genes

MDPI

*Review*

# Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population

**Karen H. Miga**

UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, CA 95064, USA; khmiga@soe.ucsc.edu; Tel.: +1-831-459-5232

Received: 2 April 2019; Accepted: 3 May 2019; Published: 8 May 2019


**Abstract:** The central goal of medical genomics is to understand the inherited basis of sequence variation that underlies human physiology, evolution, and disease. Functional association studies currently ignore millions of bases that span each centromeric region and acrocentric short arm. These regions are enriched in long arrays of tandem repeats, or satellite DNAs, that are known to vary extensively in copy number and repeat structure in the human population. Satellite sequence variation in the human genome is often so large that it is detected cytogenetically, yet due to the lack of a reference assembly and informatics tools to measure this variability, contemporary high-resolution disease association studies are unable to detect causal variants in these regions. Nevertheless, recently uncovered associations between satellite DNA variation and human disease support that these regions present a substantial and biologically important fraction of human sequence variation. Therefore, there is a pressing and unmet need to detect and incorporate this uncharacterized sequence variation into broad studies of human evolution and medical genomics. Here I discuss the current knowledge of satellite DNA variation in the human genome, focusing on centromeric satellites and their potential implications for disease.

**Keywords:** satellite DNA; centromere; sequence variation; structural variation; repeat; alpha satellite; human satellites; genome assembly

## 1. Introduction

Genome-scale initiatives, such as the Human Genome Project and the 1000 Genome (1KG) consortium [1–3], have provided a wealth of genomic information that have greatly advanced basic and biomedical research. However, in light of this progress, the millions of bases that span each human centromeric region remain largely disconnected from contemporary genetic and genomic analyses. This has historically been due to the challenge of generating and validating linear assemblies of tandemly-repeated DNA (e.g., thousands of copies of a repeat with a limited number of sequence variants to guide overlap-consensus derived assemblies), which are known to span each centromeric region [4]. Our understanding of the sequence content and organization of human centromeres improved dramatically with the release of the GRCh38 reference genome, and recent efforts to generate true linear assemblies using "ultra-long" sequencing (i.e., reads that span hundreds of kilobases [5]), wherein the centromere-assigned gaps on each chromosome assembly were updated with sequence information [6,7]. Thus, we are entering a new era in genomics where centromeric DNAs are available for detailed study, either within a single karyotype or across human populations, that will drive research aimed to understand repeat variation that contributes to genome stability, population variation, and disease.

Centromeric satellite DNA arrays are known to vary extensively in the human population, yet few genomic tools have been developed to study the full extent of this sequence variation, thereby ignoring a fraction of the human genome expected to contribute directly to cancer and human

disease [8–10]. The extent of variation has been documented at the cytogenetic level, and gross estimates of rearrangement and/or repeat expansion have been associated with cancer and infertility [11–13]. Additionally, the epigenetic regulation of satellite DNAs, as well as anomalous methylation and altered transcription of satellite DNAs, have been associated with human diseases [9,14]. However, these early observations are challenged by inconsistencies in association [15], small sample size and perhaps an incomplete (or often, low-resolution) understanding of underlying genomic structure and array variant composition.

Acknowledging the differences between the satellite arrays, there is limited utility in restricting studies to the use of a single genomic map. Rather, it is important to extend our survey of satellite DNA genomics to large panels of diverse individuals, thus enabling high-resolution maps of human sequence diversity in these regions. Such sampling efforts would be the foundation for a modern era of satellite DNA genomics: establishing allelic frequencies for satellite variants necessary to expand disease-association studies. Here I discuss our current understanding of satellite DNA variation as determined from whole-genome sequencing projects, with a focus on the largest families in the human genome and their association with disease.

## 2. What Proportion of the Human Genome is Defined by Peri/Centromeric Satellite DNAs?

The largest arrays of satellite DNAs in the human genome are organized within centromeric and pericentromeric regions [2,3,16]. Although several distinct satellite DNA families are known to contribute to pericentromeric regions (e.g., gamma, beta, and subtelomeric satellites [17–19]), this review is focused on alpha satellite and human satellites 2, 3, which are the most abundant in the human genome and most commonly associated with human disease [8]. The alpha satellite DNA family is defined by a group of related, highly divergent AT-rich repeats or 'monomers', each approximately 171 bp in length, which are found in every normal human centromere [20–22]. Previous genome-wide estimates of alpha satellite have observed that this family represents ~2.6% of the human genome [23], which roughly aligns with early hybridization-based estimates [16]. Additionally, previous physical maps of centromeric regions, and pulse-field gel electrophoresis (PFGE) southern-based estimates of chromosome-assigned satellite arrays, revealed an average (~3 Mbps) amount of alpha satellite per centromeric region [24–26]. Therefore, we can assign a very rough estimate of 72 Mbps across all 22 autosomes and two sex chromosomes (i.e., 2.4%), which remains in agreement with all previous genome-wide estimates. Human satellites 2, 3 (HSat2,3), are collectively defined by enrichment of a pentameric repeat, (CATTC)n, and represent the largest heterochromatin blocks (documented as at least 10 Mbps in length) in human pericentromeric regions; notably, on chromosomes 1, 9, 16, and Y [27–32]. In total, HSat2,3 are observed to be less abundant than alpha satellite (~1.5% of the genome) [31], yet early estimates of array lengths on the DYZ1 array suggest that abundance estimates may vary considerably in the human population [31,33–35]. Therefore, efforts to better understand the true extent of satellite subfamily overall variation would benefit from surveying a much larger panel of diverse individuals. In doing so, one can define the lower and upper bounds of satellite DNA content in the genome. For example, can one individual have 3% alpha satellite and another individual have closer to 10%? What defines these bounds? Further, do these fluctuations in overall satellite composition contribute to our understanding of chromosome segregation, genome stability, and disease?

Genome-wide estimates of alpha satellite and HSat2,3 content have relied on constructing comprehensive sequence databases using raw read data from each satellite DNA family, thereby avoiding underestimates due to assembly collapse of identical/near-identical repeats [23,31]. Alpha satellite and HSat2,3 exhibit considerable sequence heterogeneity [20,21,32], as observed most readily in the ability to hybridize specifically to divergent repeat sequences within chromosome-specific arrays [20,36–38]. Therefore, efforts to construct genome-wide libraries from short-read datasets rely on methods that are comprehensive and inclusive with respect to the potential heterogeneity within satellite families. One approach is to reformat published satellite sequence libraries [7,23,31] into

catalogs of short oligonucleotide sequences (24 bps, representing sequences in both orientations) that are specific to a given satellite family; that is, each oligo is only observed within a respective satellite database and never observed to have an exact match anywhere else in the genome. It is then possible to survey existing low-coverage, publicly available, population datasets, such as 1KG data [1], to identify what percentage of reads (as determined by exact matches with oligo libraries) are assigned to each respective satellite family (Figure 1a,b). In a study of low-coverage 1KG sequence data representing 14 diverse populations (400 male individuals and 414 female individuals) [1,7], alpha satellite has a median of 3.1% genome-wide estimate, with a range between 1% and 5% (Figure 1a). These initial HSat2,3 estimates reveal that although this satellite family is typically less abundant than alpha (median, 2.1%, range ~1–7%), it is observed in many cases to match, or surpass, alpha satellite abundance [31] (Figure 1b).

