RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

Epigenetic patterns in a complete human genome

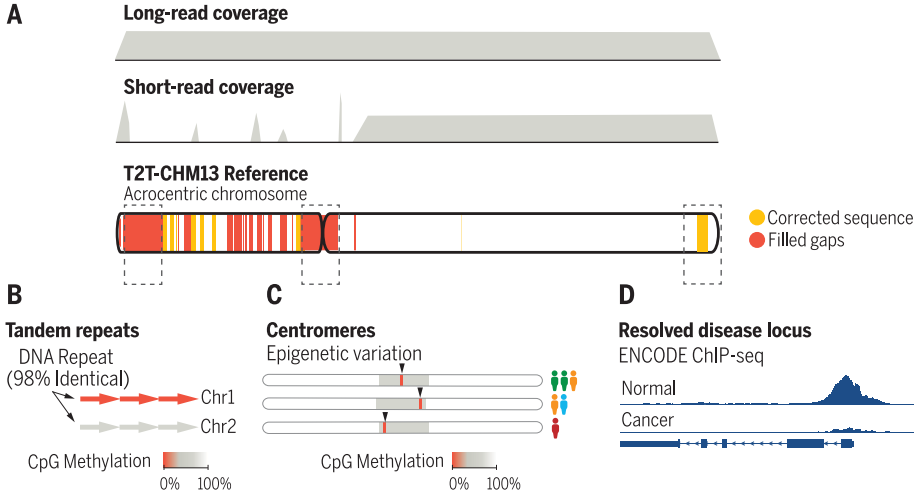
Ariel Gershman, Michael E. G. Sauria, Xavi Guitart, Mitchell R. Vollger, Paul W. Hook, Savannah J. Hoyt, Miten Jain, Alaina Shumate, Roham Razaghi, Sergey Koren, Nicolas Altemose, Gina V. Caldas, Glennis A. Logsdon, Arang Rhie, Evan E. Eichler, Michael C. Schatz, Rachel J. O'Neill, Adam M. Phillippy, Karen H. Miga*, Winston Timp*

INTRODUCTION: The human reference genome has served as the foundation for many large-scale initiatives, including the collective effort to catalog the epigenome, the set of marks and protein interactions that act to control gene activity and cellular function. However, for more than two decades, efforts to construct a complete epigenome have been hampered by an incomplete reference genome. With recent technological advances, we can now study genome structure and function comprehensively across a complete telomere-to-telomere human genome assembly, T2T-CHM13. As a result, we can now broaden the human epigenome to include 225 million base pairs (Mbp) of additional sequence.

RATIONALE: The epigenome refers to DNA modifications (e.g., CpG methylation), protein-DNA interactions, histone modifications, and chromatin organization that collectively influence gene expression, genome regulation, and genome stability. These epigenetic features are heritable upon cell division but dynamic during development, generating profiles that are unique to different tissues and cell types. Here, we present an epigenetic annotation of the human genome in which we explore previously unresolved regions, including acrocentric chromosome short arms, segmentally duplicated genes, and a diverse collection of repeat classes, including human centromeres. Generating a complete epigenetic annotation of the previously missing 8% of the human genome provides a foundation for elucidating the functional roles of these genomic elements that are critical to our understanding of genome regulation, function, and evolution.

RESULTS: Completion of the human epigenome required that we develop approaches to profiling the previously unresolved regions. Using the T2T-CHM13 reference with existing short-read epigenetic data, we identified 3 to 19% more enrichment sites for epigenetic markers. However, even with the complete reference, these short-read epigenetic methods cannot correctly resolve regions of the genome of high similarity, including segmental duplications, gene paralogs, or large repeat arrays. On the other hand, long-read epigenetic methods can resolve single-molecule epigenetic patterns within these regions by anchoring to flanking or infrequent unique regions, providing a foundational assessment of these areas. Long-read methylation calls using the T2T-CHM13 assembly increased the number of probeable CpG sites by 10% (3.2 M), revealing epigenetic patterning of genomic regions that were previously intractable. We generated long-read methylomes of distinct developmental time points and surveyed >99% of the genome's CpGs. We probed highly homologous gene families and observed paralog-specific differences in regulation between disease and nondisease states. In tandem repeats, we identified differences in epigenetic regulation between genetically identical sequences present across different genomic locations, observing locus- and single-molecule-level differences in methylation. Our analysis revealed that these regions vary in epigenetic and transcriptional activity despite high sequence identity, highlighting the importance of the local chromosome environment as a modulator of epigenetics. Finally, the T2T-CHM13 genome assembly has opened exploration of the human centromere, enabling us to probe the epigenetic elements that define centromeric chromatin. The centromere is the site of assembly of the kinetochore complex, an essential complex for eukaryotic cell division. We generated complete epigenetic maps of human centromeres, revealing epigenetic markers of centromere activity that denote active human kinetochores. We predicted kinetochore site localization within active centromeres and report variability of kinetochore localization across individuals representing diverse ancestry.

