Comparative Analysis of Tandem Repeats from Eukaryotic Genomes: Insight in Centromere Evolution

by

DANIËL PATRICK MELTERS
B.S. (Leiden University) 2003
M.S. (Leiden University) 2006

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry, Molecular, Cellular, and Developmental Biology
with a Designated Emphasis in Biotechnology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Ian F. Korf, PhD, Chair

_____

Charles H. Langley, PhD

_____

Scott C. Dawson, PhD

Committee in Charge

2013

Chapters

in loving memory of
my sister Eline
and
my mentor Simon

1. Abstract

Centromeres are the chromosomal loci where microtubule spindles bind, via the kinetochore, during mitosis and meiosis. Paradoxically the centromere, as a functional unit, is essential to guarantee faithful chromosome segregation, whereas its underlying DNA sequences and associated kinetochore proteins are fast evolving. In most animals and plants that have been studied, centromeres contain megabase-scale arrays of tandem repeats. In spite of their importance, very little is known about the degree to which centromeric tandem repeats share common properties between different species across different phyla. We used bioinformatic methods to identify high-copy tandem repeats from species using publicly available genomic sequence and our own data. We found that despite an overall lack of sequence conservation, centromeric tandem repeats from diverse species showed similar modes of evolution. Furthermore, phylogenetic analysis of sequence homology showed little evidence of sequence conservation beyond approximately 50 million years of divergence. In addition, we performed a survey of fungi genomes for the presence of high-copy tandem repeats, but found little evidence to suggest that high-copy centromeric repeats are a common feature in fungi, with the possible exception of the *Zygomycota* phylum. Finally, in most species the kinetochore assembles at a single locus, but in some cases the kinetochore forms along the entire length of the chromosomes forming holocentric chromosomes. Following a literature review we estimate that holocentricity is rather common and has evolved at least thirteen times.

2. Acknowledgements

First of all, I would like to dedicate my dissertation to my sister Eline (1986-2009) and my mentor Simon Chan (1974-2012), both who died much too early. Eline, as a sister is, challenged me when ever she could and she kept me grounded. I am missing her lively demeanor and vocal expressions. I cannot express enough how important Simon's mentorship was for my professional development. I wish I could have shared my current academic and professional success as well as future success with both of them.

The last five years have been an amazing experience thanks to all the people around me, both academically and personally. First and foremost, I would like to thank my mentors Ian Korf and Simon Chan. Without their guidance, understanding, patience, and most importantly, their friendship during my graduate studies at UC Davis, I would not have grown into the scientist I have become. Our weekly meetings were paramount in providing a well rounded experience to further my career as well as having a good laugh. They encouraged me to not only grow as a biologist and bioinformatician, but also as a mentor and independent thinker. The independence they gave me was both challenging as well as refreshing. Simon's late emails containing corrections to our papers and Ian's down-to-earth comments will be greatly missed.

While thanking Ian and Simon, I also have to thank Keith Bradnam. The one person who would not resist the temptation to correct my abstracts or papers or posters or oral presentations with an eye for the smallest of details. These valuable lessons came besides his contagious passion for programming and teaching me to program in Perl with a seasoning of UNIX.

I would like to thank the many lab members of both the Korf and Chan labs. Thank you Genis, Artem, Shahram, Ken, Yen, Ravi, Keith D, Kristen, Matt, Abby, Danielle and Paul. Thank you Ravi, Mohan, Shamoni, Han, Patrick, Joel, Natalie, Tina, Christian, Brenda, Maryam, Pak, and of course my fellow graduate student in Simon's lab: Joe. Your support and jovial conversation in the lab as well as constructive (and sometimes destructive) comments during lab meetings were a joy and made my graduate career a pleasurable experience. I would also like to thank the Comai and Britt labs for the joint lab meetings with the Chan lab every Monday afternoon.

I would also like to thank the DEB, especially Judy Kjelstrom, Denneal Jamison-McClung, and Marianne Hunter. Not only were they instrumental in my academic development, providing funding opportunities, and helping me find a great internship position at Genentech, they also created a warm environment where I felt at home.

Of course, I would like to thank my fellow graduate students for sharing the last five years.

I would also like to thank my dissertation committee members Scott Dawson and Chuck Langley for their advice and support.

Finally I would like to give a big thank you to my wife Anna and my daughter Sarah. Their unconditional love for and patients with me were pivotal during my graduate training, as I doubt I was always the most stress-free person to be around with. Where Sarah's smiles and continuing development keep me grounded to life and see life in whole new dimension. Anna, you kept me going when times got tough and it got pretty tough a few times in the last five years. I also like to thank my parents, Ruud and Jettie and my sister Nadia, Arnold, Thomas, Dinand and Sander for allowing me to be far away from home to do what I love to do while providing emotional

support. I also like to thank my best friends Peter and Mark and my family-in-law Juliana, Jaime, Tommy, Mechas, Jacobo, and Elias.

Thank you everyone who I forgot to mention for all your encouragement and support.

This dissertation was made possibly by funding from the training grant T32-GM008799 from NIH/NIGM and by the Howard Hughes Medical Institute and the the Gordon and Betty Moore Foundation (GBMF3068). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

3. Introduction

In eukaryotes the transmission of chromosomes from mother cell to the two daughter cells depend on the interaction between the microtubule spindle apparatus and each individual chromosome. The DNA underlying the attachment site is called the centromere. The interaction between the DNA and the spindles is facilitated by the kinetochore, which is comprised of hundreds of proteins (Perpelescu, & Fukagawa 2011). Besides controlling the physical separation of chromosomes (Rago, & Cheeseman 2013), the kinetochore also holds a central role in spindle checkpoint signaling pathways (Straight 1997; Lampson, & Cheeseman 2011; Lara-Gonzalez et al 2012). Although the centromere as a functional unit is essential for proper chromosome segregation, the sequences that underly the spindle attachment loci differ dramatically between divergent species. Not only does centromere DNA evolve fast, also kinetochore proteins evolve faster than expected for essential genes. This paradoxical observation (Henikoff et al 2001) remains largely unexplained.

The centromere is epigenetically defined by the replacement of the regular histone H3 with the centromere specific H3 variant cenH3 (CENP-A in humans (*Homo sapiens*), Cid in fruit-flies (*Drosophila melanogaster*), Cse4p in budding yeast *Saccharomyces cerevisiae*, Cnp1 in fission yeast *Schizosaccharomyces pombe*, HTR12 in *Arabidopsis thaliana,* and HCP-3 in *Caenorhabditis. elegans*)[1]. *De novo* incorporation of cenH3 at non-centromere loci is sufficient

[1] Recently a dispute has erupted as to which name to use for the centromere-specific histone H3 variant protein: cenH3 (Talbert et al 2012; Talbert, & Henikoff 2013) or CENP-A (Earnshaw et al 2013; Earnshaw, & Cleveland 2013). In short, the former name, cenH3, is based on the phylogenetic relationship between the various histone variants including cenH3, whereas the latter name, CENP-A, is based on historical precedent of naming proteins based on the order by which they were discovered. In this dissertation I will use cenH3 to be consistent with the names used in my publications (Chapter 4 (Melters et al 2012) and Chapter 5 (Fitzgerald-Hayes et al 1982)).

to form a functional centromere (Mendiburo et al 2011), whereas deletion of either cenH3 (Ravi, & Chan 2010) or the centromeric chromatin (Shang et al 2010) are lethal. Nevertheless, centromere DNA itself is neither sufficient nor essential (Karpen, & Allshire 1997; Sullivan, & Karpen 2001; Henikoff et al 2001).

Although there appears to be a preference for tandem repeats at the centromeres, it is unknown what drives this observed preference. Fission yeast (*Schizosaccharomyces pombe*) centromeres provide some clues, even when the structure of fission yeast centromeres (Clarke et al 1986; Steiner et al 1993) is markedly different from that of primates or grasses. Fission yeast centromeres have a 2-5 kbp core of unique DNA where about five cenH3 molecules define the functional centromere (Joglekar et al 2008). At either end of the functional core lay a mixture of inner and outer repeats (Figure 3-1). These repeats are heterochromatic, similar to the heterochromatic pericentromeres of plant and animal centromeres. This heterochromatic state is nucleated and maintained by the RNAi machinery (Grewal 2010). In addition, in fruitflies (*Drosophila melanogaster*) overexpressed cenH3 prefers to nucleate functional centromeres at heterochromatin boundaries (Olszak et al 2011). It is also believed that the pericentric heterochromatin represses meiotic recombination in fission yeast (Ellermeier et al 2010). Whereas the potential role of heterochromatin in centromere nucleation is not fully elucidated, the role of heterochromatin in sister chromatid cohesion at the centromere in fission yeast (Bernard et al 2001; Nonaka et al 2002) of vertebrates (Bernard, & Allshire 2002; Fukagawa et al 2004) is well established. The necessity of heterochromatic pericentric for centromere function can be established if neocentromeres display heterochromatic marks. Interestingly, the presence of (pericentric) heterochromatin markers, such as HP1, at neocentromeres was only slightly enriched (Saffery et al 2003). Although this region represents a relatively small domain of

heterochromatin, it is important to note that the chromatin immunoprecipitation study was a comparative one, only measuring the levels of enrichment of HP1. Thus, it is possible that an extant domain of heterochromatin already existed at this region, which was augmented for neocentromere formation. Whether or not RNAi is required to form pericentric heterochromatin which allows the nucleation and maintenance of a functional centromere remains to be determined.

Although the pericentromere is heterochromatic, the centromere core where cenH3 is found, is transcriptionally active. On various chromosomes in rice (*Oryza sativa*) genes intersperse the centromere tandem repeat arrays and are actively transcribed at levels consistent with genes located in euchromatin regions (Nagaki et al 2004; Wu et al 2004; Yan et al 2008). Neocentromeres in humans display similar behavior by genes localized in these neocentromeres (Marshall et al 2008; Marshall, & Choo 2009). In addition, transcription of the pericentric repeats in fission yeast are required for heterochromatin formation and centromere stability (Grewal 2010). The most obvious example of transcriptionally active centromeric chromatin are the chromosomes of the holocentric nematode *Caenorhabditis elegans*. CenH3-containing nucleosomes occupies about half the potential sites in the *C. elegans* genome. The sites where *C. elegans* cenH3 was not found coincides with genes required during germline development and the germline marks H3K36 methylation and Argonaute CSR-1 22G-RNA (Gassmann et al 2012). Therefore, the authors propose that H3K36 methylation and Argonaute CSR-1, which binds to short 22G-RNAs derived from germline transcripts, as a candidate mechanism for transmitting memory of germline transcription to early embryos (Claycomb et al 2009; Rechtsteiner et al 2010). One notable species where ncRNA has been implicated in centromere function is the tammar wallaby (*Macropus eugenii*) (reviewed in (O'Neill, & Carone 2009)). Furthermore, RNA

has been associated with chromosomal passenger complex, regulation of pericentromeric chromatin modifications, and cenH3 recruitment to the centromere (reviewed in (Gent, & Dawe 2012)). Thus, not only does transcription occur at the centromere, it is an important feature of a functional centromere. How these transcript regulate centromere function is subject of intense investigation.

Centromere DNA

Although hundreds of eukaryotic genomes have been sequenced (http://www.ncbi.nlm.nih.gov/genome/browse/) and are at some stage in their assembly process, centromere DNA sequences often don't make it into the assemble and therefore remain uncharacterized. Nevertheless, one of the first DNA sequences to be sequenced were the highly repetitive sequences because they were fairly easy to isolate. This even led to the fact that the centromere repeat of each individual human chromosome was published as an individual paper (Sullivan et al 1996). When genome project became commonplace at the beginning of the twenty-first century, the cloning and PCR based methods used to identify centromere DNA fell out of favor. Nevertheless, for many species the centromere DNA was determined, including the panda (*Ailuropoda melanoleuca*) (Wu et al 1990), honey bee (*Apis mellifera*) (Tarès et al 1993; Beye, & Moritz 1995), sheep (*Ovis aries*) (Buckland 1983; Novak 1984), to name a few.

Centromere DNA for most plant and animal genomes analyzed to date are characterized by the presence of a high-copy tandem repeat array (Figure 3-1; extensive analysis in Chapter 5). Human (*Homo sapiens*) centromeres consist of a 171-bp tandem repeat (Sullivan et al 1996),

historically called alpha satellites (Figure 3-1), which form a large array of tandem repeat organized in a head-to-tail fashion (Sullivan et al 1996). Although the size of the array differs between chromosomes, each centromere consists of a similar 171-bp repeat unit. The centromere core is characterized by a higher order repeat unit and these higher-order repeat structure are almost chromosome specific (Rudd et al 2006). The higher order repeat unit consists of between three and seventeen 171-bp repeat unit, which are than again repeated. The size of the human centromere tandem repeat is in the same range as that of the model plant *Arabidopsis thaliana*, where the repeat is 178-bp in length (Heslop-Harrison et al 1999; Hall et al 2003; Zhang et al 2008). In contrast, the centromere DNA of fruit-flies consists of a pentamer (Sun et al 2003). Also, the centromere repeat unit of the Japanese pufferfish (*Takifugu rubripes*) is with 118-bp (Roest Crollius et al 2000) much shorter than the human centromere repeat. Although the Japanese pufferfish repeat is similar in length to the green spotted pufferfish (*Tetraodon nigroviridis*) centromere repeat there is no sequence similarity (Fischer et al 2000). These two pufferfish species diverged about 70 million years ago. Yet, if two species are very closely related, for instance chimpanzees (*Pan troglodytes*) and humans (Haaf, & Willard 1997; Samonte et al 1997; Alkan et al 2007) or *A. thaliana* and *A. lyrata* (Hall et al 2003; Tsukahara et al 2012) or zebras (*Equus quagga*) and horses (*Equus caballus*) (Wade et al 2009; Piras et al 2010), sequence similarity between the centromere repeats was observed.

The centromere can also harbor centromere-specific transposable elements which are often of the LTR retrotranposon family (Du et al 2010). Examples are CRM in maize (*Zea mays*; Figure 3-1) (Zhong et al 2002; Sharma, & Presting 2008), CRR in rice (Cheng et al 2002; Nagaki et al 2005; Sharma, & Presting 2008) and *Tal1* in *A. lyrata (*Tsukahara et al 2012). Interestingly, the *Tal1* was introduced into the *A. thaliana* genome by transformation and it specifically incorporates

itself in the *A. thaliana* centromere (Tsukahara et al 2012). It is not known what makes a specific transposable element centromere specific.

In the fast majority of species studied the most abundant tandem repeat is also the centromere tandem repeat, but there are a few exceptions. In mice (*Mus musculus*) there are two very abundant tandem repeat and both localize to the primary constriction of the acrocentric chromosomes.The less abundant one of the two colocalizes with the functional centromere (Guenatri et al 2004; Kuznetsova et al 2006). Also in maize there are two highly abundant tandem repeats, but it is the 155-bp repeat that is actually centromeric, whereas the 180-bp repeat localizes to heterochromatic knobs (Dawe et al 2009). The cucumber genome has three highly abundant tandem repeat families, but the most abundant one does not localize to the centromere, rather it localizes to the subtelomeres (Zhao et al 2011). The second most abundant one localizes to the centromeres (Zhao et al 2011), including to a recently formed neocentromere (Han et al 2009; Koo et al 2010).

In contrast, centromere DNA in fungi look very different from that in multicellular plant or animal species. In *Saccharomyces cerevisiae* and closely related yeast species centromere DNA spans a short stretch, ranging between 109 and 203-bp (Figure 3-1; Fitzgerald-Hayes et al 1982; Meraldi et al 2006). Because of the small size of these centromeres they are named “point’ centromeres. It is also the only known centromere where proteins bind in a sequence-specific manner (McAinsh et al 2003). In the filamentous pathogenic yeast *Candida albicans* cenH3 localizes to unique sequences of 3 kbp (Sanyal et al 2004; Baum et al 2006), whereas the centromere core of fission yeast spans between 2-5 kbp of unique sequences (Clarke et al 1986;

Steiner et al 1993; Takayama et al 2008; Song et al 2008). For few other fungi has the centromere DNA been characterized.

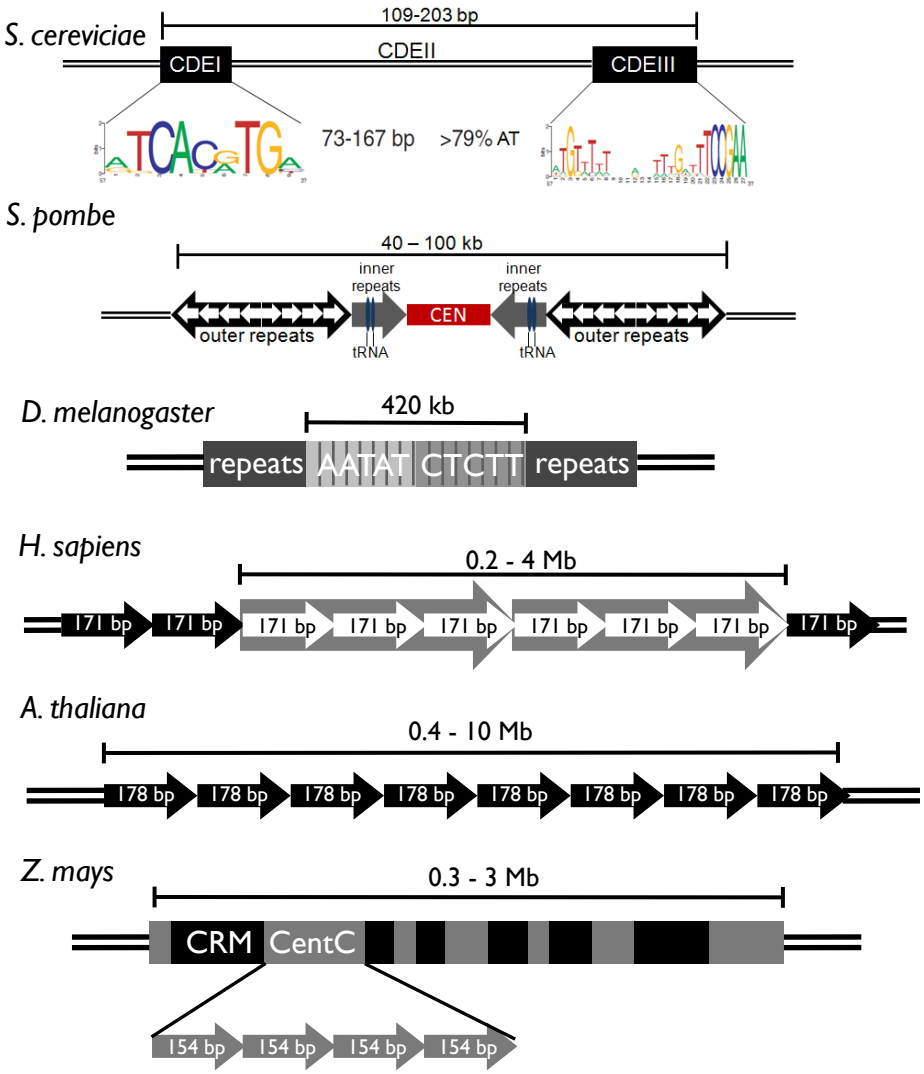


Figure 3-1. Six examples of well defined centromere DNA from model organisms. The centromere of *Saccharomyces sereviciae* (budding yeast) consists of three distinct parts: CDEI and CDEIII are bound by proteins in a sequence-dependent manner, whereas CDEII is AT rich. This centromere harbors a single cenH3 molecule The centromere of *Schizosaccharomyces pombe* (fission yeast) has a 2-5 kb core of unique DNA sequence that harbors 5 cenH3 molecules, but the heterochromatic pericentromere is comprised of a variety of inner and outer tandem repeats, whose transcription maintains its heterochromatic state. The centromere of Drosophila melanogaster (fruitlfy) consists of two types of pentamers, which are organized in head-to-tail tandem repeat arrays. The centromere of Homo sapiens (human) consists of a 171-bp tandem repeat unit that spans between 0.2 and 4 Mbp. The functional core is organized in higher order repeats, where between three and seventeen 171-bp repeat unit form a higher tandemly repeated unit. The centromere of Arabidopsis thaliana consists of a 178-bp tandem repeat unit, organized in tandem repeat arrays as large as 10 Mbp, harboring on average 375 cenH3 molecules per chromosome. The centromere of Zea mays (corn) consist of arrays of 154-bp tandem repeat unit (centc) interspersed by centromere-specific transposable elements (CRM).

The presence of unique sequences at the locus where cenH3 incorporates is not unique to these fungi or to neocentromeres (Marshall et al 2008). Chromosome 11 in horse (Equus caballus) (Piras et al 2009; Wade et al 2009) and chromosome 13 in orangutan (Pongo abelii and P. pygmaeus) (Locke et al 2011) lack tandem repeats as well. In these two cases recent evolutionary centromere repositioning events can explain the lack of tandem repeats. In contrast, only 13 of the 78 chicken (Gallus gallus) chromosomes consists of the 42-bp tandem repeat (Shang et al 2010), whereas all the other centromeres comprise of unique sequences, including the sex chromosome Z. The smallest of the centromeres with unique sequences is about 30 kbp, about the same size as the smallest neocentromere in humans (Marshall et al 2008). When this stretch of unique sequence is conditionally removed, the DT-40 chicken cell line dies (Shang et al 2010). These three cases show that in animals unique sequence can stably maintain a functional centromere.

Another interesting case if the pea (Pisum sativum). Each pea chromosome has a primary constriction that harbors between three and five cenH3 foci and the underlying DNA consists of a variety of different tandem repeats (Neumann et al 2012). These sequences differ in both distribution pattern, repeat length, and sequence composition. Each primary constriction can contain more than one variety of tandem repeats (Neumann et al 2012). It is not known if these repeat varieties are mixed or each variety consist of an uninterrupted array.

Tandem Repeat Evolution

Where genes have a fairly well defined beginning (start codon) and end (stop codon), a single tandem repeat unit lacks an obvious beginning or end. An additional complicating factor is that each individual repeat is not identical but similar to any another randomly selected repeat from

the same array. If two alpha satellite sequences were randomly selected from the human genome, they can differ up to 40% in sequence composition (Waye, & Willard 1987; Willard, & Waye 1987; Rudd, & Willard 2004). Even though it is relatively easy to identify a candidate centromere tandem repeat (see Chapter 5), in part because of their sheer abundance (2.6% in the human genome is alpha satellite DNA compared to 1.5% protein-coding genes (Levy et al 2007)), the technical difficulties resulting from sequence dissimilarity make it almost impossible to assemble tandem repeat into a contig. This results in megebase-sized gaps in any genome assembly (Henikoff 2002; Eichler et al 2004). Previously, this meant that to sequence an entire array of high-copy tandem repeats, unique primers had to be designed to be able to sequence through an array (Warburton et al 1991; Warburton, & Willard 1992; Warburton, & Willard 1995). Recently, Hayden *et al* (Hayden et al 2013a; Hayden et al 2013b) developed software (linearSat: http://github.com/JimKent/linearSat) to assemble existing centromere gaps in the human reference genome. It will be interesting to learn if this approach will also result in assembled of non-human centromeres.

Centromere DNA had been shown to vary in size and proportions of higher order repeat sequence variants within the human population (Wevrick, & Willard 1989; Warburton, & Willard 1992). Hayden and others (Hayden et al 2013a) compared fully assembled centromere regions of the human X and Y chromosomes from fourteen distinct human populations (372 individuals in total). Their findings confirmed previous report for Y chromosome repeats with distinct differences between Asian and European populations (Oakey, & Tyler-Smith 1990). As the human Y chromosome does not recombine, very low rate of sequence exchange was expected. Indeed, the sequences show evolutionary marks consistent with homogenization through either conversion or unequal crossing-over (Warburton, & Willard 1995). On the other hand, the human

X chromosome does recombine with its homologous X chromosome in females and thus may be influenced by molecular drive (Dover 1982; Ohta, & Dover 1984). Yet, Hayden *et al* (Hayden et al 2013a) only found evidence for homogenization, not sequence exchange. Analyzing the human autosomes in a similar way would be insightful in how the human centromere evolved at the genome level. By comparing this to other primates and non-primates, the general patterns of high-copy tandem repeat array evolution will become more clear.

There are two predominant hypotheses to explain the evolution of centromere tandem repeats: female centromere meiotic drive (Henikoff et al 2001) and the "library" hypothesis (Plohl et al 2008). These two hypotheses are non-mutually exclusive.

The female centromere meiotic drive hypothesis is based on the co-evolution between centromere DNA and cenH3 and other kinetochore proteins (Henikoff et al 2001). In female meiosis only one of the potential four oocytes matures, whereas in male meiosis all four spermatocytes are formed. This means that when an allele is able to distort the classic Mendelian heredity ratio in favor of the driving allele, we have female meiotic drive (Sandler, & Novitski 1957). The best studied example of centromere meiotic drive is the heterochromatic knob in maize (Dawe et al 2009; Kanizay et al 2013). Although the knob is able to physically interact with the microtubule spindles in and thereby favor its inclusion in the oocyte, there is no indication cenH3 is involved in this interaction (Dawe et al 1999).

In monkeyflowers (*Mimulus guttatus*), a species with both male and female meiosis, homozygosity for a driving centromeric locus led to positive selection (Fishman, & Willis 2005), while at the same time reducing pollen (male gametes) counts (Fishman, & Saunders 2008). The trade-off between meiotic drive in one sex versus reduced fitness for the other sex ultimately will

result in an unstable equilibrium. This problem does not seem to exist in species where only male or female meiosis happens. Fungi only have male (symmetric) meiosis and no evidence of position selection of cenH3 was observed (Bensasson et al 2008). Also, in the ciliated protozoan *Tetrahymena*, which only has female (asymmetric) meiosis, no positive selection was observed (Elde et al 2011). Why centromere meiotic drive seems unsuppressed in *Tetrahymena* is unknown.

The 'library' hypothesis states that every given genome has a variety of tandem repeat arrays. Through stochastic expansion and shrinkage, the relative abundance of each array changes between closely related species (Plohl et al 2008). Molecular drive (Dover 1982; Ohta, & Dover 1984) promotes ultimate fixation of tandem repeat arrays, but (centromere) meiotic drive cannot be excluded. The largest array is anticipated to be associated with the functional centromere. By which mechanisms the tandem repeat arrays expand or shrink and how they do homogenize remains to be experimentally shown. Nevertheless, computational modeling suggests that unequal crossing-over alone can account for array expansion and shrinkage (Smith 1976), whereas gene conversion is thought to be the mechanism for sequence homogenization (Shi et al 2010; Brown et al 2011). A predicted consequence of gene conversion is that the centromere core is more homogenized than the more proximal repeats (Schueler et al 2005). It has also been postulated that replication slippage (Ruggiero, & Topal 2004), rolling replication (Lenzmeier, & Freudenreich 2003), and replication fork collapse (Houseley, & Tollervey 2011) might contribute to tandem repeat array expansion, although no experimental evidence exist for such mechanisms of expansion (or shrinkage) for centromere tandem repeat arrays. Furthermore, *S. cerevisiae* centromeres accumulate DNA mutations at a rate three times higher than that of other non-selected genomic regions (Bensasson 2011). An observation in line with the mutation rate

observed in *Candida* centromeres (Padmanabhan et al 2008). There is no indication to assume that the mutation rate at centromere loci in other species differs dramatically from that observed in *S. cerevisiae* and *Candida* species. The effects of the stepwise mutation model from population genetics (Kimura, & Ohta 1978) has not been studied yet with respect to centromere DNA, in part due to the difficulty of assembling megabase-sized tandem repeat arrays. It would therefore be interesting to model the effect of increased mutation rates on creating novel repeat variants. In Chapter 5, I will discuss several examples where indel mutations and repeat duplication affect centromere DNA evolution.

Although the majority of species studied have monocentric chromosomes it is not yet know how common holocentric chromosomes are. The nematode *C. elegans* is the best studied species with holocentric chromosomes and cenH3 incorporation displays inverse relationship with germline genes (Gassmann et al 2012). In Chapter 4 we discuss how to discriminate holocentric chromosomes from monocentric chromosomes. With these discriminating metrics, a literature review was conducted to identify all described species with holocentric chromosomes. The identified species were analyzed for phylogenetic relationships. Using parsimonious inference, we determined how frequently holocentricity has evolved.

Conventional methods to identify centromeric tandem repeat are labor intensive and thus difficult to conduct on a large scale. In Chapter 5 a newly developed bioinformatics pipeline is described to quantify the most abundant tandem repeats from 282 animal and plant genomes. This method can utilize whole genome shotgun sequence data from various sequencing platforms. Obtained candidate centromere repeat sequences were characterized and analyzed. In Chapter 6 the

bioinformatics pipeline from Chapter 6 was used to identify, characterize, and analyze 94 fungi genomes. Finally, Chapter 7 is a general discussion on how the results from Chapter 4-6 effect the current views on how centromere DNA and centromere morphology evolves. Chapter 8 contains the primary conclusions from this dissertation.

4. Holocentric Chromosomes: Convergent Evolution, Meiotic Adaptations and Genomic Analysis

Daniël P. Melters[1,2,5], Leocadia V. Paliulis[3], Ian F. Korf[1] and Simon W. L. Chan[2,4]

1. Department of Molecular and Cell Biology and Genome Center, University of California, Davis, Davis, CA
2. Department of Plant Biology, University of California, Davis, Davis, CA
3. Biology Department, Bucknell University, Lewisburg, PA 17837
4. Howard Hughes Medical Institute
5. Corresponding author: dpmelters@me.com

Abstract

In most eukaryotes, the kinetochore protein complex assembles at a single locus termed the centromere to attach chromosomes to spindle microtubules. Holocentric chromosomes have the unusual property of attaching to spindle microtubules along their entire length. Our mechanistic understanding of holocentric chromosome function is derived largely from studies in the nematode *Caenorhabditis elegans*, but holocentric chromosomes are found over a broad range of animal and plant species. In this review, we describe how holocentricity may be identified through cytological and molecular methods. By surveying the diversity of organisms with holocentric chromosomes, we estimate that the trait has arisen at least thirteen independent times (four times in plants, and at least nine times in animals). Holocentric chromosomes have inherent problems in meiosis because bivalents can attach to spindles in a random fashion. Interestingly, there are several solutions that have evolved to allow accurate meiotic segregation of holocentric chromosomes. Lastly, we describe how extensive genome sequencing and experiments in non-model organisms may allow holocentric chromosomes to shed light on general principles of chromosome segregation.

"*Darwin answers that we must look for imperfections and oddities, because any perfection in organic design or ecology obliterates the paths of history and might have been created as we find it*." Stephen J. Gould (1986)

Introduction

The centromere is the chromosomal locus bound by kinetochore proteins that connect eukaryotic chromosomes to spindle microtubules during cell division. In most eukaryotes, kinetochore proteins assemble at a single location per chromosome. In these “monocentric” chromosomes, the centromere is visible as the primary constriction in large metaphase chromosomes. In select taxa, kinetochore proteins bind along the entire length of the chromosomes and microtubules can attach along most of the poleward facing surface (Dernburg 2001; Maddox et al 2004; Guerra et al 2010). First described in cytogenetic experiments dating from 1935 (Schrader 1935), these "holocentric" chromosomes have also been called diffuse-kinetochore chromosomes, holokinetic chromosomes, and polykinetic chromosomes. These differences in nomenclature may reflect the difficulty of distinguishing chromosomes with evenly distributed kinetochore proteins from chromosomes that contain numerous but discrete microtubule-binding sites (White 1973). For the rest of this review, we will use the most common term, holocentric chromosomes. Although holocentric chromosomes are found in a minority of eukaryotes, their prevalence may be underestimated. Many species are difficult to study cytologically. In addition, there are a large number of uncharacterized insect and nematode species whose phylogenetic position suggest that they should have holocentric centromeres.

The nematode *Caenorhabditis elegans* is by far the most well-studied holocentric organism. The function of kinetochore proteins in this organism has been reviewed elsewhere (Dernburg 2001; Maddox et al 2004). Almost nothing is known about the biology of other holocentric species. The primary goal of this review is to survey the evolution of holocentric chromosomes throughout eukaryotes, highlighting the fact that this property has arisen many independent times. A crucial step in adapting to the holocentric habit is alterations in meiosis, and we describe

how the fundamental incompatibility between distributed microtubule-binding sites and crossing-over can be resolved. Lastly, we show how widespread genome sequencing can shed light on the function of holocentric chromosomes.

Identification of holocentric chromosomes

How can we identify and confirm holocentricity? Ideally, several criteria should be met. In species with large chromosomes, cytogenetic methods are valuable and are applicable to any organism in which individual mitotic or meiotic chromosomes can be observed. First, all chromosomes in a mitotic metaphase spread must lack a primary constriction (a classic hallmark of (sub-)metacentric chromosomes) (Figure 1A). Second, during mitotic anaphase the sister chromatids must migrate in parallel to the spindle poles, in contrast to monocentric species in which pulling forces are exerted on a single chromosomal point and chromosome arms trail behind (Figure 1B). These two criteria are historically the most common methods used to identify holocentric species.

Some species are amenable to detailed cytogenetic manipulations that can diagnose holocentricity more definitively. If a holocentric chromosome is fragmented, each individual fragment retains centromere activity and can segregate to the poles (Figure 1C). Chromosomes can be broken by X-ray irradiation (Hughes-Schrader, & Schrader 1961), or more precisely by laser-dissection (Fuková et al 2007). Chromosomal fragments must persist after breakage, and micronuclei resulting from a failure to segregate chromosome fragments should not be observed. One problem of chromosome breakage techniques is that DNA damage response pathways can induce cell-cycle checkpoints or apoptosis, disallowing further studies of such cells. Laser dissection is restricted to a desired cell cycle phase (breaks are usually induced during

metaphase) which may be beneficial if checkpoint activiation or apoptosis are cell-cycle stage dependent. In select experimental systems such as grasshopper spermatocytes, individual chromosomes can be manipulated using microdissection tips, and this could serve as an ideal test of holocentricity (Paliulis, & Nicklas 2004). In principle, the ends of mitotic metaphase chromosomes should move freely in monocentric chromsomes, whereas they should be resistant to such physical stimulus in holocentric chromosomes (Doan, & Paliulis 2009) (Figure 1D).

A more precise method to identify holocentric chromosomes is is to visualize kinetochore proteins by immunofluorescence microscopy. Holocentric chromosomes will have kinetochore proteins bound along most if not the entire length of a metaphase chromosome, whereas monocentric chromosomes will have a single discrete focus of localization. Kinetochore proteins such as the centromere-specific histone variant cenH3 (CENP-A in human), or members of the NDC80 microtubule-binding complex can be found from sequence analysis of most eukaryote genomes, even though many kinetochore proteins evolve too quickly for their orthologs to be detected by protein sequence similarity (Talbert et al 2004, Meraldi et al 2006). As sequencing becomes cheaper, kinetochore proteins will be identified from more putative holocentric organisms and this approach is likely to become more widespread. Once a reference genome is available for a given organism, chromatin immunoprecipitation following by sequencing (ChIP-seq) can reveal whether kinetochore proteins are bound to a single centromere or to multiple locations on each chromosome. This method has recently been pioneered for holocentric organisms in *C. elegans* (Gassmann et al 2012). As it is independent of cytology, ChIP-seq will be an especially useful method in the many organisms whose chromosomes are too small for reliable cytogenetic assays (many deeply diverged single-celled eukaryotes fall into this category).

