosome), but are present in major satellite (Pertile et al., 2009; Komissarov et al., 2011). Variation in human HORs and the distribution of HORs in primates is examined in the context of repeat evolution in **Chapters 3** and **4**, respectively.

2.3.1 Centromeric proteins

The characterization of centromeric proteins lagged behind the study of satDNAs until the discovery in 1980 of anti-centromere antibodies in the sera of patients with the rheumatologic constellation of calcinosis, Reynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia known as CREST syndrome (Moroi et al., 1980; Earnshaw, 2015). Although more than a hundred different proteins spanning virtually every known structural and functional class make up the kinetochore and underlying centromere (Earnshaw, 2015), I focus below on three centromeric proteins that are particularly relevant to the work described in later chapters: CENP-A, the centromere-specific histone H3 variant, CENP-B, a sequence-specific DNA-binding protein,

and HJURP, the CENP-A chaperone.

CENP-A

CENP-A was first described as a small non-histone chromosomal protein recognized by autoimmune sera from a panel of scleroderma patients (Guldner et al., 1984) and was purified along with two other proteins designated CENP-B and CENP-C (Earnshaw and Rothfield, 1985). Despite its initial characterization as a non-histone protein, CENP-A was subsequently shown to co-purify with histones and nucleosome core particles (Palmer and Margolis, 1985; Palmer et al., 1987) and was, based on partial sequence analysis of protein purified from spermatozoa, declared a histone H3-like protein. Centromere-specific histones were then identified by homology in a number of species: Stoler et al. (1995) identified Cse4 in *S. cerevisiae*, Takahashi et al. (2000) reported the discovery of the *S. pombe* CENP-A homologue, Buchwitz et al. (1999) characterized a histone H3-like protein in *C. elegans*, (Henikoff et al., 2000) and Blower and Karpen (2001) described Cid in *D. melanogaster*, and Talbert et al. (2002) identified an adaptively evolving histone H3 variant in *A. thaliana*.

CENP-A, like the core histones, contains a characteristic histone fold domain (HFD; three α-helices separated by two loops designated α_{1-3} and $L_{1,2}$, respectively) at the C-terminus and an N-terminal tail region that is subject to post-translational modification (Malik and Henikoff, 2003; Talbert and Henikoff, 2010a). The CENP-A HFD is 50-60% similar to the corresponding domain in H3; however, the N-terminal tails of centromeric H3 variants tend to be highly divergent and, in some lineages including primates, show signatures of positive selection (Malik and Henikoff, 2001, 2003; Talbert et al., 2004; Malik and Henikoff, 2009; Talbert and Henikoff, 2010a; Schueler et al., 2010). The CENP-A targeting domain (CATD), which is comprised of L_1 and α_2 of the HFD, is sufficient for kinetochore formation, mitotic checkpoint signaling, and normal chromosome segregation (Black et al., 2007; Okamoto et al., 2007).

CENP-A nearly universally marks centromeres and is essential for viability and centromere function. Cse4 mutants in *S. cerevisiae* display marked chromosome segregation defects and mitotic arrest at elevated temperatures (Stoler et al., 1995). Similarly, Blower and Karpen (2001) demonstrated defects in kinetochore formation and function in embryos and cultured cells depleted of

functional Cid. In mice, the *Cenpa* gene is essential with disruption of the gene leading to lethality by 6.5 days post-conception (Howman et al., 2000). Interestingly, although CENP-As can be quite divergent, Cse4 from *S. cerevisiae* can functionally complement human CENP-A depleted using RNA interference (Wieland et al., 2004), suggesting that the general features of centromeric chromatin may be deeply conserved. There are a number of exceptions to the essentiality of centromeric histone variants including their dispensability in *C. elegans* meiosis (Monen et al., 2005) and recurrent loss of these variants associated with transitions to holocentricity in insects (Drinnenberg et al., 2014).

In *S. cerevisiae* the structure of the centromeric nucleosome has been definitively established using a number of orthogonal approaches. Centromeric nucleosomes were thought to have octameric composition (containing two copies each of CENP-A and histones H4, H2A, and H2B) similar to canonical nucleosomes until the description of an unusual tetrameric 'hemisome' containing Cid and one copy each of H4, H2A, and H2B in *D. melanogaster* cells (Dalal et al., 2007b). Furuyama and Henikoff (2009) defined positive supercoiling at the budding yeast centromere consistent with right-handed wrapping of DNA of the centromeric nucleosome (in contrast to the left-handed wrapping by canonical nucleosomes). Importantly, this topological state is incompatible with octameric composition of the *S. cerevisiae* centromeric nucleosome and suggested a DNA topology-based model for centromere specification (Furuyama and Henikoff, 2009). The application of base-pair resolution genome-wide nuclease mapping approaches revealed a highly ordered centromere structure with the CDEII region of the yeast centromere occupied by a single ~80 bp particle suggestive of a hemisome (Henikoff et al., 2011; Krassovsky et al., 2012; Skene and Henikoff, 2017), with further support for unique structure of Cse4 nucleosomes provided by atomic force microscopy (Codomo et al., 2014) and *in vitro* reconstitution of hemisomes on centromeric DNA (Furuyama et al., 2013). A chemical cleavage mapping technique, which allows precise genomic mapping of histone H4 positions (Brogaard et al., 2012), definitively established the structure of the budding yeast centromeric nucleosome as a hemisome (Henikoff et al., 2014). In contrast to budding yeast Cse4 nucleosomes, the centromeric nucleosomes in fission yeast were found to be octameric using chemical cleavage mapping (Thakur et al., 2015).

The structure of nucleosomes containing the ~20,000 CENP-A molecules (Bodor et al., 2014) distributed over the centromeres of human chromosomes is highly controversial. Human cen-

tromeres are positively supercoiled (Aze et al., 2016) and CENP-A nucleosomes may have the unusual hemisome configuration observed in budding yeast (Dimitriadis et al., 2010; Dalal et al., 2007a; Quénet and Dalal, 2012; Bui et al., 2012; Henikoff et al., 2015); however, there is also evidence that human CENP-A nucleosomes are octamers (Padeganeh et al., 2013; Hasson et al., 2013; Miell et al., 2013; Nechemia-Arbely et al., 2017). Consistent with classical restriction studies of AS in primates (see above), we recently proposed that the functional genetic unit of human centromeres is an AS dimer that precisely positions two ~100 bp CENP-A particles separated by interposed CENP-B/C (Henikoff et al., 2015) and further showed the presence of CENP-T at these alphoid dimers (Thakur and Henikoff, 2016).

The assembly of CENP-A nucleosomes is discussed in the context of the chaperone HJURP and centromere specification below.

CENP-B

CENP-B is the only characterized centromeric sequence-specific DNA-binding protein in metazoans. It has been independently exapted from *pogo*-like transposases in fission yeast, insects, and mammals (Sullivan and Glass, 1991; Tudor et al., 1992; Kipling and Warburton, 1997; Mateo and González, 2014; Drinnenberg et al., 2016). The *S. pombe* CENP-B homologues Abp1, Cbh1, and Cbh2 appear to play functionally redundant roles in chromosome segregation (Baum and Clarke, 2000) and have been localized to centromeres and pericentric heterochromatin (Lee et al., 1997; Nakagawa et al., 2002). Their involvement in nucleating heterochromatin formation in the outer centromeric repeat regions and in transposon surveillance and silencing are established (Baum and Clarke, 2000; Nakagawa et al., 2002; Cam et al., 2008; Johansen and Cam, 2015). In mammals, the CENP-B protein is highly conserved, displaying 96% amino acid identity between human and mouse (Sullivan and Glass, 1991) and a similar degree of sequence similarity between primates (Schueler et al., 2010).

CENP-B recognizes a 17-nt sequence known as the CENP-B box (Muro et al., 1992) through two N-terminal helix-turn-helix (HTH) DNA-binding domains (Yoda et al., 1992; Kitagawa et al., 1995; Tanaka et al., 2001). Although there are slight species-specific differences in binding site preferences, based on electrophoretic mobility shift assays, the consensus CENP-B box sequence

contains nine positions that are required to support DNA binding: NTTCGNNNNANNCGGGN (Masumoto et al., 1989; Jolma et al., 2013). The protein-DNA co-crystal structure of the CENP-B DNA-binding domain revealed that the HTH domains interact with adjacent major grooves and strongly kink DNA flanking the CENP-B box, creating a 59° overall bend angle (Tanaka et al., 2001).

The C-terminus of CENP-B contains a hydrophobic dimerization domain and CENP-B is thought to exist *in vivo* as a homodimer (Kitagawa et al., 1995). CENP-B also contains a DDE-like endonuclease domain typical of many transposases (Kipling and Warburton, 1997; Nesmelova and Hackett, 2010). Indeed, the detection of recombination breakpoints within human AS at a stereotyped 10-20 nt distance from CENP-B boxes led to the intriguing suggestion that CENP-B may encourage HOR formation by stabilizing out-of-register pairing through its dimerization domain and creating recombinogenic DNA breaks (Warburton et al., 1993; Kipling and Warburton, 1997). However, the CENP-B endonuclease domain contains mutations that are incompatible with metal coordination, which is required for cleavage and strand transfer reactions (Kipling and Warburton, 1997).

Although CENP-B is present in mammals, its centromeric functions are unclear and, particularly in some non-human primates, the localization of CENP-B at centromeres is disputed. In the great apes (genera *Homo*, *Pan*, *Gorilla*, and *Pongo*), CENP-B boxes were detected by *in situ* hybridization at centromeres (Haaf et al., 1995). Centromeric localization of CENP-B in gibbons, Old World Monkeys, New World Monkeys, and prosimians is murky. For example, two different studies detected CENP-B protein in African Green Monkey cells and demonstrated DNA-binding activity comparable to human CENP-B (Yoda et al., 1996; Goldberg et al., 1996). Whereas Goldberg et al. (1996) did not find CENP-B binding at centromeres, Yoda et al. (1996) localized the protein to primary constrictions of chromosomes. CENP-B binding sites were thought to be restricted to the apes and Old World Monkeys (Haaf et al., 1995), but recent studies have identified CENP-B boxes in New World Monkeys such as the common marmoset, squirrel monkey, and tamarin (Suntronpong et al., 2016; Kugou et al., 2016). Prosimians, such as lemurs, are thought to lack CENP-B boxes (Haaf et al., 1995); however, Lee et al. (2011) found diverged and presumably non-functional CENP-B boxes in the Aye-Aye (*Daubentonia madagascariensis*).

Unlike CENP-A, CENP-B is not essential (Hudson et al., 1998; Kapoor et al., 1998; Howman et al., 2000). Yet CENP-B appears to be required for centromere formation in certain contexts. In humans and other great apes, for example, CENP-B is detectable at the centromeres of all chro-

mosomes with the exception of the Y-chromosome (Earnshaw et al., 1987, 1989; Haaf et al., 1995), suggesting that centromeres can form and function without CENP-B. Puzzlingly, CENP-B boxes are required for *de novo* centromere formation on artificial chromosomes (Harrington et al., 1997; Ikeno et al., 1998; Masumoto et al., 1998; Ohzeki et al., 2002; Okada et al., 2007; Okamoto et al., 2007). Moreover, CENP-B is not just associated with functional centromeres as it is also found at the inactive centromere in dicentric chromosomes (Sullivan and Schwartz, 1995). Recently, CENP-B was shown to directly associate with CENP-C and the N-terminal tail of CENP-A (Fachinetti et al., 2015) and, relatedly, Hoffmann et al. (2016) showed that CENP-B is required for preserving CENP-C and kinetochore binding to centromeres upon CENP-A depletion. Therefore, CENP-B functions to enhance the fidelity of chromosome segregation (Fachinetti et al., 2015).

Taken together with the degree of conservation of the molecular architecture of the inner kinetochore (Goldberg et al., 1996; Drinnenberg et al., 2016), these observations raise an interesting paradox: the CENP-B protein is highly conserved among mammals and CENP-B appears to contribute to chromosome segregation, yet the presence of centromeric CENP-B boxes both across and within a species is highly variable. The genomic distribution of CENP-B boxes and the function of CENP-B in primates is explored in **Chapter 4**.

HJURP

Chromatin assembly occurs through the action of histone chaperones in replication-dependent and -independent pathways, which deposit canonical and variant histones, respectively (Henikoff and Ahmad, 2005; Hammond et al., 2017). The deposition of canonical H3-containing nucleosomes occurs coincident with DNA-replication through the CAF1 complex (Smith and Stillman, 1989), while nucleosomes containing the replication-independent variant H3.3 are incorporated into chromatin through the action of HIRA and some CAF1 complex members including Asf1 (Ahmad and Henikoff, 2002; Tagami et al., 2004; Hammond et al., 2017). In contrast to chromatinization elsewhere in the genome, CENP-A deposition occurs during early G_1 upon mitotic exit in mammals (Jansen et al., 2007) and during anaphase in *Drosophila* (Schuh et al., 2007) through specialized mechanisms that remain to be fully defined (Müller and Almouzni, 2014). In humans, amphibians, and fungi, HJURP/Scm3 family members carry out CENP-A assembly (Kato et al., 2007;

Foltz et al., 2009; Bernad et al., 2011; Shivaraju et al., 2011); however, HJURP/Scm3 homologues are not present in nematodes, insects, or fish (Sanchez-Pulido et al., 2009). Chen et al. (2014) recently identified CAL1, which is not predicted to have common ancestry with HJURP/Scm3, as the CENP-A chaperone in *Drosophila*. HJURP was first characterized as a Holliday junction-binding protein in a study of genomic instability in cancer cells (Kato et al., 2007) and subsequently shown to recruit CENP-A to centromeres via the CATD (Foltz et al., 2009). HJURP and Scm3 are structurally similar (Samoshkin et al., 2009; Zhou et al., 2011) and thought to have common ancestry (Sanchez-Pulido et al., 2009). What, if any, role the DNA-binding activity of HJURP has on its chaperone function is unknown.

2.4 Specification of centromere identity

Although the conservation of the genomic architecture of repetitive centromeres suggests an important role for DNA sequence and chromatin in centromere specification and function (Koch, 2000; Lamb and Birchler, 2003), it is challenging to interrogate these genomic regions, particularly using genomic approaches (Treangen and Salzberg, 2011; Aldrup-Macdonald and Sullivan, 2014; Miga, 2015; De Bustos et al., 2016). Combined with the intractability of large repeat arrays, the observation of extensive sequence variation at centromeres and the *de novo* formation of centromeres on non-repetitive sequences has led to an "epigenetic," *i.e.*, largely sequence-independent, model for the determination and propagation of centromere identity centered on centromeric proteins. The mechanisms of centromere specification are poorly understood in most organisms (McKinley and Cheeseman, 2016). Indeed, while CENP-A is commonly regarded the 'epigenetic' mark that specifies centromere identity, exactly *how* CENP-A is deposited at centromeres by its chaperone is unknown (Müller and Almouzni, 2014).

2.5 Centromere drive

Despite the essential and deeply conserved function of centromeres, their constituent proteins and DNA are rapidly evolving (Henikoff et al., 2001; Malik and Henikoff, 2009; Drinnenberg et al., 2016). This apparent 'centromere paradox' is resolved by considering the rapid evolution in light of genetic conflict that plays out during meiosis (Henikoff et al., 2001). In this centromere drive

model, competition between homologous chromosomes arises during female meiosis, which is intrinsically asymmetrical in many species because only one of four meiotic products develops into an oocyte. Centromeric DNA behaves selfishly to increase its odds of transmission into the oocyte through expansion of underlying repeat arrays and enhanced recruitment of kinetochore and microtubule components (*i.e.*, increased centromere 'strength'), resulting in preferential engagement of the egg's spindle apparatus. This skewing of chromosome segregation can be problematic, for example, due to the potential for spreading hitchhiking deleterious mutations (Malik and Bayes, 2006). Parity in chromosome segregation is restored by rapid evolution of centromeric proteins, which act to equalize centromere strength (Henikoff et al., 2001). Centromere drive may provide a molecular explanation for reproductive isolation underlying speciation (Henikoff et al., 2001; Henikoff and Malik, 2002). Consistent with this model, monkeyflower chromosomes harboring a putative centromeric duplication are preferentially transmitted (Fishman and Saunders, 2008). Relatedly, Robertsonian chromosomal fusions, which result in centromere expansion, are preferentially transmitted in humans and, in some cases, in mice (Pardo-Manuel de Villena and Sapienza, 2001; Underkoffler et al., 2005; Chmátal et al., 2014). However, the protein players and mechanisms involved in centromere drive remain to be clearly defined (Rosin and Mellone, 2017).

Chapter 3

GENETIC VARIATION IN HUMAN CENTROMERES

Human centromeres are composed of rapidly evolving megabase-scale arrays of α-satellite repeats and are thought to be highly dynamic genomic loci; however, polymorphism in centromeres remains largely uncharacterized due to the intractability of satellite repeats. Here, I describe a general, assembly-independent framework that uses single molecule sequencing for systematic characterization of structural variation in genomic loci embedded within tandem repeats. Applying this approach to human centromeres, I identified dominant dimeric and pentameric repeat motifs that are underrepresented in current human centromere models. I then characterized the spectrum of structural variation in human centromeres and uncovered extensive polymorphism in array length and composition in human individuals associated with the presence of the binding site for CENP-B, a sequence-specific DNA binding protein. Lastly, I describe a landscape of centromeric variation marked by aberrations in repeat structure and functional alphoid dimers in a breast cancer cell line. This approach should enable characterization of tandem repeat variation and contribute new insights into the population genetics and evolution of repetitive regions such as centromeres.

3.1 Introduction

Repetitive genetic loci composed of satellite DNAs account for a substantial fraction of complex genomes (Britten and Kohne, 1968); however, these regions are represented as gaps in genome assemblies and are typically excluded from genomic analyses due to their biological and computational intractability (Roach et al., 1999; Lander et al., 2001; Venter et al., 2001; Eichler et al., 2004). Considered a source of 'missing heritability' (Collins, 2010), polymorphism is repetitive loci has been associated with a number of diseases, particularly cancer (Zhu et al., 2011; Zhang et al., 2015; Tsurumi and Li, 2012; Ting et al., 2011; Atkin and Brito-Babapulle, 1981). Human centromeres, which serve the essential function of linking chromosomes to spindle microtubules during cell di-

vision, are embedded in tandemly repeated ~171-bp α-satellite (AS) units found in megabase-scale arrays on each chromosome. Centromeres account for ~2% of the genome, but constitute nearly 25% of unassembled sequence in the most recent human genome assembly (Eichler et al., 2004; Miga, 2015).

Centromeric AS arrays, which interact with nucleosomes containing the centromere-specific histone H3 variant CENP-A, have complex organization and are thought to harbor extensive, functionally important genetic variation. Alphoid DNA occurs in two major configurations: chromosome arm-proximal 'monomeric' arrays, which represent diverged relics of ancestral centromeres, and homogenous AS multimers called higher-order repeats (HORs) at centromeric cores, which nucleate kinetochore formation and are marked by nucleosomes containing the centromeric histone H3 variant CENP-A (Willard, 1991; Schueler et al., 2001; Rudd and Willard, 2004; Henikoff et al., 2015). HORs can be either chromosome-specific or shared between characteristic chromosomes and are thought to be the products of intra- and inter-chromosomal recombination occurring during rapid, concerted evolution of centromeres (Durfy and Willard, 1989; Greig et al., 1993; Schindelhauer and Schwarz, 2002; Roizès, 2006). Consistent with this mode of evolution, analyses of restriction and amplification polymorphism have revealed single-nucleotide, copy number, and structural variation in HORs (Wevrick and Willard, 1989; Marçais et al., 1991; Warburton and Willard, 1995) that may impact fidelity of chromosome segregation. For example, Maloney et al. (2012) identified metastable epialleles at variants of a chromosome 17-specific HOR and Aldrup-MacDonald et al. (2016) demonstrated that the differenttial propensities of variants of this HOR for interacting with CENP-A nucleosomes may underlie observed differences in mitotic stability of chromosome 17.

Comprehensive characterization of centromeric variation using next-generation sequencing remains challenging in part due to the absence of centromeres in genome assemblies and issues attendant to alignment in repetitive regions (Miga et al., 2014; Miga, 2015; Treangen and Salzberg, 2011). In contrast to short-read sequencing, single-molecule (SM) sequencing with the Pacific Biosciences or Oxford Nanopore platforms routinely produces reads exceeding 10 kb in length (Goodwin et al., 2016), providing an opportunity for assembly-independent characterization of long-range sequence organization in repetitive loci (Khost et al., 2017). Here, I describe the development of a general method for analyzing structural variation in repetitive genomic regions using

single molecule sequencing. Application of this approach enabled the comprehensive cataloguing of HORs and identified the predominance of functional AS dimers, which form a unique chromatin complex, and pentameric motifs at human centromeres. Comparison of individuals from different populations revealed extensive variation in repeat structure associated with the binding sites for CENP-B, the only known sequence-specific DNA binding protein at human centromeres. Finally, analysis of whole-genome SM sequencing of a breast cancer cell line uncovered profound aberration in the organization of centromeres characterized by reduction in functional AS dimers and enrichment of inversions. These results confirm that repetitive loci such as centromeres are rich sources of genetic variation, establish a method for the characterization of this variation, and suggest a model for HOR evolution in human centromeres dependent on the presence of CENP-B binding sites.

3.2 Results

3.2.1 Diversity of α-satellite repeat structures at human centromeres

I first sought to visualize AS-containing SM reads in order to appreciate the type and level of diversity in repeat organization observed in the raw data. I used a strategy similar to dotplotting, which is commonly used to visualize the structure of complex genetic loci, that relied on the identification of alphoid monomers on a single read followed by pairwise comparison of these monomeric units using sequence alignment (Figure 3.1). Applied to PacBio SM sequencing data from the CHM1 hydatidiform mole cell line (Chaisson et al., 2015), this approach revealed tremendous diversity in alphoid array organization (Figures **??**). Based on a random sample of AS-containing SM reads, dimeric repeats appeared to be quite common (Figure 3.2) consistent with the historical recognition of the abundance of these units (Manuelidis, 1978a,b; Manuelidis and Wu, 1978). In addition, some reads contained fragments of HORs with complex multimeric organization that included seemingly homogenous arrays and reads bearing a multitude of different array types (Figure 3.2). Reads with multiple array types were particularly interesting as the regions of transition from one array type to another (Figure 3.2) may represent recombination breakpoints. Finally, a minority of the reads contained 'monomeric' organization, *i.e.*, did not have have a regular periodic structure, and harbored inversions. Some reads even showed evidence of gradients of divergence and

disruption in repeat organization with opposite ends of reads containing HOR and monomeric structure (Figure 3.2). This initial analysis confirmed that raw SM sequencing reads can be used to analyze interesting human centromeric repeat structures.

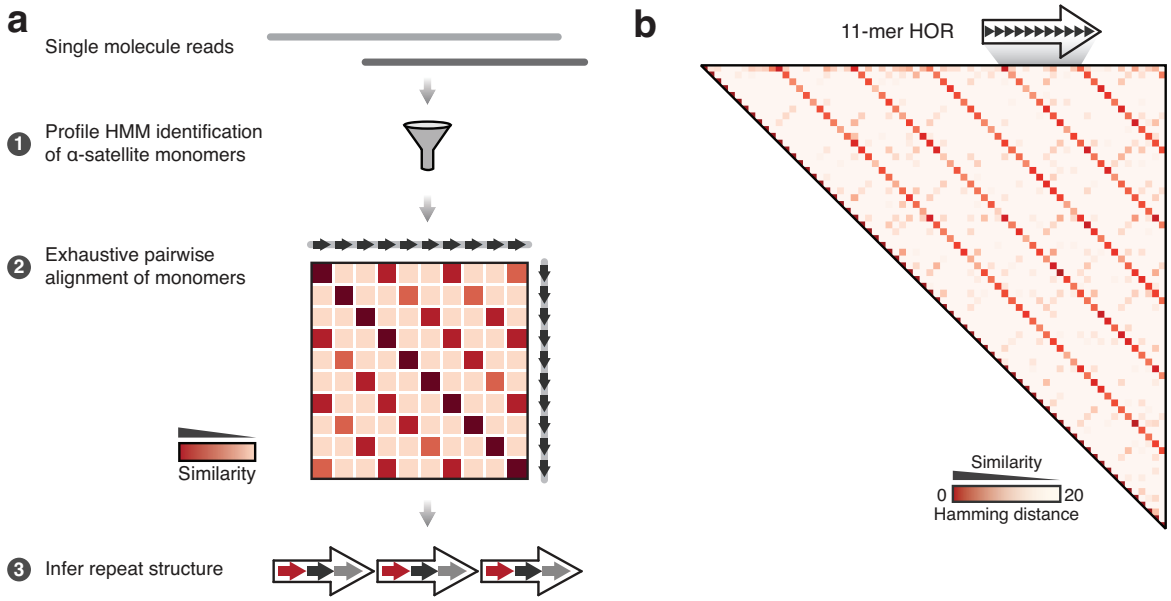

Figure 3.1 | Visualization of repeat organization in alphoid arrays. (a) A method for visualizing similarity of alphoid units akin to self-alignment dotplots. Alphoid monomers are identified on single-molecule sequencing reads using profile Hidden Markov Model homology searching and alphoid units identified on a read are subject to all-by-all pairwise alignment to generate a similarity matrix that supports inference of repeat structure based on the specific pattern of similarity scores along diagonals. Matrices are symmetric about the diagonal and each matrix represents a single sequenced molecule. **(b)** An example of an 11-mer HOR-bearing read scored using the Hamming distance.

