Complete genomic and epigenetic maps of human centromeres

One-Sentence Summary:
Deep characterization of assembled human centromeres reveals fine-scale sequence organization, variation, and evolution.

Nicolas Altemose[1], Glennis A. Logsdon[2*], Andrey V. Bzikadze[3*], Pragya Sidhwani[4*], Sasha A. Langley[1*], Gina V. Caldas[1*], Savannah J. Hoyt[5,6], Lev Uralsky[7,8], Fedor D. Ryabov[9], Colin J. Shew[10], Michael E.G. Sauria[11], Matthew Borchers[12], Ariel Gershman[13], Alla Mikheenko[14], Valery A. Shepelev[8], Tatiana Dvorkina[14], Olga Kunyavskaya[14], Mitchell R. Vollger[2], Arang Rhie[15], Ann M. McCartney[15], Mobin Asri[16], Ryan Lorig-Roach[16], Kishwar Shafin[16], Sergey Aganezov[17], Daniel Olson[18], Leonardo Gomes de Lima[12], Tamara Potapova[12], Gabrielle A. Hartley[5,6], Marina Haukness[16], Peter Kerpedjiev[19], Fedor Gusev[8], Kristof Tigyi[16,20], Shelise Brooks[21], Alice Young[21], Sergey Nurk[15], Sergey Koren[15] , Sofie R. Salama[16,20], Benedict Paten[16,34], Evgeny I. Rogaev[7,8,22,23], Aaron Streets[24,25], Gary H. Karpen[1,26], Abby F. Dernburg[1,27,20], Beth A. Sullivan[28], Aaron F. Straight[4], Travis J. Wheeler[18], Jennifer L. Gerton[12], Evan E. Eichler[2,20], Adam M. Phillippy[15], Winston Timp[13,29], Megan Y. Dennis[10], Rachel J. O'Neill[5,6], Justin M. Zook[30], Michael C. Schatz[17], Pavel A. Pevzner[31], Mark Diekhans[16], Charles H. Langley[32], Ivan A. Alexandrov[8,14,33†], Karen H. Miga[16,34†]

[1-34] Affiliations are listed at the end (page 22)

[†] Corresponding authors: Ivan A. Alexandrov, ivanalx@hotmail.com ; Karen H. Miga, khmiga@ucsc.edu
* Co-second authors contributed equally.

Abstract:

Existing human genome assemblies have almost entirely excluded highly repetitive sequences within and near centromeres, limiting our understanding of their organization, evolution, and essential role in chromosome segregation. Now, the first telomere-to-telomere human genome assembly (T2T-CHM13) has enabled us to deeply characterize peri/centromeric repeats at single-base resolution, totaling 6.2% of the genome (189.9 Mb). Mapping the inner kinetochore protein centromere protein A (CENP-A) revealed overlap with the most recently duplicated subregions within centromeric repeat arrays. A comparison of chromosome X centromeres across a diverse panel of individuals illuminated high degrees of structural, epigenetic, and sequence variation. In total, we present an atlas of human peri/centromeres to guide future studies of their complex and critical functions as well as their evolutionary dynamics.

The human genome reference sequence has remained incomplete for two decades. Genome assembly efforts to date have excluded an estimated 5-10% of the human genome, most of which is found in and around each chromosome's highly repetitive centromere, owing to a fundamental inability to assemble across long, repetitive sequences using short DNA sequencing reads (*1*, *2*). Centromeres function to ensure proper distribution of genetic material to daughter cells during cell division, making them critical for genome stability, fertility, and healthy development (*3*). Nearly everything known about the sequence composition of human centromeres and their surrounding regions, called pericentromeres, stems from individual experimental observations (*4*–*7*), low-resolution classical mapping techniques (*8*, *9*), analyses of unassembled sequencing reads (*10*–*13*), or recent studies of centromeric sequences on individual chromosomes (*14*–*16*). As a result, millions of bases in the pericentromeric and centromeric regions (hereafter peri/centromeres) remain largely uncharacterized and omitted from contemporary genetic and epigenetic studies. Recently, long-read sequencing and assembly methods enabled the Telomere-to-Telomere Consortium to produce a complete assembly of an entire human genome (T2T-CHM13) (*2*). This effort relied on careful measures to correctly assemble, polish, and validate entire peri/centromeric repeat arrays for the first time (*2*, *17*). By deeply characterizing these newly assembled sequences, we present a high-resolution, genome-wide atlas of the sequence content and organization of human peri/centromeric regions.

Centromeres provide a robust assembly point for kinetochore proteins, which physically couple each chromosome to the mitotic or meiotic spindle (*3*). Compromised centromere function can lead to nondisjunction, a major cause of somatic and germline disease (*18*, *19*). In many eukaryotes, the centromere is composed of tandemly repeated DNA sequences, called satellite DNA, but these sequences differ widely among species (*20*, *21*). In humans, centromeres are defined by alpha satellite DNA (αSat), an AT-rich repeat family composed of ~171 bp monomers, which can occur as different subtypes repeated in a head-to-tail orientation for millions of bases (*22*, *23*). In the largest αSat arrays, different monomer subtypes belong to higher order repeats (HORs); for example, monomer subtypes a,b,c can repeat as abc-abc-abc (*24*, *25*). Each array can contain thousands of nearly identical HORs, but kinetochore proteins associate with only a subset of HORs, usually within the largest HOR array on each chromosome (*25*). Distinct HOR arrays tend to differ in sequence and structure (*26*, *27*) and, like other satellite repeats, they evolve rapidly, expanding and contracting in repeat copy number over time and generating a high degree of polymorphism across individuals (*28*–*31*). Active centromeric sequences (i.e. those associated with the kinetochore) are embedded within inactive pericentromeric regions, which often include smaller arrays of diverged αSat monomers that lack HORs (*26*, *32*). Pericentromeric regions also contain transposable elements and segmental duplications, which sometimes include expressed genes (*33*, *34*), and frequently contain non-αSat satellite repeat families (*35*), which have poorly understood functions. Given the opportunity to explore these regions in a complete human genome assembly, we investigated the precise localization of inner kinetochore proteins within large αSat arrays and surveyed sequence-based trends in the structure, function, variation, and evolution of peri/centromeric DNA.

A comprehensive map of peri/centromeric satellite DNA

Human peri/centromeric satellite DNAs represent 6.2% of the T2T-CHM13v1.1 genome (~189.9 Mb) (tables S1,2 and fig. S1,2). Nearly all of this sequence remains unassembled or belongs to simulated arrays called reference models (*11*) in the current GRCh38/hg38 reference sequence (hereafter, hg38), including pericentromeric satellite DNA families that extend into each of the five acrocentric short arms. Within these satellite DNAs, an assembly evaluation pipeline found only two ~10 kb loci with strong evidence of structural misassembly issues, both within αSat on chromosome 19, representing only 0.02% of satellite DNAs in T2T-CHM13 (reported in (*17*)). From decades of individual observations, a framework for the overall organization of a typical human peri/centromeric region has been proposed (Fig. 1A). By annotating and examining the repeat content of these regions in the CHM13 assembly (Fig. 1B,C, tables S1,2, database S1), we tested and largely confirmed this broad framework genome-wide at base-pair resolution, with some notable exceptions involving large-scale structural rearrangements and previously uncharacterized satellite variants (fig. S1).

Consistent with prior studies (reviewed by (*36*)), all centromeric regions contain long tracts, or arrays, of tandemly repeated αSat monomers (85.2 Mb total genome-wide, Fig. 1B,C). Often found adjacent to αSat arrays, classical human satellites 2 and 3 (HSat2,3, totaling 28.7 and 47.6 Mb, respectively) constitute the largest contiguous satellite arrays found in the human genome, with the longest arrays on chromosomes 1, 9, and 16 (13.2, 27.6, and 12.7 Mb respectively, Fig. 1B,C). HSat2 and HSat3 are derived from a simple ancestral (CATTC)n repeat that diverged into distinct families and 14 previously characterized subfamilies (*10*, *37*), which sometimes have variable repeat unit lengths on the order of kilobases. Furthermore, two distinct satellite DNA families are known to constitute the most AT-rich regions of the genome (*37*, *38*), which we refer to as HSat1A and HSat1B. HSat1A is a 42 bp repeat spanning 13.4 Mb across chromosomes 3, 4, 8, and the acrocentrics in CHM13 (referred to as “human satellite 1” in the classic literature or “SAR” in RepBase, Fig. 1B,C, table S2) (*37*). HSat1B is a 2.4 kb composite of AT-rich sequences and Alu fragments found predominantly on the Y chromosome (*38*, *39*), with 1.2 Mb across the acrocentrics in CHM13 (referred to as a “2.5 kb male-specific repeat” in the classic literature or “HSATI” in RepBase, Fig. 1B,C, table S2) .

