



## Centromere reference models for human chromosomes X and Y satellite arrays

Karen H Miga, Yulia Newton, Miten Jain, et al.

*Genome Res.* published online February 5, 2014

Access the most recent version at doi:[10.1101/gr.159624.113](https://doi.org/10.1101/gr.159624.113)

---

<b>P&lt;P</b>	Published online February 5, 2014 in advance of the print journal.
<b>Accepted Manuscript</b>	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
<b>Creative Commons License</b>	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <a href="http://genome.cshlp.org/site/misc/terms.xhtml">http://genome.cshlp.org/site/misc/terms.xhtml</a> ). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at <a href="http://creativecommons.org/licenses/by-nc/3.0/">http://creativecommons.org/licenses/by-nc/3.0/</a> .
<b>Email Alerting Service</b>	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a> .

---

---

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Published by Cold Spring Harbor Laboratory Press

# Centromere reference models for human chromosomes X and Y satellite arrays

**SHORT TITLE:** Linear sequence models of human centromeric DNA

**Karen H. Miga<sup>1,2</sup>, Yulia Newton<sup>2</sup>, Miten Jain<sup>2</sup>, Nicolas Altemose<sup>1</sup>, Huntington F. Willard<sup>1</sup> and W. James Kent<sup>2\*</sup>**

<sup>1</sup> Duke Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina, United States of America.

<sup>2</sup> Center for Biomolecular Science & Engineering, University of California Santa Cruz, Santa Cruz, California, United States of America

\*Address for Correspondence:

W. James Kent  
(831) 459-1401  
Center for Biomolecular Science & Engineering  
University of California  
MS: CBSE-ITI  
1156 High Street  
Santa Cruz, CA 95064  
kent@soe.ucsc.edu

**KEY WORDS:** centromere, satellite DNA, assembly, population variation, tandem repeats

## ABSTRACT

The human genome sequence remains incomplete, with multi-megabase-sized gaps representing the endogenous centromeres and other heterochromatic regions. Available sequence-based studies within these sites in the genome have demonstrated a role in centromere function and chromosome pairing, necessary to ensure proper chromosome segregation during cell division. A common genomic feature of these regions is the enrichment of long arrays of near-identical tandem repeats, known as satellite DNAs, that offer a limited number of variant sites to differentiate individual repeat copies across millions of bases. This substantial sequence homogeneity challenges available assembly strategies and, as a result, centromeric regions are omitted from ongoing genomic studies. To address this problem, we utilize monomer sequence and ordering information obtained from whole-genome shotgun reads to model two haploid human satellite arrays on chromosomes X and Y, resulting in an initial characterization of 3.83 Mb of centromeric DNA within an individual genome. To further expand the utility of each centromeric reference sequence models, we evaluate sites within the arrays for short-read mappability and chromosome specificity. Because satellite DNAs evolve in a concerted manner, we use these centromeric assemblies to assess the extent of sequence variation among 366 individuals from distinct human populations. We thus identify two satellite array variants in both X and Y centromeres, as determined by array length and sequence composition. This study provides an initial sequence characterization of a regional centromere and establishes a foundation to extend genomic characterization to these sites as well as to other repeat-rich regions within complex genomes.

## INTRODUCTION

Extensive tracts of near-identical tandem repeats, known as satellite DNA arrays, are associated with constitutive heterochromatin and commonly provide the sequence definition for regional centromeres, or sites responsible for chromosome segregation (Yunis and Yasmineh 1971; Willard 1990; Schueler et al. 2001). Proper regulation of these sites is critical for cellular viability, as disruption in epigenetic maintenance often leads to genome instability and aneuploidy (Dernburg et al. 1996; Peng and Karpen 2008; Ting et al. 2011; Zhu et al. 2011). Despite their biological importance, the satellite DNA-rich regions remains incomplete and largely uncharacterized even within extensively sequenced and studied genomes, resulting in large, multi-megabase gaps within each chromosome assembly (Henikoff 2002; Eichler et al. 2004; Rudd and Willard 2004). This presents a fundamental challenge to ongoing genomic studies aimed at understanding the role of these specialized domains in cellular function and emphasizes the need for a more complete representation of sequences that comprise highly homogenized arrays.

Efforts to predict the linear sequence arrangement within satellite arrays are impeded by insufficient and sparsely arranged sites capable of distinguishing one copy of the repeat from another, resulting in an increase in read coverage and assembly collapse (Durfy and Willard 1989; Schueler et al. 2001; Schindelhauer and Schwarz 2002; Treangen and Salzberg 2011). The extreme homogeneity further exacerbates the challenges faced by BAC-based sequence assembly, resulting in only marginal satellite DNA directly adjacent to heterochromatic and centromeric gaps in most reference assemblies (Eichler et al. 2004; Rudd and Willard 2004). Although these variant sites present a methodological challenge, their representation is necessary to study mechanisms of array evolution (Willard and Waye 1987; Santos et al. 1995; Warburton and Willard 1995), improve long-range physical maps (Wevrick and Willard 1989; Luce et al. 2006), and map sites of

epigenetic enrichment that are important for centromere function (Maloney et al. 2012; Hayden et al. 2013). Thus, much effort has been devoted to designing approaches that are capable of deciphering satellite arrays.

Indeed, extensive experimental studies of human centromeric satellite arrays have provided a foundation on which to extend characterization of array sequence composition, organization, and evolution (Willard et al. 1989). Endogenous centromeres in the human genome are enriched with a single AT-rich satellite family, known as alpha satellite (Manuelidis 1978). Inherent sequence diversity between copies of the fundamental ~171-bp repeat unit enables alpha satellite monomer organization to be studied in a chromosome-specific manner (Willard 1985). High-resolution genomic characterization of alpha satellite organization reveals two general subtypes: those that appear to be highly divergent with few apparent local homology patterns, known as monomeric, and those monomers that are organized into tandemly-repeated, multi-monomer units, known as higher-order repeat (HOR) units (Willard and Waye 1987; Alexandrov et al. 1993). In contrast to monomeric alpha satellite, where near-identical tandem repeats are only occasionally observed and which provide little challenge to standard assembly efforts, HORs typically occupy multi-megabase-sized homogenized arrays that are vastly underrepresented in each chromosome assembly (International Human Genome Sequencing Consortium 2004; Rudd and Willard 2004; Hayden et al. 2013). As a result, sequences of chromosome-assigned HOR arrays are currently absent from the reference assembly (Rudd and Willard 2004; Hayden et al. 2013).

Characterization of satellite and non-satellite sequences that comprise a single array is necessary to initiate comparative analyses between individual genomes to study how these sequences change over time.

Individual alpha satellite arrays have been shown to vary considerably in array length and HOR sequence variants within the human population (Waye and Willard 1986; Durfy

and Willard 1989; Wevrick and Willard 1989; Oakey and Tyler-Smith 1990; Warburton et al. 1991; Santos et al. 1995). This level of sequence variability is found between maternally- and paternally-inherited chromosomes (Mahtani and Willard 1990; Warburton and Willard 1995; Roizès 2006), and in whole-genome data from diploid genomes; accordingly, it is difficult to infer the long-range sequence organization and assign sequence variants to a single array. The X and Y chromosomes, represented as haploid arrays in the male genome, offer a unique opportunity for high-resolution sequence definition of individual centromeric arrays. Experimental studies of these regions have provided an initial estimate of long-range sequence organization by physical mapping (Tyler-Smith and Brown 1987; Mahtani and Willard 1990; Mahtani and Willard 1998), sampled HOR homogeneity from sequencing within the array (Durfy and Willard 1989; Schueler et al. 2001; Schindelbauer and Schwarz 2002), and demonstrated the utility of a small number of array-assigned sequence variants, or experimental markers, to study centromere evolution (Durfy and Willard 1990; Laursen et al. 1992) and population genetics within a limited number of individuals (Oakey and Tyler-Smith 1990; Santos et al. 1995; Santos et al. 2000). These centromeric surveys offer a strong foundation upon which to build comprehensive satellite DNA descriptions across the entire centromeric region within and across a large number of human genomes.