CONCLUSION: The improvements in epigenetic profiling using T2T-CHM13 set the foundation for complete assemblies and long-read epigenetics for major biological advancements. Using technological advances in genome resequencing and alignment, we present a comprehensive functional assessment of previously unresolved genomic regions. This study marks the start of exploration into duplicated and repetitive portions of the epigenome, pioneering the exploration of epigenetics in a complete human genome. ■



Epigenetic characterization across a complete human genome. (**A**) The T2T-CHM13 reference contains filled gaps and corrected sequences. Using short- and long-read sequencing data, we functionally annotated these added regions. (**B**) Tandem repeats, which are nearly identical, vary in epigenetic state depending on genomic location. (**C**) The epigenetic basis of centromere identity is variable among diverse individuals. (**D**) In genes associated with disease, short reads mapped to T2T-CHM13 elucidate epigenetic dysregulation in human disease states.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: khmiga@ucsc.edu (K.H.M.); wtimp@jhu.edu (W.T.)
Cite this article as A. Gershman *et al.*, *Science* **376**, eabj5089 (2022). DOI: 10.1126/science.abj5089

RESEARCH ARTICLE

HUMAN GENOMICS

Epigenetic patterns in a complete human genome

Ariel Gershman[1], Michael E. G. Sauria[2], Xavi Guitart[3], Mitchell R. Vollger[3], Paul W. Hook[4], Savannah J. Hoyt[5,6], Miten Jain[7], Alaina Shumate[4], Roham Razaghi[4], Sergey Koren[8], Nicolas Altemose[9], Gina V. Caldas[10], Glennis A. Logsdon[3], Arang Rhie[8], Evan E. Eichler[3,11], Michael C. Schatz[2], Rachel J. O'Neill[5,6], Adam M. Phillippy[8], Karen H. Miga[7]*, Winston Timp[1,4]*

The completion of a telomere-to-telomere human reference genome, T2T-CHM13, has resolved complex regions of the genome, including repetitive and homologous regions. Here, we present a high-resolution epigenetic study of previously unresolved sequences, representing entire acrocentric chromosome short arms, gene family expansions, and a diverse collection of repeat classes. This resource precisely maps CpG methylation (32.28 million CpGs), DNA accessibility, and short-read datasets (166,058 previously unresolved chromatin immunoprecipitation sequencing peaks) to provide evidence of activity across previously unidentified or corrected genes and reveals clinically relevant paralog-specific regulation. Probing CpG methylation across human centromeres from six diverse individuals generated an estimate of variability in kinetochore localization. This analysis provides a framework with which to investigate the most elusive regions of the human genome, granting insights into epigenetic regulation.

The human reference genome has served as the foundation for many large-scale epigenetic initiatives (*1*–*3*) that aimed to catalog regulatory elements involved in gene activity and cellular function. However, efforts to construct a complete annotation of functional elements have been hampered by an incomplete reference genome. With recent technological advances, we are now able to study genome structure and function comprehensively across the finished, telomere-to-telomere human genome assembly, T2T-CHM13, which is based on the CHM13 cell line derived from a complete hydatidiform mole (*4*). As a result, we can now broaden the human epigenome to include 225 million base pairs (Mbp) of sequence, representing entire acrocentric chromosome short arms, gene family expansions, and a diverse collection of repeat classes.

The epigenome is influenced both by the specific genetic sequence and the sequence context, i.e., the flanking regions and placement of the loci within the complex structure and organization within the nucleus (*5*). The same genetic sequence can perform different functions or be regulated differently depending on the location of the sequence and its epigenetic state. This is especially relevant given possible evolutionary advantages that may be conferred by gene duplication, such as selectively silencing or activating different paralogous gene copies. These processes are hypothesized to diversify gene activity across developmental time and different tissues (*6*). Beyond evolutionary questions, epigenetic dysregulation of repetitive sequences can play a key role in development and human disease. A diverse set of repeat sequences that are difficult to probe in the human reference genome GRCh38 have been implicated in facioscapulohumeral muscular dystrophy (FSHD) (associated with deletions in *D4Z4*) (*7*); schizophrenia (associated with an expanded repeat in *TAF1I*) (*8*); neuroblastoma (associated with somatic hypomethylation of *SST1*) (*9*); lung cancer (associated with *CT47* expression) (*10*); pancreatic ductal adenocarcinomas (associated with HSat2 expression) (*11*); and immunodeficiency, centromeric region instability, and facial anomalies syndrome (ICF) (associated with heterochromatin abnormalities in HSat2,3) (*12*).

Within the improved T2T-CHM13 reference, the previously unresolved areas are highly repetitive, containing only infrequent sites of unique, mappable regions. This presents a limitation to short-read sequence mapping strategies, even with a more accurate reference and unique k-mer anchored alignments (*13*, *14*). Emerging long-read technologies (*15*) offer sequence lengths capable of spanning infrequent unique markers and provide a direct measurement of the base sequence and epigenetic state on single molecules (*16*, *17*).