Two additional large families, Beta satellite (βSat, 7.7 Mb total, 52% GC) and Gamma satellite (γSat, 630 kb total, 72% GC), are more GC-rich than αSat (39% GC) and contain dense CpG methylation (fig. S3). βSat can be further subdivided into simple arrays (68-bp repeat unit) and β-composite arrays, in which βSat repeats are interspersed with LSau elements (*40*–*42*). All remaining annotated pericentromeric satellite DNAs (collectively referenced as ‘p-censat’) total 5.55 Mb, with 1.19 Mb representing previously uncharacterized types of satellite DNA found across many chromosomes in CHM13 (table S2 and fig. S2) (*42*). Non-satellite bases observed between adjacent arrays and extending into the p-arms and q-arms are considered ‘centric transition’ regions, which largely represent long tracts of segmental duplications, including expressed genes (Fig. 1C, fig. S1) (*2*, *43*, *44*).

Complete assessment of αSat substructure and genomic organization

A genome-wide assessment of CHM13 αSat monomers revealed a broad range of pairwise sequence identity (*44*), with limited deviation in repeat unit length (median 171 bp, with 99.5% within the range 140-187 bp). Using previously described methods (*32*, *45*, *46*), we performed complete, monomer-by-monomer classification of all αSat into 20 distinct suprachromosomal families (SFs) each identified by its own monomeric classes ((*44*), table S2,3, database S2). Within each peri/centromeric region, we identify between 1 and 9 HOR arrays, totaling 80 different HOR arrays and >1000 different monomers in HORs across the genome (70 Mb total; table S4, database S3) (*38*). Although 18 out of 23 chromosomes contain multiple, distinct HOR arrays, only one HOR array per chromosome contains sequences that are consistently associated with inner kinetochore proteins across individuals (*25*), and this array is designated as “active” with respect to centromere function (total of 62 Mb in active HOR arrays genome-wide) (table S4, Fig. 1B,C). The active array on each chromosome ranges in size from 5 Mb on chr18 down to 340 kb on chr21, which is near the low end of the previously reported size range for this centromere among healthy individuals (*47*). All other HOR arrays on the same chromosome are considered inactive (8 Mb inactive HORs, Fig. 1B,C). Adjacent to many highly homogeneous arrays are regions of highly divergent αSat HORs, in which HOR periodicity is somewhat or even completely eroded (*46*), as well as highly divergent αSat monomeric layers (*32*), together totaling 15.2 Mb in CHM13.

Distinct repeat arrays from the same satellite family show varying degrees of similarity with each other. For example, centromeres on chromosomes 13/21, 14/22 and 1/5/19 have near-identical HORs that have confounded studies in the past (26, 36–38). The completeness and quality of the T2T-CHM13 assemblies allowed us to successfully assign each of these active arrays to a specific chromosome, which we validated using flow-sorted chromosome libraries for the cen 1/5/19 arrays, and to discern their evolutionary history ((*44*), fig. S4). To provide a genome-wide view of the overall sequence similarity between different αSat arrays, we obtained the full set of 75-mer sequences within each array and searched for exact matches to the rest of the genome (Fig. 1D), readily identifying the hierarchical evolutionary relationships between subsets of αSat arrays (which can be organized into SFs and sub-SFs; (reviewed in (*36*)). Both HSat and βSat show evidence of hierarchical subfamily organization with chromosome-specific features like those of αSat, however at the resolution of shared 75-mers, their inter-array divergence is lower than for αSats overall (Fig. 1D-F).

Large structural rearrangements in peri/centromeric regions

These maps of peri/centromeric regions provide an opportunity to comprehensively examine the large-scale organization of satellite DNAs and their embedded non-satellite sequences, including transposable elements (TEs) and genes (Fig. 2 A-F). While divergent αSats are known to contain many inversions (*48*) and TE insertions (*49*), such events within active HOR arrays are unexpected, as they were considered to be homogeneous (*30*, *50*). Quantifying strand inversions across entire

satellite arrays (Fig. 2A,B,F, fig. S1, table S5, database S4,5) revealed unexpected anomalies, such as a 1.7 Mb inversion inside the active αSat HOR array on chr1 (Fig. 2A), along with inversions in inactive HORs on chromosomes 3, 16, and 20 (Fig. 2B and fig. S1,5). Surprisingly, the large pericentromeric HSat3 array on chr9 and the βSat arrays on chr1 and the acrocentrics contain over 200 inversion breakpoints (Fig. 2A and fig. S5).

Apart from inversions, two multi-megabase HSat1A arrays appear to have been inserted and expanded within the active HOR arrays of chromosomes 3 and 4 (Fig. 2B; Fig. 1C, table S6). We also found evidence for an ancient duplication event that predated African ape divergence and involved a large segment of the ancient chr6 centromere plus about 1 Mb of adjacent p-arm sequence ((*44*), database S6). This duplication created a new centromere locus that hosts the current active cen6 HOR array.

Using previously developed standards (*10*), we also assigned HSat2 and HSat3 arrays into their respective sequence subfamilies (Fig. 2A-C) and found the chromosomal localizations of these subfamilies to be largely concordant with previous predictions (*10*). However, we identified a 280 kb HSat3 array on chr17 that belongs to subfamily B1 (Fig. 2C), which had not previously been localized to a particular chromosome (*10*). While we found novel chromosomal localizations of several additional HSat3 subfamilies (Fig. 2E), we also noticed a conspicuous lack of HSat3 subfamily B2 (HSat3B2) on chr1, contrary to expectations based on different cell lines (*10*), implying a large deletion of this subfamily on chr1 in CHM13.

To better understand if these structural rearrangements are common outside of the CHM13 genome, we searched for evidence of these insertion/inversion/deletion breakpoints across 16 haplotype-resolved draft diploid assemblies from genetically diverse individuals from the Human Pangenome Reference Consortium (HPRC) (*51*) ((*44*)). We found that the chr1 active HOR inversion is polymorphic across individuals, evident in about half of ascertainable haplotypes (11/24; fig. S6). However, the HSat1A insertions on chr3 and chr4 were evident in all ascertainable haplotypes (32/32 and 33/33, respectively; fig. S7). Furthermore, CHM13's missing chr1 HSat3B2 array is, in fact, contained within a 400 kb polymorphic deletion, which we detected in 29% (8/28) of haplotypes examined (Fig. 2A, fig. S7). These analyses demonstrate the utility of a complete reference sequence for investigating the organization and variation of these formerly missing regions of the human genome assembly.

Transposable elements and genes in peri/centromeric regions

Like inversions and insertions, transposable elements (TEs) are virtually absent from homogeneous HOR arrays but are enriched in divergent αSat in CHM13 (Fig. 2F, database S7) (*49*, *52*). The CHM13 assembly also revealed regions where combinations of TE sequences have been tandemly duplicated, forming previously uncharacterized satellite sequences which we refer to as "composite satellites" (described in (*42*)). Consistent with individual published observations (*38*, *40*, *53*), we also found that other satellites, such as HSat1, HSat3, and βSat, often include fragments of ancient TEs

as part of their repeating units—a phenomenon we rarely observe in αSat HOR arrays (Fig. 2A,F, fig. S8).

We also compared our pericentromeric maps to gene annotations reported for T2T-CHM13, revealing 676 gene/pseudogene annotations embedded between large satellite arrays, including 23 protein coding genes and 141 lncRNAs (when excluding the acrocentric short arms; table S7 and database S8; (*2*)). One region on chr17, located between the large HSat3 and αSat arrays (Fig. 2D), contains two protein-coding genes: *KCNJ17*, which encodes a disease-associated potassium channel in muscle cells (*54*); and *UBBP4*, which encodes a functional ubiquitin variant that may play a role in regulating nuclear lamins (*55*). Notably, *KCNJ17* is missing from GRCh38, which likely has caused inaccurate and missed variant calls in homologous genes *KCNJ12* and *KCNJ18* (*56*). This region also contains a long non-coding RNA annotation (*LINC02002*), which starts inside an SST1 element and continues into an adjacent 33 kb array of divergent αSat (Fig. 2D). We also identified a processed paralog of an apoptosis-related protein-coding gene, *BCLAF1* (BCL2 Associated Transcription Factor 1), as part of a segmental duplication embedded within an inactive HOR array on chr16 (fig. S9).

The fine repeat structure of satellite DNA arrays

To further chart the structure of peri/centromeric regions at high resolution, we compared individual repeat units within and between different satellite arrays. We decomposed each αSat HOR array first into individual monomers and then into entire HORs, revealing the positions of full-size canonical HORs and structural variant HORs resulting from insertions or deletions ((*44*), database S9). While some chromosomes, such as chr7, are composed almost entirely of canonical HOR units, other chromosomes, such as chr10, contain many structural variant HOR types, with high variation in the relative frequency of these variants across individuals (Fig. 3A and fig. S10).