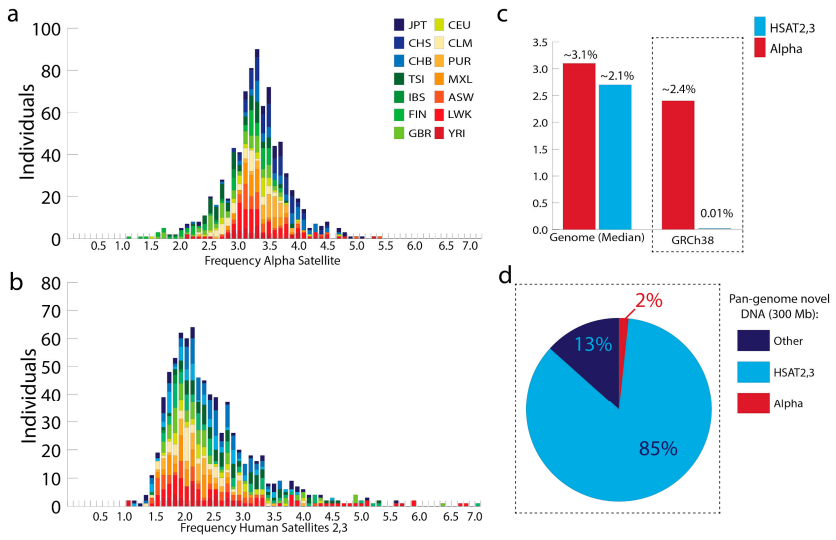


**Figure 1.** Proportion of alpha satellite and human satellites 2,3 in the human population. Using 1KG [1] data representing 14 diverse populations (400 male individuals and 414 female individuals) (a) the frequency of 24-mers that have an exact match with alpha satellite [23], (b) the frequency of 24-mers that have an exact match with human satellite 2,3 [31]. (c) Median frequencies (from panel (a) alpha and (b) HSat2,3) are listed relative to the observed frequency in the human reference genome assembly (GRCh38; GCA_000001405.15). (d) Evaluation of 300 Mb of DNA from the collective genomes of 910 people of African descent, previously determined to be missing or unaligned to GRCh38 [39]. Key for human subpopulations: CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; CHS: Southern Han Chinese; CEU: Utah Residents (CEPH) with Northern and Western European Ancestry; TSI: Toscani in Italia; FIN: Finnish in Finland; GBR: British in England and Scotland; IBS: Iberian Population in Spain; YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya; GWD: Gambian in Western Divisions in the Gambia; ASW: Americans of African Ancestry in SW USA; MXL: Mexican Ancestry from Los Angeles USA; PUR: Puerto Ricans from Puerto Rico; CLM: Colombians from Medellin, Colombia.

Proper representation of satellite DNAs in human reference assemblies will be critical to ensure faithful short read mapping and accurate assessments of satellite family variability in the future. Notably, the addition of millions of bases of alpha satellite "reference models" with the release of the GRCh38 reference genome has provided initial short read mapping targets [7,40]. This has proven to be useful in decreasing off-target alignments and has enabled high-resolution studies aimed to study the epigenetic structure in centromeres [41,42]. In contrast, HSat2,3 are woefully underrepresented in all human assemblies, with only ~0.01% representation in GRCh38 (Figure 1c). This can lead to pronounced differences in the way we annotate and study human variation. For example, a recent study of 910 individuals of African descent identified roughly 300 Mbp of sequences not present in

the human reference (GRCh38) [39], of which the largest proportion have exact oligo matches with HSat2,3 (Figure 1d). This demonstrates the importance of proper satellite DNA representation in the reference assembly in shaping our interpretation of novel sequences in the population. Notably, alpha satellite, a satellite family with great representation in the current reference assembly, still has candidate sequences that are not aligned to the reference models derived from the HuRef genome (Figure 1d, red). This result emphasizes a second important point: because satellites are expected to vary between genomes, the use of a single individual's genome as a reference, in this case HuRef [43], is not sufficient to capture the sequence diversity in the human population. Expanding the representation of human sequence diversity has been previously shown to improve mapping of variants [44,45], highlighting the need for a 'pan-human genome reference' to improve mapping efficiency and satellite variation studies in the future.

## 3. What is the Nature of Sequence Variation within a Single Satellite Array?

The variation of satellite DNAs genome-wide abundance is driven by repeat expansion and contraction, commonly attributed to mechanisms of non-homologous crossover and/or conversion [46,47]. Genomic-based studies of satellite DNA evolution have greatly benefited from the advancement of software designed to study tandem repeat variation in unassembled reads (reviewed [48]). The advancement of such high-resolution studies across a broad number of species is expected to dramatically advance our knowledge of the rates and mechanisms driving satellite array evolution. Previous studies of comprehensive studies of satellite DNA classes. Efforts in the past to study satellite repeat variation have focused on shorter microsatellites and tandem repeat classes that are amenable to complete assembly using long read technologies. However, recent efforts to study tandem repeat variation in human rDNA arrays revealed a high level of heterogeneity (i.e., an average rate of 7.5 variants per kb). Each rDNA unit is 45 kb with roughly 500 copies per diploid cell, and much like the satellite arrays, rDNA array length can vary significantly in size from just a few units to >100 between individuals [49,50].

The relationship between repeat units from different alpha satellite arrays would suggest that the rate of intra-chromosomal exchange (i.e., sister chromatid exchange) is higher than inter-homologue exchange [51]. As a result, most satellite arrays in the human genome can be defined by highly homogeneous arrays that can be often typified by chromosome-specific multi-monomeric repeat units, or higher-order repeats (HOR) [20,52]. Although the chromosome-assignment of HORs is largely invariant between individuals, as demonstrated by the effectiveness of commercially available satellite Fluorescence in situ hybridization (FISH) markers for chromosome labeling in clinical cytogenetics, the thousands of copies of the HOR that comprise a single array are expected to represent a mixture of expansion/contraction of repeat variants, shifts in orientation, and mobile element insertions (Figure 2a) [7,20,37,53].

This ever-changing genomic landscape guides our understanding of centromere function and chromosome stability. For example, the repeat structure and array length are expected to change the frequency of and spacing of the 17-bp centromere protein B binding motif, or CENP-B box. The functional role of CENP-B at human centromeres is not yet fully understood [54,55], yet recent studies suggest that the periodicity may contribute to kinetochore function and centromere fidelity [56,57]. Rearrangement in canonical HOR units, i.e., insertions and/or deletions presumably due to unequal crossing-over events, are observed at different frequencies between spatially distinct arrays. The Chr17-specific alpha satellite HOR (D17Z1) is characterized by arrays containing approximately 1000 repeat units that range in length from 11–16 monomers [36,56,58]. The frequency and ordering of these variants have been shown to influence the centromere location on human chromosomes with metastable epialleles [59,60]. Ultimately, sequence composition within each satellite array is thought to influence expression of the repeats [10,61,62], transcription factor binding [63,64], and replication efficiency [65–67]. Therefore, the high-resolution and comprehensive study of array sequence composition and structure is key to our understanding of how these specialized loci function.