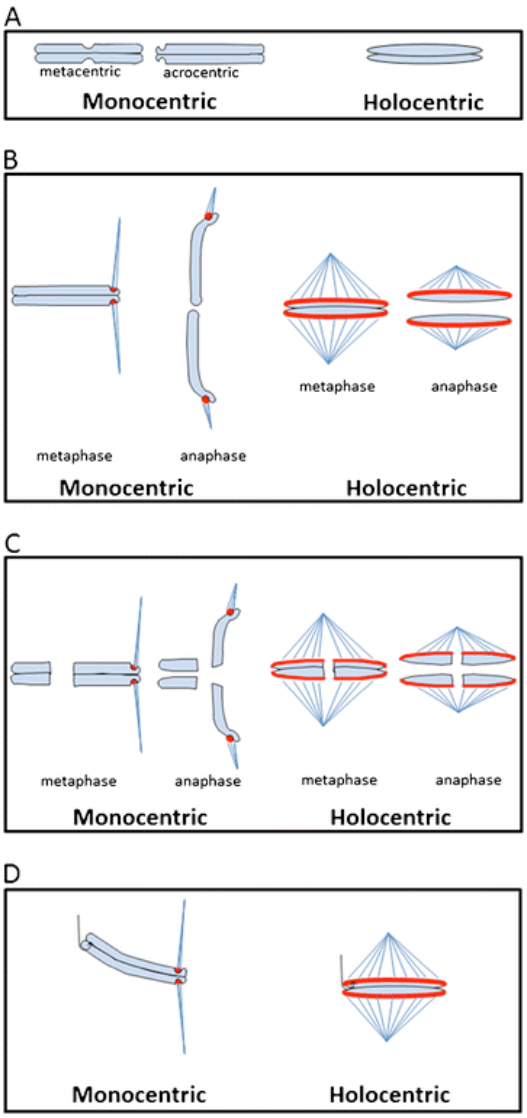


Figure 4-1 How to diagnose holocentricity. **a** Monocentric chromosomes have a primary constriction. In metacentric chromosomes, the primary constriction is approximately equidistant from both chromosomal ends, whereas the primary constriction on acrocentric chromosomes is close to one of the ends. In contrast, holoncentric chromosomes lack a primary constriction. **b** During anaphase, monocentric chromosome are pulled by spindle microtubules from a single point where the kinetochore is assembled. In holocentric chromosomes, the kinetochores are assembled along the length of the chromosome, and during anaphase, the two sister chromatids maintain their parallel relationship. **c** Upon chromosome fragmention, only two fragments of a monocentric chromosome will retain kinetochore function and segregate to the spindle poles. In holocentric chromosomes, all fragments migrate to the poles. **d** Micromanipulation of monocentric chromosome ends during metaphase results in the movement of just the chromosome end. In contrast, in holocentric chromosomes, where spindle microtubles display pulling forces over the length chromosome, the entire chromosome will be perturbed

Broad phylogenetic distribution of holocentric chromosomes

Holocentric chromosomes have radically different patterns of kinetochore protein deposition compared to monocentric chromosomes, and require substantial changes in chromosome behavior to ensure accurate meiosis. Given these facts, it is interesting to ask how often this unusual chromosome structure has arisen during eukaryotic evolution. Most holocentric organisms were identified using cytology before molecular methods became available, and only a small subset of these studies used the more stringent method of chromosome fragmentation. As mentioned above, cytogenetic identification of holocentricity is much easier for large chromosomes. In all, 768 species have been reported to have holocentric chromosomes (including 472 insects, 228 plants, 50 arachnids, and 18 nematodes) (Supplementary Table 1). There are several cases in which reports of holocentric chromosomes have been later corrected by more careful cytological studies, such as the moss *Pleurozium schreberi* and the marine alga *Spirogyra* (Godward 1954; Vaarama 1954; Mughal, & Godward 1973; Kuta et al 1998; Kuta et al 2000). To reduce the possibility of false positives, we focus on well-substantiated cases in this review, especially those in clades with more than one holocentric species. As molecular methods are applied to eukaryotes with small chromosomes, many additional holocentric clades may be discovered. We cannot be sure whether the last common ancestor of all eukaryotes had monocentric or holocentric chromosomes. However, we infer by parsimony that monocentricity was the ancestral state. The sporadic and phylogenetically widespread occurrence of holocentricity in the tree of life suggests that the habit evolved from monocentric chromosomes at least 13 independent times (Figure 2). We summarize the evolution of holocentric chromosomes in plants and animals in separate sections below.

Holocentric chromosomes evolved at least 4 independent times in plants

All known holocentric plant-species belong to the flowering plants (phylum *Angiosperma*) and include both monocots and eudicots (Figure 2). The holocentric monocots predominantly belong to the rush grasses (family *Juncaceae*) and sedges (family *Cyperaceae*) (Luceño et al 1998; Kuta et al 2004). These include the snowy woodrush *Luzula nivea* (*Juncaceae*), the most well studied holocentric plant. Not all genera in Cyperaceae and Juncaceae are holocentric. For example localized centromeres were reported in the aquatic grass genus *Scirpus* (Nijalingappa 1974). Flowering perennial herbs from the genus *Chionographis* (family *Melanthiaceae*) (Tanaka, & Tanaka 1977) also contain holocentric chromosomes.

Holocentric eudicots are limited to two genera; *Drosera* or sundews (family *Droseraceae*) and *Cuscuta* or dodders (family *Convulvulaceae*) (oddly, both are parasitic in nature). *Cuscuta* contains three subgenera, *Cuscuta* subgenus *Cuscuta*, *C*. subgenus *Grammica*, and *C*. subgenus *Monogyna*. Only the subgenus *Cuscuta* and one species in the subgenus *Grammica* are holocentric, offering an attractive opportunity for comparative genomics. (Pazy, & Plitmann 1994; Sheikh, & Kondo 1995; Guerra et al 2010).

Holocentric chromosomes are likely to have evolved at least 9 independent times in animals

In the animal kingdom, holocentric chromosomes have been found in two phyla: *Nematoda* and *Arthropoda*. We estimate that holocentric chromosomes arose once in the *Nematoda*, and 8 times in *Arthropoda*.

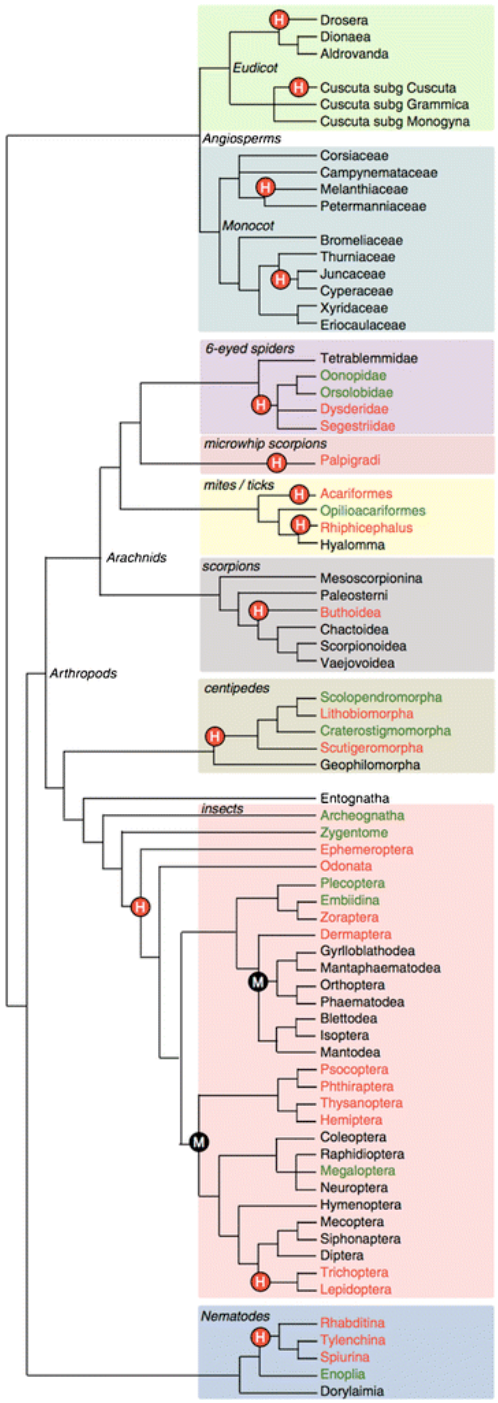


Figure 4-2 Holocentricity has evolved at least 13 times in animals and plants. A depicition of the phylogenetic relationship of species with reported holocentric chromosomes and their close relatives with monocentric chromosomes are shown (Bremer 2002; Grimaldi and Engel 2005; Mitreva et al. 2005; De Ley 2006; Bradley et al. 2009; Regier et al. 2010; Maddison). The most parsimonious estimates are that holocentricity has evolved 13 times (*red circles*) and monocentricity has evolved form holocentric ancestor two times (*black circles*). The clades that consist of species with holocentric chromosomes are depicted by *red names*, whereas clades that consist of species with monocentric chromosomes are depicted by *black names*. The*green names* represent clades for which no karyotype data exists

The most well known group of holocentric species can be found in the Secernentea class of the nematodes, which includes *C. elegans*. Other nematodes are usually described as holocentric because of their phylogenetic relationship to *C. elegans* rather than because of karyotypic evidence. The parasitic roundworms *Trichinella* and *Trichuris* (order *Trichurida*) (Mutafova et al 1982; Spakulová et al 1994), have been described as having monocentric chromosomes, whereas conflicting experimental data is available for *Onchocerca volvulus*, the causative agent of river blindness (Procunier, & Hirai 1986; Hirai et al 1987; Post 2005)

Holocentric chromosomes are found in many clades within the phylum *Arthropoda* (invertebrate animals with an exoskeleton). Notably, true bugs (*Hemiptera*) include the first well-characterized holocentric chromosomes. Diffuse binding of spindle microtubules along chromosomes was noted in as early 1935, and mitotic segregation of chromosome fragments was used to confirm holocentricity in scale insects soon afterward (Schrader 1935; Hughes-Schrader, & Ris 1941; Hughes-Schrader, & Schrader 1961). We have estimated a pattern for how the distribution of holocentric arthropods could have arisen from a monocentric ancestor, inferring by parsimony that the ancestor of all arthropods was monocentric. (Figure 2).

Holocentricity in the class *Insecta* (insects) is relatively common, being found in mayflies (order *Ephemeroptera*), dragonflies (order *Odonata*), angel insects (order *Zoraptera*), earwigs (order *Dermaptera*), caddisflies (order *Trichoptera*), moths and butterflies (order *Lepidoptera*), and the superorder *Paraneoptera* (encompassing lice, thrips and true bugs). Large clades of monocentric insects are nested between these orders (it should be noted that insect phylogenetic relationships are still debated) (Grimaldi, & Engel 2005; Regier et al 2010). Therefore, we propose that the ancestor of most insect orders was holocentric, and that monocentricity returned twice during

evolution of modern insects (Figure 2). A second instance of holocentric chromosome evolution occurred during the divergence of *Trichoptera* and *Lepidoptera* from a monocentric ancestor. An isolated report of a monocentric hemipteran insect may represent a further reversion to the ancestral monocentric habit (Desai 1969; Desai, & Deshpande 1969).

In contrast to *Insecta*, the class *Arachnida* (scorpions, spiders, mites and ticks) has few families or subfamilies with holocentric chromosomes. Holocentric arachnids include spiders (order *Araneae*), microwhip scorpions (order *Palpigradi*), isolated mites and ticks (genuses *Prostigmata* and *Radiicephalidae*), and primitive scorpions (order *Scorpiones*). Holocentricity is also found in centipedes (class *Chilopoda*, closely related to arachnids). Karyotypic studies of mites and ticks suggest that many species may be monocentric, so we assume that holocentric chromosomes evolved twice in this order (Oliver 1972; Oliver et al 1974; Oliver 1977). It is reasonable to assume that holocentric chromosomes arose at least 6 independent times during arthropod evolution, and at least 9 times overall in animals (Figure 2).

Holocentric chromosomes face a kinetochore geometry problem in meiosis

The aim of meiosis is to reduce the chromosome number so haploid gametes are produced from a diploid parent cell. Reduction of chromosome number in meiosis happens because a single round of DNA replication is followed by two rounds of cell division. Correct chromosome segregation in meiosis requires changes in kinetochore geometry and differences in release of sister chromatid cohesion relative to mitosis. Holocentric chromosomes create unique problems during meiosis that organisms with monocentric chromosomes do not face. We review several diverse mechanisms that have arisen in holocentric organisms to allow correct distribution of chromosomes during meiosis.

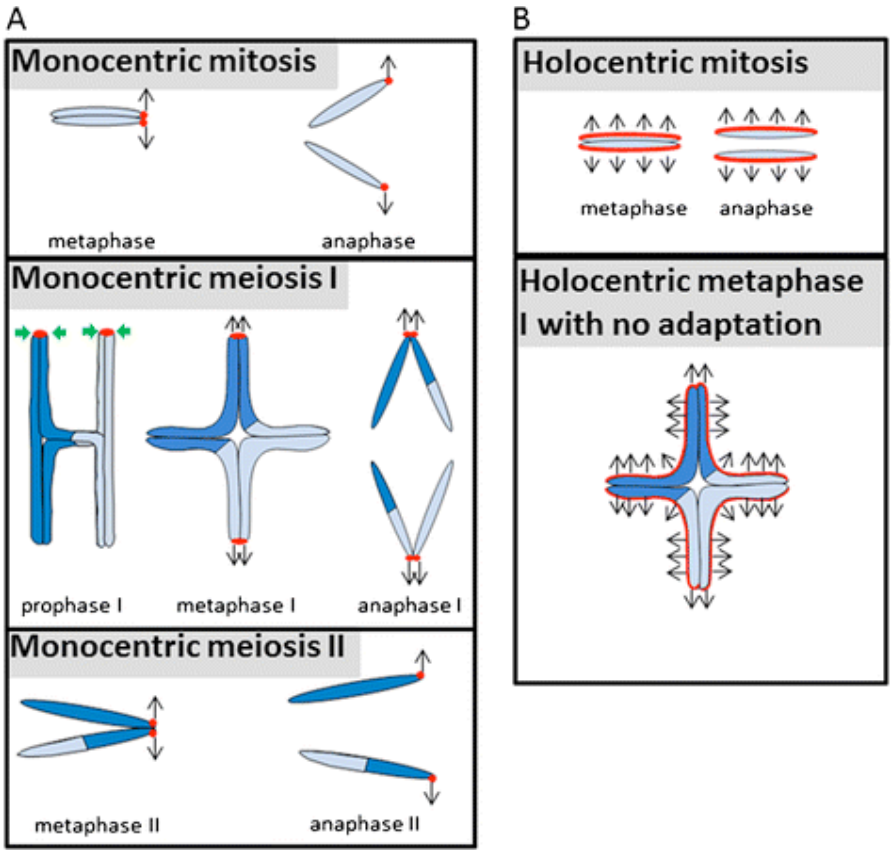


Figure 4-3 Mitosis and meiosis with monocentric and holocentric chromosomes. **a** In mitosis, sister kinetochores of monocentric chromosomes face opposite spindle poles. This leads to separation of sister chromatids in anaphase of mitosis. In meiosis I, sister kinetochores are fused and face in the same direction. This leads to homologous-chromosome separation in anaphase I. In meiosis II, the sister kinetochores face in opposite directions in metaphase, and sister chromatids separate in anaphase II. **b** In mitosis, kinetochores are distributed along the length of the holocentric chromosome arm, so chromosomes connect to the spindle all along their entire length. In anaphase, because the spindle-attachment sites face in opposite directions, sister chromatids separate from one another. In meiosis I, if there is no alteration of chromosome structure or attachment surface position, attachment sites can face in all directions, leading to problems in chromosome segregation

The way a chromosome divides is based on its geometry. In mitosis, both monocentric and holocentric chromosomes have kinetochores whose spindle-microtubule capture surfaces face in opposite directions, so sister chromatids are separated at anaphase (Paliulis, & Nicklas 2004; John 1990a). Although mitosis in holocentric chromosomes is straightforward, there is potential for major problems in meiosis. In prophase I, recombination and sister chromatid cohesion link homologous chromosomes together as a bivalent (Figure 3A). Bivalents with a single chiasma

(recombination site) are cruciform (Figure 3A). In monocentric chromosomes, sister kinetochores fuse and face in the same direction in meiosis I. In anaphase I, cohesion between sister-chromatid arms is released while centromeric cohesion is protected, so sister chromatids remain together and homologues are separated. In meiosis II, sister kinetochores face opposite spindle poles, so sister chromatids separate in anaphase II when centromeric cohesion is released (Paliulis, & Nicklas 2005; John 1990b). Holocentric chromosomes can theoretically attach to the meiosis I spindle at many positions along their length. Therefore, if a holocentric bivalent has no modification to its chromosome structure or kinetochore positioning, its microtubule capture surfaces will face in all directions (Figure 3B). Depending on how cohesion is released, chromosomes could segregate randomly or not at all. Obviously, holocentric organisms require special adaptations to allow correct segregation of one chromatid to each gamete.

Holocentric meiosis in *Caenorhabditis elegans*

The nematode *C. elegans* is the holocentric organism in which meiosis has been best studied. *C. elegans* bivalents tend to have a single chiasma placed closer to one end of the chromosome than the other (the end is chosen randomly). Therefore they are cruciform in late prophase I, and outwardly resemble monocentric chromosomes at the same stage (Monen et al 2005; Albertson, & Thomson 1993) (Figure 4A). Holocentric chromosomes in mitosis do not have a single centromere to act as a site for maintenance of sister chromatid cohesion. In *C. elegans*, cohesion is maintained distal to the chiasma, meaning that this structure has an important role in ensuring that homologues separate in anaphase I, but sister chromatids stay together until anaphase II. Before nuclear envelope breakdown of meiosis I, bivalents condense very tightly, so the short arms of the cruciform are no longer visible and the bivalent takes on a capsule shape

(Schvarzstein et al 2010). Each end of the bivalent (Figure 4A) faces a spindle pole, and moves toward that pole in anaphase. The kinetochore proteins CENP-C, KNL-1, BUB-1, HIM-10, NDC-80, Nuf2, and MIS-12 form a cup around each end of the bivalent (Monen et al 2005). Intriguingly, this structure is independent of the centromere-specific histone HCP-3 (the ortholog of cenH3/CENP-A/Cid). In oocytes, lateral interactions between the bivalent and spindle microtubules, which form a sheath around each bivalent, facilitate congression to the metaphase plate (Monen et al 2005). The chromokinesin KLP 19, which forms a ring around the equator of the bivalent, exerts a polar ejection force that aids congression (Wignall, & Villeneuve 2009).

Because both oocyte and spermatocyte meiosis can be visualized in *C. elegans*, interesting mechanistic differences between the two have been observed. In spermatocytes, spindle-attachment is restricted to each end of a bivalent, as microtubules are embedded in focused areas on each chromosome (Albertson, & Thomson 1993; Shakes et al 2009; Wignall, & Villeneuve 2009). Despite this limited attachment, kinetochore proteins appear to form a cup around each entire half-bivalent as they do in oocytes (Shakes et al 2009; Wignall, & Villeneuve 2009). In spermatocytes, each chromosome end acts functionally as a kinetochore, binding to spindle microtubules and allowing congression to the metaphase plate (Shakes et al 2009). Anaphase chromosome separation occurs because cohesion distal to the chiasma is released (Figure 4A). In oocytes, homolog segregation appears to be driven by growth of microtubules between separating homologues (Dumont et al 2010). In spermatocytes, however, homologues move toward their associated spindle pole along microtubules attached to the "kinetochore" at each end of the bivalent (Shakes et al 2009).

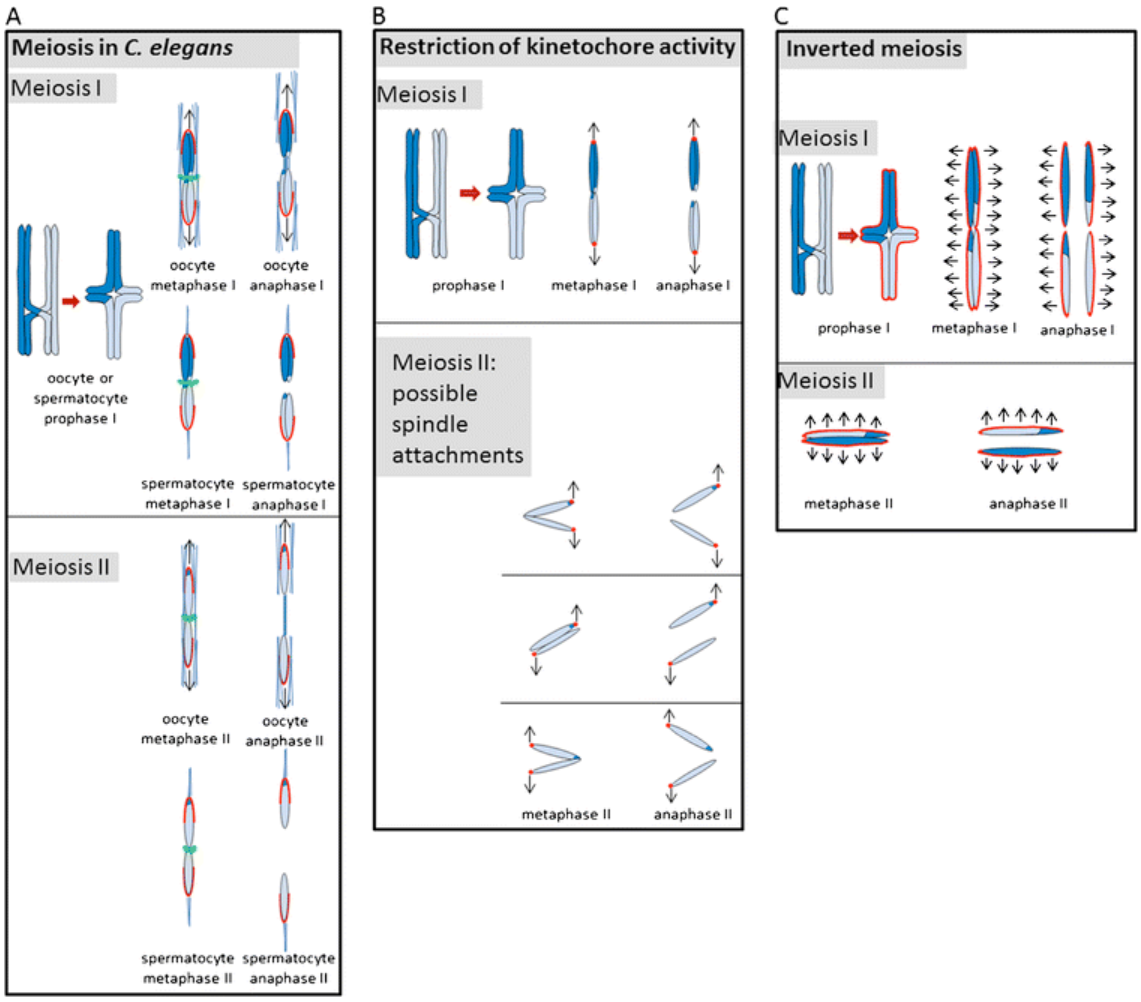


Figure 4-4 Solutions to the problem of holocentric chromosomes in meiosis. **a** During meiosis in the nematode *C. elegans* in both the oocyte and spermatocyte the kinetochore forms a cup-shaped structure in meiosis I. In oocytes, the spindles form a sheet around the bivalents, whereas in spermatocytes the spindle binding is restricted to the ends of the bivalents. In meiosis II, the sister chromatids separate similarly as in meiosis I.**b** In some insects, holocentric chromosomes in mitosis behave as monocentric chromosomes in meiosis. Kinetochore activity is limited to one end of the chromosome (randomly chosen). **c** In inverted meiosis, sister chromatids separate in the first meiotic division. It is speculated that this separation happens because chiasmata are terminalized in prophase of meiosis I, such that chromosomes are arranged in metaphase I with sister chromatids facing opposite poles. In anaphase I, all chromatids separate. Then, homologous chromatids re-pair prior to meiosis II

Restriction of kinetochore activity in the true bugs (Hemiptera)

Meiosis in the true bugs (Hemiptera*:* Heteroptera) resembles that of *C. elegans* spermatocytes, in that kinetochore activity is restricted to a small region of a chromosome (notably, meiosis in Hemiptera has been mainly studied in spermatocytes) (Figure 4B). Electron microscopy shows

that the mitotic holocentric kinetochore in the milkweed bug *Oncopeltus fasciatus* has a layered appearance with visible kinetochore plates, while in meiosis, the areas of kinetic activity are more diffuse (Comings, & Okada 1972). After crossover, bivalents condense very strongly and terminalize their chiasmata so they take on a capsule shape similar to that of *C. elegans* chromosomes (Wolfe, & John 1965). Kinetochore activity in meiosis is restricted to the ends of the capsule-shaped bivalent, and chromosome cohesion is lost in a two-step pattern. This ensures that homologous chromosomes separate in anaphase I and sister chromatids separate in anaphase II, leading to the formation of haploid gametes. Mechanisms for forming the temporary kinetochore and for differential retention of cohesion between anaphase I and anaphase II are not well understood. It will be interesting to determine if bugs use a similar mechanism to *C. elegans* for selecting which end of the chromosome faces the spindle pole and has kinetochore activity. One interesting finding is that the chromosome end chosen to act as a temporary kinetochore can switch between the two meiotic divisions, as seen in *Triatoma infestans* (Heteroptera) (Pérez et al 1997). Amazingly, meiosis II in *T. infestans* can feature sister chromatids with kinetochores at opposite ends (Pérez et al 2000).

Inverted meiosis

An alternative solution to the problem of holocentric chromosomes in meiosis is to invert the meiotic divisions, so that sister chromatids separate in meiosis I and homologs separate in meiosis II (Figure 4C). Inverted meiosis appears in both animals and plants, including a suborder of true bugs (Homoptera), some dragonflies and damselflies, some arachnids, and woodrushes of the genus *Luzula* (Viera et al 2009). It is best studied in the true bugs (Homoptera), where most species have at least one sex with inverted meiosis (John 1990b). In meiotic prophase of these

species, chromosomes recombine, and chiasmata are terminalized as in the Heteroptera (Hughes-Schrader 1944; John, & Claridge 1974; John 1990b) (Figure 4C). In inverted meiosis, chromosomes align differently than in cases of restricted kinetochore activity. Sister chromatids face opposite poles, and separate from one another in anaphase I (Figure 4C). All cohesion appears to be lost between chromatids by telophase I (Hughes-Schrader 1944; John 1990b). Homologous chromatids re-pair prior to the second meiotic division, and separate from one another in anaphase II. The molecular mechanisms underlying inverted meiosis have not been studied.

Restricted kinetochore activity and inverted meiosis can co-exist in the same cell. In the Heteroptera, autosomes restrict kinetic activity (Comings, & Okada 1972; Wolfe, & John 1965), while sex chromosomes undergo inverted meiosis (Schrader 1935; Pérez et al 2000; Viera et al 2009). The sex chromosomes of Heteroptera differ from autosomes of Homoptera. Either they lack a pairing partner (in XX-X0 sex determination), or the X and Y chromosomes do not recombine. Thus, in meiosis I, the sex chromosomes align on the spindle and behave like holocentric chromosomes in mitosis, separating sister chromatids (Schrader 1935; Pérez et al 2000; Viera et al 2009). In meiosis II in organisms with XX-X0 sex determination, the lone X moves to one spindle pole in anaphase II (Schrader 1935). In organisms with XX-XY sex determination, sister chromatids of the X and Y chromosomes separate in anaphase I. In meiosis II, the X and Y chromatids re-pair, align on the metaphase plate, then separate in anaphase II (Pérez et al 2000; Viera et al 2009). Again, little is known about proteins that regulate these behaviors.

Other meiotic adaptations in holocentric organisms

In some Homoptera, formation of gametes in males does not require meiosis at all. In species with haploid males and diploid females, males form sperm via typical mitotic divisions and females have inverted meiosis (Hughes-Schrader, & Tremblay 1966). Other Homoptera (e.g. *Phenacoccus*) have diploid males who generate haploid sperm starting with a mitotic division. To achieve this, one haploid set of chromosomes is inactivated and maintained as heterochromatin (the haploid chromosome sets are separately marked by imprinting). During meiosis II, the active chromosome set segregates to one end of the cell and is incorporated into a sperm, while the inactive chromosome set is sequestered in a separate nucleus and eventually ejected (Hughes-Schrader 1935).

Meiosis in some holocentric organisms remains poorly characterized. In addition, it remains unclear why very different meiotic adaptations exist and why the closely related clades (e.g. the two suborders of Hemiptera) utilize such different mechanisms. Molecular characterization of meiosis in holocentric organisms is likely to illuminate these questions.

Genomic tools to study genomes of holocentric species

For most species with holocentric chromosomes, the DNA sequence underlying the kinetochore is unknown. Recently, the chromosomal localization of the centromere-specific histone cenH3 was determined in the nematode *C. elegans* by ChIP-chip analysis (Gassmann et al 2012). It was found that ~50% of the genome can be associated with cenH3, showing that particular DNA sequences are unlikely to control cenH3 incorporation (furthermore, up to 90% of the nucleosomes in these centromeric regions may contain conventional H3). Importantly, the distribution of cenH3-containing regions was inversely correlated with genes transcribed in the

germline and early embryo when the pattern of cenH3 incorporation is established (Gassmann et al 2012). This suggests that transcription excludes cenH3 incorporation.

The chromosomes of most animal and plant species are monocentric, and the centromeres of these species are usually characterized by high copy tandem repeat arrays (Henikoff et al 2001; Melters et al in preparation). The genome of *C. elegans* contains few tandem repeats (Hillier et al 2007), but this is not the case for all genomes of species with holocentric chromosomes (Table 1). In the snowy woodrush *Luzula nivea* and its close relative *L. elegans* (Nagaki et al 2005; Heckmann et al 2011). In contrast to *C.* elegans, a high copy 178 bp tandem repeat was found in the genome of *L. nivea* (Haizel et al 2005). This repeat formed at least five distinct arrays per chromosome when studied by FISH. Whether cenH3 is preferentially localized to these large tandem repeat arrays remains to be determined.

With advances in genome sequencing technology, the genomes of dozens of holocentric species are available. Bioinformatic tools can find and analyze high copy tandem repeats from shotgun genome sequences (Alkan et al 2011; Melters et al in preparation). One recent study surveyed tandem repeats in 282 animal and plant species, including 32 holocentric species (11 arthropods and 21 nematodes) (Melters et al in preparation). The genomic abundance of tandem repeats in these 32 species differs greatly. The majority have only low-copy tandem repeats similar to *C. elegans*, but some holocentric species have high copy tandem repeats with an overall genomic abundance comparable to that of the monocentric *Arabidopsis thaliana* and human genomes. These may represent cases like *Luzula nivea* in which tandem repeats have spread to several locations along each chromosome.

Species	Common name	Length	Localization	References
Juncaceae; Monocots				
Luzula flaccida	Pale woodrush	127 bp	Interspersed	(Collet 1984)
Luzula flaccida	Pale woodrush	190 bp	N/A	(Collet 1984)
Luzula flaccida	Pale woodrush	184 bp	N/A	(Collet 1984)
Luzula nivea	Snowy woodrush	178 bp	Interspersed	(Haizel et al 2005)
Lepidoptera; Insecta				
Mamestra brassicae	Cabbage moth	234 bp	Subtelomeric	(Mandrioli et al 2003)
Spodoptera frugiperda	Fall armyworm	189 bp	N/A	(Lu et al 1994)
Hemiptera; Insecta				
Myzus persicae	Peach potato aphid	169 bp	Subtelomeric	(Spence et al 1998)
Myzus persicae	Peach potato aphid	189 bp	Subtelomeric	(Mandrioli et al 1999)
Ixodidae; Arachnida				
Rhipicephalus microplus	Southern cattle tick	149 bp	Subtelomeric	(Hill et al 2009)
Rhicicephalus microplus	Southern cattle tick	178 bp	Subtelomeric	(Hill et al 2009)
Rhicicephalus microplus	Southern cattle tick	177 bp	Subtelomeric	(Hill et al 2009)
Rhicicephalus microplus	Southern cattle tick	216 bp	Subtelomeric	(Hill et al 2009)
Ascarididae; Nematoda				
Ascaris suum	Pig roundworm	123 bp	Subtelomeric†	(Niedermaier, & Moritz 2000; Streeck et al 1982)
Ascaris lumbricoides	Roundworm	121 bp	N/A	(Müller et al 1982)
Parascaris univalens	Roundworm	5 bp	Subtelomeric†	(Niedermaier, & Moritz 2000; Teschke et al 1991)
Parascaris univalens	Roundworm	10 bp	Subtelomeric†	(Niedermaier, & Moritz 2000)
Rhabditida; Nematoda				
Steinernema glaseri	Roundworm	174 bp	N/A	(Grenier et al 1996)
Heterorhabditis bacteriophora	Roundworm	168 bp	N/A	(Grenier et al 1996)
Heterorhabditis indicus	Roundworm	174 bp	N/A	(Abadon et al 1998)
Panagrellus redivivus	Roundworm	155 bp	N/A	(de Chastonay et al 1990)
Panagrellus redivivus	Roundworm	167 bp	N/A	(de Chastonay et al 1990)
Ostertagia circumcincta	Red stomach worm	218 bp	N/A	(Callaghan, & Beh 1996)
Aphelenchida; Nematoda				
Bursaphelenchus xylophilus	Pine wood nematode	160 bp	N/A	(Tarès et al 1993)
Tylenchida; Nematoda				
Meloidogyne hapla	Root-knot nematode	169 bp	N/A	(Piotte et al 1994)
Meloidogyne incognita	Root-knot nematode	295 bp	N/A	(Piotte et al 1994)
Meloidogyne chitwoodi	Root-knot nematode	180 bp	N/A	(Castagnone-Sereno et al 1998)
Meloidogyne arenaria	Root-knot nematode	172 bp	N/A	(Castagnone-Sereno et al 2000; Mestrović et al 2005)

Table 4-1 List of species with holocentric chromosomes and literature reports of tandem repeats and the chromosomal localization of the tandem repeat if known. † Species with chromatin diminution (Müller et al 1996).

High copy tandem repeats have also been studied experimentally in holocentric organisms (Table 1). These tandem repeats were often localized to subtelomeric regions (Spence et al 1998; Malik, & Henikoff 2009; Mandrioli et al 2003; Hill et al 2009). With the restricted kinetochore activity in meiosis observed in some holocentric species, it is tempting to ask if tandem repeat arrays have a role in specifying centromere activity in meiosis.