RESULTS

Epigenetic profiles from a T2T genome in disease-relevant loci

The T2T-CHM13 assembly resolves gaps and corrects misassembled or patched regions in GRCh38, leading to the introduction of nearly 225 Mbp (*4*). Using existing short-read epigenetic data from the ENCODE project (*1*), we probed previously unidentified areas of the genome. To ensure accurate mapping to these regions, we intersected ENCODE chromatin immunoprecipitation–sequencing (ChIP-seq) alignments with unique k-mers of varying size of k (range, k = 50 to 100; fig. S1 and tables S1 and S2) (*1*, *18*). On average, 2.35% more reads mapped to T2T-CHM13 than GRCh38 across six different histone marks and CTCF, an important regulator of chromatin architecture (fig. S2). Reads filtered out of GRCh38 due to non-unique mapping were largely confined to the satellite DNA and segmental duplications (SDs) (fig. S3). Although the total number of peaks called per sample was variable because of differences in cell type, all samples had an increase in the number of peaks called when comparing T2T-CHM13 with GRCh38 (Fig. 1A). As expected, we saw the most substantial increase in H3K9me3 (19.4%) and H3K27me3 (15.2%) enrichment compared with GRCh38 (Table 1), consistent with the introduced peri/centromeric satellites (CenSat), SDs, and other repetitive sequences in T2T-CHM13 (Fig. 1A) that are associated with constitutive heterochromatin (*19*). The number of called peaks in activating marks increased as well; most notably, there was a 4.9% increase in H3K36me3, a mark present across active gene bodies. Previously unresolved activating histone peaks (H3K27ac, H3K4me1, H3K36me3, and H3K4me3) and CTCF were primarily enriched in unique genic regions and in SDs (Fig. 1A).

T2T-CHM13 increased the number of annotated genes by 5.7% (*4*), revealing 2680 genes exclusive to T2T-CHM13 with no assigned ortholog in GRCh38 (*18*). These gene predictions require detailed study for functionality and validation. Here, we generated a functional annotation of the previously unresolved genes using activating peaks (H3K4me3 or H3K27ac) from ENCODE cell lines. We annotated activating peaks from at least two ENCODE cell lines at the transcriptional start site at 57 of these previously unresolved genes (table S3). Of these loci, most (n = 20) were long noncoding RNAs (lncRNAs), including *LINC01666*, which is known for its associations with gastric cancer (*20*). Many (n = 19) were pseudogenes, including FSHD region gene 1 (*FRG1*), which is a poorly understood candidate gene for FSHD

[1]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Biology and Computer Science, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 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]UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [8]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [9]Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA. [10]Department of Molecular and Cell Biology, University of California Berkeley, Berkeley CA, USA. [11]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.
*Corresponding author. Email: khmiga@ucsc.edu (K.H.M.); wtimp@jhu.edu (W.T.)