Here we model the sequence composition and local monomer organization across two alpha satellite haploid arrays on chromosomes X and Y, presenting an initial characterization of sequences currently represented by centromeric gaps in the chromosome assemblies. To accomplish this, we have designed and implemented software (linearSat) that utilizes monomer sequence and relative order as observed from a whole-genome shotgun (WGS) read database (Levy et al. 2007b), to present a linear sequence model that describes sequence content within each satellite array. To demonstrate the utility of these reference sequences we have indicated those sequences that are represented only on the X or Y chromosome, defining chromosome-specific

markers useful in short-read mapping (Hayden et al. 2013). Additionally, we have provided annotation of satellite sequence copy number estimates to determine sites in the array that are low-copy and thus useful for extending long-range array characterization. Acknowledging the expected variability of satellite DNA arrays in the human population, we study this singular reference relative to 366 male, low-coverage genomes to perform a high-resolution analysis of sequence profiles and array length estimates (The 1000 Genomes Project Consortium 2012). This study provides an initial sequence characterization of regional centromeres from chromosomes X and Y and establishes a foundation for extending this method to study other satellite DNA arrays within complex genomes.

## RESULTS

### Algorithmic overview

Centromeric satellite DNAs are composed of tandem repeats that, apart from a limited number of variant sites, are identical across multi-megabase-sized arrays. This excess of sequence identity and an inability to determine the correct biological ordering of repeats has challenged previous assembly algorithms. Here, we provide an alternate approach to characterizing satellite DNA arrays; abandoning the need to determine the “true” linear order, we rather aimed to generate a linear sequence that models the observed variation and repeat order as identified in an initial database of high-quality, Sanger WGS reads.

This general sequence-processing pipeline is subdivided into three steps, as depicted for the centromere X alpha satellite array in Figure 1. The HOR sequence on the X chromosome (DXZ1) is described by a highly-homogenized, 12-monomer tandem repeat that spans the length of the centromeric gap in Figure 1a (Waye and Willard

1985; Mahtani and Willard 1990). A limited number of HOR sequence variants are observed across the entirety of the array, defined by single-copy nucleotide variants (as indicated in Figure 1 by a single-base change resulting in two HORs that are 99% identical within the pink and blue boxes), rearrangements resulting in a different monomer number and organization from the canonical repeat unit (highlighted in orange), and insertion of non-satellite sequences [as shown for the long interspersed element (long interspersed element, or LINE) in green]. To study the occurrence and frequency of such variation within a given array, it is first necessary to create a read database specific to the DXZ1 array (Figure 1b, step 1). Each HOR can be defined as an ordered arrangement of alpha satellite monomers (labeled m1–m12, with average length of 171 bp) that is repeated in a head-to-tail organization [as illustrated by the red line connecting m12 back to the start of the repeat (m1)]. Due to high sequence identity among copies of a given HOR, the majority of individual monomers appear identical to the corresponding consensus monomer sequence, noted as grey ovals. However, variant sites of single nucleotide change are shown in blue, rearrangement in monomer ordering in orange (where m7 is observed adjacent to m9, omitting m8 found in the canonical organization), and sequence insertion and deletion (as illustrated in green to mark the site of LINE insertion adjacent to m10) can be readily detected within the dataset. Thus, by generating an initial description of sequence content within each centromeric array one is able to document variant sites and the relative frequency of their occurrence (Durfy and Willard 1989).

Second, the DXZ1 sequence database can be reformatted as a bidirectional multi-graph (Medvedev and Brudno 2009), as illustrated in Figure 1b, step 2. All identical full-length monomers are compressed to represent a single node, thereby emphasizing those variant sites within the array. For example, m2, containing a single-base change from the consensus, can be subdivided into two monomer groups: m2v1 (containing five identical monomers) and m2v2 (defined by one monomer). Edges between nodes are

defined by local monomer adjacency and relative orientation, as observed in the initial read database. Edge weights provide the frequency of observed read adjacencies from each node. Junctions with non-alpha satellite sequence, as provided for the example of the LINE element, are catalogued including the partial, interrupted alpha satellite monomer.

The final step uses a second-order Markov chain model to generate sequences of monomers. This is built up by looking at the order of monomers in each read, and for each possible pair of adjacent monomers in the data, counting up the type and amount of third monomers that follow this pair (Rambaut and Grassly 1997). The algorithm generates a linearized monomer order, producing the first two monomers as a special case, and thereafter picking the next monomer randomly from all of the monomers observed to directly follow these two monomers, with the probability weighed according to the frequency of these following monomers observed in the read database. This provides a complete traversal of the sequence graph of size N, or the estimated array length provided by sequence coverage. The algorithm is designed to include each edge at least once and in proportion to the provided edge weights. When a monomer is randomly picked, the probability that it will be picked in subsequent rounds is reduced to increase the chance that rare monomers will be included in the output.

Given the homogeneity of the array, sites that differ from the consensus are expected to have low read coverage and occasionally present premature path termination. To ensure that such low-frequency variants are fully represented by the algorithm, artificial edge assignments link monomers relative to consensus-based ordering, relying on first-order information. Modeled arrangement of a centromeric array is intended to provide a more complete sequence definition of sequence variant and monomer ordering proportional to an initial unassembled read database (as shown in Figure 1b, step 3).

This approach while preserving local order and including both common and rare variation, does not necessarily depict the true long-range HOR ordering across the length of the inferred centromere array, nor the repeat orientation, and is not guided by the current reference assembly on the adjacent p- and q-arm. The resulting centromere array models do provide a linearized description of the sequences within a given satellite array read database, resulting in a genomic reference useful in extending mapping tools and functional annotation.

### **A comprehensive study of chromosome X and Y centromeric sequences**

To provide centromeric reference models of alpha satellite arrays on chromosomes X and Y, we prepared HOR (DXZ1 and DYZ3) read databases from a single male reference genome (HuRef) (Levy et al. 2007b)(see Methods). DXZ1 has a 12-mer HOR (Willard et al. 1983; Waye and Willard 1985), represented by 15,563 reads (totaling 13.9 Mb) and DYZ3 has a 34-mer HOR (Wolfe et al. 1985), represented by 1,008 reads (totaling 0.89 Mb). The array length estimates as determined by read depth in the HuRef genome, fall into an expected distribution of previous high-molecular weight pulse-field gel electrophoresis (PFGE) studies across a variety of cell lines for the DXZ1 array (with array lengths that vary between 1.3 and 3.7 Mb)(Mahtani and Willard 1990) and the DYZ3 array (ranging from 0.2 to 1.2 Mb)(Wevrick and Willard 1989; Oakey and Tyler-Smith 1990). In addition, we have validated PFGE array size estimates from the donor-matched cell line (data not shown), providing further support for the general findings within our study.

A study of alpha satellite sequences within each higher-order array read database provided evidence that the arrays are indeed highly homogenized (monomer global alignments to consensus; DXZ1: 97.4% average, with range 92.2-100%; DYZ3: 99.6%, with range 97.2-100%) and that interruptions in the array by non-satellite DNA are

exceedingly rare (six events in the 3.6 Mb DXZ1 array, and no detectable events across the DYZ3 array). Additionally, monitoring the directionality of the monomers on both single reads and between paired reads, we find no evidence of shifts in polarity, suggesting that the HORs are organized in a single orientation across the length of the entire array in the HuRef genome. Paired-read assessment across both the X and Y arrays would suggest that the majority of paired reads contain only HOR satellite sequences, with only a small fraction (< 1.0%) assigned to uniquely mapping sequences that can be confidently assigned to p- or q-arm. Thus, these data provide evidence, in line with previous studies (Tyler-Smith and Brown 1987; Mahtani and Willard 1998), for a single alpha-satellite array that spans the length of each centromere-assigned gap for chromosomes X and Y.

To study the occurrence and frequency of array HOR sequence variants, we reformatted the read database into full-length, high-quality monomers with notation relative to the consensus monomer ordering for the DXZ1 and DYZ3 repeat units. Total monomer libraries across the repeat were consistent in both DXZ1—with an average 3,583 monomers across the 12-mer HOR—and DYZ3—with an average 58 monomers across the 34-mer HOR (SFig1a). Compression of monomers into groupings based on strict identity (or 100% identity) revealed ~10-fold compression for DXZ1 (average 300 unique monomer types) and ~20-fold compression for DYZ3 (average 3 unique monomer types) (SFig1b). Sites that vary from each HOR-derived consensus appear to be relatively equal between transitions and transversions (as shown for DXZ1 in Figure 2), although we detect that the bases that differ from the consensus appear to deplete the total number of GC base pairs, thereby increasing AT-richness. Within the array, 8% of single-site changes are characterized by insertions or deletions relative to the consensus (insertion and deletion tracks, Figure 2). The majority (68%) of these sites are associated with the expansion or contraction of homopolymer sequences, with relatively equal changes associated with A and T nucleotides. Single- or multiple-base changes within the array

provide little evidence for variation in the length of each individual alpha satellite monomer, thus maintaining the individual alpha satellite repeat unit length.