Evolutionary implications of holocentric chromosomes

In theory, holocentric chromosomes encourage rapid karyotype evolution. Fission of a holocentric chromosome should create two fragments that both retain centromere activity. Similarly, fusion of holocentric chromosomes would not create the problems faced by a dicentric chromosome in a monocentric organism. These predictions are based on mitotic chromosome segregation, and some types of meiotic adaptation (e.g. kinetochore restriction) may be fatally affected by chromosome fission. In general, holocentric clades do not show the predicted increase in karyotypic diversity (Panzera et al 1996; Gokhman, & Kuznetsova 2006). In contrast, sedges (genus *Carex*), scale insects (genus *Apiomorpha*), and scorpions from the family *Buthidae* have extremely labile karyotypes (Cook 2000; Hipp 2007; Schneider et al 2009). *Carex* has inverted meiosis, and *Buthidae* have achiasmate meiosis (Davies 1956; Schneider et al 2009). These features may have allowed karyotypes to change without compromising holocentric meiosis (Schneider et al 2009).

The holocentric chromosome characteristic is also likely to affect the evolution of centromere DNA sequences and of the kinetochore proteins that bind to them. Despite the essential function

of centromeres in chromosome segregation, centromere DNAs and many kinetochore proteins evolve very rapidly (Talbert et al 2004, Meraldi et al 2006). Henikoff and Malik have proposed a female meiotic drive hypothesis to explain this paradox (Henikoff et al 2001; Malik, & Henikoff 2009). Asymmetric meiosis in females generates only one functional egg cell from four gametes, and centromere DNA sequence polymorphisms that encourage segregation into the surviving egg cell will have a huge selective advantage. Correspondingly, kinetochore proteins such as cenH3 would co-evolve to equalize binding between all centromeres in a population, ensuring that unequal centromere DNA/kinetochore protein interactions do not cause chromosome segregation errors. Tandem repeats are likely to facilitate rapid evolution of centromere DNA in monocentric chromosomes. A sequence polymorphism that shows preferential segregation could spread from one chromosome to another by gene conversion, because tandem repeats on all chromosomes have similar sequences.

Genomic evidence suggests that many holocentric chromosomes lack tandem repeats, and have cenH3 binding sites distributed over a wide variety of unique sequences (*C. elegans* is the exemplar of such organisms) (Gassmann et al 2012; Melters et al in preperation). If cenH3 binds to such a diverse range of sites, it may be much more difficult for holocentric chromosomes to acquire DNA sequence changes that favor segregation into the surviving egg cell during female meiosis. Changes in centromere DNA evolution might relax the pressure on cenH3 to evolve rapidly , a prediction which should be testable in holocentric clades. Interesting, *Caenorhabditis* cenH3 continues to show signs of positive selection despite the fact that it binds to diverse DNA sequences in holocentric chromosomes (Zedek & Bureš 2012).

Another factor which is likely to influence centromere DNA evolution is the mechanism of kinetochore assembly during meiosis in holocentric organisms. We do not know if particular DNA sequences are important for assembling the cup-shaped kinetochore protein structures seen in *C. elegans* meiosis, and chromosome segregation is cenH3 independent in meiosis (Monen et al 2005). Similarly, inverted meiosis seems to feature spindle attachment sites that are spread along the length of meiotic chromosomes and may assemble on unique sequences in many holocentric organisms. Only in holocentric organisms such as *Luzula nivea*, which has tandem repeats underlying cenH3 binding sites, is it possible for centromere DNA to evolve rapidly in the same manner as it does in monocentric chromosomes (Haizel et al 2005). *Luzula* has inverted meiosis with distributed kinetochore activity. Centromere repeat evolution may well be different in inverted meiosis organisms when compared with holocentric organisms that restrict kinetochore activity to particular chromosome regions (especially if these regions contain high copy tandem repeats).

Final remarks

How do holocentric chromosomes inform our general understanding of the chromosome segregation machinery? Given the extreme meiotic adaptations necessary for organisms to adopt holocentric chromosomes, it is surprising and fascinating that distributed kinetochores have arisen so many times during eukaryotic evolution. It is likely that kinetochore location is controlled epigenetically in almost all eukaryotes (Stimpson, & Sullivan 2010). Convergent evolution of holocentric chromosomes in extremely disparate eukaryotes suggests that cenH3 recruitment and maintenance can be altered relatively easily to greatly alter the placement of kinetochore proteins on mitotic chromosomes. Recent results from *A. thaliana* suggest that

meiotic kinetochores have a unique assembly pathway, possibly because sister kinetochores must be mono-oriented towards the same side of the spindle (Ravi et al 2011). Meiosis in holocentric organisms has several different solutions to the problem of kinetochore geometry, further suggesting that centromere assembly in meiosis can be altered without affecting the nature of mitotic cenH3 recruitment.

A flurry of recent papers describing cenH3 from deeply diverged eukaryotes shows how inexpensive genome sequencing is changing our ability to address questions in comparative cell biology (Maruyama et al 2008; Dawson et al 2007; Dubin et al 2010; Brooks et al 2011). It will be very interesting to ask whether holocentric chromosomes are seen in any single celled eukaryotes whose genomes are too small for classical cytology. Similarly, it is now much easier to generate molecular reagents and to inactivate genes in non-model organisms (for example, with TALE nucleases) (Bogdanove, & Voytas 2011). These tools should allow investigators to make great inroads in our understanding of centromere assembly and meiotic adaptations in organisms with holocentric chromosomes.

Acknowledgements

We are indebted to P. Ward, P. Gullan and P. Cranston for their advice on insect phylogeny. D.M. was supported by training grant T32-GM08799 from NIH-NIGMS. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH. S.C. is a Howard Hughes Medical Institute and Gordon and Betty Moore Foundation investigator.

References

Abadon, M., Grenier, E., Laumond, C. & Abad, P., 1998, A species-specific satellite DNA from the entomopathogenic nematode Heterorhabditis indicus, Genome / National Research Council Canada = Génome / Conseil national de recherches Canada, 41(2), pp. 148-53.

Albertson, D.G. & Thomson, J.N., 1993, Segregation of holocentric chromosomes at meiosis in the nematode, Caenorhabditis elegans, Chromosome Research, 1(1), pp. 15-26.

Alkan, C., Cardone, M.F., Catacchio, C.R., Antonacci, F., O'Brien, S.J., Ryder, O.A., Purgato, S., Zoli, M., Della Valle, G., Eichler, E.E. & Ventura, M., 2011, Genome-wide characterization of centromeric satellites from multiple mammalian genomes, Genome research, 21(1), pp. 137-45.

Bogdanove, A.J. & Voytas, D.F., 2011, TAL effectors: customizable proteins for DNA targeting, Science (New York, N.Y.), 333(6051), pp. 1843-6.

Bradley, T.J., Briscoe, A.D., Brady, S.G., Contreras, H.L., Danforth, B.N., Dudley, R., Grimaldi, D., Harrison, J.F., Kaiser, J.A. & Merlin, C., 2009, Episodes in insect evolution, Integrative and Comparative Biology, 49(5), pp. 590-606.

Bremer, K., 2002, Gondwanan evolution of the grass alliance of families (Poales), Evolution; international journal of organic evolution, 56(7), pp. 1374-87.

Brooks, C.F., Francia, M.E., Gissot, M., Croken, M.M., Kim, K. & Striepen, B., 2011, Toxoplasma gondii sequesters centromeres to a specific nuclear region throughout the cell cycle, Proceedings of the National Academy of Sciences of the United States of America, 108(9), pp. 3767-72.

Callaghan, M.J. & Beh, K.J., 1996, A tandemly repetitive DNA sequence is present at diverse locations in the genome of Ostertagia circumcincta, Gene, 174(2), pp. 273-9.

Castagnone-Sereno, P., Leroy, F. & Abad, P., 2000, Cloning and characterization of an extremely conserved satellite DNA family from the root-knot nematode Meloidogyne arenaria, Genome / National Research Council Canada = Génome / Conseil national de recherches Canada, 43(2), pp. 346-53.

Castagnone-Sereno, P., Leroy, H., Semblat, J.P., Leroy, F., Abad, P. & Zijlstra, C., 1998, Unusual and strongly structured sequence variation in a complex satellite DNA family from the nematode Meloidogyne chitwoodi, Journal of molecular evolution, 46(2), pp. 225-33.

de Chastonay, Y., Müller, F. & Tobler, H., 1990, Two highly reiterated nucleotide sequences in the low C-value genome of Panagrellus redivivus, Gene, 93(2), pp. 199-204.

Collet, C., 1984, Highly repeated DNA and holocentric chromosomes of the woodrush Luzula flaccida (Juncaceae), Canadian journal of genetics and cytology, 26(3), pp. 288-95.

Comings, D.E. & Okada, T.A., 1972, Holocentric chromosomes in Oncopeltus: kinetochore plates are present in mitosis but absent in meiosis, Chromosoma, 37(2), pp. 177-92.

Cook, L.G., 2000, Extraordinary and extensive karyotypic variation: a 48-fold range in chromosome number in the gall-inducing scale insect Apiomorpha (Hemiptera: Coccoidea: Eriococcidae), Genome / National Research Council Canada = Génome / Conseil national de recherches Canada, 43(2), pp. 255-63.

Davies, E.W., 1956, Cytology, evolution and origin of the aneuploid series in the genus Carex, Hereditas, 42, pp. 349-65.

Dawson, S.C., Sagolla, M.S. & Cande, W.Z., 2007, The cenH3 histone variant defines centromeres in Giardia intestinalis, Chromosoma, 116(2), pp. 175-84.

Dernburg, A.F., 2001, Here, there, and everywhere: kinetochore function on holocentric chromosomes, The Journal of cell biology, 153(6), pp. F33-8.

Desai, R.N., 1969, Monocentric nature of the chromosomes ofRanatra (heteroptera) verified by the induced fragmentation experiments, Cellular and Molecular Life Sciences, 25(11), pp. 1170-1.

Desai, R.N. & Deshpande, S.B., 1969, Centromere nature of the chromosomes ofRanatra (Heteroptera), Cellular and Molecular Life Sciences, 25(4), pp. 386-7.

Doan, R.N. & Paliulis, L.V., 2009, Micromanipulation reveals an XO-XX sex determining system in the orb-weaving spider Neoscona arabesca (Walckenaer), Hereditas, 146(4), pp. 180-2.

Dubin, M., Fuchs, J., Gräf, R., Schubert, I. & Nellen, W., 2010, Dynamics of a novel centromeric histone variant CenH3 reveals the evolutionary ancestral timing of centromere biogenesis, Nucleic acids research, 38(21), pp. 7526-37.
Dumont, J., Oegema, K. & Desai, A., 2010, A kinetochore-independent mechanism drives anaphase chromosome separation during acentrosomal meiosis, Nature cell biology, 12(9), pp. 894-901.
Fuková, I., Traut, W., Vítková, M., Nguyen, P., Kubícková, S. & Marec, F., 2007, Probing the W chromosome of the codling moth, Cydia pomonella, with sequences from microdissected sex chromatin, Chromosoma, 116(2), pp. 135-45.
Gassmann, R., Rechtsteiner, A., Yuen, K.W., Muroyama, A., Egelhofer, T., Gaydos, L., Barron, F., Maddox, P., Essex, A., Monen, J., Ercan, S., Lieb, J.D., Oegema, K., Strome, S. & Desai, A., 2012, An inverse relationship to germline transcription defines centromeric chromatin in C. elegans, Nature.
Godward, M.B.E., 1954, Irradiation of Spirogyra chromosomes, Heredity, 8(2), p. 293.
Gokhman, V.E. & Kuznetsova, V.G., 2006, Comparative insect karyology: current state and applications, Entomological Review, 86(3), pp. 352-68.
Grenier, E., Laumond, C. & Abad, P., 1996, Molecular characterization of two species-specific tandemly repeated DNAs from entomopathogenic nematodes Steinernema and Heterorhabditis (Nematoda:Rhabditida), Molecular and biochemical parasitology, 83(1), pp. 47-56.
Grimaldi, D. & Engel, M.S., 2005, Evolution of the insects, Cambridge University Press,.
Guerra, M., Cabral, G., Cuacos, M., González-García, M., González-Sánchez, M., Vega, J. & Puertas, M.J., 2010, Neocentrics and holokinetics (holocentrics): chromosomes out of the centromeric rules, Cytogenetic and genome research, 129(1-3), pp. 82-96.
Haizel, T., Lim, Y.K., Leitch, A.R. & Moore, G., 2005, Molecular analysis of holocentric centromeres of Luzula species, Cytogenetic and genome research, 109(1-3), pp. 134-43.
Heckmann, S., Schroeder-Reiter, E., Kumke, K., Ma, L., Nagaki, K., Murata, M., Wanner, G. & Houben, A., 2011, Holocentric chromosomes of Luzula elegans are characterized by a longitudinal centromere groove, chromosome bending, and a terminal nucleolus organizer region, Cytogenetic and genome research, 134(3), pp. 220-8.
Henikoff, S., Ahmad, K. & Malik, H.S., 2001, The centromere paradox: stable inheritance with rapidly evolving DNA, Science, 293(5532), pp. 1098-102.
Hill, C.A., Guerrero, F.D., Van Zee, J.P., Geraci, N.S., Walling, J.G. & Stuart, J.J., 2009, The position of repetitive DNA sequence in the southern cattle tick genome permits chromosome identification, Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 17(1), pp. 77-89.
Hillier, L.W., Miller, R.D., Baird, S.E., Chinwalla, A., Fulton, L.A., Koboldt, D.C. & Waterston, R.H., 2007, Comparison of C. elegans and C. briggsae genome sequences reveals extensive conservation of chromosome organization and synteny, PLoS biology, 5(7), p. e167.
Hipp, A.L., 2007, Nonuniform processes of chromosome evolution in sedges (Carex: Cyperaceae), Evolution; international journal of organic evolution, 61(9), pp. 2175-94.
Hirai, H., Tada, I., Takahashi, H., Nwoke, B.E. & Ufomadu, G.O., 1987, Chromosomes of Onchocerca volvulus (Spirurida: Onchocercidae): a comparative study between Nigeria and Guatemala, J Helminthol, 61(1), pp. 43-6.
Hughes-Schrader, S., 1935, The chromosome cycle of Phenacoccus (Coccidae), Biological Bulletin, 69(3), pp. 462-8.
Hughes-Schrader, S., 1944, A primitive coccid chromosome cycle in Puto sp, The Biological Bulletin, 87(3), pp. 167-76.
Hughes-Schrader, S. & Ris, H., 1941, The diffuse spindle attachment of coccids, verified by the mitotic behavior of induced chromosome fragments, Journal of Experimental Zoology, 87(3), pp. 429-56.
Hughes-Schrader, S. & Schrader, F., 1961, The kinetochore of the Hemiptera, Chromosoma, 12(1), pp. 327-50.
Hughes-Schrader, S. & Tremblay, E., 1966, Gueriniella and the cytotaxonomy of iceryine coccids (Coccoidea: Margarodidae), Chromosoma, 19(1), pp. 3-13.
John, B., 1990a, Meiosis, Cambridge University Press,.

John, B., 1990b, Meiosis, Cambridge University Press,.
John, B. & Claridge, M.F., 1974, Chromosome variation in British populations on Oncopsis (Hemiptera: Cicadellidae), Chromosoma, 46(1), pp. 77-89.
Kuta, E., Bohanec, B., Dubas, E., Vizintin, L. & Przywara, L., 2004, Chromosome and nuclear DNA study on Luzula - a genus with holokinetic chromosomes, Genome / National Research Council Canada = Génome / Conseil national de recherches Canada, 47(2), pp. 246-56.
Kuta, E., Przywara, L. & Hejmej, J., 1998, Karyotype variability in Pleurozium schreberi (Brid.) Mitt, order.
Kuta, E., Przywara, L. & Ilnicki, T., 2000, Heterochromatin in Pleurozium schreberi (Brid.) Mitt, order.
De Ley, P., 2006, A quick tour of nematode diversity and the backbone of nematode phylogeny, WormBook : the online review of C. elegans biology, pp. 1-8.
Lu, Y.J., Kochert, G.D., Isenhour, D.J. & Adang, M.J., 1994, Molecular characterization of a strain-specific repeated DNA sequence in the fall armyworm Spodoptera frugiperda (Lepidoptera: Noctuidae), Insect molecular biology, 3(2), pp. 123-30.
Luceño, M., Vanzela, A.L.L. & Guerra, M., 1998, Cytotaxonomic studies in Brazilian Rhynchospora (Cyperaceae), a genus exhibiting holocentric chromosomes, Canadian journal of botany, 76(3), pp. 440-9.
Maddison, D.R. & Schulz, K.S., Tree of Life Web Project, Tree of Life Web Project.
Maddox, P.S., Oegema, K., Desai, A. & Cheeseman, I.M., 2004, Holo'er than thou: Chromosome segregation and kinetochore function in C. elegans, Chromosome Research, 12(6), pp. 641-53.
Malik, H.S. & Henikoff, S., 2009, Major evolutionary transitions in centromere complexity, Cell, 138(6), pp. 1067-82.
Mandrioli, M., Bizzaro, D., Manicardi, G.C., Gionghi, D., Bassoli, L. & Bianchi, U., 1999, Cytogenetic and molecular characterization of a highly repeated DNA sequence in the peach potato aphid Myzus persicae, Chromosoma, 108(7), pp. 436-42.
Mandrioli, M., Manicardi, G.C. & Marec, F., 2003, Cytogenetic and molecular characterization of the MBSAT1 satellite DNA in holokinetic chromosomes of the cabbage moth, Mamestra brassicae (Lepidoptera), Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 11(1), pp. 51-6.
Maruyama, S., Matsuzaki, M., Kuroiwa, H., Miyagishima, S.Y., Tanaka, K., Kuroiwa, T. & Nozaki, H., 2008, Centromere structures highlighted by the 100%-complete Cyanidioschyzon merolae genome, Plant Signal Behav, 3(2), pp. 140-1.
Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J.F., DeRisi, J.L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I.F. & Chan, S.W., Patterns of Centromere Tandem Repeat Evolution in 282 Animal and Plant Genomes, in preparation.
Meraldi, P., McAinsh, A.D., Rheinbay, E., Sorger, P.K., 2006, Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. 7(3), p. R23
Mestrović, N., Randig, O., Abad, P., Plohl, M. & Castagnone-Sereno, P., 2005, Conserved and variable domains in satellite DNAs of mitotic parthenogenetic root-knot nematode species, Gene, 362, pp. 44-50.
Mitreva, M., Blaxter, M.L., Bird, D.M. & McCarter, J.P., 2005, Comparative genomics of nematodes, Trends in genetics : TIG, 21(10), pp. 573-81.
Monen, J., Maddox, P.S., Hyndman, F., Oegema, K. & Desai, A., 2005, Differential role of CENP-A in the segregation of holocentric C. elegans chromosomes during meiosis and mitosis, Nature cell biology, 7(12), pp. 1248-55.
Mughal, S. & Godward, M.B.E., 1973, Kinetochore and microtubules in two members of Chlorophyceae, Cladophora fracta and Spirogyra majuscula, Chromosoma, 44(2), pp. 213-29.
Mutafova, T., Dimitrova, Y. & Komandarev, S., 1982, The karyotype of four Trichinella species, Z Parasitenkd, 67(1), pp. 115-20.
Müller, F., Bernard, V. & Tobler, H., 1996, Chromatin diminution in nematodes, Bioessays, 18(2), pp. 133-8.

Müller, F., Walker, P., Aeby, P., Neuhaus, H., Felder, H., Back, E. & Tobler, H., 1982, Nucleotide sequence of satellite DNA contained in the eliminated genome of Ascaris lumbricoides, Nucleic acids research, 10(23), pp. 7493-510.
Nagaki, K., Kashihara, K. & Murata, M., 2005, Visualization of diffuse centromeres with centromere-specific histone H3 in the holocentric plant Luzula nivea, The Plant cell, 17(7), pp. 1886-93.
Niedermaier, J. & Moritz, K.B., 2000, Organization and dynamics of satellite and telomere DNAs in Ascaris: implications for formation and programmed breakdown of compound chromosomes, Chromosoma, 109(7), pp. 439-52.
Nijalingappa, B.H.M., 1974, Cytological studies inScirpus (Cyperaceae), Proceedings: Plant Sciences, 80(3), pp. 134-8.
Oliver, J.H., 1972, Cytogenetics of Ticks (Acari: Ixodoidea). 8. Chromosomes of six species of Egyptian Hyalomma (Ixodidae), The Journal of parasitology, 58(3), pp. 611-3.
Oliver, J.H., 1977, Cytogenetics of mites and ticks, Annu Rev Entomol, 22, pp. 407-29.
Oliver, J.H., Tanaka, K. & Sawada, M., 1974, Cytogenetics of ticks (Acari: Ixodoidea). 14. Chromosomes of nine species of Asian Haemaphysalines, Chromosoma, 45, pp. 445-56.
Paliulis, L.V. & Nicklas, R.B., 2004, Micromanipulation of chromosomes reveals that cohesion release during cell division is gradual and does not require tension, Current biology, 14(23), pp. 2124-9.
Paliulis, L.V. & Nicklas, R.B., 2005, Kinetochore rearrangement in meiosis II requires attachment to the spindle, Chromosoma, 113(8), pp. 440-6.
Panzera, F., Pérez, R., Hornos, S., Panzera, Y., Cestau, R., Delgado, V. & Nicolini, P., 1996, Chromosome numbers in the Triatominae (Hemiptera-Reduviidae): a review, Mem Inst Oswaldo Cruz, 91(4), pp. 515-8.
Pazy, B. & Plitmann, U., 1994, Holocentric chromosome behaviour in Cuscuta (Cuscutaceae), Plant systematics and evolution, 191(1), pp. 105-9.
Pérez, R., Panzera, F., Page, J., Suja, J.A. & Rufas, J.S., 1997, Meiotic behaviour of holocentric chromosomes: orientation and segregation of autosomes in Triatoma infestans (Heteroptera), Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 5(1), pp. 47-56.
Pérez, R., Rufas, J.S., Suja, J.A., Page, J. & Panzera, F., 2000, Meiosis in holocentric chromosomes: orientation and segregation of an autosome and sex chromosomes in Triatoma infestans (Heteroptera), Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 8(1), pp. 17-25.
Piotte, C., Castagnone-Sereno, P., Bongiovanni, M., Dalmasso, A. & Abad, P., 1994, Cloning and characterization of two satellite DNAs in the low-C-value genome of the nematode Meloidogyne spp, Gene, 138(1-2), pp. 175-80.
Post, R., 2005, The chromosomes of the Filariae, Filaria J, 4, p. 10.
Procunier, W.S. & Hirai, H., 1986, The chromosomes of Onchocerca volvulus, Parasitol Today, 2(11), pp. 307-9.
Ravi, M., Shibata, F., Ramahi, J.S., Nagaki, K., Chen, C., Murata, M. & Chan, S.W., 2011, Meiosis-specific loading of the centromere-specific histone CENH3 in Arabidopsis thaliana, PLoS genetics, 7(6), p. e1002121.
Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W. & Cunningham, C.W., 2010, Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences, Nature, 463(7284), pp. 1079-83.
Schneider, M.C., Zacaro, A.A., Pinto-da-Rocha, R., Candido, D.M. & Cella, D.M., 2009, Complex meiotic configuration of the holocentric chromosomes: the intriguing case of the scorpion Tityus bahiensis, Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 17(7), pp. 883-98.
Schrader, F., 1935, Notes an the Mitotic Behavior of Long Chromosomes, Cytologia, 6(4), pp. 422-30.
Schvarzstein, M., Wignall, S.M. & Villeneuve, A.M., 2010, Coordinating cohesion, co-orientation, and congression during meiosis: lessons from holocentric chromosomes, Genes & development, 24(3), pp. 219-28.
Shakes, D.C., Wu, J.C., Sadler, P.L., Laprade, K., Moore, L.L., Noritake, A. & Chu, D.S., 2009, Spermatogenesis-specific features of the meiotic program in Caenorhabditis elegans, PLoS genetics, 5(8), p. e1000611.

Sheikh, S.A. & Kondo, K., 1995, Differential Staining with Orcein, Giemsa, CMA, and DAPI for Comparative Chromosome Study of 12 Species of Australian Drosera (Droseraceae), American Journal of Botany, 82(10), pp. 1278-86.
Spakulová, M., Králová, I. & Cutillas, C., 1994, Studies on the karyotype and gametogenesis in Trichuris muris, J Helminthol, 68(1), pp. 67-72.
Spence, J.M., Blackman, R.L., Testa, J.M. & Ready, P.D., 1998, A 169-base pair tandem repeat DNA marker for subtelomeric heterochromatin and chromosomal rearrangements in aphids of the Myzus persicae group, Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 6(3), pp. 167-75.
Stimpson, K.M. & Sullivan, B.A., 2010, Epigenomics of centromere assembly and function, Curr Opin Cell Biol, 22(6), pp. 772-80.
Streeck, R.E., Moritz, K.B. & Beer, K., 1982, Chromatin diminution in Ascaris suum: nucleotide sequence of the eliminated satellite DNA, Nucleic acids research, 10(11), pp. 3495-502.
Talbert, P.B., Bryson, T.D., Henikoff, S., 2004, Adaptive evolution of centromere proteins in plants and animals. J. Biol. 3(4), p.18
Tanaka, T. & Tanaka, T., 1977, Chromosome studies in Chionographis (Liliaceae): 1. On the holokinetic nature of chromosomes in Chionographis japonica Maxim, Cytologia (Japan), 42(3-4), pp. 753-63.
Tarès, S., Lemontey, J.M., de Guiran, G. & Abad, P., 1993, Cloning and characterization of a highly conserved satellite DNA sequence specific for the phytoparasitic nematode Bursaphelenchus xylophilus, Gene, 129(2), pp. 269-73.
Teschke, C., Solleder, G. & Moritz, K.B., 1991, The highly variable pentameric repeats of the AT-rich germline limited DNA in Parascaris univalens are the telomeric repeats of somatic chromosomes, Nucleic acids research, 19(10), pp. 2677-84.
Vaarama, A., 1954, Cytological observations of Pleurozium schreberi, with special reference to centromere evolution, Ann. Bot. Soc. Zool. Bot. Fenn., 28, pp. 1-59.
Viera, A., Page, J. & Rufas, J.S., 2009, Inverted meiosis: the true bugs as a model to study, Genome Dyn, 5, pp. 137-56.
White, M.J.D., 1973, Animal cytology and evolution, Cambridge University Press,.
Wignall, S.M. & Villeneuve, A.M., 2009, Lateral microtubule bundles promote chromosome alignment during acentrosomal oocyte meiosis, Nature cell biology, 11(7), pp. 839-44.
Wolfe, S.L. & John, B., 1965, The organization and ultrastructure of male meiotic chromosomes n Oncopeltus fasciatus, Chromosoma, 17(2), pp. 85-103.
Zedek, F & Bureš P, 2012, Evidence for centromere drive in holocentric chromosomes of Caenorhabditis, PLoS ONE, 7(1), p. e30496

Supplementary files

Additional file 1 (table):

http://link.springer.com/content/esm/art:10.1007/s10577-012-9292-1/file/MediaObjects/

10577_2012_9292_MOESM1_ESM.pdf

5. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution

Daniël P Melters*[1,2], Keith R Bradnam*[1], Hugh A Young[3], Natalie Telis[1,2], Michael R May[4], J Graham Ruby[5], Robert Sebra[6], Paul Peluso[6], John Eid[6], David Rank[6], José Fernando Garcia[7], Joseph L DeRisi[5,8], Timothy Smith[10], Christian Tobias[3], Jeffrey Ross-Ibarra[#9], Ian Korf[#1] and Simon W-L Chan[2,8]

* contributed equally

corresponding author: JR-I: rossibarra@gmail.com; IK: ifkorf@ucdavis.edu

1. Department of Molecular and Cell Biology and Genome Center, University of California, Davis
2. Department of Plant Biology, University of California, Davis
3. USDA-ARS, Western Regional Research Center
4. Department of Evolution and Ecology, University of California, Davis
5. Department of Biochemistry and Biophysics, University of California, San Francisco
6. Pacific Biosciences
7. Department of Animal Production and Health, Universidade Estadual Paulista, IAEA Collaborating Centre in Animal Genomics and Bioinformatics
8. Howard Hughes Medical Institute
9. Department of Plant Sciences, Center for Population Biology, and Genome Center, University of California, Davis
10. USDA-ARS, U.S. Meat Animal Research Center

Abstract

Background: Centromeres are essential for chromosome segregation, yet their DNA sequences evolve rapidly. In most animals and plants that have been studied, centromeres contain megabase-scale arrays of tandem repeats. Despite their importance, very little is known about the degree to which centromere tandem repeats share common properties between different species across different phyla. We used bioinformatic methods to identify high-copy tandem repeats from 282 species using publicly available genomic sequence and our own data.

Results: Our methods are compatible with all current sequencing technologies. Long Pacific Biosciences sequence reads allowed us to find tandem repeat monomers up to 1,419 bp. We assumed that the most abundant tandem repeat is the centromere DNA, which was true for most species whose centromeres have been previously characterized, suggesting this is a general property of genomes. High-copy centromere tandem repeats were found in almost all animal and plant genomes, but repeat monomers were highly variable in sequence composition and length. Furthermore, phylogenetic analysis of sequence homology showed little evidence of sequence conservation beyond approximately 50 million years of divergence. We find that despite an overall lack of sequence conservation, centromere tandem repeats from diverse species showed similar modes of evolution.

Conclusions: While centromere position in most eukaryotes is epigenetically determined, our results indicate that tandem repeats are highly prevalent at centromeres of both animal and plant genomes. This suggests a functional role for such repeats, perhaps in promoting concerted evolution of centromere DNA across chromosomes.

Introduction

Faithful chromosomal segregation in mitosis and meiosis requires that chromosomes attach to spindle microtubules in a regulated manner via the kinetochore protein complex. As the site of kinetochore assembly, the centromere is the genetic locus that facilitates accurate inheritance. Deletion of the centromere or mutation of critical kinetochore proteins results in chromosome loss (Stoler et al 1995; Shang et al 2010). Proteins and DNA sequences involved in most essential cellular functions are characterized by their high degree of conservation. Given their conserved function, the observed rapid evolution of kinetochore proteins (Talbert et al 2004) and lack of homology of centromere repeats thus poses somewhat of a paradox (Henikoff et al 2001).

Centromeres differ greatly in their sequence organization among species. In the budding yeast *Saccharomyces cerevisiae* a 125-bp sequence is sufficient to confer centromere function, and essential kinetochore proteins bind to this "point centromere" in a sequence-dependent manner (Meraldi et al 2006). Point centromeres are a derived evolutionary characteristic, as ascomycete fungi more distantly related to *S. cerevisiae* have much longer centromere DNAs and do not rely on specific sequences to recruit kinetochore proteins (McAinsh et al 2003; Meraldi et al 2006). In the limited set of plant and animal species that have been previously analyzed, centromere DNAs consist of megabase-sized arrays of simple tandem repeats (or satellite DNA), sometimes interspersed with long terminal repeat transposons (Birchler et al 2009; Wang et al 2009; Wade et al 2009). Some taxa exhibit higher order repeat structures (HORs), in which multiple polymorphic monomers make up a larger repeating unit (Waye et al 1987; Rudd et al 2006). When centromeric tandem repeat sequences of different species are compared, sequence

similarity appears limited to short evolutionary distances (Henikoff et al 2001; Meraldi et al 2006). In fact, specific DNA sequences are probably dispensable for centromere function in most eukaryotes, as kinetochore proteins in diverse organisms can assemble on non-centromeric sequences (Amor, & Choo 2002; Nasuda et al 2005; Lam et al 2006; Wade et al 2009; Shang et al 2010; Locke et al 2011). In humans, these "neocentromeres" have been found through karyotype analysis and can arise at many different loci (Marshall et al 2008). In some animals and plants, individual chromosomes — or even the entire chromosome complement — may lack high-copy tandem repeat arrays (Nasuda et al 2005; Wade et al 2009; Shang et al 2010; Locke et al 2011) and in rare cases centromere repeat sequences differ between chromosomes (Neumann et al 2012; Gong et al 2012) The epigenetic nature of centromere location may be explained by the fact that kinetochores assemble on nucleosomes containing a centromere-specific histone H3 variant, CENH3 (CENP-A in human). Extreme cases of kinetochore protein assembly on diverse sequences are seen in polycentric (Neumann et al 2012) and holocentric chromosomes (Dernburg 2001). The former has a single very large primary constriction that contains three-to-five CENH3 foci (Neumann et al 2012), whereas the latter has CENH3-bound sequences and microtubule attachment sites along the entire length of mitotic chromosomes (Gassmann et al 2012). Despite their dispensable nature, the presence of tandem repeats at the centromere locus of most animals and plants suggests that they serve a function.

centromere DNA in animals and plants have so far focused on single organisms or on small clades (Elder, & Turner 1994; Meraldi et al 2006) and few review articles have been dedicated to a broad survey of tandem repeats (Palomeque, & Lorite 2008). No conserved motif has been found for centromere DNA except in small clades (for example, the CENP-B box found in

mammalian centromeres (Masumoto et al 2004)). Are there shared properties among centromeric tandem repeats from diverse animals and plants? In *Saccharomyces cerevisiae* and closely related yeast species, short centromere DNA sequences evolve three times faster than other intergenic regions of its genome (Bensasson et al 2008; Bensasson 2011). How rapidly do centromere tandem repeats evolve and which molecular processes govern their evolution? We performed a survey of tandem repeats in a large and phylogenetically diverse set of animal and plant species in order to address these questions.

Conventional methods used to identify centromeric tandem repeats, particularly CENH3 chromatin immunoprecipitation, are labor intensive and thus difficult to do on a large scale. In this paper, we identified and quantified the most abundant tandem repeats from 282 animal and plant species using a newly developed bioinformatic pipeline. Our method can utilize shotgun whole genome sequence (WGS) data from various sequencing platforms with varying read lengths, including Sanger, Illumina, 454, and Pacific Biosciences. Candidate centromere repeat sequences were characterized by a seemingly unbiased nature. Repeat monomers varied widely in length, GC composition and genomic abundance. Despite great differences in sequence composition, centromere DNAs appeared to evolve by expansion and shrinkage of arrays of related repeat variants (the “library” hypothesis (Plohl et al 2008)). Using Pacific Biosciences (PacBio) single molecule real-time sequencing to span many contiguous monomers, we characterized the mixing of repeat variants within a single array and the presence of higher-order repeating units. Our data greatly broaden the phylogenetic sampling of centromere DNA, putting evolutionary conclusions about this fast evolving chromosome region on a firmer footing.

Results

A bioinformatic pipeline to identify candidate centromere tandem repeats.