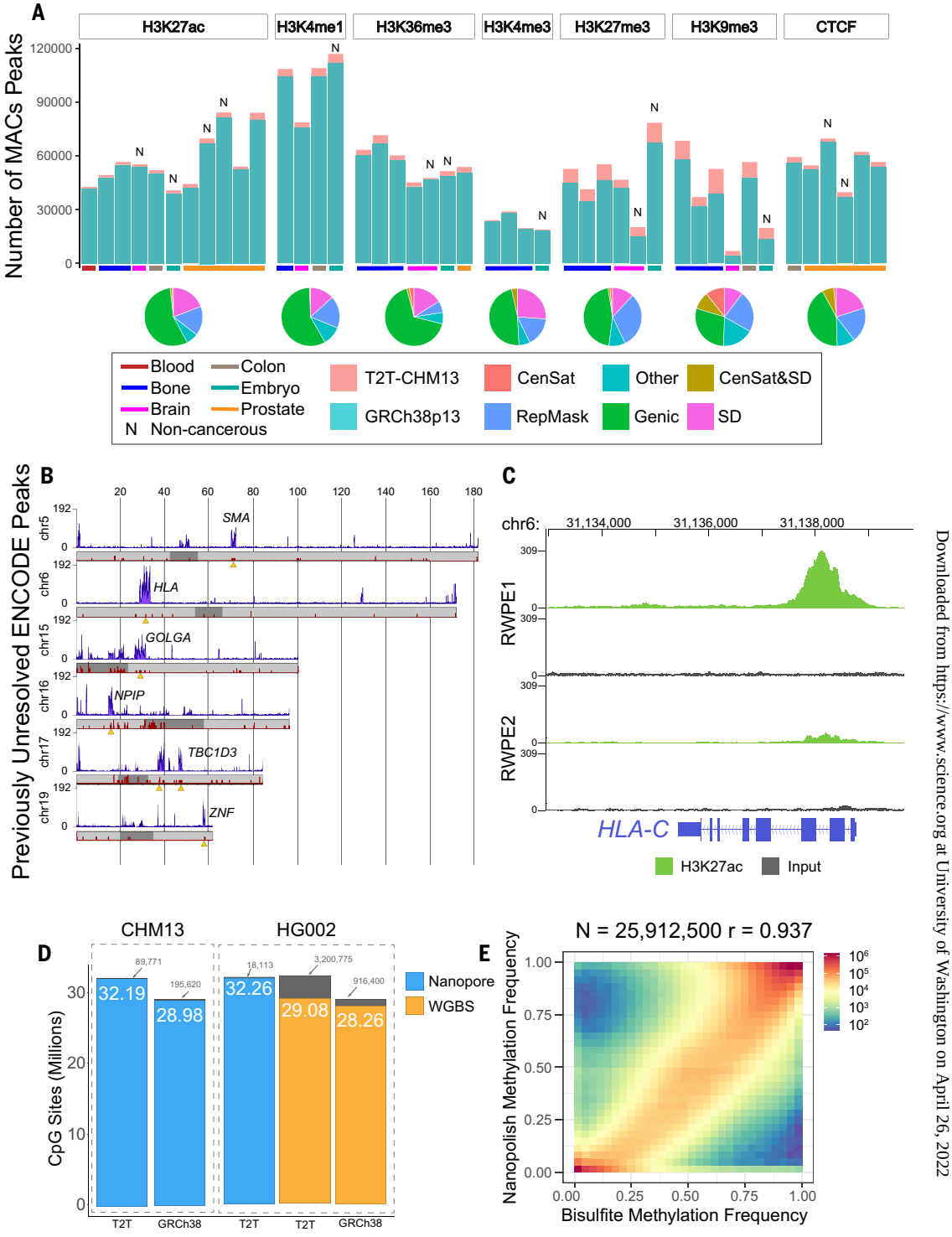


Fig. 1. Epigenetics in previously unresolved genome regions. (**A**) Top: bar plots of the number of peaks called per ENCODE sample using dynamic k-mer mapping to GRCh38 (blue) or T2T-CHM13 (salmon). Bottom: pie charts indicating the genomic localization of peaks found only in T2T-CHM13. (**B**) Number of T2T-CHM13 unique ENCODE peaks across chromosomes 5, 6, 15, 16, 17, and 19 in 50-kb bins (purple). Chromosome ideograms show the density of previously unannotated genes (red) with the centromere annotated as dark gray. Orange triangles denote regions of interest with a high density of previously uncalled peaks. (**C**) ENCODE ChIP-seq read coverage at the HLA-C gene locus on chromosome 6. (**D**) Number of CpGs with methylation profiled comparing sequencing method and reference assembly. (**E**) Correlation of HG002 WGBS and Nanopolish methylation calls aligned to T2T-CHM13.

(*21*). Three were protein-coding genes, including *BOLBA2B*, one of the most common genes associated with autism (*22*).

Our analysis of previously unresolved ENCODE peaks revealed enrichment of peaks for high-copy-number gene families (e.g., *GOLGA*, *NPIP*, *ZNF*, and *TBC1D3*) (Fig. 1B). Large structural variants resolved in T2T-CHM13 explain the additional ChIP-seq mapping events (fig. S4). Epigenetic annotation at these genetic loci may lead to insights of paralog-specific function in evolution (e.g., human-specific neural genes) and disease (*23*, *24*). For instance, *SMN1/2* is associated with spinal muscular atrophy (SMA) and was historically one of the most difficult regions to assemble (*25*). At the *SMN2* gene, we observed peaks of the activating H3K4me3 mark at the promoter in all four ENCODE cell lines analyzed (fig. S5), indicating high transcriptional activity of the gene across tissues. SMA is a leading cause of childhood death (*26*) and has the potential to be treated by regulating expression through histone deacetylase inhibitors, but understanding the disease-specific epigenetic differences between paralogs has been challenging (*27*).

Table 1. Peaks called using ENCODE datasets. Shown is a summary of ENCODE peak analysis including the mark profiled, summed peak calls per mark across all datasets, difference in peak number between references, and number of datasets.

Mark	Peaks called in GRCh38	Peaks called in CHM13	Difference in no. of peaks	Increase in peaks	No. of datasets
H3K9me3	194,681	241,497	46,816	19.4%	6
H3K27me3	249,945	294,819	44,874	15.2%	6
H3K36me3	373,933	393,224	19,291	4.9%	7
CTCF	327,713	342,284	14,571	4.3%	6
H3K4me1	396,332	412,907	16,575	4.0%	4
H3K27ac	611,645	632,837	21,192	3.3%	11
H3K4me3	88,985	91,724	2739	3.0%	4

Another previously intractable region of the genome, the HLA locus, is critical for understanding a wide range of biology from immunity to neuropsychiatric disorders (*28*, *29*). Our results reveal enrichment of ENCODE peaks across a variety of histone marks at the HLA locus (Fig. 1B and fig. S6A). Decreasing expression of HLA genes is associated with soft tissue cancers, particularly prostate cancer, and can even be indicative of chemotherapy resistance (*30*). Comparing non-neoplastic adult human prostate epithelial cells (RWPE-1) and the c-Ki-ras–transformed prostate cancer model cells from the same donor (RWPE-2) (*31*), we observed a decline in H3K27ac, an activating mark, at HLA gene promoters, concomitant with an increase in CTCF binding in RWPE-2 (Fig. 1C and fig. S6B). The differences in histone marks in this region indicate epigenetic dysregulation of the HLA locus in prostate cancer that may warrant further studies and inform upon potential therapies (*32*).