Although the majority of reads support a canonical monomer ordering, we identify ten sites of HOR rearrangement in DXZ1 (of which two have been previously described) (Willard et al. 1983; Warburton et al. 1991), (Figure 2), and three sites in DYZ3 (not shown). Investigation of the DYZ3 HOR repeat identified a low-frequency, two-monomer insertion previously determined to be represented in a smaller proportion of the array in European genomes (Wolfe et al. 1985; Santos et al. 1995). Additionally, we detect evidence for local duplication of m13, and a rearrangement involving m11 and m13 (SFig1c).

The read databases for DXZ1 and DYZ3 were reformatted into a sequence graph (as shown in Figure 1), where nodes describe a grouping of identical full-length monomers, and edges between nodes are provided based on observed local monomer ordering within a single read. In summary, we have used this data structure to describe a read database, representing the census of sequences that comprise DXZ1 and DYZ3 arrays in the HuRef genome.

### **Centromere reference models for DXZ1 and DYZ3 alpha satellite arrays**

To generate a centromere reference model of each alpha satellite higher-order array, we designed the linearSat software (see Methods for link and description). This software utilizes a second-order Markov model to traverse each respective centromeric sequence graph, resulting in monomer ordering that is proportional to that observed within the initial read database. The software is sensitive to include low-coverage, variant sites within the array and employs a consensus-informed ordering that guides the extension of monomer ordering relative to the canonical repeat organization, within ambiguous

regions defined by low read depth. Junctions between alpha satellite monomers and non-alpha satellite sequences described within each sequence graph are represented as those high-quality bases ( $\geq$  a *phred* score of 20) adjacent to the full-length monomer, defined by partial alpha and non-alpha satellite sequences with an appended 100-bp gap (as first depicted in Figure 1b). Read-depth estimated array sizes for both DXZ1 and DYZ3 were used to set the threshold for Markov chain termination. In doing so, we determined the content of a 3.6-Mb DXZ1 array and a 0.23-Mb DYZ3 array, representing a full listing of all monomers in the proportion expected from the input.

To evaluate the accuracy of these results, we performed a comparison of each generated array sequence to the original read database. The linearSat software operates at the level of full-length monomers, thereby omitting information from partial monomers commonly found at the 5' and 3' ends of individual sequence reads. To account for the representation of these sequences in the final projection, we reformatted the generated linear sequence and those WGS reads from the original unassembled sequence database into windows of size *k* (where *k*=50–400, with a 1-bp slide in both strands), demonstrating an average positive predictive value of 94% and 95% across all lengths *k* for DXZ1 and DYZ3, respectively (Figure 3a). To evaluate each inferred centromere projection for a predicted monomer ordering that is not observed in the initial read dataset, we performed inverse analysis and determined a predictive value of 84% averaged across windows in DXZ1 and 94% in DYZ3 (Figure 3b).

Given the stochastic approach of using a generative Markov process, we do not expect to generate the true long-range linear order across the entire array. However, we hypothesize that this model is capable of correctly predicting regional sequence organization (defined as greater than the length of a single read) within the resulting linear sequence. To evaluate the long-range prediction, we studied concordant paired-read support between ordered HOR within each array, demonstrating that roughly 74%

of small plasmid inserts (with average insert size of 2 kb) in the DXZ1 array and 95% within the DYZ3 array have at least one concordant arrangement within the generated array. We hypothesize that longer reads, thereby increasing the model order, would greatly improve our confidence in true HOR ordering within the array. To test this, we simulated long reads from each DXZ1 and DYZ3 linearized array generated in this study and increased the model order accordingly. Blocks of correctly ordered monomers were determined against the initial DXZ1 and DYZ3 array representations. Evaluation of the maximum block length, or longest string that has an exact match with the initial array, is shown to increase with model order (within the range of 3 to 24 monomers, as shown in Figure 3c); demonstrating the ability to correctly predict the ordering of a megabase of the array (about a quarter of the estimated DXZ1 array size, 0.9 Mb) using a monomer model order of 22 (or the monomer order described on read lengths ~4 kb). Additionally, we show that extending the model order to ten (as observed within a ~2-kb read) we observe ~40% of all available 10 kb windows within the array are correctly represented (SFig2a). Identifying a discrete list of monomer blocks we determined the equivalent N50 values to indicate a linear improvement with an increased model order (SFig2b). Therefore, this method is currently capable of representing local monomer ordering and sequence composition within a satellite DNA array and is expected to improve long-range organization prediction with only an increase in read length.

### **Assessment of short read mappability across centromeric satellite arrays**

Alpha satellite DNAs are expected to share a basic sequence definition across all subsets and higher number of exact sequence alignments among closely-related HOR arrays (Alexandrov et al. 1988; Hayden et al. 2013). Such sequence homology is expected to challenge accurate mapping and interpretation of short-read datasets, common to epigenomic and population-based whole-genome sequencing studies (The 1000 Genomes Project Consortium 2012; The ENCODE Project Consortium 2012). To establish

array mappability [that is, to characterize those sequences specific to the DXZ1 and DYZ3 arrays, as described earlier (Hayden et al., 2013)], we reformatted each linearized centromeric array into a k-mer library (where  $k=24, 36, 50, 100$ -bp windows with a 1-bp slide) and identified those sequences that are found only within the DXZ1 or DYZ3 read database and lack an exact match with all remaining sequences in the HuRef genome. At the resolution of 24 bp, we detect 78% of the DXZ1 array and 49% of the DYZ3 array to be specific to those arrays, with a gain in array mappability with increasing k-mer length (SFig3). Additionally, a survey against 814 low-coverage genomes (The 1000 Genomes Project Consortium 2012) by sex demonstrates the specificity of DYZ3 24-mers to male individuals and a relative doubling of DXZ1 specific 24-mers when compared between females and males (SFig4), as expected. As a result, we have qualified those sites along the length of each centromeric array that are both present in the original dataset and are capable of ensuring array-specific mapping and annotation.

Short k-mers within each DXZ1 and DYZ3 array can be further studied within their relative abundance, thereby indicating those sites represented in the majority of HOR repeats as well as sites of low-frequency array variants. To annotate the prevalence of each k-mer library with respect to either the DXZ1 or DYZ3 sequence libraries, we estimated the frequency profiles across each linearized centromeric array. We determined 2.1% of DXZ1 24mers and 0.8% of DYZ3 24mers to be equivalent to single copy [present in original-read dataset at or below single-copy read depth estimates (see Methods) and observed less than three times in the inferred array]. Due to the expected high level of homogeneity within each array, the vast majority of sequences in both arrays are defined by consensus sequence (representing 89–92.3% of all k-mers for DXZ1 and DYZ3, respectively), with all remaining sequences representing intermediate variants that are not yet fixed within the context of the entire array. This initial sequence characterization—defining array specificity and sequence copy number—is

intended to qualify the interpretation and mapping capabilities in these highly repetitive regions, thereby allowing these sequences to be useful as a reference for ongoing genomic studies.

### **A study of centromeric array variation within human populations**

Satellite arrays are known to expand and contract through mechanisms underlying concerted evolution, resulting in substantial differences in array length and quantitative differences in higher-order repeat variants between individuals in the human population (Wevrick and Willard 1989; Mahtani and Willard 1990). To assess this level of variation in a larger set of individuals, we surveyed 366 male genomes from the 1000 Genomes Project to study sequence abundance and population-based signatures of haploid X and Y centromeric arrays (The 1000 Genomes Project Consortium 2012). Due to low-coverage, haploid sequence representation of X and Y arrays, we restricted our analysis to the top 75<sup>th</sup> percentile of the most frequent array-specific markers in the HuRef genome (1,546 unique 24-mers for DXZ1 and 1,837 unique 24-mers for DYZ3). To estimate pairwise similarity between individual arrays, we calculated the Euclidean distance between frequency vectors (24-mer normalized frequency profile for either DXZ1 or DYZ3 array), resulting in an  $n \times n$  (366 x 366) affinity matrix. We performed unsupervised clustering (see Methods) to predict two distinct array groups (group 1 and group 2), as illustrated in heat-map clustered matrices in Figure 4a, for both DXZ1 and DYZ3. To expand our study to identify those sequence features useful in classifying each array group (i.e., capable of distinguishing DYZ3 group 1 from group 2), we performed supervised learning models (Support vector machine, SVM) with leave-one-out cross-validation (Wang et al. 2006). As a result, we identified 138 24-mers within the 90<sup>th</sup> percentile that were capable of discriminating DXZ1 groups (selected features with accuracy scores in range of 0.96 – 0.99) and 166 for DYZ3 (accuracy scores in range of 0.86 – 0.91) (STbl1, SFig5). It is expected that higher sequence coverage and

longer reads will offer an increased resolution of mixing between each array group classification, as many low-frequency alleles, describing relatively new mutations that or intermediate signatures between the two groups, that are not currently included in this analysis.

