Conflict begets complexity: the evolution of centromeres
Harmit S Malik†‡ and Steven Henikoff*†§

Centromeres mediate the faithful segregation of eukaryotic chromosomes. Yet they display a remarkable range in size and complexity across eukaryotes, from ~125 bp in budding yeast to megabases of repetitive satellites in human chromosomes. Mapping the fine-scale structure of complex centromeres has proven to be daunting, but recent studies have provided a first glimpse into this unexplored bastion of our genomes and the evolutionary pressures that shape it. Evolutionary studies of proteins that bind centromeric DNA suggest genetic conflict as the underlying basis of centromere complexity, drawing interesting parallels with the myriad selfish elements that employ centromeric activity for their own survival.

Addresses
*Howard Hughes Medical Institute, †Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA
‡e-mail: hmalki@fhcrc.org
§e-mail: steveh@fhcrc.org

0959-437X/02/$ — see front matter © 2002 Elsevier Science Ltd. All rights reserved.

Published online 1 October 2002

Centromeres might not be nearly as complex if they hadn't started competing during sex.

Introduction
The post-genomic era has fueled a surge of interest in how mutation and selection have shaped our genes and genomes. However, the study of centromeres has lagged behind. This is largely a result of technical difficulties in accurately assembling the highly repetitive arrays of satellite sequences that constitute centromeres in most organisms. Despite these difficulties, recent studies have begun to provide ‘evolutionary snapshots’ of the centromere. They suggest that different sequence variants jockey for evolutionary dominance, even as homogeneous arrays of satellite repeats are destroyed by the insertion of a variety of mobile elements. Parallel studies of centromere-binding proteins also suggest that competition may drive the sequence complexity at centromeres, and may be responsible for rapidly changing karyotypes throughout evolution.

Centromeres: simple and complex
Centromeres play an essential role in every eukaryote — that is, faithful segregation of chromosomes when cells divide. Yet the sequences underlying centromere function show a remarkably wide range of complexity across different phyla (Figure 1). The simplest centromeres are the ‘point centromeres’ of budding yeast, followed by the ‘regional centromeres’ of plants and animals that consist of hundreds of kilobases of repetitive satellites. Finally, in holokinetic organisms such as nematodes, centromeric determinants are dispersed throughout the length of the chromosome. It is perplexing to find such diverse solutions to the same problem, suggesting that different evolutionary pressures are responsible for this range in complexity. In spite of this diversity, all centromere types are exclusively packaged in nucleosomes containing specialized centromeric histone H3 (CenH3) molecules, allowing an unambiguous assignment of centromere identity.

The simple, point centromeres of the budding yeast, Saccharomyces cerevisiae, are ~125 bp long, consisting of conserved DNA elements: CDE I, CDE II and CDE III. CDE I and CDE III are binding sites for proteins that recruit other components of the centromere–kinetochore apparatus [1] and, not surprisingly, are subject to strong purifying selection. CDE II, on the other hand, is under relatively few constraints except a minimum length and AT bias, which might contribute to preferential binding of nucleosomes containing the centromeric histone [2]. This simple solution to centromere function is atypical, even for other fungi. Centromeres in the fission yeast Schizosaccharomyces pombe are considerably more complex [3,4]. They vary between 35–110 kb in length and have a common organization (Figure 1), consisting of a central core region of non-repetitive sequence, cnt, flanked by two classes of repeats. This definition of centromeric regions may be too broad; the centromeres (packaged by CenH3) are likely to be restricted to the cnt sequences with the surrounding repeats serving important pericentric roles, including binding of sister cohesins [5*,6*].