Unlike αSat, some families like HSat2 and HSat3 have inconsistent or unknown repeat unit lengths and often contain an irregular hierarchy of smaller repeating units. We propose calling these repeat units nested tandem repeats (NTRs), a more general term than HORs, which are composed of discrete numbers of monomers of similar lengths. To expand our ability to annotate repeat structure within assembled satellite DNA arrays, we developed NTRprism, an algorithm to discover and visualize satellite repeat periodicity (Fig. 3B and fig. S11). Using this tool, we discovered HORs in HSat1 and βSat arrays, as well as NTRs in multiple HSat2,3 arrays (Fig. 3B and fig. S11). We also applied this tool in smaller windows across individual arrays, showing that repeat periodicity can vary across an array, consistent with NTRs evolving and expanding hyper-locally in some cases (fig. S11).

Genome-wide evidence of layered expansions in centromeric arrays

Previous studies have hypothesized a “layered expansion model” for centromeric αSat arrays based on limited available sequence information (reviewed in (*36*)). This model postulates that distinct new repeats periodically emerge and expand within an active array, displacing the older repeats sideways

and becoming the new site of kinetochore assembly. As this process repeats itself over time, the displaced sequences form distinct layers flanking the active centromere with mirror symmetry (Fig. 3C). These defunct flanks rapidly shrink and accumulate mutations, inversions, TE insertions, and other satellite expansions (*16*, *32*, *46*). Efforts to document this layered expansion pattern have focused on divergent αSats that surround HOR arrays (*32*). Here, we studied active αSat arrays in their entirety along with adjacent diverged αSat, providing detailed, genome-wide evidence in support of this model (*44*).

First, in agreement with prior studies, we observed a symmetrical flanking arrangement of two types of divergent αSat: divergent HORs (dHORs) (database S10), and monomeric αSats (table S8), which represent ancient, decayed centromeres of primate ancestors (*32*). We classified divergent αSat into distinct SFs and dHOR families, and demonstrated how these sequences accumulate mutations, inversions, TE insertions, and non-αSat satellite expansions over time (Fig. 3C, table S5,6,9, databases S4,5,7). In agreement with previous studies (*16*, *32*, *46*), we show gradients of size and intra-array divergence (17 to 26%) in monomeric αSat layers, a steep (~10%) divergence increase between HORs and dHORs, and a gradient of L1 element quantity and age which parallels the age of monomeric layers (Fig. 2F, Fig. 3C, table S9, database S7).

We next asked if the layered expansion pattern overlaps the active arrays themselves. As shown in Fig. 3D, the sequences seeding the expanding satellite array can be either introduced from within (intra-array seeding) (32) or from an external HOR (or non-HOR) array (inter-array seeding) (*57*, *58*). In total, we document four cases of symmetry consistent with inter-array seeding (on chrs 1, 2, 16, 18; Fig. 3D) of which only one was known before (*59*). In some cases of the inter-array model, the active HOR array originates from a different SF than the flanking inactive array (chrs 1 and 16) (Fig. 3D). This provides evidence of how entire arrays have been recently displaced in favor of an externally introduced sequence.

Moreover, classification of HORs by their shared sequence variants revealed symmetry within active HOR arrays. Such variants have been known for decades (*30*, *60*, *61*) and were noted in the completely assembled centromeres from chromosomes X (*62*, *63*) and 8 (*16*). In these studies the central part of the active array was found to contain HOR variants slightly different from those on the flanks. To test if this array structure is typical, we aligned individual HOR units within the same array and clustered them into “HOR-haplotypes” or “HOR-haps” (*44*). Initial broad classifications of entire arrays into 2-4 distinct HOR-haps revealed that active HOR arrays are also composed of distinct layers, which typically expand from the middle (dark red versus grey, Fig. 3D). Further classification into a larger number of HOR-haps (5-10) found additional evidence for symmetric patterns (see the next section, Fig. 3E and (*44*)).

To examine whether the middle HOR-haps are the most recently evolved, we built phylogenetic trees of consensus HOR-haps (Fig. 3E) and rooted them using reconstructed “ancestral” sequences built from consensus monomers for each SF. We also performed complete phylogenetic analysis of all

HORs. The identification of evolutionarily younger and older HOR-haps was supported by both methods, as shown for chr3 (Fig. 3E). In addition, the intra-array divergence in central HOR-haps is often slightly lower than in the flanking arrays, indicating that the central HOR-haps have expanded more recently (Fig. 3E, materials and methods section 6). Together, these findings present genome-wide evidence for a layered expansion pattern within active arrays.

Precise mapping of sites of kinetochore assembly

Human centromeres are defined epigenetically as the specific subregion bound by inner kinetochore proteins within each active αSat HOR array (*21*, *64*). Previous studies were able to broadly determine which HOR arrays are associated with kinetochore proteins (*65*), and therefore active, but T2T assemblies have now enabled us to map the precise positions of centromeres within active arrays (*16*). Centromeres contain a combination of epigenetic marks distinguishing them from the surrounding pericentromeric heterochromatin, including the presence of the histone variant CENP-A (*66*, *67*), "centrochromatin"-associated modifications to canonical histones (*68*), consistently high CpG methylation compared to neighboring inactive arrays (*69*), and regions of reduced CpG methylation called Centromere Dip Regions (CDRs) (*15*, *16*, *69*). To study HOR organization at sites of kinetochore assembly, we identified discrete regions of CENP-A enrichment within each active array using published native CHM13 ChIP-seq (NChIP) data (*16*) along with CUT&RUN (*70*) data generated in this study ((*44*), table S10). To precisely map these short-read data, we developed a repeat-sensitive alignment method using a collection of informative markers across each HOR array (Fig. 4A). Consistent with previous studies, CENP-A binding is almost exclusively localized within αSat HOR arrays, with one active array per chromosome (*25*) (table S4,11).

In agreement with previous studies on individual chromosomes (*16*, *69*), we found the strongest CENP-A enrichment near and within CDRs on all chromosomes. Notably, some chromosomes show evidence for multiple CENP-A peaks within each CDR. These multiple peaks could represent interspersed domains, variation in the organization of CENP-A nucleosomes across the two homologous chromosomes, or polymorphic organization across the population of cells (*69*). We found that the complete span of each centromere position, defined as a window with strong CENP-A enrichment, extends outside of the CDR and totals 150-500 kb on each chromosome (Fig. 4B,C and table S11). The lengths of these CENP-A windows bear no apparent relationship to the total length of each HOR array (table S11), contrary to predictions from previous work (*65*). However, we note that we cannot exclude the possibility that lower levels of CENP-A extend beyond these windows of strong enrichment, or that the sizes of these windows vary among cells or cell types. We detected smaller regions of CENP-A enrichment outside of the primary CDR, with some overlapping a minor, secondary CDR (chromosome 4, 16, 22) or no CDR at all (chromosome 18) (Fig. 4C,D, fig. S12, table S11). Furthermore, similar dips in CpG methylation, although infrequent, do occur outside of αSat and CENP-A associated regions, as observed in a 5S RNA composite satellite array (*42*), and in two regions (less than 10 kb) where CpG methylation drops below 25% within the active HOR arrays on chromosomes 5 and 8 (fig. S12).

We also found that CENP-A is typically enriched in young, recently expanded HOR-haplotypes within active HOR arrays (except on chr6, described below, and chr21, which has a very small array enriched across its full length; table S11, Fig. 4B-F). For example, in the active array on chromosome 12, we identified CENP-A enrichment on one of two large macro-repeat structures, both presenting similar HOR-hap sequences (Fig. 4B and fig. S13). Further investigation into the region of CENP-A enrichment on chr12 revealed a zone of very recent HOR expansions (i.e. eight sites of nearly identical duplications within a ~365 kb region; (*44*); fig. S14) that coincides with the CDR and distinguishes one macro-repeat region from the other (Fig. 4B,F).

On most other chromosomes, we observed similar zones of recently expanded, younger HOR-haps that overlap CENP-A (8 more examples shown in Fig. 4C, table S11), although we identified a few notable exceptions to this general trend. On chr4, which has two CENP-A regions occurring on either side of a 1.7 Mb HSat1A array, we found that the larger CENP-A region spans a slightly younger HOR-hap and the smaller CENP-A region spans an older HOR-hap (Fig. 4D,F). On chr7, CENPA overlaps young HOR-haps, but these do not appear to have been recently duplicated (fig. S15). Inversely, on chr2, CENPA overlaps older HOR-haps, but these do appear to have been recently duplicated (fig. S15). On chr6, we observe CENP-A enrichment within an older HOR-hap layer, over a megabase away from the site of recent duplications and expansions (Fig. 4E,F). In summary, we observe that human centromeres and CDRs are typically, though not universally, positioned over young and/or recently expanded layers within each HOR array in CHM13.