To study the distribution of array lengths in the human population, we provide DXZ1 and DYZ3 size estimates for each haploid array (see Methods, SFig6), resulting in distributions that are largely concordant with previous experimental estimates from smaller numbers of individuals (Mahtani and Willard 1990; Oakey and Tyler-Smith 1990)(Table 1). The DYZ3 array is determined to have a mean size of 0.81 Mb and is observed to vary by over an order of magnitude (range, 0.1 – 2.2 Mb) (Wevrick and Willard 1989; Oakey and Tyler-Smith 1990). Likewise, we observe a mean array size of 3.2 Mb for DXZ1 (range: 0.5 – 4.9 Mb), in line with previous estimates (Mahtani and Willard 1990). When applying the previous array classification labels (group 1 and 2 for both DXZ1 and DYZ3) based on sequence signatures within each array, we identify groups of array lengths for both DXZ1 and DYZ3 that fall into two distinct bimodal distributions (t-test, DYZ3 p-value < 0.01 and DXZ1 p-value < 0.05)(Figure 4b, Table 1). Thus, we provide evidence for two predominant satellite array types in each of DXZ1 and DYZ3 that are defined by sequence composition and associated array length distribution (STbl2).

To investigate population-based patterns of satellite array inheritance, we subdivided the DYZ3 groups 1 and 2 within the context of 1000 Genomes population assignments (Figure 4c). In support of previous findings (Oakey and Tyler-Smith 1990), we determine that DYZ3 group 1 arrays are observed in high frequency in Western Europeans (GBR, TSI, IBS) and that group 2 arrays are observed to be more prevalent in Asians (CHB, CHS, JPT). When monitoring population assignment of the DXZ1 array groups we observe a higher prevalence of group 1 in Asian populations (CHB, JPT, CHS), with a frequency of

~0.5 within individuals tested, and within South American populations (MXL: 0.6, CLM: ~0.5). In line with the hypothesis that the inheritance of the X and Y centromeric arrays are independent of one another, we do not observe a statistical correlation between X and Y types relative to a simulated null.

In total, these data support the hypothesis that DXZ1 and DYZ3 arrays in early human populations could be subdivided into two general groups whose genetic signatures and array size are largely maintained in modern human populations. This suggests that mechanisms of conversion and unequal crossover greatly outweigh the influence of novel mutation or inter-homologue/chromosomal exchange.

## **DISCUSSION**

Addressing a long-standing technical problem for sequence assembly across genomic regions of highly homogenous repetitive DNA, here we provide an initial linear sequence using locally-ordered read assemblies of haploid human centromeric regions on the X and Y chromosomes. Within this analysis, we convert a comprehensive array sequence library into a sequence graph, thereby permitting documentation of the occurrence and frequency of sequence variants across the entirety of the array. To convert this data structure to a linear reference sequence model, we traverse a path through these centromeric sequence graphs to present repeat local ordering and array sequence variants in a manner proportional to that observed in the initial sequence read database. Further, we demonstrate by simulation the utility of this method to improve long-range ordering with a modest increase in read length. It is important to note that each linear representation provides an approximation of the true array sequence organization (as defined by the initial graph structure); however, the inferred array sequence is capable of providing a biologically rich description of array variants and local monomer organization as observed in the initial read dataset and is useful as a reference for

further genomic studies. Thus, this sequence characterization and linear representation of a regional centromere address a fundamental challenge in the genome sciences—the inability to generate a reference sequence across regions of homogenized satellite DNA. This work was intended to provide a high-resolution study of two haploid arrays. However, the method should be useful for generating a reference that represents pooled centromeric array sequence libraries from diploid chromosomes.

Genomic descriptions of human centromeric regions are necessary to promote studies of array sequence evolution and function. This requires not only single, robust array reference sequences, but also tools and annotations to guide confident and biologically meaningful alignments across highly repetitive regions. Satellite DNA arrays are identical across the majority of tandem repeats and, in addition, share stretches of identity with related satellite arrays distributed throughout the genome (Hayden et al. 2013). To address these issues and strengthen the utility of our reference array sequences, we identified all sites that could be localized with confidence to only the X (DXZ1) and Y (DYZ3) arrays. By thus establishing array-specific mappability, it is possible to study sequence maps within the shared HOR definition assigned to a chromosomally assigned array. The majority of these sites are shared among most, if not all, copies of the repeat within the array. To improve the resolution within each array, we have provided an index of each array-specific marker to include HOR frequency or an estimate of copy number within the HuRef genome. Such array annotation strengthens the utility of this reference database and will enable studies to extend from this singular read database to perform comparative estimates of array sequence organization within the human population.

Satellite DNA in centromeric regions had been previously shown to vary in size and proportion of HOR sequence variants within the human population (Wevrick and Willard 1989; Warburton and Willard 1992). Our evaluation of evolutionary patterns of the X and Y arrays across 366 male individuals from 14 distinct human populations (The 1000 Genomes Project Consortium 2012), reveals that DXZ1 and DYZ3 satellite arrays in

modern humans can each be classified into one of two groups defined by sequence composition and array size. The results for DYZ3 are concordant with previous experimental estimates of two, bimodal array types identified at different frequencies within Asian and European individuals (Oakey and Tyler-Smith 1990). Here we extend that initial characterization to provide array group frequencies throughout available population-assigned genomic datasets. These data suggest that the rates of homogenization—conversion and unequal crossing over—are sufficiently high to maintain the array sequence states and sequence composition, and that introduction of novel sequences by chance mutation and/or inter-array exchange is exceedingly rare (Warburton and Willard 1995). This decrease of sequence exchange is likely expected for the Y centromere, due to the lack of homologous pairing at this site. In contrast, the DXZ1 array pairs with a homologous X chromosome and is expected to have a slightly elevated probability of sequence exchange; it therefore may be more readily influenced by molecular drive (Dover 1982; Ohta and Dover 1984). It is important to note that, due to limitations of sequence coverage, many low-frequency sequence variants capable of detecting low-proportional mixing between groups may not have been discovered in this analysis.

In total this work presents an initial centromere reference database useful for promoting additional functional and evolutionary studies to study these regions in a comparative and rigorous manner. To make these data fully accessible and integrated into current genomic studies, we have introduced an annotated reference (as shown in Figure 5) that builds upon three central results from our study: 1) the biological arrangement of repeats presented in the singular reference database; 2) mappability indexing to empower additional studies to map and further characterize these regions in an array-specific manner; and 3) a rich sequence definition across X and Y arrays in the human population. Collectively, these efforts lead to a useful genomic reference enabling

studies in centromere function, satellite stability, and sequence evolution in these repetitive sites in the genome.

## **METHODS**

### **Alpha satellite sequence graph: DXZ1 and DYZ3**

Complete HuRef WGS read libraries for both DXZ1 and DYZ3 array were obtained from alignment to full-length HOR sequences that were previously described (Hayden et al. 2013). Alignments of DXZ1 and DYZ3 HOR sequence (consensus HOR derived from sequence read library) to all remaining HuRef genomic reads (with subtraction of DXZ1 and DYZ3 reads) revealed no additional alignments across 100 consecutive bp with a threshold percent identity of 90% or greater. Reads paired to DXZ1 and DYZ3 assigned sequences were assigned to p- or q-arm using uniquely mapping read assignments to the HuRef assembly (Levy et al. 2007a) (using Burrows-Wheeler Aligner, Smith-Waterman alignment BWA-SW) (Li and Durbin 2010). Repeats were identified in array-assigned reads, and high-quality paired reads using RepeatMasker (v4.0; cross\_match, sensitive settings) (Smit et al. 1996–2010). Full-length monomer predictions were initially determined using hmmer software (Eddy 2009) and a model of consensus alpha satellite (Waye and Willard 1987). Limited events of spacing between monomers, where the end of one monomer does not directly precede the base of the following full-length monomer, were evaluated and corrected manually. Monomer libraries of sufficient quality (with *phred* score of 20 across the entirety of the full-length monomer) were organized in the same strand orientation relative to orientation of published consensus. Global alignments (EMBOSS, needle software) (Rice P 2000), provided monomer libraries relative to a reference consensus DXZ1 and DYZ3 HOR repeat unit. Sequence library viewer across DXZ1 used in Figure 2 was created using Circos software (Krzywinski et al. 2009). Monomer groups are included in the study if support for the given sequence variant is observed in donor-matched, targeted re-sequencing (Hayden et al. 2013) or

observed in available flow-sorted chromosome alpha satellite sequences from X (Bentley et al. 2008) and whole chromosome flow-sorted datasets for the X (Trace Archive query terms: center\_project:'CHR\_X\_10470') and Y chromosomes (Trace Archive query term: center\_project='S228'). Identical, full-length monomers were classified under a single label (i.e., m1v1) and read descriptions were reformatted relative to the order of identical monomer groups (i.e., read id m1v1 m2v1 m3v1, etc.). Adjacency and second-order monomer information were obtained from ordered arrangements of monomer labels on each read in the array dataset.