3.2.2 A method for characterization of higher order repeats using single-molecule sequencing

Although this approach based on comparisons of predefined AS monomers is suitable for visualizing repeat organization contained in SM reads, it is intrinsically biased by requiring *a priori* identification of alphoid monomers, represents only part of the information contained in reads by focusing solely on the alphoid content of reads, and scales poorly because of the reliance on exhaustive pairwise alignment. Further, the relatively high error rate of raw single molecule reads (Carneiro et al., 2012; Roberts et al., 2013) and the inherent challenge of alignment-based error correction in

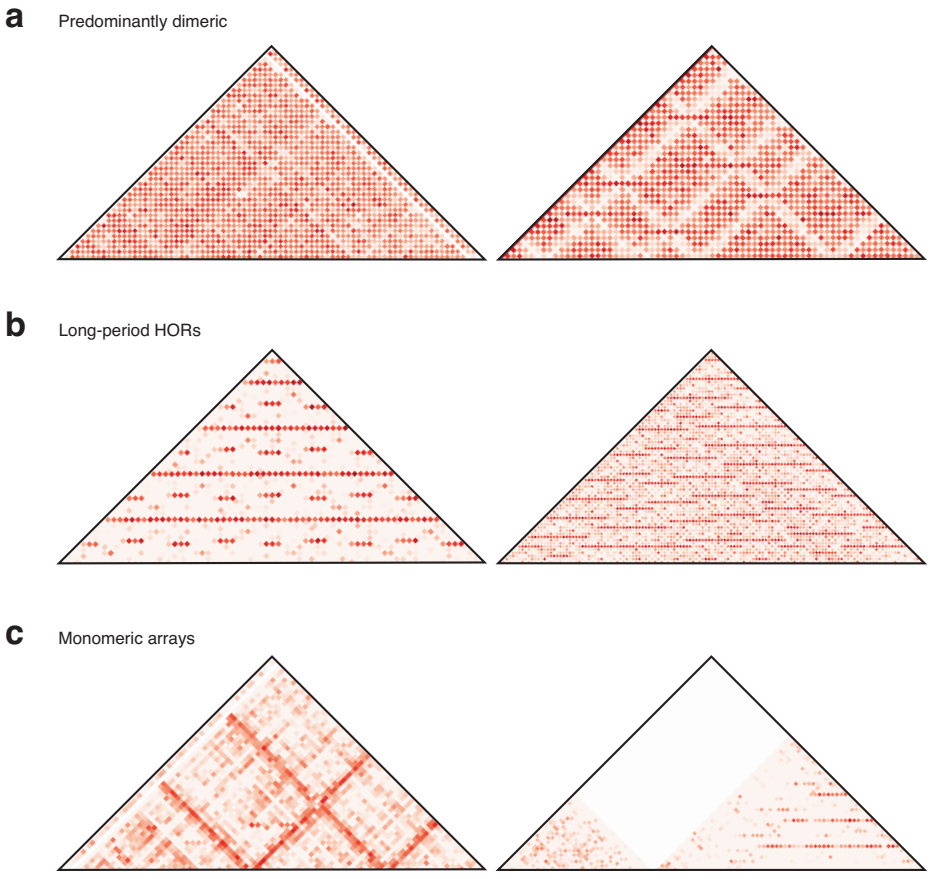

Figure 3.2 | Diversity in alphoid repeat structure at human centromeres. Examples of monomer similarity matrices from single-molecule sequencing reads containing predominant dimeric structure **(a)**, HORs with period >2 alphoid units **(b)**, and 'monomeric' arrays containing highly diverged alphoid units **(c)**. Only one half of the matrix is plotted in each case due to symmetry about the diagonal. Note that the read represented on the right in **(c)** contains an inversion, which manifests as a 'two-peak' appearance because alignments performed to generate matrices are directional (*i.e.*, the reverse complement alignment is not performed for any pair of repeat units).

repetitive regions necessitated the development of a sequencing noise-resilient approach for characterizing periodicities. I therefore sought to develop a more general approach for analyzing the repeat content of SM sequencing reads. This new approach, called ASTRL (Analysis of Satellites and Tandem Repeats in Long reads), is based on enumerating positions of repeating exact k-mers and subsequent analysis of the distributions of k-mer offsets coupled with pairwise alignment-based validation of candidate repeats. Importantly, ASTRL enables characterization of both regular and irregular alphoid and non-alphoid tandem arrays in raw data from the Pacific Biosciences and Oxford Nanopore platforms (Figure 3.3). We validated ASTRL by simulating single-molecule reads from alphoid arrays of known periodicity and confirmed the robustness of the method for a variety of error profiles (Figure 3.3). ASTRL outperformed frequency-domain methods (Suvorova et al., 2014) and a recently described graph-based approach for analyzing tandem repeats in single molecule reads (Sevim et al., 2016), particularly in annotating irregular arrays. Compared with these other approaches, ASTRL does not strictly rely on *a priori* identification of monomeric repeat units, makes no assumptions about the number of periodic segments in a read, and is resilient to indel error. I conclude that ASTRL is a general method for assembly-independent characterization of tandem repeats in raw single-molecule sequencing data.

3.2.3 Functional alphoid dimers dominate human centromeres

HORs have traditionally been defined using restriction digestion coupled with Southern blotting (Southern, 1975b; Willard et al., 1986); however, this method is labor-intensive and does not allow comprehensive cataloguing of HORs at centromeres. Using ASTRL, I carried out *ab initio* characterization of the spectrum of periodicities in SM reads containing alphoid sequence from the CHM1 and CHM13 hydatidiform mole cell lines (Chaisson et al., 2015; Huddleston et al., 2017), which harbor a single haplotype. To facilitate comparisons between datasets, I iteratively sampled read lengths (amounting to 10X coverage of the genome) from a log-normal approximation of the typical Pacific Biosciences read length distribution (Figure 3.4). Approximately two-thirds of reads contained HORs, with the remaining third of reads containing no detectable periodicity (Figure 3.4). I next quantified the relative amounts of sequence contained within HORs of different periodicities (Figure 3.4), which were roughly comparable between the two cell lines. Consistent

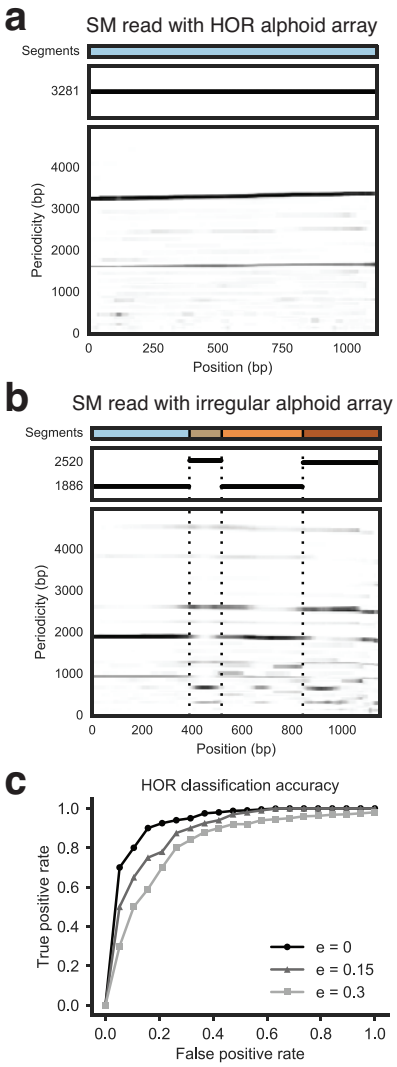

Figure 3.3 | Robust detection of repeat periodicities in error-prone single-molecule sequencing reads. (**a**) Position-periodicity representation of predominant periodicity detected along a single-molecule (SM) read containing a single, homogenous ~20-mer higher-order alphoid repeat array with characteristic periodicity of 3,281 bp. (**b**) An SM read with an irregular alphoid array containing multiple periodicities; dotted lines indicate regions of transition between periodicities that may represent recombination breakpoints. (**c**) Receiver operating characteristic curve for ASTRL-based classification of simulated SM reads from previously characterized HORs under a variety of indel error regimes ranging from no error ($e = 0$) to 30% error ($e = 0.3$).

with previous models for the evolutionary origins of hominoid AS arrays (Alexandrov et al., 1988, 2001), alphoid dimers and pentamers and their integer multiples dominate human centromeres (Figure 3.4).

The most recent release of the human genome assembly (GRCh38/hg38) contains centromere reference models produced using a graph-based approximation dependent on monomer adjacency relationships observed in Sanger sequencing data (Miga et al., 2014; Rosenbloom et al., 2015). In order to evaluate the similarity of this model to the raw sequencing reads, I simulated SM reads from the hg38 centromere reference models with a read length distribution (Figure 3.4) and error profile similar to the CHM1 and CHM13 datasets and analyzed the spectrum of HOR periodicities observed. In contrast to the periodicities observed in the raw data, AS hexamers and decamers were over-represented in the hg38 reference models and larger periodicities were under-represented (Figure 3.4). I also evaluated the potential functional significance of the different HORs by quantifying the number of CENP-B boxes, which are short 17-bp recognition sites for the sequence-specific centromeric protein CENP-B associated with functional centromeres (Muro et al., 1992). Dimeric HORs were four-fold more enriched for CENP-B boxes than other HORs (Figure 3.4), reminiscent of functional Cen1-like AS dimers that associate with CENP-A nucleosomes, which we previously identified (Henikoff et al., 2015). To further characterize these dimeric sequences, I determined whether they were similar to the Cen1-like dimeric AS unit. Approximately 75% of dimeric arrays had high-stringency alignments to the Cen1-like dimer, with a smaller proportion containing matches to other dimeric sequences (Figure 3.4). Taken together, these results suggest that functional dimeric AS arrays capable of binding CENP-B dominate human centromeres and are under-represented in current models of centromeric sequence organization.

3.2.4 A catalogue of inter-individual variation in human centromeres

ASTRL was applied next to delineate the spectrum of HOR periodicity variation in human individuals using available PacBio whole-genome sequencing of lymphoblastoid cell lines (LCLs) (Pendleton et al., 2015; Zook et al., 2016; Shi et al., 2016). I sampled reads from high-coverage datasets employing the same approach used to analyze the CHM1 and CHM13 datasets. The rel-

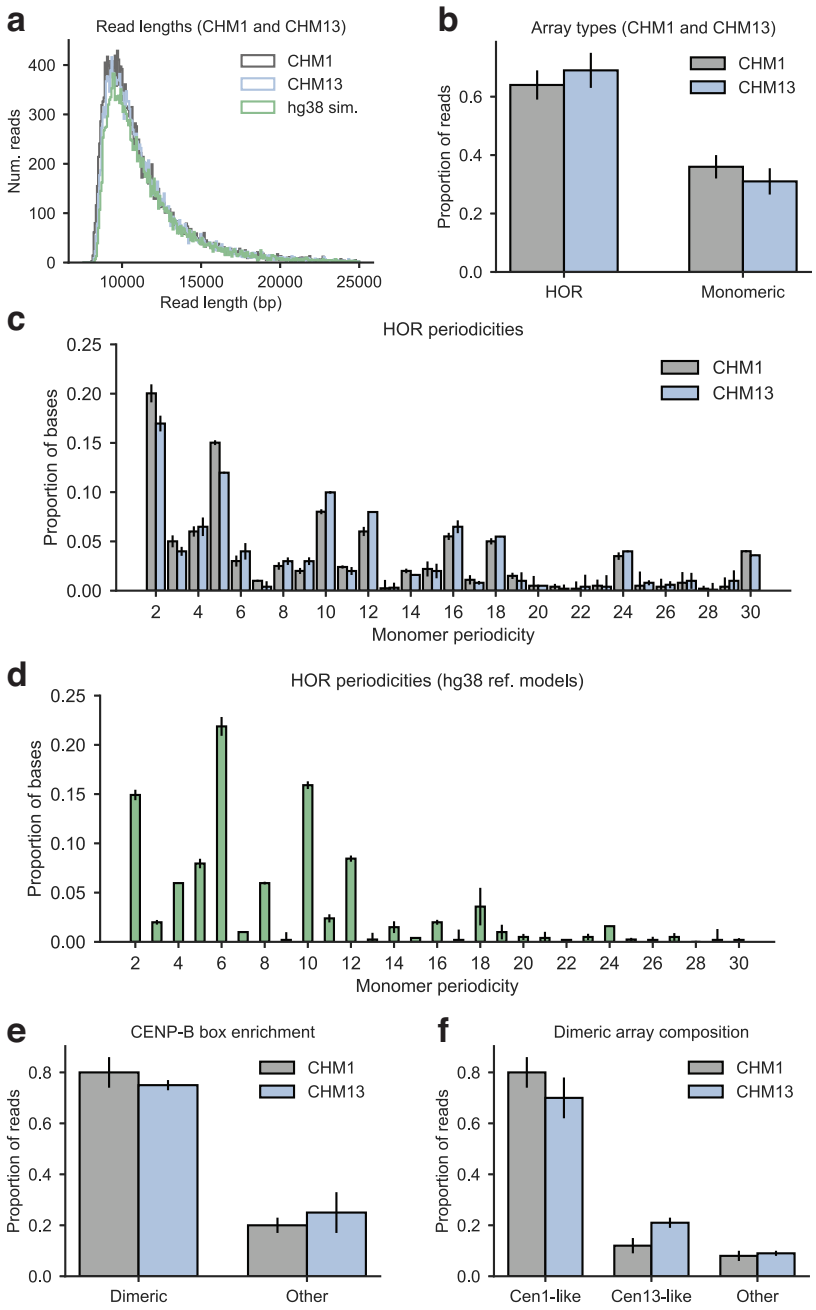

Figure 3.4 | Comprehensive characterization of higher-order repeat periodicities in haploid human cell lines. SM sequencing datasets from the haploid CHM1 and CHM13 cell lines were analyzed by iterative resampling of ~10X genomic coverage partitions with reads selected based on lengths drawn from a log-normal distribution **(a)**. **(b)** Proportion of CHM1 and CHM13 reads containing HORs and monomeric organization. Proportion of bases in CHM1 and CHM13 alphoid reads **(c)** and simulated reads from hg38 centromere reference models **(d)** contained in HORs of differing periodicities. **(e)** Enrichment of CENP-B boxes in dimeric vs. other, non-dimeric HORs. **(f)** Proportion of dimeric arrays belonging to previously defined Cen1- or Cen13-like functional dimeric sequences. Error bars represent mean ± standard deviation for analyses performed on ten different random partitions of the raw sequencing data.

ative fraction of HOR and monomeric reads was stable across different individuals (Figure 3.5). However, the relative proportion of HORs with specific periodicities varied between individuals, with closely related individuals sharing more HOR spectrum similarities (Figure 3.5). Interestingly, the most variation seemed to be in non-dimeric HORs (such as 10-mer HORs in the Yoruba individual relative to the other individuals analyzed, see Figure 3.5). ASTRL can therefore be used to identify gross differences in HOR periodicities between human individuals.

3.2.5 Disease-associated variation in alphoid arrays

Karyotypic changes, especially aneuploidies, are pathologic hallmarks of cancer and centromeres have been proposed to play a role in cancer-associated instability (Marshall et al., 2008). To gain insight into centromeric structural variation in cancer, I applied ASTRL to AS-containing PacBio reads from whole-genome sequencing of SK-BR-3, a $HER2^{+}$ breast cancer cell line with a complex, abnormal karyotype (Nattestad et al., 2016; Rondón-Lagos et al., 2014). Compared to the hydatidiform mole and LCL datasets, SK-BR-3 showed a roughly equal proportion of HOR and monomeric reads (Figure 3.6) and, consistent with genomic instability, a substantially larger fraction of SK-BR-3 reads contained HOR inversions (Figure 3.6). The spectrum of periodicities observed in SK-BR-3 was also substantially different than in the LCLs and was marked by a reduction in the abundance of alphoid dimers (Figure 3.6). I compared these dimeric units to the CENP-A associated functional dimer we previously identified (Henikoff et al., 2015) and detected a substantial depletion of these functional sequences (Figure 3.6), hinting at possible changes in centromere function. Lastly, based on reports of widespread transcription in satellite regions important for normal centromere function (Blower, 2016; Quénet et al., 2016; Quénet and Dalal, 2014; Chan et al., 2012), I compared the abundance of alphoid sequences in SM PacBio transcriptome sequence (IsoSeq) data and detected a reduction in RNA molecules containing AS in another breast cancer cell line (MCF7) relative to the CEPH/Utah individual (Figure 3.6). These analyses hint at extensive aberration in centromeres in cancer and identified alterations to functional centromeric sequences that may play a role in disease-associated genomic instability.

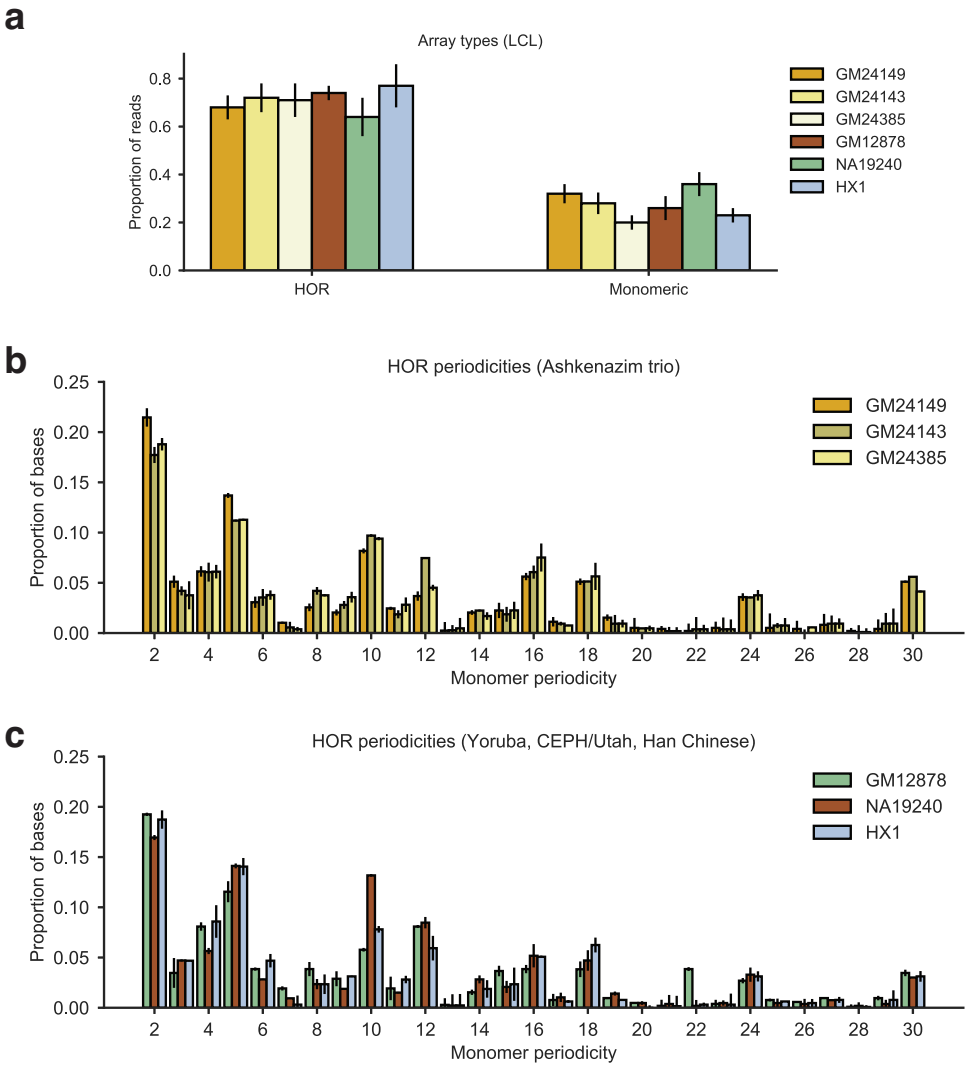

Figure 3.5 | Higher order repeat structure in human individuals. **(a)** Proportion of reads from lymphoblastoid cell lines (LCLs) whole-genome sequencing from an Ashkenazim trio (GM24149, GM24143, GM24385), a CEPH/Utah individual (GM12878), a Yoruba individual (NA19240), and Han Chinese individual (HX1) containing higher-order repeat (HOR) or monomeric repeat structure. Spectrum of observed HOR periodicities in the Ashkenazim trio **(b)** and in human individuals from diverse populations **(c)**.

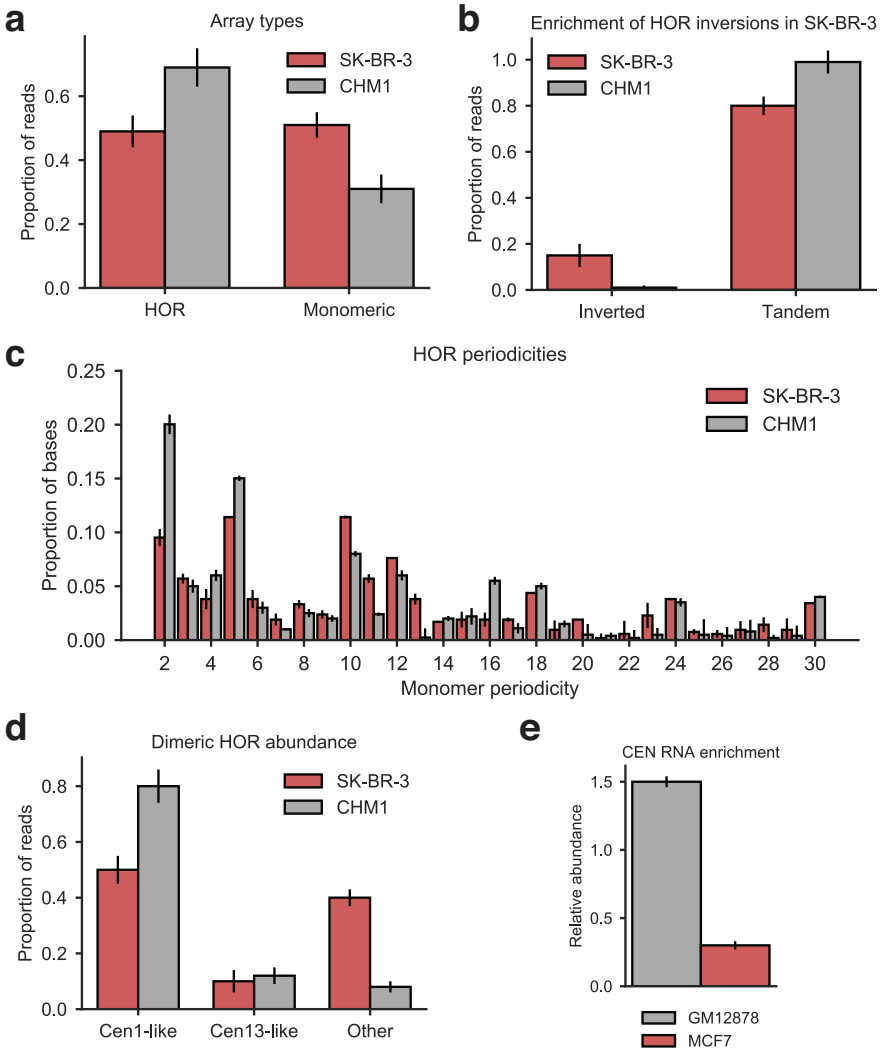

Figure 3.6 | Aberrant centromeric repeat structure in a breast cancer cell line. Proportion of reads containing higher order repeats (HOR) vs. monomers **(a)** and inversions vs. tandem arrays **(b)** from the SK-BR-3 breast cancer or CHM1 hydatidiform mole cell lines. Spectrum of HOR periodicities **(c)** and abundance of CENP-A associated Cen1- and Cen13-like functional dimers **(d)** observed in SK-BR-3 vs. CHM1. **(e)** Enrichment of putative centromeric RNAs containing alphoid monomers in GM12878 (lymphoblastoid cell line) and MCF7 (breast cancer cell line) long-read transcriptome sequencing data.