Genetic variation across human X centromeres

Satellite DNA arrays are highly variable in size across individuals. In fact, the extremes of satellite size variation are often plainly visible under the microscope in chromosomal karyotypes (*71*, *72*), yet the clinical significance of these variants remains unknown and largely unexplored. Studies have provided low-resolution sequencing-based evidence for variability in both satellite array lengths and in the frequency of certain sequence and structural variants within human populations (*10*–*12*). This suggests accelerated sequence evolution in these regions compared to the rest of the genome. However, satellite array variation and evolution have remained poorly understood at base-level resolution due to the lack of complete centromere assemblies.

Therefore, we characterized and compared centromere array assemblies from chrX across seven XY individuals with diverse genetic ancestry (lymphoblastoid cell lines from (*73*), Fig. 5A, fig. S16, table S12). We assigned repeats in the cenX active array to seven HOR-haps, revealing both localized and broad variation within each array (*44*). For example, we identified duplications spanning hundreds of kilobases in two assemblies relative to CHM13 (HG01109 and HG03492, Fig. 5A, fig. S17). Four of the seven arrays contain zones of recent expansion in the younger HOR-hap (CHM13, HG01109, HG02145, HG03098). The remaining three assemblies (HG03492, HG01243, HG02055) show a trend of recent duplication within a shared region closer to the p-arm (spanning different subsets of more divergent and less derived older HOR-haps). Notably, we found evidence for a recently

expanded HOR-hap type (HOR-hap 6) present in three individuals with recent African ancestry but absent in the other individuals, including CHM13 (Fig. 5A, dark red).

Next, we studied how this variation within αSats relates to variation across single-nucleotide markers that tend to be co-inherited with the centromere. Because meiotic crossover rates are low in peri/centromeric regions (*74*), centromeres are embedded in long haplotypes, called cenhaps (Fig 5B; (*75*)). Cenhaps are identified by clustering pericentromeric single-nucleotide variants into phylogenetic trees, and then splitting them into large clades of shared descent. Here, we divided a group of 1599 XY individuals genotyped using published short-read sequencing data (*76*) into 12 cenhaps (with 98 individuals remaining unclassified; Fig. 5C, fig. S18 and database S11). We also defined array-specific and HOR-hap-specific k-mer markers in order to utilize short-read sequencing data to estimate the absolute size of each individual's chrX centromere array (fig. S19, database S11, (*11*, *75*)), along with the relative proportion of that individual's array belonging to each HOR-hap (*44*). The results revealed that different cenhaps have different αSat array size distributions as well as different average HOR-hap compositions (Fig. 5C and fig. S18). For example, HOR arrays belonging to cenhaps 1 and 2 tend to be larger overall than those belonging to cenhaps 3-12. We found a recent duplication in the chrX HOR array, representing hundreds of kilobases, that is common in cenhap 1 and can explain the relatively larger average HOR array sizes in this cenhap (Fig. 5C).

As shown in Fig. 5D, two of the 12 cenhaps, 1 and 2, are very common in non-African populations (overall, 49% and 47%, respectively) and rare in African populations (1.7% and 3.5%, respectively). The remaining 10 cenhaps are almost exclusive to African populations as well as those with recent African admixture (ASW, PUR, CLM, ACB). The relatively low cenhap diversity in non-African populations is consistent with their lower overall genetic diversity, attributable to demographic bottlenecks during early human migrations out of Africa (*73*). This analysis also revealed that HOR-hap 6 (shown for three individuals in Fig. 5A) appears to be almost exclusively found in cenhaps 10-12, which form an anciently diverged clade within African populations (Fig. 5C). These findings demonstrate that centromere-linked SNVs can be used to tag and track the evolution of αSat, and they underline the need for greater representation of African genomes in pan-genome assembly efforts.

To explore the variation within one of the large cenhap groups (cenhap 2), we compared fine-scale cenhap phylogenies and HOR-hap assignments across 558 individual X chromosomes, revealing a degree of further substructure and variation in the αSat array on a more recent evolutionary timescale (Fig. 5E). To dissect this further, we compared two finished centromere assemblies from CHM13 and HG002, a cell line whose chrX array had been constructed using T2T assembly methods, and whose array structure had been experimentally validated by both pulse-field gel electrophoresis Southern blots and by digital droplet PCR (*2*). CHM13 and HG002 have similar array sizes and both belong to cenhap 2 (Fig. 5E). We found both genomes to be highly concordant across the array, apart from three regions, where we observe recent amplifications and/or deletions of repeats (Fig. 5F, fig. S20).

Epigenetic variation across human X centromeres

To examine how centromere positioning varies among individuals at high resolution, we compared patterns of CENP-A CUT&RUN enrichment on the fully assembled chrX centromeres from HG002 and CHM13 (*69*). Notably, the region with the most pronounced structural differences between CHM13 and HG002 coincides with the strongest CENP-A enrichment in both arrays. Therefore, even though inner kinetochore proteins are present in both arrays over CDRs and young HOR-haps, the HOR sequences enriched with CENP-A represent local duplication events that are not shared and distinguish the two arrays (marked in yellow, Fig. 5G and fig. S20).

Finally, we asked if CENP-A enrichment patterns were consistently found in the younger HOR-haps, as observed in CHM13 and HG002, across 7 additional cell lines with publicly available CENP-A ChIP-seq and CUT&RUN datasets (fig. S21). Using the T2T-CHM13 X array as a reference, we determined CENP-A enrichment for each cenX HOR-hap relative to the matched input DNA. Unlike CHM13, in three XY individuals we observed CENP-A enrichment within the older HOR-hap subregion, proximal to the p-arm, indicating the presence of an epiallele (as shown for HuRef (*77*) in Fig. 5H, and for HT1080b (*78*) and MS4221 (*79*) in fig. S21). Further, we examined two independent CUT&RUN experiments from the RPE-1 cell line (XX) (*80*) and found enrichment on both older and younger HOR-haps, which could be explained if the two chrX homologs carry different functional epialleles. Three additional XX cell lines (IMS13q, PDNC4, K562 (*81*)) were consistent with CHM13, providing evidence that the same CENP-A-enriched HOR-hap is shared across both chrX homologs in each line. In total, these findings uncover frequent variation in the position of the chrX centromere, with some XX individuals potentially harboring heterozygous epialleles. Further, these observations highlight the need to study both epigenetic and genetic variation in centromeric regions, across both related and unrelated individuals and across populations of cells over time, to better define the trends and exceptions regarding centromeric epiallele positioning and inheritance.

Discussion

Here we provide detailed maps of previously unassembled centromeric and pericentromeric regions to facilitate further analysis of these relatively unexplored loci in the human genome. Using this resource, we characterized satellite array variation, uncovering a 400 kb polymorphic deletion of an entire HSat3 array and a 1.7 Mb polymorphic inversion in an active HOR array, both on chr1, along with an expansion of a particular αSat HOR-haplotype on chrX in individuals with recent African ancestry. The high degree of polymorphism in these regions, even within a single cenhap (Fig. 5E), underlines the need to produce T2T assemblies from many diverse individuals, to fully capture the extent of human variation in these regions and to shed light on their recent evolution and the functional consequences of this evolution. Achieving this goal requires an ability to produce accurate, complete, phased assemblies from diploid individuals. Centromeric regions have presented a challenge for phased assembly due to their repetitive nature, but their high degree of variation may assist these efforts. Now, equipped with the T2T-CHM13 assembly and the ability to compare challenging repetitive regions in the genome, we are optimistic that future high-quality, phased, diploid, T2T assemblies are within reach. Apart from genetic variation in these regions, we identified

epigenetic variation in the location of CENP-A within an array, as has been observed on other chromosomes (*82*–*85*). We acknowledge the need to study centromeric protein localization at fine scales across many individuals to better understand centromeric establishment and propagation, and how this relates to the underlying genetic variation found within each array.

In light of our observation that CENP-A tends to localize to the most recently expanded HORs genome-wide, many questions remain about the evolutionary and molecular mechanisms responsible for the relationship between the kinetochore and the layered expansion patterns of αSat. One possibility, which we call the "independent expansion hypothesis," is that αSat expansions occur independently of the kinetochore yet somehow attract it by virtue of their homogeneity or some other property. Kinetochore-independent expansion is feasible in light of the fact that we observe large duplications and localized repeat expansions in non-centromeric satellites like HSat3 arrays, which are not associated with kinetochores (fig. S11). Another possibility is what we refer to as the "kinetochore selection hypothesis," in which kinetochore proteins, or their associated loading and/or replication factors, play a causal role in the expansion of particular HOR variants (*36*, *86*). This aligns with the proposed recombination-based homogenization process in Arabidopsis (*87*), that is hypothesized to maintain a satellite consensus optimal for kinetochore recruitment. Further, experiments in model organisms have demonstrated that extreme array sequence variants increase meiotic and mitotic nondisjunction rates and can promote both mutational drive and/or (female) meiotic drive (*20*, *88*–*90*). Similar drive mechanisms, along with selection for variants that promote high-fidelity chromosome transmission, may also play a role in shaping centromeric sequence diversity in the human population.