Previous methods have used unassembled reads from whole-genome sequencing projects to evaluate chromosome-specific satellite overall abundance, or copy number, and the frequency of variants (e.g., HOR rearrangements, inversions, transposition, and single nucleotide variants (SNVs)) within the array [7,31,68]. Specifically, the centromeric regions on the X and Y chromosomes in male genomes offer a unique opportunity to study the variation in haploid array length within the human population. Altemose et al. [31], estimated the array size using low-coverage, short-read sequencing data from 396 male 1KG individuals [1], showing that the DYZ1 array varies over an order of magnitude (7–98 Mbps, with a mean of 24 Mb), consistent with previous experimental observations of Y-chromosome length variability [34,69,70]. Similarly, 1KG read-depth-based estimates of alpha satellite array lengths on the X and Y chromosomes (DXZ1 and DYZ3) agree with prior PFGE Southern experiments [7,25,71]. Although the X array has been predicted to have a 10-fold size range (800 kb to 8 Mbps), the medians of predicted X array lengths per human population, are observed to fall within experimentally validated lengths of 2.2–3.7 Mbps (mean 3010 kb) [7,25] (Figure 2b). This further corroborates the accuracy of short-read-based array length estimates applied to diverse groups of people.

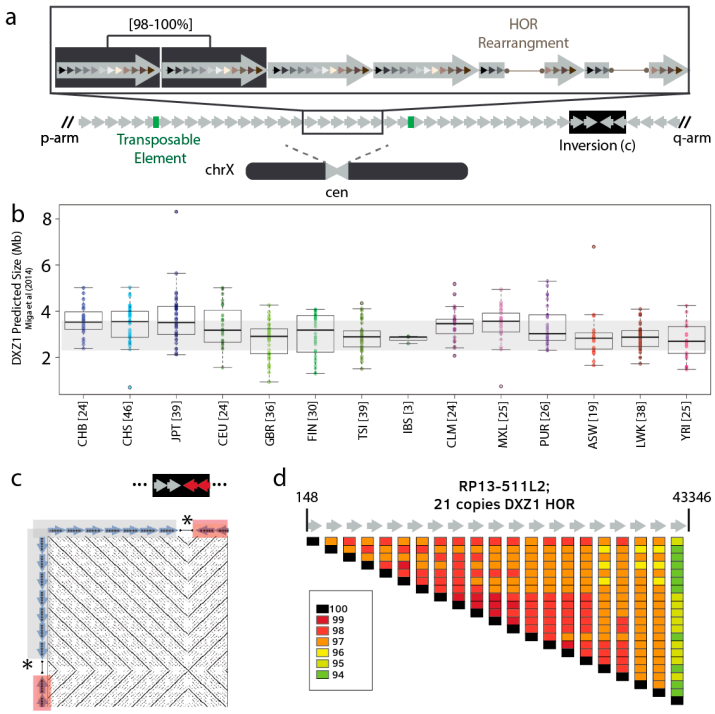


**Figure 2.** Intra-array satellite sequence variation. (a) All normal human centromeric regions contain at least one alpha satellite array, shown in grey, which is tandemly organized in a head-to-tail orientation with occasionally transposable element interruptions (green) and shifts in directionality (black box). The fundamental alpha satellite repeat unit, or ~171 bp monomer, is shown in a variation of shaded colors to illustrate the heterogeneity of the sequencing identity. Multi-monomer repeat units, or 'higher-order repeats (HORs), are shown by the larger grey arrows that encompass the collection of smaller repeats. In contrast to the individual monomers, these repeats are shown to be identical, or near-identical (98–100%). In addition to single nucleotide differences between the HORs, larger rearrangements (shown as a deletion of five monomers) are observed to occur and expand and contract within the array. (b) Satellite array length predictions on the X chromosome (DXZ1) [7], grey shading marks the previously observed PFGE Southern length range [25]. (c) Inversion detected using error-corrected PacBio reads [68]. (d) RP13-511L2 is an X-specific BAC that represents the transition from core alpha satellite to the edge of the array. HOR pair-wise repeat identity (muscle alignment [72]) showing increased divergence approaching the chromosome arm (43,346 bp), as typically observed at the edge of the array.

Use of error-corrected long reads (e.g., Pacific Biosciences, PacBio) prior to assembly provide an automated method to identify larger structural variation (SV) in satellite arrays, such as: HOR rearrangement (insertion and/or deletion), inversions, and interruption by transposons [68]. In addition to changes in the HOR structure, one can monitor precise sites where shifts in orientation or inversions take place within the array (Figure 2c). Further, when tracking sites of transposable element insertion, LINE1 is documented to be the most prevalent, consistent with the literature of alpha satellite DNA [73]. In addition to advancing our understanding of sequence organization and centromere function, such low-copy sequence variants that interrupt the uniformity of the satellite array are also expected to also guide linear assembly efforts [4,6]. Likewise, low-copy SNVs have been shown to be useful in overlap-consensus assembly, but they depend on high sequencing accuracy often obtained from Illumina reads and/or high-coverage of long-read data [6]. Satellite DNA studies using bacterial artificial chromosome (BAC) data provide a snapshot of local SNV spacing, where increased divergence is expected at the edges of the array (closest to the transition with the chromosome arms) with sparse, and infrequent informative sites within the array (Figure 2d) [73]. Ultimately, efforts to construct robust databases of satellite-associated SVs and SNVs will benefit from additional high-coverage, long read (PacBio or nanopore sequencing) datasets from diverse individuals. Such databases would provide allele-frequency data needed to guide future disease associations of variants.

## 4. Centromeric Regions Span Variants Associated with Disease.

Entire multi-megabase-sized centromeric regions, including the heterochromatic regions in the pericentromere, suppress meiotic recombination and are commonly observed as a single haplotype block, or 'cenhap' (Figure 3a) [74]. Little is known about the unique evolution and regulatory properties of those sequences that are associated with these highly specialized regions. Position effect variegation (PEV), or the mosaic pattern of gene expression when placed within or near heterochromatic environments, has been observed in organisms from yeast to humans [71]. The extensive range of satellite array sizes observed within the human population may contribute to studies of PEV variability and gene regulation in the human genome. Sequences directly adjacent to the centromeric satellite arrays have been documented as hypermutable, with a speculation that the increased mutation rate may be attributed to centromere activity [73,74]. Further, genes that are largely excluded by recombination may influence the efficacy of selection and create a 'protected' environment for gene mutations, inheritance, and disease.