3.2.6 Variant HORs are enriched for CENP-B boxes

CENP-B, a highly conserved, sequence-specific DNA binding protein that recognizes a 17-nt motif called the CENP-B box, has been proposed to generate HORs through mechanisms that favor unequal exchange (Kipling and Warburton, 1997; Warburton et al., 1993). I therefore examined the distribution of CENP-B boxes relative to array features observed in SM reads. A substantial fraction of reads contained multiple HOR types, suggestive of the presence of recombination breakpoints (Figure 3.2). I compared SM reads containing these 'junctional' HORs across the different human individuals analyzed earlier and defined two array types: 1,204 'variant' arrays, which show differential enrichment in abundance in at least one pairwise comparison of individuals, and 830 'invariant arrays,' which show no evidence of differential enrichment in any pairwise comparison. Variant arrays were substantially more enriched for CENP-B boxes compared to invariant arrays, suggesting that the presence of the CENP-B box may be associated with the generation of variation as proposed earlier (Kipling and Warburton, 1997; Warburton et al., 1993). Further, in support of a potential role for CENP-B in actively generating variation, the variants detected in the setting of genomic instability in SK-BR-3 were not enriched for CENP-B boxes (Figure 3.7). Lastly, analysis of homology in regions upstream and downstream of CENP-B boxes in variant arrays revealed a signature of inter-array recombination (Figure 3.7): high identity upstream of the CENP-B box in the region of homology and low identity downstream (Warburton et al., 1993). These results suggest a role for CENP-B in inter-individual variation in centromeres and, perhaps, in shaping the evolution of HORs.

3.3 Discussion

ASTRL allows the characterization of structural variation in tandemly repeated genomic regions using SM sequencing. ASTRL is applicable to raw/uncorrected data from Pacific Biosciences and Oxford Nanopore platforms and is robust over a range of sequencing errors. When applied to centromeres, this method revealed that functional alphoid dimers, which associate with CENP-A nucleosomes, and pentamers are the major repeat structures at human centromeres. This study also generated, to the best of my knowledge, the first genome-wide catalogues of alphoid repeat structure in human individuals and in cancer. Further, AS variants were found to be enriched

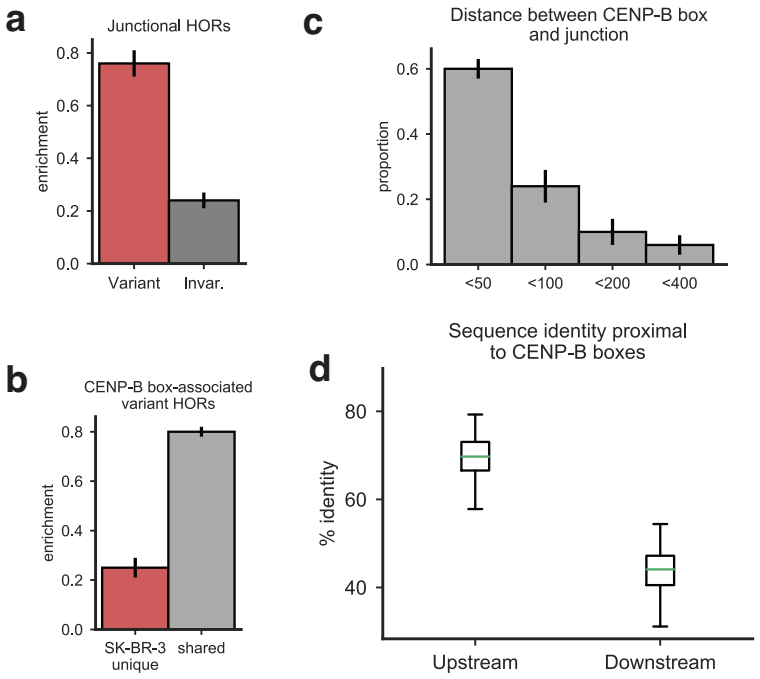

Figure 3.7 | Centromeric structural variants occur proximal to CENP-B boxes. (a) Enrichment of CENP-B boxes in reads containing 'junctional' HORs (defined by the presence of multiple HORs on the same read) in 'invariant' arrays, which show no variation in abundance in human individuals, and 'variant' arrays that show a differential abundance in a pairwise comparison of indivduals. **(b)** Enrichment of CENP-B box-associated variant HORs (defined based on differential abundances across the datasets examined and based on presence of the CENP-B box) in the SK-BR-3 cancer cell line vs. shared between LCLs. **(c)** Distribution of observed distances between a junction between two HORs and the most proximal CENP-B box. **(d)** Sequence identity in 1-kb windows up- and down-stream of oriented CENP-B boxes.

proximal to CENP-B boxes consistent with the importance of this protein in directing evolution of HORs.

Investigating structural variation in satellite sequences using conventional next generation sequencing is challenging for a variety of reasons. Notably, repetitive regions tend to fall in assembly gaps (Roach et al., 1999; Henikoff, 2002; Eichler et al., 2004) and short read alignment profiles are difficult to interpret in tandemly repeated regions (Treangen and Salzberg, 2011). In contrast to previous short-read sequencing-based approaches (Miga et al., 2014), ASTRL leverages long-read SM sequencing and assembly-independent analysis to enable detailed characterization of repeat structure. However, although ASTRL is robust to sequencing error, it can currently only be used to detect large (monomer scale) variation due to the high intrinsic error rate of SM sequencing. Variation at a smaller scale (especially single-nucleotide polymorphism) is known to occur in alphoid arrays and has been proposed to play an important role in the evolution of human centromeres (Alexandrov et al., 2001; Marçais et al., 1991; Roizès, 2006). Investigating this type of variation with SM sequencing data necessarily requires assembly-independent error correction. Hybrid error correction, which involves the alignment of high-quality short reads to SM read scaffolds (Koren et al., 2012), remains challenging for the same reasons that short, repetitive reads are difficult to map to genome assemblies. However, because adjacent HORs are known to be highly identical (Willard, 1991), ASTRL may provide the opportunity to exploit the repetitive nature of alphoid reads to correct SM sequencing errors through multiple alignment-based consensus calling approaches (Chin et al., 2013).

Previous studies detected HORs as restriction fragment length polymorphisms in population scale analyses, typically examining a few HORs (Southern, 1975b; Willard et al., 1989). Importantly, ASTRL enabled comprehensive identification structural variation in AS arrays (limited only by sequencing depth) and highlighted evolution of HORs along haplotypic lineages (Waye and Willard, 1986a; Wevrick and Willard, 1989; Warburton and Willard, 1995). Consistent with concerted AS evolution by unequal crossover, there was more variation between human individuals from different populations than individuals from the same population. Further, in addition to detecting polymorphism at known AS arrays such as the D17Z1 sequence recently shown to harbor functional structural variants (Aldrup-MacDonald et al., 2016), ASTRL identified hundreds of new variants, some of which may be functional on the basis of differential enrichment of CENP-A ChIP-seq sig-

nal and Cen1-like dimeric sequence abundance. In addition to analyzing polymorphism in human individuals, ASTRL identified changes in the gross structure of centromeric repeats in the SK-BR-3 cancer cell line that were consistent with extensive genomic instability revealed by spectral karyotyping (Rondón-Lagos et al., 2014). Interestingly, functional Cen1-like dimeric sequences were depleted in SK-BR-3, suggesting possible changes in centromere function that may be a result of chromosome breakage events (Schvartzman et al., 2010; Bunting and Nussenzweig, 2013), may be adaptive or, alternatively, that could drive formation of neocentromeres (Marshall et al., 2008).

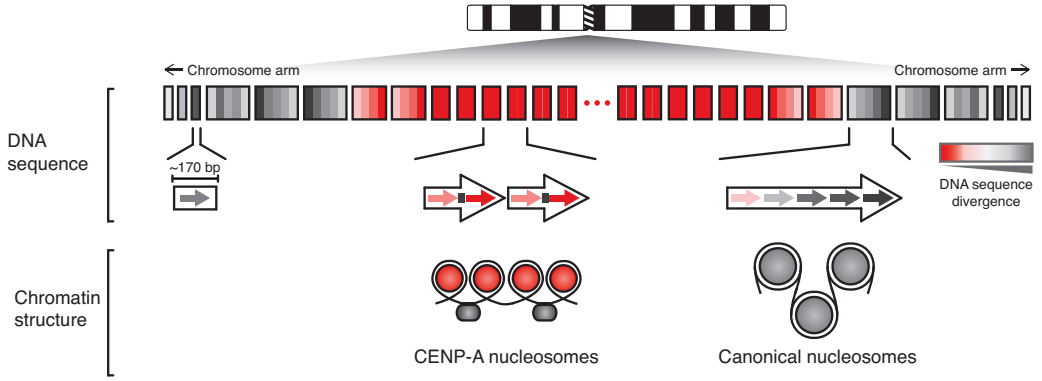

Figure 3.8 | A model for sequence organization at human centromeres. Chromosome arm-proximal centromeric regions are composed of diverged alphoid monomers that give way to increasingly homogenous HORs towards the centromeric core. This pattern of organization is largely supported by the edges of assembled chromosome arms. The centromeric core is composed of abundant alphoid dimers, which interact with CENP-A nucleosomes and, in some cases CENP-B, while peripheral alphoid units wrap canonical H3-containing nucleosomes.

This study also revealed a difference in centromere reference models produced from Sanger sequencing data, which are now part of the human reference genome assembly (Rosenbloom et al., 2015), and raw sequencing data. I speculate that these differences could be due to assumptions in the underlying second-order Markov model used to generate the reference models (Miga et al., 2014), *e.g.*, that the observed monomer adjacencies are reflective of the *in vivo* state of centromeres. This compositional variation could arise from recombination as tandem repeat-containing sequences are known to be unstable in bacteria (Thapana et al., 2014; Treangen and Salzberg, 2011), which were used to amplify genomic DNA for shotgun sequencing (Venter et al., 2001). In contrast, the PacBio datasets were produced without passing DNA through bacterial or yeast hosts and were further not subjected to PCR amplification (Huddleston et al., 2017; Zook

et al., 2016; Chaisson et al., 2015). Consistent with our previous identification of functional dimeric units with unique chromatin structure (Henikoff et al., 2015), ASTRL detected a high proportion of dimeric sequences in all of the datasets analyzed. These Cen1-like dimers, which were particularly enriched in the raw data, contain CENP-B boxes, which appeared to be associated with junctions between different HORs. Previous studies have proposed that CENP-B promotes recombination, potentially by homodimerizing and stabilizing out-of-register pairing (Warburton et al., 1993; Kipling and Warburton, 1997). Based on these observations, I propose a model for the human centromere (Figure 3.8) that modifies the current view (Schueler et al., 2001). In this model, the centromeric core is composed of dimeric sequences such as the Cen1- or Cen13-like functional units, which interact with CENP-A. The edges of the centromere harbor HORs, which are produced by random mutation accumulation in dimers and CENP-B-facilitated unequal exchange. As CENP-B boxes are lost and more mutations accumulate, repeat units cease to undergo exchange and homogenization and become targets of transposon insertion at the chromosome arm-proximal periphery of centromeres.

As long read sequencing becomes more commonplace, I expect techniques such as ASTRL, complemented with base-pair resolution experimental methods, will allow the delineation of the structure of human centromeres and will contribute new insights into mechanisms by which variation in repetitive regions shapes organismal evolution and disease risk.

Chapter 4

EVOLUTIONARY DYNAMICS OF PRIMATE CENTROMERES

Centromeric chromatin is nearly universally marked by the presence of the histone H3 variant CENP-A and rapidly evolving repetitive genetic elements. Despite the high degree of conservation of kinetochore architecture, extensive turnover of centromeric DNA and the apparent absence of sequence specificity in *de novo* centromerization events has led to a model wherein centromeres are defined epigenetically; however, the precise mechanism by which centromere identity is specified remains obscure. Here DNA sequence determinants at centromeres are reconsidered from an evolutionary perspective in a dense sampling of the primate lineage. I identified putative centromeric satellites in prosimians and confirm the abundance of α-satellite units in the centromeres of simian primates. I also find that higher-order repeats, initially thought to be found only in the centromeres of hominoids, are a general feature of primate centromeres and that their abundance correlates with presence of binding sites for the sequence-specific DNA-binding protein CENP-B. Further, CENP-B box abundance is inversely correlated with palindromes in alphoid monomers and the tendency for the formation of DNA stem loops and hairpins. I speculate that centromere identity is specified by a deeply conserved genetic mechanism dependent on the recognition of cruciate DNA structures by CENP-A histone chaperones.

4.1 Introduction

Eukaryotic centromeres are nearly universally marked by the presence of the histone H3 variant CENP-A, which is thought to be sufficient for DNA sequence-independent, 'epigenetic' centromere specification. However, DNA sequences found at endogenous centromeres are required to support centromerization and faithful transmission of artificial chromosomes (Ohzeki et al., 2002; Ebersole et al., 2000; Henning et al., 1999; Masumoto et al., 1998; Ikeno et al., 1998; Harrington et al., 1997) and the nature of centromeric DNA tends to be conserved, suggesting an important genetic contribution to centromere identity (Koch, 2000).

Over nearly 60 million years of evolution, the centromeres of simian primates have remained embedded within megabase-scale arrays of tandemly repeated ~170 bp α-satellite (AS) elements, which undergo rapid, concerted evolution along species-specific trajectories through unequal crossover (Willard, 1991; Alexandrov et al., 2001). While this fundamental AS repeat unit is shared among simian primates, the long-range architecture of repeats is variable (Alexandrov et al., 2001). In apes, the centromeric periphery is characterized by 'monomeric' arrays of highly diverged alphoid units and the functional centromeric core is composed of highly homogenous multimeric AS arrays (Schueler and Sullivan, 2006). These multimers, called higher order repeats (HORs), are specific to chromosomes and are thought to be the product of inter- and intra-chromosomal exchange (Willard, 1991). Curiously, while New World Monkeys (NWMs) were recently shown to have HORs consistent with this model, Old World Monkeys have a repeat structure marked by homogenous monomers or dimers (Alexandrov et al., 2001). Another difference in simian AS arrays is the presence of the 17-nt CENP-B box, which is recognized by the sequence-specific DNA binding protein CENP-B (Muro et al., 1992). Domesticated from *pogo*-like transposases, CENP-B is deeply conserved in mammals at the amino acid level (~96% sequence identity between human and mouse) (Kipling and Warburton, 1997; Mateo and González, 2014). CENP-B protein is present in primates and, in apes, CENP-B boxes are detectable by hybridization at the primary constrictions of all chromosomes excepting the Y chromosome (Haaf et al., 1995). In humans, CENP-B interacts with other centromeric proteins to help nucleate kinetochore formation and enhance the fidelity of chromosome segregation (Hoffmann et al., 2016; Fachinetti et al., 2015, 2013). Paradoxically, while CENP-B and its binding site are present at some NWM centromeres (Kugou et al., 2016), OWMs do not have centromeric CENP-B boxes or protein localization (Goldberg et al., 1996).

In order to better define evolutionary transitions in primate centromeric DNAs, I densely sampled species from the primate lineage and compared their centromeric satellites using data from whole-genome sequencing. This analysis confirmed and extended previously characterized variation to include the identification of putative centromeric repeats in lemurs. Further, we found presence of the CENP-B box to be limited to hominoids, excepting gibbons, and some New World Monkeys, suggesting a potential role for CENP-B in satellite replacement and lending further support to a role for CENP-B in repeat evolution (**Chapter 3**). Analysis of global sequence features revealed an anti-correlation between presence of the CENP-B box and dyad symmetries in satel-

lite monomers. Computational predictions suggested that regions of dyad symmetry could form stem-loop structures. Based on these data, I propose a genetic model for the specification of centromere identity based on the formation of cruciform structures recognized by the CENP-A chaperone HJURP.

4.2 Results

4.2.1 Identification of candidate centromeric satellites in primates

Historically, ~170-bp alphoid satellites are thought to be present at the centromeres of all primates; however, recent work has identified non-alphoid repeat expansions in the centromeres of gibbons and highly divergent satellites in New World Monkeys (Baicharoen et al., 2012; Hara et al., 2012; Carbone et al., 2014; Sujiwattanarat et al., 2015; Kugou et al., 2016); sequences at prosimian centromeres remain to be clearly defined (Lee et al., 2011). Therefore, I sought to identify dominant, candidate centromeric repeats in each species using an unbiased approach. I carried out the computational equivalent of density gradient ultracentrifugation or restriction analyses classically used to define satellite DNAs by searching for tandem repeats in ~1-kb long Sanger sequencing reads (similar to the approach used by Melters et al. (2013)) and contigs from primate sequencing projects. This analysis recovered the expected ĩ70 bp AS sequences in the simian primates (Figure 4.1); however, detection of a strong signal corresponding to AS was dependent on sequencing depth. Consistent with previous reports (Prakhongcheep et al., 2013; Sujiwattanarat et al., 2015), the major AT-rich complex satellite in NWMs is a 340-bp dimeric sequence (Figure 4.1) containing one monomer that is highly divergent from AS. SVA elements, which constitute novel class derived from interspersed repeats at gibbon centromeres (Hara et al., 2012), were not recovered as a highly reiterated satellite. Candidate satellites could not be identified in the genomes of two prosimians (*Lemur catta* and *Otolemur garnettii*) due to low read depth; however, a ~50-bp GA-rich repeat with no detectable homology to AS was discovered in *Microcebus murinus* (Figure 4.1).

4.2.2 HORs are present in all simian primates and in the sifaka lemur

In order to determine whether HORs are present in OWMs and NWMs as predicted by AS evolution by unequal crossover, contigs from genome assemblies were queried for the presence of

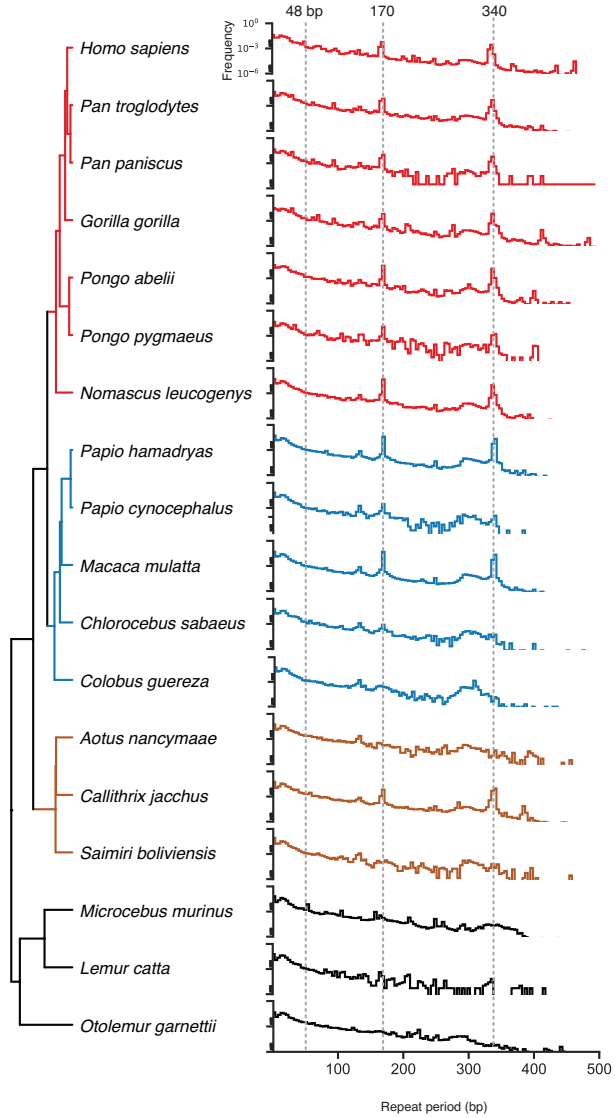

Figure 4.1 | Putative centromeric repeats dominate the tandem repeat spectrum of primates. Tandem repeat period histograms for a sampling of primates based on raw Sanger sequencing data deposited in the NCBI Trace Archive. Dashed lines indicate the location of peaks detected in simian primates (~170 and ~340 bp) and a prosimian, *Microcebus murinus* (~48 bp). Note that y-axes are all on the same logarithmic scale.

candidate repeats identified above. Because HORs are, by definition, composed of relatively divergent monomers, these sequences should be amenable to assembly. The construction of self-alignment dotplots for a random selection of contigs revealed that higher-order structure is present in both OWMs and NWMs (Figure 4.2). I used PacBio data available for some OWMs and lemurs to confirm the tandem repeat spectra with dominant ~170-bp repeats (Figure 4.3). Interestingly, the sifaka lemur (*Propithecus coquereli*) was found to have a 170-bp satellite, although this repeat unit was not homologous to α-satellite. ASTRL, the method for repeat characterization described in **Chapter 3**, was used to estimate the abundance HOR periodicities in one great ape, three OWMs, and one lemur (*Propithecus coquereli*) with available single-molecule sequencing data. While HOR periodicities in gorilla are similar to the spectrum observed in humans (**Chapter 3**), OWMs have centromeres were highly enriched for dimeric units with higher periodicities accounting for $< 5\%$ of total alphoid sequence (Figure 4.3). This likely explains why HORs have gone undetected in these species. Based on these analyses, I conclude that HORs are present in all simian primates, but are substantially enriched in hominoids.

4.2.3 Extensive variation in primate α-satellite inferred from short-read sequencing

Previous studies of AS sequence diversity in primates relied on analysis of restriction fragments, limited whole-genome Sanger sequencing data, or end-sequenced BAC or fosmid clones (Willard et al., 1986; Alkan et al., 2007; Terada et al., 2013; Sujiwattanarat et al., 2015). In order to broadly sample alphoid sequences from representative hominoids, NWMs, and OWMs, we took advantage of publicly available short-read, ~100$\times$100-bp paired-end sequencing datasets. Candidate centromeric sequences were detected by Hidden Markov Model-based homology searching using AS models deposited in DFAM (Hubley et al., 2016); therefore, 'alphoid' as used here refers to sequences that are homologous to the classical α-satellite unit defined in hominoids and OWMs. I first used principal components analysis of 5-mer vectors to verify that 100-bp reads, which are shorter than the AS monomer length (~170 bp), contained sufficient information to distinguish the three clades and species within each clade (Figure 4.4). The short read datasets analyzed here recapitulate previous phylogenetic studies of AS sequence variation in primates (Schueler et al., 2005; Alkan et al., 2007). Estimation of alphoid sequence abundance revealed extensive variation

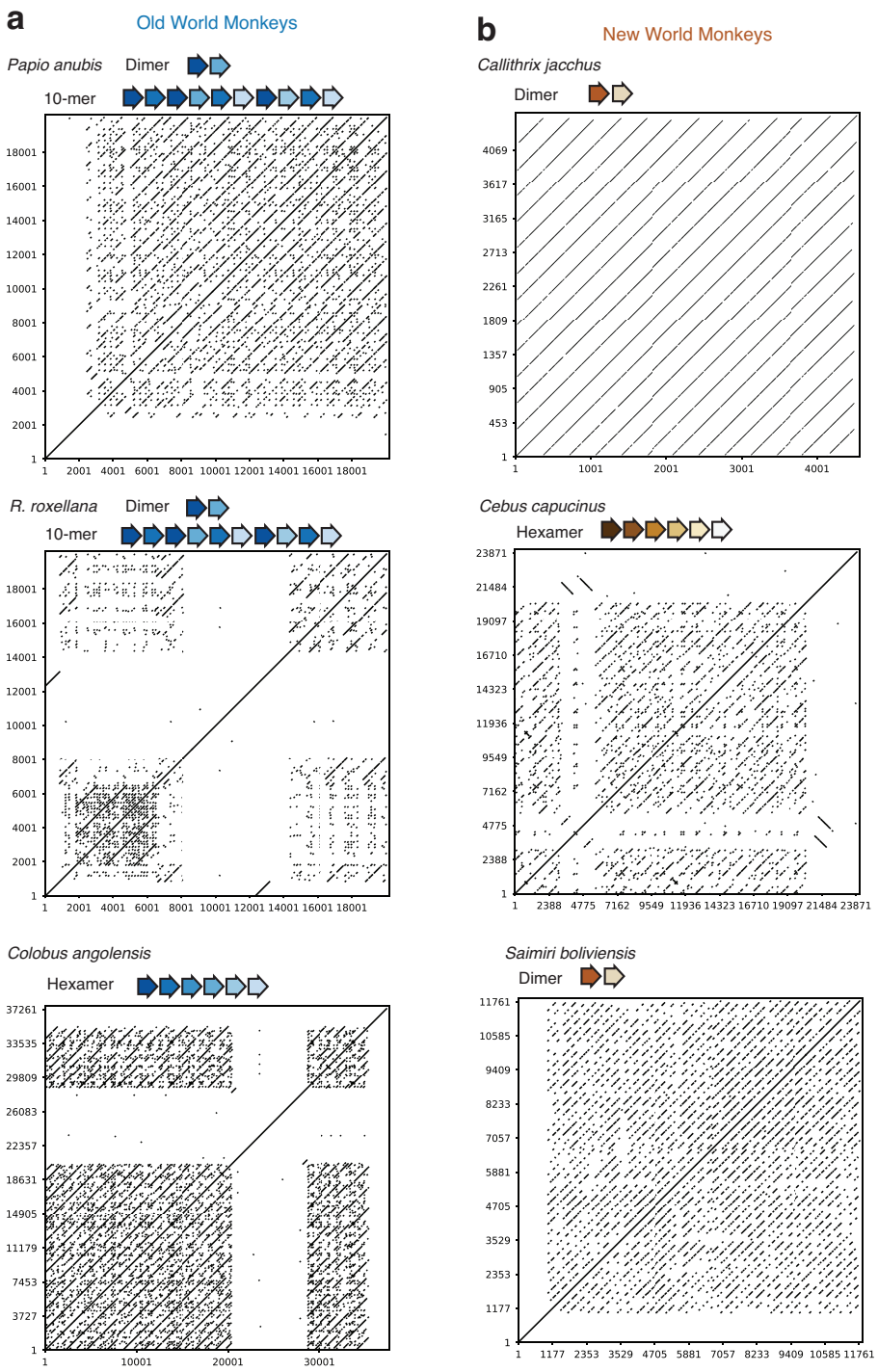

Figure 4.2 | Higher order repeats are a general feature of Old and New World Monkey centromeres. Conventional self-alignment dotplots for α-satellite-containing contigs from genome assembly projects for the indicated Old World Monkey **(a)** and New World Monkey **(b)** species. Repeat structure inferred from distances between dominant diagonals is schematized with arrows. Note that dotplots are not scaled to reflect the different sizes of the analyzed contigs.