Exploring these models will require careful experimental systems and methods for precisely measuring interactions between kinetochore proteins and repetitive DNA, as well as measuring how these interactions affect the fidelity of chromosome transmission. While the short-read mapping methods that we developed enable the use of existing protocols like NChIP (*91*) and CUT&RUN (*70*) to provide sensitive protein-DNA interaction information at broad scales within satellite arrays, new long-read methods for mapping protein-DNA interactions will be essential for providing high-resolution binding footprint information, including in regions that lack single-copy or region-specific markers (*92*). Furthermore, forthcoming pan-genome and pan-epigenome references promise to make human peri/centromeric regions even more accessible for careful study using modern genomic tools.

Figure Captions

Fig. 1. Overview of all peri/centromeric regions in CHM13. (**A**) Schematic of a generalized human peri/centromeric region identifying major sequence components (not to scale) and their canonical repeat structure, repeat unit length (note: HSat2,3 lengths vary by genomic region), and GC%. (**B**) Barplots of the total lengths of each major satellite family genome-wide. (**C**) Micrographs of representative DAPI-stained chromosomes from CHM13 metaphase spreads, next to a color-coded map of peri/centromeric satellite DNA arrays (available as a browser track, database S1). Large satellite arrays are labeled. (**D-F**) Circos plots showing genome-wide relationships in sequence content for 3 different satellite

families. Connecting line widths indicate the proportion of 75-mers that are shared between arrays. αSat lines are colored by their SF assignment. Radial barplots indicate specificity of 75-mers proportionally, with white indicating 75-mers unique to the array, light gray indicating 75-mers shared with other centromeric regions, and black indicating 75-mers shared with regions outside of centromeres.

Fig. 2. Discoveries in three peri/centromeric regions. (**A**) The peri/centromeric region of chr1 (cylindrical schematic at top), zooming into the transition region between the large αSat and HSat2 arrays (tracks 1-4). Track 1: satellite families (see color key); placement above or below the line indicates which strand contains the canonical polarity for each satellite family. Track 2: positions of TEs overlapping αSat or HSat1,2,3, colored by TE type. Track 3: transcription start sites of pericentromeric gene annotations, colored by gene type. Track 4: HSat2,3 subfamily assignments (as in (*10*)), and αSat SF assignments, with large arrays labeled. (**B**) As in (A) but for the peri/centromeric region of chr3. (**C**) As in (A) but for the peri/centromeric region of chr17, with the newly localized HSat3B1 array indicated with an asterisk. (**D**) Zoomed-in view of gene annotations between the αSat and HSat3 arrays on chr17. (**E**) Heatmap showing the major and minor localizations of each αSat HOR SF (upper, red) and each HSat2,3 subfamily (lower, blue). Novel localizations relative to (*10*) are indicated with the letter N. The polymorphic chr1 HSat3B2 array shown in (A) is marked with a '-'. Note: HSat3A3 and 3A6 are almost entirely found on chrY, which is not present in CHM13. (**F**) Barplots illustrating the number of inversion breakpoints (strand switches) or the number and type of TEs detected per megabase within different satellite families genome-wide. "div" = divergent αSat (dHORs + monomeric).

Fig. 3. **Genome-wide evidence of layered expansions in centromeric arrays.** (**A**) Left: HOR structural variant positions across the active αSat arrays on chr7 and chr10 (gray=canonical HORs; other colors=variants). Right: proportions of the same structural variant types among HORs identified on HiFi sequencing reads from 16 HPRC cell lines. (**B**) Illustration of NTRprism used to identify tandem repeat periodicities, with results for 3 arrays below. (**C**) Comparison of the age and divergence of LINE TEs embedded in different αSat SF layers. (**D**) Centromere displacement modes supported by T2T-CHM13. (i) Four centromeres where an active HOR array of distinct origin appears to have expanded within a now-inactive HOR array. (ii-iii) Monomeric SFs (rainbow colors) surrounding active HOR arrays on 8 chrs, with major HOR-haps shown within each array (k=2 clusters; red: younger, emphasized below with red rectangles; gray: older, emphasized below with asterisks). (**E**) Zoomed-in view of major αSat HOR arrays on chr3, divided into finer HOR-haps (k=7 clusters) showing further symmetry. Bottom left: radial tree showing the phylogenetic relationships between all HORs, colored by fine (k=7) HOR-hap assignments as in the track above. Red/gray ellipses indicate major (k=2) HOR-hap assignments of each subclade. Bottom right: phylogenetic tree built from HOR-hap consensus sequences, rooted with a reconstructed ancestral cen3 active HOR sequence (ANC) (*44*). Branch lengths represent base substitutions per position.

Fig. 4. Kinetochore proteins tend to bind recently expanded centromeric subregions. (**A**) Two approaches to define NChIP/CUT&RUN enrichment in centromeric regions. (**B**) Active αSat HOR array on chr12 (coordinates at top). Track 1: CENP-A NChIP marker-assisted mapping coverage. Track 2: reference-free region-specific marker enrichment (bins with no region-specific markers are shown in black). Track 3: percent of CpG sites methylated. Tracks 4/5: HOR-haps (k=5 or 2 clusters, respectively). Track 6: number of identical copies of each HOR unit occurring within the adjacent 10 HOR units. Bottom: self-alignment dotplot (exact-match word size 2000), with arrows pointing to a zone of recent duplication. Small inset on left: smaller dotplot of the entire array (word size 500, allowing for detection of older duplications), with positions of two large macro-repeats indicated with blue lines. (**C**) As in (B) but for 8 different centromeres, with the full span of CENP-A enrichment framed by black windows. (**D**) As in (B) for chr4, with an inset highlighting a secondary CENP-A enrichment site and minor CDR on the other side of the interrupting HSat1 array. (**E**) As in (B) for chr6, with CENP-A enrichment over an older HOR-hap region. (**F**) HOR-hap consensus trees as in Fig. 3E, with the location of the CENP-A enriched region(s) indicated with arrows.

Fig. 5. Evidence for substantial genetic and epigenetic variation in a human centromere. (**A**) Comparing the active αSat HOR array on chrX (DXZ1) between CHM13 (top) and 6 HPRC cell line HiFi read assemblies. Tracks indicate

HOR-haps (top: k=7; bottom: k=2) and recent HOR duplication events (bottom, as in Fig. 4B). (**B**) Low recombination rate around centromeres produces large, centromere-spanning haplotypes with many linked single-nucleotide variants (SNVs) flanking the centromere. (**C**) Left: tree illustrating the relationships of 12 cenhaps defined using short-read data from 1599 XY genomes from (*73*, *76*) plus HG002, CHM13, and HuRef. Triangle vertical length is proportional to the number of individuals in that cenhap (note: 98 individuals, labelled NA and colored dark gray, belong to small clades not among the 12 major cenhaps). Middle: barplots illustrating the average HOR-hap compositions for all individuals within each cenhap, colored as in (A). Right: ridgeline plots indicating the distribution of estimated total array sizes for all individuals within each cenhap, with individual values represented as jittered points. (**D**) Populations represented among the 1599 XY genomes, with pie charts indicating the proportion of cenhap assignments within each population, with the same colors as in (C). Population descriptions are in (*44*). (**E**) A detailed tree showing the relationships of all cenhap 2 individuals inferred from their SNV data, along with each individual's superpopulation (AFR: African, AMR: American, EAS: East Asian, EUR: European, SAS: South Asian) and their HOR-hap barplot as in (C). (**F**) Comparison of the DXZ1 assembly for CHM13 and HG002, which are both in cenhap 2. Tracks are as in (A), with the addition of gray shading to the top track to indicate regions that align closely between the two individuals, and yellow bars indicating regions of increased divergence between the two individuals. Vertical dotted line indicates the homologous site of a CHM13 expansion on the HG002 array. Bottom: Dotplots representing the percent identity of self-alignments within the array, with a color-key and histogram below. (**G**) Left: a zoom-in of the CHM13 kinetochore region with a self-alignment dotplot (exact match word size 2000) revealing patterns of recent nested duplication. Right: a model for the recent evolution in this region. (**H**) Left: comparison of CENP-A NChIP enrichment from CHM13 and HuRef cell lines in DXZ1 HOR-haps, colored as in (A), for 2 replicates (gray/black). Right: comparison of HOR-hap assignments from input controls. Bottom: asterisks highlighting the CENP-A enriched regions from HuRef and CHM13, with respect to the CHM13 array.