Long-read sequencing to derive complete human methylomes

Methylation profiling has traditionally faced difficulties in mapping success rates to repetitive regions of the genome, and such mapping inefficiencies are exaggerated by the bisulfite conversion of unmethylated cytosine to uracil, sequenced as thymine (*33*). Methylation profiles in T2T-CHM13 using long-read nanopore data demonstrate an increase in the genome coverage (32.8 M compared with 29.17 M in GRCh38, omitting chromosome Y) and surveyed more CpGs (10%, 3.18 M) compared with short-read whole-genome bisulfite sequencing (WGBS) (Fig. 1D). We called nanopore methylation data with Nanopolish (*34*), finding a high correlation (R = 0.937) both to WGBS results in regions mappable by both data types (Fig. 1E) and to the alternative nanopore methylation caller, Megalodon (R = 0.952) (fig. S7). Examining the difference between mapping of WGBS and nanopore methylation data, we generated short-read mappability scores in 200-bp windows, with a score of 0 being unmappable and 200 being highly mappable (*18*). We found that the 165 Mbp of sequence with a score of 0 (highly unmappable) is enriched in SDs and satellite DNA. Stratifying the nanopore data by read length, we found reads longer than 50 kilo–base pairs (kb) were capable of accurately determining methylation in these regions (figs. S8 and S9).

We sequenced the CHM13 cell line, representing an early developmental state, and HG002, a terminally differentiated lymphoblast cell line. The sequenced cell line CHM13 and HG002 nanopore datasets surveyed 32.19 M CpGs (99.7% of total) and 32.26 M CpGs (99.9% of total). As expected for differentiated cell lines, most of the HG002 genome is methylated (75% median methylation), with a secondary peak of unmethylated CpGs largely reflecting unmethylated CpG islands (CGIs) (fig. S10). By contrast, CHM13 is markedly hypomethylated (36.8% median methylation), as expected from a trophoblastic cell line (*35*). Comparing CHM13's methylation state with existing DNA reduced representation bisulfite-sequencing data on early human embryos (fig. S11 and table S4) (*35*), we observed that CHM13 clusters closely with cleavage and blastocyst-stage embryos as well as trophectoderm tissue.

To probe chromatin state in repetitive DNA, we generated long-read nanoNOMe data on HG002 using M.CviPI methyltransferase to decorate accessible chromatin with exogenous GpC methylation (*16*) and called CpG and GpC methylation with Nanopolish to measure chromatin accessibility (figs. S12 and S13). With the combination of long-read epigenetic data and the complete human reference, we now describe a complete human epigenome, providing a foundation for further study.

Paralog-specific epigenetic regulation

The neuroblastoma breakpoint family (*NBPF*) family of genes has been implicated in the expansion of the human prefrontal cortex since our lineage diverged from apes (*36*). One of its copies, *NBPF1*, has been reported to act as a tumor suppressor in neuroblastoma, in which hypomethylation of CGIs has been associated with astrocytoma formation (*37*). Understanding the regulation of this gene family, however, has been particularly challenging because the *NBPF* genes correspond to large, high-identity duplications (>98%) that are copy number polymorphic among humans and map to gaps in the existing reference sequences (*38*). The fully resolved nature of T2T-CHM13 allowed us to remap ENCODE data to discover regulatory elements associated with this gene family (Fig. 2A). When comparing the balance between H3K36me3, a mark of active exons and gene bodies, and H3K27me3, a repressive mark, in samples including the BE2C cell line (neuroblastoma) and primary brain microvascular tissue (normal brain), we found that BE2C shows a higher proportion of H3K27me3 peaks (BE2C 38, brain 8) and a lower proportion of H3K36me3 peaks (BE2C 36, brain 89) at *NPBF* loci (fig. S14 and table S5). Taking advantage of the increased resolution and more accurate *NBPF* copy number provided by T2T-CHM13 (*39*), we assayed paralog-specific epigenetic changes occurring in neuroblastoma (Fig. 2B). Among the different *NBPF* gene copies, the largest shifts in epigenetic regulation occurred at *NBPF26* and *NBPF10*, moving from active marks in primary brain microvascular tissue to repressive marks in BE2C. These specific *NBPF* copies are noteworthy because they associate with the human-specific duplicate genes *NOTCH2NLA* and *NOTCH2NLR*, determinants of the size and complexity of the human neocortex (*40*). This association identifies the functional *NBPF* copies, emphasizing the importance of studying paralog-specific epigenetics for the discovery of potential drug targets.