Centromere DNAs in most animals and plants share two distinctive properties: the presence of tandem repeats, and their extremely high repeat abundance (often >10,000 copies per chromosome). Therefore, we hypothesized that the most abundant tandem repeat in a given genome would be the prime candidate for the centromere repeat (our method is not designed to find centromere-specific retrotransposons or chromosome-specific repeat sequences). To find such sequences *de novo* from whole genome shotgun (WGS) sequence data, we developed a bioinformatic pipeline that identifies tandem repeats from a variety of sequencing technologies with different read lengths (see Methods) (Fig. 1A). For example, the 171 bp human centromere repeats (Rudd et al 2006) were identified from Sanger reads, the ~1,400 bp Bovidae repeats (Gaillard et al 1981; Płucienniczak et al 1982; Taparowsky, & Gerbi 1982) were identified from PacBio reads (Supplementary Fig. S5), and the 728 bp monkeyflower (*Mimulus guttatus* (Fishman, & Saunders 2008) were identified from assemblies of Illumina reads. In each case,the most abundant tandem repeat unit was considered to be the candidate centromere DNA.

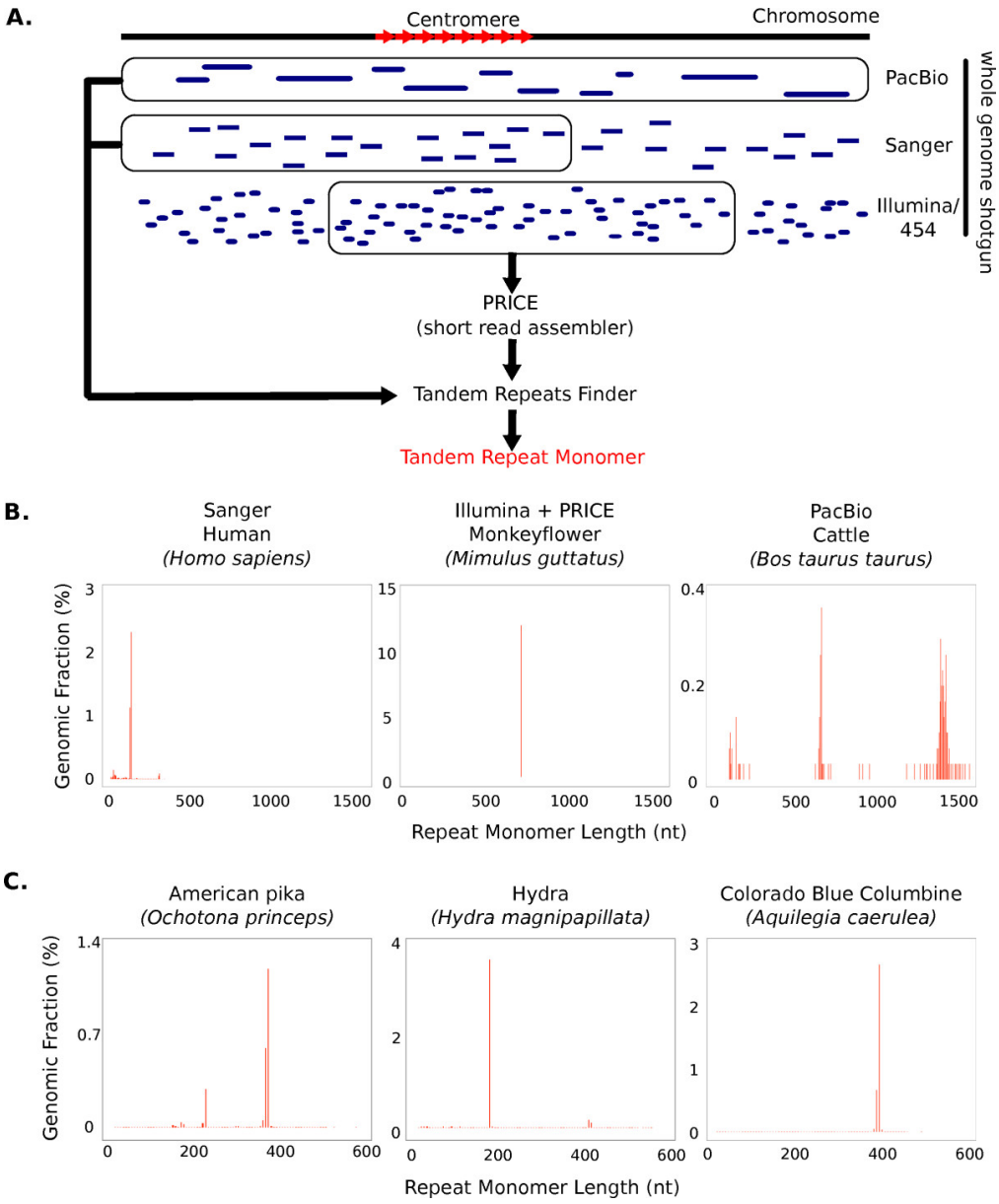


Figure 5-1. A bioinformatic pipeline to identify candidate centromere DNAs based on their tandem repeat nature and abundance.

A. Random shotgun sequences from a variety of platforms can be used to identify the most common tandem repeat monomer. Sanger and PacBio reads are usually long enough to contain multiple copies of a tandem repeat. Illumina and 454 reads are generally too short, and must be assembled to create longer sequences. Tandem repeat monomers were identified by Tandem Repeats Finder (TRF).

B. Identification of known centromere tandem repeats from three species. The human centromere repeat is 171 bp in length. The 728 bp monkeyflower centromere repeat is too long to be found in Sanger reads, but a PRICE assembly of Illumina reads reveals the known repeat . The 1,419 bp cattle centromere repeat and a less abundant 680 bp tandem repeat were directly identified from PacBio reads. Note that the graph for monkeyflower has no background of low abundance tandem repeats because these were not assembled by PRICE.

C. Three examples of *de novo* identification of centromere tandem repeats. Sanger WGS reads from the American pika, Hydra, and Colorado Blue Columbine revealed 253 bp, 183 bp, and 329 bp repeat monomers, respectively.

Validating the bioinformatic pipeline by identifying known centromere tandem repeats

To validate candidate centromere tandem repeats, we compared our results to sequences described in the literature (Supplementary Table S2). Centromere DNAs have been characterized by restriction enzyme-based methods (e.g. laddering on ethidium bromide-stained gels) combined with fluorescence in situ hybridization (FISH), and by chromatin immunoprecipitation (ChIP) with antibodies raised against a kinetochore protein (typically the centromere-specific histone CENH3). Overall, centromere DNA sequences have been described from 43 of the 282 species in this study. In 38 out of 43 cases, we identified a similar repeat to that reported in the literature (Supplementary Table S2). In the case of opossum (*Monodelphis domestica*) and elephant (*Loxodonta africana*), centromere repeat monomers are believed to be very long (528 and 936 bp respectively) (Alkan et al 2011) and therefore cannot be found using Sanger reads. We lacked suitable Illumina or 454 data to allow assembly of long tandem repeats from these species, and did not have PacBio data to find long repeats directly. Potato and pea are unusual in that centromere repeats differ across chromosomes (Neumann et al 2012; Gong et al 2012), with some potato chromosomes lacking tandem centromere repeats entirely (Gong et al 2012). These repeats are too diverse and too long to be identified by our pipeline (upper limit of 2 kbp). Other discrepancies between our candidate centromere repeats and published sequences may be explained by the fact that many previous studies used experimental methods that did not quantify all tandem repeats in the genome (see Supplementary Table S2 for a per species explanation).

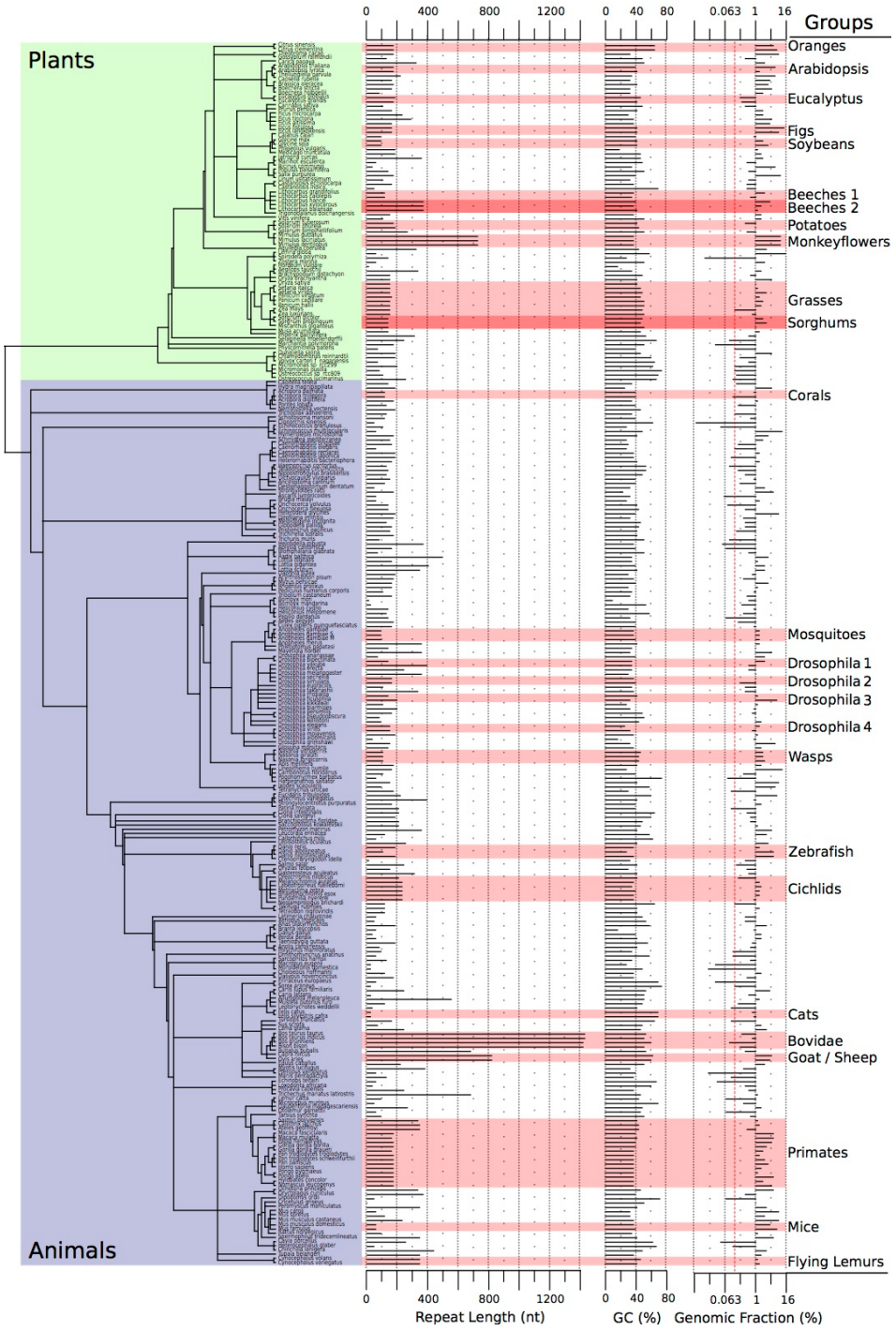


Figure 5-2. Centromere tandem repeat details from diverse animal and plant genomes.
The phylogenetic relationships between 282 species (204 Animalia and 78 Plantae) are shown. For each species, the figure shows tandem repeat length, GC content, and genomic fraction (log 2 scale) for the (candidate) centromere repeat monomer. Taxonomic relationships were derived from the NCBI taxonomy website. Approximately one third of the species (84 out of 282) could be clustered into 26 groups (light red horizontal bars) that exhibited sequence similarity of the tandem repeat monomer within each group. No sequence similarity was found outside these groups, or between them. The most distantly related species within a group diverged about 50 million years ago.

In limited cases, an assembled reference genome can assist in identifying a bona fide centromere tandem repeat. As expected for a true centromere DNA sequence, the 1,419 bp repeat from cattle is generally clustered into one large array on all 30 chromosomes in the UMD3.0 genome assembly (Zimin et al 2009). These putative centromere arrays contain hundreds of repeat copies (notably, secondary arrays elsewhere in this genome assembly contain only 5–10 copies of the monomer).

CENH3 ChIP followed by sequencing is the most definitive method to confirm that a given sequence underlies the functional kinetochore. Only 13 species out of the 43 had CENH3 ChIP-seq data, and our method correctly identified the published centromere tandem repeat in 10 out of 13 of these cases. The three exceptions were opossum, elephant, and potato where we lacked appropriate sequencing reads to find long tandem repeats (opossum and elephant) or the tandem repeats were too diverse (potato). In summary, our bioinformatic pipeline identified the correct centromere tandem repeat in the large majority of cases where experimental data was available.

In two cases, the most abundant tandem repeat was not the known centromere DNA sequence. In the sequenced maize strain B73 (*Zea mays*) (Schnable et al 2009), heterochromatic “knobs” contain highly abundant tandem repeats that outnumber the centromere tandem repeat CentC (Ananiev et al 2000). Knob number, size, repeat abundance and distribution can differ depending on the particular maize variety analyzed, as repeat abundance is variable between isolates (Dawe et al 2009; Chia et al). A 178 bp tandem repeat is present at the centromere of the Tammar wallaby (*Macropus eugenii*), but this sequence was only the third most abundant tandem repeat in our analysis (Carone et al 2009). By mammalian standards, Tammar wallaby centromeres are unusually small (~450 kbp per chromosome), and tandem repeats make up a minority of this

chromosome region because it is also populated by a centromere-specific retroelement (Carone et al 2009).

Candidate centromere tandem repeats from many uncharacterized animal and plant species

To detect candidate centromere repeats**,** we analyzed a total of 282 species, comprising 78 plants and 204 animals spanning 16 phyla (Fig. 2, Supplementary Table S1). Sanger, Illumina, and 454 sequences were obtained from public databases, and we also performed our own PacBio sequencing. The WGS data included 171 species from Sanger sequencing, 132 from Illumina, 13 from 454, and 9 from PacBio (Fig. 1C). For the 37 species which had both Sanger and assembled Illumina data, both data types yielded the same candidate centromere repeat in the majority of cases (28 out of 37) (Fig. 1C). In most cases where analysis of unassembled Sanger reads revealed a different repeat to Illumina data, individual Sanger reads were too short to find the long repeat monomers (see Supplementary Table S4 for a per species explanation).

Many species whose centromere DNAs had not been previously characterized showed a single tandem repeat whose abundance was much greater than all other tandem repeats in the genome. For example, the American pika (*Ochotona princeps*), *Hydra* (*Hydra magnipapillata*), and Colorado Blue Columbine (*Aquilegia caerulea*) had candidate centromere DNAs of 341 bp, 183 bp, and 329 bp respectively (Fig. 1C).

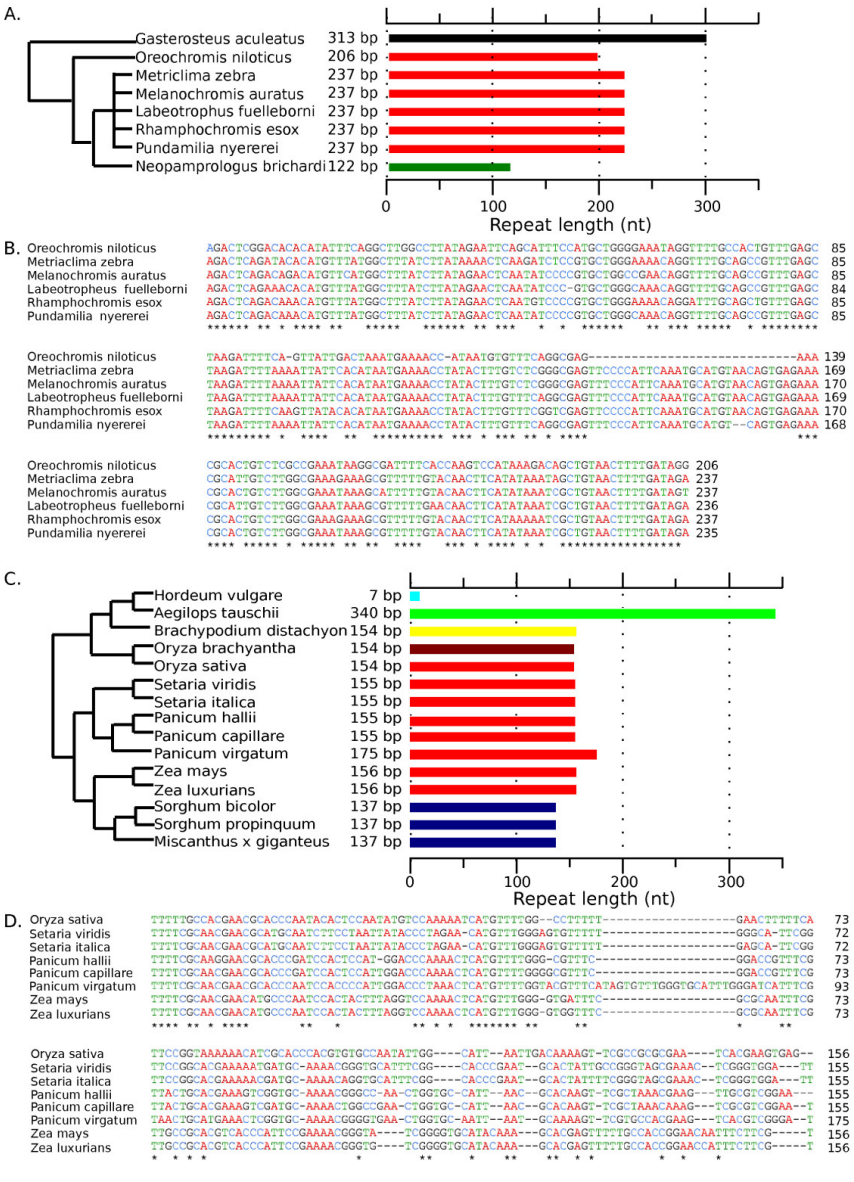


Figure 5-3 Evolution by indel acquisition and coexistence of repeat variants support the 'library' hypothesis. **(a)** Candidate centromere repeat sequences of eight cichlids were analyzed for interspecies sequence similarity. The Princess cichlid *Neolamprologus brichardi* lacked centromere repeat similarity with its sister clade of Lake Malawi cichlids (shown in orange, and also including Nile tilapia). **(b)** Sequence alignment of candidate centromere repeats shows that Nile tilapia (*Oreochromis niloticus*) has a deletion relative to other cichlid species. **(c)** Candidate centromere repeat sequences of 15 grass species were analyzed for interspecies sequence similarity. We found two groups of species with centromere repeat sequences that were similar. The closely related *Sorghum* and *Miscanthus* species have similar 137 bp repeats (blue bars). The clade shown by red bars contains *Oryza sativa* (rice), which is relatively distant from the other species that have similar centromere tandem repeats (red bars). Although the centromere repeats of *Oryza brachyantha* and *Brachypodium distachyon* have repeat monomer length similar to the orange-highlighted group, no sequence similarity was found between them. Interestingly, no sequence similarity was found between the closely related *Zea* species and *Sorghum* species or between *Oryza* species and *Brachypodium, Aegilops*, or *Hordeum*. **(d)**Sequence alignment of candidate centromere repeats from eight grass species. Switchgrass (*Panicum virgatum*) is distinguished by the presence of a short insertion relative to the other species.

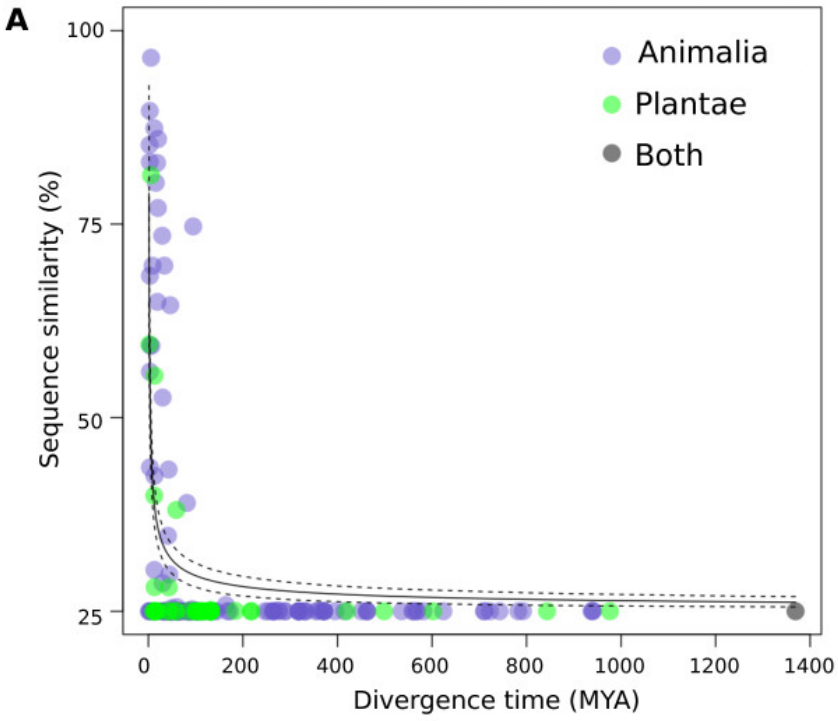
The most accurate measurements of centromere tandem repeat array size in animals and plants are generally in the range of ~500 kbp to several Mbp (Waye et al 1987; Hosouchi et al 2002; Lee et al 2005; Carone et al 2009). Although estimated repeat abundance is subject to several experimental biases, we calculated the average amount of repeat per chromosome, and most organisms in our survey were estimated to contain hundreds of kbp (Supplementary Table S1). Since our analysis was based on WGS data, it is not possible to detect chromosome-to-chromosome variation (Wade et al 2009; Shang et al 2010; Locke et al 2011).

How rapidly do centromere DNA sequences evolve?

An all-vs-all BLAST search of our consensus repeats revealed that sequence conservation was limited to only very closely related species. We found 26 groups of species that showed sequence similarity between centromere tandem repeats (Fig. 2, Supplement Fig. S1). Notable groupings of species with substantial sequence similarity included the primates (Supplement Fig. S2), cichlids (Fig. 6A-B) and grasses (Fig. 6C-D).

The well-studied nature of human centromeres, and the availability of many closely related species, make primates an excellent clade to illustrate the evolution of centromere DNAs (Rudd et al 2006; Horvath, & Willard 2007; Lee et al 2011). Candidate centromeric tandem repeats in primates showed similarity between monkeys and apes (Fig. 3A, Fig. 4A), but these candidate centromere DNAs were unrelated to those in more basal primates (tarsiers and prosimians). We inspected lower abundance tandem repeat sequences from the Tandem Repeats Finder (TRF) output, and no tandem repeat in tarsiers or prosimians was found to have sequence similarity to the primate candidate centromeric tandem repeat. These results reinforce recent findings showing

that the aye-aye (*Daubentonia madagascariensis*) has centromere repeats with no similarity to monkeys and apes (Lee et al 2011).



B Blomberg's K

Clade	Monomer length	GC content	Genomic fraction
Primates	1.852265 (p = 0.001)	1.615503 (p = 0.001)	0.7551643 (p = 0.004)
Grasses	0.414999 (p = 0.427)	0.6775028 (p = 0.158)	0.361093 (p = 0.645)

K > 1 means more conserved than expected
K < 1 means less conserved than expected

Figure 5-4 Centromere tandem repeat monomers are conserved only between closely related species. **(a)**Percentage identity between candidate centromere repeat sequences plotted against estimated divergence time. We averaged percentage identity between comparisons to generate a single value for each node in the phylogenetic tree (Figure 2). To accommodate unresolved relationships, we repeated the analysis on random resolutions of the tree. One such analysis is shown (quantitative results were very similar between analyses).**(b)** For primates and grasses, the phylogenetic signal was tested using Blomberg's *K* analysis for three different parameters: repeat monomer length, repeat monomer GC content and genomic abundance. In primates both repeat length and GC content were more conserved than expected ($K > 1$), whereas genomic abundance was less conserved than expected by a model of Brownian evolution ($K < 1$). Though $K < 1$ for all three traits in the grasses, none were significantly different from 1. *P*-values are shown in brackets.

Cichlid fish are another clade in which we identified both conservation and rapid divergence of centromere repeats. Lake Malawi cichlids and the Nile tilapia (*Oreochromis niloticus*) had candidate centromere DNAs that shared 78% sequence similarity, although tilapia diverged from other cichlids 45 MYA. The Princess cichlid *Neolamprologus brichardi* (from Lake Tanganyika) had a candidate centromere repeat with no sequence similarity to either the Lake Malawi cichlids or Nile tilapia, though *Neolamprologus* diverged from Lake Malawi cichlids only 30 MYA. Similar patterns of both conservation and rapid change can be seen in the grasses (Fig. 6C-D). A maize-like centromere repeat can be found in *Panicum*, *Setaria*, and even in a species as distant as rice (*Oryza*), which diverged from maize ~41 MYA. In contrast, sorghum-maize (9 MYA) and *Hordeum-Aegilops* (14 MYA) comparisons show little to no sequence similarity.

To evaluate the rate of sequence evolution across the entirety of our sampled taxa, we assessed the conservation of sequence identity across the phylogeny using a node-averaged comparative analysis (Fig. 3A). We fit a model of exponential decay with divergence, finding that on average sequence identity falls rapidly to background levels (i.e. random 25% identity) after ~50 MYA.

Candidate centromere tandem repeats from 282 animals and plants display no readily apparent conserved characteristics

If centromere DNAs are fast evolving, do their repeat monomers at least possess other conserved properties? As our survey is the broadest phylogenetic analysis of tandem repeats to date, we asked if candidate centromere DNAs from 282 species shared common characteristics. Our analyses showed that this was not the case.

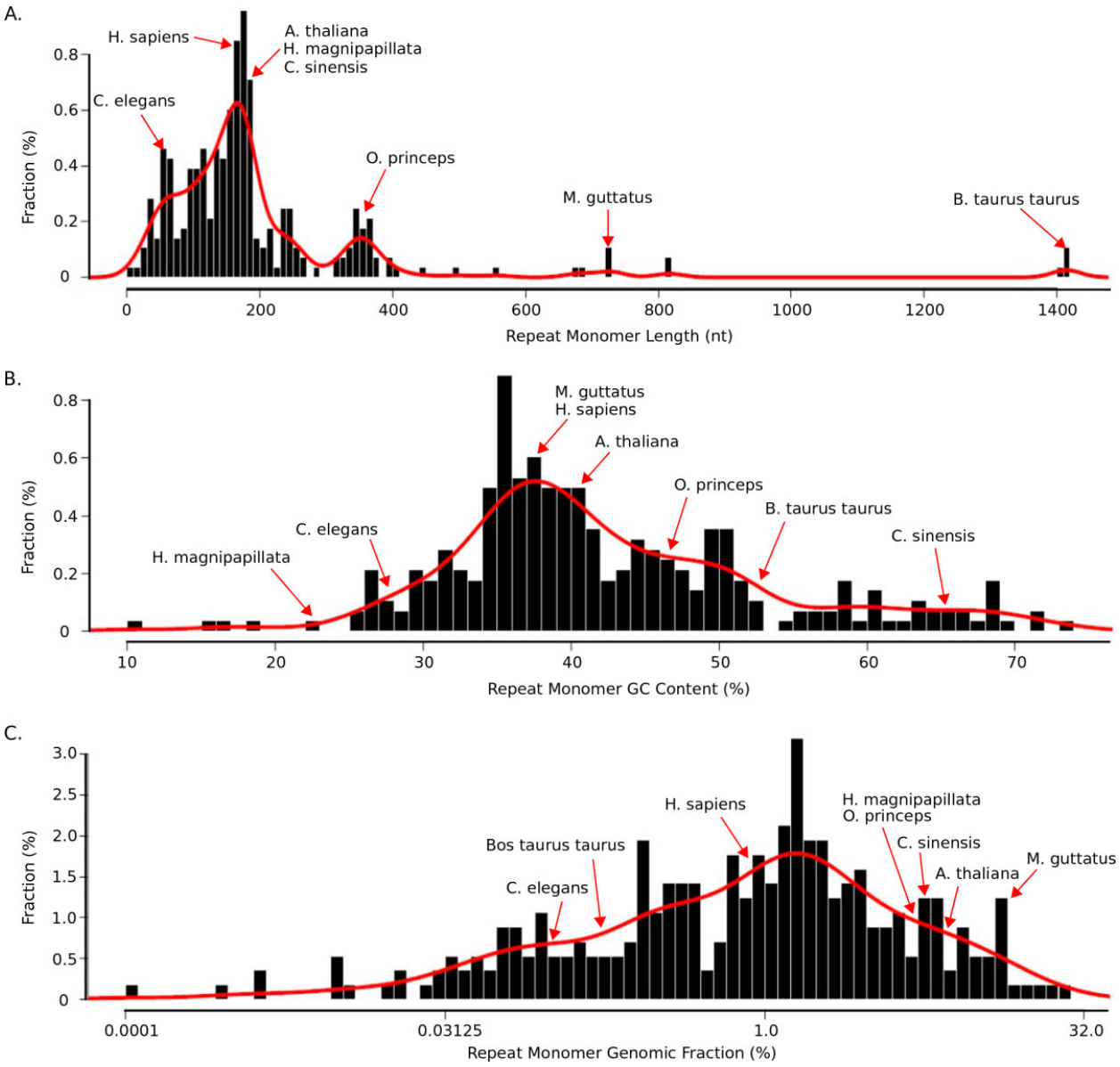


Figure 5-5 Centromere tandem repeats lack conserved sequence properties. **(a-c)** No strong bias was observed in distribution of centromere repeat monomer length (a), GC content (b), or genomic fraction (c).

First, centromere tandem repeat monomer length is not conserved. As CENH3 is essential for kinetochore nucleation, it has been hypothesized that centromere repeat monomers may tend to be about the size of one nucleosomal DNA (Willard 1990; Shelby et al 1997), as is seen with human (171 bp), *Arabidopsis thaliana* (178 bp), and maize (156 bp) centromere DNAs. This is clearly not a universal rule, as some centromere tandem repeat monomers are much shorter and

longer than nucleosomal sizes (e.g. soybean at 92 bp (Gill et al 2009; Tek et al 2010) and cattle at 1,419 bp (Taparowsky, & Gerbi 1982)) (Fig. 4A). Plant species tended to have repeat sequences with lengths of approximately 180 bp, whereas we found a broader length distribution in animals. Modest trends in our data, however, may reflect sampling bias in the species for which WGS data was available in public archives rather than biologically meaningful preferences in centromere tandem repeat length.

Second, GC content of centromere tandem repeats is not conserved. Based on limited analysis of animal centromere repeats, it was suggested that centromeric DNA is AT-rich (Henikoff et al 2001). Our analysis of 282 species revealed that centromeric DNA can be very GC-rich (Fig. 4B), although a slight preference for AT-rich tandem repeats was observed in animals. Plant species do not appear to have a preference for AT- or GC-rich centromere tandem repeats.

Third, the abundance of centromere tandem repeats varies widely (Fig. 4C). We calculated repeat abundance by finding the proportion of reads that matched the repeat monomer (using a set of randomly sampled reads, see Methods and Supplemental Methods). Tandem repeat abundance can be compared between species, but is subject to variability introduced by different library construction protocols at particular sequencing centers, and by biases in the way different sequencing technologies capture high-copy repeats. We compared repeat abundance of 40 species for which there was sequence data from multiple sequencing technologies. On average, sequences derived from Illumina sequencing had higher estimated repeat abundances compared to Sanger, 454 or PacBio data. For most species we estimated that at least 0.5% of the genome was comprised of the candidate centromeric tandem repeat, but the overall percentage was highly variable (Fig. 2, Fig. 4C).

Simple nonphylogenetic correlations found no relation between repeat length, GC content, and genomic fraction of candidate centromere tandem repeats (Supplementary Fig. S3). Similarly, we did not find a correlation between these factors and genome size, genome-wide GC content or chromosome number.

To explicitly test for conservation of sequence characteristics at a finer phylogenetic resolution, we searched for signals in phylogenetic trees that represented the grass and primate clades (Fig. 3B). Both clades are of a similar age (40–45 MYA for the most divergent species) and show substantial sequence similarity among taxa. We calculated Blomberg's *K* statistic (Blomberg et al 2003), a measure of phylogenetic conservation, for various tandem repeat characteristics. The *K* statistic indicates the amount of phylogenetic signal in the data. Values of *K* greater than one suggest that related taxa resemble each other more than would be expected given a null model in which the trait evolves along the tree according to Brownian motion. Values of *K* less than one are observed when related taxa are less similar than expected under the null model. Although repeat monomer length, GC content, and genome fraction all had values of $K<1$ in the grasses, none were statistically significant. In contrast, values for all three characteristics were significantly different from the null model in primates, with GC content and repeat monomer length showing $K>1$ and repeat abundance $K<1$. These data suggest that individual clades likely differ in terms of their tendency for closely related species to have centromere repeats that share conserved sequence characteristics.

Which species lack candidate high-copy tandem repeats at their centromeres?

Which animal and plant genomes lack high-copy centromere tandem repeats? The nematode *Caenorhabditis elegans* is a useful negative control for measuring tandem repeat abundance (see red dashed line in the genomic fraction column of Fig. 2), because it has holocentric chromosomes and has been reported to lack centromere tandem repeat arrays in its genome (Gassmann et al 2012). In total 41 species had a lower abundance of tandem repeats than in *C. elegans*, and these could be assumed to lack high-copy centromere tandem repeats. Nine of these species are known to be holocentric (Melters et al 2012) and are not expected to have large tandem arrays. Fungi such as *Saccharomyces cerevisiae*, *Candida albicans*, and *Schizosaccharomyces pombe* have small genomes and do not contain high-copy tandem repeat arrays at their centromeres (Meraldi et al 2006). Many of the other genomes that exhibited low tandem repeat abundance also had small genomes including 7 species of basal plants (green algae, moss, and liverwort) and 11 animals. A few species exhibited low tandem repeat abundance despite possessing large genomes (hedgehog (*Erinaceus europaeus*), tenrec (*Echinops telfairi*), seal (*Leptonychotes weddellii*) and dolphin (*Tursiops truncatus*). This may be due to these species having large repeat units that could not be identified in the available Sanger reads. While a definitive answer is not possible yet, it appears that species lacking large tandem arrays tend to have holocentric centromeres or small genomes.