The S. pombe centromere organization resembles ‘regional’ centromeres of plants and animals, where a centromeric ‘core’ is flanked by pericentric heterochromatin. Unlike S. pombe, ‘cores’ in more complex centromeres are themselves composed almost entirely of repetitive satellite elements, including the α-satellite in humans and the 180-bp repeat in Arabidopsis thaliana [7]. Studies on the centromere of a minichromosome Dp (1;f) 1187 from Drosophila melanogaster reveal a mosaic structure, consisting of distinct blocks of simple satellite repeats interspersed with blocks of more complex sequences, which are frequently transposable elements [8]. The demarcation of the ‘core’ from flanking heterochromatin is more difficult in these cases as both are frequently composed of similar satellite sequences, and because both are important for chromosome segregation.

Evolution of centromeric satellites
Their highly repetitive nature makes centromeric satellites subject to the typical evolutionary forces that govern tandem repeats in genomes — that is, mutation, recombination (unequal crossing over and gene conversion), deletion and translocation [9]. The best-studied centromeric satellites...
Centromeres: simple to complex. (a) Centromeres in *S. cerevisiae* are ~125 bp long, and consist of three DNA elements: CDEI, CDEII and CDEIII. CDEI and III encode protein binding sites and flank CDEII, which is AT-rich and is likely packaged by the centromeric nucleosome. A variation on this theme is found in *Kluyveromyces lactis* [50]. Its centromeres include not only CDEI and CDEII (that is twice as long) but also an additional conserved DNA segment, CDE0, of as yet unknown function. (b) *S. pombe* centromeric regions [3,4] are 35–110 kb long and consist of a central core, cnt, that is 4.1 kb long, of which a 1.4 kb region is 48% identical across the 3 chromosomes. cnt is flanked by the 5.6 kb inner repeats (imr) that are often interrupted by tRNA genes, followed by the outer repeats (otr) that consist of the dg (4.4 kb), cen253 (0.3 kb) and dh (4.8 kb) repeats. The dg repeats are the most well conserved elements across the three centromeric regions, being ~97% identical. Notably, along with the cnt, dg repeats are the only component essential for centromere function in *S. pombe* [51]. (c) The human X centromeric region is 2–4 Mb long and consists of a mostly homogeneous array of DXZ1 α-satellite [23**]. Flanking the DXZ1 array, sequence homogeneity increases in step-wise increments from ~70% (light shading) to ~98% (dark) over a 40 kb stretch. Line-1 insertions provide a useful ‘paleontological’ marker for the age of the various satellites; whereas pericentric satellites have both ancient (light shading) and young (dark) Line1 insertions, the DXZ1 satellite has only one young Line1 insertion. Although the DXZ1 homogeneous satellite array is 2–4 Mb long, only a 500 kb uninterrupted array of DXZ1 is sufficient for centromeric function [23**].

- α-satellite DNA is formed by tandem head-to-tail reiteration of a ~170 bp unit that can be found in monomer or higher-order arrangements [10]. Chromosome-specific arrangements of α-satellites can be identified in great apes (although not in lower primates) indicating some recombinational isolation between non-homologous chromosomes that has retarded infiltration of higher-order repeats across chromosomes. Consistent with this, the Y chromosome, which is most recombinationally isolated, does not share any α-satellite arrays with other chromosomes. Sequence analysis of the α-satellites on various chromosomes from related species reveals cases where orthology is maintained — that is, related satellite variants are found on homologous chimpanzee and human chromosomes [11]. Cases are also found where this orthology has not been maintained clearly [12–14].

What are the selective constraints on the α-satellite? Some segments of the α-satellite repeat are especially well-conserved; these have been proposed to be the binding sites for the DNA-binding protein, CENP-B, and are hence called the CENP-B boxes [15]. Although the CENP-B box is not preserved in α-satellites of primates [16], the CENP-B protein is well conserved! Genetic knockouts of the CENP-B gene in mice display no centromeric defects [17]; thus CENP-B’s general role in centromere function is unclear. CENP-B is homologous to the integrase domains of transposable elements [18,19]. This has led to the tantalizing model that they represent two components of what was an active transposon system, with CENP-B transposons recognizing themselves at the CENP-B transposons recognizing themselves at the CENP-B integrase domain.