Regulatory regions are excluded because of low short-read mappability scores among high identity paralogs as in the *NBPF* gene family (Fig. 2C and fig. S15) (*18*). We found that genome-wide methylation, H3K4me2 (a mark of active promoters), and H3K27me3 (a repressive mark) correlate with Iso-Seq coverage (transcription) and together can be used to systematically evaluate the functional activity of this gene family (fig. S16). We correlated this activity with the evolutionary age of the paralogs, estimated using *NBPF* gene paralogs

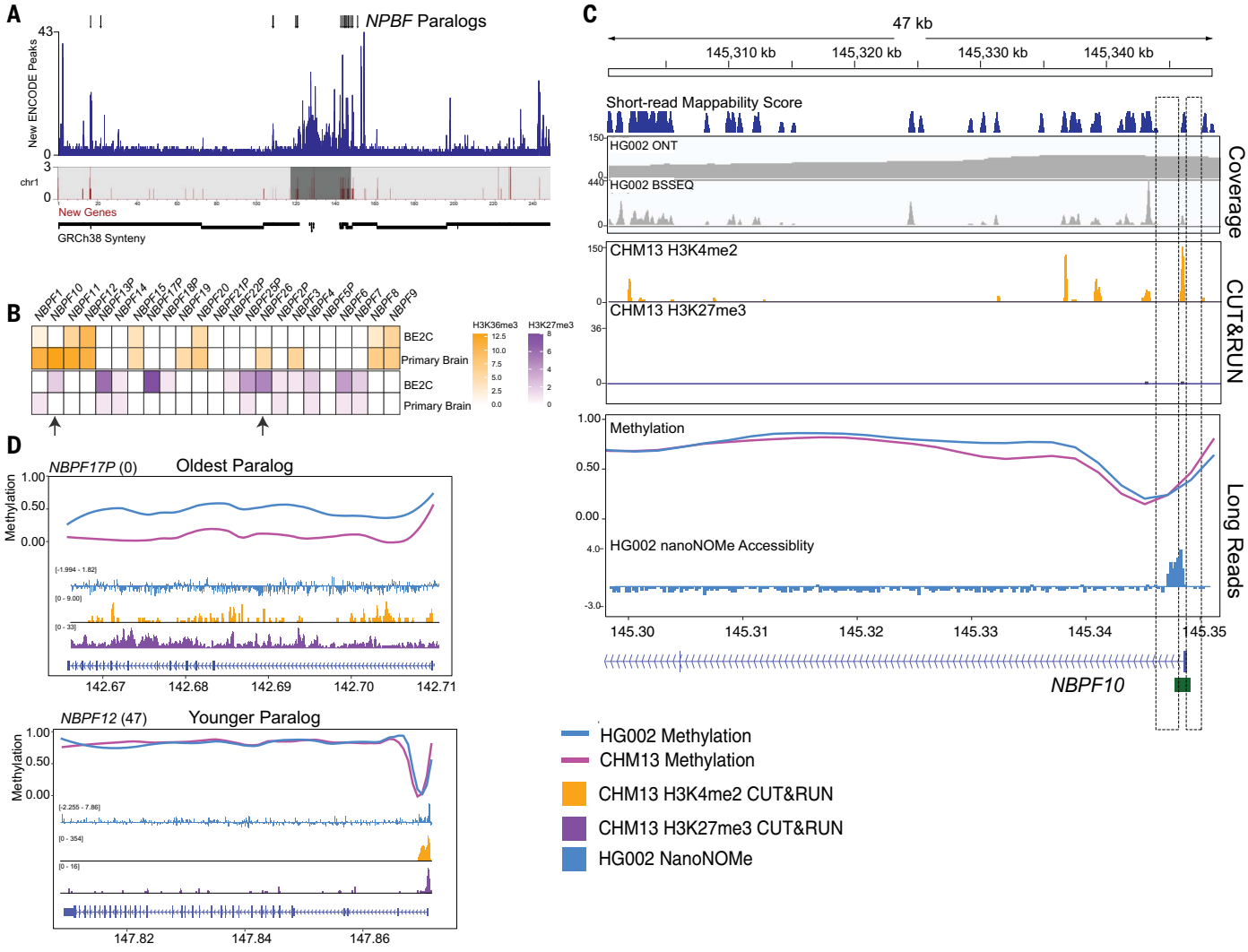


Fig. 2. Paralog-specific epigenetic regulation of the NBPF gene family. (**A**) Location of T2T-CHM13 previously uncalled ENCODE peaks across chromosome 1 in 50-kb bins (purple). Chromosome ideograms contain the density of previously unannotated genes (red) and centromere annotations (dark gray). NBPF paralogs are indicated by black arrows (top). (**B**) Heatmap illustrating the number of peaks for H3K36me3 (orange) and H3K27me3 (purple) per NBPF paralog in the ENCODE cell line BE2C (neuroblastoma) and brain tissue (primary brain microvascular tissue). Arrows indicate NBPF10 and NBPF26. (**C**) Epigenetic data at the NBPF10 promoter and first intron (chromosome 1: 145,300,425 to 145,348,763). Short-read mappability score from 0 to 200 calculated as a 200-bp region, with a score of 200 being the most mappable and 0 being the least mappable. Coverage tracks (Illumina WGBS and ONT) and CUT&RUN tracks display read pileups. Long-read methylation tracks show base-level methylation frequency, with 0 as unmethylated and 1 as fully methylated. The long-read HG002 accessibility track is a 200-bp binned *Z*-score of nanoNOMe GpC methylation frequency. Dashed boxes highlight the promoter region that is largely unmappable with short-reads. (**D**) Bottom: younger NBPF12 gene paralog displaying CHM13 and HG002 nanopore methylation, CHM13 H3K4me2 and H3K27me3 CUT&RUN coverage, and HG002 nanoNOMe. Top: older NBPF17P gene paralog displaying CHM13 and HG002 nanopore methylation, CHM13 H3K4me2 and H3K27me3 CUT&RUN, and HG002 nanoNOMe. Numbers in parentheses refer to the number of PacBio Iso-seq transcripts mapped to this paralog.

from six nonhuman primates from local genome assembly of the *NBPF* gene family from each primate (*39*) (fig. S17). The oldest paralog, *NBPF17P*, has low Iso-Seq coverage correlated with an epigenetic signature consistent with a repressive state, including promoter hypermethylation and inaccessibility, enrichment of H3K27me3, and decline of H3K4me2 (Fig. 2D). By contrast, the younger paralogs, including human-specific copies, have higher Iso-Seq coverage and epigenetic signatures consistent with active genes, including hypomethylated and accessible promoters and enrichment of H3K4me2. Activity in the younger paralogs is more variable, with *NBPF10* and *NBPF20* displaying high functional activity and sharing promoters with *NOTCH2NLA* and *NOTCH2NLB*. Taken together, our results illustrate the role of epigenetics in the regulation of gene paralogs, silencing evolutionarily older paralogs while activating newer copies. This provides mechanistic insight into potentially functional genes related to human-specific cortical expansion and dysregulation in neoplasia.