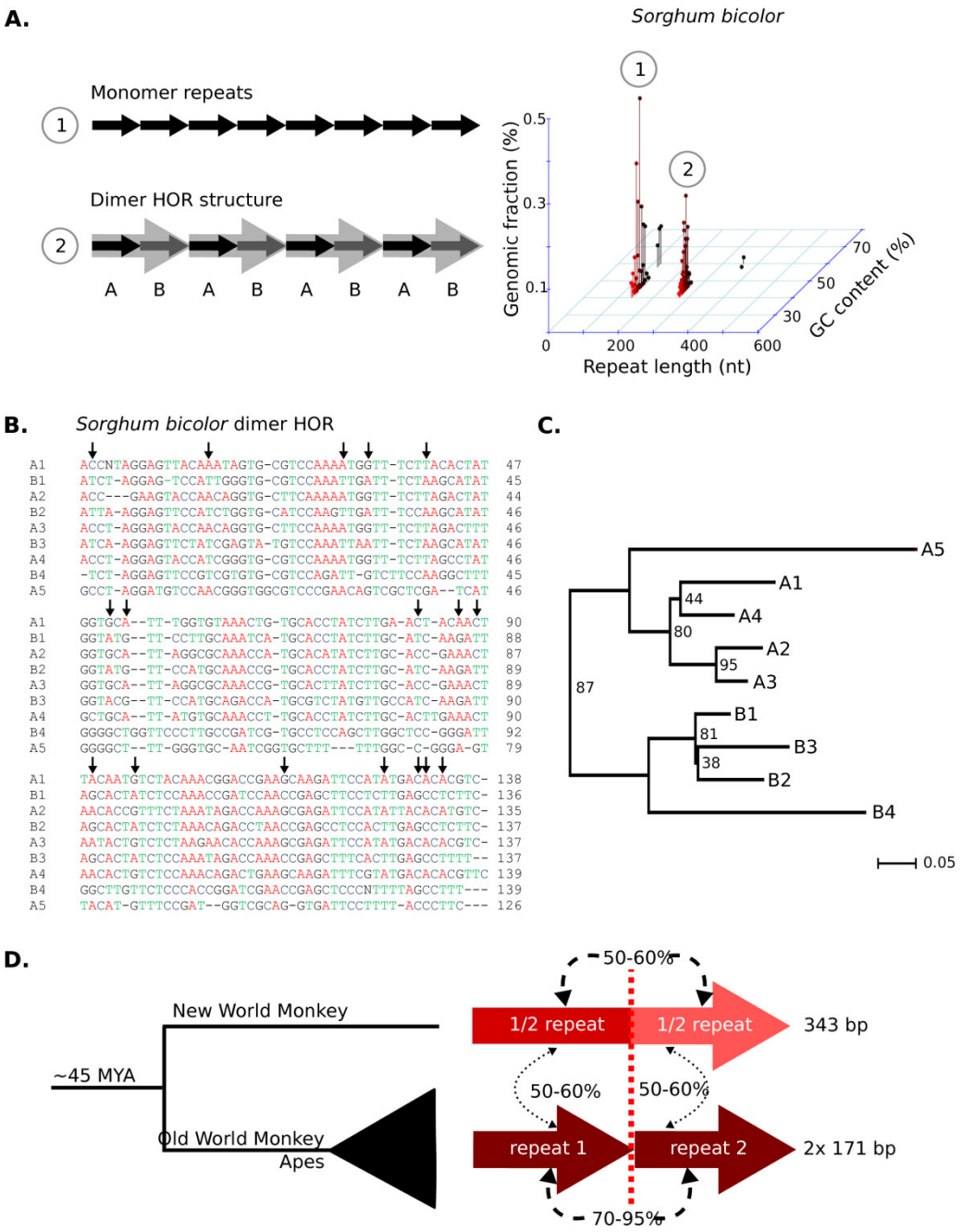


Figure 5-6 Higher order repeat structures are prevalent in diverse animals and plants. **(a)** Graphical representation of higher order repeat structure compared to simple monomer repeats. In the higher order repeat, two variants, A and B, form a single dimer repeat that is repeated in tandem. When plotting repeat monomer length by GC content by genomic fraction, two distinct peaks are seen for *Sorghum bicolor*. The second peak (2) is exactly double the length of the first peak (1). **(b)** Sequence alignment of repeat units from a single *Sorghum bicolor* Sanger read that exhibits a higher order repeat structure consisting of an AB dimer. The arrows point to SNPs unique for either the A or B repeat of the dimer. **(c)**Neighbor joining analysis showing grouping of A and B repeats from sequence alignment in B. Bootstrap numbers are shown. **(d)** Higher order repeat structures can lead to novel centromere repeats. In New World monkeys, the two halves of the 343-bp monomer are weakly related to each other and to the 171-bp repeat in Old World monkeys and apes.

Higher order repeat structure and evolution of novel repeat monomers.

Primate centromeres contain higher order repeat (HOR) structures (Warburton et al 1993; Rudd et al 2006), in which multiple repeat monomers with specific polymorphisms form a unit that itself is repeated (Fig. 5A). HOR structure was easiest to observe in Sanger data, which combines relatively long reads with high sequence accuracy. We used the output from Tandem Repeats Finder (TRF) (Benson 1999) to identify higher order repeat structures among Sanger sequences from the NCBI trace archive. TRF reports both the repeat monomer, as well as repeating units carrying multimers of the monomer that may represent HOR structure. TRF-defined repeats that occupied approximately the same coordinates within a single read were compared to identify whether longer repeats were dimers of the basic monomer. In true HOR structures, the percentage identity between adjacent multimers should be much higher than between individual monomers (TRF should also report higher scores for the repeats with the longer monomer). Therefore, we filtered TRF output to detect these multimers that had both a higher percentage identity and a higher TRF score compared tothe monomeric repeat that spanned the same coordinates.

Clear cases of HOR structure were identified in 76 of the 171 species with Sanger data. Phylogenetic trees constructed with individual monomers extracted from a single read showed that the “A” monomers and “B” monomers from a dimeric “AB” structure that clustered separately (Fig. 5B-C), confirming that the AB structure indeed represented a HOR unit. HOR structure has been previously described in primates, but our analysis show that it is widespread across both plant and animal kingdoms. The capability to detect HOR units is limited by Sanger read length, so shorter repeat monomers were more likely to display HOR structures. We rarely

identified HOR structures that had three or more copies of a repeat monomer, because such structures require at least six monomers to be found in a single Sanger read.

Can HOR structure result in evolution of a new centromere tandem repeat? The centromere repeat monomer has only been reported for one New World Monkey and its length (343 bp) is essentially double the size of human alpha satellite (Cellamare et al 2009). We extended this analysis to three New World Monkeys and fifteen Old World Monkeys and apes (Fig. 5D, Supplementary Fig. 2). All Old World Monkeys and apes had a 171 bp candidate centromeric tandem repeat, whereas New World Monkeys had a 343 bp candidate centromeric tandem repeat. If the 343 bp repeat is split into two equal halves and aligned to the 171 bp repeat, both halves align, but each has specific polymorphisms and indels (Fig. 5D, Supplementary Fig. S1). These data suggest that in the New World Monkey clade, a doubled version or dimer of the ancestral 171 bp repeat became the dominant centromeric tandem repeat. Such patterns of evolution are likely to be general, as they depend only on acquisition of polymorphism and a particular pattern of recombination within a repeat array (Smith 1976).

Where HOR structure is present, it means that our calculated values for the abundance of candidate centromere repeats are most likely underestimates. A notable example of this occurs in the gorilla genome (*Gorilla gorilla gorilla*). We correctly identify the 171 bp centromere repeat as the most abundant repeat and this accounts for approximately 1.3% of the genome. However, we also identify a separate, but related, 340 bp repeat which represents a doubled version of the 171 bp repeat. This second repeat accounts for a further 1.2% of the genome, showing that dimeric HOR structure may be especially common among gorilla centromere repeats.

Coexistence of related repeats support the "library" hypothesis

The "library" hypothesis aims to explain how centromere DNA evolves so rapidly (Plohl et al 2008). This hypothesis assumes that variants of centromere tandem repeats co-exist within the same tandem arrays. Over time, the abundance of particular variants stochastically changes through both expansion and shrinkage (Ma, & Jackson 2006; Ma, & Bennetzen 2006), resulting in replacement of the most abundant variant with a different variant. Centromere repeat variants could arise by point mutation, deletion, insertion or by mixing of different parental sequences during allopolyploid formation (in all cases, a process such as gene conversion would be required to transfer variants between chromosomes) (Hemleben et al 2007). Are there cases in our data set that support the "library" hypothesis? Specifically, do repeat variants differ and are there cases where such a repeat was able to colonize a genome and replace the original monomer?

Several Lake Malawi cichlids contained a 237 bp candidate centromeric tandem repeat, whereas the closely related Nile tilapia contained a shorter repeat of 206 bp (Fig. 6A-B). However, the Nile tilapia did contain a less abundant, 237 bp repeat that was similar to the Lake Malawi cichlid repeat (Supplementary Fig. S4). This suggests that the centromere tandem repeat in the common ancestor of Lake Malawi cichlids and Nile tilapia was replaced by a related sequence (having either an insertion or deletion of 29 bp) in one of the two modern clades.

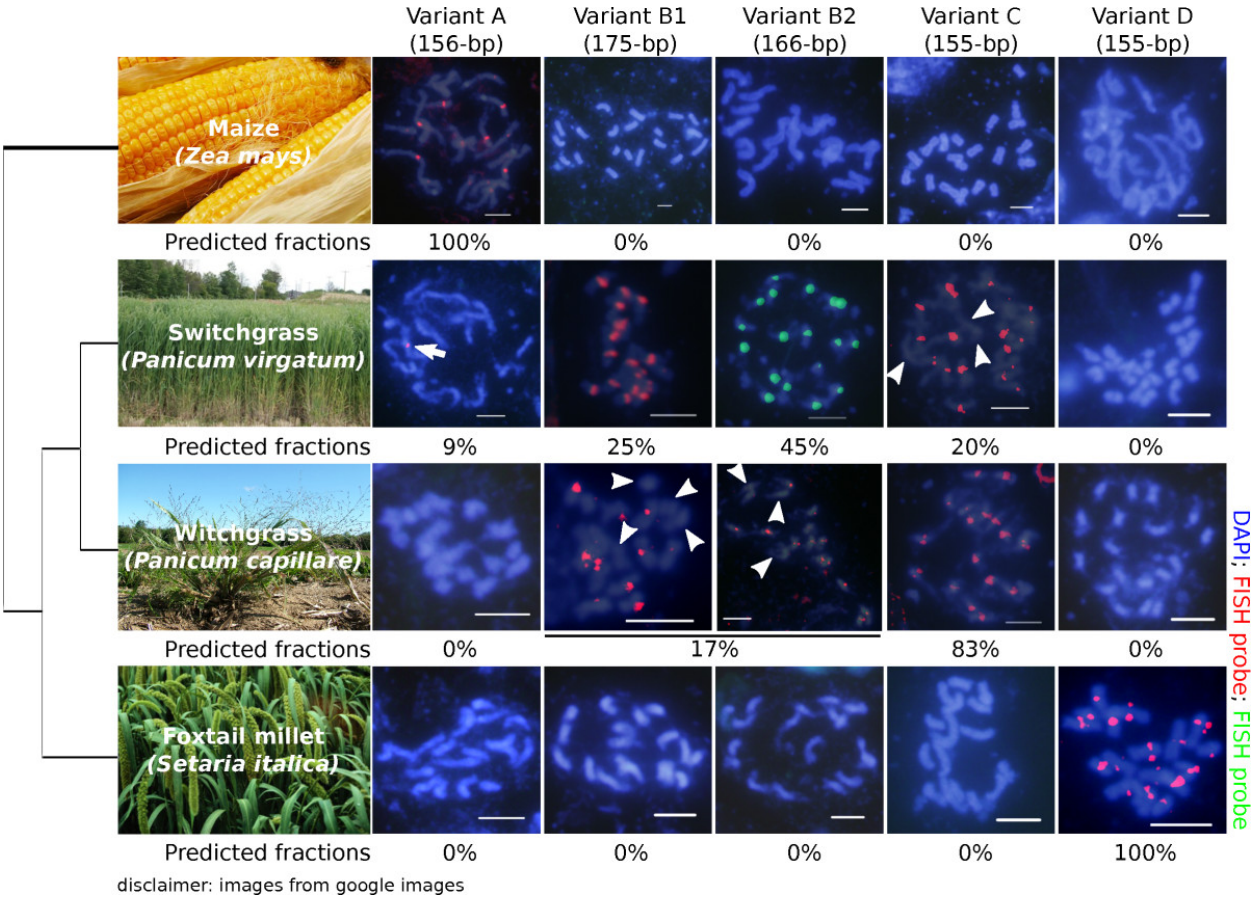


Figure 5-7 Chromosomal localization of repeat variants in grasses is consistent with repeat abundance measured by our bioinformatic pipeline. Chromosomal localization of the different grass repeat variants (maize variant A, switchgrass variants B1 and B2, witchgrass variant C, and foxtail millet variant D) was determined by FISH on metaphase chromosomes of maize (*Zea mays*), switchgrass (*Panicum virgatum*), witchgrass (*Panicum capillare*), and foxtail millet (*Setaria italica*). Switchgrass variants B1 and B2 differ by a 9-bp deletion, whereas both variants differ from maize, witchgrass and foxtail millet by a 20-bp insertion. Maize and foxtail millet chromosomes hybridized only to variants A and D, respectively. Only one switchgrass chromosome hybridized to variant A (arrow), but variants B1, B2 and C labeled most chromosomes (arrowheads indicate chromosomes that showed weaker hybridization to variant C). Witchgrass chromosomes were most consistently labeled by variant C, but showed chromosome-specific hybridization to variants B1 and B2, consistent with their lower abundance in the genome. In all cases the FISH probes hybridized to the primary constriction, which is indicative of centromere localization. The percentages below the panels represent computational predictions of repeat variant ratios in each species.

More support for the "library" hypothesis was seen in the grasses (family *Poaceae*); this was the largest plant clade in our dataset that exhibited sequence similarity among most of its members. The modal length of repeat monomers in grasses was 156 bp, but deletions and insertions were found in several species (an 80 bp conserved motif between rice and maize was previously noted

within this sequence (Lee et al 2005)). Eight of the fifteen grass species had candidate centromere repeats that displayed no similarity to the common 156 bp sequence (Fig. 6C-D). We then searched our data for less abundant tandem repeats related to the dominant repeat monomer. Sanger sequence data for four grass species revealed distinct centromere tandem repeat variants. Maize and foxtail millet (*Setaria italica*) only contained one variant each (variants A and D respectively), witchgrass (*Panicum capillare*) had two variants (B and C) and the switchgrass genome contained three variants (A, B, and C). Variant B itself consists of two distinct repeats, one of 175 bp (variant B1) and another of 166 bp (variant B2). B2 differs from B1 by the deletion of 9 bp, but these two subvariants are otherwise very similar in sequence, so we consider them as one variant (variant B). The existence of related repeat variants in switchgrass and witchgrass is similar to our observations in Lake Malawi cichlids and Nile tilapia, and both these cases further support the "library" hypothesis (Plohl et al 2008).

Next we asked if switchgrass repeat variants occupied the same tandem repeat arrays by using computationally derived repeat monomers as probes in fluorescent in situ hybridization (FISH) experiments. FISH analysis confirmed that these repeat variants were found at centromeres (Fig. 7). Variants not found in a given genome did not stain chromosomes from that species, showing that our hybridization conditions were specific. The variant A probe only hybridized strongly to one switchgrass chromosome. Variant B in switchgrass was comprised of two repeats (B1 = 175 bp and B2 = 166 bp) and FISH experiments revealed that all switchgrass chromosomes showed hybridization to variants B1, B2 and C, but with differing hybridization intensities (Fig. 8A-B). These data indicate that specific chromosomes harbor different amounts of particular repeat variants, again suggesting that repeat arrays can grow and shrink over evolutionary time.

Pacific Biosciences sequencing reveals that switchgrass repeat arrays are homogeneous and contain long higher order repeat structures.

Centromere repeat variants in switchgrass were found on the same chromosomes using FISH (Fig. 8A-B), but the resolution of these experiments could not distinguish large homogeneous arrays of two variants (in close proximity) from arrays that showed more significant mixing of repeat variants. Theoretical simulations predict that an array of polymorphic repeats can become rapidly homogenized by unequal crossing over (Smith 1976). Conversely, gene conversion can introduce novel variants into the middle of a repeat array. To determine the degree to which variants were mixed in a given array, we used the PacBio sequencing platform, which yields much longer reads (up to 16.5 kbp) than other sequencing technologies (Fig. 8C) (Schadt et al 2010). As PacBio sequencing has a very high indel rate, we focused on repeat variants that differ by indels of at least 9 bp. Switchgrass genomic DNA was sequenced on four runs of the PacBio RS system using the C2 chemistry and a ~10 kbp insert library (see methods for details). All switchgrass chromosomes stained positive for both variant B1 and B2 FISH probes and both repeat variants were present in the PacBio sequence data. However, individual PacBio sequencing reads never contained a mixture of the two variants..This shows that centromere repeat arrays in switchgrass are composed of long homogeneous arrays variants, but that these arrays are mixed together on the same chromosome.

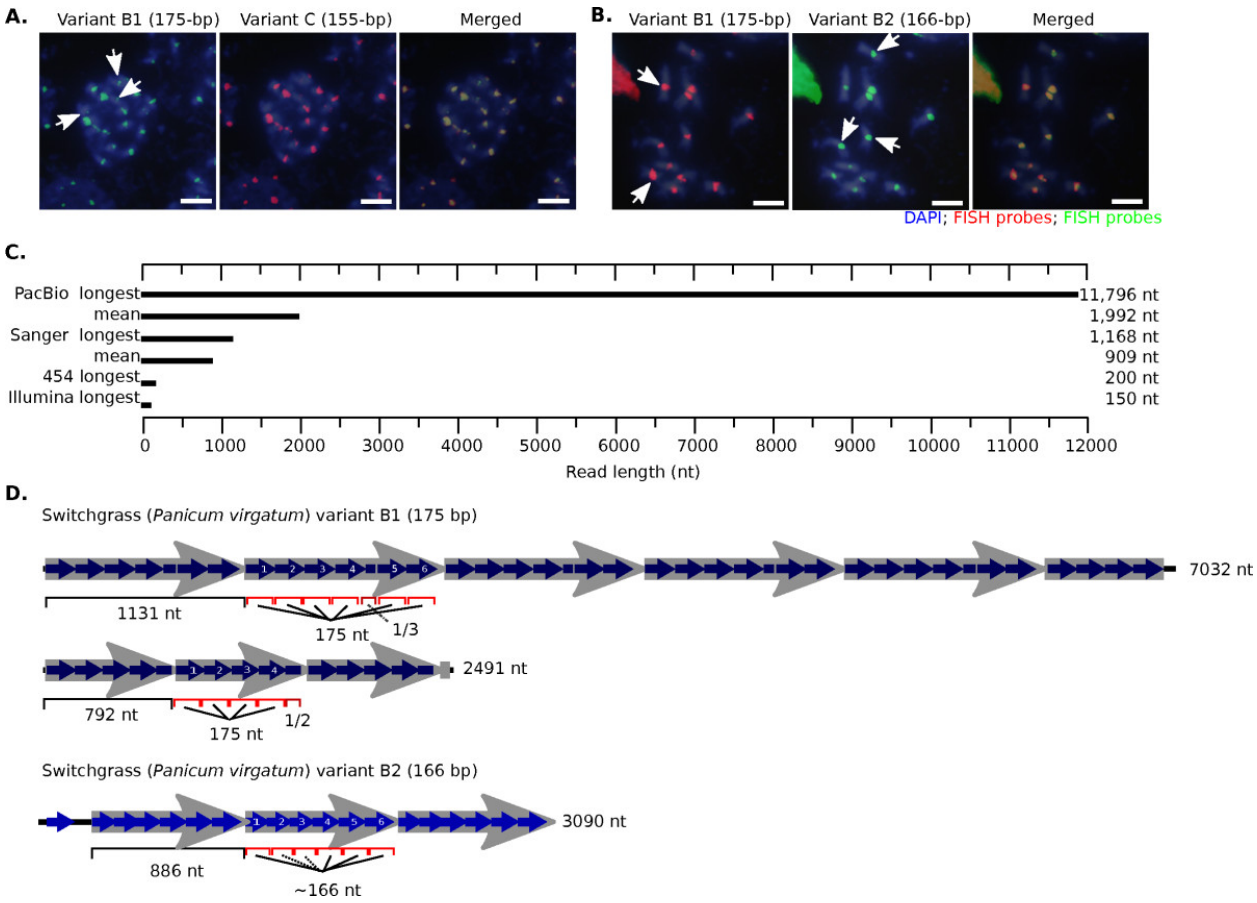


Figure 5-8 Pacific Biosciences sequencing shows homogeneity of repeat arrays and detects long higher order repeat structures. **(a)** Switchgrass variant B1 hybridized to all switchgrass chromosomes, whereas witchgrass variant C hybridized to all but three switchgrass chromosomes. The three chromosomes that only showed hybridization of variant B1 (arrows) were stained green (see merged). **(b)** Although both switchgrass variants B1 and B2 co-hybridize to all switchgrass chromosomes, the hybridization signal showed a chromosome-specific pattern. The arrows highlight chromosomes with stronger hybridization signal for one sub-variant over the other. **(c)** The strength of PacBio sequencing is the extreme length of a small fraction of the reads. In the AP13 switchgrass PacBio sequencing run, the longest inserted sequence was almost 12 kbp in length, although the mean of all the PacBio reads was about 2 kbp. Sanger reads are shorter, but have a more consistent length, whereas both Illumina and 454 reads are very short and very homogeneous in length (longest reads in our study only shown). **(d)** Although no repeat variant mixing was detected in the PacBio reads, several HOR structures were found in longer PacBio reads. These HOR structures consisted of a mixture of complete and trunctated repeats. Two switchgrass variant B1 centromere reads with higher order structure and one switchgrass variant B2 centromere repeat are shown. The 1,131-bp HOR structure consisted of six repeat monomers and a truncated repeat (about one-third the size of 175 bp repeat). In total, five-and-half copies of the 1,131-bp repeat were found within the 7 kbp read. One variant B2-containing read is shown, containing three copies of a 886-bp HOR structure (composed of six 166-bp repeats).

Another benefit of PacBio sequence reads is their ability to detect HOR structure that extends beyond the dimer and trimer structures typically visible in shorter Sanger reads (Fig. 5). We

found a novel pattern of HOR structure in switchgrass centromeres using PacBio sequencing: large repeating units that contain deleted versions of a canonical centromere repeat (Fig. 8D). A 2,491 bp read contained a higher order repeat comprised of 4 B1-type monomers followed by a truncated variant approximately half the size of the B1 repeat. The B1 repeat is 175 bp long, and the HOR repeat is 792 bp, too long to be detected by Sanger sequencing. Similarly, a 7,032 bp PacBio read contained a 1,131 bp HOR repeat made of 6 B1-type monomers and a truncated B1 repeat of 53 bp. In this case, the HOR repeat itself is longer than almost all Sanger sequence reads. This application shows that long reads have the benefit of directly revealing long repeat structures that could previously only be seen through painstaking and indirect assembly strategies or by chromosome-specific cytogenetic methods (Warburton et al 1993; Warburton et al 1996; Rudd et al 2006; Lee et al 2011).

Discussion

The ready availability of whole genome shotgun sequence from a wide variety of eukaryote genomes makes comparative genomics an appealing way to study rapidly-evolving tandem repeat sequences, such as those commonly associated with centromeres. Animals and plants are evolutionarily distant, so previous studies showing the presence of high-copy centromere tandem repeats in these organisms raised the question of whether this was indeed a general property. Recently, bioinformatic methods for identifying centromere tandem repeats have been described, and applied to several previously uncharacterized mammals (Swaminathan et al 2007; Alkan et al 2011; Hayden, & Willard 2012) and plants (Navajas-Pérez, & Paterson 2009). We have performed the largest survey of animal and plant tandem repeats to date, encompassing every

species with sufficient whole genome shotgun sequence in the NCBI trace archive and DDBJ sequence read archive. The bioinformatic methods we used are amenable to every available DNA sequencing technology, making our study expandable as future DNA sequences are generated. In species with previously reported centromere repeats, the most abundant tandem repeat identified in our analysis matched the published sequence in almost every case. The presence of highly abundant tandem repeats in the large majority of species that we analyzed suggests that tandem repeats likely underlie the functional centromere in most animals and plants. Candidate centromere tandem repeats did not share conserved properties such as monomer length, GC content, or common sequence motifs. We found that higher-order tandem repeat structures were prevalent across a broad phylogenetic distribution, as wastheoretical predictions that the tandem repeat nature of centromere DNA in animals and in plants can facilitate the rapid evolution of these sequences (Smith 1976).

As centromeres can form on non-centromeric DNA sequences in both animals and plants, the function of tandem repeats at centromeres is enigmatic (Amor, & Choo 2002; Nagaki et al 2004; Nasuda et al 2005; Marshall et al 2008). Our finding that centromere tandem repeats are common reinforces the argument that they have a functional, albeit subtle role, although careful experiments may be required to detect this *in vivo*. Further evidence for this comes from both evolutionary and functional experiments. Neocentromeres formed during evolution eventually acquire tandem repeats (Ventura et al 2007), and neocentromeres lacking tandem repeats are subtly defective in one human cell culture assay (Bassett et al 2010). It is possible that centromere specification will be a balance between epigenetic and genetic factors in most plants

and animals, although it is clear that epigenetic memory provided by the centromere-specific histone CENH3 is the most important factor.

High-copy tandem repeats have a propensity to form heterochromatin (Pikaard, & Pontes 2007), but it is unlikely that this property alone explains their presence at centromeres. Transposons in pericentromeric regions are also highly heterochromatic, and there is little in the chromatin landscape of large repeat-rich genomes such as maize that distinguishes centromeres from similarly gene-poor regions. Transposons inserted into the tandem repeat arrays of cereals and other plant genomes have not been shown to have a function in centromere biology, although they are bound by CENH3 (Zhong et al 2002; Houben et al 2007; Tek et al 2010) and centromere-specific transposons localize exclusively to the centromeres of close relatives (Tsukahara et al 2012). Most interestingly, the tandem repeats within the CENH3-binding domain of the centromere have significantly different chromatin modifications from typical heterochromatin (Sullivan, & Karpen 2004). In *A. thaliana* and maize, tandem repeats at the functional centromere have been observed to have lower DNA methylation that those at the edge of the repeat array (Zhang et al 2008). Extended chromatin fiber microscopy has shown that centromeres in *Drosophila melanogaster* and humans contain some modifications typical of euchromatin (e.g. lack of H3K9 di- or trimethylation), in addition to those associated with gene silencing (hypoacetylation of H3 and H4) (Sullivan, & Karpen 2004). Tethering a transcriptional silencer to a human artificial chromosome or altering its acetylation/methylation balance can lead to centromere inactivation (Nakano et al 2008; Ohzeki et al 2012). Lastly, it is possible that non-coding RNAs may have a role in centromere function, and transcription of such molecules may

not be compatible with heterochromatic marks (Topp et al 2004; Carone et al 2009; O'Neill, & Carone 2009; Du et al 2010).

dimers (rotational positioning) within CentC repeats, which suggests that CentC repeats could contribute to a highly stable nucleosome arrangment in centromeres (Gent et al 2011). Nucleosome phasing over the entire centromere should be dominated by nucleosomes containing conventional histone H3, as CENH3 nucleosomes bind to only a small fraction of the tandem repeat array. In a phasing model, the aquisition and accumulation of tandem repeat arrays would be fostered by the chromatin arrangement of centromeres. The phenomenon of centromere reactivation, in which a centromere first loses kinetochore-nucleating activity, then regains it, could suggest that tandem repeats encourage centromeric chromatin states. Notably, centromere reactivation has been observed in both maize (Han et al 2009; Sato et al 2012) and possibly in humans (Mackinnon, & Campbell 2011).

Rapid evolution itself may explain the fact that centromere DNA in so many animals and plants is comprised of tandem repeats. A prevailing model to explain fast evolution of centromere DNA sequences and CENH3 is that asymmetric meiosis during oogenesis encourages centromeric drive (Henikoff et al 2001; Malik, & Henikoff 2009). In this model, competition of centromeres for preferential segregation into the single meiotic cell that survives to become the egg can drive rapid sequence evolution. Eventually, centromere DNA and CENH3 differences could introduce reproductive barriers, causing speciation. CENH3 binding domains in animal and plant chromosomes cover many kbps of DNA. How is it possible that these large stretches of DNA could co-evolve with a histone H3 variant? Similarly, how do centromere DNA sequences on different chromosomes co-evolve? In a tandem repeat array, CENH3 is necessarily binding to the

same sequences throughout the centromere, and all chromosomes in the cell typically share versions of the same repeat monomer (Kawabe, & Charlesworth 2007). In addition, tandem repeats foster rapid evolution, and this property may be favored by meiotic drive (Smith 1976; Henikoff et al 2001). A mutation that arises in any copy of a tandem repeat can be amplified and spread throughout the array by unequal crossing over (Smith 1976) or by replication fork collapse (Houseley, & Tollervey 2011). Repeat variants can move between different chromosomes in the cell via gene conversion, or possibly through the mobilization of retrotransposons inserted into tandem repeat arrays (Shi et al 2010; Birchler, & Presting 2012). As we have shown, the centromere tandem repeat array can be a "library" of sequence variants that show expansion and shrinkage (Ma, & Jackson 2006; Ma, & Bennetzen 2006), creating opportunities for new variants to colonize a chromosome, likely via concerted evolution or molecular drive (Dover 1982). Centromeres with sequence differences would be immediately exposed to selection in organisms with asymmetric female meiosis. Thus, the ability of tandem repeats to facilitate concerted evolution may explain their prevalence at animal and plant centromeres. Yeast species with symmetrical meiosis lack high copy tandem repeats at centromeres (Meraldi et al 2006). Similarly, the centromere-specific histone does not show positive selection in *Tetrahymena* species with symmetrical meiosis (Elde et al 2011). In the future, it will be interesting to test whether tandem repeats are found at centromeres of diverse eukaryotes that lack asymmetric meiosis.

Methods

Obtaining sequence data from online archives

Only whole genome shotgun (WGS) or whole chromosome shotgun (WCS) data were used in our analysis. Sanger DNA sequences (FASTA and corresponding ancillary files) were downloaded from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/home/). For each of the 170 species with WGS or WCS Sanger data, we downloaded up to 5 randomly selected FASTA files (up to 500,000 sequences/file). Illumina and 454 data were downloaded from the DDBJ Sequence Read Archive (http://trace.ddbj.nig.ac.jp/DRASearch/). As of April 1, 2012, 146 species had WGS Illumina or 454 data. For these species, two random FASTQ files were downloaded (one per direction, on average 2Gb/file). For 37 species both Sanger and Illumina data were obtained. A complete list of species, and associated sequence data, that were used in our study can be found in Supplementary Table S1.

Bioinformatics pipeline for Sanger and Pacific Biosciences data

WGS or WCS data were processed using a Perl-based bioinformatics pipeline. First, Sanger sequences were clipped for quality and/or vector contamination. Subsequently, sequences that had >5% Ns were removed, as were any sequences shorter than 100 bp (Sanger) or 1,000 bp (PacBio). Low complexity sequences were then masked using the DUST filter. The remaining sequences were analyzed by Tandem Repeats Finder (TRF) (http://tandem.bu.edu/trf/trf.html) (Benson 1999) to identify tandem repeats. We assumed that candidate centromeric tandem repeat arrays should be continuous and occupy the majority — if not all — of any individual read. We therefore excluded repeats that accounted for <80% of the entire read. TRF sometimes predicted

multiple tandem repeats occupying the same span within a read (with different repeat monomer lengths). In these situations we only retained the shortest repeat for further analysis. Very short repeats, with monomer lengths <50 bp, were also excluded from further analysis.

After producing a set of tandem repeats for each species of interest (using the consensus repeat sequence from TRF), we then used WU-BLASTN ([NO STYLE for: Gish]) with parameters M=1 N=-1 Q=3 R=3 W=10 (with post-processing from various Perl scripts) to produce a set of 'global' and 'local' clusters of repeats in each species (see Supplemental Methods for full details). Global clusters contained repeats with very similar sequences that also had near-identical lengths. This clustering step used just a sample of the total number of tandem repeats produced by TRF and we identifed the source reads of all of the sample repeats. This allowed us to identify what fraction of the input sample reads were represented by each global or local cluster. Repeats in the top clusters are presumed to be the candidate centromeric repeat.

Bioinformatics pipeline for Illumina and 454 data

centromere tandem repeats described so far (Henikoff et al 2001)). Some short read assemblers do not work well with sequences containing polymorphisms. To assemble polymorphic centromere tandem repeats, we used the short read assembler PRICE (Paired-Read Iterative Contig Extension) (http://derisilab.ucsf.edu/software/price/index.html). For most of the Illumina and 454 data we used PRICE beta version 0.6. This version could only handle paired-end Illumina and 454 data. The later PRICE beta version 0.13 and subsequent versions also allowed for use of single end Illumina and 454 data. For each species, we used 200,000 randomly selected reads which were assembled on 20,000 seed sequences (see PRICE manual) with at

least 85% sequence similarity. The contigs were analyzed for the presence of tandem repeats by TRF, allowing for a tandem repeat monomer of 2,000 bp (upper limit of TRF). To determine genomic fraction, 1,000,000 short reads were aligned to the obtained tandem repeat monomers (see Supplementary Methods for more details).

Data analysis of centromere tandem repeats

To compare candidate centromeres from all species to each other, we performed a BLASTN (Altschul et al 1990) search. We used WU-BLAST version 2.0 with parameters M=1 N=-1 Q=2 R=2 W=8. Since tandem repeat boundaries are arbitrary, it is possible for related repeats to align in a staggered fashion and align over only a fraction of their true length. We therefore aligned a file of repeats to a file of duplicated repeats. Since BLAST produces local alignments and we were interested in overall similarity, we calculated a global percent identity by adding additional alignment length assuming a 25% match rate in unaligned regions.

To assess the rate at which sequence similarity decays on phylogenetic timescales, we performed node-averaged phylogenetically independent contrasts (Coyne, & Orr 1989; Fitzpatrick 2002). In order to account for shared history in comparisons of sequence similarity, this method calculates the average sequence similarity between each pair of taxa spanning a node to generate a single value for each node in the tree. Since the taxa of interest span a wide range of eukaryotes and our analyses are relatively insensitive to branch length estimates, we used a tree based on the NCBI taxonomy (Sayers et al 2012) and repeated our analyses on ten random resolutions of the tree in order to accommodate unresolved relationships. As most unresolved nodes were shallow, these random resolutions had little effect on the quantitative results of the analyses performed (data not

shown). All phylogenetic analyses were conducted using the R package APE (Paradis et al 2004). We then performed regression analysis in order to determine the relationship between node age (as determined with TimeTree, (Hedges et al 2006)) and node-averaged sequence similarity. We used the R package bbmle2 to fit the simple exponential model $H \sim \alpha t^{\lambda}$, where H is the node-averaged homology and t is node age, and α is the intercept.

To determine the conservation of several repeat characteristics on a finer scale, we performed phylogenetic comparative analysis using the R packages GEIGER (http://cran.r-project.org/web/packages/geiger/index.html) (Harmon et al 2008) and picante (Kembel et al 2010; Popescu et al 2012). We estimated Blomberg's *K* measure of phylogenetic conservation for repeat length, GC content, and repeat abundance using chronograms estimated for primates (http://10ktrees.fas.harvard.edu/index.html) and grasses (Bouchenak-Khelladi et al 2010).