The enigmatic relationships of centromeric satellites and a variety of mobile elements can be extended to other species. Among plants, for example, a clade of Ty3-gypsy like retrotransposons, including *Athila* in *A. thaliana*, is found in close proximity with the centromeric 180 bp repeats [20,21*]. It may be impossible to ascertain the evolutionary origins of the 180 bp repeats themselves because of multiple episodes of mutation and homogenization. Segments of mobile elements are represented in the centromeric regions of many plant species. Whether these segments represent an intermediate stage between intact mobile element and centromeric satellites is still...
unclear. In *Drosophila*, a Het-A/TART retrotransposon-related array, typically found at telomeric ends, is now found at the Y centromere of both *D. melanogaster* and *D. simulans* [22], although it is unclear whether this is a centromeric or pericentric satellite. It should, therefore, not be surprising if it turns out that different classes of mobile elements have been readily recruited for centromeric function, on the basis of their abundance in individual lineages.

**Evolutionary landscape of the human X centromere**

One of the main problems in accurately assembling large arrays of repetitive satellites is the lack of unique features that serve as ‘marker buoys’ in a sea of homogeneous sequence. To overcome this problem, the Willard group have employed unique sequence-tagged sites to successfully order the satellite-containing bacterial artificial chromosomes using *in situ* hybridization, despite an overall low sequence diversity of the human X centromeric region [23••]. This assembly reveals an exquisite evolutionary landscape comprising a gradient of sequence conservation at the human X centromere (Figure 1). The most homogeneous sequences of the DXZ1 satellite are found at the middle of the array, whereas more degenerate copies of the satellite are found towards the edges of the array, close to the euchromatin–heterochromatin boundary. The DXZ1 homogeneous subfamily in the middle is 2–4 Mb long and is also the youngest; it arose at the orangutan–gorilla split, ~7 million years ago. Furthermore, whereas the α-satellites towards the edges of the array have been heavily peppered by ancient and young Line-1 insertions, the DXZ1 array contains only one, very recent Line-1 insertion. A 500 kb homogeneous array of DXZ1 satellite is sufficient to support segregation of artificial chromosomes [23••]. Thus, the DXZ1 subfamily that has arisen most recently has expanded to become the central core of the human X centromere.

It is apparent that recombinogenic events are more frequent in the middle of the array compared to the ends, keeping the central DXZ1 subfamily at high-sequence homogeneity and relatively devoid of insertions (Figure 2). The surrounding satellites, which may have served centromeric function in a primate ancestor, are now relegated to serve a pericentric role, suffering constant insertions of retroelements [24]. The study by Schueler *et al.* thus provides an invaluable snapshot of the constantly changing satellite sequences at centromeres. On the one hand, new sequence variants arise by mutation and are either fixed or eliminated stochastically; on the other, transposable element insertions constantly disrupt uninterrupted arrays of satellite sequences, with selection for centromere function imposing a minimum length threshold for an uninterrupted array. These two evolutionary forces can quickly reshape the sequence architecture of centromeric regions [25].

The human X centromere provides valuable insights for the delineation of centromeric regions in other organisms. For example, in *A. thaliana*, the non-recombining regions encompass both centromeres and pericentric heterochromatin [26]. Efforts to assemble the *Arabidopsis* centromere have not progressed enough to elucidate the structure of its centromere ‘core’. Indeed, the size estimates of the centromeric and pericentric regions have changed dramatically (e.g. from 2.2 to 5.3 Mb for Chromosome 4 [21*]), highlighting the problems of assembling highly repetitive regions using bacterial and yeast artificial chromosomes. Probably, final elucidation of the centromere ‘core’ will come from a combination of different experimental approaches including restriction-mapping, assembly with unique sequence features (as in the human X centromere study [23••]), and cytological analysis [21*,27*].
Neocentromeres