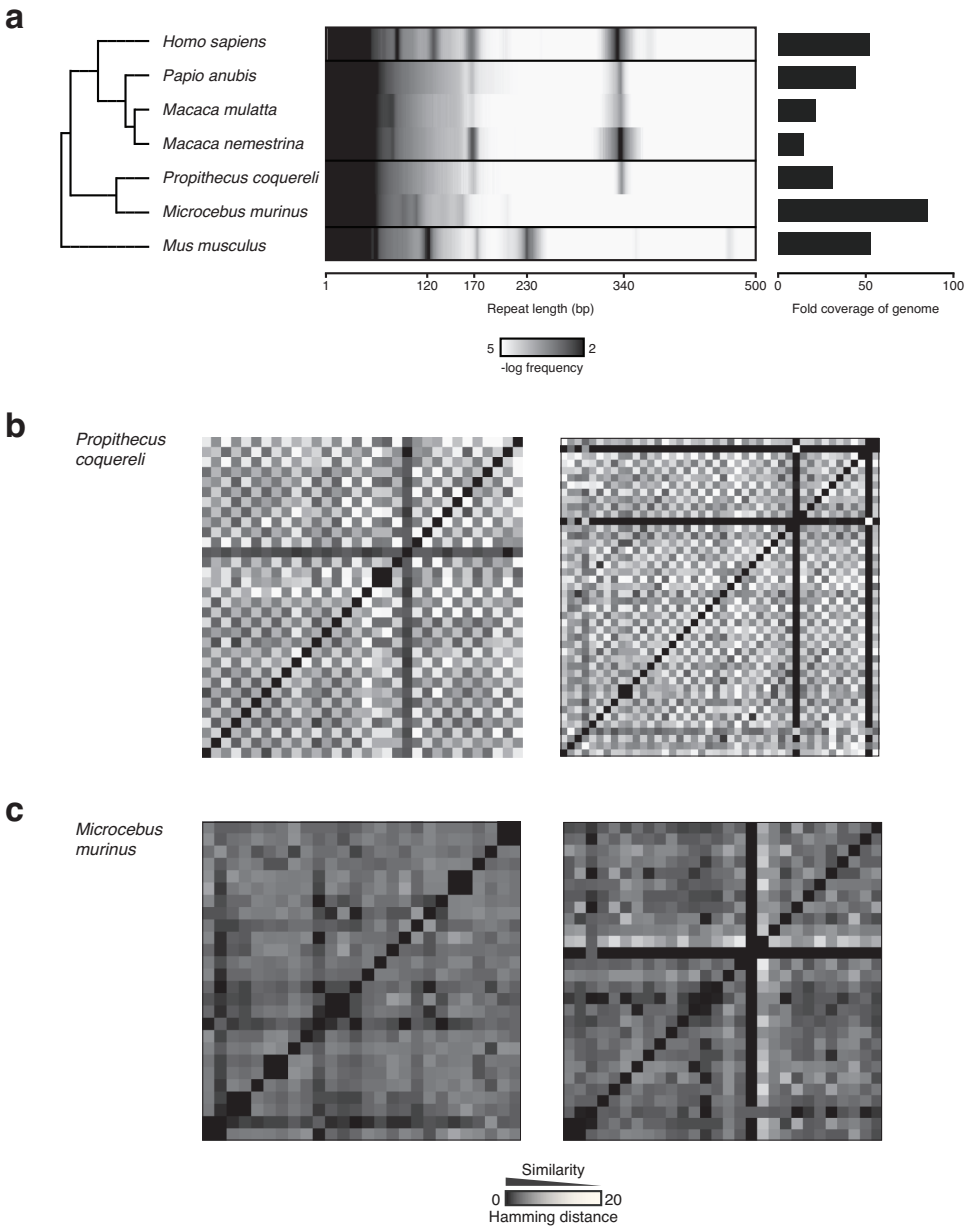

Figure 4.3 | Characterization of the structure of putative centromeric repeats from two prosimians. (**a**) Tandem repeat abundance histograms based on PacBio single-molecule sequencing datasets for human (CHM1 hydatidiform mole cell line), Old World Monkeys (*Papio anubis*, *Macaca mulatta*, and *Macaca nemestrina*), lemurs (*Propithecus coquereli* and *Microcebus murinus*), and mouse (*Mus musculus*). Example similarity matrices (see **Chapter 3**) for HMM-defined ~170-bp putative centromeric repeat units from *Propithecus coquereli* **(b)** and ~48-bp putative centromeric repeat unit from *Microcebus murinus* **(c)**.

between and within clades (Figure 4.4). In general, in hominoids and OWMs, alphoid sequence accounted for approximately 2-10% of the genome; whereas, only one NWM (*A. nancymaae*) contained an appreciable AS fraction (Figure 4.4). The relative paucity of satellite homologous to hominoid/OWM AS is consistent with reports of highly diverged or derivative AS at the centromeres of NWMs (Prakhongcheep et al., 2013; Sujiwattanarat et al., 2015). Within hominoids, the most variability in AS abundance was in gibbons (Figure 4.4), which have undergone rapid karyotype evolution accompanied by the innovation of new composite transposon-derived centromeric repeats (Hara et al., 2012; Carbone et al., 2014). We next queried datasets in their entirety and alphoid reads specifically for the 17-nt core CENP-B motif required to support DNA-binding (`NTTCGTNNNANNCGGGN`) and found a highly restricted distribution of these sequences (Figure 4.4). The hominid genera (*Homo, Pan, Gorilla*) contained the most matches to the core CENP-B box at both the genomic and alphoid sequence levels, whereas gibbons had few detectable CENP-B motifs (Figure 4.4). OWMs had negligible amounts of CENP-B at either the genomic or AS levels (Figure 4.4), consistent with a previous report (Goldberg et al., 1996). Interestingly, three of the NWMs (*Callithrix jacchus, Cebus capucinus, Saimiri boliviensis*) contained an appreciable quantity of CENP-B boxes (Figure 4.4) as suggested by one previous study (Kugou et al., 2016); however, these sequences were not in the fraction homologous to hominoid and OWM AS (Figure 4.4). Combined with the analysis of HORs above, I conclude that the abundance of centromeric CENP-B boxes correlates with the fraction of centromeric satellite contained in HORs.

4.2.4 Presence of CENP-B boxes is inversely correlated with hairpin-forming dyad symmetries

In addition to variation in CENP-B boxes, previous studies identified dyad symmetries in alphoid DNA and the tendency for some alphoid DNA to adopt hairpin structures (Koch, 2000; Jonstrup et al., 2008). Visual examination of randomly selected monomeric units from various primates identified regions of dyad symmetry (Figure 4.5). Comprehensive enumeration of dyad symmetries over varying loop and palindrome lengths (Figure 4.5) of monomer fragments detected using Illumina sequencing revealed a striking clade-specific enrichment of particular symmetries in Old World Monkeys and New World Monkeys and a depletion of closely-spaced long-palindrome dyads in great apes. Given the predilection for dyad symmetries to give rise to hairpin and stem

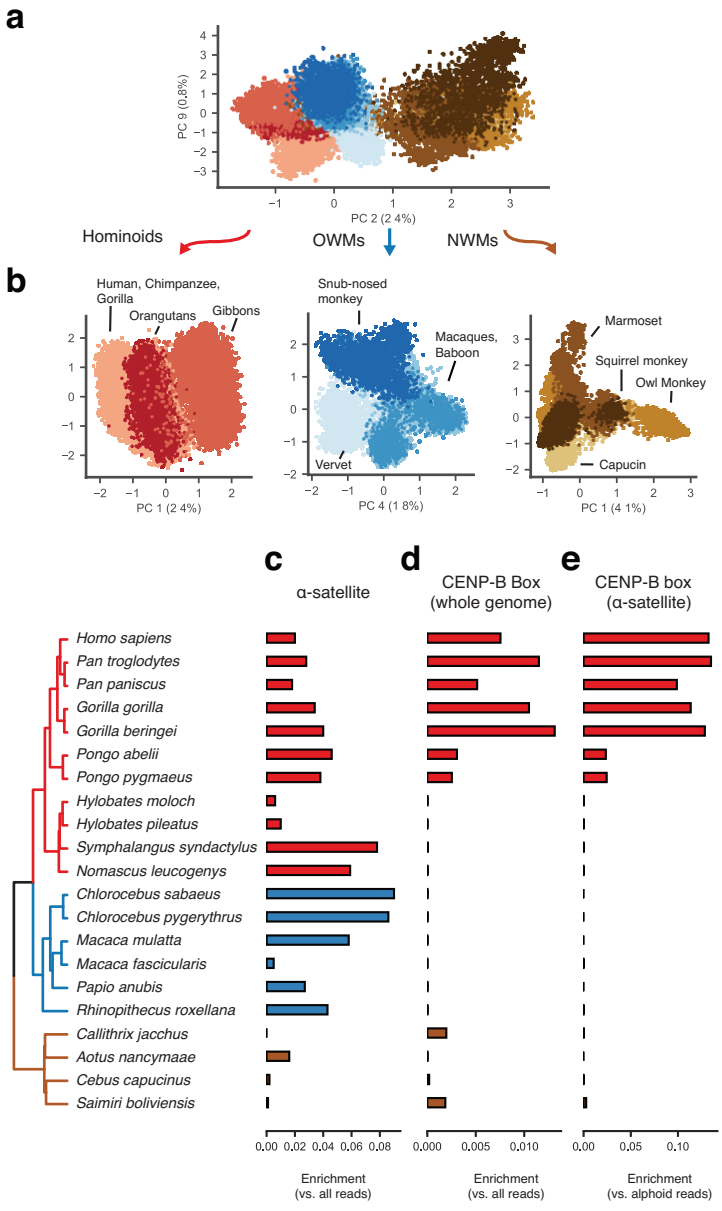

Figure 4.4 | Characterization of α-satellite in primates using short-read sequencing. (**a**) Principal components analysis of 5-mer vectors for a sampling of hominoids, Old World Monkeys (OWMs), and New World Monkeys (NWMs) separates these three clades. (**b**) Principal components analysis of 5-mer vectors for the species in each of the three clades separately robustly differentiates genera and/or species. (**c**) Estimated abundance of alphoid sequence (defined via profile hidden Markov model homology searching against human and OWM-defined alphoid sequence) and CENP-B boxes in the whole genome sequencing dataset or in alphoid reads specifically (**d,e**).

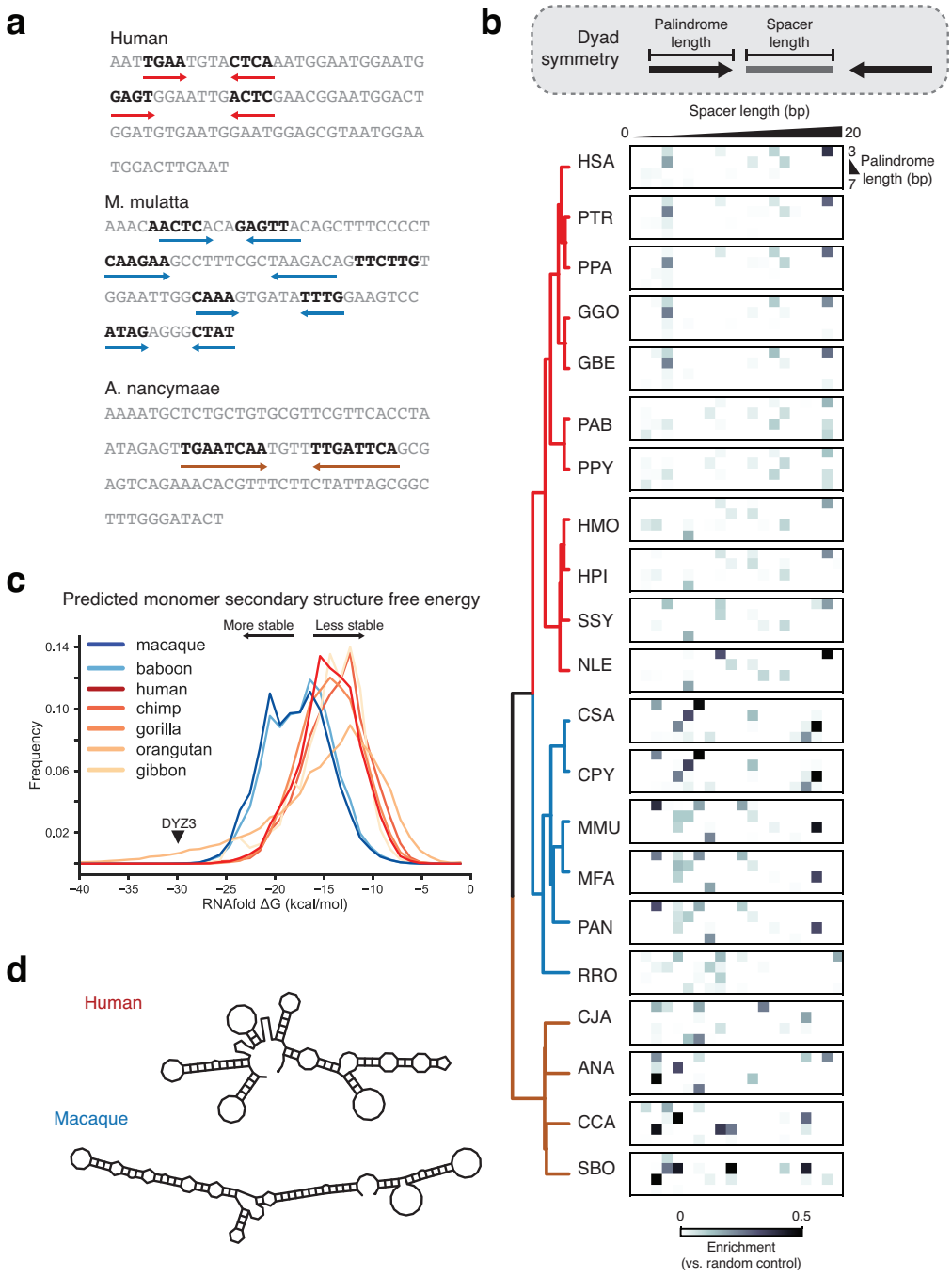

Figure 4.5 | Differential enrichment of dyad symmetries with the potential to form secondary structure in primate alphoid satellites. **(a)** Examples of dyad symmetries in segments of alphoid monomers from human, an Old World Monkey (*Macaca mulatta*), and a New World Monkey (*Aotus nancymaae*). **(b)** Enrichment of dyad symmetries of varying spacer and palindrome lengths in alphoid monomers from a sampling of simian primates with available next-generation sequencing data. **(c)** Predicted secondary structure formation free energies for 10^5 randomly sampled alphoid monomers from hominoid and Old World Monkey species with high-quality Sanger sequencing data. **(d)** Examples of predicted secondary structures for randomly chosen alphoid monomers from human and the rhesus macaque, an Old World Monkey (*Macaca mulatta*).

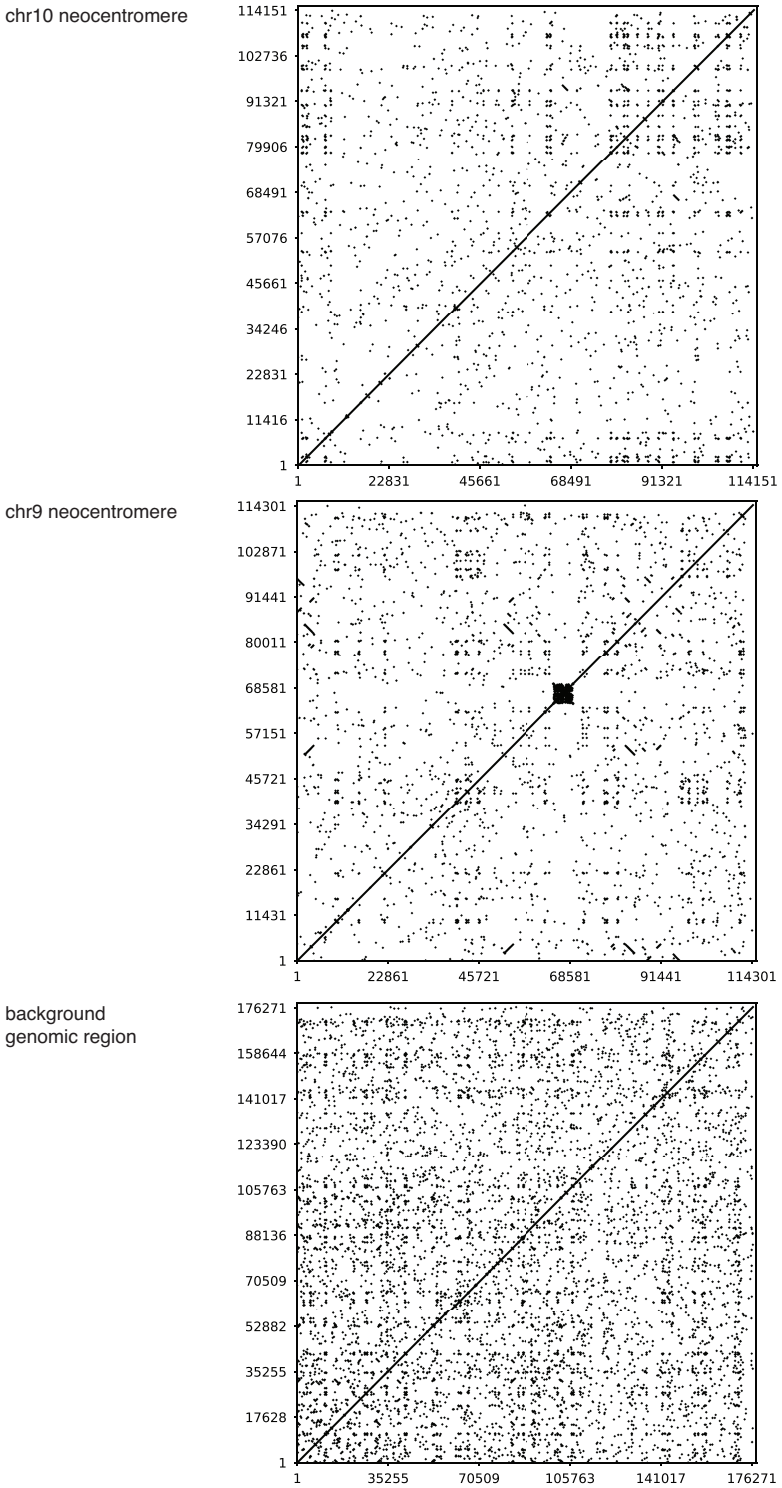

Figure 4.6 | Examples of neocentromeres with dyad symmetries. Conventional self-alignment dotplots for BAC sequences hybridizing to sequences from previously published CENP-A chromatin immunoprecipitation in cell lines harboring neocentromeres. Note that dyad symmetries manifest as diagonal lines that are orthogonal to the main diagonal and that dotplots are not scaled to reflect differences in the length of the BAC sequences analyzed.

loop structures, I turned to computational prediction of single-stranded DNA folding free energy (Lorenz et al., 2011) to determine whether alphoid monomers from different primates form favorable secondary structures. Consistent with the pattern of enrichment of dyad symmetries, the distributions of folding free energies for two OWMs are left-shifted relative to the distributions for hominoids (Figure 4.5). Interestingly, this tendency for alphoid sequence to form secondary structure inversely correlated with the observed distribution of centromeric CENP-B boxes (Figure 4.4). To further confirm this trend, I examined the DYZ3 satellite from the human Y chromosome, which does not contain CENP-B boxes, and found that, of all human AS monomers, it has among the lowest predicted folding free energies (Figure 4.5). I also examined BACS containing human neocentromere sequences (Alonso et al., 2007, 2003) relative to a region adjacent to one of these BAC sequences not known to form neocentromeres (Figure 4.6) and identified a slight enrichment for inversions consistent with a possible role for these sequences in specifying centromere identity (Koch, 2000; Jonstrup et al., 2008).

4.3 Discussion

I examined sequence variation in centromeric satellites in a broad sampling of species from the primate lineage. This analysis confirmed previously identified variation in centromeric satellite sequences for simian primates and identified novel non-alphoid putative centromeric repeats in prosimians. Further, this study demonstrates that HORs are a universal feature of primate centromeres and are not just restricted to hominoids and NWMs. The abundance of HORs, however, appears to correlate with fraction of alphoid sequence containing CENP-B boxes. Further, the presence of CENP-B boxes is inversely correlated with enrichment for dyad symmetries and predicted tendency to form hairpin/stem-loop DNA secondary structure.

Taken together, these lines of evidence suggest a model for specification of centromere identity based on the recognition of cruciform structures (Figure 4.7). In this model, centromere identity is determined by the formation of cruciform secondary structures in alphoid monomers, which are recognized by the CENP-A chaperone Holliday junction binding protein (HJURP) (Kato et al., 2007). In species with alphoid sequences that do not favorably adopt hairpin/stem-loop structures, CENP-B functions to create conditions that favor the formation of these structures, possibly by

deforming or locally melting the DNA backbone (Tanaka et al., 2001). Consistent with this model, alphoid sequence has been shown to form stem/loop structures *in vitro* (Jonstrup et al., 2008) and alphoid DNA in humans forms long single-stranded DNA domains (Aze et al., 2016).

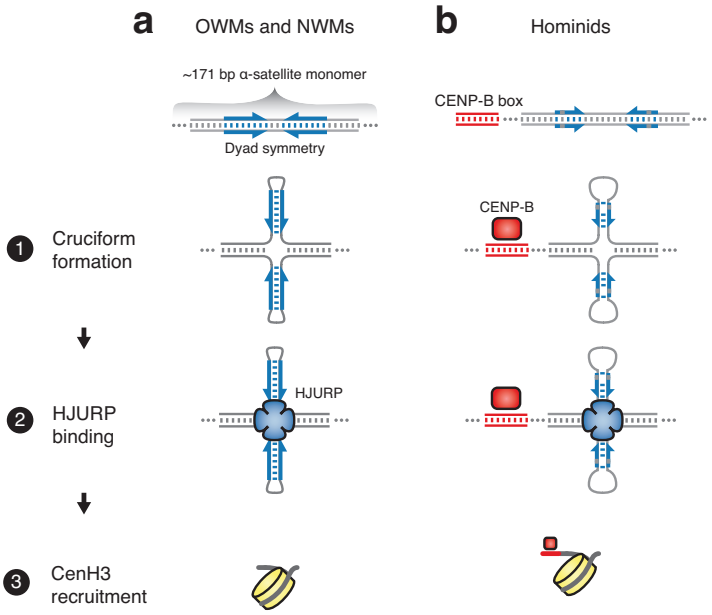

Figure 4.7 | A model for centromere specification. Repetitive centromeres vary in their predilection for forming cruciform structures exemplified by the α-satellite sequences of Old World Monkeys, which are predicted to form remarkably stable hairpins and stem-loops, and the hominids, which form less thermodynamically stable structures.

This model provides parsimonious explanations for a number of puzzling phenomena and is compatible with proposed functions for CENP-B in enhancing the fidelity of chromosome segregation (Fachinetti et al., 2015, 2013; Hoffmann et al., 2016). Importantly, it resolves the CENP-B paradox (Goldberg et al., 1996; Kipling and Warburton, 1997) by identifying a necessary role for CENP-B in facilitating the formation of secondary structures in species with alphoid monomers that do not favorably form cruciforms. This may also explain why CENP-B, although not required for native centromere specification in which centromere identity is maintained by the presence of CENP-A, is required for *de novo* centromerization of artificial chromosomes (Ohzeki et al., 2002; Ebersole et al., 2000; Henning et al., 1999; Masumoto et al., 1998; Ikeno et al., 1998; Harrington et al., 1997). The recognition of cruciforms and the possible enrichment of these sequences in the

few well-characterized human neocentromeres may also support this model. A structural mechanism is compatible the conservation of HJURP/Scm3 CENP-A chaperones in many eukaryotes (Sanchez-Pulido et al., 2009), while still allowing for a large sequence space that can be sampled during rapid evolution of centromeric DNA. Therefore, the recognition of cruciform DNA structures in satellite DNAs may be a general mechanism for the specification of centromere identity.