### **LinearSat algorithm**

All documentation and source code for the linearSat is freely available at [github.com/JimKent/linearSat](https://github.com/JimKent/linearSat). In summary, the program requires an input of monomer chains where each vector lists monomer order as observed on sequence reads relative to a model of a given higher-order repeat and monomer variant unique identifier (e.g., Read#1 m1v1 m2v1 m3v1, etc.). Although the program is capable of generating Markov models of arbitrary order, a model order of two was used for linear representation of DXZ1 and DYZ3 arrays in this study as three full-length, high-quality monomers were observed on average per WGS Sanger read in the HuRef alpha satellite database. The program outputs a Markov model of monomer variants and sequence generated following model probabilities. When the higher-order Markov model has no data, a lower order Markov model is used. If there is no data to support a first-order Markov model, the higher-order repeat model is used instead. The probabilities within a model are reduced by the weight of the monomer output. When a variant is present only once, there is only an expectation of one copy of the variant in the output, a final pass is necessary to insert rare variants that are present in input but not output; such insertions are constrained by the higher order model.

### **Intra- and inter-array mappability**

To identify those sequences that are array specific, DXZ1 and DYZ3 read databases were filtered to high-quality (where all bases within the defined  $k$  are greater than a *phred* score of 20)  $k$ -mer library (where  $k=24, 36, 48, 100$  bp) in forward and reverse orientation. Reads that collectively defined each array database were subtracted from the total WGS HuRef dataset, resulting in two sequence libraries: a read library to query for specificity (either DXZ1 or DYZ3) and a remaining library of reads useful in identifying the presence and relative abundance of any given  $k$ -mer from the query as background. Each individual  $k$ -mer was considered specific to a given array if it was found in less than  $< 0.1\%$  of all alpha satellite reads, and  $< 0.01\%$  of all reads in the HuRef genome. To determine copy-number estimates of those  $k$ -mers that are specific to either the X or Y array, we provided raw relative frequency values (or the number of observed counts/total number of sequences within the array) and copy number estimates as derived from single copy read depth distributions across non-satellite DNA on chromosome matched datasets.

### **Alpha satellite DXZ1 and DYZ3 array length estimates**

Low-coverage Illumina sequencing reads were obtained from the 1000 Genomes Project representing 362 male individuals from 14 populations (The 1000 Genomes Project Consortium 2012). High-quality 24-mers (*phred* score  $>20$  for all bases) were compared with array specific 24-mers for DXZ1 and DYZ3. For each individual the total number of high-quality 24-mers across all reads and the total number of high-quality 24-mers exactly matching an alpha array-specific 24-mer were enumerated. For each male individual, the DYZ3 array size was first estimated as follows:

Where the proportion of all high quality 24-mers matching a DYZ3-specific 24-mer, or “a”, is normalized by the proportion of all high quality 24-mers on DYZ3-containing HuRef reads matching one of the specific 24-mers, or “h”, multiplied by the estimated

size of the diploid male genome from hg19):  $a/(h \cdot g)$ . To correct for any coverage bias, a set of 49,994 unique 24-mers was obtained from the chrY reference sequence in hg19 and enumerated across all individuals tested. These control 24-mers were matched to have the same distribution of AT-richness as the DYZ3 24-mers. To estimate the total error in the array size estimates, the size of the control region was estimated by the same calculation for each male individual and compared with the actual value. After correction to the mean coverage bias across samples, 95% of samples had control size estimates within 12% of the actual value (for DYZ3). Similarly, the DXZ1 array estimates used 49,993 unique GC-matched chromosome X 24-mers, corrected down by the mean error rate of 1.7%.

### **Unsupervised clustering of array k-mer profiles**

Each array-specific satellite sequence (1,546 unique 24-mers DXZ1 and 1,837 unique 24-mers DYZ3) was surveyed for normalized abundance (total count of sequence observed normalized to the total number of high quality 24-mers within each whole genomic dataset). The resulting matrices for DXZ1 and DYZ3 ( $n \times m$ ), where 'n' represents each individual and 'm' provides the normalized frequency for each 24-mer queried, provided a vector profile of shared array sequence abundance with the initial HuRef reference. An affinity matrix was constructed by pairwise calculations of Euclidean distance between individual 24-mer profiles, providing a final  $n \times n$  matrix of (366 individuals  $\times$  366 individuals). To provide an initial assessment of cluster number, this matrix was evaluated by hierarchical clustering displaying a heat map, or clustergram object (MATLAB, 2009b). Unsupervised, spectral clustering of two groups (Luxburg 2007) was determined by principal component analysis (MATLAB, 2009b) and consequently, K-means clustering (MATLAB, 2009b, The MathWorks; squared euclidean distance measure) to predict two clusters in both the DXZ1 and DYZ3 datasets. The

optimal number of groups, or  $k=2$ , was determined as the greatest average measure of cluster proximity, or mean silhouette values (MATLAB, silhouette plot).

### **Array group k-mer classification and recursive feature selection**

Similar to a previously-published method (Wang et al. 2006), individuals were classified with labels from spectral clustering (group 1 and group 2 for both DXZ1 and DYZ3 arrays). Initial matrices ( $n \times m$ ) were provided for both DXZ1 and DYZ3 datasets, where 'n' corresponded to each feature (or array specific 24-mer), and 'm' provided the normalized frequency (average of observed forward and reverse 24-mer) for each individual. For each feature, or 24-mer, we iteratively applied a linear Support vector machine (SVM)(R, libsvm)(Chang and Lin 2011) classifier. Leave-one-out cross-validation was used to evaluate the classifier, by which all but one individual in the dataset was used to initially train the SVM classifier to predict the class of the one held out individual. We then compare the predicted class with the actual class membership for the held out individual. Average accuracy scores across SVM cross validation were provided for each 24-mer. Features are then ranked based on the accuracy of the SVM classifier and the top 90th percentile k-mers are selected as the most informative features. Supplemental Figure 6a shows the diagram of this feature selection method.

### **DATA ACCESS**

The sequences and assemblies generated in this study have been submitted to NCBI BioProject (<http://www.ncbi.nlm.nih.gov/bioproject>) under accession number PRJNA193213 and GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) under accession numbers GK000058 and GK000059.

All documentation and source code for the linearSat is freely available at [github.com/JimKent/linearSat](https://github.com/JimKent/linearSat).

## ACKNOWLEDGMENTS

We thank B. Paten, A. Ewing, G. Hickey, M. Haeussler, and D. Haussler for helpful review of this work. This work was supported by grants to J.K. from NHGRI 5U41HG002371 and 3U41HG004568-09S1.

## DISCLOSURE DECLARATION

The authors have nothing to declare.

## FIGURE LEGENDS

Figure 1: An algorithmic overview of satellite characterization and linear representation. (a) Cartoon depiction of centromeric array spanning the complete centromere assigned gap on chromosome X. The multi-megabase-sized DXZ1 array is comprised of tandemly arranged higher-order repeats, shown as dark grey arrows. Examples of array sequence variants are indicated: between pink and blue boxes, single-nucleotide change, illustrated in the second monomer of the HOR; orange box provides a description of monomer rearrangement with a deletion in HOR monomer order; and green box demonstrating a site of transposable element insertion interrupting the repeat. (b) To generate linear representation of these sequences the algorithm uses three key steps: First, generating an array sequence database, where full length monomers that are identified on each WGS read are organized relative to the DXZ1 HOR canonical repeat, with sites of variation as indicated. Second, read databases are reformatted into sequence graphs, wherein nodes are defined by identical monomers and edge weights are defined by the normalized read counts that define each observed adjacency in the WGS read database. Finally, traversal of the graph using a second-order Markov model provides a linear description of the original read database: presenting variant sequences in the proportion and preserving the local-monomer ordering (defined by length of read database ~500bp) as observed in the initial read database.