Acknowledgements

This work was supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (SN, SK, AR, AMM, SB, AY and AMP), the Intramural funding at the National Institute of Standards and Technology (JMZ), HHMI Hanna H. Gray Fellowship (NA), Damon Runyon Postdoctoral Fellowship, Pew Latin American Fellowship (GVC). Grants from the U.S. National Institutes of Health (NIH/NIGMS F32 GM134558 to GAL; NIH/NHGRI R01 HG00990905 to PS and AFS; NIH R01GM123312-02 to SJH, GAH, RJO; NIH R21CA240199, NSF 1643825 to RJO; NIH/NHGRI F31HG011205 to CJS; NIH/NHGRI R01 HG009190 to AG and WT; NIH/NHGRI R01HG010485, U41HG010972, U01HG010961 to KS; NIH/NIGMS R01GM132600, P20GM103546 and NIH/NHGRI U24HG010136 to DO and TJW; NIH/NHGRI R01 HG010329 to SRS; NIH/NHGRI R01HG010485, U41HG010972, U01HG010961, U24HG011853, OT2OD026682 to BP; NIH/GM R35 GM139653 and R01 GM117420 to GHK; NIH R01 GM124041, R01 GM129263, R21 CA238758 to BAS; NIH/NHGRI R01HG002385, R01HG010169 NIH/NHGRI U01 1U01HG010971 to EEE; NIH/OD/NIMH DP2 OD025824 to MYD; NIH/NHGRI U24HG010263; NHGRI U24HG006620; NCI U01CA253481; NIDDK R24 DK106766-01A1 to MCS; NIH/NHGRI U41HG007234 to MD; NIH/NHGRI R01 1R01HG011274-01, NIH/NHGRI R21 1R21HG010548-01 to KHM), National Science Foundation (NSF 1613806 to SJH, GAH, RJO; NSF DBI-1627442, NSF IOS-1732253, NSF IOS-1758800 to MCS); Mark Foundation for Cancer Research to SA and MCS (19-033-ASP); Russian Science Federation RSF 19-75-30039 (analysis of genomic repeats) to IAA; St. Petersburg State University ((grant ID PURE 73023573) to AM, TD, IAA and (grant ID PURE 51555639) to OK); NIH R01AG054712 to EIR; Ministry of Science and Higher Education of the

Russian Federation (075-10-2020-116 (13.1902.21.0023)) to FG; Connecticut Innovations to RJO; Stowers Institute for Medical Research to JLG; AS is a Chan Zuckerberg Biohub Investigator; EEE and AFD are investigators of the Howard Hughes Medical Institute. Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

Author contributions

αSat sequence characterization: AVB, LU, FDR, AM, VAS, TD, OK, FG, EIR, PAP, NA, IAA, KHM; Pericentromeric satellite characterization: NA, GAL, SJH, MEGS, DO, TJW, LGDL, AMP, RJO, KHM CUT&RUN experiments, mapping, and enrichment analyses: GVC, NA, GAL, PS, SJH, AMM, AR, MEGS, KT, SRS, AS, AFS, BAS, AFD, GHK, AG, WT, KHM; cenhap analysis and interpretation: SAL, NA, CHL, IAA, KHM; array length prediction: MB, JLG, MCS, JMZ; Methylation Analysis: AG, WT; Chromosome imaging and flow sorting: TP, JLG, SB, AY, AMP; Dotplot analysis: LU, FDR, MRV, RL, PK, AMP, IAA; Transposable element analysis: NA, SJH, GAH, RJO, LU, IAA; CHM13 satellite assembly and het analysis: NA, GAL, SN, SK, AR, AMM, AMP; UCSC genome browser and annotation workflow: MD; HiFi assemblies and quality assessment of diverse panel: NA, MA, RL, KS, AM, AVB, SA, JMZ, MCS, BP, EEE, AMP, gene annotation and expression: CJS, MRV, MH, MYD, MD; manuscript writing: NA, IAA, KHM, with input from all authors

Competing interests

SK and KHM have received travel funds to speak at symposia organized by Oxford Nanopore. W.T. has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore Technologies. KHM is a SAB member of Centaura.

Data and materials availability

Sequence data are available through https://www.ncbi.nlm.nih.gov/bioproject/559484

Human Pangenome Reference Consortium (HPRC) generated long and accurate HiFi reads for sixteen human samples HG002, HG003, HG004, HG005, HG006, HG007, HG01243, HG02055, HG02109, HG02723, HG03492, HG01109, HG01442,HG02080,HG02145, and HG03098. We refer to these datasets as HPRC samples. (https://github.com/human-pangenomics/hpgp-data).

Data tracks and satellite annotations can be visualized on the UCSC Genome Browser (*93*, *94*): http://genome.ucsc.edu/cgi-bin/hgTracks?genome=t2t-chm13-v1.0&hubUrl=http://t2t.gi.ucsc.edu/chm13/hub/hub.txt

Annotation data tables and other supporting data and analysis workflows: https://github.com/kmiga/t2t_censat/

Affiliations

[1] Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA
[2] Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
[3] Graduate Program in Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA, USA
[4] Department of Biochemistry, Stanford University, Stanford, California, USA
[5] Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA
[6] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
[7] Sirius University of Science and Technology, Sochi, Russia
[8] Vavilov Institute of General Genetics, Moscow, Russia
[9] Moscow Polytechnic University, Moscow, Russia
[10] Genome Center, MIND Institute, and Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, Davis, Davis, CA, USA
[11] Department of Biology, Johns Hopkins University, Baltimore, MD, USA
[12] Stowers Institute for Medical Research, Kansas City, MO, USA
[13] Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA
[14] Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia
[15] Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[16] UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA
[17] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
[18] Department of Computer Science, University of Montana, Missoula, MT
[19] Reservoir Genomics LLC, Oakland, CA
[20] Howard Hughes Medical Institute, Chevy Chase, MD, USA
[21] NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[22] Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA
[23] Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia
[24] Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA
[25] Chan Zuckerberg Biohub, San Francisco, CA, USA
[26] BioEngineering & BioMedical Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[27] Institute for Quantitative Biosciences (QB3), University of California, Berkeley, Berkeley, CA, US.
[28] Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC, USA
[29] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[30] Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD
[31] Department of Computer Science and Engineering, University of California at San Diego, San Diego, CA, USA
[32] Department of Evolution and Ecology, University of California Davis, Davis, CA, USA
[33] Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia

[34] Department of Biomolecular Engineering, University of California Santa Cruz, CA, USA

References

1. E. E. Eichler, R. A. Clark, X. She, An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004).

2. S. Nurk *et al.*, The complete sequence of a human genome. *bioRxiv* (2021), p. 2021.05.26.445798.

3. K. L. McKinley, I. M. Cheeseman, The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).

4. R. Wevrick, H. F. Willard, Physical map of the centromeric region of human chromosome 7: relationship between two distinct alpha satellite arrays. *Nucleic Acids Res.* **19**, 2295–2301 (1991).

5. M. S. Jackson, P. Slijepcevic, B. A. Ponder, The organisation of repetitive sequences in the pericentromeric region of human chromosome 10. *Nucleic Acids Res.* **21**, 5865–5874 (1993).

6. H. E. Trowell, A. Nagy, B. Vissel, K. H. Choo, Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. *Hum. Mol. Genet.* **2**, 1639–1649 (1993).

7. C. Tyler-Smith, Structure of repeated sequences in the centromeric region of the human Y chromosome. *Development*. **101 Suppl**, 93–100 (1987).

8. I. Tagarro, A. M. Fernández-Peralta, J. J. González-Aguilera, Chromosomal localization of human satellites 2 and 3 by a FISH method using oligonucleotides as probes. *Hum. Genet.* **93**, 383–388 (1994).

9. N. Archidiacono *et al.*, Comparative mapping of human alphoid sequences in great apes using fluorescence in situ hybridization. *Genomics*. **25**, 477–484 (1995).

10. N. Altemose, K. H. Miga, M. Maggioni, H. F. Willard, Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput. Biol.* **10**, e1003628 (2014).

11. K. H. Miga *et al.*, Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).

12. Y. Suzuki, E. W. Myers, S. Morishita, Rapid and ongoing evolution of repetitive sequence structures in human centromeres. *Sci Adv*. **6** (2020), doi:10.1126/sciadv.abd9230.

13. C. Alkan *et al.*, Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS Comput. Biol.* **3**, 1807–1818 (2007).

14. M. Jain *et al.*, Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).

15. K. H. Miga *et al.*, Telomere-to-telomere assembly of a complete human X chromosome. *Nature*.

585, 79–84 (2020).

16. G. A. Logsdon *et al.*, The structure, function and evolution of a complete human chromosome 8. *Nature*. **593**, 101–107 (2021).

17. A. M. M. Cartney *et al.*, Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *bioRxiv* (2021), , doi:10.1101/2021.07.02.450803.

18. R. M. Naylor, J. M. van Deursen, Aneuploidy in Cancer and Aging. *Annu. Rev. Genet.* **50**, 45–66 (2016).

19. S. I. Nagaoka, T. J. Hassold, P. A. Hunt, Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat. Rev. Genet.* **13**, 493–504 (2012).