Chapter 5

CONCLUSIONS & PERSPECTIVES

> The existing hypotheses concerning the functions of satellite DNA usually leave the general reader with the impression of an insoluble morass. However, if one simply considers the hard data... it is soon apparent that the difficulties are more imaginary than real.
>
> —Miklos and John, ***American Journal of Human Genetics***, (1979)

Despite the important role played by satellite DNA in early studies of complex genomes, repetitive loci have been arguably understudied in the post-genomic era. In the centromere field, for example, the focus apparently shifted in the 1980s due in part to the identification of reagents allowing the characterization of centromeric proteins. However, the pendulum seems to be swinging back: advances in experimental and computational techniques for sequence analysis are increasingly bringing satellite DNA back into focus, especially in the context of chromatin biology. Below, I outline what I consider important important areas for advancing our understanding of repetitive centromeres.

The methodological contributions of this work are important from the perspective of better understanding the extent of variation in satellite DNAs at centromeres, which is critical given the evolutionary plasticity of these loci and their implication in disease. In addition to suggesting ways in which existing short-read sequencing data could be analyzed to explore single-nucleotide variants, the tools described here can be used to analyze the growing trove of single molecule sequencing data to interrogate large-scale polymorphism in repeat organization. The application of single-molecule sequencing to characterize disease-associated variation in individual patients (Merker et al., 2016) is especially exciting because it provides an opportunity for investigating the role centromeric DNA plays in hallmarks of disease such as genome instability. I also expect that these and other methods, when applied to single-molecule data generated from a variety of organisms through ongoing large scale sequencing projects, will yield new insights into the evolution of centromeres and satellite DNAs.

In addition to identifying extensive polymorphism in human centromeres, this work advances

testable models positing roles for CENP-B in the evolution of centromeric DNA and non-B-form DNA structures such as cruciforms or stem-loops in specification of centromere identity. Given the facility of manipulation of mammalian cells with advances in genome editing, I anticipate these ideas can be experimentally tested. Specification of centromeres by genetically-encoded DNA structures, in particular, could be easily tested using routinely used plasmid/chromosome maintenance assays (Harrington et al., 1997). If validated, a deeply conserved, genetic basis for centromere specification would have important implications for the creation of artificial chromosomes with a range of applications in biotechnology and, potentially, therapeutics.

The computational challenges inherent to sequence analysis in repetitive regions have precluded assembly using short-read sequencing and conventional algorithms (Treangen and Salzberg, 2011). However, the maturation of single-molecule sequencing technologies coupled with improved experimental approaches for isolating high-molecular weight DNA and new computational tools suggests that the goal of completing genome projects may be within reach (Miga, 2015). Current single-molecule sequencing technologies routinely produce average read lengths of ~10-20 kb, although there have been recent reports of ~1-Mb reads produced using nanopore sequencing (Jain et al., 2017a). Indeed, this technology has already been leveraged to sequence whole bacterial artificial chromosome inserts from the human Y chromosome centromere, permitting their linear ordering and thus the first assembly of a human centromeric array (Jain et al., 2017b). Centromere assemblies should further enable studies of genetic variation and, in conjunction with approaches for the precise manipulation of centromere components with genome editing tools and high-resolution approaches for analyzing chromatin, contribute to a clearer understanding of the role that satellite DNAs and centromeres play in evolution, human health, and disease.

BIBLIOGRAPHY

Mysterious satellites. *Nature*, 225(5236):899–900, Mar 1970. doi: 10.1038/225899a0.

Kami Ahmad and Steven Henikoff. The histone variant h3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol Cell*, 9(6):1191–200, Jun 2002.

Megan E Aldrup-Macdonald and Beth A Sullivan. The past, present, and future of human centromere genomics. *Genes (Basel)*, 5(1):33–50, Jan 2014.

Megan E Aldrup-MacDonald, Molly E Kuo, Lori L Sullivan, Kimberline Chew, and Beth A Sullivan. Genomic variation within alpha satellite dna influences centromere location on human chromosomes with metastable epialleles. *Genome Res*, 26(10):1301–1311, Oct 2016. doi: 10.1101/gr.206706.116.

I Alexandrov, A Kazakov, I Tumeneva, V Shepelev, and Y Yurov. Alpha-satellite dna of primates: old and new families. *Chromosoma*, 110(4):253–66, Aug 2001.

I A Alexandrov, S P Mitkevich, and Y B Yurov. The phylogeny of human chromosome specific alpha satellites. *Chromosoma*, 96(6):443–53, 1988.

I A Alexandrov, T D Mashkova, T A Akopian, L I Medvedev, L L Kisselev, S P Mitkevich, and Y B Yurov. Chromosome-specific alpha satellites: two distinct families on human chromosome 18. *Genomics*, 11(1):15–23, Sep 1991.

Can Alkan, Jeffrey A Bailey, Evan E Eichler, S Cenk Sahinalp, and Eray Tuzun. An algorithmic analysis of the role of unequal crossover in alpha-satellite dna evolution. *Genome Inform*, 13: 93–102, 2002.

Can Alkan, Evan E Eichler, Jeffrey A Bailey, S Cenk Sahinalp, and Eray Tüzün. The role of unequal crossover in alpha-satellite dna evolution: a computational analysis. *J Comput Biol*, 11(5):933–44, 2004. doi: 10.1089/cmb.2004.11.933.

Can Alkan, Mario Ventura, Nicoletta Archidiacono, Mariano Rocchi, S Cenk Sahinalp, and Evan E Eichler. Organization and evolution of primate centromeric dna from whole-genome shotgun sequence data. *PLoS Comput Biol*, 3(9):1807–18, Sep 2007. doi: 10.1371/journal.pcbi.0030181.

Alicia Alonso, Radma Mahmood, Shulan Li, Fanny Cheung, Kinya Yoda, and Peter E Warburton. Genomic microarray analysis reveals distinct locations for the cenp-a binding domains in three human chromosome 13q32 neocentromeres. *Hum Mol Genet*, 12(20):2711–21, Oct 2003. doi: 10.1093/hmg/ddg282.

Alicia Alonso, Björn Fritz, Dan Hasson, György Abrusan, Fanny Cheung, Kinya Yoda, Bernhard Radlwimmer, Andreas G Ladurner, and Peter E Warburton. Co-localization of cenp-c and cenp-h to discontinuous domains of cenp-a chromatin at human neocentromeres. *Genome Biol*, 8(7): R148, 2007. doi: 10.1186/gb-2007-8-7-r148.

G Alves, H N Seuánez, and T Fanning. Alpha satellite dna in neotropical primates (platyrrhini). *Chromosoma*, 103(4):262–7, Jul 1994.

G Alves, H N Seuánez, and T Fanning. A clade of new world primates with distinctive alphoid satellite dnas. *Mol Phylogenet Evol*, 9(2):220–4, Apr 1998. doi: 10.1006/mpev.1997.0462.

N B Atkin and V Brito-Babapulle. Heterochromatin polymorphism and human cancer. *Cancer Genet Cytogenet*, 3(3):261–72, Apr 1981.

Antoine Aze, Vincenzo Sannino, Paolo Soffientini, Angela Bachi, and Vincenzo Costanzo. Centromeric dna replication reconstitution reveals dna loops and atr checkpoint suppression. *Nat Cell Biol*, 18(6):684–91, Jun 2016. doi: 10.1038/ncb3344.

Sudarath Baicharoen, Visit Arsaithamkul, Yuriko Hirai, Toru Hara, Akihiko Koga, and Hirohisa Hirai. In situ hybridization analysis of gibbon chromosomes suggests that amplification of alpha satellite dna in the telomere region is confined to two of the four genera. *Genome*, 55(11):809–12, Nov 2012. doi: 10.1139/gen-2012-0123.

Sudarath Baicharoen, Takako Miyabe-Nishiwaki, Visit Arsaithamkul, Yuriko Hirai, Kwanruen Duangsa-ard, Boripat Siriaroonrat, Hiroshi Domae, Kornsorn Srikulnath, Akihiko Koga, and Hirohisa Hirai. Locational diversity of alpha satellite dna and intergeneric hybridization aspects in the nomascus and hylobates genera of small apes. *PLoS One*, 9(10):e109151, 2014. doi: 10.1371/journal.pone.0109151.

A Baldini, D I Smith, M Rocchi, O J Miller, and D A Miller. A human alphoid dna clone from the ecori dimeric family: genomic and internal organization and chromosomal assignment. *Genomics*, 5(4):822–8, Nov 1989.

A Baldini, D A Miller, O J Miller, O A Ryder, and A R Mitchell. A chimpanzee-derived chromosome-specific alpha satellite dna sequence conserved between chimpanzee and human. *Chromosoma*, 100(3):156–61, Mar 1991.

A Baldini, T Ried, V Shridhar, K Ogura, L D'Aiuto, M Rocchi, and D C Ward. An alphoid dna sequence conserved in all human and great ape chromosomes: evidence for ancient centromeric sequences at human chromosomal regions 2q21 and 9q13. *Hum Genet*, 90(6):577–83, Feb 1993.

M Baum and L Clarke. Fission yeast homologs of human cenp-b have redundant functions affecting cell growth and chromosome segregation. *Mol Cell Biol*, 20(8):2852–64, Apr 2000.

G Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic Acids Res*, 27(2): 573–80, Jan 1999.

T. Beridze. *Satellite DNA*. Springer-Verlag, 1986. ISBN 9783540158769. URL https://books.google.com/books?id=5ttqAAAAMAAJ.

Rafael Bernad, Patricia Sánchez, Teresa Rivera, Miriam Rodríguez-Corsino, Ekaterina Boyarchuk, Isabelle Vassias, Dominique Ray-Gallet, Alexei Arnaoutov, Mary Dasso, Geneviève Almouzni, and Ana Losada. Xenopus hjurp and condensin ii are required for cenp-a assembly. *J Cell Biol*, 192(4):569–82, Feb 2011. doi: 10.1083/jcb.201005136.

James A Birchler and Gernot G Presting. Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes. *Genes Dev*, 26(7): 638–40, Apr 2012. doi: 10.1101/gad.191049.112.

Ben E Black, Lars E T Jansen, Paul S Maddox, Daniel R Foltz, Arshad B Desai, Jagesh V Shah, and Don W Cleveland. Centromere identity maintained by nucleosomes assembled with histone h3 containing the cenp-a targeting domain. *Mol Cell*, 25(2):309–22, Jan 2007. doi: 10.1016/j.molcel.2006.12.018.

M D Blower and G H Karpen. The role of drosophila cid in kinetochore formation, cell-cycle progression and heterochromatin interactions. *Nat Cell Biol*, 3(8):730–9, Aug 2001. doi: 10.1038/35087045.

Michael D Blower. Centromeric transcription regulates aurora-b localization and activation. *Cell Rep*, 15(8):1624–33, May 2016. doi: 10.1016/j.celrep.2016.04.054.

Dani L Bodor, João F Mata, Mikhail Sergeev, Ana Filipa David, Kevan J Salimian, Tanya

Panchenko, Don W Cleveland, Ben E Black, Jagesh V Shah, and Lars Et Jansen. The quantitative architecture of centromeric chromatin. *Elife*, 3:e02137, Jul 2014.

M R Botchan. Bovine satellite i dna consists of repetitive units 1,400 base pairs in length. *Nature*, 251(5473):288–92, Sep 1974.

R J Britten and D E Kohne. Repeated sequences in dna. hundreds of thousands of copies of dna sequences have been incorporated into the genomes of higher organisms. *Science*, 161(3841): 529–40, Aug 1968.

D Broccoli, O J Miller, and D A Miller. Relationship of mouse minor satellite dna to centromere activity. *Cytogenet Cell Genet*, 54(3-4):182–6, 1990.

Kristin Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, Jun 2012. doi: 10.1038/nature11142.

D D Brown, P C Wensink, and E Jordan. A comparison of the ribosomal dna's of xenopus laevis and xenopus mulleri: the evolution of tandem genes. *J Mol Biol*, 63(1):57–73, Jan 1972.

B J Buchwitz, K Ahmad, L L Moore, M B Roth, and S Henikoff. A histone-h3-like protein in c. elegans. *Nature*, 401(6753):547–8, Oct 1999. doi: 10.1038/44062.

Minh Bui, Emilios K Dimitriadis, Christian Hoischen, Eunkyung An, Delphine Quénet, Sindy Giebe, Aleksandra Nita-Lazar, Stephan Diekmann, and Yamini Dalal. Cell-cycle-dependent structural transitions in the human cenp-a nucleosome in vivo. *Cell*, 150(2):317–26, Jul 2012. doi: 10.1016/j.cell.2012.05.035.

Samuel F Bunting and Andre Nussenzweig. End-joining, translocations and cancer. *Nat Rev Cancer*, 13(7):443–54, Jul 2013. doi: 10.1038/nrc3537.

Hugh P Cam, Ken-ichi Noma, Hirotaka Ebina, Henry L Levin, and Shiv I S Grewal. Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature*, 451(7177):431–6, Jan 2008. doi: 10.1038/nature06499.

Lucia Carbone, R. Alan Harris, Sante Gnerre, Krishna R. Veeramah, Belen Lorente-Galdos, John Huddleston, Thomas J. Meyer, Javier Herrero, Christian Roos, Bronwen Aken, Fabio Anaclerio, Nicoletta Archidiacono, Carl Baker, Daniel Barrell, Mark A. Batzer, Kathryn Beal, Antoine Blancher, Craig L. Bohrson, Markus Brameier, Michael S. Campbell, Oronzo Capozzi, Claudio Casola, Giorgia Chiatante, Andrew Cree, Annette Damert, Pieter J. de Jong, Laura Dumas, Marcos Fernandez-Callejo, Paul Flicek, Nina V. Fuchs, Ivo Gut, Marta Gut, Matthew W. Hahn, Jessica Hernandez-Rodriguez, LaDeana W. Hillier, Robert Hubley, Bianca Ianc, Zsuzsanna Izsvak, Nina G. Jablonski, Laurel M. Johnstone, Anis Karimpour-Fard, Miriam K. Konkel, Dennis Kostka, Nathan H. Lazar, Sandra L. Lee, Lora R. Lewis, Yue Liu, Devin P. Locke, Swapan Mallick, Fernando L. Mendez, Matthieu Muffato, Lynne V. Nazareth, Kimberly A. Nevonen, Majesta O'Bleness, Cornelia Ochis, Duncan T. Odom, Katherine S. Pollard, Javier Quilez, David Reich, Mariano Rocchi, Gerald G. Schumann, Stephen Searle, James M. Sikela, Gabriella Skollar, Arian Smit, Kemal Sonmez, Boudewijn ten Hallers, Elizabeth Terhune, Gregg W. C. Thomas, Brygg Ullmer, Mario Ventura, Jerilyn A. Walker, Jeffrey D. Wall, Lutz Walter, Michelle C. Ward, Sarah J. Wheelan, Christopher W. Whelan, Simon White, Larry J. Wilhelm, August E. Woerner, Mark Yandell, Baoli Zhu, Michael F. Hammer, Tomas Marques-Bonet, Evan E. Eichler, Lucinda Fulton, Catrina Fronick, Donna M. Muzny, Wesley C. Warren, Kim C. Worley, Jeffrey Rogers, Richard K. Wilson, and Richard A. Gibbs. Gibbon genome and the fast karyotype evolution of small apes. *Nature*, 513(7517):195–201, 09 2014. URL http://dx.doi.org/10.1038/nature13679.

Mauricio O Carneiro, Carsten Russ, Michael G Ross, Stacey B Gabriel, Chad Nusbaum, and Mark A

DePristo. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, 13:375, Aug 2012. doi: 10.1186/1471-2164-13-375.

B Cazaux, J Catalan, F Justy, C Escudé, E Desmarais, and J Britton-Davidian. Evolution of the structure and composition of house mouse satellite dna sequences in the subgenus mus (rodentia: Muridea): a cytogenomic approach. *Chromosoma*, 122(3):209–20, Jun 2013. doi: 10.1007/s00412-013-0402-4.

A Cellamare, C R Catacchio, C Alkan, G Giannuzzi, F Antonacci, M F Cardone, G Della Valle, M Malig, M Rocchi, E E Eichler, and M Ventura. New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset. *Mol Biol Evol*, 26(8):1889–900, Aug 2009. doi: 10.1093/molbev/msp101.

Mark J P Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M Landolin, John A Stamatoyannopoulos, Michael W Hunkapiller, Jonas Korlach, and Evan E Eichler. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–11, Jan 2015. doi: 10.1038/nature13907.

Geoffrey K Chambers, Caitlin Curtis, Craig D Millar, Leon Huynen, and David M Lambert. Dna fingerprinting in zoology: past, present, future. *Investig Genet*, 5(1):3, Feb 2014. doi: 10.1186/2041-2223-5-3.

F Lyn Chan, Owen J Marshall, Richard Saffery, Bo Won Kim, Elizabeth Earle, K H Andy Choo, and Lee H Wong. Active transcription and essential role of rna polymerase ii at the centromere during mitosis. *Proc Natl Acad Sci U S A*, 109(6):1979–84, Feb 2012. doi: 10.1073/pnas.1108705109.

B Charlesworth, P Sniegowski, and W Stephan. The evolutionary dynamics of repetitive dna in eukaryotes. *Nature*, 371(6494):215–20, Sep 1994. doi: 10.1038/371215a0.

Iain M Cheeseman and Arshad Desai. Molecular architecture of the kinetochore-microtubule interface. *Nat Rev Mol Cell Biol*, 9(1):33–46, Jan 2008. doi: 10.1038/nrm2310.

Chin-Chi Chen, Mekonnen Lemma Dechassa, Emily Bettini, Mary B Ledoux, Christian Belisario, Patrick Heun, Karolin Luger, and Barbara G Mellone. Cal1 is the drosophila cenp-a assembly factor. *J Cell Biol*, 204(3):313–29, Feb 2014. doi: 10.1083/jcb.201305036.

Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nat Methods*, 10(6):563–9, Jun 2013. doi: 10.1038/nmeth.2474.

Lukáš Chmátal, Sofia I Gabriel, George P Mitsainas, Jessica Martínez-Vargas, Jacint Ventura, Jeremy B Searle, Richard M Schultz, and Michael A Lampson. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr Biol*, 24(19):2295–300, Oct 2014. doi: 10.1016/j.cub.2014.08.017.

K H Choo. Why is the centromere so cold? *Genome Res*, 8(2):81–2, Feb 1998.

L Clarke and J Carbon. The structure and function of yeast centromeres. *Annu Rev Genet*, 19:29–55, 1985. doi: 10.1146/annurev.ge.19.120185.000333.

Don W Cleveland, Yinghui Mao, and Kevin F Sullivan. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell*, 112(4):407–21, Feb 2003.

Christine A Codomo, Takehito Furuyama, and Steven Henikoff. Cenp-a octamers do not confer a reduction in nucleosome height by afm. *Nat Struct Mol Biol*, 21(1):4–5, Jan 2014. doi: 10.1038/nsmb.2743.

E Coen, T Strachan, and G Dover. Dynamics of concerted evolution of ribosomal dna and histone gene families in the melanogaster species subgroup of drosophila. *J Mol Biol*, 158(1):17–35, Jun 1982.

E S Coen and G A Dover. Unequal exchanges and the coevolution of x and y rdna arrays in drosophila melanogaster. *Cell*, 33(3):849–55, Jul 1983.

Sarit Cohen, Neta Agmon, Olga Sobol, and Daniel Segal. Extrachromosomal circles of satellite repeats and 5s ribosomal dna in human cells. *Mob DNA*, 1(1):11, Mar 2010. doi: 10.1186/1759-8753-1-11.

Francis Collins. Has the revolution arrived? *Nature*, 464(7289):674–5, Apr 2010. doi: 10.1038/464674a.

G P Copenhaver, W E Browne, and D Preuss. Assaying genome-wide recombination and centromere functions with arabidopsis tetrads. *Proc Natl Acad Sci U S A*, 95(1):247–52, Jan 1998.

G P Copenhaver, K Nickel, T Kuromori, M I Benito, S Kaul, X Lin, M Bevan, G Murphy, B Harris, L D Parnell, W R McCombie, R A Martienssen, M Marra, and D Preuss. Genetic definition and sequence analysis of arabidopsis centromeres. *Science*, 286(5449):2468–74, Dec 1999.

G Corneo, E Ginelli, C Soave, and G Bernardi. Isolation and characterization of mouse and guinea pig satellite deoxyribonucleic acids. *Biochemistry*, 7(12):4373–9, Dec 1968.

A K Csink and S Henikoff. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet*, 14(5):200–4, May 1998.

Yamini Dalal, Takehito Furuyama, Danielle Vermaak, and Steven Henikoff. Structure, dynamics, and evolution of centromeric nucleosomes. *Proc Natl Acad Sci U S A*, 104(41):15974–81, Oct 2007a. doi: 10.1073/pnas.0707648104.

Yamini Dalal, Hongda Wang, Stuart Lindsay, and Steven Henikoff. Tetrameric structure of centromeric nucleosomes in interphase drosophila cells. *PLoS Biol*, 5(8):e218, Aug 2007b. doi: 10.1371/journal.pbio.0050218.

S M Darling, J M Crampton, and R Williamson. Organization of a family of highly repetitive sequences within the human genome. *J Mol Biol*, 154(1):51–63, Jan 1982.

Alfredo De Bustos, Angeles Cuadrado, and Nicolás Jouve. Sequencing of long stretches of repetitive dna. *Sci Rep*, 6:36665, Nov 2016. doi: 10.1038/srep36665.

Emilios K Dimitriadis, Christian Weber, Rajbir K Gill, Stephan Diekmann, and Yamini Dalal. Tetrameric organization of vertebrate centromeric nucleosomes. *Proc Natl Acad Sci U S A*, 107 (47):20317–22, Nov 2010. doi: 10.1073/pnas.1009563107.

L Donehower, C Furlong, D Gillespie, and D Kurnit. Dna sequence of baboon highly repeated dna: evidence for evolution by nonrandom unequal crossovers. *Proc Natl Acad Sci U S A*, 77(4): 2129–33, Apr 1980.

G Dover. Molecular drive: a cohesive mode of species evolution. *Nature*, 299(5879):111–7, Sep 1982.

Ines A Drinnenberg, Dakota deYoung, Steven Henikoff, and Harmit Singh Malik. Recurrent loss of cenh3 is associated with independent transitions to holocentricity in insects. *Elife*, 3, Sep 2014. doi: 10.7554/eLife.03676.

Ines A Drinnenberg, Steven Henikoff, and Harmit S Malik. Evolutionary turnover of kinetochore proteins: A ship of theseus? *Trends Cell Biol*, 26(7):498–510, Jul 2016. doi: 10.1016/j.tcb.2016.01.005.

S J Durfy and H F Willard. Patterns of intra- and interarray sequence variation in alpha satellite

from the human x chromosome: evidence for short-range homogenization of tandemly repeated dna sequences. *Genomics*, 5(4):810–21, Nov 1989.

S J Durfy and H F Willard. Concerted evolution of primate alpha satellite dna. evidence for an ancestral sequence shared by gorilla and human x chromosome alpha satellite. *J Mol Biol*, 216 (3):555–66, Dec 1990. doi: 10.1016/0022-2836(90)90383-W.

W C Earnshaw and N Rothfield. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma*, 91(3-4):313–21, 1985.

W C Earnshaw, K F Sullivan, P S Machlin, C A Cooke, D A Kaiser, T D Pollard, N F Rothfield, and D W Cleveland. Molecular cloning of cdna for cenp-b, the major human centromere autoantigen. *J Cell Biol*, 104(4):817–29, Apr 1987.

W C Earnshaw, H Ratrie, 3rd, and G Stetten. Visualization of centromere proteins cenp-b and cenp-c on a stable dicentric chromosome in cytological spreads. *Chromosoma*, 98(1):1–12, Jun 1989.

William C Earnshaw. Discovering centromere proteins: from cold white hands to the a, b, c of cenps. *Nat Rev Mol Cell Biol*, 16(7):443–9, 07 2015. doi: 10.1038/nrm4001.

T A Ebersole, A Ross, E Clark, N McGill, D Schindelhauer, H Cooke, and B Grimes. Mammalian artificial chromosome formation from circular alphoid input dna does not require telomere repeats. *Hum Mol Genet*, 9(11):1623–31, Jul 2000.

Evan E Eichler, Royden A Clark, and Xinwei She. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet*, 5(5):345–54, May 2004. doi: 10.1038/nrg1322.

Daniele Fachinetti, H Diego Folco, Yael Nechemia-Arbely, Luis P Valente, Kristen Nguyen, Alex J Wong, Quan Zhu, Andrew J Holland, Arshad Desai, Lars E T Jansen, and Don W Cleveland. A two-step mechanism for epigenetic specification of centromere identity and function. *Nat Cell Biol*, 15(9):1056–66, Sep 2013. doi: 10.1038/ncb2805.

Daniele Fachinetti, Joo Seok Han, Moira A McMahon, Peter Ly, Amira Abdullah, Alex J Wong, and Don W Cleveland. Dna sequence-specific binding of cenp-b enhances the fidelity of human centromere function. *Dev Cell*, 33(3):314–27, May 2015. doi: 10.1016/j.devcel.2015.03.020.