Pacific Biosciences single molecule real time sequencing

Switchgrass (tetraploid *Panicum virgatum* AP13) DNA was isolated using a modified protocol for Chen and Ronald (Chen, & Ronald 1999) (Supplementary methods). Library preparation and sequencing was performed according to manufacturer's instructions (Pacific Biosciences, Menlo Park, CA). In short, 3-10 µg of genomic DNA was isolated and fragmented to 7-10 kbp fragments using HydroShear for 15 minutes (switchgrass), or Covaris G-tube (cattle, yak, water buffalo). The first of five Ampure XP bead purifications was performed (0.45X Ampure beads added to DNA dissolved in 200 µL EB, vortexed for 10 minutes at 2,000 rpm, followed by two washes with 70% alcohol and finally diluted in EB) After each Ampure XP purification step a quality control was performed comprised of DNA concentration determination by nanodrop and

fragment size distribution by bioanalyzer. Next, the DNA fragments were repaired using DNA Damage Repair solution (1X DNA Damage Repeat Buffer, 1X NAD+, 1 mM ATP high, 0.1 mM dNTP, and 1X DNA Damage Repeat Mix with a final volume of 85.3 µL) with a volume of 21.1 µL and incubated at 37ºC for 20 minutes. DNA ends were repaired next by adding 1X End Repeat Mix to the solution, which was incubated at 25ºC for 5 minutes, followed by the second Ampure XP purification step. Next, 0.75 µM of Blunt Adapter was added to the DNA, followed by 1X template Prep Buffer, 0.05 mM ATP low and 0.75 U/µL T4 ligase to ligate (final volume of 47.5 µL) the bell adapters to the DNA fragments. This solution was incubated at 25ºC for 30 minutes, followed by a 65ºC 10 minute heat-shock. The exonuclease treatment to remove unligated DNA fragments consists of 1.81 U/µL Exo III and 0.18 U/µL Exo IV (final volume of 3.8 µL) which is incubated at 37ºC for 1 hour. Next, three Ampure XP purifications steps were performed. Finally, the bell primer is annealed to the PacBio bell with inserted DNA fragment (80ºC for 2 minute 30 followed by decreasing the temperature by 0.1º/s to 25Cº). This complex was loaded into PacBio RS SMRT cells, which were loaded onto the machine for either 2 x 30, 2 x 45, 1 x 75, or 1 x 90 minute runs. Four cells each were used for *Zea mays*, *Zea luxurians, Panicum virgatum, Bos taurus taurus, Bos taurus indicus, Bos grunniens, Bison bison* and *Bubalus bubalis*, while two cells were sufficient for *Panicum capillare*.

Fluorescence in situ hybridization (FISH)

Mitotic chromosome spreads were generated following a protocol by Zhang and Friebe (Zhang et al 2010) with a few modifications (Supplementary methods). Plasmid vectors containing a single copy of each repeat variant (A, B1, B2, C, or D) were synthesized by Bio Basic Inc.

(Ontario, Canada) and used as probes for FISH analyses. Probe hybridization signals were detected using anti-digoxigenin (dig) conjugated FITC (green), anti-dig conjugated Rhodamine (red), or Streptavidin conjugated Rhodamine (red) antibodies (Roche Applied Sciences). Chromosomes were counter-stained with 4',6-diamidino-2-phenylindole (DAPI). Digital images were recorded using an Olympus BX51 epifluorescence microscope (Olympus Corporation, Center Valley, PA) (Suplementary methods for more details).

Data Access

Pacific Biosciences sequences for *Panicum capillare* and *Panicum virgatum* were deposited in the NCBI SRA under accession number SRA052051. A list of GenBank and Sequence Read Archive accession numbers for all sequences used in this study are provided in the supplementary material. A spreadsheet containing all of the tandem repeat information for each species in this study, along with copies of all Perl scripts used are available to download online (http://korflab.ucdavis.edu/Datasets/).

Acknowledgments

This work would not have been possible without the guidence, vision, and boundless enthusiasm of Simon Chan who sadly passed away on August 22th, 2012 at the age of 38.

We thank Reneé Godtel for technical assistance with bovid DNA sequencing, the International Nellore Genome Sequencing Consortium for prepublication access to sequence data, Michael Heaton for access to DNA samples of bison, yak, and water buffalo, Lauren Sagara and Henriette O'Geen for technical assistance with Pacific Biosciences sequencing, and Felicia Tsang, Alan Raetz and Alex Han for technical assistance with bioinformatic analysis.

The U.S. Department of Agriculture, Agricultural Research Service, is an equal opportunity/ affirmative action employer and all agency services are available without discrimination. Mention of commercial products and organizations in this manuscript is solely to provide specific information. It does not constitute endorsement by USDA-ARS over other products and organizations not mentioned.

This work was supported by a Joint USDA/DOE Office of Science Feedstock genomics grant DE-AI02-09ER64829 (to C.T.), and by National Science Foundation grants IOS-0922703 to J.R.-I. and IOS-1026094 to S.C. D.M. was supported by training grant T32-GM008799 from NIH-NIGMS. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

J.D. is an investigator of the Howard Hughes Medical Institute. S.C. is a Howard Hughes Medical Institute and Gordon and Betty Moore Foundation investigator.

References

Alkan, C., Cardone, M.F., Catacchio, C.R., Antonacci, F., O'Brien, S.J., Ryder, O.A., Purgato, S., Zoli, M., Della Valle, G., Eichler, E.E. & Ventura, M., 2011, Genome-wide characterization of centromeric satellites from multiple mammalian genomes, *Genome research*, 21(1), pp. 137-45.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., 1990, Basic local alignment search tool, *Journal of molecular biology*, 215(3), pp. 403-10.

Amor, D.J. & Choo, K.H., 2002, Neocentromeres: role in human disease, evolution, and centromere study, *Am J Hum Genet*, 71(4), pp. 695-714.

Ananiev, E.V., Phillips, R.L. & Rines, H.W., 2000, Complex structure of knobs and centromeric regions in maize chromosomes, *Tsitol Genet*, 34(2), pp. 11-5.

Bassett, E.A., Wood, S., Salimian, K.J., Ajith, S., Foltz, D.R. & Black, B.E., 2010, Epigenetic centromere specification directs aurora B accumulation but is insufficient to efficiently correct mitotic errors, *The Journal of cell biology*, 190(2), pp. 177-85.

Bensasson, D., 2011, Evidence for a high mutation rate at rapidly evolving yeast centromeres, *BMC evolutionary biology*, 11, p. 211.

Bensasson, D., Zarowiecki, M., Burt, A. & Koufopanou, V., 2008, Rapid evolution of yeast centromeres in the absence of drive, *Genetics*, 178(4), pp. 2161-7.

Benson, G., 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic acids research*, 27(2), pp. 573-80.

Birchler, J.A. & Presting, G.G., 2012, Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes, *Genes & development*, 26(7), pp. 638-40.

Birchler, J.A., Gao, Z. & Han, F., 2009, A tale of two centromeres--diversity of structure but conservation of function in plants and animals, *Functional & integrative genomics*, 9(1), pp. 7-13.

Blomberg, S.P., Garland, T. & Ives, A.R., 2003, Testing for phylogenetic signal in comparative data: behavioral traits are more labile, *Evolution*, 57(4), pp. 717-45.

Bouchenak-Khelladi, Y., Verboom, G.A., Savolainen, V. & Hodkinson, T.R., 2010, Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal evolutionary history in geographical space and geological time, *Botanical Journal of the Linnean Society*, 162(4), pp. 543-57.

Carone, D.M., Longo, M.S., Ferreri, G.C., Hall, L., Harris, M., Shook, N., Bulazel, K.V., Carone, B.R., Obergfell, C., O'Neill, M.J. & O'Neill, R.J., 2009, A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres, *Chromosoma*, 118(1), pp. 113-25.

Cellamare, A., Catacchio, C.R., Alkan, C., Giannuzzi, G., Antonacci, F., Cardone, M.F., Della Valle, G., Malig, M., Rocchi, M., Eichler, E.E. & Ventura, M., 2009, New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset, *Molecular biology and evolution*, 26(8), pp. 1889-900.

Chen, D.- & Ronald, P.C., 1999, A Rapid DNA Minipreparation Method Suitable for AFLP and Other PCR Applications, *Plant Molecular Biology Reporter*, 17(1), pp. 53-7.

Chia, J.M., Song, C., Bradbury, P., Costich, D., de Leon, N., Doebley, J.C., Elshire, R.J., Gaunt, B.S., Geller, L., Glaubitz, J.C., Gore, M., Guill, K.E., Holland, J., Hufford, M.B., Lai, J., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B.M., Pyhäjärvi, T., Rong, T., Shekhon, R.S., Sun Q, Tenaillon, M., Tian, F., Wang, J., Xu X, Zhang Z, Kaeppler, S., Ross-Ibarra, J.,

McMullen, M., Buckler, E.S., Zhang, G., Xu, Y. & Ware, D., Capturing extant variation from a genome in flux: maize HapMap II, *Nature Genetics*, pp. in press.
Coyne, J.A. & Orr, H.A., 1989, Patterns of speciation in Drosophila, *Evolution*, pp. 362-81.
Dawe, R.K., Bennetzen, J.L. & Hake, S. 2009, Maize Centromeres and Knobs (neocentromeres) Handbook of Maize, in Springer New York, pp. 239-50.
Dernburg, A.F., 2001, Here, there, and everywhere: kinetochore function on holocentric chromosomes, *The Journal of cell biology*, 153(6), pp. F33-8.
Dover, G., 1982, Molecular drive: a cohesive mode of species evolution, *Nature*, 299(5879), pp. 111-7.
Du, Y., Topp, C.N. & Dawe, R.K., 2010, DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA, *PLoS genetics*, 6(2), p. e1000835.
Elde, N.C., Roach, K.C., Yao, M.C. & Malik, H.S., 2011, Absence of positive selection on centromeric histones in Tetrahymena suggests unsuppressed centromere: drive in lineages lacking male meiosis, *Journal of molecular evolution*, 72(5-6), pp. 510-20.
Elder, J.F. & Turner, B.J., 1994, Concerted evolution at the population level: pupfish HindIII satellite DNA sequences, *Proceedings of the National Academy of Sciences of the United States of America*, 91(3), pp. 994-8.
Fishman, L. & Saunders, A., 2008, Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers, *Science (New York, N.Y.)*, 322(5907), pp. 1559-62.
Fitzpatrick, B.M., 2002, Molecular correlates of reproductive isolation, *Evolution*, 56(1), pp. 191-8.
Gaillard, C., Doly, J., Cortadas, J. & Bernardi, G., 1981, The primary structure of bovine satellite 1.715, *Nucleic acids research*, 9(22), pp. 6069-82.
Gassmann, R., Rechtsteiner, A., Yuen, K.W., Muroyama, A., Egelhofer, T., Gaydos, L., Barron, F., Maddox, P., Essex, A., Monen, J., Ercan, S., Lieb, J.D., Oegema, K., Strome, S. & Desai, A., 2012, An inverse relationship to germline transcription defines centromeric chromatin in C. elegans, *Nature*.
Gent, J.I., Schneider, K.L., Topp, C.N., Rodriguez, C., Presting, G.G. & Dawe, R.K., 2011, Distinct influences of tandem repeats and retrotransposons on CENH3 nucleosome positioning, *Epigenetics & chromatin*, 4, p. 3.
Gill, N., Findley, S., Walling, J.G., Hans, C., Ma, J., Doyle, J., Stacey, G. & Jackson, S.A., 2009, Molecular and chromosomal evidence for allopolyploidy in soybean, *Plant Physiol*, 151(3), pp. 1167-74.
Gish, Whttp://blast.wustl.edu, *http://blast.wustl.edu*. Retrieved 2009,
Gong, Z., Wu, Y., Koblízková, A., Torres, G.A., Wang, K., Iovene, M., Neumann, P., Zhang, W., Novák, P., Buell, C.R., Macas, J. & Jiang, J., 2012, Repeatless and Repeat-Based Centromeres in Potato: Implications for Centromere Evolution, *The Plant cell*.
Han, F., Gao, Z. & Birchler, J.A., 2009, Reactivation of an inactive centromere reveals epigenetic and structural components for centromere specification in maize, *The Plant cell*, 21(7), pp. 1929-39.
Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E. & Challenger, W., 2008, GEIGER: investigating evolutionary radiations, *Bioinformatics*, 24(1), pp. 129-31.

Hayden, K. & Willard, H., 2012, Composition and organization of active centromere sequences in complex genomes, *BMC genomics*, 13(1), p. 324.
Hedges, S.B., Dudley, J. & Kumar, S., 2006, TimeTree: a public knowledge-base of divergence times among organisms, *Bioinformatics*, 22(23), pp. 2971-2.
Hemleben, V., Kovarik, A., Torres-Ruiz, R.A., Volkov, R.A. & Beridze, T., 2007, Plant highly repeated satellite DNA: molecular evolution, distribution and use for identification of hybrids, *Systematics and Biodiversity*, 5(03), pp. 277-89.
Henikoff, S., Ahmad, K. & Malik, H.S., 2001, The centromere paradox: stable inheritance with rapidly evolving DNA, *Science (New York, N.Y.)*, 293(5532), pp. 1098-102.
Horvath, J.E. & Willard, H.F., 2007, Primate comparative genomics: lemur biology and evolution, *Trends in genetics : TIG*, 23(4), pp. 173-82.
Hosouchi, T., Kumekawa, N., Tsuruoka, H. & Kotani, H., 2002, Physical map-based sizes of the centromeric regions of Arabidopsis thaliana chromosomes 1, 2, and 3, *DNA Res*, 9(4), pp. 117-21.
Houben, A., Schroeder-Reiter, E., Nagaki, K., Nasuda, S., Wanner, G., Murata, M. & Endo, T.R., 2007, CENH3 interacts with the centromeric retrotransposon cereba and GC-rich satellites and locates to centromeric substructures in barley, *Chromosoma*, 116(3), pp. 275-83.
Houseley, J. & Tollervey, D., 2011, Repeat expansion in the budding yeast ribosomal DNA can occur independently of the canonical homologous recombination machinery, *Nucleic acids research*, 39(20), pp. 8778-91.
Kawabe, A. & Charlesworth, D., 2007, Patterns of DNA variation among three centromere satellite families in Arabidopsis halleri and A. lyrata, *Journal of molecular evolution*, 64(2), pp. 237-47.
Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P. & Webb, C.O., 2010, Picante: R tools for integrating phylogenies and ecology, *Bioinformatics*, 26(11), pp. 1463-4.
Lam, A.L., Boivin, C.D., Bonney, C.F., Rudd, M.K. & Sullivan, B.A., 2006, Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA, *Proceedings of the National Academy of Sciences of the United States of America*, 103(11), pp. 4186-91.
Lee, H.R., Hayden, K.E. & Willard, H.F., 2011, Organization and molecular evolution of CENP-A--associated satellite DNA families in a basal primate genome, *Genome biology and evolution*, 3, pp. 1136-49.
Lee, H.R., Zhang, W., Langdon, T., Jin, W., Yan, H., Cheng, Z. & Jiang, J., 2005, Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species, *Proceedings of the National Academy of Sciences of the United States of America*, 102(33), pp. 11793-8.
Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.P., Wang, Z., Chinwalla, A.T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K.D., Fronick, C., Schmidt, H., Fulton, L.A., Fulton, R.S., Nelson, J.O., Magrini, V., Pohl, C., Graves, T.A., Markovic, C., Cree, A., Dinh, H.H., Hume, J., Kovar, C.L., Fowler, G.R., Lunter, G., Meader, S., Heger, A., Ponting, C.P., Marques-Bonet, T., Alkan, C., Chen, L., Cheng, Z., Kidd, J.M., Eichler, E.E., White, S., Searle, S., Vilella, A.J., Chen, Y., Flicek, P., Ma, J., Raney, B., Suh, B., Burhans,

R., Herrero, J., Haussler, D., Faria, R., Fernando, O., Darré, F., Farré, D., Gazave, E., Oliva, M., Navarro, A., Roberto, R., Capozzi, O., Archidiacono, N., Della Valle, G., Purgato, S., Rocchi, M., Konkel, M.K., Walker, J.A., Ullmer, B., Batzer, M.A., Smit, A.F., Hubley, R., Casola, C., Schrider, D.R., Hahn, M.W., Quesada, V., Puente, X.S., Ordoñez, G.R., López-Otín, C., Vinar, T., Brejova, B., Ratan, A., Harris, R.S., Miller, W., Kosiol, C., Lawson, H.A., Taliwal, V., Martins, A.L., Siepel, A., Roychoudhury, A., Ma, X., Degenhardt, J., Bustamante, C.D., Gutenkunst, R.N., Mailund, T., Dutheil, J.Y., Hobolth, A., Schierup, M.H., Ryder, O.A., Yoshinaga, Y., de Jong, P.J., Weinstock, G.M., Rogers, J., Mardis, E.R., Gibbs, R.A. & Wilson, R.K., 2011, Comparative and demographic analysis of orang-utan genomes, *Nature*, 469(7331), pp. 529-33.
Ma, J. & Bennetzen, J.L., 2006, Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice, *Proceedings of the National Academy of Sciences of the United States of America*, 103(2), pp. 383-8.
Ma, J. & Jackson, S.A., 2006, Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice, *Genome research*, 16(2), pp. 251-9.
Mackinnon, R.N. & Campbell, L.J., 2011, The role of dicentric chromosome formation and secondary centromere deletion in the evolution of myeloid malignancy, *Genet Res Int*, 2011, p. 643628.
Malik, H.S. & Henikoff, S., 2009, Major evolutionary transitions in centromere complexity, *Cell*, 138(6), pp. 1067-82.
Marshall, O.J., Chueh, A.C., Wong, L.H. & Choo, K.H., 2008, Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution, *Am J Hum Genet*, 82(2), pp. 261-82.
Masumoto, H., Nakano, M. & Ohzeki, J., 2004, The role of CENP-B and alpha-satellite DNA: de novo assembly and epigenetic maintenance of human centromeres, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 12(6), pp. 543-56.
McAinsh, A.D., Tytell, J.D. & Sorger, P.K., 2003, Structure, function, and regulation of budding yeast kinetochores, *Annu Rev Cell Dev Biol*, 19, pp. 519-39.
Melters, D.P., Paliulis, L.V., Korf, I.F. & Chan, S.W., 2012, Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 20(5), pp. 579-93.
Meraldi, P., McAinsh, A.D., Rheinbay, E. & Sorger, P.K., 2006, Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins, *Genome biology*, 7(3), p. R23.
Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, C.R. & Jiang, J., 2004, Sequencing of a rice centromere uncovers active genes, *Nature genetics*, 36(2), pp. 138-45.
Nakano, M., Cardinale, S., Noskov, V.N., Gassmann, R., Vagnarelli, P., Kandels-Lewis, S., Larionov, V., Earnshaw, W.C. & Masumoto, H., 2008, Inactivation of a human kinetochore by specific targeting of chromatin modifiers, *Developmental cell*, 14(4), pp. 507-22.

Nasuda, S., Hudakova, S., Schubert, I., Houben, A. & Endo, T.R., 2005, Stable barley chromosomes without centromeric repeats, *Proceedings of the National Academy of Sciences of the United States of America*, 102(28), pp. 9842-7.
Navajas-Pérez, R. & Paterson, A.H., 2009, Patterns of tandem repetition in plant whole genome assemblies, *Molecular genetics and genomics : MGG*, 281(6), pp. 579-90.
Neumann, P., Navrátilová, A., Schroeder-Reiter, E., Koblížková, A., Steinbauerová, V., Chocholová, E., Novák, P., Wanner, G. & Macas, J., 2012, Stretching the rules: monocentric chromosomes with multiple centromere domains, *PLoS genetics*, 8(6), p. e1002777.
Ohzeki, J.I., Bergmann, J.H., Kouprina, N., Noskov, V.N., Nakano, M., Kimura, H., Earnshaw, W.C., Larionov, V. & Masumoto, H., 2012, Breaking the HAC Barrier: Histone H3K9 acetyl/ methyl balance regulates CENP-A assembly, *The EMBO journal*.
O'Neill, R.J. & Carone, D.M., 2009, The role of ncRNA in centromeres: a lesson from marsupials, *Progress in molecular and subcellular biology*, 48, pp. 77-101.
Palomeque, T. & Lorite, P., 2008, Satellite DNA in insects: a review, *Heredity*, 100(6), pp. 564-73.
Paradis, E., Claude, J. & Strimmer, K., 2004, APE: Analyses of Phylogenetics and Evolution in R language, *Bioinformatics*, 20(2), pp. 289-90.
Pikaard, C. & Pontes, O., 2007, Heterochromatin: condense or excise, *Nature cell biology*, 9(1), pp. 19-20.
Plohl, M., Luchetti, A., Mestrović, N. & Mantovani, B., 2008, Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin, *Gene*, 409(1-2), pp. 72-82.
Popescu, A.A., Huber, K.T. & Paradis, E., 2012, ape 3.0: new tools for distance based phylogenetics and evolutionary analysis in R, *Bioinformatics*.
Płucienniczak, A., Skowroński, J. & Jaworski, J., 1982, Nucleotide sequence of bovine 1.715 satellite DNA and its relation to other bovine satellite sequences, *J Mol Biol*, 158(2), pp. 293-304.
Rudd, M.K., Wray, G.A. & Willard, H.F., 2006, The evolutionary dynamics of alpha-satellite, *Genome research*, 16(1), pp. 88-96.
Sato, H., Masuda, F., Takayama, Y., Takahashi, K. & Saitoh, S., 2012, Epigenetic inactivation and subsequent heterochromatinization of a centromere stabilize dicentric chromosomes, *Current biology : CB*, 22(8), pp. 658-67.
Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I.M., Geer, L.Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, W.J., Yaschenko, E. & Ye, J., 2012, Database resources of the National Center for Biotechnology Information, *Nucleic acids research*, 40(Database issue), pp. D13-25.
Schadt, E.E., Turner, S. & Kasarskis, A., 2010, A window into third-generation sequencing, *Hum Mol Genet*, 19(R2), pp. R227-40.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A. & Wilson, R.K., 2009, The B73 maize genome: complexity, diversity, and dynamics, *Science (New York, N.Y.)*, 326(5956), pp. 1112-5.

Shang, W.H., Hori, T., Toyoda, A., Kato, J., Popendorf, K., Sakakibara, Y., Fujiyama, A. & Fukagawa, T., 2010, Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences, *Genome research*, 20(9), pp. 1219-28.

Shelby, R.D., Vafa, O. & Sullivan, K.F., 1997, Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites, *The Journal of cell biology*, 136(3), pp. 501-13.

Shi, J., Wolf, S.E., Burke, J.M., Presting, G.G., Ross-Ibarra, J. & Dawe, R.K., 2010, Widespread gene conversion in centromere cores, *PLoS biology*, 8(3), p. e1000327.

Smith, G.P., 1976, Evolution of repeated DNA sequences by unequal crossover, *Science*, 191(4227), pp. 528-35.

Stoler, S., Keith, K.C., Curnick, K.E. & Fitzgerald-Hayes, M., 1995, 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 & development*, 9(5), pp. 573-86.

Sullivan, B.A. & Karpen, G.H., 2004, Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin, *Nature structural & molecular biology*, 11(11), pp. 1076-83.

Swaminathan, K., Varala, K. & Hudson, M., 2007, Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey, *BMC genomics*, 8(1), p. 132.

Talbert, P.B., Bryson, T.D. & Henikoff, S., 2004, Adaptive evolution of centromere proteins in plants and animals, *J Biol*, 3(4), p. 18.

Taparowsky, E.J. & Gerbi, S.A., 1982, Structure of 1.71 lb gm/cm(3) bovine satellite DNA: evolutionary relationship to satellite I, *Nucleic acids research*, 10(18), pp. 5503-15.
Tek, A.L., Kashihara, K., Murata, M. & Nagaki, K., 2010, Functional centromeres in soybean include two distinct tandem repeats and a retrotransposon, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 18(3), pp. 337-47.
Topp, C.N., Zhong, C.X. & Dawe, R.K., 2004, Centromere-encoded RNAs are integral components of the maize kinetochore, *Proceedings of the National Academy of Sciences of the United States of America*, 101(45), pp. 15986-91.
Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-I, T., Toyoda, A., Fujiyama, A., Tarutani, Y. & Kakutani, T., 2012, Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata, *Genes & development*, 26(7), pp. 705-13.
Ventura, M., Antonacci, F., Cardone, M.F., Stanyon, R., D'Addabbo, P., Cellamare, A., Sprague, L.J., Eichler, E.E., Archidiacono, N. & Rocchi, M., 2007, Evolutionary formation of new centromeres in macaque, *Science (New York, N.Y.)*, 316(5822), pp. 243-6.
Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blocker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Della Valle, G., Fryc, S., Guerin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Roed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvanen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Lander, E.S. & Lindblad-Toh, K., 2009, Genome sequence, comparative analysis, and population genetics of the domestic horse, *Science (New York, N.Y.)*, 326(5954), pp. 865-7.
Wang, G., Zhang, X. & Jin, W., 2009, An overview of plant centromeres, *Journal of genetics and genomics = Yi chuan xue bao*, 36(9), pp. 529-37.
Warburton, P.E., Haaf, T., Gosden, J., Lawson, D. & Willard, H.F., 1996, Characterization of a chromosome-specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chromosomes, *Genomics*, 33(2), pp. 220-8.
Warburton, P.E., Waye, J.S. & Willard, H.F., 1993, Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin, *Molecular and cellular biology*, 13(10), pp. 6520-9.
Waye, J.S., Durfy, S.J., Pinkel, D., Kenwrick, S., Patterson, M., Davies, K.E. & Willard, H.F., 1987, Chromosome-specific alpha satellite DNA from human chromosome 1: hierarchical structure and genomic organization of a polymorphic domain spanning several hundred kilobase pairs of centromeric DNA, *Genomics*, 1(1), pp. 43-51.
Willard, H.F., 1990, Centromeres of mammalian chromosomes, *Trends in genetics : TIG*, 6(12), pp. 410-6.
Zhang, W., Friebe, B., Gill, B.S. & Jiang, J., 2010, Centromere inactivation and epigenetic modifications of a plant chromosome with three functional centromeres, *Chromosoma*, 119(5), pp. 553-63.

Zhang, W., Lee, H.-R., Koo, D.-H. & Jiang, J., 2008, Epigenetic Modification of Centromeric Chromatin: Hypomethylation of DNA Sequences in the CENH3-Associated Chromatin in Arabidopsis thaliana and Maize, *The Plant Cell Online*, 20(1), pp. 25-34.
Zhong, C.X., Marshall, J.B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J.A., Jiang, J. & Dawe, R.K., 2002, Centromeric retroelements and satellites interact with maize kinetochore protein CENH3, *The Plant cell*, 14(11), pp. 2825-36.
Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A. & Salzberg, S.L., 2009, A whole-genome assembly of the domestic cow, Bos taurus, *Genome biology*, 10(4), p. R42.

Supplementary files

Additional file 1 (methods):

http://genomebiology.com/content/supplementary/gb-2013-14-1-r10-s1.pdf

Additional file 2 (figures):

http://genomebiology.com/content/supplementary/gb-2013-14-1-r10-s2.pdf

Additional file 3 (tables):

http://genomebiology.com/content/supplementary/gb-2013-14-1-r10-s3.xls

6. Comparative Analysis of Tandem Repeats in 94 Fungi Genomes

In the previous chapter, 282 plant and animal genomes were analyzed for the presence of tandem repeats. In this study the most abundant was assumed to be the candidate centromere repeat based on published centromere DNAs in plants and animals (Chapter 5). The logic was based on the observation of fungal centromeres, such as *Saccharomyces cerevisiae*, *Candida albicans*, and *Schizosaccharomyces pombe*. The centromeres of *S cerevisiae* cover a stretch of 125-bp which can be subdivided in three domains: CDEI, CDEII, and CDEIII (Figure 3-1). CDEIII is absolutely essential for centromere function, as deletion of CDEIII results in complete centromere inactivation and even a single base substitution in the central CCG sequence results in complete centromere inactivation (Furuyama, & Biggins 2007). Close relatives to *S. cerevisiae* also have "point" centromeres, where both CDEI and CDEIII show sequence conservation, but both the length and sequence composition of CDEII differ (Meraldi et al 2006). Even though the *Saccharomyces* centromeres are very small, they accumulate mutations at a rate three times that of non-selected parts of its genome (Bensasson et al 2008; Bensasson 2011). Similarly, centromeres of *Candida* species also evolve faster than expected by "neutral" mutation rates (Padmanabhan et al 2008). Even tandem repeat elements in the pericentromeres of *Schizosaccharomyces* species lack sequence similarity, despite shared synteny around the centromeres (Rhind et al 2011). Although the pericentromeres of *Schizosaccharomyces* species are comprised of a variety of tandem repeats, the centromere core consists of unique sequences (Rhind et al 2011), similar in uniqueness (not sequence) to *Candida* centromeres (Padmanabhan et al 2008). The centromeres of the filamentous fungi *Neurospora crassa* were described to consist of degenerated retrotransposons and simple repeats (Smith et al 2012).

Previously, a motif finding exercise was performed to identify how common the "point" centromere of *S. cerevisiae* is (Meraldi et al 2006). Nevertheless, it remains unknown how common centromere tandem repeat are in the fungi kingdom. In this chapter I set out to identify, characterize and analyze tandem repeat in the genomes of 94 fungi.

Material & methods

Using the same bioinformatics pipeline described in Chapter 5, 94 fungi genomes were analyzed for the presence of highly abundant tandem repeats. These 94 fungi genomes were sequenced by Sanger technology and per species up to 500,000 whole genome shotgun (WGS) sequence data was downloaded from the central public repository NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?view=list_arrivals). The sequences were analyzed for the presence of tandem repeat sequences (Melters et al 2013) and subsequently these tandem repeats were characterized for repeat length, GC composition and genome fraction. Next, the sequences were analyzed for sequence similarity using BLASTn (http://blast.ncbi.nlm.nih.gov/) with the following settings M=1, N=-1, Q=3, R=3 W=10, hspmax=1. These results were compared to the phylogenetic relationships between the 94 fungi species. The phylogeny was obtained from NCBI taxonomy website (http://www.ncbi.nlm.nih.gov/taxonomy). The genome size of each fungi was obtained from the Fungal Genome Size Database (http://www.zbi.ee/fungal-genomesize/) (Kullman et al 2005) for further correlation analysis.

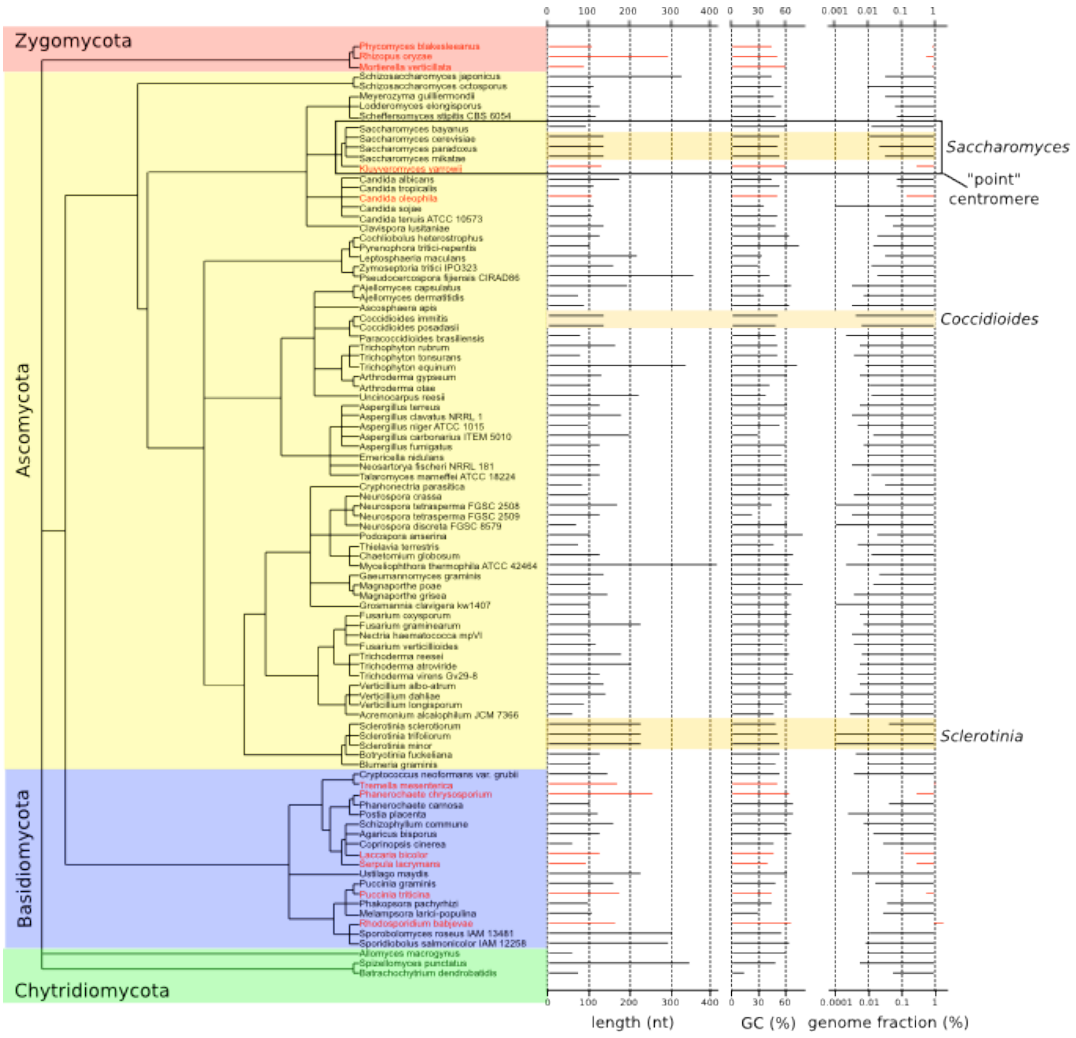


Figure 6-1 Centromere tandem repeat details from diverse fungi genomes.
The phylogenetic relationships between 94 fungi species are shown. For each species, the figure shows tandem repeat length, GC content, and genomic fraction (log 10 scale) for the most abundant tandem repeat monomer. Taxonomic relationships were derived from the NCBI taxonomy website. Only three groups of species were found to show sequence similarity between the most abundant tandem repeat. Each of these three groups consists of species from the same genus. No sequence similarity was found outside these groups, or between them. In red are the species where the most abundant tandem repeat is at least as abudant as in the holocentric nematode *C. elegans* (see dashed red line in Figure 5-2). The box around the *Saccharomyces* and *Kluyveromyces* species represent species known to have "point" centromeres.