Figure 2: A complete array sequence database across centromeric regions. Monomer sequence identity across each monomer with average percent identity across a 10-bp window, with red color increasing to 100% as provided in key. Transitions (green) and transversions (blue) relative to the consensus sequence are provided for each 10-bp window (where the sum of each paired transition frequency window and transversion frequency window is 1). Sites of single base-pair insertion (white tracks with dark grey background) and deletion (dark grey on light grey background) are provided as observed in monomer library. Junctions that describe insertions of RepeatMasker-identified transposable elements are shown in purple with numbers indicating read depth. Consensus links (>3000 read support) between individual monomers are shown in black, non-consensus links describing rearrangements in the HOR repeat structure ordering are shown in shades of blue, with color intensity increase with estimated copy number. Image was created using the Circos software (Krzywinski et al. 2009).

Figure 3: Evaluation of linear representation of centromeric arrays. (a) Estimate of accurate WGS sequences in processed linear representation of X (black) and Y (grey) linearized centromeric arrays. Read libraries and linearized centromere arrays X and Y are reformatted into k-mer libraries (where k=50-400 bp with 1-bp slide in both strand orientations) and the proportion of sequences observed in the initial read database observed in the final database. (b) Estimate of sequences observed in linearized centromeric array relative to the initial WGS sequence database, where proportions less than one reflect the gain of novel sequence windows due to the Markov chain model. (c) To determine the improvement of array long-range prediction given an increase of model order, simulated long reads were generated at random from each linearized centromeric array (with length defined by monomer order 3-23, with an average monomer of 171 bp) and the longest arrangement of correctly ordered monomers were normalized to the total length of the array.

Figure 4: Assessment of array variation in the human population. (a) Hierarchical clustering and heatmap representation of affinity matrices for array-specific 24-mer frequencies across the X and Y centromeres provide evidence for two array groups (1 and 2). (b) Classification labels from spectral clustering of array 24-mer profiles for each individual array demonstrate a bimodal distribution with observed array size (DYZ3 group 1 in blue, group 2 in red; DXZ1 group1 in yellow, group 2 in purple). Population-based labels assign array groups to particular geographic locations (c).

Figure 5: Centromeric reference database and sequence annotation. Linear representation of the DYZ3 array is shown to completely replace the centromere gap placeholder in the chromosome Y reference assembly. Evaluation of monomer ordering across the array predicts 40 higher-order repeat units within a generated array of 227 kb. Increased resolution in the linearized centromeric array demonstrates the monomer sequence order along the bottom in blue shading (labels m1v-m34v) that defines the particular HOR arrangement, and the variant sites and base changes observed in the dataset (shown in purple). Each 24-bp sliding window across this region demonstrates the representation of these sequences within the HuRef WGS database, with peaks indicating sites that are overrepresented and likely due to exact homology with satellites outside of the Y array. The top 75<sup>th</sup> percentile mappable sites are provided to extend survey across other individuals. Six individual array profiles are provided as an example of population-based data, where DYZ3 array group 1 (three individuals from the CEU population) is shown in blue, and array group 2 (three individuals from the CHS population) is shown in red.

Table 1: Array length estimate of DXZ1 and DYZ3.

Chr (locus)	Array Group	No. of Arrays	Mean Size (Mb)	Range (Mb)
X (DXZ1)	1	242	2.698	0.705–4.166
	2	130	4.062	3.043–8.313
Y (DYZ3)	1	181	0.399	0.126–0.823
	2	191	1.186	0.774–2.389

## REFERENCES

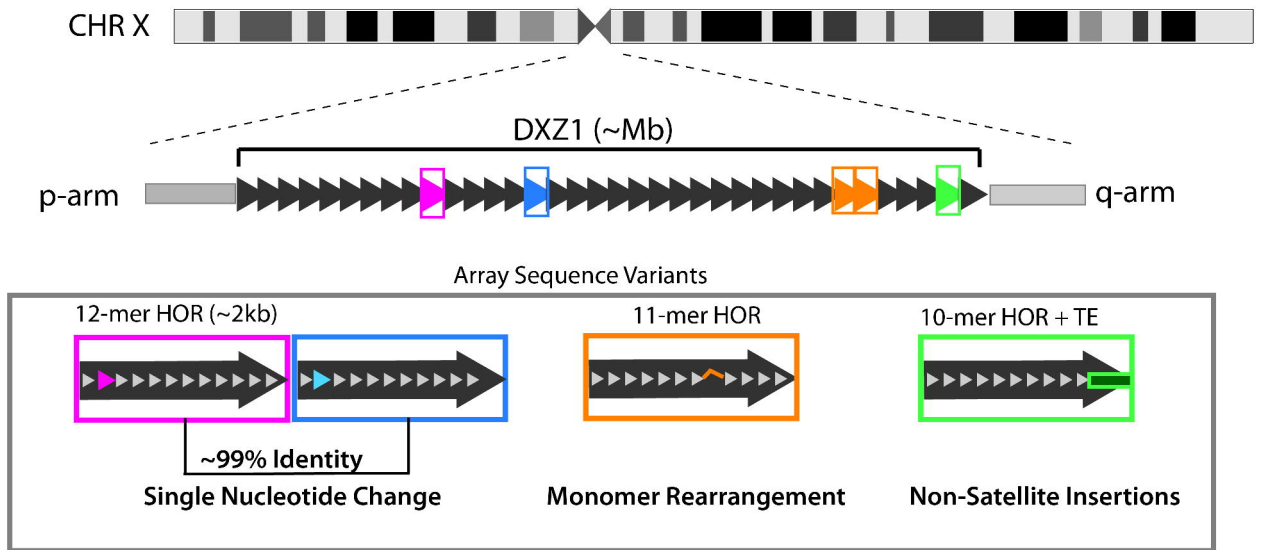
- Alexandrov IA, Medvedev LI, Mashkova TD, Kisselev LL, Romanova LY, Yurov YB. 1993. Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic Acids Res* **21**(9): 2209-2215.
- Alexandrov IA, Mitkevich SP, Yurov YB. 1988. The phylogeny of human chromosome specific alpha satellites. *Chromosoma* **96**(6): 443-453.
- Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.
- Chang C, Lin C. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**(27): 1-27.
- Dernburg AF, Sedat JW, Hawley RS. 1996. Direct evidence of a role for heterochromatin in meiotic chromosome segregation. In *Cell*, Vol 86, pp. 135-146.
- Dover GA. 1982. Molecular drive: a cohesive mode of species evolution. *Nature* **299**(5879): 111-117.
- Durfy SJ, Willard HF. 1989. Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences. *Genomics* **5**(4): 810-821.
- Durfy SJ, Willard HF. 1990. Concerted evolution of primate alpha satellite DNA. Evidence for an ancestral sequence shared by gorilla and human X chromosome alpha satellite. *J Mol Biol* **216**(3): 555-566.
- Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome informatics International Conference on Genome Informatics* **23**(1): 205-211.
- Eichler EE, Clark RA, She X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet* **5**(5): 345-354.
- Hayden KE, Strome ED, Merrett SL, Lee HR, Rudd MK, Willard HF. 2013. Sequences associated with centromere competency in the human genome. *Mol Cell Biol* **33**(4): 763-772.