These immense linkage blocks encompass satellite DNAs, segmental duplications [75], and a collection of well-annotated genes [76], many of which have been previously attributed to human clinical and disease phenotypes. Although the functional implications of gene-level associations are difficult to infer due to the large region of linkage disequilibrium, it may be useful for studies to recognize and bin these centromere-associated genomic regions as it is likely that they are share a compartment of the genome with specialized inheritance and evolution. The Xq cenhap region contains eight genes that are documented in the Online Mendelian Inheritance in Man (OMIM) noting the potential for allelic variants to represent disease-causing mutations (Figure 3b) [77,78]. Additionally, genome-wide association studies (GWAS) have identified SNPs in cenhap regions as associated with human disease (as shown in Figure 3b for NHGRI-EBI GWAS data, each selected with p-values $< 1.0 \times 10–5$), many of which do not overlap with genes or annotated sequence features [79,80]. Studies of variants directly adjacent to centromeric regions have been associated with chromosome instability and disease. For example, multiple independent signals associated with chromosome X loss around the centromere of chromosome X have been reported in a study of mosaic chromosomal alterations in clonal hematopoiesis [81], with a strong association ($P = 6.6 \times 10–27$, with an observed 1.9:1 bias in the lost haplotype) near the centromere array (DXZ1, Xp11.1). Further examples of centromere-adjacent or associated SNPs have been used to predict a significant association with multiple sclerosis risk around the chromosome 1 (lod = 4.9; with initial scan of 484 cases and 1043 controls; genotyped at 1082 SNPs) [82]. It is likely that many other disease association loci

exist in these centromere-proximal regions, as association with centromeric SNPs (defined as within 2 Mbps of an alpha satellite reference model in GRCh38 that do not overlap with a known gene or segmental duplication) have been observed for a variety of clinical studies, including various cancers, neurodegenerative disorders and cardiovascular diseases (Table 1) [80,83].

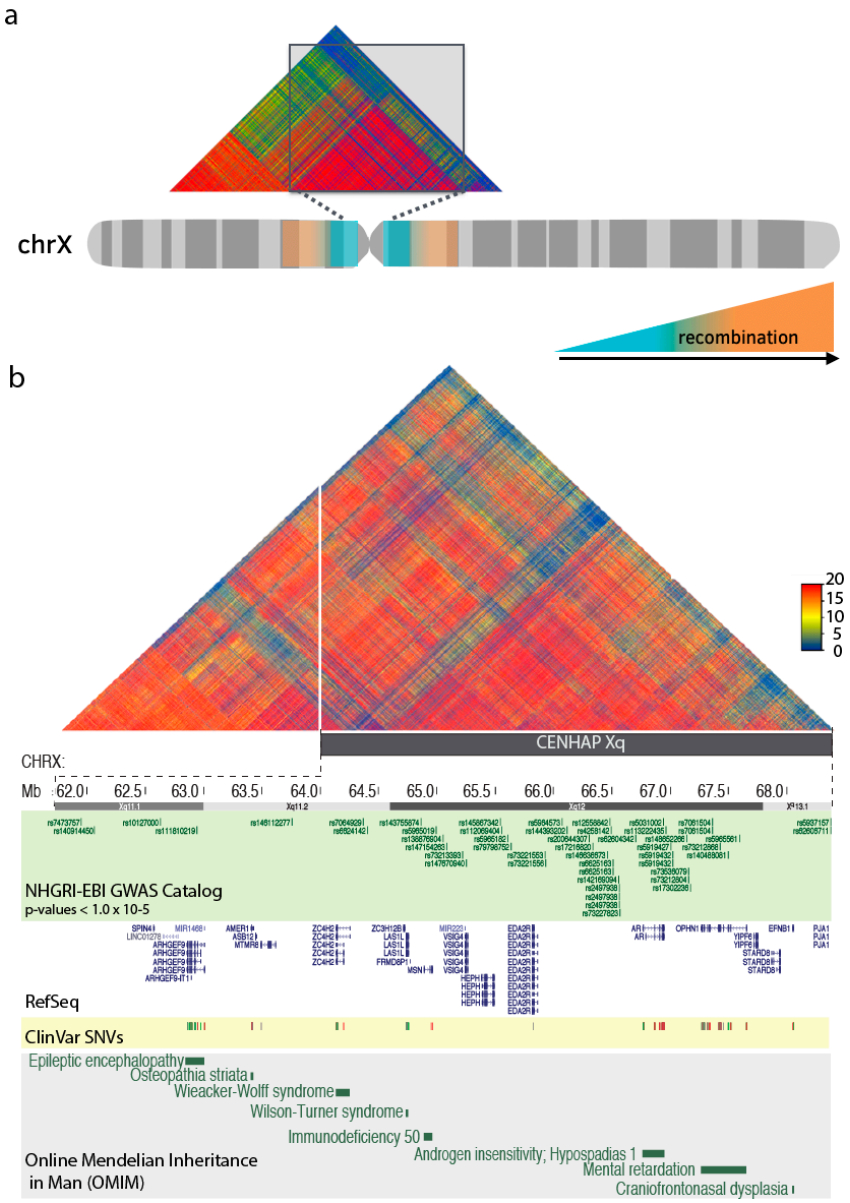


**Figure 3.** Disease-associated variants in centromere-associated haplotypes. (a) Centromeres act as the primary constriction of chromosomes, and are historically defined by the reduction of meiotic recombination (indicated by blue). Therefore, sequences in these regions are commonly inherited in large linkage blocks, or cenhaps (shown in the linkage disequilibrium heat map) [70]. (b) Study of disease and clinically associated single nucleotide variants (GWAS Catalog (green), ClinVar SNVs (yellow) in the Xq cenhap region (with Linkage Disequilibrium heat map from (a) enlarged) and a collection of annotated genes (RefSeq, white), of which variation have been attributed to a human disease (OMIM data, grey).

In addition to studies that involve the cenhap associated regions, extensive sequence variation within the satellite array is expected to contribute to our understanding of centromere instability and disease. Cytogenetic staining has revealed the constitutive heterochromatin in human centromeric

regions has a highly heteromorphic structure. Given the critical importance of centromeres in ensuring proper chromosome segregation, such genomic variation is hypothesized to drive genome stability, and have been linked with human disease and cancers (reviewed [84]). Nevertheless, in the case of cancers where cells are expected to present increased genomic rearrangements, altered regulation and localization of kinetochore proteins, it will be important to estimate the rate of neocentromere formation with respect to native centromere sequence stability to test functionally relevant satellite DNA variants. We are only beginning to understand the sequence organization, allelic frequency, and evolution of satellite DNAs in the human population. Indeed, an analysis of optical genome maps of 154 individuals from 26 populations provided evidence for a large proportion of structural variants in satellite DNAs [85]. Such high-resolution diversity maps are expected to guide studies aimed to characterize satellite array structures that are associated with disease from those that have little functional consequence.