20. S. Henikoff, K. Ahmad, H. S. Malik, The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*. **293**, 1098–1102 (2001).

21. G. H. Karpen, R. C. Allshire, The case for epigenetic effects on centromere identity and function. *Trends Genet.* **13**, 489–496 (1997).

22. J. C. Wu, L. Manuelidis, Sequence definition and organization of a human repeated DNA. *J. Mol. Biol.* **142**, 363–386 (1980).

23. H. F. Willard, The genomics of long tandem arrays of satellite DNA in the human genome. *Genome*. **31**, 737–744 (1989).

24. H. F. Willard, J. S. Waye, Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987).

25. S. M. McNulty, B. A. Sullivan, Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018).

26. I. Alexandrov, A. Kazakov, I. Tumeneva, V. Shepelev, Y. Yurov, Alpha-satellite DNA of primates: old and new families. *Chromosoma*. **110**, 253–266 (2001).

27. H. F. Willard, Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37**, 524–532 (1985).

28. K. H. Miga, Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes* . **10** (2019), doi:10.3390/genes10050352.

29. R. Oakey, C. Tyler-Smith, Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. *Genomics*. **7**, 325–330 (1990).

30. P. E. Warburton, H. F. Willard, Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages. *J. Mol. Evol.* **41**, 1006–1015 (1995).

31. M. M. Mahtani, H. F. Willard, Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: high-frequency polymorphisms and array size estimate. *Genomics*. **7**,

607–613 (1990).

32. V. A. Shepelev, A. A. Alexandrov, Y. B. Yurov, I. A. Alexandrov, The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet.* **5**, e1000641 (2009).

33. X. She *et al.*, The structure and evolution of centromeric transition regions within the human genome. *Nature*. **430**, 857–864 (2004).

34. G. Genovese *et al.*, Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406–14, 414e1–2 (2013).

35. C. Lee, R. Wevrick, R. B. Fisher, M. A. Ferguson-Smith, C. C. Lin, Human centromeric DNAs. *Hum. Genet.* **100**, 291–304 (1997).

36. Karen H. Miga and Ivan A. Alexandrov, Variation and evolution of human centromeres: A field guide and perspective. *Annu. Rev. Genet.* (2021).

37. J. Prosser, M. Frommer, C. Paul, P. C. Vincent, Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* **187**, 145–155 (1986).

38. M. Frommer, J. Prosser, P. C. Vincent, Human satellite I sequences include a male specific 2.47 kb tandemly repeated unit containing one Alu family member per repeat. *Nucleic Acids Res.* **12**, 2887–2900 (1984).

39. M. Babcock, S. Yatsenko, P. Stankiewicz, J. R. Lupski, B. E. Morrow, AT-rich repeats associated with chromosome 22q11.2 rearrangement disorders shape human genome architecture on Yq12. *Genome Res.* **17**, 451–460 (2007).

40. A. Agresti *et al.*, Linkage in human heterochromatin between highly divergent Sau3A repeats and a new family of repeated DNA sequences (HaeIII family). *J. Mol. Biol.* **205**, 625–631 (1989).

41. R. Meneveri *et al.*, Molecular organization and chromosomal location of human GC-rich heterochromatic blocks. *Gene*. **123**, 227–234 (1993).

42. S. J. Hoyt, J. M. Storer, G. A. Hartley, P. G. S. Grady, From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *bioRxiv* (2021) (available at https://www.biorxiv.org/content/10.1101/2021.07.12.451456.abstract).

43. M. R. Vollger *et al.*, Segmental duplications and their variation in a complete human genome. *bioRxiv* (2021), p. 2021.05.26.445678.

44. Materials and methods are available as supplementary materials at the Science website.

45. V. A. Shepelev *et al.*, Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genom Data*. **5**, 139–146 (2015).

46. L. I. Uralsky *et al.*, Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome

assembly. *Data Brief*. **24**, 103708 (2019).

47. A. W. Lo, G. C. Liao, M. Rocchi, K. H. Choo, Extreme reduction of chromosome-specific alpha-satellite array is unusually common in human chromosome 21. *Genome Res.* **9**, 895–908 (1999).

48. M. K. Rudd, H. F. Willard, Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **20**, 529–533 (2004).

49. A. E. Kazakov *et al.*, Interspersed repeats are found predominantly in the “old” α satellite families. *Genomics*. **82**, 619–627 (2003).

50. P. E. Warburton, H. F. Willard, Genomic analysis of sequence variation in tandemly repeated DNA. Evidence for localized homogeneous sequence domains within arrays of alpha-satellite DNA. *J. Mol. Biol.* **216**, 3–16 (1990).

51. HPRC, Human Pangenome Reference Consortium. *T2T Diversity Panel* (2021), (available at https://github.com/human-pangenomics/hpgp-data).

52. M. G. Schueler, A. W. Higgins, M. K. Rudd, K. Gustashaw, H. F. Willard, Genomic and genetic definition of a functional human centromere. *Science*. **294**, 109–115 (2001).

53. R. Bandyopadhyay, C. McQuillan, S. L. Page, K. H. Choo, L. G. Shaffer, Identification and characterization of satellite III subfamilies to the acrocentric chromosomes. *Chromosome Res.* **9**, 223–233 (2001).

54. D. P. Ryan *et al.*, Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell*. **140**, 88–98 (2010).

55. M.-L. Dubois *et al.*, UBB pseudogene 4 encodes functional ubiquitin variants. *Nat. Commun.* **11**, 1306 (2020).

56. Sergey Aganezov, Stephanie M. Yan, Daniela C. Soto, Melanie Kirsche, Samantha Zarate, A complete reference genome improves analysis of human genetic variation. *bioRxiv (in review)* (2021).

57. I. A. Alexandrov, S. P. Mitkevich, Y. B. Yurov, The phylogeny of human chromosome specific alpha satellites. *Chromosoma*. **96**, 443–453 (1988).

58. H. F. Willard, J. S. Waye, Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* **25**, 207–214 (1987).

59. P. Finelli *et al.*, Structural organization of multiple alphoid subsets coexisting on human chromosomes 1, 4, 5, 7, 9, 15, 18, and 19. *Genomics*. **38**, 325–330 (1996).

60. S. J. Durfy, H. F. Willard, Concerted evolution of primate alpha satellite DNA. Evidence for an ancestral sequence shared by gorilla and human X chromosome alpha satellite. *J. Mol. Biol.* **216**, 555–566 (1990).

61. P. E. Warburton, R. Wevrick, M. M. Mahtani, H. F. Willard, Pulsed-Field and Two-Dimensional Gel

Electrophoresis of Long Arrays of Tandemly Repeated DNA: Analysis of Human Centromeric Alpha Satellite. *Pulsed-Field Gel Electrophoresis*, pp. 299–318.

62. A. V. Bzikadze, P. A. Pevzner, Automated assembly of centromeres from ultra-long error-prone reads. *Nat. Biotechnol.* **38**, 1309–1316 (2020).

63. K. H. Miga, Centromere studies in the era of “telomere-to-telomere” genomics. *Exp. Cell Res.* **394**, 112127 (2020).

64. M. D. Blower, B. A. Sullivan, G. H. Karpen, Conserved organization of centromeric chromatin in flies and humans. *Dev. Cell*. **2**, 319–330 (2002).

65. L. L. Sullivan, C. D. Boivin, B. Mravinac, I. Y. Song, B. A. Sullivan, Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosome Res.* **19**, 457–470 (2011).

66. A. A. Van Hooser *et al.*, Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. *J. Cell Sci.* **114**, 3529–3542 (2001).

67. B. E. Black, D. W. Cleveland, Epigenetic centromere propagation and the nature of CENP-a nucleosomes. *Cell*. **144**, 471–479 (2011).

68. B. A. Sullivan, G. H. Karpen, Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat. Struct. Mol. Biol.* **11**, 1076–1083 (2004).

69. A. Gershman *et al.*, Epigenetic Patterns in a Complete Human Genome. *bioRxiv* (2021), p. 2021.05.26.443420.

70. P. J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*. **6** (2017), doi:10.7554/eLife.21856.

71. Y. B. Yurov, S. P. Mitkevich, I. A. Alexandrov, Application of cloned satellite DNA sequences to molecular-cytogenetic analysis of constitutive heterochromatin heteromorphisms in man. *Hum. Genet.* **76**, 157–164 (1987).

72. B. Erdtmann, Aspects of evaluation, significance, and evolution of human C-band heteromorphism. *Hum. Genet.* **61**, 281–294 (1982).

73. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

74. M. Nambiar, G. R. Smith, Repression of harmful meiotic recombination in centromeric regions. *Semin. Cell Dev. Biol.* **54**, 188–197 (2016).