Human neocentromeres represent previously non-centromeric chromosomal segments that have fully developed centromeric competence and are able to sponsor mitotic stability of chromosomal fragments that lack endogenous centromeres. Strikingly, many human neocentromeres are almost entirely non-repetitive and completely devoid of α-satellite [28•]. They highlight what has become increasingly evident from several studies: centromere inheritance relies on a chromatin state [29], most likely conferred by the presence of CenH3 [30••]. The comparison of non-centromeric ancestral sequences and neocentromeric ones are beginning to reveal general commonalities about neocentromeric origins [28•,31]. Human neocentromeres, which are typically sampled from patient-derived cell lines, are not randomly distributed. In particular, chromosomal segment 13q is a ‘hotspot’ for neocentromere formation [32]. It is still unclear whether a ‘hotspot’ is an ancestral centromere location that is simply re-activated, or, more likely, a chromosomal region with sequence features that make it especially conducive to centromere function. This remains an exciting area of active study.

From an evolutionary standpoint, neocentromeres are not expected to contribute significantly to the evolution of species karyotypes — they arise in chromosomal rearrangements that could be extremely deleterious, and often have to compete with existing centromeres for centromeric status. However, there is at least one case in which a neocentromere may have replaced an endogenous centromere. In a recent comparison of the X chromosomes from humans, black lemurs and ring-tailed lemurs, Rocchi and co-workers used in situ hybridization to show that the three chromosomes are isosequential throughout their lengths [33•]. However, each of these X chromosomes have their centromeres in different locations. As no evidence has been found suggesting that the new centromeres have ‘transposed’ from other centromeric regions, neocentromere emergence together with degeneration of the old centromere is the most parsimonious explanation for this repositioning [33•].

Evolution of centromere-binding proteins

It is apparent that the low-throughput study of individual centromeres will not lend itself easily to analyses of centromere evolution. As a surrogate to directly investigating centromeric DNA evolution, recent studies have begun to examine the evolution of proteins that specifically bind at centromeric sequences. Prominent among these are studies of the centromeric histones, or CenH3s. CenH3s are
members of the H3 family of histones that are found at the centromeres of every eukaryote [30••]. CenH3 evolution in *Drosophila* and *Arabidopsis* [34•,35] reveals that these proteins are subject to positive selection, apparently driven by changing DNA-specificity. This is an unexpected result given that histones are among the most highly constrained proteins in eukaryotes. Indeed, it is not clear at first glance why two essential components of the chromosome segregation machinery (CenH3s and centromeres) have not co-evolved to an optimum state. This is the ‘centromere paradox’ in a nutshell [30••].

The positive selection of CenH3s in both plants and flies is best explained by likening the CenH3–centromere system to a host–parasite relationship [36]. The centromeric satellites have an opportunity to either ‘drive’ or increase their proportion over random segregation where there is asymmetric transmission of homologous chromosomes (i.e. female meiosis). In both metazoans and plants, a preferred position in female meiosis (except the Y), karyotypes can change dramatically solely on the basis of female meiosis transmission. Closely related species display major ‘switches’ in karyotype, rapidly attaining divergent configurations. It can be inferred that metacentric chromosomes have been rapidly ‘switched’ from preferred to non-preferred, and vice-versa. Regardless of which way the polarity goes, ‘mixed karyotypes’ would be rare and transient in nature, as is the case in mammalian karyotypes [38••].

The components of meiotic polarity have not been defined, but involve some level of hormonal control [39].
For example, in several bird species it may be possible to influence offspring sex on the basis of which sex chromosome survives female meiosis: either the male-determining Z or the female-determining W. This decision is made pre-fertilization (in early meiosis), appears to be correlated with hormone levels in the yolk, and does not involve selective abortion or reabsorption of eggs after ovulation [40].