- Henikoff S. 2002. Near the edge of a chromosome's "black hole". *Trends Genet* **18**(4): 165-167.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**(9): 1639-1645.
- Laursen HB, Jørgensen AL, Jones C, Bak AL. 1992. Higher rate of evolution of X chromosome alpha-repeat DNA in human than in the great apes. *EMBO J* **11**(7): 2367-2372.
- Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al. 2007a. The diploid genome sequence of an individual human. *PLoS Biol* **5**(10): e254.
- Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al. 2007b. The diploid genome sequence of an individual human. *PLoS Biol* **5**(10): e254.
- Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5): 589-595.
- Luce AC, Sharma A, Mollere OSB, Wolfgruber TK, Nagaki K, Jiang J, Presting GG, Dawe RK. 2006. Precise centromere mapping using a combination of repeat junction markers and chromatin immunoprecipitation-polymerase chain reaction. *Genetics* **174**(2): 1057-1061.
- Luxburg Uv. 2007. A tutorial on spectral clustering. *Statistics and computing* **17**(4): 395-416.
- Mahtani MM, Willard HF. 1990. Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate. *Genomics* **7**(4): 607-613.
- Mahtani MM, Willard HF. 1998. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res* **8**(2): 100-110.
- Maloney KA, Sullivan LL, Matheny JE, Strome ED, Merrett SL, Ferris A, Sullivan BA. 2012. Functional epialleles at an endogenous human centromere. *Proc Natl Acad Sci U S A* **109**(34): 13704-13709.
- Manuelidis L. 1978. Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* **66**(1): 23-32.
- Medvedev P, Brudno M. 2009. Maximum likelihood genome assembly. *J Comput Biol* **16**(8): 1101-1116.
- 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): 325-330.
- Ohta T, Dover GA. 1984. The cohesive population genetics of molecular drive. *Genetics* **108**(2): 501-521.
- Peng JC, Karpen GH. 2008. Epigenetic regulation of heterochromatic DNA stability. *Curr Opin Genet Dev* **18**(2): 204-211.

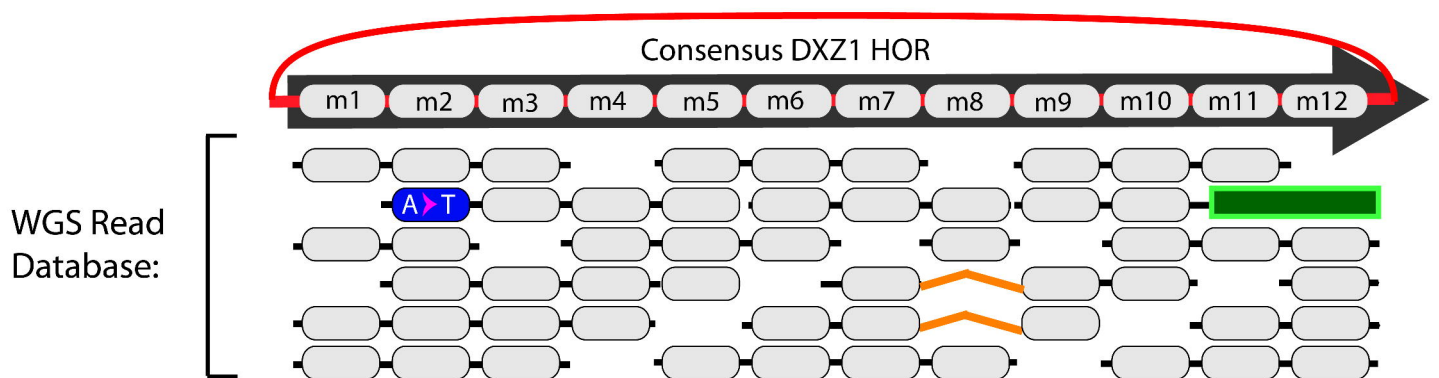
- Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* **13**(3): 235-238.
- Rice P LI, and Bleasby A 2000. EMBOS: The European Molecular Biology Open Software Suite *Trends in Genetics* **16**(6): 276-277.
- 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 Res* **34**(6): 1912-1924.
- Rudd MK, Willard HF. 2004. Analysis of the centromeric regions of the human genome assembly. *Trends Genet* **20**(11): 529-533.
- Santos F, Pandya A, Kayser M. 2000. A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. *Human molecular ...*
- Santos FR, Pena SD, Tyler-Smith C. 1995. PCR haplotypes for the human Y chromosome based on alphoid satellite DNA variants and heteroduplex analysis. *Gene* **165**(2): 191-198.
- Schindelhauer D, Schwarz T. 2002. Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous alpha-satellite DNA array. *Genome Res* **12**(12): 1815-1826.
- Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294**(5540): 109-115.
- Smit AFA, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.
- The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.
- Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S et al. 2011. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science (New York, NY)* **331**(6017): 593-596.
- Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. In *Nat Rev Genet*, Vol 13, pp. 36-46.
- Tyler-Smith C, Brown WR. 1987. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J Mol Biol* **195**(3): 457-470.
- Wang Z, Willard HF, Mukherjee S, Furey TS. 2006. Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comput Biol* **2**(9): e113.
- Warburton P, Willard H. 1995. Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages. *J Mol Evol* **41**(6): 1006-1015.
- Warburton PE, Greig GM, Haaf T, Willard HF. 1991. PCR amplification of chromosome-specific alpha satellite DNA: definition of centromeric STS markers and polymorphic analysis. *Genomics* **11**(2): 324-333.
- Warburton PE, Willard HF. 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 Res* **20**(22): 6033-6042.
- Waye J, Willard H. 1987. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res* **15**(18): 7549-7569.
- Waye JS, Willard HF. 1985. Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. In *Nucleic Acids Res*, Vol 13, pp. 2731-2743.
- Waye JS, Willard HF. 1986. Molecular analysis of a deletion polymorphism in alpha satellite of human chromosome 17: evidence for homologous unequal crossing-over and subsequent fixation. *Nucleic Acids Res* **14**(17): 6915-6927.
- Wevrick R, Willard HF. 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. *Proc Natl Acad Sci USA* **86**(23): 9394-9398.
- Willard HF. 1985. Chromosome-specific organization of human alpha satellite DNA. *Am J Hum Genet* **37**(3): 524-532.
- Willard HF. 1990. Centromeres of mammalian chromosomes. *Trends Genet* **6**(12): 410-416.
- Willard HF, Smith KD, Sutherland J. 1983. Isolation and characterization of a major tandem repeat family from the human X chromosome. *Nucleic Acids Res* **11**(7): 2017-2033.
- Willard HF, Waye JS. 1987. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends in Genetics* **3**(7): 192-198.
- Willard HF, Wevrick R, Warburton PE. 1989. Human centromere structure: organization and potential role of alpha satellite DNA. *Prog Clin Biol Res* **318**: 9-18.
- Wolfe J, Darling SM, Erickson RP, Craig IW, Buckle VJ, Rigby PW, Willard HF, Goodfellow PN. 1985. Isolation and characterization of an alphoid centromeric repeat family from the human Y chromosome. *J Mol Biol* **182**(4): 477-485.
- Yunis J, Yasmineh WG. 1971. Heterochromatin, Satellite DNA, and Cell Function. *Science* **174**(4015): 1200-1209.
- Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, Gage FH, Verma IM. 2011. BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* **477**: 179-184.

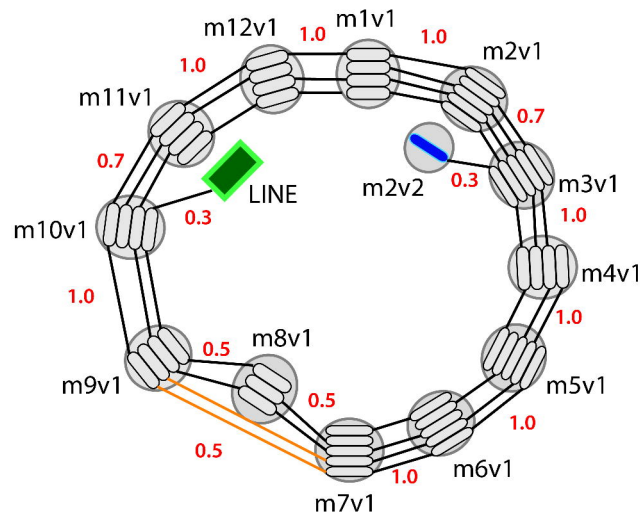
a



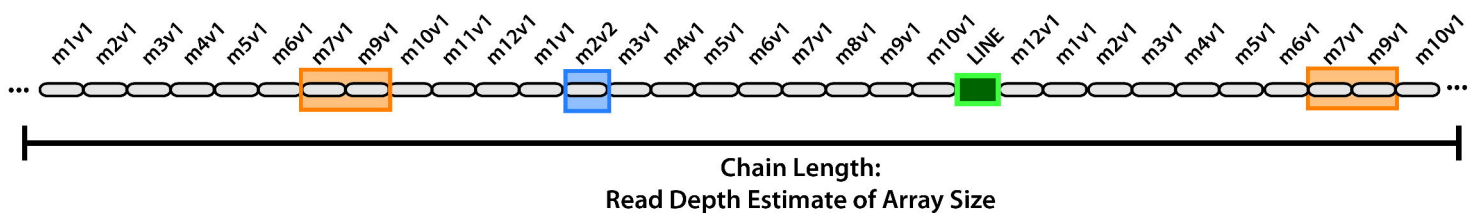
## b (1) HOR Array Sequence Database to Catalogue Sequence Variants