**Table 1.** Description of centromere-adjacent single nucleotide polymorphisms (SNPs) identified by published Genome-Wide Association Studies (GWAS), collected in the NHGRI-EBI GWAS Catalog published jointly by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI) [80]. SNPs are included if found within a two-megabase window of an alpha satellite reference model (GRCh38) and do not overlap with annotated genes or segmental duplication).

| Trait | SNPs | CEN adjacent (2Mb) Regions | Citation |
|---|---|---|---|
| Cancer | rs930395, rs2241024, rs142427110, rs35951924, rs199501877, rs11146838, rs6490525, rs2050203, rs7278690, rs35505947 | 4p12; 5p12; 5q11; 10p11; 13q12; 18p11; 19q11; 20p11; 21q11 | [86–89] |
| Cardiovascular disease | rs10132760, rs12186641, rs9367716, rs71566846, rs223290, rs144961578, rs3813127, rs1657346, rs1254531, rs10793514 | 5q11.2; 6p11.2; 6q11.1; 10q11.21; 14q11.2; 18q11.2 | [90–92] |
| Neurodegenerative diseases | rs11826064, rs13168838, rs62365447, rs140996952, rs1480597, rs10783624, rs7989524, rs6822736, rs13110633, rs2424635 | 4p11; 4q12; 5p12; 5q11.1; 6q11.1; 10q11; 11p11; 12q12; 13q12; 20p11 | [93–100] |
| Scoliosis/Bone Density (Spine) | rs8111296, rs11652527, rs1436931, rs6061081, rs17599071, rs10136383, rs9288898, rs10772040, rs4562194, rs810967, rs6050182, rs6511621, rs11229654, rs6551418, rs1006899 | 3p11.1; 3q11.2; 6q12; 7q11.21; 10q11.21; 11q11; 12p11.21; 14q11.2; 17q11.2; 19p12; 20p11.21; 21q11.2 | [101,102] |
| Digestive system disease | rs4243971, rs2342002, rs4800353, rs6058869, rs6087990 | 6q11.1; 18q11.2; 20q11.21 | [103,104] |

## 5. Concluding Remarks

In conclusion, human satellite DNAs provide a new, largely uncharted source of sequence variation in the human population. Chromosome-specific satellite arrays are expected to vary considerably in the human population, and measuring the overall range in the abundance and frequencies of repeat variants will contribute to ongoing studies of centromere biology and genome instability. Efforts to identify and study these variants will rely on improved, comprehensive genomic methods capable of mapping the full extent of satellite sequence heterogeneity that cannot be captured using a single reference genome. Such maps are necessary to direct future biomedical research to variants that are associated with disease, rather than natural sequence variation, which may have little or no clinical consequence.

**Funding:** The research was made possible by the generous financial support of the W.M. Keck Foundation.

**Acknowledgments:** I would like to thank Nicolas Altemose, Charles Langley, and Sasha Langley for valuable comments on the manuscript. The results presented herein were obtained at the Genomics Institute at the University of California, Santa Cruz.

**Conflicts of Interest:** The author declares no conflict of interest.

## References


1. 1000 Genomes Project Consortium; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74.
2. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [PubMed]
3. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; et al. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351. [CrossRef]
4. Miga, K.H. Completing the human genome: The progress and challenge of satellite DNA assembly. *Chromosome Res.* **2015**, *23*, 421–426. [CrossRef]
5. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338. [CrossRef]
6. Jain, M.; Olsen, H.E.; Turner, D.; Stoddart, D.; Paten, B.; Haussler, D.; Willard, H.F.; Akeson, M.; Miga, K.H. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **2018**, *36*, 321–323. [CrossRef] [PubMed]
7. Miga, K.H.; Newton, Y.; Jain, M.; Altemose, N.; Willard, H.F.; Kent, W.J. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **2014**, *24*, 697–707. [CrossRef] [PubMed]
8. Black, E.M.; Giunta, S. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes* **2018**, *9*, 615. [CrossRef] [PubMed]
9. Ferreira, D.; Meles, S.; Escudeiro, A.; Mendes-da-Silva, A.; Adega, F.; Chaves, R. Satellite non-coding RNAs: The emerging players in cells, cellular pathways and cancer. *Chromosome Res.* **2015**, *23*, 479–493. [CrossRef]
10. Enukashvily, N.I.; Donev, R.; Waisertreiger, I.S.-R.; Podgornaya, O.I. Human chromosome 1 satellite 3 DNA is decondensed, demethylated and transcribed in senescent cells and in A431 epithelial carcinoma cells. *Cytogenet. Genome Res.* **2007**, *118*, 42–54. [CrossRef] [PubMed]
11. Atkin, N.B.; Brito-Babapulle, V. Heterochromatin polymorphism and human cancer. *Cancer Genet. Cytogenet.* **1981**, *3*, 261–272. [CrossRef]
12. Berger, R.; Bernheim, A.; Kristoffersson, U.; Mitelman, F.; Olsson, H. C-band heteromorphism in breast cancer patients. *Cancer Genet. Cytogenet.* **1985**, *18*, 37–42. [CrossRef]
13. Sahin, F.I.; Yilmaz, Z.; Yuregir, O.O.; Bulakbasi, T.; Ozer, O.; Zeyneloglu, H.B. Chromosome heteromorphisms: An impact on infertility. *J. Assist. Reprod. Genet.* **2008**, *25*, 191–195. [CrossRef]
14. Ting, D.T.; Lipson, D.; Paul, S.; Brannigan, B.W.; Akhavanfard, S.; Coffman, E.J.; Contino, G.; Deshpande, V.; Iafrate, A.J.; Letovsky, S.; et al. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **2011**, *331*, 593–596. [CrossRef]
15. Atkin, N.B.; Brito-Babapulle, V. Chromosome 1 heterochromatin variants and cancer: A reassessment. *Cancer Genet. Cytogenet.* **1985**, *18*, 325–331. [CrossRef]
16. Wu, J.C.; Manuelidis, L. Sequence definition and organization of a human repeated DNA. *J. Mol. Biol.* **1980**, *142*, 363–386. [CrossRef]
17. Lee, C.; Wevrick, R.; Fisher, R.B.; Ferguson-Smith, M.A.; Lin, C.C. Human centromeric DNAs. *Hum. Genet.* **1997**, *100*, 291–304. [CrossRef]
18. Rudd, M.K.; Willard, H.F. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **2004**, *20*, 529–533. [CrossRef]
19. Eichler, E.E.; Clark, R.A.; She, X. An assessment of the sequence gaps: Unfinished business in a finished human genome. *Nat. Rev. Genet.* **2004**, *5*, 345–354. [CrossRef]
20. Willard, H.F. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **1985**, *37*, 524–532.
21. Waye, J.S.; Willard, H.F. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: A survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res.* **1987**, *15*, 7549–7569. [CrossRef]
22. Manuelidis, L.; Wu, J.C. Homology between human and simian repeated DNA. *Nature* **1978**, *276*, 92–94. [CrossRef]