This transmission advantage in female meiosis has not escaped the attention of conventional selfish genetic elements. The earliest selfish genetic elements to be characterized were the extra-centromeric satellite arrays in maize, called knobs [41]. Under the ‘driving’ genetic background, Ab10, preferential segregation occurs when crossing over between the knob and centromere produces a heteromorphic dyad. Knobs bind microtubules, as a result of which knob-bearing chromatids are pulled toward the spindle poles in meiosis I and eventually toward the outermost megaspores, one of which will become the gametophyte and produce gametes [42]. Instead of a 50% expected ratio of transmission, knob transmission varies from 59 to 82% and correlates with the size of the satellite array. The presence of maize knobs and their chromosomal locations are best explained under models wherein the knobs are ‘driving’ in female meiosis despite being deleterious in a ‘non-driving’ background [43]. The satellite repeats found at knobs share significant homology with maize centromeric satellites [44]. It appears that at least two independent satellite arrays and their binding proteins may have co-evolved to exploit this advantage [45•].

Several maize races also have a supernumerary B chromosome that typically undergoes nondisjunction during the second microspore division resulting in two kinds of sperm, one with two B chromosomes and one with no B chromosomes. The sperm with the two B chromosomes preferentially fertilizes the egg cell whereas the other fuses to form the endosperm, leading to a transmission advantage for the B chromosome in male meiosis. However, the size of the centromere-related satellite array is correlated with female meiotic transmission, suggesting that a selective threshold maintains a minimum length satellite array at maize B chromosomes, just as it is likely to at endogenous centromeres [46]. The B chromosome centromere contains sequences homologous to the maize chromosome 4 centromeric sequence, suggesting that a purely selfish centromere has derived from a conventional one [44,47•].

Conclusions
In yeast that have symmetric meioses, centromere competition is not expected to occur at all; removal of this genetic conflict may have allowed the optimal co-evolution of centromeric histones and centromeres, along with the gradual simplification of the centromeric sequences themselves. Under this model, S. cerevisiae centromeres, which are believed to consist of one nucleosome each, represent the ultimate stage of centromere optimization, whereas other genomes, including our own, constantly struggle with the consequences of unfair advantages in female meiosis.

Update
In humans, the bias in favor of transmitting Robertsonian fusions in female meiosis has been documented [37,36••], but Daniel et al. [48,49] also reiterate another dramatic effect of Robertsonians — reduced male fertility. Among families with Robertsonian arrangements coming to prenatal diagnosis, there are 2.4 fold fewer male parent carriers compared to female parents. This is despite the fact that in their progeny there is an ~1:1 ratio of male:female transmission of Robertsonian rearrangements. This points to a significant decline in fertility in male carriers of Robertsonian fusions, compared to female carriers. This duality (i.e. increased chromosomal transmission in female meiosis offset by lowered male fertility) provides strong support for the centromere-drive model.

Acknowledgements
We thank Kumi Ahmad, Sue Biggins, Jenny Cooper, Erin McKintrick, Paul Talbert and Danielle Vermaak for helpful discussions and comments on the manuscript. We are supported by the Helen Hay Whitney Foundation (HS Malik) and the Howard Hughes Medical Institute (S Henikoff).

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:
• of special interest
**of outstanding interest
Conflict begets complexity: the evolution of centromeres

Malik and Henikoff


16. Haaf T, Mater AG, Wienberg J, Ward DC: Increase estimates by forces that constantly shape centromeric regions. The authors used refined restriction mapping to re-size the centromeres to lower estimates of repeat number in the chimpanzee. Proc Natl Acad Sci USA 1996, 93:1443-1448.


Two different repeat classes are found in maize knobs. This study reveals that both of these, together with their respective binding proteins, constitute independent drive systems. Experiments with drugs that arrest microtubule dynamics but do not affect knob neocentromeric activity also suggest that the binding proteins are most likely responsible for motor activity.


This study shows that the primary constriction of chromosome 4 has a major representation of a repeat that is related to the B-specific centromere sequence, confirming the evolutionary relationship between the centromeres of chromosome 4 and the B chromosome. It also discusses alternative models for the origin of B chromosomes in maize.