T G Fanning, H N Seuánez, and L Forman. Satellite dna sequences in the neotropical marmoset callimico goeldii (primates, platyrrhini). *Chromosoma*, 98(6):396–401, Dec 1989.

T G Fanning, H N Seuánez, and L Forman. Satellite dna sequences in the new world primate cebus apella plaatyrrhini, primates). *Chromosoma*, 102(5):306–11, May 1993.

Lila Fishman and Arpiar Saunders. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science*, 322(5907):1559–62, Dec 2008. doi: 10.1126/science.1161406.

Walther Flemming. *Zellsubstanz, kern und zelltheilung*. Vogel, 1882.

Daniel R Foltz, Lars E T Jansen, Aaron O Bailey, John R Yates, 3rd, Emily A Bassett, Stacey Wood, Ben E Black, and Don W Cleveland. Centromere-specific assembly of cenp-a nucleosomes is mediated by hjurp. *Cell*, 137(3):472–84, May 2009. doi: 10.1016/j.cell.2009.02.039.

Takehito Furuyama and Steven Henikoff. Centromeric nucleosomes induce positive dna supercoils. *Cell*, 138(1):104–13, Jul 2009. doi: 10.1016/j.cell.2009.04.049.

Takehito Furuyama, Christine A Codomo, and Steven Henikoff. Reconstitution of hemisomes on budding yeast centromeric dna. *Nucleic Acids Res*, 41(11):5769–83, Jun 2013. doi: 10.1093/nar/gkt314.

C Gaff, D du Sart, P Kalitsis, R Iannello, A Nagy, and K H Choo. A novel nuclear protein binds

centromeric alpha satellite dna. *Hum Mol Genet*, 3(5):711–6, May 1994.

Dongying Gao, Ning Jiang, Rod A Wing, Jiming Jiang, and Scott A Jackson. Transposons play an important role in the evolution and diversification of centromeres among closely related species. *Front Plant Sci*, 6:216, 2015. doi: 10.3389/fpls.2015.00216.

Reto Gassmann, Andreas Rechtsteiner, Karen W Yuen, Andrew Muroyama, Thea Egelhofer, Laura Gaydos, Francie Barron, Paul Maddox, Anthony Essex, Joost Monen, Sevinc Ercan, Jason D Lieb, Karen Oegema, Susan Strome, and Arshad Desai. An inverse relationship to germline transcription defines centromeric chromatin in c. elegans. *Nature*, 484(7395):534–7, Apr 2012. doi: 10.1038/nature10973.

Jennifer R Gatchel and Huda Y Zoghbi. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*, 6(10):743–55, Oct 2005. doi: 10.1038/nrg1691.

Simona Giunta and Hironori Funabiki. 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(8):1928–1933, Feb 2017. doi: 10.1073/pnas.1615133114.

I G Goldberg, H Sawhney, A F Pluta, P E Warburton, and W C Earnshaw. Surprising deficiency of cenp-b binding sites in african green monkey alpha-satellite dna: implications for cenp-b function at centromeres. *Mol Cell Biol*, 16(9):5156–68, Sep 1996.

Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–51, 05 2016. doi: 10.1038/nrg.2016.49.

G M Greig, P E Warburton, and H F Willard. Organization and evolution of an alpha satellite dna subset shared by human chromosomes 13 and 21. *J Mol Evol*, 37(5):464–75, Nov 1993.

Mounia Guenatri, Delphine Bailly, Christèle Maison, and Geneviève Almouzni. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J Cell Biol*, 166(4):493–505, Aug 2004. doi: 10.1083/jcb.200403109.

H H Guldner, H J Lakomek, and F A Bautz. Human anti-centromere sera recognise a 19.5 kd non-histone chromosomal protein from hela cells. *Clin Exp Immunol*, 58(1):13–20, Oct 1984.

T Haaf and D C Ward. Rabl orientation of cenp-b box sequences in tupaia belangeri fibroblasts. *Cytogenet Cell Genet*, 70(3-4):258–62, 1995.

T Haaf and H F Willard. Chromosome-specific alpha-satellite dna from the centromere of chimpanzee chromosome 4. *Chromosoma*, 106(4):226–32, Sep 1997.

T Haaf and H F Willard. Orangutan alpha-satellite monomers are closely related to the human consensus sequence. *Mamm Genome*, 9(6):440–7, Jun 1998.

T Haaf, P E Warburton, and H F Willard. Integration of human alpha-satellite dna into simian chromosomes: centromere protein binding and disruption of normal chromosome segregation. *Cell*, 70(4):681–96, Aug 1992.

T Haaf, A G Mater, J Wienberg, and D C Ward. Presence and abundance of cenp-b box sequences in great ape subsets of primate-specific alpha-satellite dna. *J Mol Evol*, 41(4):487–91, Oct 1995.

Colin M Hammond, Caroline B Strømme, Hongda Huang, Dinshaw J Patel, and Anja Groth. Histone chaperone networks shaping chromatin function. *Nat Rev Mol Cell Biol*, 18(3):141–158, Mar 2017. doi: 10.1038/nrm.2016.159.

Toru Hara, Yuriko Hirai, Israt Jahan, Hirohisa Hirai, and Akihiko Koga. Tandem repeat sequences evolutionarily related to sva-type retrotransposons are expanded in the centromere region of the western hoolock gibbon, a small ape. *J Hum Genet*, 57(12):760–5, Dec 2012. doi: 10.1038/jhg.2012.

107.

J J Harrington, G Van Bokkelen, R W Mays, K Gustashaw, and H F Willard. Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. *Nat Genet*, 15(4):345–55, Apr 1997. doi: 10.1038/ng0497-345.

Dan Hasson, Tanya Panchenko, Kevan J Salimian, Mishah U Salman, Nikolina Sekulic, Alicia Alonso, Peter E Warburton, and Ben E Black. The octamer is the major form of cenp-a nucleosomes at human centromeres. *Nat Struct Mol Biol*, 20(6):687–95, Jun 2013. doi: 10.1038/nsmb.2562.

Jorja G Henikoff, Jason A Belsky, Kristina Krassovsky, David M MacAlpine, and Steven Henikoff. Epigenome characterization at single base-pair resolution. *Proc Natl Acad Sci U S A*, 108(45):18318–23, Nov 2011. doi: 10.1073/pnas.1110731108.

Jorja G Henikoff, Jitendra Thakur, Sivakanthan Kasinathan, and Steven Henikoff. A unique chromatin complex occupies young α-satellite arrays of human centromeres. *Sci Adv*, 1(1), Feb 2015. doi: 10.1126/sciadv.1400234.

S Henikoff, K Ahmad, J S Platero, and B van Steensel. Heterochromatic deposition of centromeric histone h3-like proteins. *Proc Natl Acad Sci U S A*, 97(2):716–21, Jan 2000.

S Henikoff, K Ahmad, and H S Malik. The centromere paradox: stable inheritance with rapidly evolving dna. *Science*, 293(5532):1098–102, Aug 2001. doi: 10.1126/science.1062939.

Steven Henikoff. Near the edge of a chromosome's "black hole". *Trends Genet*, 18(4):165–7, Apr 2002.

Steven Henikoff and Kami Ahmad. Assembly of variant histones into chromatin. *Annu Rev Cell Dev Biol*, 21:133–53, 2005. doi: 10.1146/annurev.cellbio.21.012704.133518.

Steven Henikoff and Harmit S Malik. Centromeres: selfish drivers. *Nature*, 417(6886):227, May 2002. doi: 10.1038/417227a.

Steven Henikoff, Srinivas Ramachandran, Kristina Krassovsky, Terri D Bryson, Christine A Codomo, Kristin Brogaard, Jonathan Widom, Ji-Ping Wang, and Jorja G Henikoff. The budding yeast centromere dna element ii wraps a stable cse4 hemisome in either orientation in vivo. *Elife*, 3:e01861, Apr 2014. doi: 10.7554/eLife.01861.

W Hennig and P M Walker. Variations in the dna from two rodent families (cricetidae and muridae). *Nature*, 225(5236):915–9, Mar 1970.

K A Henning, E A Novotny, S T Compton, X Y Guan, P P Liu, and M A Ashlock. Human artificial chromosomes generated by modification of a yeast artificial chromosome containing both human alpha satellite and single-copy dna sequences. *Proc Natl Acad Sci U S A*, 96(2):592–7, Jan 1999.

J S Pat Heslop-Harrison and Trude Schwarzacher. Nucleosomes and centromeric dna packaging. *Proc Natl Acad Sci U S A*, 110(50):19974–5, Dec 2013. doi: 10.1073/pnas.1319945110.

Sebastian Hoffmann, Marie Dumont, Viviana Barra, Peter Ly, Yael Nechemia-Arbely, Moira A McMahon, Solène Hervé, Don W Cleveland, and Daniele Fachinetti. Cenp-a is dispensable for mitotic centromere function after initial centromere/kinetochore assembly. *Cell Rep*, 17(9):2394–2404, Nov 2016. doi: 10.1016/j.celrep.2016.10.084.

W Hörz and W Altenburger. Nucleotide sequence of mouse satellite dna. *Nucleic Acids Res*, 9(3):683–96, Feb 1981.

Roger A Hoskins, Christopher D Smith, Joseph W Carlson, A Bernardo Carvalho, Aaron Halpern, Joshua S Kaminker, Cameron Kennedy, Chris J Mungall, Beth A Sullivan, Granger G Sutton,

Jiro C Yasuhara, Barbara T Wakimoto, Eugene W Myers, Susan E Celniker, Gerald M Rubin, and Gary H Karpen. Heterochromatic sequences in a drosophila whole-genome shotgun assembly. *Genome Biol*, 3(12):RESEARCH0085, 2002.

E V Howman, K J Fowler, A J Newson, S Redward, A C MacDonald, P Kalitsis, and K H Choo. Early disruption of centromeric chromatin organization in centromere protein a (cenpa) null mice. *Proc Natl Acad Sci U S A*, 97(3):1148–53, Feb 2000.

Robert Hubley, Robert D Finn, Jody Clements, Sean R Eddy, Thomas A Jones, Weidong Bao, Arian F A Smit, and Travis J Wheeler. The dfam database of repetitive dna families. *Nucleic Acids Res*, 44(D1):D81–9, Jan 2016. doi: 10.1093/nar/gkv1272.

John Huddleston, Mark J P Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A Graves-Lindsay, Katherine M Munson, Zev N Kronenberg, Laura Vives, Paul Peluso, Matthew Boitano, Chen-Shin Chin, Jonas Korlach, Richard K Wilson, and Evan E Eichler. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*, 27(5):677–685, May 2017. doi: 10.1101/gr.214007.116.

D F Hudson, K J Fowler, E Earle, R Saffery, P Kalitsis, H Trowell, J Hill, N G Wreford, D M de Kretser, M R Cancilla, E Howman, L Hii, S M Cutts, D V Irvine, and K H Choo. Centromere protein b null mice are mitotically and meiotically normal but have lower body and testis weights. *J Cell Biol*, 141(2):309–19, Apr 1998.

M Ikeno, B Grimes, T Okazaki, M Nakano, K Saitoh, H Hoshino, N I McGill, H Cooke, and H Masumoto. Construction of yac-based mammalian artificial chromosomes. *Nat Biotechnol*, 16(5): 431–9, May 1998. doi: 10.1038/nbt0598-431.

E W Jabs, S F Wolf, and B R Migeon. Characterization of a cloned dna sequence that is present at centromeres of all human autosomes and the x chromosome and shows polymorphic variation. *Proc Natl Acad Sci U S A*, 81(15):4884–8, Aug 1984a.

E W Jabs, S F Wolf, and B R Migeon. Characterization of reiterated human dna with respect to mammalian x chromosome homology. *Somat Cell Mol Genet*, 10(1):93–103, Jan 1984b.

Isabel Jaco, Andrés Canela, Elsa Vera, and Maria A Blasco. Centromere mitotic recombination in mammalian cells. *J Cell Biol*, 181(6):885–92, Jun 2008. doi: 10.1083/jcb.200803042.

Miten Jain, Sergey Koren, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Karen H Miga, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M. Phillippy, Jared T Simpson, Nicholas James Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*, 2017a. doi: 10.1101/128835. URL https://www.biorxiv.org/content/early/2017/04/20/128835.

Miten Jain, Hugh E. Olsen, Daniel J. Turner, David Stoddart, Kira V. Bulazel, Benedict Paten, David Haussler, Huntington Willard, Mark Akeson, and Karen H. Miga. Linear assembly of a human y centromere using nanopore long reads. *bioRxiv*, 2017b. doi: 10.1101/170373. URL https://www.biorxiv.org/content/early/2017/07/31/170373.

Lars E T Jansen, Ben E Black, Daniel R Foltz, and Don W Cleveland. Propagation of centromeric chromatin requires exit from mitosis. *J Cell Biol*, 176(6):795–805, Mar 2007. doi: 10.1083/jcb.200701066.

Peter Johansen and Hugh P Cam. Suppression of meiotic recombination by cenp-b homologs in schizosaccharomyces pombe. *Genetics*, 201(3):897–904, Nov 2015. doi: 10.1534/genetics.115.

179465.

Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–39, Jan 2013. doi: 10.1016/j.cell.2012.12.009.

K W Jones. Chromosomal and nuclear location of mouse satellite dna in individual cells. *Nature*, 225(5236):912–5, Mar 1970.

K W Jones. Satellite dna. *J Med Genet*, 10(3):273–81, Sep 1973.

Anette Thyssen Jonstrup, Tina Thomsen, Yong Wang, Birgitta R Knudsen, Jørn Koch, and Anni H Andersen. Hairpin structures formed by alpha satellite dna of human centromeres are cleaved by human topoisomerase iialpha. *Nucleic Acids Res*, 36(19):6165–74, Nov 2008. doi: 10.1093/nar/gkn640.

A L Jørgensen, C J Bostock, and A L Bak. Chromosome-specific subfamilies within human alphoid repetitive dna. *J Mol Biol*, 187(2):185–96, Jan 1986.

A L Jørgensen, C Jones, C J Bostock, and A L Bak. Different subfamilies of alphoid repetitive dna are present on the human and chimpanzee homologous chromosomes 21 and 22. *EMBO J*, 6(6):1691–6, Jun 1987.

A L Jørgensen, S Kølvraa, C Jones, and A L Bak. A subfamily of alphoid repetitive dna shared by the nor-bearing human chromosomes 14 and 22. *Genomics*, 3(2):100–9, Aug 1988.

A Joseph, A R Mitchell, and O J Miller. The organization of the mouse satellite dna at centromeres. *Exp Cell Res*, 183(2):494–500, Aug 1989.

Jerzy Jurka, Vladimir V Kapitonov, Oleksiy Kohany, and Michael V Jurka. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet*, 8:241–59, 2007. doi: 10.1146/annurev.genom.8.080706.092416.

M Kapoor, R Montes de Oca Luna, G Liu, G Lozano, C Cummings, M Mancini, I Ouspenski, B R Brinkley, and G S May. The cenpb gene is not essential in mice. *Chromosoma*, 107(8):570–6, Dec 1998.

Tatsuya Kato, Nagato Sato, Satoshi Hayama, Takumi Yamabuki, Tomoo Ito, Masaki Miyamoto, Satoshi Kondo, Yusuke Nakamura, and Yataro Daigo. Activation of holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. *Cancer Res*, 67(18):8544–53, Sep 2007. doi: 10.1158/0008-5472.CAN-07-1307.

Alexei E Kazakov, Valery A Shepelev, Irina G Tumeneva, Alexander A Alexandrov, Yuri B Yurov, and Ivan A Alexandrov. Interspersed repeats are found predominantly in the "old" alpha satellite families. *Genomics*, 82(6):619–27, Dec 2003.

Daniel E Khost, Danna G Eickbush, and Amanda M Larracuente. Single-molecule sequencing resolves the detailed structure of complex satellite dna loci in drosophila melanogaster. *Genome Res*, 27(5):709–721, May 2017. doi: 10.1101/gr.213512.116.

D Kipling and P E Warburton. Centromeres, cenp-b and tigger too. *Trends Genet*, 13(4):141–5, Apr 1997.

D Kipling, H E Ackford, B A Taylor, and H J Cooke. Mouse minor satellite dna genetically maps to the centromere and is physically linked to the proximal telomere. *Genomics*, 11(2):235–41, Oct 1991.

D Kipling, A R Mitchell, H Masumoto, H E Wilson, L Nicol, and H J Cooke. Cenp-b binds a novel

centromeric sequence in the asian mouse mus caroli. *Mol Cell Biol*, 15(8):4009–20, Aug 1995.

S Kit. Equilibrium sedimentation in density gradients of dna preparations from animal tissues. *J Mol Biol*, 3:711–6, Dec 1961.

K Kitagawa, H Masumoto, M Ikeda, and T Okazaki. Analysis of protein-dna and protein-protein interactions of centromere protein b (cenp-b) and properties of the dna-cenp-b complex in the cell cycle. *Mol Cell Biol*, 15(3):1602–12, Mar 1995.

J Koch. Neocentromeres and alpha satellite: a proposed structural code for functional human centromere dna. *Hum Mol Genet*, 9(2):149–54, Jan 2000.

Aleksey S Komissarov, Ekaterina V Gavrilova, Sergey Ju Demin, Alexander M Ishov, and Olga I Podgornaya. Tandemly repeated dna families in the mouse genome. *BMC Genomics*, 12:531, Oct 2011. doi: 10.1186/1471-2164-12-531.

Sergey Koren, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis, and Adam M Phillippy. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*, 30(7):693–700, Jul 2012. doi: 10.1038/nbt.2280.

Kristina Krassovsky and Steven Henikoff. Distinct chromatin features characterize different classes of repeat sequences in drosophila melanogaster. *BMC Genomics*, 15:105, Feb 2014. doi: 10.1186/1471-2164-15-105.

Kristina Krassovsky, Jorja G Henikoff, and Steven Henikoff. Tripartite organization of centromeric chromatin in budding yeast. *Proc Natl Acad Sci U S A*, 109(1):243–8, Jan 2012. doi: 10.1073/pnas.1118898109.

Kazuto Kugou, Hirohisa Hirai, Hiroshi Masumoto, and Akihiko Koga. Formation of functional cenp-b boxes at diverse locations in repeat units of centromeric dna in new world monkeys. *Sci Rep*, 6:27833, Jun 2016. doi: 10.1038/srep27833.

R M Kuhn, L Clarke, and J Carbon. Clustered trna genes in schizosaccharomyces pombe centromeric dna sequence repeats. *Proc Natl Acad Sci U S A*, 88(4):1306–10, Feb 1991.

Jonathan C Lamb and James A Birchler. The role of dna sequence in centromere formation. *Genome Biol*, 4(5):214, 2003.

E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls,

E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001. doi: 10.1038/35057062.

A M Laurent, J Puechberty, and G Roizès. Hypothesis: for the worst and for the best, l1hs retrotransposons actively participate in the evolution of the human centromeric alphoid sequences. *Chromosome Res*, 7(4):305–17, 1999.

Hye-Ran Lee, Karen E Hayden, and Huntington F Willard. Organization and molecular evolution of cenp-a–associated satellite dna families in a basal primate genome. *Genome Biol Evol*, 3:1136–49, 2011. doi: 10.1093/gbe/evr083.

J K Lee, J A Huberman, and J Hurwitz. Purification and characterization of a cenp-b homologue protein that binds to the centromeric k-type repeat dna of schizosaccharomyces pombe. *Proc Natl Acad Sci U S A*, 94(16):8427–32, Aug 1997.

A R Lohe, A J Hilliker, and P A Roberts. Mapping simple repeated dna sequences in heterochromatin of drosophila melanogaster. *Genetics*, 134(4):1149–74, Aug 1993.

Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms Mol Biol*, 6:26, Nov 2011. doi: 10.1186/1748-7188-6-26.

J J Maio. Dna strand reassociation and polyribonucleotide binding in the african green monkey, cercopithecus aethiops. *J Mol Biol*, 56(3):579–95, Mar 1971.

J J Maio and C L Schildkraut. Isolated mammalian metaphase chromosomes. ii. fractionated chromosomes of mouse and chinese hamster cells. *J Mol Biol*, 40(2):203–16, Mar 1969.

J J Maio, F L Brown, and P R Musich. Toward a molecular paleontology of primate genomes. i. the hindiii and ecori dimer families of alphoid dnas. *Chromosoma*, 83(1):103–25, 1981.

H S Malik and J J Bayes. Genetic conflicts during meiosis and the evolutionary origins of centromere complexity. *Biochem Soc Trans*, 34(Pt 4):569–73, Aug 2006. doi: 10.1042/BST0340569.

H S Malik and S Henikoff. Adaptive evolution of cid, a centromere-specific histone in drosophila. *Genetics*, 157(3):1293–8, Mar 2001.

Harmit S Malik and Steven Henikoff. Phylogenomics of the nucleosome. *Nat Struct Biol*, 10(11): 882–91, Nov 2003. doi: 10.1038/nsb996.

Harmit S Malik and Steven Henikoff. Major evolutionary transitions in centromere complexity. *Cell*, 138(6):1067–82, Sep 2009. doi: 10.1016/j.cell.2009.08.036.

Kristin A Maloney, Lori L Sullivan, Justyne E Matheny, Erin D Strome, Stephanie L Merrett, Alyssa Ferris, and Beth A Sullivan. Functional epialleles at an endogenous human centromere. *Proc Natl Acad Sci U S A*, 109(34):13704–9, Aug 2012. doi: 10.1073/pnas.1203126109.

L Manuelidis. Chromosomal localization of complex and simple repeated human dnas. *Chromosoma*, 66(1):23–32, Mar 1978a.

L Manuelidis. Complex and simple sequences in human repeated dnas. *Chromosoma*, 66(1):1–21, Mar 1978b.

L Manuelidis and J C Wu. Homology between human and simian repeated dna. *Nature*, 276(5683): 92–4, Nov 1978.

B Marçais, J P Charlieu, B Allain, E Brun, M Bellis, and G Roizès. On the mode of evolution of alpha satellite dna in human populations. *J Mol Evol*, 33(1):42–8, Jul 1991.

Marmoset Genome Sequencing and Analysis Consortium. The common marmoset genome provides insight into primate biology and evolution. *Nat Genet*, 46(8):850–7, Aug 2014. doi: 10.1038/ng.3042.

Owen J Marshall, Anderly C Chueh, Lee H Wong, and K H Andy Choo. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. *Am J Hum Genet*, 82(2):261–82, Feb 2008. doi: 10.1016/j.ajhg.2007.11.009.

H Masumoto, H Masukata, Y Muro, N Nozaki, and T Okazaki. A human centromere antigen (cenp-b) interacts with a short specific sequence in alphoid dna, a human centromeric satellite. *J Cell Biol*, 109(5):1963–73, Nov 1989.

H Masumoto, M Ikeno, M Nakano, T Okazaki, B Grimes, H Cooke, and N Suzuki. Assay of centromere function using a human artificial chromosome. *Chromosoma*, 107(6-7):406–16, Dec 1998.

Lidia Mateo and Josefa González. Pogo-like transposases have been repeatedly domesticated into cenp-b-related proteins. *Genome Biol Evol*, 6(8):2008–16, Jul 2014. doi: 10.1093/gbe/evu153.

Ramsay J McFarlane and Timothy C Humphrey. A role for recombination in centromere function. *Trends Genet*, 26(5):209–13, May 2010. doi: 10.1016/j.tig.2010.02.005.

Kara L McKinley and Iain M Cheeseman. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol*, 17(1):16–29, Jan 2016. doi: 10.1038/nrm.2015.5.

Daniël P Melters, Keith R Bradnam, Hugh A Young, Natalie Telis, Michael R May, J Graham Ruby, Robert Sebra, Paul Peluso, John Eid, David Rank, José Fernando Garcia, Joseph L DeRisi, Timothy Smith, Christian Tobias, Jeffrey Ross-Ibarra, Ian Korf, and Simon W L Chan. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol*, 14(1):R10, Jan 2013. doi: 10.1186/gb-2013-14-1-r10.

Jason Merker, Aaron M Wenger, Tam Sneddon, Megan Grove, Daryl Waggott, Sowmi Utiramerur, Yanli Hou, Christine C Lambert, Kevin S Eng, Luke Hickey, Jonas Korlach, James Ford, and Euan A Ashley. Long-read whole genome sequencing identifies causal structural variation in a mendelian disease. *bioRxiv*, 2016. doi: 10.1101/090985. URL https://www.biorxiv.org/content/early/2016/12/02/090985.

M Meselson, F W Stahl, and J Vinograd. Equilibrium sedimentation of macromolecules in density gradients. *Proc Natl Acad Sci U S A*, 43(7):581–8, Jul 1957.

Matthew D D Miell, Colin J Fuller, Annika Guse, Helena M Barysz, Andrew Downes, Tom Owen-Hughes, Juri Rappsilber, Aaron F Straight, and Robin C Allshire. Cenp-a confers a reduction in

height on octameric nucleosomes. *Nat Struct Mol Biol*, 20(6):763–5, Jun 2013. doi: 10.1038/nsmb.2574.