Results and Discussion

The 94 fungi species covered four phyla: *Zygomycota* (n=3), *Chytridiomycota* (n=3), *Basidiomycota* (n=18), and *Ascomycota* (n=70). Following the identification of the most

abundant tandem repeat, an all-vs-all BLAST search of the consensus repeats was performed. This revealed that sequence conservation was limited to only very closely related species. I found ony three groups of species that showed sequence similarity between tandem repeats (Figure 6-1), namely three *Saccharomyces* species, two *Coccicoides* species and three *Sclerotinia* species. The former group is known to have "point" centromeres (Meraldi et al 2006), which excludes the possibility that these most abundant tandem repeats are candidate centromere repeats. The abundance of the tandem repeats in the two remaining groups were similar to the genome abundance of the tandem repeats in the *Saccharomyces* genus. It would therefore be unlikely that these conserved tandem repeats represent candidate centromere repeats.

The length distribution of the most abundant tandem repeat from fungi genomes (Figure 6-2A) was substantially shorter compared to that of plants and animals (Figure 5-5A). This was emphasized by the average length of the most abundant tandem repeat in fungi (147-bp; standard error (se) = 7.5), compared to average length of the most abundant tandem repeats in plants and animals (200-bp; se = 11.6).

The most important parameter that was used in the previous chapter was the genomic abundance of the most abundant tandem repeat. The assumption was that the most abundant tandem repeat would be the candidate centromere tandem repeat. Yet, in some cases the abundance was extremely low, which led us to doubt our assumption that for in those specific cases a candidate centromere tandem repeat was identified. As an initial cut-off value, the genomic abundance of the holocentric nematode *C. elegans* was used. As a result for 41 species it could not be determined what the candidate centromere sequences was (Chapter 5). When the same cut-off

value was used for the 94 fungi genomes, only eleven species were found to have candidate centromere tandem repeat sequences. This also included *Kluyveromyces yarrowia*, which is predicted to have a "point" centromere based on its close phylogenetic relationship to *K. lactis* (Heus et al 1990; Heus et al 1993; Heus et al 1994; Mattei et al 2002) and *S. cerevisiae* (Meraldi et al 2006). In addition, the GC content of most abundant tandem repeats (Figure 6-2B) was relatively high compared to the observed low GC content of studies fungi centromeres (Lynch et al 2010), which further emphasized that most of the identified tandem repeat were not candidate centromere sequences.

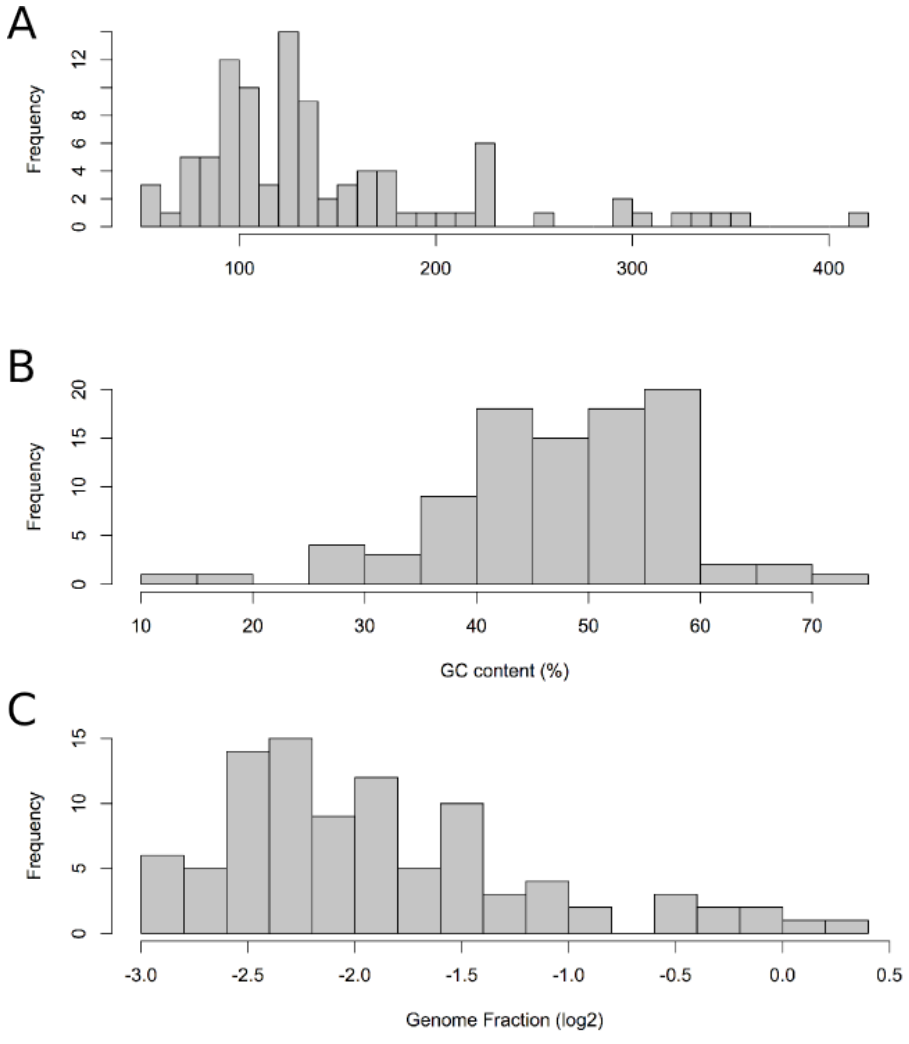


Figure 6-2 Tandem repeats lack conserved sequence properties. **(a-c)** No strong bias was observed in distribution of centromere repeat monomer length (a), GC content (b), or genomic fraction (c).

One specific clade did jump out with the potential of having candidate centromere tandem repeats: the three species from the *Zygomycota* phylum (Figure 6-1). Also, six of the eighteen (33%) species from the *Basidiomycota* phylum had genomic abundances higher than *C. elegans*, but these six species didn't cluster together, but rather covered the entire phylum, which complicates the interpretation. One possibility is that for all six *Basidiomycota* species the most abundant tandem repeat indeed represent the respective candidate centromere repeat, whereas the alternative possibility would be that . Chromatin-immunoprecipitation experiments with an antibody targeting cenH3 would conclusively determine if these tandem repeats are indeed centromere repeats. Also, expanding the bioinformatics approach used in this study to include next-generation sequence data would create a picture with better coverage of the different fungi phyla and the position of candidate centromere tandem repeats.

For very few fungi genomes a candidate centromere tandem repeat was identified, but why would so few fungi have predicted candidate centromere repeats? The small genome size of fungi species could potentially explain this (average size = 35.2 Mbp; se = 1.8; range = 11 - 82 Mbp). Small genomes dedicate a relative large fraction of their genome to genes and regulatory sequences, which leaves relatively little space for repetitive DNA. In contrast, the average genome size of the plants and animals from Chapter 5 was about forty times larger (average size = 1,453.6 Mbp; se = 89.6; range = 16 - 1550 Mbp) than the average fungal genome. If genome size was indeed a factor in why so few fungal centromere most likely lack high-copy tandem repeat arrays, a strong correlation between genome fraction and genome size would be expected. Yet, no correlation was found, identical to the lack of such a correlation between plant and animal genome sizes and genome fraction (Figure 6-3).

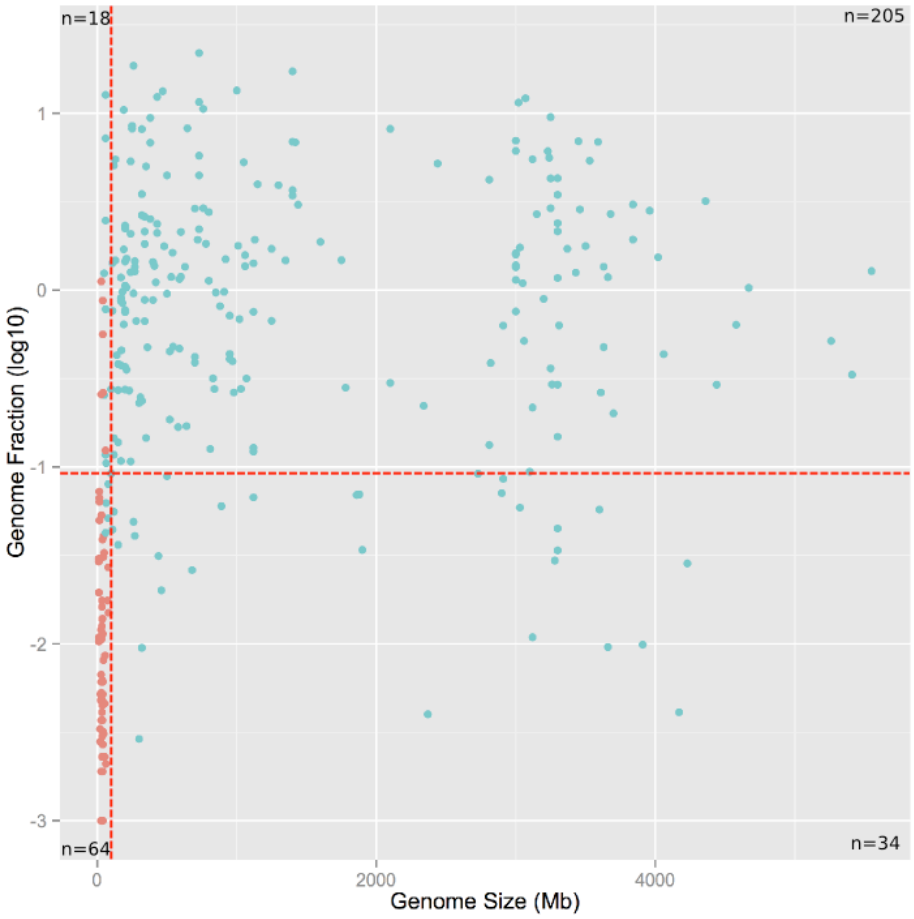


Figure 6-3 Comparing genome size and genome fraction of most abundant tandem repeat between plant and animal genomes (251/282 species; teal dots) and fungi genomes (70/94 species, red dots). Where the two red dashed lines intersect marks the genome size (100 Mb) and genome fraction (0.092%) of *C. elegans*. The lower left quadrant (n=64) contains species with smaller genomes and a lower genome fraction than *C. elegans*. The upper left quadrant (n=18) contains species with smaller genomes, but higher abundant genome fraction of the most abundant tandem repeat than *C. egelans*. The lower right quadrant (n=34) contains species with larger genomes, but lower abundant genome fraction than *C. elegans*. The upper right quadrant (n=205) contains species with larger genomes and more abundant genome fractions than *C. elegans*.

In summary, although all fungi genomes contained tandem repeats, only the genomes of *Zygomycota* species contain candidate centromere tandem repeats, based on genomic abundance of the respective tandem repeats. In addition, six of the eighteen *Basidiomycota* species might have centromeres containing high-copy tandem repeat arrays. Overall, the average repeat monomer length was substantially shorter than the average repeat monomer length of the most abundance tandem repeat in the plant and animal genomes from Chapter 5.

Acknowledgments

This work would not have been possible without the guidence, vision, and boundless enthusiasm of Simon Chan who died on August 22th, 2012 at the age of 38. I would also like to thank Keith Bradnam and Ian Korf for their technical assistance with bioinformatics and I am indebted to Scott Dawson for his advice on fungi phylogeny. This work was supported by training grant T32-GM008799 from NIH-NIGMS and by the Howard Hughes Medical Institute and the the Gordon and Betty Moore Foundation (GBMF3068). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or NIH.

Supplementary files

Additional file (table):

https://dl.dropboxusercontent.com/u/935849/Fungi_dissertation.xls

7. Discussion

The centromere paradox is the observation that centromere DNA and associated proteins evolve faster than expected for essential proteins and DNA sequences (Henikoff et al 2001). In addition, the chromosomal localization of centromeres does not follow the evolution of shared synteny blocks (Marshall et al 2008), despite the reduced recombination rate at centromeres (Nachman 2002). The lack of synteny is the result of evolutionary centromere repositioning events (or neocentromere formation), which result in the acquistion of centromere chromatin modifcation of the new site, whereas the old site quickly loses its centromere marks, including rapid loss of centromere tandem repeat (Piras et al 2009; Wade et al 2009; Piras et al 2010; Locke et al 2011; Rocchi et al 2012). Despite that centromere tandem repeat are neither sufficient nor essential (Karpen, & Allshire 1997; Sullivan, & Karpen 2001; Henikoff et al 2001), tandem repeat are commonly found at centromeres (Chapter 5). Here we discuss a possible mechanism how holocentricity could have evolved so often (Chapter 4) and why "point" centromeres and regional centromeres can be so different in sequence composition while still performing the same essential function. Finally, we will present a model for how novel repeats can be formed from existing tandem repeat arrays.

How can holocentricity have emerged so frequently?

The vast majority of what we know about holocentric chromosomes is based on studying *C. elegans*. This does not mean that *C. elegans* is an incidental case of a species with holocentric chromosomes. In Chapter 4 we estimated that holocentricity has evolved at least thirteen times. Various grasses (*Juncaceae* and *Cyperaceae*), dodders (*Cuscuta*), sundew (*Drosera*), nematodes,

arthropods, and arachnids (Figure 4-2) were described to have holocentric chromosomes. The most obvious morphological difference that differentiates holocentric chromosomes from monocentric chromosomes is the lack of a single specialized centromere locus per chromosome, rather, the kinetochore assembles along the entire length of the chromosome. This could indicate a change in the capacity of holocentric chromosomes to form centromere-specific chromatin domains. In the *Drosophila* S2 cell-line five distinct types of chromatin were identified based on the genome-wide binding maps of 53 chromatin components (Filion et al 2010). This includes a centromere-specific chromatin type (Filion et al 2010). When cenH3 (Cid) was modestly overexpressed *de novo* functional centromeres formed at heterochromatin boundaries (Lunyak 2008; Olszak et al 2011). Chromatin boundaries are formed by insulator proteins, such as CTCF, Mediator, and cohesin, creating chromatin loops that are transcriptionally isolated from the rest of the chromosome/genome (Phillips-Cremins et al 2013). The largest loops are stable throughout differentiation and in humans commonly include CTCF and cohesin (Phillips-Cremins et al 2013). Interestingly, the CTCF homologue was lost during nematode evolution, but remains present in the basal nematode clade *Dorylaimia* (Heger et al 2009), the one clade of nematodes that monocentric (Mutafova et al 1982; Spakulová et al 1994). This pattern matches the phylogeny of the evolution of holocentric chromosomes in nematodes (Figure 4-2). It is tempting to speculate that the loss of chromatin insulation results in losing the capacity to form centromere-specific chromatin domains, thereby allowing cenH3 to stably replace H3 across the entire genome. Comparative analysis between closely related species with holocentric and monocentric chromosomes for the presence or absence of insulator proteins will help shad light on this correlation.

Different types of mitosis requires different types of centromere?

A remarkable difference exist in sequences that make up of centromere DNA typesbetween *Saccharomyces cerevisiae*, and *Schizosaccharamyces pombe*, *Candida albicans*, and most plants and animals (see analyses from Chapters 5 and 6). These centromeres can be divided in three types: 1) "point" centromeres in *S. cerevisiae* (100-250 bp), 2) regional centromeres with unique sequences such as *S. pombe* and *C. albicans* centromeres (2-10 kbp) and 3) regional centromeres with tandem repeat arrays (>30 kbp) as can be found in most plant and animal centromeres.

Besides different types of centromeres, there is another factor that differentiates these three groups of species: how they go about their nucleus during mitosis. In *S. cerevisiae* spindle formation already starts during S phase and both the nuclear envelope and nuclear pore complexes remain intact during mitosis. This type of mitosis is called *closed mitosis* (De Souza, & Osmani 2007; Boettcher, & Barral 2013). In contrast, *Schizosaccharomyces japonicus* (Gu et al 2012; Aoki et al 2013) and *Aspergillus nidulans* (De Souza et al 2003; De Souza et al 2004) disassemble their nuclear pore complexes during the first phases of mitosis allowing the spindles to enter the nucleus. As the rest of the nuclear envelope remains intact, this type of mitosis has been dubbed *partially closed mitosis* (De Souza, & Osmani 2007; Boettcher, & Barral 2013). On the other hand, during mitosis in plants and animals the nuclear pore complexes as well as the nuclear envelope are disassembled, exposing the chromosomes to the cytoplasm. This is known as *open mitosis* (Boettcher, & Barral 2013).

Despite a lack of evidence supporting this correlation, there is an argument to be made in favor of it. When the spindles already form during S phase in closed mitosis, there is sufficient time for the spindles to find the centromeres. Ideally, the centromeres have already been positioned in

such a way that the spindles can find them relatively easy. As the chromosomes maintain association with the nuclear envelope, positioning of the centromeres can be facilitated.

An advantage of disassembling the nuclear pore complex is a larger gap for the spindle to enter the nucleus. The larger gap increases the chance of successful entry of the spindles into the nucleus and more lateral movement is permitted to allow the spindles to find the centromere . For optimal access of the spindles to the centromeres, also the nuclear envelope is disassembled. This is what happens in open mitosis in plants and animals. This allows for unrestricted access to the centromeres, but it does come at a significant metabolic cost, as the entire nuclear envelope and nuclear pore complexes have to be reassembled and the nucleoplasmic-cytoplasmic concentration gradient have to be reestablished (Maske, & Vaux 2000). Just disassembling the nuclear pore complexes leaves specific gaps in the nuclear envelope, but all that needs to be done post-mitosis is the reassembly of the nuclear pore complexes. For as long as the nuclear envelope is still intact, limited metabolic resources need to be directed to its reassembly of the envelope and the reestablishment of nucleoplasmic-cytoplasmic concentration gradient. At the same time it allows for the maintenance of chromosome anchoring to the same nuclear envelope. Anchoring of chromosomes to both the envelope and nuclear pore complexes have been associated with nuclear positioning of heterochromatin domains and genes (Akhtar, & Gasser 2007; Schneider, & Grosschedl 2007; Dechat et al 2008; Van de Vosse et al 2011). Positioning of "point" or regional centromeres (or hetrochromatic pericentromeres) in (partially) closed mitosis in close proximity to the (disassembled) nuclear pore complex could facilitate spindle-to-centromere attachment. In *S. pombe* (peri)centromeres have been shown to be clustered and bound to the nuclear envelope. When the physical linkage between centromeres and nuclear envelope were lost, chromosomes missegregated during mitosis (Funabiki et al 1993; Ding et al 1997; Hou et al

2012). There is also evidence to suggest that centromeres in *S. pombe* are released from the nuclear envelope and than recaptured by intranuclear microtubules, which subsequently align and segregate the chromosomes during mitosis (Funabiki et al 1993). Whether intranuclear microtubules (Zimmerman et al 2004) are a requirement for partially closed mitosis remains to be determined.

When sufficient time is available to attach the spindles to the centromeres, a very small "point" centromere would suffice. A short burst of tubulin molecules production has to rapidly perform to two functions: assemble spindles and attach to the centromeres. A larger regional centromeres would create a platform to which the spindles can attach. When there is no nuclear envelope left to spatially facilitate centromere positioning relative to the gaps left by disassembled nuclear complexes, chromosomes are lined up at the metaphase plate. A megabase-sized centromere might be favored to increase the change of for a successful attachment by the spindle. Therefore, one would predict that "point' centromere require few cenH3 molecules per centromere, whereas regional centromeres would require larger quantities of cenH3 molecules, especially those with open mitosis.

Overall, not much data is available on the quantity of cenH3 molecules per centromere because these experiments are both technically difficult and they require spadework. The data that is available fits the prediction described earlier. *S. cerevisiae* centromeres are marked by a single cenH3 molecule per centromere (Joglekar et al 2006; Furuyama, & Biggins 2007), which is substantially less than a *S. pombe* centromere (five cenH3 molecules) or a *C. albicans* centromere (8 cenH3 molecules) (Joglekar et al 2008). Even less data is available for species whose regional centromeres consistent of large tandem repeat arrays. In *D. melanogaster* on

average a centromere consists of 84 cenH3 molecules (Raychaudhuri et al 2012) and an average *Arabidopsis thaliana* centromere counts about 375 cenH3 molecules (Ramahi *et al*, submitted). Future studies focussing on the correlation between number of cenH3 molecules per centromere, types of mitosis and centromere DNAs will be needed.

Where do novel centromere tandem repeat variants come from?

When two random centromere tandem repeat sequences from two random animal or plant species are compared, odds are, little to now sequence similarity are going to be found. This means that one tandem repeat frequently outcompetes the other at the centromere. Do these sequence evolve rapidly in *cis* or are *tran*s mechanism at play? Thus, where do these new repeats come from? One possibility is that they arise from within (*cis*) an existing tandem repeat array.

In grasses several varients of a 155-bp tandem repeat were found to colocalize with the primary constrictions (Figure 5-7). More interestingly, in switchgrass (*Panicum virgatum*) three different repeat varients were identified. One showed similarity to the maize (*Zea mays*) centC sequences (Ananiev et al 2000; Zhong et al 2002; Sharma, & Presting 2008) and only hybridized to a single switchgrass centromere, whereas the other switchgrass centromeres hybridized to two longer repeat variants (B1 and B2). Variant B1 has a 20-bp insertion (Figure 5-3) from its ancestral repeat unit (155-bp) and variant B2 has a 9-bp deletion from the B1 variant. Single molecule sequencing revealed that an array consisted only of one repeat variant. This suggests that when a new repeat variant emerges it expands from that point onward, which is in line with the "library" hypothesis (Plohl et al 2008). Stochastic growth and shrinkage events could be facilitated by various molecular mechanisms, but it is generally assumed that unequal crossing-over is the major driver (Smith 1976). For unequal crossing-over to happen a single Holliday junction has to

form somewhere along the array of thousands of copies of short tandem repeats. Besides unequal crossing-over, gene conversion also requires Holliday junctions and gene conversion has been observed at centromere cores in maize (Shi et al 2010). In both cases there is a strong deletion bias (Smith 1976; Assis, & Kondrashov 2012), which would prefer continious shrinkage of tandem repeat arrays over tandem repeat array expansion. Yet, high-copy tandem repeat arrays are commonly found at centromeres (Chapter 5). The heterochromatic state of pericentromeres might help repress gene conversion (Cummings et al 2007), whereas the non-heterochromatic centromere core (Filion et al 2010) would allow for gene conversion to occur. On the other hand unequal crossing over appears to be less dependent on chromatin states (Kurnit 1979; Sudman, & Greenbaum 1990; Cabot et al 1993) and thus could affect repeats in the heterochromatic pericentromeres for as long as there is sufficient sequence homology to form Holliday junctions.

In addition, the functional centromere is epigenetically defined by the incorporation of cenH3 to form very stable cenH3-containing nucleosomes (Dimitriadis et al 2010; Lopes da Rosa et al 2011). CenH3 is not the only centromere chromatin mark; also CENP-C and other constitutive centromere-specific protein form the constitutive centromere-associated network (CCAN) (Perpelescu, & Fukagawa 2011; McAinsh, & Meraldi 2011). The stablity of the CCAN could limit the progression of Holliday junction along the array. The region on the tandem repeat array where CCAN can be found is expected to be primarily the target of gene conversion.

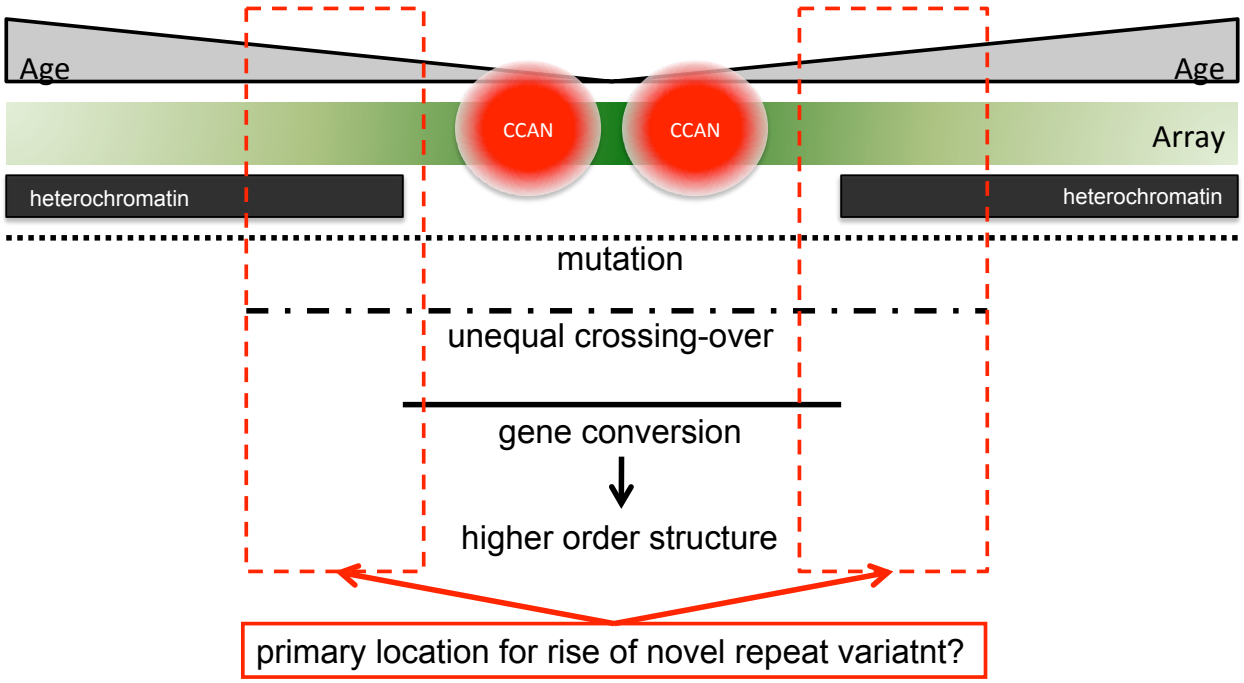


Figure 7-1. This is a hypothetical model proposing that novel repeats can be derived from existing high copy tandem repeat arrays, but the relative position of the repeat susceptible to novel repeat evolution is limited. Although the mutation pressure is approximately equal over the entire array, a difference in the likelihood of events between gene conversion (double Holliday junction) and unequal crossover (single Holliday junction). Those regions of the array that are more susceptible to unequal crossing over, but less so to gene conversion. The green gradient represents sequence homology, where greener equals greater homology. The gray rectangles indicate the age of the repeat within the array, where the most outer repeat tend to be the oldest. The black bars represent the hetrochromatic pericentromeres. The red dots represent the constitutive centromere-associated network (CCAN), which consists of 15 proteins, including cenH3 and CENP-C. The dotted line marks the region susceptible to the accumulation of mutations. The dotted-dashed line marks the region susceptible to unequal crossing over and the black line marks the region susceptible to gene conversion which in humans in believed to account for the observation of higher order repeat structures (Roizès 2006).

Combining these template-based mechanisms that allow repeat expansion and shrinkage with the physical limitation imposed by CCAN leads to the model described in Figure 7-1. This model explains where novel repeat variants could appear in an existing centromere tandem repeat arrays. A relatively small window within the entire array could exihibit the potential to create new repeat variants which can include indel (Figure 5-3) as a result from the increased mutation rate at the centromere (Bensasson et al 2008; Bensasson 2011; Padmanabhan et al 2008). Particullary unequal crossing-over would drive the expansion, but also gene conversion cannot

be excluded. The doubles variant of 171-bp primate centromere repeat sequence found in New World Monkeys (Figure 5-6) suggest that some level of gene conversion could have contributed to the formation of the new repeat. Additional stochastic expantion and shrinkage of the various repeat variants with enough time can lead to the disappearance of one variant and the conquest of another.

8. Conclusion

The centromere poses a paradox: it is functionally essential for all eukaryotes, yet its molecular components evolve fast. To better understand how centromere DNA evolved, we sat out to identify, characterize and analyze tandem repeat from hundreds of eukaryotic genomes using publicly available whole genome shotgun sequence data. In addition, we wanted to determine how frequently holocentricity has evolved to better understand its importance in chromosome biology

The four main findings from the studies described in this dissertation are: 1) holocentricity has evolved at least thirteen times (Chapter 4); 2) most plant and animal centromeres consist of high-copy tandem repeat arrays, but these repeat sequences lack conserved characteristics (Chapter 5); 3) no repeat sequences similarity was observed for species that diverged 50 million years ago (Chapter 5); and 4) maybe with the exception of the *Zygomycota* phylum, no fungi species is predicted to have centromere DNAs comprised of high-copy tandem repeat arrays (Chapter 6).

9. References

Abadon, M., Grenier, E., Laumond, C. & Abad, P., 1998, A species-specific satellite DNA from the entomopathogenic nematode Heterorhabditis indicus, *Genome*, 41(2), pp. 148-53.

Akhtar, A. & Gasser, S.M., 2007, The nuclear envelope and transcriptional control, *Nature Reviews Genetics*, 8(7), pp. 507-17.

Alkan, C., Ventura, M., Archidiacono, N., Rocchi, M., Sahinalp, S.C. & Eichler, E.E., 2007, Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data, *PLoS Comput Biol*, 3(9), pp. 1807-18.

Ananiev, E.V., Phillips, R.L. & Rines, H.W., 2000, Complex structure of knobs and centromeric regions in maize chromosomes, *Tsitol Genet*, 34(2), pp. 11-5.

Aoki, K., Shiwa, Y., Takada, H., Yoshikawa, H. & Niki, H., 2013, Regulation of nuclear envelope dynamics via APC/C is necessary for the progression of semi-open mitosis in Schizosaccharomyces japonicus, *Genes to cells : devoted to molecular & cellular mechanisms*.

Assis, R. & Kondrashov, A.S., 2012, A strong deletion bias in nonallelic gene conversion, *PLoS genetics*, 8(2), p. e1002508.

Baum, M., Sanyal, K., Mishra, P.K., Thaler, N. & Carbon, J., 2006, Formation of functional centromeric chromatin is specified epigenetically in Candida albicans, *Proceedings of the National Academy of Sciences of the United States of America*, 103(40), pp. 14877-82.

Bensasson, D., 2011, Evidence for a high mutation rate at rapidly evolving yeast centromeres, *BMC evolutionary biology*, 11, p. 211.

Bensasson, D., Zarowiecki, M., Burt, A. & Koufopanou, V., 2008, Rapid evolution of yeast centromeres in the absence of drive, *Genetics*, 178(4), pp. 2161-7.

Bernard, P. & Allshire, R., 2002, Centromeres become unstuck without heterochromatin, *Trends in cell biology*, 12(9), pp. 419-24.

Bernard, P., Maure, J.F., Partridge, J.F., Genier, S., Javerzat, J.P. & Allshire, R.C., 2001, Requirement of heterochromatin for cohesion at centromeres, *Science (New York, N.Y.)*, 294(5551), pp. 2539-42.

Beye, M. & Moritz, R.F., 1995, Characterization of honeybee (Apis mellifera L.) chromosomes using repetitive DNA probes and fluorescence in situ hybridization, *The Journal of heredity*, 86(2), pp. 145-50.

Boettcher, B. & Barral, Y., 2013, The cell biology of open and closed mitosis, *Nucleus (Austin, Tex.)*, 4(3), pp. 160-5.

Brown, J.D., Carone, D.M., Flynn, B.L., Finn, C.E., Mlynarski, E.E. & O'Neill, R.J., 2011, Centromere conversion and retention in somatic cell hybrids, *Cytogenetic and genome research*, 134(3), pp. 182-90.

Buckland, R.A., 1983, Comparative structure and evolution of goat and sheep satellite I DNAs, *Nucleic acids research*, 11(5), pp. 1349-60.

Cabot, E.L., Doshi, P., Wu, M.L. & Wu, C.I., 1993, Population genetics of tandem repeats in centromeric heterochromatin: unequal crossing over and chromosomal divergence at the Responder locus of Drosophila melanogaster, *Genetics*, 135(2), pp. 477-87.

Callaghan, M.J. & Beh, K.J., 1996, A tandemly repetitive DNA sequence is present at diverse locations in the genome of Ostertagia circumcincta, *Gene*, 174(2), pp. 273-9.

Castagnone-Sereno, P., Leroy, F. & Abad, P., 2000, Cloning and characterization of an extremely conserved satellite DNA family from the root-knot nematode Meloidogyne arenaria, *Genome*, 43(2), pp. 346-53.

Castagnone-Sereno, P., Leroy, H., Semblat, J.P., Leroy, F., Abad, P. & Zijlstra, C., 1998, Unusual and strongly structured sequence variation in a complex satellite DNA family from the nematode Meloidogyne chitwoodi, *Journal of molecular evolution*, 46(2), pp. 225-33.
de Chastonay, Y., Müller, F. & Tobler, H., 1990, Two highly reiterated nucleotide sequences in the low C-value genome of Panagrellus redivivus, *Gene*, 93(2), pp. 199-204.
Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R. & Jiang, J., 2002, Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon, *The Plant cell*, 14(8), pp. 1691-704.
Clarke, L., Amstutz, H., Fishel, B. & Carbon, J., 1986, Analysis of centromeric DNA in the fission yeast Schizosaccharomyces pombe, *Proceedings of the National Academy of Sciences of the United States of America*, 83(21), pp. 8253-7.
Claycomb, J.M., Batista, P.J., Pang, K.M., Gu, W., Vasale, J.J., van Wolfswinkel, J.C., Chaves, D.A., Shirayama, M., Mitani, S., Ketting, R.F., Conte, D. & Mello, C.C., 2009, The Argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation, *Cell*, 139(1), pp. 123-34.
Collet, C., 1984, Highly repeated DNA and holocentric chromosomes of the woodrush Luzula flaccida (Juncaceae), *Canadian journal of genetics and cytology*, 26(3), pp. 288-95.
Cummings, W.J., Yabuki, M., Ordinario, E.C., Bednarski, D.W., Quay, S. & Maizels, N., 2007, Chromatin structure regulates gene conversion, *PLoS biology*, 5(10), p. e246.
Dawe, R.K., Bennetzen, J.L. & Hake, S. 2009, Maize Centromeres and Knobs (neocentromeres) Handbook of Maize, in Springer New York, pp. 239-50.
Dawe, R.K., Reed, L.M., Yu, H.G., Muszynski, M.G. & Hiatt, E.N., 1999, A maize homolog of mammalian CENPC is a constitutive component of the inner kinetochore, *The Plant cell*, 11(7), pp. 1227-38.
Dechat, T., Pfleghaar, K., Sengupta, K., Shimi, T., Shumaker, D.K., Solimando, L. & Goldman, R.D., 2008, Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin, *Genes & development*, 22(7), pp. 832-53.
Dimitriadis, E.K., Weber, C., Gill, R.K., Diekmann, S. & Dalal, Y., 2010, Tetrameric organization of vertebrate centromeric nucleosomes, *Proceedings of the National Academy of Sciences of the United States of America*, 107(47), pp. 20317-22.
Ding, R., West, R.R., Morphew, D.M., Oakley, B.R. & McIntosh, J.R., 1997, The spindle pole body of Schizosaccharomyces pombe enters and leaves the nuclear envelope as the cell cycle proceeds, *Molecular biology of the cell*, 8(8), pp. 1461-79.
Dover, G., 1982, Molecular drive: a cohesive mode of species evolution, *Nature*, 299(5879), pp. 111-7.
Du, J., Tian, Z., Hans, C.S., Laten, H.M., Cannon, S.B., Jackson, S.A., Shoemaker, R.C. & Ma, J., 2010, Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison, *The Plant journal : for cell and molecular biology*, 63(4), pp. 584-98.
Earnshaw, W.C. & Cleveland, D.W., 2013, CENP-A and the CENP nomenclature: response to Talbert and Henikoff, *Trends in Genetics*.
Earnshaw, W.C., Allshire, R.C., Black, B.E., Bloom, K., Brinkley, B.R., Brown, W., Cheeseman, I.M., Choo, K.H.A., Copenhaver, G.P. & DeLuca, J.G., 2013, Esperanto for histones: CENP-A, not CenH3, is the centromeric histone H3 variant, *Chromosome Research*, pp. 1-6.
Eichler, E.E., Clark, R.A. & She, X., 2004, An assessment of the sequence gaps: unfinished business in a finished human genome, *Nature Reviews Genetics*, 5(5), pp. 345-54.