75. S. A. Langley, K. H. Miga, G. H. Karpen, C. H. Langley, Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *Elife*. **8** (2019), doi:10.7554/eLife.42989.

76. M. Byrska-Bishop *et al.*, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* (2021), p. 2021.02.06.430068.

77. J. G. Henikoff, J. Thakur, S. Kasinathan, S. Henikoff, A unique chromatin complex occupies young α-satellite arrays of human centromeres. *Sci Adv*. **1** (2015), doi:10.1126/sciadv.1400234.

78. J. Thakur, S. Henikoff, CENPT bridges adjacent CENPA nucleosomes on young human α-satellite dimers. *Genome Res.* **26**, 1178–1187 (2016).

79. D. Hasson *et al.*, The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat. Struct. Mol. Biol.* **20**, 687–695 (2013).

80. M. Dumont *et al.*, Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. *EMBO J.* **39**, e102924 (2020).

81. S. J. Falk *et al.*, CENP-C reshapes and stabilizes CENP-A nucleosomes at the centromere. *Science*. **348**, 699–703 (2015).

82. K. A. Maloney *et al.*, Functional epialleles at an endogenous human centromere. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 13704–13709 (2012).

83. M. E. Kuo, L. L. Sullivan, K. Chew, B. A. Sullivan, Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome* (2016) (available at http://genome.cshlp.org/content/26/10/1301.short).

84. M. E. Aldrup-MacDonald, M. E. Kuo, L. L. Sullivan, K. Chew, B. A. Sullivan, Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* **26**, 1301–1311 (2016).

85. K. E. Hayden *et al.*, Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33**, 763–772 (2013).

86. W. R. Rice, A Game of Thrones at Human Centromeres II. A new molecular/evolutionary model. *Cold Spring Harbor Laboratory* (2019), p. 731471.

87. M. Naish *et al.*, The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science*. **374**, eabi7489 (2021).

88. L. Chmátal *et al.*, Centromere Strength Provides the Cell Biological Basis for Meiotic Drive and Karyotype Evolution in Mice. *Curr. Biol.* **24**, 2295–2300 (2014).

89. L. Fishman, J. K. Kelly, Centromere-associated meiotic drive and female fitness variation in Mimulus. *Evolution*. **69** (2015), pp. 1208–1218.

90. L. E. Kursel, H. S. Malik, The cellular mechanisms and consequences of centromere drive. *Curr. Opin. Cell Biol.* **52**, 58–65 (2018).

91. P. J. Park, ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

92. N. Altemose *et al.*, DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome-wide. *bioRxiv* (2021), p. 2021.07.06.451383.

93. B. J. Raney *et al.*, Track data hubs enable visualization of user-defined genome-wide annotations

on the UCSC Genome Browser. *Bioinformatics*. **30**, 1003–1005 (2014).

94. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

95. O. Kunyavskaya, T. Dvorkina, A. V. Bzikadze, I. A. Alexandrov, P. A. Pevzner, HORmon: automated annotation of human centromeres. *bioRxiv* (2021), p. 2021.10.12.464028.

96. O. K. Smith *et al.*, Identification and characterization of centromeric sequences in Xenopus laevis. *Genome Res.* **31**, 958–967 (2021).

97. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker (1996).

98. D. Olson, T. Wheeler, ULTRA: A Model Based Tool to Detect Tandem Repeats. *ACM BCB*. **2018**, 37–46 (2018).

99. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).

100. T. J. Wheeler, S. R. Eddy, nhmmer: DNA homology search with profile HMMs. *Bioinformatics*. **29**, 2487–2489 (2013).

101. J. S. Waye, H. F. Willard, Chromosome-specific alpha satellite DNA: nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. *Nucleic Acids Res.* **13**, 2731–2743 (1985).

102. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

103. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

104. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

105. E. M. Black, S. Giunta, Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes* . **9** (2018), doi:10.3390/genes9120615.

106. V. Sevim, A. Bashir, C.-S. Chin, K. H. Miga, Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics*. **32**, 1921–1924 (2016).

107. S. Levy *et al.*, The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).

108. C. Jain *et al.*, Weighted minimizer sampling improves long read mapping. *Bioinformatics*. **36**, i111–i118 (2020).

109. L. Manuelidis, Repeating restriction fragments of human DNA. *Nucleic Acids Res.* **3**,

3063–3076 (1976).

110. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. **14**, 417–419 (2017).

111. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).

112. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. **31**, 166–169 (2015).

113. H.-D. Li, GTFtools: a Python package for analyzing various modes of gene models. *bioRxiv* (2018), p. 263517.

114. T. Lassmann, E. L. L. Sonnhammer, Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. **6**, 298 (2005).

115. F. Madeira *et al.*, The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).

116. R. Ihaka, R. Gentleman, R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).

117. C. Ginestet, ggplot2: elegant graphics for data analysis. *JOURNAL-ROYAL STATISTICAL SOCIETY SERIES A*. **174**, 245–245 (2011).

118. K. Tamura *et al.*, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).

119. W. N. Venables, B. D. Ripley, in *Modern Applied Statistics with S*, W. N. Venables, B. D. Ripley, Eds. (Springer New York, New York, NY, 2002), pp. 271–300.

120. C. R. Beck *et al.*, LINE-1 retrotransposition activity in human genomes. *Cell*. **141**, 1159–1170 (2010).

121. A. Damert *et al.*, 5’-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res.* **19**, 1992–2008 (2009).

122. B. L. Ng, N. P. Carter, Factors affecting flow karyotype resolution. *Cytometry A*. **69**, 1028–1036 (2006).

123. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018).

124. M. Kokot, M. Dlugosz, S. Deorowicz, KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*. **33**, 2759–2761 (2017).

125. S. R. Eddy, Profile hidden Markov models. *Bioinformatics*. **14**, 755–763 (1998).

126. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

127. B. J. Raney *et al.*, Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*. **30**, 1003–1005 (2014).

128. W. Ziccardi *et al.*, Clusters of alpha satellite on human chromosome 21 are dispersed far onto the short arm and lack ancient layers. *Chromosome Res.* **24**, 421–436 (2016).

129. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

130. M. Ventura *et al.*, Evolutionary formation of new centromeres in macaque. *Science*. **316**, 243–246 (2007).

131. O. Capozzi *et al.*, Evolutionary descent of a human chromosome 6 neocentromere: a jump back to 17 million years ago. *Genome Res.* **19**, 778–784 (2009).

132. K. H. Miga, T. Wang, The Need for a Human Pangenome Reference Sequence. *Annu. Rev. Genomics Hum. Genet.* (2021), doi:10.1146/annurev-genom-120120-081921.

133. J. Thakur, S. Henikoff, Unexpected conformational variations of the human centromeric chromatin complex. *Genes Dev.* **32**, 20–25 (2018).

134. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10–12 (2011).

135. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. **27**, 764–770 (2011).

136. W. Huang, L. Li, J. R. Myers, G. T. Marth, ART: a next-generation sequencing read simulator. *Bioinformatics*. **28**, 593–594 (2012).

137. A. Mikheenko, A. V. Bzikadze, A. Gurevich, K. H. Miga, P. A. Pevzner, TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics*. **36**, i75–i83 (2020).

138. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*. **18**, 170–175 (2021).

139. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).

140. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).

141. A. Rzhetsky, M. Nei, Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* **10**, 1073–1095 (1993).

142. M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics* (Oxford University Press, 2000).

143. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

144. H. Rosenberg, M. Singer, M. Rosenberg, Highly Reiterated Sequences of SIMIANSIMIANSIMIANSIMIANSIMIAN. *Science*. **200** (1978), pp. 394–402.

145. P. R. Musich, F. L. Brown, J. J. Maio, Highly repetitive component alpha and related alphoid DNAs in man and monkeys. *Chromosoma*. **80**, 331–348 (1980).

146. H. J. Cooke, J. Hindley, Cloning of human satellite III DNA: different components are on different chromosomes. *Nucleic Acids Res.* **6**, 3177–3197 (1979).

147. J. S. Waye, H. F. Willard, Human beta satellite DNA: genomic organization and sequence definition of a class of highly repetitive tandem DNA. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6250–6254 (1989).

148. R. Metzdorf, E. Göttert, N. Blin, A novel centromeric repetitive DNA from human chromosome 22. *Chromosoma*. **97**, 154–158 (1988).

149. C. C. Lin, R. Sasi, C. Lee, Y. S. Fan, D. Court, Isolation and identification of a novel tandemly repeated DNA sequence in the centromeric region of human chromosome 8. *Chromosoma*. **102**, 333–339 (1993).

150. C. Lee, X. Li, E. W. Jabs, D. Court, C. C. Lin, Human gamma X satellite DNA: an X chromosome specific centromeric DNA sequence. *Chromosoma*. **104**, 103–112 (1995).

151. P. E. Warburton *et al.*, Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics*. **9**, 533 (2008).