23. Hayden, K.E.; Strome, E.D.; Merrett, S.L.; Lee, H.-R.; Rudd, M.K.; Willard, H.F. Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **2013**, *33*, 763–772. [CrossRef]
24. Wevrick, R.; Willard, H.F. 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* **1989**, *86*, 9394–9398. [CrossRef]
25. Mahtani, M.M.; Willard, H.F. Pulsed-field gel analysis of alpha-satellite DNA at the human X chromosome centromere: High-frequency polymorphisms and array size estimate. *Genomics* **1990**, *7*, 607–613. [CrossRef]
26. Marçais, B.; Bellis, M.; Gérard, A.; Pagès, M.; Boublik, Y.; Roizès, G. Structural organization and polymorphism of the alpha satellite DNA sequences of chromosomes 13 and 21 as revealed by pulse field gel electrophoresis. *Hum. Genet.* **1991**, *86*, 311–316. [CrossRef]
27. Jones, K.W.; Prosser, J.; Corneo, G.; Ginelli, E. The chromosomal location of human satellite DNA III. *Chromosoma* **1973**, *42*, 445–451. [CrossRef]
28. Jones, K.W.; Corneo, G. Location of satellite and homogeneous DNA sequences on human chromosomes. *Nat. New Biol.* **1971**, *233*, 268–271. [CrossRef]
29. Gosden, J.R.; Mitchell, A.R.; Buckland, R.A.; Clayton, R.P.; Evans, H.J. The location of four human satellite DNAs on human chromosomes. *Exp. Cell Res.* **1975**, *92*, 148–158. [CrossRef]
30. Tagarro, I.; Fernández-Peralta, A.M.; González-Aguilera, J.J. Chromosomal localization of human satellites 2 and 3 by a FISH method using oligonucleotides as probes. *Hum. Genet.* **1994**, *93*, 383–388. [CrossRef]
31. Altemose, N.; Miga, K.H.; Maggioni, M.; Willard, H.F. Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput. Biol.* **2014**, *10*, e1003628. [CrossRef]
32. Prosser, J.; Frommer, M.; Paul, C.; Vincent, P.C. Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* **1986**, *187*, 145–155. [CrossRef]
33. Cooke, H. Repeated sequence specific to human males. *Nature* **1976**, *262*, 182–186. [CrossRef]
34. Kunkel, L.M.; Smith, K.D.; Boyer, S.H.; Borgaonkar, D.S.; Wachtel, S.S.; Miller, O.J.; Breg, W.R.; Jones, H.W., Jr.; Rary, J.M. Analysis of human Y-chromosome-specific reiterated DNA in chromosome variants. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 1245–1249. [CrossRef]
35. Nakahori, Y.; Mitani, K.; Yamada, M.; Nakagome, Y. A human Y-chromosome specific repeated DNA family (DYZ1) consists of a tandem array of pentanucleotides. *Nucleic Acids Res.* **1986**, *14*, 7569–7580. [CrossRef]
36. Willard, H.F.; Waye, J.S. 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.* **1987**, *25*, 207–214. [CrossRef]
37. Alexandrov, I.; Kazakov, A.; Tumeneva, I.; Shepelev, V.; Yurov, Y. Alpha-satellite DNA of primates: Old and new families. *Chromosoma* **2001**, *110*, 253–266. [CrossRef]
38. Jeanpierre, M.; Weil, D.; Gallano, P.; Creau-Goldberg, N.; Junien, C. The organization of two related subfamilies of a human tandemly repeated DNA is chromosome specific. *Hum. Genet.* **1985**, *70*, 302–310. [CrossRef]
39. Sherman, R.M.; Forman, J.; Antonescu, V.; Puiu, D.; Daya, M.; Rafaels, N.; Boorgula, M.P.; Chavan, S.; Vergara, C.; Ortega, V.E.; et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **2019**, *51*, 30. [CrossRef]
40. Schneider, V.A.; Graves-Lindsay, T.; Howe, K.; Bouk, N.; Chen, H.-C.; Kitts, P.A.; Murphy, T.D.; Pruitt, K.D.; Thibaud-Nissen, F.; Albracht, D.; et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **2017**, *27*, 849–864. [CrossRef]
41. Miga, K.H.; Eisenhart, C.; Kent, W.J. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res.* **2015**, *43*, e133. [CrossRef]
42. Nechemia-Arbely, Y.; Fachinetti, D.; Miga, K.H.; Sekulic, N.; Soni, G.V.; Kim, D.H.; Wong, A.K.; Lee, A.Y.; Nguyen, K.; Dekker, C.; et al. Human centromeric CENP-A chromatin is a homotypic, octameric nucleosome at all cell cycle points. *J. Cell Biol.* **2017**, *216*, 607–621. [CrossRef]
43. Levy, S.; Sutton, G.; Ng, P.C.; Feuk, L.; Halpern, A.L.; Walenz, B.P.; Axelrod, N.; Huang, J.; Kirkness, E.F.; Denisov, G.; et al. The diploid genome sequence of an individual human. *PLoS Biol.* **2007**, *5*, e254. [CrossRef]
44. Audano, P.A.; Sulovari, A.; Graves-Lindsay, T.A.; Cantsilieris, S.; Sorensen, M.; Welch, A.E.; Dougherty, M.L.; Nelson, B.J.; Shah, A.; Dutcher, S.K.; et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **2019**, *176*, 663–675.e19. [CrossRef]