Array-specific epigenetic regulation of tandem repeats

Using k-mer–directed ENCODE alignments to the T2T-CHM13 reference, we report epigenetic features from human centromeric regions, subtelomeres, and acrocentric short arms, which represent previously unresolved regions of the genome that are dominated by CenSat DNAs (fig. S18). Five different ENCODE lines had an enrichment of H3K9me3 in CenSat DNA, notably observed in short-read mappable regions of the acrocentric short arms (fig. S19).

SJCRH30 (a rhabdomyosarcoma-derived line) had lower H3K9me3 enrichment in CenSat compared with the rest of the chromosome, suggesting satellite epigenetic dysregulation as a clinically relevant pathology in rhabdomyosarcoma (figs. S19 and S20A and B). This trend can be observed with more detail in an HSat3 repeat on the acrocentric arm of chromosome 15, where H3K9me3 in SJCRH30 is clearly depleted compared with HAP-1 (fig. S20C).

In contrast to these heterochromatic marks, we found enrichment of activating marks, including H3K27ac, H3K4me3, and CTCF in the telomere-associated repeat (TAR) region, typically located 2 kb upstream from the canonical telomeric repeat. A CTCF site in the TAR loci drives transcription of the TERRA lncRNA (*41*), a negative regulator of telomerase-mediated telomere elongation. We observed enrichment of CTCF in all ENCODE cell lines at the TAR loci (fig. S21A), but the subtelomeric regions were rich in SDs, resulting in the TAR sequence being dispersed throughout the genome (*42*). When comparing telomeric TAR sequences with nontelomeric TAR sequences, we did not observe statistically significant differences (Kruskal-Wallis, $P = 0.12$) in sequence divergence (fig. S21B). Although both telomeric and nontelomeric TAR sequences are enriched for CTCF, the nontelomeric TAR sequences are more enriched for the activating chromatin marks H3K27ac and H3K4me3, suggesting differences in TERRA activity.

Examining nanopore CpG methylation in tandemly repeated satellite DNA elements in CHM13 and HG002 revealed hypomethylation in CHM13 compared with HG002 (Fig. 3A) (*43*). To assess the chromatin profile of satellite repeats, we called accessibility peaks from the HG002 nanoNOMe data (*18*). We found that, when corrected for the size of the region, repeats have lower peak density than the genome as a whole. The number of nanoNOMe peaks per megabase of sequence was lower in satellite DNA (1.5), LINEs (8), SINEs (15), and LTRs (13.4) compared with the whole genome (31.8) (Fig. 3B and table S6) (*44*). The human satellites (HSat 2,3) and monomeric alpha satellites (MON) were largely devoid of accessibility peaks. Repetitive DNA is typically associated with densely packed heterochromatin (*45*), and our findings are consistent with this association and transcriptional profiles from (*44*). However, our data allow us to investigate accessibility profiles within previously unmappable satellite repeats.

