A Genetic Basis for Human Centromeres

May 18, 2015

GE Zentner

Accurate chromosome segregation relies upon the kinetochore, a protein complex that attaches to specific sites on chromatin known as centromeres. In humans, centromeres lie within arrays of repeated a-satellite sequences, making analysis of their function, identity, and evolution difficult. Most annotated a-satellite arrays are higher-order repeats (HORs), which comprise tandem copies of an a-satellite array, itself composed of multiple tandem ~170-base pair (bp) units with varying degrees of divergence from one another. For instance, the X chromosome DXZ1 HOR is comprised of tandem copies of a 12-repeat array, wherein each repeat unit is on average only 77% identical to its neighbors. As a-satellite DNA is estimated to comprise 2-3% of the human genome, or ~500,000 a-satellite repeats per haploid genome, ~20,000 a-satellites are expected on each chromosome. However, the assembled a-satellite sequences represent nearly 100 times less a-satellite sequence than expected.

To identify functional centromeres, data analyst Jorja Henikoff, postdoctoral fellow Jitendra Thakur, and graduate student Sivakanthan Kasinathan in the laboratory of Dr. Steven Henikoff (Basic Sciences Division) identified DNA sequences associated with kinetochore proteins and introduced an unbiased computational approach to characterize functional centromeric sequences. "Conventional sequence assembly methods are unable to map the most homogeneous repetitive sequences, leaving gaps in the ~170-bp a-satellite repeat arrays, 'black holes' that each of our centromeres is embedded in," said Dr. Henikoff. "We realized that by isolating the sequences bound to centromere-specific chromatin proteins we could identify centromere sequences by function. We found that two dimeric alpha satellite units dominate the most homogeneous, and therefore the youngest, a-satellite repeats bound by centromere-specific chromatin proteins. Remarkably, these dimeric units showed precise positioning of chromatin particles, which protect 100 base pairs centered over each of the units of the dimer, whereas a-satellite units that have been mapped to the edges the gaps showed little if any precise positioning."

To enrich for presumably functional centromeric sequences, the authors performed chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) for the kinetochore proteins CENP-A and CENP-C. Because highly repetitive sequences are difficult to map unambiguously, most genome assembly methods skip over them, resulting in missing sequences in the reference

genome. As a-satellite arrays are not assembled into the human genome, the authors first aligned the DNA associated with these kinetochore proteins to bacterial artificial chromosomes (BACs) known to contain a-satellite arrays. The most homogenous, and therefore youngest, a-satellite sequences displayed the most robust enrichment of CENP-A and CENP-C, and decreased occupancy by these proteins was associated with increased degeneracy (evolutionary age) of a-satellite repeats. The authors thus hypothesized that young, homogenous sequences were functional centromere sequences.

The authors next used a phylogenetic approach to identify the most abundant sequences recovered in the CENP-A ChIP-seq to further define functional centromeric sequences. One group of these sequences matched a-satellite arrays from chromosomes 1, 5, and 19, and another group matched sequences from chromosomes 13, 14, 21, and 22. The authors then generated a consensus sequence for each group: Cen1-like and Cen13-like. Several sequences also matched unplaced clones from the human genome assembly. These abundant sequences were composed of two a-satellite repeats flanking a CENP-B box, the binding site of the only identified mammalian sequence-specific centromere-binding protein.

To analyze the effect of a-satellite dimers on chromatin organization, the authors aligned their CENP-A and CENP-C ChIP-seq data to the Cen1-like and Cen13-like dimeric consensus sequences. This revealed a striking pattern of CENP-A enrichment, wherein two ~100 bp 'pillars' of CENP-A flank a 'pedestal' containing the CENP-B box. In contrast, alignment of these ChIP-seq data to previously the most abundant consensus sequence from old, degenerate HORs revealed a different arrangement, with a single poorly positioned particle adjacent to a CENP-B pedestal. From this, the authors concluded that young, but not old, a-satellite arrays precisely position CENP-A nucleosomes.

This work strongly supports the idea that young, homogeneous a-satellite arrays are functional centromeric sequences and also challenges an established idea about how human centromeres are defined. "Our finding that functional human centromeres are dominated by specific sequences with a unique chromatin conformation argues that human centromeres are defined by DNA sequence, in contrast to the common perception that they are epigenetically defined," said Dr. Henikoff.

Henikoff JG, Thakur J, Kasinathan S, Henikoff S. 2015. A unique chromatin complex occupies young a-satellite arrays of human centromeres. *Sci Adv* 1(1):e1400234.
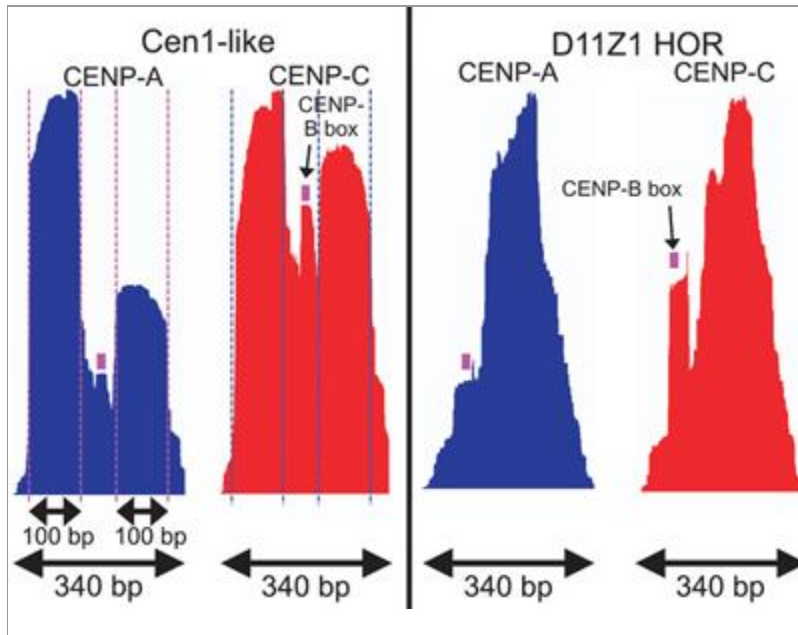
Image from the publication

(Left) At young a-satellite dimers, such as the Cen1-like consensus, two CENP-A containing nucleosomes are precisely positioned, protecting ~100 bp of DNA on either side of a CENP-B pedestal. (Right) At old a-satellite dimers, such as the D11Z1 HOR, only a single poorly positioned CENP-A-containing nucleosome is found adjacent to a CENP-B pedestal.