45. Chaisson, M.J.P.; Wilson, R.K.; Eichler, E.E. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **2015**, *16*, 627–640. [CrossRef] [PubMed]
46. Marçais, B.; Charlieu, J.P.; Allain, B.; Brun, E.; Bellis, M.; Roizès, G. On the mode of evolution of alpha satellite DNA in human populations. *J. Mol. Evol.* **1991**, *33*, 42–48. [CrossRef] [PubMed]
47. Smith, G.P. Evolution of repeated DNA sequences by unequal crossover. *Science* **1976**, *191*, 528–535. [CrossRef]
48. Lower, S.S.; McGurk, M.P.; Clark, A.G.; Barbash, D.A. Satellite DNA evolution: old ideas, new approaches. *Curr. Opin. Genet. Dev.* **2018**, *49*, 70–78. [CrossRef] [PubMed]
49. Stults, D.M.; Killen, M.W.; Pierce, H.H.; Pierce, A.J. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* **2008**, *18*, 13–18. [CrossRef] [PubMed]
50. Kim, J.-H.; Dilthey, A.T.; Nagaraja, R.; Lee, H.-S.; Koren, S.; Dudekula, D.; Wood, W.H., III; Piao, Y.; Ogurtsov, A.Y.; Utani, K.; et al. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res.* **2018**, *46*, 6712–6725. [CrossRef]
51. Warburton, P.E.; Willard, H.F. Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: Evidence for concerted evolution along haplotypic lineages. *J. Mol. Evol.* **1995**, *41*, 1006–1015. [CrossRef]
52. Willard, H.F.; Waye, J.S. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **1987**, *3*, 192–198. [CrossRef]
53. Hayden, K.E. Human centromere genomics: Now it's personal. *Chromosome Res.* **2012**, *20*, 621–633. [CrossRef] [PubMed]
54. Pluta, A.F.; Saitoh, N.; Goldberg, I.; Earnshaw, W.C. Identification of a subdomain of CENP-B that is necessary and sufficient for localization to the human centromere. *J. Cell Biol.* **1992**, *116*, 1081–1093. [CrossRef]
55. Hudson, D.F.; Fowler, K.J.; Earle, E.; Saffery, R.; Kalitsis, P.; Trowell, H.; Hill, J.; Wreford, N.G.; de Kretser, D.M.; Cancilla, M.R.; et al. Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. *J. Cell Biol.* **1998**, *141*, 309–319. [CrossRef]
56. Warburton, P.E.; Waye, J.S.; Willard, H.F. Nonrandom localization of recombination events in human alpha satellite repeat unit variants: Implications for higher-order structural characteristics within centromeric heterochromatin. *Mol. Cell. Biol.* **1993**, *13*, 6520–6529. [CrossRef] [PubMed]
57. Fachinetti, D.; Han, J.S.; McMahon, M.A.; Ly, P.; Abdullah, A.; Wong, A.J.; Cleveland, D.W. DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function. *Dev. Cell* **2015**, *33*, 314–327. [CrossRef] [PubMed]
58. Waye, J.S.; Willard, H.F. Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: Evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. *Mol. Cell. Biol.* **1986**, *6*, 3156–3165. [CrossRef] [PubMed]
59. Maloney, K.A.; Sullivan, L.L.; Matheny, J.E.; Strome, E.D.; Merrett, S.L.; Ferris, A.; Sullivan, B.A. Functional epialleles at an endogenous human centromere. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 13704–13709. [CrossRef]
60. Aldrup-MacDonald, M.E.; Kuo, M.E.; Sullivan, L.L.; Chew, K.; Sullivan, B.A. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res.* **2016**, *26*, 1301–1311. [CrossRef]
61. McNulty, S.M.; Sullivan, L.L.; Sullivan, B.A. Human Centromeres Produce Chromosome-Specific and Array-Specific Alpha Satellite Transcripts that Are Complexed with CENP-A and CENP-C. *Dev. Cell* **2017**, *42*, 226–240.e6. [CrossRef]
62. Hall, L.L.; Byron, M.; Carone, D.M.; Whitfield, T.W.; Pouliot, G.P.; Fischer, A.; Jones, P.; Lawrence, J.B. Demethylated HSATII DNA and HSATII RNA Foci Sequester PRC1 and MeCP2 into Cancer-Specific Nuclear Bodies. *Cell Rep.* **2017**, *18*, 2943–2956. [CrossRef]
63. Cobb, B.S.; Morales-Alcelay, S.; Kleiger, G.; Brown, K.E.; Fisher, A.G.; Smale, S.T. Targeting of Ikaros to pericentromeric heterochromatin by direct DNA binding. *Genes Dev.* **2000**, *14*, 2146–2160. [CrossRef]
64. Nishibuchi, G.; Déjardin, J. The molecular basis of the organization of repetitive DNA-containing constitutive heterochromatin in mammals. *Chromosome Res.* **2017**, *25*, 77–87. [CrossRef]
65. Delpu, Y.; McNamara, T.F.; Griffin, P.; Kaleem, S.; Narayan, S.; Schildkraut, C.; Miga, K.H.; Tahiliani, M. Chromosomal rearrangements at hypomethylated Satellite 2 sequences are associated with impaired replication efficiency and increased fork stalling. *bioRxiv* **2019**. [CrossRef]

66. Erliandri, I.; Fu, H.; Nakano, M.; Kim, J.-H.; Miga, K.H.; Liskovykh, M.; Earnshaw, W.C.; Masumoto, H.; Kouprina, N.; Aladjem, M.I.; et al. Replication of alpha-satellite DNA arrays in endogenous human centromeric regions and in human artificial chromosome. *Nucleic Acids Res.* **2014**, *42*, 11502–11516. [CrossRef]
67. Bersani, F.; Lee, E.; Kharchenko, P.V.; Xu, A.W.; Liu, M.; Xega, K.; MacKenzie, O.C.; Brannigan, B.W.; Wittner, B.S.; Jung, H.; et al. Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 15148–15153. [CrossRef]
68. Sevim, V.; Bashir, A.; Chin, C.-S.; Miga, K.H. Alpha-CENTAURI: Assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics* **2016**, *32*, 1921–1924. [CrossRef]
69. Pathak, D.; Premi, S.; Srivastava, J.; Chandy, S.P.; Ali, S. Genomic instability of the DYZ1 repeat in patients with Y chromosome anomalies and males exposed to natural background radiation. *DNA Res.* **2006**, *13*, 103–109. [CrossRef]
70. Rahman, M.M.; Bashamboo, A.; Prasad, A.; Pathak, D.; Ali, S. Organizational variation of DYZ1 repeat sequences on the human Y chromosome and its diagnostic potentials. *DNA Cell Biol.* **2004**, *23*, 561–571. [CrossRef]
71. Oakey, R.; Tyler-Smith, C. Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. *Genomics* **1990**, *7*, 325–330. [CrossRef]
72. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
73. Schueler, M.G.; Higgins, A.W.; Rudd, M.K.; Gustashaw, K.; Willard, H.F. Genomic and genetic definition of a functional human centromere. *Science* **2001**, *294*, 109–115. [CrossRef]
74. Langley, S.A.; Miga, K.; Karpen, G.H.; Langley, C.H. Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *BioRxiv* **2018**. [CrossRef]
75. She, X.; Horvath, J.E.; Jiang, Z.; Liu, G.; Furey, T.S.; Christ, L.; Clark, R.; Graves, T.; Gulden, C.L.; Alkan, C.; et al. The structure and evolution of centromeric transition regions within the human genome. *Nature* **2004**, *430*, 857–864. [CrossRef]
76. Pruitt, K.D.; Tatusova, T.; Maglott, D.R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2005**, *33*, D501–D504. [CrossRef]
77. Amberger, J.; Bocchini, C.A.; Scott, A.F.; Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res.* **2009**, *37*, D793–D796. [CrossRef]
78. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **2005**, *33*, D514–D517. [CrossRef]
79. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Hoover, J.; et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **2016**, *44*, D862–D868. [CrossRef]
80. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [CrossRef]
81. Loh, P.-R.; Genovese, G.; Handsaker, R.E.; Finucane, H.K.; Reshef, Y.A.; Palamara, P.F.; Birmann, B.M.; Talkowski, M.E.; Bakhoum, S.F.; McCarroll, S.A.; et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **2018**, *559*, 350–355. [CrossRef]
82. Reich, D.; Patterson, N.; De Jager, P.L.; McDonald, G.J.; Waliszewska, A.; Tandon, A.; Lincoln, R.R.; DeLoa, C.; Fruhan, S.A.; Cabre, P.; et al. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat. Genet.* **2005**, *37*, 1113–1118. [CrossRef] [PubMed]
83. Karolchik, D.; Hinrichs, A.S.; Furey, T.S.; Roskin, K.M.; Sugnet, C.W.; Haussler, D.; Kent, W.J. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **2004**, *32*, D493–D496. [CrossRef]
84. Barra, V.; Fachinetti, D. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat. Commun.* **2018**, *9*, 4340. [CrossRef]
85. Levy-Sakin, M.; Pastor, S.; Mostovoy, Y.; Li, L.; Leung, A.K.Y.; McCaffrey, J.; Young, E.; Lam, E.T.; Hastie, A.R.; Wong, K.H.Y.; et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **2019**, *10*, 1025. [CrossRef]