Karen H Miga. Completing the human genome: the progress and challenge of satellite dna assembly. *Chromosome Res*, 23(3):421–6, Sep 2015. doi: 10.1007/s10577-015-9488-2.

Karen H Miga, Yulia Newton, Miten Jain, Nicolas Altemose, Huntington F Willard, and W James Kent. Centromere reference models for human chromosomes x and y satellite arrays. *Genome Res*, 24(4):697–707, Apr 2014. doi: 10.1101/gr.159624.113.

G L Miklos and B John. Heterochromatin and satellite dna in man: properties and prospects. *Am J Hum Genet*, 31(3):264–80, May 1979.

Joost Monen, Paul S Maddox, Francie Hyndman, Karen Oegema, and Arshad Desai. Differential role of cenp-a in the segregation of holocentric c. elegans chromosomes during meiosis and mitosis. *Nat Cell Biol*, 7(12):1248–55, Dec 2005. doi: 10.1038/ncb1331.

Y Moroi, C Peebles, M J Fritzler, J Steigerwald, and E M Tan. Autoantibody to centromere (kinetochore) in scleroderma sera. *Proc Natl Acad Sci U S A*, 77(3):1627–31, Mar 1980.

S L Mowbray and A Landy. Generation of specific repeated fragments of eukaryote dna. *Proc Natl Acad Sci U S A*, 71(5):1920–4, May 1974.

Sebastian Müller and Geneviève Almouzni. A network of players in h3 histone variant deposition and maintenance at centromeres. *Biochim Biophys Acta*, 1839(3):241–50, Mar 2014. doi: 10.1016/j.bbagrm.2013.11.008.

Y Muro, H Masumoto, K Yoda, N Nozaki, M Ohashi, and T Okazaki. Centromere protein b assembles human centromeric alpha-satellite dna at the 17-bp sequence, cenp-b box. *J Cell Biol*, 116(3):585–96, Feb 1992.

T D Murphy and G H Karpen. Localization of centromere function in a drosophila minichromosome. *Cell*, 82(4):599–609, Aug 1995.

P R Musich, F L Brown, and J J Maio. Highly repetitive component alpha and related alphoid dnas in man and monkeys. *Chromosoma*, 80(3):331–48, 1980.

Kiyotaka Nagaki, Zhukuan Cheng, Shu Ouyang, Paul B Talbert, Mary Kim, Kristine M Jones, Steven Henikoff, C Robin Buell, and Jiming Jiang. Sequencing of a rice centromere uncovers active genes. *Nat Genet*, 36(2):138–45, Feb 2004. doi: 10.1038/ng1289.

Hiromi Nakagawa, Joon-Kyu Lee, Jerard Hurwitz, Robin C Allshire, Jun-Ichi Nakayama, Shiv I S Grewal, Katsunori Tanaka, and Yota Murakami. Fission yeast cenp-b homologs nucleate centromeric heterochromatin by promoting heterochromatin-specific histone tail modifications. *Genes Dev*, 16(14):1766–78, Jul 2002. doi: 10.1101/gad.997702.

Mridula Nambiar and Gerald R Smith. Repression of harmful meiotic recombination in centromeric regions. *Semin Cell Dev Biol*, 54:188–97, Jun 2016. doi: 10.1016/j.semcdb.2016.01.042.

Maria Nattestad, Marley C. Alford, Fritz J. Sedlazeck, and Michael C. Schatz. Splitthreader: Exploration and analysis of rearrangements in cancer genomes. *bioRxiv*, 2016. doi: 10.1101/087981. URL http://www.biorxiv.org/content/early/2016/11/15/087981.

Yael Nechemia-Arbely, Daniele Fachinetti, Karen H Miga, Nikolina Sekulic, Gautam V Soni, Dong Hyun Kim, Adeline K Wong, Ah Young Lee, Kristen Nguyen, Cees Dekker, Bing Ren, Ben E Black, and Don W Cleveland. Human centromeric cenp-a chromatin is a homotypic, octameric nucleosome at all cell cycle points. *J Cell Biol*, 216(3):607–621, Mar 2017. doi: 10.1083/jcb.201608083.

Irina V Nesmelova and Perry B Hackett. Dde transposases: Structural similarity and diversity.

Adv Drug Deliv Rev, 62(12):1187–95, Sep 2010. doi: 10.1016/j.addr.2010.06.006.

Pavel Neumann, Alice Navrátilová, Andrea Koblížková, Eduard Kejnovský, Eva Hřibová, Roman Hobza, Alex Widmer, Jaroslav Doležel, and Jiří Macas. Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob DNA*, 2(1):4, Mar 2011. doi: 10.1186/1759-8753-2-4.

Wei Niu, Zhi John Lu, Mei Zhong, Mihail Sarov, John I Murray, Cathleen M Brdlik, Judith Janette, Chao Chen, Pedro Alves, Elicia Preston, Cindie Slightham, Lixia Jiang, Anthony A Hyman, Stuart K Kim, Robert H Waterston, Mark Gerstein, Michael Snyder, and Valerie Reinke. Diverse transcription factor binding features revealed by genome-wide chip-seq in c. elegans. *Genome Res*, 21(2):245–54, Feb 2011. doi: 10.1101/gr.114587.110.

Jun-ichirou Ohzeki, Megumi Nakano, Teruaki Okada, and Hiroshi Masumoto. Cenp-b box is required for de novo centromere chromatin assembly on human alphoid dna. *J Cell Biol*, 159(5): 765–75, Dec 2002. doi: 10.1083/jcb.200207112.

Teruaki Okada, Jun-ichirou Ohzeki, Megumi Nakano, Kinya Yoda, William R Brinkley, Vladimir Larionov, and Hiroshi Masumoto. Cenp-b controls centromere formation depending on the chromatin context. *Cell*, 131(7):1287–300, Dec 2007. doi: 10.1016/j.cell.2007.10.045.

Yasuhide Okamoto, Megumi Nakano, Jun-ichirou Ohzeki, Vladimir Larionov, and Hiroshi Masumoto. A minimal cenp-a core is required for nucleation and maintenance of a functional human centromere. *EMBO J*, 26(5):1279–91, Mar 2007. doi: 10.1038/sj.emboj.7601584.

Abbas Padeganeh, Joël Ryan, Jacques Boisvert, Anne-Marie Ladouceur, Jonas F Dorn, and Paul S Maddox. Octameric cenp-a nucleosomes are present at human centromeres throughout the cell cycle. *Curr Biol*, 23(9):764–9, May 2013. doi: 10.1016/j.cub.2013.03.037.

D K Palmer and R L Margolis. Kinetochore components recognized by human autoantibodies are present on mononucleosomes. *Mol Cell Biol*, 5(1):173–86, Jan 1985.

D K Palmer, K O'Day, M H Wener, B S Andrews, and R L Margolis. A 17-kd centromere protein (cenp-a) copurifies with nucleosome core particles and with histones. *J Cell Biol*, 104(4):805–15, Apr 1987.

T Palomeque and P Lorite. Satellite dna in insects: a review. *Heredity (Edinb)*, 100(6):564–73, Jun 2008. doi: 10.1038/hdy.2008.24.

F Pardo-Manuel de Villena and C Sapienza. Transmission ratio distortion in offspring of heterozygous female carriers of robertsonian translocations. *Hum Genet*, 108(1):31–6, Jan 2001.

M L Pardue and J G Gall. Chromosomal localization of mouse satellite dna. *Science*, 168(3937): 1356–8, Jun 1970.

Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi-Yang Fritz, Han Cao, Ariella Cohain, Gintaras Deikus, Russell E Durrett, Scott C Blanchard, Roger Altman, Chen-Shan Chin, Yan Guo, Ellen E Paxinos, Jan O Korbel, Robert B Darnell, W Richard McCombie, Pui-Yan Kwok, Christopher E Mason, Eric E Schadt, and Ali Bashir. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods*, 12(8):780–6, Aug 2015. doi: 10.1038/nmeth.3454.

Mark D Pertile, Alison N Graham, K H Andy Choo, and Paul Kalitsis. Rapid evolution of mouse y centromere repeat dna belies recent sequence stability. *Genome Res*, 19(12):2202–13, Dec 2009. doi: 10.1101/gr.092080.109.

L M Pike, A Carlisle, C Newell, S B Hong, and P R Musich. Sequence and evolution of rhesus monkey alphoid dna. *J Mol Evol*, 23(2):127–37, 1986.

Luca Pozzi, Jason A Hodgson, Andrew S Burrell, Kirstin N Sterner, Ryan L Raaum, and Todd R Disotell. Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. *Mol Phylogenet Evol*, 75:165–83, Jun 2014. doi: 10.1016/j.ympev.2014.02.023.

C Prades, A M Laurent, J Puechberty, Y Yurov, and G Roizés. Sine and line within human centromeres. *J Mol Evol*, 42(1):37–43, Jan 1996.

Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, August E. Woerner, Timothy D. O/'Connor, Gabriel Santpere, Alexander Cagan, Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch, Asger Hobolth, Anders E. Halager, Maika Malig, Jessica Hernandez-Rodriguez, Irene Hernando-Herraez, Kay Prufer, Marc Pybus, Laurel Johnstone, Michael Lachmann, Can Alkan, Dorina Twigg, Natalia Petit, Carl Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo, Marc Dabad, Michael L. Wilson, Laurie Stevison, Cristina Camprubi, Tiago Carvalho, Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela Kondova, Ronald E. Bontrop, Anne Pusey, Felix Lankester, John A. Kiyang, Richard A. Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pascal Gagneux, David Comas, Hans Siegismund, Julie Blanc, Lidia Agueda-Calpena, Marta Gut, Lucinda Fulton, Sarah A. Tishkoff, James C. Mullikin, Richard K. Wilson, Ivo G. Gut, Mary Katherine Gonder, Oliver A. Ryder, Beatrice H. Hahn, Arcadi Navarro, Joshua M. Akey, Jaume Bertranpetit, David Reich, Thomas Mailund, Mikkel H. Schierup, Christina Hvilsom, Aida M. Andres, Jeffrey D. Wall, Carlos D. Bustamante, Michael F. Hammer, Evan E. Eichler, and Tomas Marques-Bonet. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 07 2013. URL http://dx.doi.org/10.1038/nature12228.

Ornjira Prakhongcheep, Yuriko Hirai, Toru Hara, Kornsorn Srikulnath, Hirohisa Hirai, and Akihiko Koga. Two types of alpha satellite dna in distinct chromosomal locations in azara's owl monkey. *DNA Res*, 20(3):235–40, Jun 2013. doi: 10.1093/dnares/dst004.

D Quénet, D Sturgill, and Y Dalal. Identifying centromeric rnas involved in histone dynamics in vivo. *Methods Enzymol*, 573:445–66, 2016. doi: 10.1016/bs.mie.2016.01.010.

Delphine Quénet and Yamini Dalal. The cenp-a nucleosome: a dynamic structure and role at the centromere. *Chromosome Res*, 20(5):465–79, Jul 2012. doi: 10.1007/s10577-012-9301-4.

Delphine Quénet and Yamini Dalal. A long non-coding rna is required for targeting centromeric protein a to the human centromere. *Elife*, 3:e03254, Aug 2014. doi: 10.7554/eLife.03254.

E J Richards, H M Goodman, and F M Ausubel. The centromere region of arabidopsis thaliana chromosome 1 contains telomere-similar sequences. *Nucleic Acids Res*, 19(12):3351–7, Jun 1991.

J C Roach, A F Siegel, G van den Engh, B Trask, and L Hood. Gaps in the human genome project. *Nature*, 401(6756):843–5, Oct 1999. doi: 10.1038/44684.

Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. The advantages of smrt sequencing. *Genome Biol*, 14(7):405, Jul 2013. doi: 10.1186/gb-2013-14-6-405.

Gérard Roizès. Human centromeric alphoid domains are periodically homogenized so that they vary substantially between homologues. mechanism and implications for centromere functioning. *Nucleic Acids Res*, 34(6):1912–24, 2006. doi: 10.1093/nar/gkl137.

L Y Romanova, G V Deriagin, T D Mashkova, I G Tumeneva, A R Mushegian, L L Kisselev, and I A Alexandrov. Evidence for selection in evolution of alpha satellite dna: the central role of cenp-b/pj alpha binding region. *J Mol Biol*, 261(3):334–40, Aug 1996. doi: 10.1006/jmbi.1996.0466.

Milena Rondón-Lagos, Ludovica Verdun Di Cantogno, Caterina Marchiò, Nelson Rangel, Cesar Payan-Gomez, Patrizia Gugliotta, Cristina Botta, Gianni Bussolati, Sandra R Ramírez-Clavijo,

Barbara Pasini, and Anna Sapino. Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis. *Mol Cytogenet*, 7(1):8, Jan 2014. doi: 10.1186/1755-8166-7-8.

H Rosenberg, M Singer, and M Rosenberg. Highly reiterated sequences of simiansimiansimian-simiansimian. *Science*, 200(4340):394–402, Apr 1978.

Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A Harte, Steve Heitner, Glenn Hickey, Angie S Hinrichs, Robert Hubley, Donna Karolchik, Katrina Learned, Brian T Lee, Chin H Li, Karen H Miga, Ngan Nguyen, Benedict Paten, Brian J Raney, Arian F A Smit, Matthew L Speir, Ann S Zweig, David Haussler, Robert M Kuhn, and W James Kent. The ucsc genome browser database: 2015 update. *Nucleic Acids Res*, 43(Database issue):D670–81, Jan 2015. doi: 10.1093/nar/gku1177.

Leah F Rosin and Barbara G Mellone. Centromeres drive a hard bargain. *Trends Genet*, 33(2): 101–117, Feb 2017. doi: 10.1016/j.tig.2016.12.001.

M Katharine Rudd and Huntington F Willard. Analysis of the centromeric regions of the human genome assembly. *Trends Genet*, 20(11):529–33, Nov 2004. doi: 10.1016/j.tig.2004.08.008.

M Katharine Rudd, Gregory A Wray, and Huntington F Willard. The evolutionary dynamics of alpha-satellite. *Genome Res*, 16(1):88–96, Jan 2006. doi: 10.1101/gr.3810906.

Alexander Samoshkin, Alexei Arnaoutov, Lars E T Jansen, Ilia Ouspenski, Louis Dye, Tatiana Karpova, James McNally, Mary Dasso, Don W Cleveland, and Alexander Strunnikov. Human condensin function is essential for centromeric chromatin assembly and proper sister kinetochore orientation. *PLoS One*, 4(8):e6831, Aug 2009. doi: 10.1371/journal.pone.0006831.

Luis Sanchez-Pulido, Alison L Pidoux, Chris P Ponting, and Robin C Allshire. Common ancestry of the cenp-a chaperones scm3 and hjurp. *Cell*, 137(7):1173–4, Jun 2009. doi: 10.1016/j.cell.2009.06.010.

Stefano Santaguida and Andrea Musacchio. The life and miracles of kinetochores. *EMBO J*, 28(17): 2511–31, Sep 2009. doi: 10.1038/emboj.2009.173.

C L Schildkraut and J J Maio. Studies on the intranuclear distribution and properties of mouse satellite dna. *Biochim Biophys Acta*, 161(1):76–93, Jun 1968.

Dirk Schindelhauer and Tobias Schwarz. Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite dna array. *Genome Res*, 12(12):1815–26, Dec 2002. doi: 10.1101/gr.451502.

M G Schueler, A W Higgins, M K Rudd, K Gustashaw, and H F Willard. Genomic and genetic definition of a functional human centromere. *Science*, 294(5540):109–15, Oct 2001. doi: 10.1126/science.1065042.

Mary G Schueler and Beth A Sullivan. Structural and functional dynamics of human centromeric chromatin. *Annu Rev Genomics Hum Genet*, 7:301–13, 2006. doi: 10.1146/annurev.genom.7.080505.115613.

Mary G Schueler, John M Dunn, Christine P Bird, Mark T Ross, Luigi Viggiano, NISC Comparative Sequencing Program, Mariano Rocchi, Huntington F Willard, and Eric D Green. Progressive proximal expansion of the primate x chromosome centromere. *Proc Natl Acad Sci U S A*, 102(30): 10563–8, Jul 2005. doi: 10.1073/pnas.0503346102.

Mary G Schueler, Willie Swanson, Pamela J Thomas, NISC Comparative Sequencing Program, and Eric D Green. Adaptive evolution of foundation kinetochore proteins in primates. *Mol Biol Evol*,

27(7):1585–97, Jul 2010. doi: 10.1093/molbev/msq043.

Melina Schuh, Christian F Lehner, and Stefan Heidmann. Incorporation of drosophila cid/cenp-a and cenp-c into centromeres during early embryonic anaphase. *Curr Biol*, 17(3):237–43, Feb 2007. doi: 10.1016/j.cub.2006.11.051.

Juan-Manuel Schvartzman, Rocio Sotillo, and Robert Benezra. Mitotic chromosomal instability and cancer: mouse modelling of the human disease. *Nat Rev Cancer*, 10(2):102–15, Feb 2010. doi: 10.1038/nrc2781.

Volkan Sevim, Ali Bashir, Chen-Shan Chin, and Karen H Miga. Alpha-centauri: assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics*, 32(13):1921–4, Jul 2016. doi: 10.1093/bioinformatics/btw101.

Valery A Shepelev, Alexander A Alexandrov, Yuri B Yurov, and Ivan A Alexandrov. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet*, 5(9):e1000641, Sep 2009. doi: 10.1371/journal.pgen.1000641.

Jinghua Shi, Sarah E Wolf, John M Burke, Gernot G Presting, Jeffrey Ross-Ibarra, and R Kelly Dawe. Widespread gene conversion in centromere cores. *PLoS Biol*, 8(3):e1000327, Mar 2010. doi: 10.1371/journal.pbio.1000327.

Lingling Shi, Yunfei Guo, Chengliang Dong, John Huddleston, Hui Yang, Xiaolu Han, Aisi Fu, Quan Li, Na Li, Siyi Gong, Katherine E Lintner, Qiong Ding, Zou Wang, Jiang Hu, Depeng Wang, Feng Wang, Lin Wang, Gholson J Lyon, Yongtao Guan, Yufeng Shen, Oleg V Evgrafov, James A Knowles, Francoise Thibaud-Nissen, Valerie Schneider, Chack-Yung Yu, Libing Zhou, Evan E Eichler, Kwok-Fai So, and Kai Wang. Long-read sequencing and de novo assembly of a chinese genome. *Nat Commun*, 7:12065, Jun 2016. doi: 10.1038/ncomms12065.

Manjunatha Shivaraju, Raymond Camahort, Mark Mattingly, and Jennifer L Gerton. Scm3 is a centromeric nucleosome assembly factor. *J Biol Chem*, 286(14):12016–23, Apr 2011. doi: 10.1074/jbc.M110.183640.

Gregory E Sims, Se-Ran Jun, Guohong Albert Wu, and Sung-Hou Kim. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proc Natl Acad Sci U S A*, 106(40):17077–82, Oct 2009. doi: 10.1073/pnas.0909377106.

D Singer and L Donehower. Highly repeated dna of the baboon: organization of sequences homologous to highly repeated dna of the african green monkey. *J Mol Biol*, 134(4):835–42, Nov 1979.

Peter J Skene and Steven Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of dna binding sites. *Elife*, 6, Jan 2017. doi: 10.7554/eLife.21856.

R Keith Slotkin. The epigenetic control of the athila family of retrotransposons in arabidopsis. *Epigenetics*, 5(6):483–90, Aug 2010.

G P Smith. Evolution of repeated dna sequences by unequal crossover. *Science*, 191(4227):528–35, Feb 1976.

S Smith and B Stillman. Purification and characterization of caf-i, a human cell factor required for chromatin assembly during dna replication in vitro. *Cell*, 58(1):15–25, Jul 1989.

Martin Šošic and Mile Šikic. Edlib: a c/c ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, May 2017. doi: 10.1093/bioinformatics/btw753.

E M Southern. Base sequence and evolution of guinea-pig alpha-satellite dna. *Nature*, 227(5260):794–8, Aug 1970.

E M Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *J Mol Biol*, 98(3):503–17, Nov 1975a.

E M Southern. Long range periodicities in mouse satellite dna. *J Mol Biol*, 94(1):51–69, May 1975b.

Florian A Steiner and Steven Henikoff. Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. *Elife*, 3:e02025, Jan 2014.

Florian A Steiner and Steven Henikoff. Diversity in the organization of centromeric chromatin. *Curr Opin Genet Dev*, 31:28–35, Apr 2015. doi: 10.1016/j.gde.2015.03.010.

Bianca K Stöcker, Johannes Köster, and Sven Rahmann. Simlord: Simulation of long read data. *Bioinformatics*, 32(17):2704–6, Sep 2016. doi: 10.1093/bioinformatics/btw286.

S Stoler, K C Keith, K E Curnick, and M Fitzgerald-Hayes. A mutation in cse4, an essential gene encoding a novel chromatin-associated protein in yeast, causes chromosome nondisjunction and cell cycle arrest at mitosis. *Genes Dev*, 9(5):573–86, Mar 1995.

T Strachan, E Coen, D Webb, and G Dover. Modes and rates of change of complex dna families of drosophila. *J Mol Biol*, 158(1):37–54, Jun 1982.

T Strachan, D Webb, and G A Dover. Transition stages of molecular drive in multiple-copy dna families in drosophila. *EMBO J*, 4(7):1701–8, Jul 1985.

Penporn Sujiwattanarat, Watcharaporn Thapana, Kornsorn Srikulnath, Yuriko Hirai, Hirohisa Hirai, and Akihiko Koga. Higher-order repeat structure in alpha satellite dna occurs in new world monkeys and is not confined to hominoids. *Sci Rep*, 5:10315, May 2015. doi: 10.1038/srep10315.

B A Sullivan and S Schwartz. Identification of centromeric antigens in dicentric robertsonian translocations: Cenp-c and cenp-e are necessary components of functional centromeres. *Hum Mol Genet*, 4(12):2189–97, Dec 1995.

K F Sullivan and C A Glass. Cenp-b is a highly conserved mammalian centromere protein with homology to the helix-loop-helix family of proteins. *Chromosoma*, 100(6):360–70, Jul 1991.

X Sun, J Wahlstrom, and G Karpen. Molecular structure of a functional drosophila centromere. *Cell*, 91(7):1007–19, Dec 1997.

Xiaoping Sun, Hiep D Le, Janice M Wahlstrom, and Gary H Karpen. Sequence analysis of a functional drosophila centromere. *Genome Res*, 13(2):182–94, Feb 2003. doi: 10.1101/gr.681703.

Aorarat Suntronpong, Kazuto Kugou, Hiroshi Masumoto, Kornsorn Srikulnath, Kazuhiko Ohshima, Hirohisa Hirai, and Akihiko Koga. Cenp-b box, a nucleotide motif involved in centromere formation, occurs in a new world monkey. *Biol Lett*, 12(3):20150817, Mar 2016. doi: 10.1098/rsbl.2015.0817.

Yulia M Suvorova, Maria A Korotkova, and Eugene V Korotkov. Comparative analysis of periodicity search methods in dna sequences. *Comput Biol Chem*, 53 Pt A:43–8, Dec 2014. doi: 10.1016/j.compbiolchem.2014.08.008.

Hideaki Tagami, Dominique Ray-Gallet, Geneviève Almouzni, and Yoshihiro Nakatani. Histone h3.1 and h3.3 complexes mediate nucleosome assembly pathways dependent or independent of dna synthesis. *Cell*, 116(1):51–61, Jan 2004.

K Takahashi, E S Chen, and M Yanagida. Requirement of mis6 centromere connector for localizing a cenp-a-like protein in fission yeast. *Science*, 288(5474):2215–9, Jun 2000.

Paul B Talbert and Steven Henikoff. Histone variants–ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol*, 11(4):264–75, Apr 2010a. doi: 10.1038/nrm2861.

Paul B Talbert and Steven Henikoff. Centromeres convert but don't cross. *PLoS Biol*, 8(3):e1000326, Mar 2010b. doi: 10.1371/journal.pbio.1000326.

Paul B Talbert, Ricardo Masuelli, Anand P Tyagi, Luca Comai, and Steven Henikoff. Centromeric localization and adaptive evolution of an arabidopsis histone h3 variant. *Plant Cell*, 14(5):1053–66, May 2002.

Paul B Talbert, Terri D Bryson, and Steven Henikoff. Adaptive evolution of centromere proteins in plants and animals. *J Biol*, 3(4):18, 2004. doi: 10.1186/jbiol11.

Y Tanaka, O Nureki, H Kurumizaka, S Fukai, S Kawaguchi, M Ikuta, J Iwahara, T Okazaki, and S Yokoyama. Crystal structure of the cenp-b protein-dna complex: the dna-binding domains of cenp-b induce kinks in the cenp-b box dna. *EMBO J*, 20(23):6612–8, Dec 2001. doi: 10.1093/emboj/20.23.6612.

Shoko Terada, Yuriko Hirai, Hirohisa Hirai, and Akihiko Koga. Higher-order repeat structure in alpha satellite dna is an attribute of hominoids rather than hominids. *J Hum Genet*, 58(11):752–4, Nov 2013. doi: 10.1038/jhg.2013.87.