Elde, N.C., Roach, K.C., Yao, M.C. & Malik, H.S., 2011, Absence of positive selection on centromeric histones in Tetrahymena suggests unsuppressed centromere: drive in lineages lacking male meiosis, *Journal of molecular evolution*, 72(5-6), pp. 510-20.
Ellermeier, C., Higuchi, E.C., Phadnis, N., Holm, L., Geelhood, J.L., Thon, G. & Smith, G.R., 2010, RNAi and heterochromatin repress centromeric meiotic recombination, *Proceedings of the National Academy of Sciences of the United States of America*, 107(19), pp. 8701-5.
Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J. & van Steensel, B., 2010, Systematic protein location mapping reveals five principal chromatin types in Drosophila cells, *Cell*, 143(2), pp. 212-24.
Fischer, C., Ozouf-Costaz, C., Roest Crollius, H., Dasilva, C., Jaillon, O., Bouneau, L., Bonillo, C., Weissenbach, J. & Bernot, A., 2000, Karyotype and chromosome location of characteristic tandem repeats in the pufferfish Tetraodon nigroviridis, *Cytogenet Cell Genet*, 88(1-2), pp. 50-5.
Fishman, L. & Saunders, A., 2008, Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers, *Science (New York, N.Y.)*, 322(5907), pp. 1559-62.
Fishman, L. & Willis, J.H., 2005, A novel meiotic drive locus almost completely distorts segregation in mimulus (monkeyflower) hybrids, *Genetics*, 169(1), pp. 347-53.
Fitzgerald-Hayes, M., Clarke, L. & Carbon, J., 1982, Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs, *Cell*, 29(1), p. 235.
Fukagawa, T., Nogami, M., Yoshikawa, M., Ikeno, M., Okazaki, T., Takami, Y., Nakayama, T. & Oshimura, M., 2004, Dicer is essential for formation of the heterochromatin structure in vertebrate cells, *Nature cell biology*, 6(8), pp. 784-91.
Funabiki, H., Hagan, I., Uzawa, S. & Yanagida, M., 1993, Cell cycle-dependent specific positioning and clustering of centromeres and telomeres in fission yeast, *The Journal of cell biology*, 121(5), pp. 961-76.
Furuyama, S. & Biggins, S., 2007, Centromere identity is specified by a single centromeric nucleosome in budding yeast, *Proceedings of the National Academy of Sciences of the United States of America*, 104(37), pp. 14706-11.
Gassmann, R., Rechtsteiner, A., Yuen, K.W., Muroyama, A., Egelhofer, T., Gaydos, L., Barron, F., Maddox, P., Essex, A., Monen, J., Ercan, S., Lieb, J.D., Oegema, K., Strome, S. & Desai, A., 2012, An inverse relationship to germline transcription defines centromeric chromatin in C. elegans, *Nature*.
Gent, J.I. & Dawe, R.K., 2012, RNA as a structural and regulatory component of the centromere, *Annual review of genetics*, 46, pp. 443-53.
Grenier, E., Laumond, C. & Abad, P., 1996, Molecular characterization of two species-specific tandemly repeated DNAs from entomopathogenic nematodes Steinernema and Heterorhabditis (Nematoda:Rhabditida), *Mol Biochem Parasitol*, 83(1), pp. 47-56.
Grewal, S.I., 2010, RNAi-dependent formation of heterochromatin and its diverse functions, *Current opinion in genetics & development*, 20(2), pp. 134-41.
Gu, Y., Yam, C. & Oliferenko, S., 2012, Divergence of mitotic strategies in fission yeasts, *Nucleus (Austin, Tex.)*, 3(3), pp. 220-5.
Guenatri, M., Bailly, D., Maison, C. & Almouzni, G., 2004, Mouse centric and pericentric satellite repeats form distinct functional heterochromatin, *The Journal of cell biology*, 166(4), pp. 493-505.
Haaf, T. & Willard, H.F., 1997, Chromosome-specific alpha-satellite DNA from the centromere of chimpanzee chromosome 4, *Chromosoma*, 106(4), pp. 226-32.

Haizel, T., Lim, Y.K., Leitch, A.R. & Moore, G., 2005, Molecular analysis of holocentric centromeres of Luzula species, *Cytogenetic and genome research*, 109(1-3), pp. 134-43.
Hall, S.E., Kettler, G. & Preuss, D., 2003, Centromere satellites from Arabidopsis populations: maintenance of conserved and variable domains, *Genome research*, 13(2), pp. 195-205.
Han, Y., Zhang, Z., Liu, C., Liu, J., Huang, S., Jiang, J. & Jin, W., 2009, Centromere repositioning in cucurbit species: implication of the genomic impact from centromere activation and inactivation, *Proceedings of the National Academy of Sciences of the United States of America*, 106(35), pp. 14937-41.
Hayden, K.E., Newton, Y., Jain, M., Altemose, N., Willard, H.F. & Kent, J., 2013a, Complete sequence representation across human X and Y centromeric regions, *arXiv preprint arXiv: 1307.0035*.
Hayden, K.E., Strome, E.D., Merrett, S.L., Lee, H.R., Rudd, M.K. & Willard, H.F., 2013b, Sequences associated with centromere competency in the human genome, *Molecular and cellular biology*, 33(4), pp. 763-72.
Heger, P., Marin, B. & Schierenberg, E., 2009, Loss of the insulator protein CTCF during nematode evolution, *BMC molecular biology*, 10, p. 84.
Henikoff, S., 2002, Near the edge of a chromosome's "black hole", *Trends in genetics : TIG*, 18(4), pp. 165-7.
Henikoff, S., Ahmad, K. & Malik, H.S., 2001, The centromere paradox: stable inheritance with rapidly evolving DNA, *Science (New York, N.Y.)*, 293(5532), pp. 1098-102.
Heslop-Harrison, J.S., Murata, M., Ogura, Y., Schwarzacher, T. & Motoyoshi, F., 1999, Polymorphisms and genomic organization of repetitive DNA from centromeric regions of Arabidopsis chromosomes, *The Plant cell*, 11(1), pp. 31-42.
Heus, J.J., Zonneveld, B.J., de Steensma, H.Y. & van den Berg, J.A., 1993, The consensus sequence of Kluyveromyces lactis centromeres shows homology to functional centromeric DNA from Saccharomyces cerevisiae, *Molecular & general genetics : MGG*, 236(2-3), pp. 355-62.
Heus, J.J., Zonneveld, B.J., Steensma, H.Y. & Van den Berg, J.A., 1990, Centromeric DNA of Kluyveromyces lactis, *Current genetics*, 18(6), pp. 517-22.
Heus, J.J., Zonneveld, B.J., Steensma, H.Y. & Van den Berg, J.A., 1994, Mutational analysis of centromeric DNA elements of Kluyveromyces lactis and their role in determining the species specificity of the highly homologous centromeres from K. lactis and Saccharomyces cerevisiae, *Molecular & general genetics : MGG*, 243(3), pp. 325-33.
Hill, C.A., Guerrero, F.D., Van Zee, J.P., Geraci, N.S., Walling, J.G. & Stuart, J.J., 2009, The position of repetitive DNA sequence in the southern cattle tick genome permits chromosome identification, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 17(1), pp. 77-89.
Hou, H., Zhou, Z., Wang, Y., Wang, J., Kallgren, S.P., Kurchuk, T., Miller, E.A., Chang, F. & Jia, S., 2012, Csi1 links centromeres to the nuclear envelope for centromere clustering, *The Journal of cell biology*, 199(5), pp. 735-44.
Houseley, J. & Tollervey, D., 2011, Repeat expansion in the budding yeast ribosomal DNA can occur independently of the canonical homologous recombination machinery, *Nucleic acids research*, 39(20), pp. 8778-91.
Joglekar, A.P., Bouck, D., Finley, K., Liu, X., Wan, Y., Berman, J., He, X., Salmon, E.D. & Bloom, K.S., 2008, Molecular architecture of the kinetochore-microtubule attachment site is conserved between point and regional centromeres, *The Journal of cell biology*, 181(4), pp. 587-94.

Joglekar, A.P., Bouck, D.C., Molk, J.N., Bloom, K.S. & Salmon, E.D., 2006, Molecular architecture of a kinetochore-microtubule attachment site, *Nature cell biology*, 8(6), pp. 581-5.
Kanizay, L.B., Albert, P.S., Birchler, J.A. & Dawe, R.K., 2013, Intragenomic conflict between the two major knob repeats of maize, *Genetics*, 194(1), pp. 81-9.
Karpen, G.H. & Allshire, R.C., 1997, The case for epigenetic effects on centromere identity and function, *Trends in genetics : TIG*, 13(12), pp. 489-96.
Kimura, M. & Ohta, T., 1978, Stepwise mutation model and distribution of allelic frequencies in a finite population, *Proceedings of the National Academy of Sciences*, 75(6), pp. 2868-72.
Koo, D.H., Nam, Y.W., Choi, D., Bang, J.W., de Jong, H. & Hur, Y., 2010, Molecular cytogenetic mapping of Cucumis sativus and C. melo using highly repetitive DNA sequences, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 18(3), pp. 325-36.
Kullman, B., Tamm, H. & Kullman, K., 2005, Fungal genome size database, *Institute of Agricultural and Environmental Sciences, Estonian Agricultural University, Tartu, Estonia. Available on the internet: http://www. zbi. ee/fungal-genomesize. Reference*, 2, p. 2005.
Kurnit, D.M., 1979, Satellite DNA and heterochromatin variants: the case for unequal mitotic crossing over, *Human genetics*, 47(2), pp. 169-86.
Kuznetsova, I., Podgornaya, O. & Ferguson-Smith, M.A., 2006, High-resolution organization of mouse centromeric and pericentromeric DNA, *Cytogenetic and genome research*, 112(3-4), pp. 248-55.
Lampson, M.A. & Cheeseman, I.M., 2011, Sensing centromere tension: Aurora B and the regulation of kinetochore function, *Trends in cell biology*, 21(3), pp. 133-40.
Lara-Gonzalez, P., Westhorpe, F.G. & Taylor, S.S., 2012, The spindle assembly checkpoint, *Current biology : CB*, 22(22), pp. R966-80.
Lenzmeier, B.A. & Freudenreich, C.H., 2003, Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair, *Cytogenetic and genome research*, 100(1-4), pp. 7-24.
Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L. & Venter, J.C., 2007, The diploid genome sequence of an individual human, *PLoS biology*, 5(10), p. e254.
Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.P., Wang, Z., Chinwalla, A.T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K.D., Fronick, C., Schmidt, H., Fulton, L.A., Fulton, R.S., Nelson, J.O., Magrini, V., Pohl, C., Graves, T.A., Markovic, C., Cree, A., Dinh, H.H., Hume, J., Kovar, C.L., Fowler, G.R., Lunter, G., Meader, S., Heger, A., Ponting, C.P., Marques-Bonet, T., Alkan, C., Chen, L., Cheng, Z., Kidd, J.M., Eichler, E.E., White, S., Searle, S., Vilella, A.J., Chen, Y., Flicek, P., Ma, J., Raney, B., Suh, B., Burhans, R., Herrero, J., Haussler, D., Faria, R., Fernando, O., Darré, F., Farré, D., Gazave, E., Oliva, M., Navarro, A., Roberto, R., Capozzi, O., Archidiacono, N., Della Valle, G., Purgato, S., Rocchi, M., Konkel, M.K., Walker, J.A., Ullmer, B., Batzer, M.A., Smit, A.F., Hubley, R., Casola, C., Schrider, D.R., Hahn, M.W., Quesada, V., Puente, X.S., Ordoñez, G.R., López-Otín, C., Vinar, T., Brejova, B., Ratan, A., Harris, R.S., Miller, W., Kosiol, C., Lawson, H.A., Taliwal, V., Martins, A.L., Siepel, A., Roychoudhury, A., Ma, X., Degenhardt, J., Bustamante, C.D., Gutenkunst, R.N., Mailund, T., Dutheil, J.Y., Hobolth, A., Schierup, M.H., Ryder, O.A., Yoshinaga, Y., de

Jong, P.J., Weinstock, G.M., Rogers, J., Mardis, E.R., Gibbs, R.A. & Wilson, R.K., 2011, Comparative and demographic analysis of orang-utan genomes, *Nature*, 469(7331), pp. 529-33.
Lopes da Rosa, J., Holik, J., Green, E.M., Rando, O.J. & Kaufman, P.D., 2011, Overlapping regulation of CenH3 localization and histone H3 turnover by CAF-1 and HIR proteins in Saccharomyces cerevisiae, *Genetics*, 187(1), pp. 9-19.
Lu, Y.J., Kochert, G.D., Isenhour, D.J. & Adang, M.J., 1994, Molecular characterization of a strain-specific repeated DNA sequence in the fall armyworm Spodoptera frugiperda (Lepidoptera: Noctuidae), *Insect Mol Biol*, 3(2), pp. 123-30.
Lunyak, V.V., 2008, Boundaries. Boundaries...Boundaries??? *Current opinion in cell biology*, 20(3), pp. 281-7.
Lynch, D.B., Logue, M.E., Butler, G. & Wolfe, K.H., 2010, Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres, *Genome biology and evolution*, 2, pp. 572-83.
Mandrioli, M., Bizzaro, D., Manicardi, G.C., Gionghi, D., Bassoli, L. & Bianchi, U., 1999, Cytogenetic and molecular characterization of a highly repeated DNA sequence in the peach potato aphid Myzus persicae, *Chromosoma*, 108(7), pp. 436-42.
Mandrioli, M., Manicardi, G.C. & Marec, F., 2003, Cytogenetic and molecular characterization of the MBSAT1 satellite DNA in holokinetic chromosomes of the cabbage moth, Mamestra brassicae (Lepidoptera), *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 11(1), pp. 51-6.
Marshall, O.J. & Choo, K.H., 2009, Neocentromeres come of age, *PLoS genetics*, 5(3), p. e1000370.
Marshall, O.J., Chueh, A.C., Wong, L.H. & Choo, K.H., 2008, Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution, *Am J Hum Genet*, 82(2), pp. 261-82.
Maske, C. & Vaux, D.J., 2000, Structure-Function relationships of the nuclear envelope, *Advances in Structural Biology*, 6, pp. 261-98.
Mattei, S., Sampaolese, B., De Santis, P. & Savino, M., 2002, Nucleosome organization on Kluyveromyces lactis centromeric DNAs, *Biophysical chemistry*, 97(2-3), pp. 173-87.
McAinsh, A.D. & Meraldi, P., 2011, The CCAN complex: linking centromere specification to control of kinetochore-microtubule dynamics, *Seminars in cell & developmental biology*, 22(9), pp. 946-52.
McAinsh, A.D., Tytell, J.D. & Sorger, P.K., 2003, Structure, function, and regulation of budding yeast kinetochores, *Annu Rev Cell Dev Biol*, 19, pp. 519-39.
Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J.F., Derisi, J.L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I. & Chan, S.W., 2013, Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution, *Genome biology*, 14(1), p. R10.
Melters, D.P., Paliulis, L.V., Korf, I.F. & Chan, S.W., 2012, Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 20(5), pp. 579-93.
Mendiburo, M.J., Padeken, J., Fülöp, S., Schepers, A. & Heun, P., 2011, Drosophila CENH3 is sufficient for centromere formation, *Science (New York, N.Y.)*, 334(6056), pp. 686-90.
Meraldi, P., McAinsh, A.D., Rheinbay, E. & Sorger, P.K., 2006, Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins, *Genome biology*, 7(3), p. R23.

Mestrović, N., Randig, O., Abad, P., Plohl, M. & Castagnone-Sereno, P., 2005, Conserved and variable domains in satellite DNAs of mitotic parthenogenetic root-knot nematode species, *Gene*, 362, pp. 44-50.
Mutafova, T., Dimitrova, Y. & Komandarev, S., 1982, The karyotype of four Trichinella species, *Zeitschrift für Parasitenkunde (Berlin, Germany)*, 67(1), pp. 115-20.
Müller, F., Walker, P., Aeby, P., Neuhaus, H., Felder, H., Back, E. & Tobler, H., 1982, Nucleotide sequence of satellite DNA contained in the eliminated genome of Ascaris lumbricoides, *Nucleic acids research*, 10(23), pp. 7493-510.
Nachman, M.W., 2002, Variation in recombination rate across the genome: evidence and implications, *Current opinion in genetics & development*, 12(6), pp. 657-63.
Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, C.R. & Jiang, J., 2004, Sequencing of a rice centromere uncovers active genes, *Nature genetics*, 36(2), pp. 138-45.
Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C.R., Cheng, Z. & Jiang, J., 2005, Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice, *Molecular biology and evolution*, 22(4), pp. 845-55.
Neumann, P., Navrátilová, A., Schroeder-Reiter, E., Koblížková, A., Steinbauerová, V., Chocholová, E., Novák, P., Wanner, G. & Macas, J., 2012, Stretching the rules: monocentric chromosomes with multiple centromere domains, *PLoS genetics*, 8(6), p. e1002777.
Niedermaier, J. & Moritz, K.B., 2000, Organization and dynamics of satellite and telomere DNAs in Ascaris: implications for formation and programmed breakdown of compound chromosomes, *Chromosoma*, 109(7), pp. 439-52.
Nonaka, N., Kitajima, T., Yokobayashi, S., Xiao, G., Yamamoto, M., Grewal, S.I. & Watanabe, Y., 2002, Recruitment of cohesin to heterochromatic regions by Swi6/HP1 in fission yeast, *Nature cell biology*, 4(1), pp. 89-93.
Novak, U., 1984, Structure and properties of a highly repetitive DNA sequence in sheep, *Nucleic acids research*, 12(5), pp. 2343-50.
Oakey, R. & Tyler-Smith, C., 1990, Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males, *Genomics*, 7(3), pp. 325-30.
Ohta, T. & Dover, G.A., 1984, The cohesive population genetics of molecular drive, *Genetics*, 108(2), pp. 501-21.
Olszak, A.M., van Essen, D., Pereira, A.J., Diehl, S., Manke, T., Maiato, H., Saccani, S. & Heun, P., 2011, Heterochromatin boundaries are hotspots for de novo kinetochore formation, *Nature cell biology*, 13(7), pp. 799-808.
O'Neill, R.J. & Carone, D.M., 2009, The role of ncRNA in centromeres: a lesson from marsupials, *Progress in molecular and subcellular biology*, 48, pp. 77-101.
Padmanabhan, S., Thakur, J., Siddharthan, R. & Sanyal, K., 2008, Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, Candida albicans and Candida dubliniensis, *Proceedings of the National Academy of Sciences of the United States of America*, 105(50), pp. 19797-802.
Perpelescu, M. & Fukagawa, T., 2011, The ABCs of CENPs, *Chromosoma*, 120(5), pp. 425-46.
Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., Bland, M.J., Wagstaff, W., Dalton, S., McDevitt, T.C., Sen, R., Dekker, J., Taylor, J. & Corces, V.G., 2013, Architectural protein subclasses shape 3D organization of genomes during lineage commitment, *Cell*, 153(6), pp. 1281-95.

Piotte, C., Castagnone-Sereno, P., Bongiovanni, M., Dalmasso, A. & Abad, P., 1994, Cloning and characterization of two satellite DNAs in the low-C-value genome of the nematode Meloidogyne spp, *Gene*, 138(1-2), pp. 175-80.
Piras, F.M., Nergadze, S.G., Magnani, E., Bertoni, L., Attolini, C., Khoriauli, L., Raimondi, E. & Giulotto, E., 2010, Uncoupling of satellite DNA and centromeric function in the genus Equus, *PLoS genetics*, 6(2), p. e1000845.
Piras, F.M., Nergadze, S.G., Poletto, V., Cerutti, F., Ryder, O.A., Leeb, T., Raimondi, E. & Giulotto, E., 2009, Phylogeny of horse chromosome 5q in the genus Equus and centromere repositioning, *Cytogenetic and genome research*, 126(1-2), pp. 165-72.
Plohl, M., Luchetti, A., Mestrović, N. & Mantovani, B., 2008, Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin, *Gene*, 409(1-2), pp. 72-82.
Rago, F. & Cheeseman, I.M., 2013, Review series: The functions and consequences of force at kinetochores, *The Journal of cell biology*, 200(5), pp. 557-65.
Ravi, M. & Chan, S.W., 2010, Haploid plants produced by centromere-mediated genome elimination, *Nature*, 464(7288), pp. 615-8.
Raychaudhuri, N., Dubruille, R., Orsi, G.A., Bagheri, H.C., Loppin, B. & Lehner, C.F., 2012, Transgenerational propagation and quantitative maintenance of paternal centromeres depends on Cid/Cenp-A presence in Drosophila sperm, *PLoS biology*, 10(12), p. e1001434.
Rechtsteiner, A., Ercan, S., Takasaki, T., Phippen, T.M., Egelhofer, T.A., Wang, W., Kimura, H., Lieb, J.D. & Strome, S., 2010, The histone H3K36 methyltransferase MES-4 acts epigenetically to transmit the memory of germline gene expression to progeny, *PLoS genetics*, 6(9), p. e1001091.
Rhind, N., Chen, Z., Yassour, M., Thompson, D.A., Haas, B.J., Habib, N., Wapinski, I., Roy, S., Lin, M.F., Heiman, D.I., Young, S.K., Furuya, K., Guo, Y., Pidoux, A., Chen, H.M., Robbertse, B., Goldberg, J.M., Aoki, K., Bayne, E.H., Berlin, A.M., Desjardins, C.A., Dobbs, E., Dukaj, L., Fan, L., FitzGerald, M.G., French, C., Gujja, S., Hansen, K., Keifenheim, D., Levin, J.Z., Mosher, R.A., Müller, C.A., Pfiffner, J., Priest, M., Russ, C., Smialowska, A., Swoboda, P., Sykes, S.M., Vaughn, M., Vengrova, S., Yoder, R., Zeng, Q., Allshire, R., Baulcombe, D., Birren, B.W., Brown, W., Ekwall, K., Kellis, M., Leatherwood, J., Levin, H., Margalit, H., Martienssen, R., Nieduszynski, C.A., Spatafora, J.W., Friedman, N., Dalgaard, J.Z., Baumann, P., Niki, H., Regev, A. & Nusbaum, C., 2011, Comparative functional genomics of the fission yeasts, *Science (New York, N.Y.)*, 332(6032), pp. 930-6.
Rocchi, M., Archidiacono, N., Schempp, W., Capozzi, O. & Stanyon, R., 2012, Centromere repositioning in mammals, *Heredity*, 108(1), pp. 59-67.
Roest Crollius, H., Jaillon, O., Dasilva, C., Ozouf-Costaz, C., Fizames, C., Fischer, C., Bouneau, L., Billault, A., Quetier, F., Saurin, W., Bernot, A. & Weissenbach, J., 2000, Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis, *Genome research*, 10(7), pp. 939-49.
Roizès, G., 2006, Human centromeric alphoid domains are periodically homogenized so that they vary substantially between homologues. Mechanism and implications for centromere functioning, *Nucleic acids research*, 34(6), pp. 1912-24.
Rudd, M.K. & Willard, H.F., 2004, Analysis of the centromeric regions of the human genome assembly, *Trends in genetics : TIG*, 20(11), pp. 529-33.
Rudd, M.K., Wray, G.A. & Willard, H.F., 2006, The evolutionary dynamics of alpha-satellite, *Genome research*, 16(1), pp. 88-96.

Ruggiero, B.L. & Topal, M.D., 2004, Triplet repeat expansion generated by DNA slippage is suppressed by human flap endonuclease 1, *The Journal of biological chemistry*, 279(22), pp. 23088-97.
Saffery, R., Sumer, H., Hassan, S., Wong, L.H., Craig, J.M., Todokoro, K., Anderson, M., Stafford, A. & Choo, K.H., 2003, Transcription within a functional human centromere, *Molecular cell*, 12(2), pp. 509-16.
Samonte, R.V., Ramesh, K.H. & Verma, R.S., 1997, Comparative mapping of human alphoid satellite DNA repeat sequences in the great apes, *Genetica*, 101(2), pp. 97-104.
Sandler, L. & Novitski, E., 1957, Meiotic drive as an evolutionary force, *American Naturalist*, pp. 105-10.
Sanyal, K., Baum, M. & Carbon, J., 2004, Centromeric DNA sequences in the pathogenic yeast Candida albicans are all different and unique, *Proceedings of the National Academy of Sciences of the United States of America*, 101(31), pp. 11374-9.
Schneider, R. & Grosschedl, R., 2007, Dynamics and interplay of nuclear architecture, genome organization, and gene expression, *Genes & development*, 21(23), pp. 3027-43.
Schueler, M.G., Dunn, J.M., Bird, C.P., Ross, M.T., Viggiano, L., Rocchi, M., Willard, H.F. & Green, E.D., 2005, Progressive proximal expansion of the primate X chromosome centromere, *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), pp. 10563-8.
Shang, W.H., Hori, T., Toyoda, A., Kato, J., Popendorf, K., Sakakibara, Y., Fujiyama, A. & Fukagawa, T., 2010, Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences, *Genome research*, 20(9), pp. 1219-28.
Sharma, A. & Presting, G.G., 2008, Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity, *Molecular genetics and genomics : MGG*, 279(2), pp. 133-47.
Shi, J., Wolf, S.E., Burke, J.M., Presting, G.G., Ross-Ibarra, J. & Dawe, R.K., 2010, Widespread gene conversion in centromere cores, *PLoS biology*, 8(3), p. e1000327.
Smith, G.P., 1976a, Evolution of repeated DNA sequences by unequal crossover, *Science*, 191(4227), pp. 528-35.
Smith, K.M., Galazka, J.M., Phatale, P.A., Connolly, L.R. & Freitag, M., 2012, Centromeres of filamentous fungi, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 20(5), pp. 635-56.
Song, J.S., Liu, X., Liu, X.S. & He, X., 2008, A high-resolution map of nucleosome positioning on a fission yeast centromere, *Genome research*, 18(7), pp. 1064-72.
De Souza, C.P. & Osmani, S.A., 2007, Mitosis, not just open or closed, *Eukaryotic Cell*, 6(9), pp. 1521-7.
De Souza, C.P., Horn, K.P., Masker, K. & Osmani, S.A., 2003, The SONB(NUP98) nucleoporin interacts with the NIMA kinase in Aspergillus nidulans, *Genetics*, 165(3), pp. 1071-81.
De Souza, C.P., Osmani, A.H., Hashmi, S.B. & Osmani, S.A., 2004, Partial nuclear pore complex disassembly during closed mitosis in Aspergillus nidulans, *Current biology : CB*, 14(22), pp. 1973-84.
Spakulová, M., Králová, I. & Cutillas, C., 1994, Studies on the karyotype and gametogenesis in Trichuris muris, *Journal of helminthology*, 68(1), pp. 67-72.
Spence, J.M., Blackman, R.L., Testa, J.M. & Ready, P.D., 1998, A 169-base pair tandem repeat DNA marker for subtelomeric heterochromatin and chromosomal rearrangements in aphids of

the Myzus persicae group, *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 6(3), pp. 167-75.
Steiner, N.C., Hahnenberger, K.M. & Clarke, L., 1993, Centromeres of the fission yeast Schizosaccharomyces pombe are highly variable genetic loci, *Molecular and cellular biology*, 13(8), pp. 4578-87.
Straight, A.F., 1997, Cell cycle: checkpoint proteins and kinetochores, *Current biology : CB*, 7(10), pp. R613-6.
Streeck, R.E., Moritz, K.B. & Beer, K., 1982, Chromatin diminution in Ascaris suum: nucleotide sequence of the eliminated satellite DNA, *Nucleic acids research*, 10(11), pp. 3495-502.
Sudman, P.D. & Greenbaum, I.F., 1990, Unequal crossing over and heterochromatin exchange in the X-Y bivalents of the deer mouse, Peromyscus beatae, *Chromosoma*, 99(3), pp. 183-9.
Sullivan, B. & Karpen, G., 2001, Centromere identity in Drosophila is not determined in vivo by replication timing, *The Journal of cell biology*, 154(4), pp. 683-90.
Sullivan, B.A., Schwartz, S. & Willard, H.F., 1996, Centromeres of human chromosomes, *Environmental and molecular mutagenesis*, 28(3), pp. 182-91.
Sun, X., Le, H.D., Wahlstrom, J.M. & Karpen, G.H., 2003, Sequence analysis of a functional Drosophila centromere, *Genome research*, 13(2), pp. 182-94.
Takayama, Y., Sato, H., Saitoh, S., Ogiyama, Y., Masuda, F. & Takahashi, K., 2008, Biphasic incorporation of centromeric histone CENP-A in fission yeast, *Molecular biology of the cell*, 19(2), pp. 682-90.
Talbert, P.B. & Henikoff, S., 2013, Phylogeny as the basis for naming histones, *Trends in genetics: TIG*.
Talbert, P.B., Ahmad, K., Almouzni, G., Ausió, J., Berger, F., Bhalla, P.L., Bonner, W.M., Cande, W.Z., Chadwick, B.P., Chan, S.W., Cross, G.A., Cui, L., Dimitrov, S.I., Doenecke, D., Eirin-López, J.M., Gorovsky, M.A., Hake, S.B., Hamkalo, B.A., Holec, S., Jacobsen, S.E., Kamieniarz, K., Khochbin, S., Ladurner, A.G., Landsman, D., Latham, J.A., Loppin, B., Malik, H.S., Marzluff, W.F., Pehrson, J.R., Postberg, J., Schneider, R., Singh, M.B., Smith, M.M., Thompson, E., Torres-Padilla, M.E., Tremethick, D.J., Turner, B.M., Waterborg, J.H., Wollmann, H., Yelagandula, R., Zhu, B. & Henikoff, S., 2012, A unified phylogeny-based nomenclature for histone variants, *Epigenetics & chromatin*, 5, p. 7.
Tarès, S., Cornuet, J.M. & Abad, P., 1993, Characterization of an unusually conserved AluI highly reiterated DNA sequence family from the honeybee, Apis mellifera, *Genetics*, 134(4), pp. 1195-204.
Teschke, C., Solleder, G. & Moritz, K.B., 1991, The highly variable pentameric repeats of the AT-rich germline limited DNA in Parascaris univalens are the telomeric repeats of somatic chromosomes, *Nucleic acids research*, 19(10), pp. 2677-84.
Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-I, T., Toyoda, A., Fujiyama, A., Tarutani, Y. & Kakutani, T., 2012, Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata, *Genes & development*, 26(7), pp. 705-13.
Van de Vosse, D.W., Wan, Y., Wozniak, R.W. & Aitchison, J.D., 2011, Role of the nuclear envelope in genome organization and gene expression, *Wiley interdisciplinary reviews. Systems biology and medicine*, 3(2), pp. 147-66.
Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blocker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Della Valle, G., Fryc, S.,

Guerin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Roed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvanen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Lander, E.S. & Lindblad-Toh, K., 2009, Genome sequence, comparative analysis, and population genetics of the domestic horse, *Science (New York, N.Y.)*, 326(5954), pp. 865-7.
Warburton, P.E. & Willard, H.F., 1992, PCR amplification of tandemly repeated DNA: analysis of intra- and interchromosomal sequence variation and homologous unequal crossing-over in human alpha satellite DNA, *Nucleic acids research*, 20(22), pp. 6033-42.
Warburton, P.E. & Willard, H.F., 1995, Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages, *Journal of molecular evolution*, 41(6), pp. 1006-15.
Warburton, P.E., Greig, G.M., Haaf, T. & Willard, H.F., 1991, PCR amplification of chromosome-specific alpha satellite DNA: definition of centromeric STS markers and polymorphic analysis, *Genomics*, 11(2), pp. 324-33.
Waye, J.S. & Willard, H.F., 1987, Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes, *Nucleic acids research*, 15(18), pp. 7549-69.
Wevrick, R. & Willard, H.F., 1989, Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability, *Proceedings of the National Academy of Sciences of the United States of America*, 86(23), pp. 9394-8.
Willard, H.F. & Waye, J.S., 1987, Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat, *Journal of molecular evolution*, 25(3), pp. 207-14.
Wu, J., Yamagata, H., Hayashi-Tsugane, M., Hijishita, S., Fujisawa, M., Shibata, M., Ito, Y., Nakamura, M., Sakaguchi, M., Yoshihara, R., Kobayashi, H., Ito, K., Karasawa, W., Yamamoto, M., Saji, S., Katagiri, S., Kanamori, H., Namiki, N., Katayose, Y., Matsumoto, T. & Sasaki, T., 2004, Composition and structure of the centromeric region of rice chromosome 8, *The Plant cell*, 16(4), pp. 967-76.
Wu, Z.A., Liu, W.X., Murphy, C. & Gall, J., 1990, Satellite 1 DNA sequence from genomic DNA of the giant panda Ailuropoda melanoleuca, *Nucleic acids research*, 18(4), p. 1054.
Yan, H., Talbert, P.B., Lee, H.R., Jett, J., Henikoff, S., Chen, F. & Jiang, J., 2008, Intergenic locations of rice centromeric chromatin, *PLoS biology*, 6(11), p. e286.
Zhang, W., Lee, H.-R., Koo, D.-H. & Jiang, J., 2008, Epigenetic Modification of Centromeric Chromatin: Hypomethylation of DNA Sequences in the CENH3-Associated Chromatin in Arabidopsis thaliana and Maize, *The Plant Cell Online*, 20(1), pp. 25-34.
Zhao, X., Lu, J., Zhang, Z., Hu, J., Huang, S. & Jin, W., 2011, Comparison of the distribution of the repetitive DNA sequences in three variants of Cucumis sativus reveals their phylogenetic relationships, *Journal of genetics and genomics = Yi chuan xue bao*, 38(1), pp. 39-45.
Zhong, C.X., Marshall, J.B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J.A., Jiang, J. & Dawe, R.K., 2002, Centromeric retroelements and satellites interact with maize kinetochore protein CENH3, *The Plant cell*, 14(11), pp. 2825-36.
Zimmerman, S., Daga, R.R. & Chang, F., 2004, Intra-nuclear microtubules and a mitotic spindle orientation checkpoint, *Nature cell biology*, 6(12), pp. 1245-6.