86. Michailidou, K.; Lindström, S.; Dennis, J.; Beesley, J.; Hui, S.; Kar, S.; Lemaçon, A.; Soucy, P.; Glubb, D.; Rostamianfar, A.; et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **2017**, *551*, 92–94. [CrossRef]
87. O'Donnell, P.H.; Stark, A.L.; Gamazon, E.R.; Wheeler, H.E.; McIlwee, B.E.; Gorsic, L.; Im, H.K.; Huang, R.S.; Cox, N.J.; Dolan, M.E. Identification of novel germline polymorphisms governing capecitabine sensitivity. *Cancer* **2012**, *118*, 4063–4073. [CrossRef] [PubMed]
88. Moore, K.N.; Tritchler, D.; Kaufman, K.M.; Lankes, H.; Quinn, M.C.J.; Ovarian Cancer Association Consortium; Van Le, L.; Berchuck, A.; Backes, F.J.; Tewari, K.S.; et al. Genome-wide association study evaluating single-nucleotide polymorphisms and outcomes in patients with advanced stage serous ovarian or primary peritoneal cancer: An NRG Oncology/Gynecologic Oncology Group study. *Gynecol. Oncol.* **2017**, *147*, 396–401. [CrossRef] [PubMed]
89. Hofer, P.; Hagmann, M.; Brezina, S.; Dolejsi, E.; Mach, K.; Leeb, G.; Baierl, A.; Buch, S.; Sutterlüty-Fall, H.; Karner-Hanusch, J.; et al. Bayesian and frequentist analysis of an Austrian genome-wide association study of colorectal cancer and advanced adenomas. *Oncotarget* **2017**, *8*, 98623–98634. [CrossRef]
90. Deng, X.; Sabino, E.C.; Cunha-Neto, E.; Ribeiro, A.L.; Ianni, B.; Mady, C.; Busch, M.P.; Seielstad, M. REDSII Chagas Study Group from the NHLBI Retrovirus Epidemiology Donor Study-II Component International Genome wide association study (GWAS) of Chagas cardiomyopathy in Trypanosoma cruzi seropositive subjects. *PLoS ONE* **2013**, *8*, e79629. [CrossRef] [PubMed]
91. Cordell, H.J.; Bentham, J.; Topf, A.; Zelenika, D.; Heath, S.; Mamasoula, C.; Cosgrove, C.; Blue, G.; Granados-Riveron, J.; Setchfield, K.; et al. Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat. Genet.* **2013**, *45*, 822–824. [CrossRef]
92. van der Harst, P.; Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **2018**, *122*, 433–443. [CrossRef]
93. Nagel, M.; Jansen, P.R.; Stringer, S.; Watanabe, K.; de Leeuw, C.A.; Bryois, J.; Savage, J.E.; Hammerschlag, A.R.; Skene, N.G.; Muñoz-Manchado, A.B.; et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **2018**, *50*, 920–927. [CrossRef]
94. Turley, P.; Walters, R.K.; Maghzian, O.; Okbay, A.; Lee, J.J.; Fontana, M.A.; Nguyen-Viet, T.A.; Wedow, R.; Zacher, M.; Furlotte, N.A.; et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **2018**, *50*, 229–237. [CrossRef]
95. Herold, C.; Hooli, B.V.; Mullin, K.; Liu, T.; Roehr, J.T.; Mattheisen, M.; Parrado, A.R.; Bertram, L.; Lange, C.; Tanzi, R.E. Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. *Mol. Psychiatry* **2016**, *21*, 1608–1612. [CrossRef]
96. Fung, H.-C.; Scholz, S.; Matarin, M.; Simón-Sánchez, J.; Hernandez, D.; Britton, A.; Gibbs, J.R.; Langefeld, C.; Stiegert, M.L.; Schymick, J.; et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: First stage analysis and public release of data. *Lancet Neurol.* **2006**, *5*, 911–916. [CrossRef]
97. Goes, F.S.; McGrath, J.; Avramopoulos, D.; Wolyniec, P.; Pirooznia, M.; Ruczinski, I.; Nestadt, G.; Kenny, E.E.; Vacic, V.; Peters, I.; et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **2015**, *168*, 649–659. [CrossRef]
98. Li, Z.; Chen, J.; Yu, H.; He, L.; Xu, Y.; Zhang, D.; Yi, Q.; Li, C.; Li, X.; Shen, J.; et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **2017**, *49*, 1576–1583. [CrossRef]
99. Beecham, G.W.; Hamilton, K.; Naj, A.C.; Martin, E.R.; Huentelman, M.; Myers, A.J.; Corneveaux, J.J.; Hardy, J.; Vonsattel, J.-P.; Younkin, S.G.; et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. *PLoS Genet.* **2014**, *10*, e1004606. [CrossRef]
100. Wang, K.-S.; Liu, X.-F.; Aragam, N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophr. Res.* **2010**, *124*, 192–199. [CrossRef]
101. Styrkarsdottir, U.; Halldorsson, B.V.; Gretarsdottir, S.; Gudbjartsson, D.F.; Walters, G.B.; Ingvarsson, T.; Jonsdottir, T.; Saemundsdottir, J.; Snorradóttir, S.; Center, J.R.; et al. New sequence variants associated with bone mineral density. *Nat. Genet.* **2009**, *41*, 15–17. [CrossRef] [PubMed]

102. Liu, J.; Zhou, Y.; Liu, S.; Song, X.; Yang, X.-Z.; Fan, Y.; Chen, W.; Akdemir, Z.C.; Yan, Z.; Zuo, Y.; et al. The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Hum. Genet.* **2018**, *137*, 553–567. [CrossRef] [PubMed]
103. Liu, J.Z.; van Sommeren, S.; Huang, H.; Ng, S.C.; Alberts, R.; Takahashi, A.; Ripke, S.; Lee, J.C.; Jostins, L.; Shah, T.; et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **2015**, *47*, 979–986. [CrossRef] [PubMed]
104. Levine, D.M.; Ek, W.E.; Zhang, R.; Liu, X.; Onstad, L.; Sather, C.; Lao-Sirieix, P.; Gammon, M.D.; Corley, D.A.; Shaheen, N.J.; et al. A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett's esophagus. *Nat. Genet.* **2013**, *45*, 1487–1493. [CrossRef] [PubMed]