Jitendra Thakur and Steven Henikoff. Cenpt bridges adjacent cenpa nucleosomes on young human α-satellite dimers. *Genome Res*, 26(9):1178–87, Sep 2016. doi: 10.1101/gr.204784.116.

Jitendra Thakur, Paul B Talbert, and Steven Henikoff. Inner kinetochore protein interactions with regional centromeres of fission yeast. *Genetics*, 201(2):543–61, Oct 2015. doi: 10.1534/genetics.115.179788.

Watcharaporn Thapana, Penporn Sujiwattanarat, Kornsorn Srikulnath, Hirohisa Hirai, and Akihiko Koga. Reduction in the structural instability of cloned eukaryotic tandem-repeat dna by low-temperature culturing of host bacteria. *Genet Res (Camb)*, 96:e13, Oct 2014. doi: 10.1017/S0016672314000172.

R E Thayer, M F Singer, and T F McCutchan. Sequence relationships between single repeat units of highly reiterated african green monkey dna. *Nucleic Acids Res*, 9(1):169–81, Jan 1981.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 10 2015. URL http://dx.doi.org/10.1038/nature15393.

David T Ting, Doron Lipson, Suchismita Paul, Brian W Brannigan, Sara Akhavanfard, Erik J Coffman, Gianmarco Contino, Vikram Deshpande, A John Iafrate, Stan Letovsky, Miguel N Rivera, Nabeel Bardeesy, Shyamala Maheswaran, and Daniel A Haber. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science*, 331(6017):593–6, Feb 2011. doi: 10.1126/science.1200801.

Todd J Treangen and Steven L Salzberg. Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, 13(1):36–46, Nov 2011. doi: 10.1038/nrg3117.

Amy Tsurumi and Willis X Li. Global heterochromatin loss: a unifying theory of aging? *Epigenetics*, 7(7):680–8, Jul 2012. doi: 10.4161/epi.20540.

M Tudor, M Lobocka, M Goodell, J Pettitt, and K O'Hare. The pogo transposable element family of drosophila melanogaster. *Mol Gen Genet*, 232(1):126–34, Mar 1992.

Lara A Underkoffler, Laura E Mitchell, Zaki S Abdulali, Joelle N Collins, and Rebecca J Oakey. Transmission ratio distortion in offspring of mouse heterozygous carriers of a (7.18) robertsonian translocation. *Genetics*, 169(2):843–8, Feb 2005. doi: 10.1534/genetics.104.032755.

J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fa-

sulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–51, Feb 2001. doi: 10.1126/science.1058040.

Nadine Vincenten, Lisa-Marie Kuhl, Isabel Lam, Ashwini Oke, Alastair Rw Kerr, Andreas Hochwagen, Jennifer Fung, Scott Keeney, Gerben Vader, and Adèle L Marston. The kinetochore prevents centromere-proximal crossover recombination during meiosis. *Elife*, 4, Dec 2015. doi: 10.7554/eLife.10850.

P M Walker. Origin of satellite dna. *Nature*, 229(5283):306–8, Jan 1971.

Lily Hui-Ching Wang, Thomas Schwarzbraun, Michael R Speicher, and Erich A Nigg. Persistence of dna threads in human anaphase cells suggests late completion of sister chromatid decatenation. *Chromosoma*, 117(2):123–35, Apr 2008. doi: 10.1007/s00412-007-0131-7.

P E Warburton and H F Willard. Interhomologue sequence variation of alpha satellite dna from human chromosome 17: evidence for concerted evolution along haplotypic lineages. *J Mol Evol*, 41(6):1006–15, Dec 1995.

P E Warburton, J S Waye, and H F Willard. Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin. *Mol Cell Biol*, 13(10):6520–9, Oct 1993.

P E Warburton, T Haaf, J Gosden, D Lawson, and H F Willard. Characterization of a chromosome-specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chro-

mosomes. *Genomics*, 33(2):220–8, Apr 1996. doi: 10.1006/geno.1996.0187.

M Waring and R J Britten. Nucleotide sequence repetition: a rapidly reassociating fraction of mouse dna. *Science*, 154(3750):791–4, Nov 1966.

Wesley C Warren, Anna J Jasinska, Raquel García-Pérez, Hannes Svardal, Chad Tomlinson, Mariano Rocchi, Nicoletta Archidiacono, Oronzo Capozzi, Patrick Minx, Michael J Montague, Kim Kyung, LaDeana W Hillier, Milinn Kremitzki, Tina Graves, Colby Chiang, Jennifer Hughes, Nam Tran, Yu Huang, Vasily Ramensky, Oi-Wa Choi, Yoon J Jung, Christopher A Schmitt, Nikoleta Juretic, Jessica Wasserscheid, Trudy R Turner, Roger W Wiseman, Jennifer J Tuscher, Julie A Karl, Jörn E Schmitz, Roland Zahn, David H O'Connor, Eugene Redmond, Alex Nisbett, Béatrice Jacquelin, Michaela C Müller-Trutwin, Jason M Brenchley, Michel Dione, Martin Antonio, Gary P Schroth, Jay R Kaplan, Matthew J Jorgensen, Gregg W C Thomas, Matthew W Hahn, Brian J Raney, Bronwen Aken, Rishi Nag, Juergen Schmitz, Gennady Churakov, Angela Noll, Roscoe Stanyon, David Webb, Francoise Thibaud-Nissen, Magnus Nordborg, Tomas Marques-Bonet, Ken Dewar, George M Weinstock, Richard K Wilson, and Nelson B Freimer. The genome of the vervet (chlorocebus aethiops sabaeus). *Genome Res*, 25(12):1921–33, Dec 2015. doi: 10.1101/gr.192922.115.

J S Waye and H F Willard. Molecular analysis of a deletion polymorphism in alpha satellite of human chromosome 17: evidence for homologous unequal crossing-over and subsequent fixation. *Nucleic Acids Res*, 14(17):6915–27, Sep 1986a.

J S Waye and H F Willard. Structure, organization, and sequence of alpha satellite dna from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human x chromosome. *Mol Cell Biol*, 6(9):3156–65, Sep 1986b.

J S Waye and H F Willard. Concerted evolution of alpha satellite dna: evidence for species specificity and a general lack of sequence conservation among alphoid sequences of higher primates. *Chromosoma*, 98(4):273–9, Oct 1989.

R Wevrick and H F Willard. Long-range organization of tandem arrays of alpha satellite dna at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. *Proc Natl Acad Sci U S A*, 86(23):9394–8, Dec 1989.

Travis J Wheeler and Sean R Eddy. nhmmer: Dna homology search with profile hmms. *Bioinformatics*, 29(19):2487–9, Oct 2013. doi: 10.1093/bioinformatics/btt403.

Gerhard Wieland, Sandra Orthaus, Sabine Ohndorf, Stephan Diekmann, and Peter Hemmerich. Functional complementation of human centromere protein a (cenp-a) by cse4p from saccharomyces cerevisiae. *Mol Cell Biol*, 24(15):6620–30, Aug 2004. doi: 10.1128/MCB.24.15.6620-6630.2004.

H F Willard. Evolution of alpha satellite. *Curr Opin Genet Dev*, 1(4):509–14, Dec 1991.

H F Willard, K D Smith, and J Sutherland. Isolation and characterization of a major tandem repeat family from the human x chromosome. *Nucleic Acids Res*, 11(7):2017–33, Apr 1983.

H F Willard, J S Waye, M H Skolnick, C E Schwartz, V E Powers, and S B England. Detection of restriction fragment length polymorphisms at the centromeres of human chromosomes by using chromosome-specific alpha satellite dna probes: implications for development of centromere-based genetic linkage maps. *Proc Natl Acad Sci U S A*, 83(15):5611–5, Aug 1986.

H F Willard, R Wevrick, and P E Warburton. Human centromere structure: organization and potential role of alpha satellite dna. *Prog Clin Biol Res*, 318:9–18, 1989.

A K Wong and J B Rattner. Sequence organization and cytological localization of the minor satellite

of mouse. *Nucleic Acids Res*, 16(24):11645–61, Dec 1988.

Cheng Xue, Muthuswamy Raveendran, R Alan Harris, Gloria L Fawcett, Xiaoming Liu, Simon White, Mahmoud Dahdouli, David Rio Deiros, Jennifer E Below, William Salerno, Laura Cox, Guoping Fan, Betsy Ferguson, Julie Horvath, Zach Johnson, Sree Kanthaswamy, H Michael Kubisch, Dahai Liu, Michael Platt, David G Smith, Binghua Sun, Eric J Vallender, Feng Wang, Roger W Wiseman, Rui Chen, Donna M Muzny, Richard A Gibbs, Fuli Yu, and Jeffrey Rogers. The population genomics of rhesus macaques (macaca mulatta) based on whole-genome sequences. *Genome Res*, 26(12):1651–1662, Dec 2016. doi: 10.1101/gr.204255.116.

Huihuang Yan, Paul B Talbert, Hye-Ran Lee, Jamie Jett, Steven Henikoff, Feng Chen, and Jiming Jiang. Intergenic locations of rice centromeric chromatin. *PLoS Biol*, 6(11):e286, Nov 2008. doi: 10.1371/journal.pbio.0060286.

T P Yang, S K Hansen, K K Oishi, O A Ryder, and B A Hamkalo. Characterization of a cloned repetitive dna sequence concentrated on the human x chromosome. *Proc Natl Acad Sci U S A*, 79 (21):6593–7, Nov 1982.

W G Yasmineh and J J Yunis. Satellite dna in mouse autosomal heterochromatin. *Biochem Biophys Res Commun*, 35(6):779–82, Jun 1969.

K Yoda, K Kitagawa, H Masumoto, Y Muro, and T Okazaki. A human centromere protein, cenp-b, has a dna binding domain containing four potential alpha helices at the nh2 terminus, which is separable from dimerizing activity. *J Cell Biol*, 119(6):1413–27, Dec 1992.

K Yoda, T Nakamura, H Masumoto, N Suzuki, K Kitagawa, M Nakano, A Shinjo, and T Okazaki. Centromere protein b of african green monkey cells: gene structure, cellular expression, and centromeric localization. *Mol Cell Biol*, 16(9):5169–77, Sep 1996.

Weiqi Zhang, Jingyi Li, Keiichiro Suzuki, Jing Qu, Ping Wang, Junzhi Zhou, Xiaomeng Liu, Ruotong Ren, Xiuling Xu, Alejandro Ocampo, Tingting Yuan, Jiping Yang, Ying Li, Liang Shi, Dee Guan, Huize Pan, Shunlei Duan, Zhichao Ding, Mo Li, Fei Yi, Ruijun Bai, Yayu Wang, Chang Chen, Fuquan Yang, Xiaoyu Li, Zimei Wang, Emi Aizawa, April Goebl, Rupa Devi Soligalla, Pradeep Reddy, Concepcion Rodriguez Esteban, Fuchou Tang, Guang-Hui Liu, and Juan Carlos Izpisua Belmonte. Aging stem cells. a werner syndrome stem cell model unveils heterochromatin alterations as a driver of human aging. *Science*, 348(6239):1160–3, Jun 2015. doi: 10.1126/science.aaa1356.

Xuming Zhou, Boshi Wang, Qi Pan, Jinbo Zhang, Sudhir Kumar, Xiaoqing Sun, Zhijin Liu, Huijuan Pan, Yu Lin, Guangjian Liu, Wei Zhan, Mingzhou Li, Baoping Ren, Xingyong Ma, Hang Ruan, Chen Cheng, Dawei Wang, Fanglei Shi, Yuanyuan Hui, Yujing Tao, Chenglin Zhang, Pingfen Zhu, Zuofu Xiang, Wenkai Jiang, Jiang Chang, Hailong Wang, Zhisheng Cao, Zhi Jiang, Baoguo Li, Guang Yang, Christian Roos, Paul A Garber, Michael W Bruford, Ruiqiang Li, and Ming Li. Whole-genome sequencing of the snub-nosed monkey provides insights into folivory and evolutionary history. *Nat Genet*, 46(12):1303–10, Dec 2014. doi: 10.1038/ng.3137.

Zheng Zhou, Hanqiao Feng, Bing-Rui Zhou, Rodolfo Ghirlando, Kaifeng Hu, Adam Zwolak, Lisa M Miller Jenkins, Hua Xiao, Nico Tjandra, Carl Wu, and Yawen Bai. Structural basis for recognition of centromere histone variant cenh3 by the chaperone scm3. *Nature*, 472(7342):234–7, Apr 2011. doi: 10.1038/nature09854.

Quan Zhu, Gerald M Pao, Alexis M Huynh, Hoonkyo Suh, Nina Tonnu, Petra M Nederlof, Fred H Gage, and Inder M Verma. Brca1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature*, 477(7363):179–84, Sep 2011. doi: 10.1038/nature10371.

Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, Elizabeth Henaff, Alexa B R McIntyre, Dhruva Chandramohan, Feng Chen, Erich Jaeger, Ali Moshrefi, Khoa Pham, William Stedman, Tiffany Liang, Michael Saghbini, Zeljko Dzakula, Alex Hastie, Han Cao, Gintaras Deikus, Eric Schadt, Robert Sebra, Ali Bashir, Rebecca M Truty, Christopher C Chang, Natali Gulbahce, Keyan Zhao, Srinka Ghosh, Fiona Hyland, Yutao Fu, Mark Chaisson, Chunlin Xiao, Jonathan Trow, Stephen T Sherry, Alexander W Zaranek, Madeleine Ball, Jason Bobe, Preston Estep, George M Church, Patrick Marks, Sofia Kyriazopoulou-Panagiotopoulou, Grace X Y Zheng, Michael Schnall-Levin, Heather S Ordonez, Patrice A Mudivarti, Kristina Giorda, Ying Sheng, Karoline Bjarnesdatter Rypdal, and Marc Salit. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*, 3:160025, Jun 2016. doi: 10.1038/sdata.2016.25.

Appendix A

ANALYSIS OF GENETIC VARIATION IN HUMAN CENTROMERES

This Appendix provides the methodological details for the work presented in **Chapter 3**.

A.1 Datasets

Pacific Bioscence Single-Molecule (SM), Real-Time sequencing datasets used in this study were from the following NIH SRA projects: PRJNA246220 and PRJNA253496 (CHM1)11, PRJNA269593 (CHM13), PRJNA200694 (Ashkenazim Trio), PRJNA288807 (Yoruban), PRJNA200694 and PRJEB7353 (GM12878), PRJNA301527 (Han Chinese individual). GM12878 Oxford Nanopore data are available from a public release by Oxford Nanopore, Inc. (https://nanoporetech.com/publications/na12878-human-reference-oxford-nanopore-minion).

A.2 Identification of SM reads containing repeat monomers

Reads containing alphoid satellite (AS) repeat monomers were identified using profile hidden Markov model searching with HMMSEARCH (Wheeler and Eddy, 2013). HMMs used for read annotation were produced using seed alignments from DFAM(Hubley et al., 2016) (accession DF0000029) and HMMSEARCH was run with default parameters using the forward and reverse complements of DFAM seed alignments. Alternatively, BLAST searching against a sequence or database containing sequences of interest can be used. In cases requiring de novo discovery of tandemly repeated sequences, the input sequence database for defining repeat monomers can be derived using a tandem repeats detection approach such as Tandem Repeats Finder. Note that the periodicity detection and read segmentation approaches are independent of precise identification of repeat monomers, which was performed only to define a set of reads containing the repeats of interest and for visualization of reads.

A.3 ASTRL: Characterization of array periodicities

For a read s of length L, I define a segment to be periodic if, for $x \in [0, L)$, positive integer n and an error parameter ϵ, $s(x) = s(nx \pm \epsilon)$. Conventional periodicity/frequency detection approaches such as Fourier- and wavelet-based methods are not robust to indels without additional computationally expensive steps such as dynamic time warping. Given the large number of reads ($> 10^6$) and their relatively long lengths (10^4), I used a fast, indel-resilient suffix array-based periodicity detection algorithm. In outline, for each read, a suffix array S was computed using the SAIS linear-time induced sorting algorithm and repeating k-mers were identified using binary search over the suffix array. Periodicities in $[1, L/2)$, *i.e.*, the less than the Nyquist limit, were subsequently analyzed. Given that indels cause variation in the distance d between matches that is non-linear in d, we used a random walk approximation of the error parameter ϵ described by Benson (1999) ($\epsilon = 2.3\sqrt{p_i \cdot d}$ for indel probability p_i). A two-step approach was employed to detect and refine periodicities. In the first step, all occurrences of repeating k-mers under the edit distance were discovered, generating a set of, typically, tens of thousands of periodicities for each read. In the final step, candidate periods were validated by computing the edit distance. This step relied on the Edlib library for fast sequence alignment (Šošic and Šikic, 2017).

A.4 ASTRL: Read segmentation

To account for the possibility that a read could contain arrays with differing characteristics (including non-periodic components), reads were segmented based on periodicity. For each read, we created a two-dimensional position-periodicity representation akin to a time-frequency spectrogram generated from short-time Fourier transformation. For a read of length L and maximum periodicity of interest P_{max}, we defined a $L \times P_{max}$ matrix S in which the number of times a single-base position $x \in [0, L)$ is in a segment with periodicity $p \leq P_{max}$ is stored at $S(x, p)$. Subharmonic summation was used to emphasize fundamental periodicities and suppress noise inherent in periodicity detection by exploiting the expectation for 'ringing' at higher harmonics of fundamental periodicities. In this formulation, $\alpha \in (0, 1]$ is a decay factor, which weights the harmonics as a geometric progression and effectively suppresses detection of integer multiples of the fundamental period. Maxima in S correspond to the optimal segmentation. A fifth order median filter was

used to smooth the signal and minimize point discrepancies. Finally, change points (transitions from one periodicity to another) were naively defined to occur at a position x_i if $|x_{i+1} - x_i| \geq t$ for an empirically determined threshold value t.

A.5 False discovery rate (FDR) estimation and validation of approach

For each dataset, FDR estimates were derived using the Benjamini-Hochberg procedure from a corresponding set of randomly permuted reads. Simulated PacBio reads from previously characterized alphoid arrays were used to validate the periodicity detection and read segmentation approach. Simulated reads were produced using SimLoRD (Stöcker et al., 2016).

Appendix B

ANALYSIS OF CENTROMERIC SATELLITE IN PRIMATES

This Appendix describes the methodology employed for characterization of variation in primate centromeric satellites described in **Chapter 4**.

B.1 Datasets

B.1 lists the NCBI Sequence Read Archive (SRA) accession numbers for the publicly available whole genome sequencing datasets used in the comparative analysis of primate alpha satellite. **??** lists the SRA accessions for ssDNA-seq, PIP-seq, and ChIP-seq datasets that were analyzed.

B.2 Definition of putative centromeric satellites

Putative centromeric satellites were defined using raw whole-genome Sanger sequencing data and assembled contigs from primate genome sequencing projects. Satellite units of length 1-1,000 bp were identified using Tandem Repeats Finder (TRF) (Benson, 1999) similar to Melters et al. (2013). TRF was run with the following parameters `2 7 7 80 10 50 1000 -h -ngs` (corresponding to: match weight and mismatch and indel penalties of 2, 7, and 7, respectively; match and indel probabilities of 80 and 10, respectively; minimum reportable alignment score of 50; maximum period of 1,000 bp; with HTML output suppressed and compact 'ngs' output). Lengths of putative centromeric repeats were defined based on peak locations in repeat length histograms and sequences corresponding to these peaks were recovered from the TRF output and deduplicated; common 'background' sequences such as Alu and LINE elements were filtered out at this stage. Simian sequences generally aligned to published AS consensus sequences and, to facilitate downstream analyses, were shifted to occupy a common register. Simian sequences were sufficiently divergent that this was not performed. Reads and contigs were then re-scanned using BLASTN (parameters: `-task dc-megablast -outfmt 6`) to recover in-register monomeric units that were used in subsequent analyses.

Species	SRA Accession(s)	Reference(s)
Homo sapiens	SRR794330, SRR794336	The 1000 Genomes Project Consortium (2015)
Gorilla gorilla	SRR747963, SRR747964	Prado-Martinez et al. (2013)
Gorilla berengei	SRR747651, SRR747652	Prado-Martinez et al. (2013)
Pan troglodytes	SRR726233, SRR726241	Prado-Martinez et al. (2013)
Pan paniscus	SRR726612, SRR726613, SRR726614	Prado-Martinez et al. (2013)
Pongo pygmaeus	SRR747999, SRR748001	Prado-Martinez et al. (2013)
Pongo abelii	SRR748024, SRR748025	Prado-Martinez et al. (2013)
Hylobates pileatus	SRR1391006, SRR1391007	Carbone et al. (2014)
Hylobates moloch	SRR1390731-SRR1390766,SRR1522073,SRR1522074	Carbone et al. (2014)
Nomascus leucogenys	SRR1425539-SRR1425591	Carbone et al. (2014)
Symphalangus syndactylus	SRR1390940-SRR1391005	Carbone et al. (2014)
Macaca mulatta	SRR1952214, SRR1952216	Xue et al. (2016)
Macaca fascicularis	SRR445659,SRR445662,SRR445666	Ref. genome (PRJNA20409)
Papio anubis	SRR927654, SRR927659	Ref. genome (PRJNA169345)
Chlorocebus aethiops	SRR1660260, SRR1660261, SRR1660288	Warren et al. (2015)
Chlorocebus pygerythrus	SRR556129, SRR556168	Warren et al. (2015)
Rhinopithecus roxellana	SRR1040959, SRR1040960	Zhou et al. (2014)
Callithrix jacchus	SRR1282348, SRR1282349	Marmoset Genome Sequencing and Analysis Consortium (2014)
Cebus capucinus	SRR3136953	Ref. genome (PRJNA328123)
Saimiri boliviensis	SRR315549, SRR315555	Ref. genome (PRJNA169636)
Aotus nancymaae	SRR1692993, SRR1692996	Ref. genome (PRJNA282745)
Mus musculus	SRR091273, SRR091274	Ref. genome (PRJNA60381)

Table B.1 | Primate short read whole-genome sequencing datasets. Paired-end (~100x100 bp) Illumina datasets used in this study. Datasets without associated publications were the source data for publicly available RefSeq genome assemblies; the NCBI BioProject numbers are provided in place of a reference.

B.3 Centromeric satellite HMMs

HMMs specific to hominoids, OWMs, and NWMs were produced by combining monomers from Sanger reads and contigs from the respective species and randomly sampling 1000 for multiple alignment using MUSCLE (with default parameters). These multiple alignments were used to construct HMMs using HMMBUILD (with default parameters). A similar procedure was performed with the prosimian sequences, except HMMs constructed were species-specific.

In addition to these HMMs, the three α-satellite DFAM HMMs (accession numbers DF000029, DF000014, DF000015) were also used to identify AS units in Illumina and PacBio datasets.

B.4 Data processing and identification of satellite monomers in Illumina and PacBio datasets

Illumina sequencing reads were pre-processed (adapter trimming and quality filtering) using bbduk (with default parameters). Alphoid reads were identified via homology searching using HMMSEARCH (with default parameters). Only illumina reads containing matches with length $>= 85$ bp were retained and, in instances where only one end of a paired-end read was flagged as a match, both ends were retained for subsequent analyses.

B.5 Alignment-free comparison of AS monomers

Given the success of alignment-free comparison of whole-genomes using feature frequency profiles (Sims et al., 2009), k-mer decomposition with $k = 5$ was used to cluster 10,000 randomly chosen Illumina reads from each species. Note that sequences were reverse complemented as needed to ensure all sequences were on the same strand relative to the corresponding HMM consensus sequence. For each sequence, a 1024-dimensional feature vector was constructed and used to store 5-mer counts. Prior to principal components analysis (PCA), the collection of feature vectors was transformed such that counts for a given 5-mer across the dataset had zero mean and unit standard deviation. PCA was performed using scikit-learn, keeping the first 25 components.

B.6 Detection of dyad symmetries

In order to efficiently detect dyad symmetries in large databases of sequences (short reads from Illumina sequencing and longer reads from Sanger sequencing), an approach based on suffix arrays

was used. Each sequence of interest was concatenated with its reverse complement and the suffix and longest common prefix (LCP) arrays were constructed. Regions exhibiting dyad symmetry should occupy adjacent positions in the suffix array, which maintains the suffixes of the sequence and its reverse complement in lexicographic order. The LCP array provides a fast way to determine the number of identical positions in the prefixes of adjacent suffixes. Therefore, traversal of the suffix and LCP arrays allows the rapid identification of dyad symmetries.

B.7 DNA secondary structure prediction

The ViennaRNA package v2.2 (Lorenz et al., 2011) was used to predict the minimum free energies of secondary structure formation in satellite sequences. `RNAfold` was used with the following parameters for DNA secondary structure prediction: `'–noGU –noconv –noPS –paramFile=dna_mathews2004.par'`. Folding predictions were performed on random samples of 10^5 HMM-defined monomers (Sanger data) or 10^4 reads containing HMM matches (Illumina data).