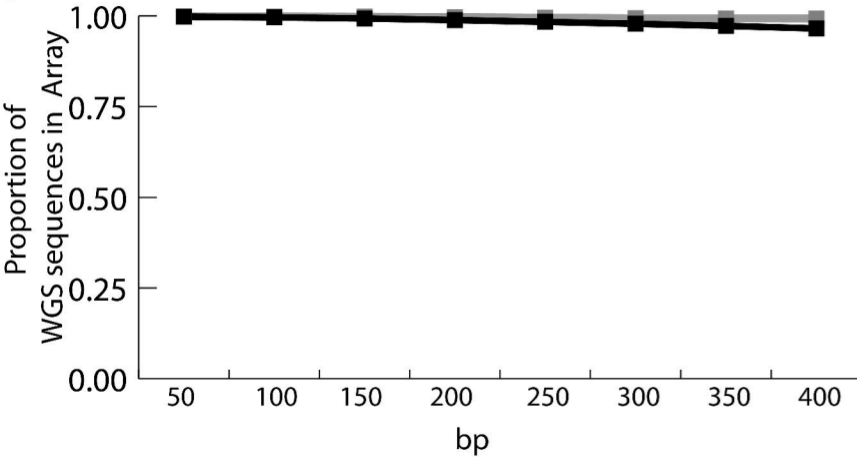
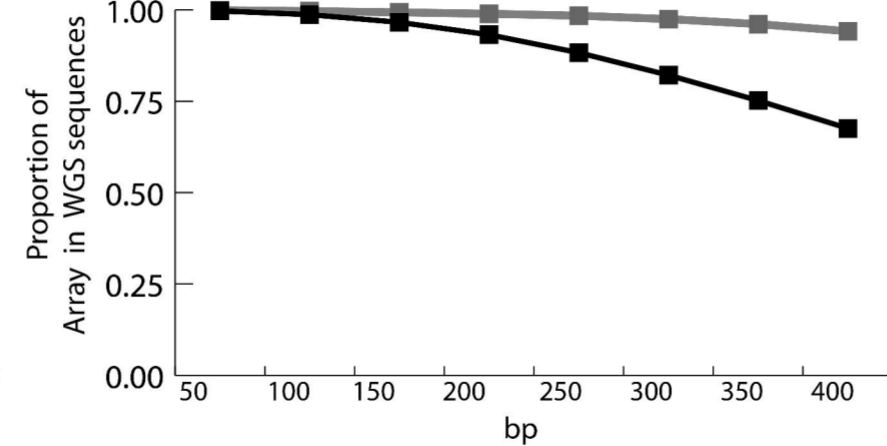
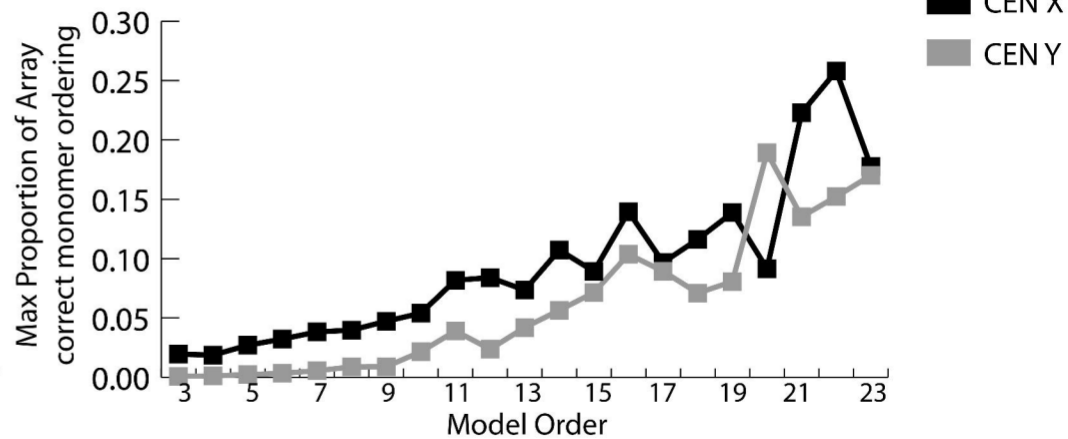
## (2) Reformat into Array Sequence Graph



## (3) Generate Centromere Array Models: Second-Order Markov Chain





**a****b****c**

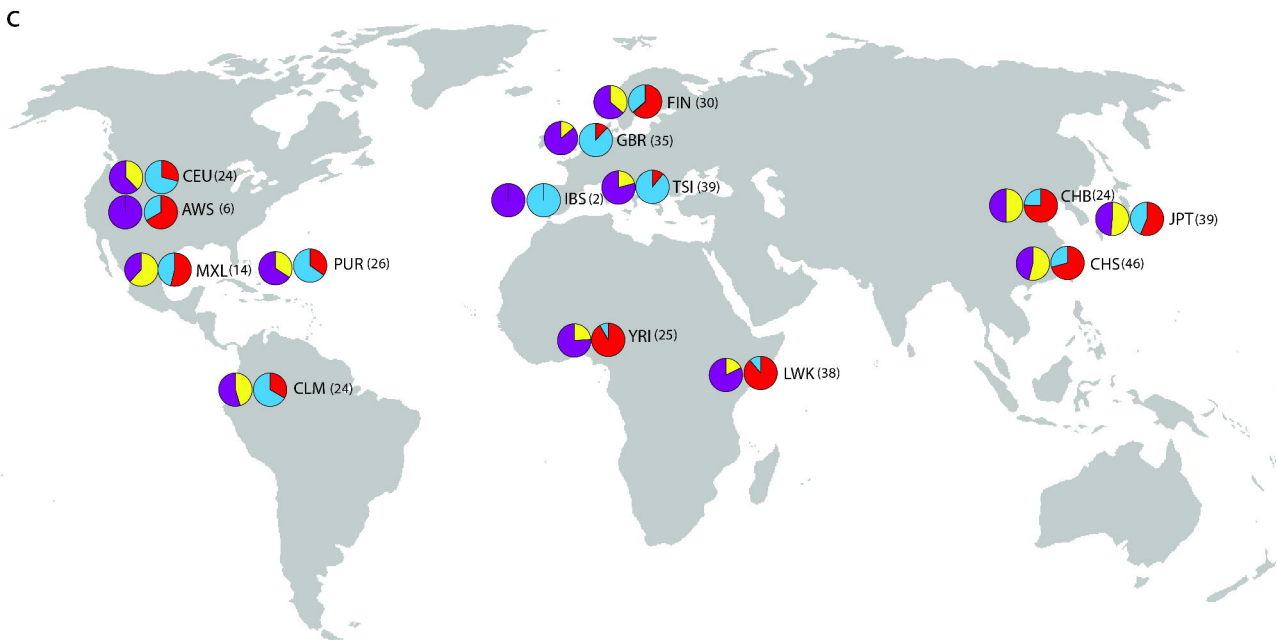
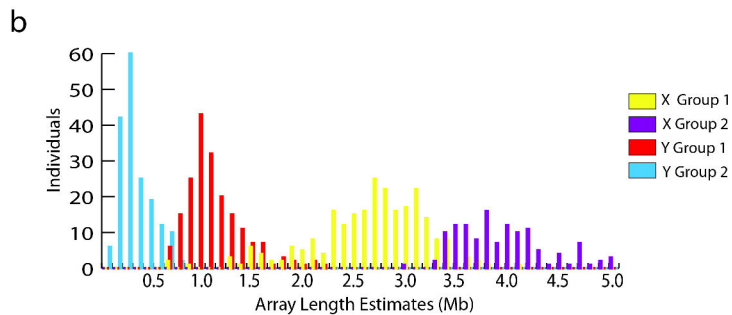
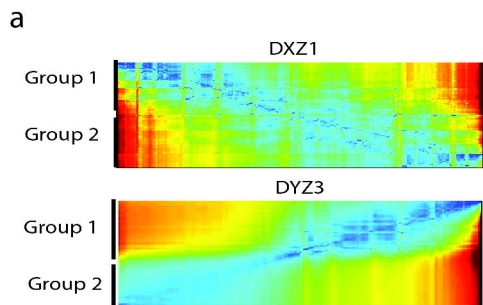


Figure 5