Contrary to the expectation of compact chromatin and satellite DNA, we discovered enrichment of accessibility peaks in the SST1 satellite both inside the CenSat (41.4 peaks/Mbp) and in the chromosome arms (198.1 peaks/Mbp). Our peak annotations in HG002 were consistent with (*44*), which showed higher activity in CHM13 at noncentromeric arrays on chromosomes 4 and 19 compared with other SST1 arrays (table S7). After the SST1, the satellite repeat with the second highest peak enrichment was the ACRO_Composite, a 7-kb repeat found across 12 chromosomes, including as tandemly arrayed sequences across the five acrocentrics with high sequence identity across composite units (*44*). The tandemly arrayed promoter elements in the ACRO_composite give rise to a periodic bimodal methylation structure across the array (Fig. 3C). This epigenetic pattern has been proposed to be important for both the efficient transcription of noncoding RNAs and the maintenance of the nearly perfect tandem arrays (*46*). The array has regions of increased CpG methylation that were associated with nanoNOMe peaks and transcription (CHM13 PRO-seq) (Fig. 3C). We quantified nanoNOMe peak densities across the ACRO_Composite between chromosomes and found that chromosome 21 has the highest (4.5 peaks/100 kb) and chromosomes 13 and 15 have the lowest (0 peaks/100 kb) (Fig. 3D). The absence of nanoNOMe peaks in chromosomes 13 and 15 is correlated with low transcriptional activity (fig. S22). This high-resolution look within the acrocentric repeats indicates chromosome-specific activity of the ACRO_Composite across both CHM13 and HG002, suggesting a persistent functional role for the ACRO noncoding RNA throughout early- and late-stage development.

By contrast, we also observed methylation periodicity in untranscribed satellite repeats such as the HSat2, and these regions were largely inaccessible as measured by nanoNOMe (Fig. 3E) (*44*). This periodicity in methylation corresponds to the underlying chromatin structure and echoes the genetic repeat size, suggesting the presence of functional genomic elements. Our initial epigenetic assessments of these assembled satellite sequences indicate a complicated regulatory structure stretching beyond the accepted notion that the repetitive fraction of mammalian genomes is entirely methylated and repressed by a highly condensed chromatin state (*47*).

Single-read-level analysis in satellite arrays reveals array heterogeneity

Long reads, coupled to a complete reference assembly, confer the ability to explore methylation patterns of single molecules, each of which represents the methylation pattern of a single allele from a single cell. The X chromosome provides a unique opportunity to study these patterns because of the role of allele-specific methylation in X chromosome inactivation (XCI). Female somatic tissues have a mixture of paternal or maternal X expression because the same X chromosome is not always repressed; therefore, the active X (Xa) and inactive X (Xi) cannot be distinguished with heterozygous single-nucleotide polymorphisms alone. Examining methylation state at CGIs, we clustered reads on the CHM13 X chromosome as hyper- or hypomethylated (*18*). To determine whether the clusters represent the Xa and Xi, we first focused on genes known to be subject to XCI (XCI genes) or known to escape inactivation (escape genes) and compared our results with the clonal female lymphoblast cell line GM12878, in which the Xi is always the paternal allele (fig. S23A and B) (*48*). There, we found the Xa to have hypomethylated promoters and hypermethylated gene bodies compared with the Xi (*49*). However, in CHM13, we discovered that not all genes (e.g., *TAF9B* and *PRKX)* were properly regulated, with *TAF9B* escaping XCI and *PRKX* being subject to XCI, contrary to expectation. This is likely due to failure of X chromosome inactivation in androgenetic CHMs (fig. S23C and D and table S8) (*50*).

Moving this analysis into repetitive regions, we analyzed DXZ4, a satellite that acts as a major epigenetic regulator of XCI (*51*). This 165-kb macrosatellite repeat contains 3-kb monomeric units, each with a bidirectional CGI promoter and a CTCF site that is hypomethylated on the Xi and hypermethylated on the Xa in healthy cells (*52*, *53*). Single-read clustering revealed two distinct clusters of reads, one with higher methylation across the repeat and the other with lower methylation (Fig. 3F). This analysis revealed an unexpected level of heterogeneity in methylation of monomers within the array. We hypothesize that this variation is a result of the aberrant XCI state of CHM13, because intra-array variation was not observed in the Xa at DXZ4 in HG002 (fig. S24). Observing epigenetic differences between monomers of satellite repeats could grant insights into human disease, providing a detailed mechanistic understanding of satellite dysregulation. From this analysis, we demonstrate that we can cluster reads using methylation alone to identify heterogeneous populations and intra-array epigenetic variation even in the absence of heterozygous genetic variants.

Methylation maps of human centromeres reveal complex epigenetic patterns

Human centromeres are composed of alpha satellite DNA, with an AT-rich, ~171-bp repeat unit or "monomer." The largest arrays of alpha satellites in the human genome are further organized in chromosome-specific, higher-order repeats (HORs) or larger, multimonomeric repeat units (*54*). Centromeres can contain multiple distinct alpha satellite HOR arrays that can be classified into active and inactive HORs (*55*, *56*). The HORs within active arrays have specialized epigenetic regulation that are important in establishing and maintaining centromere identity (*56*, *57*).

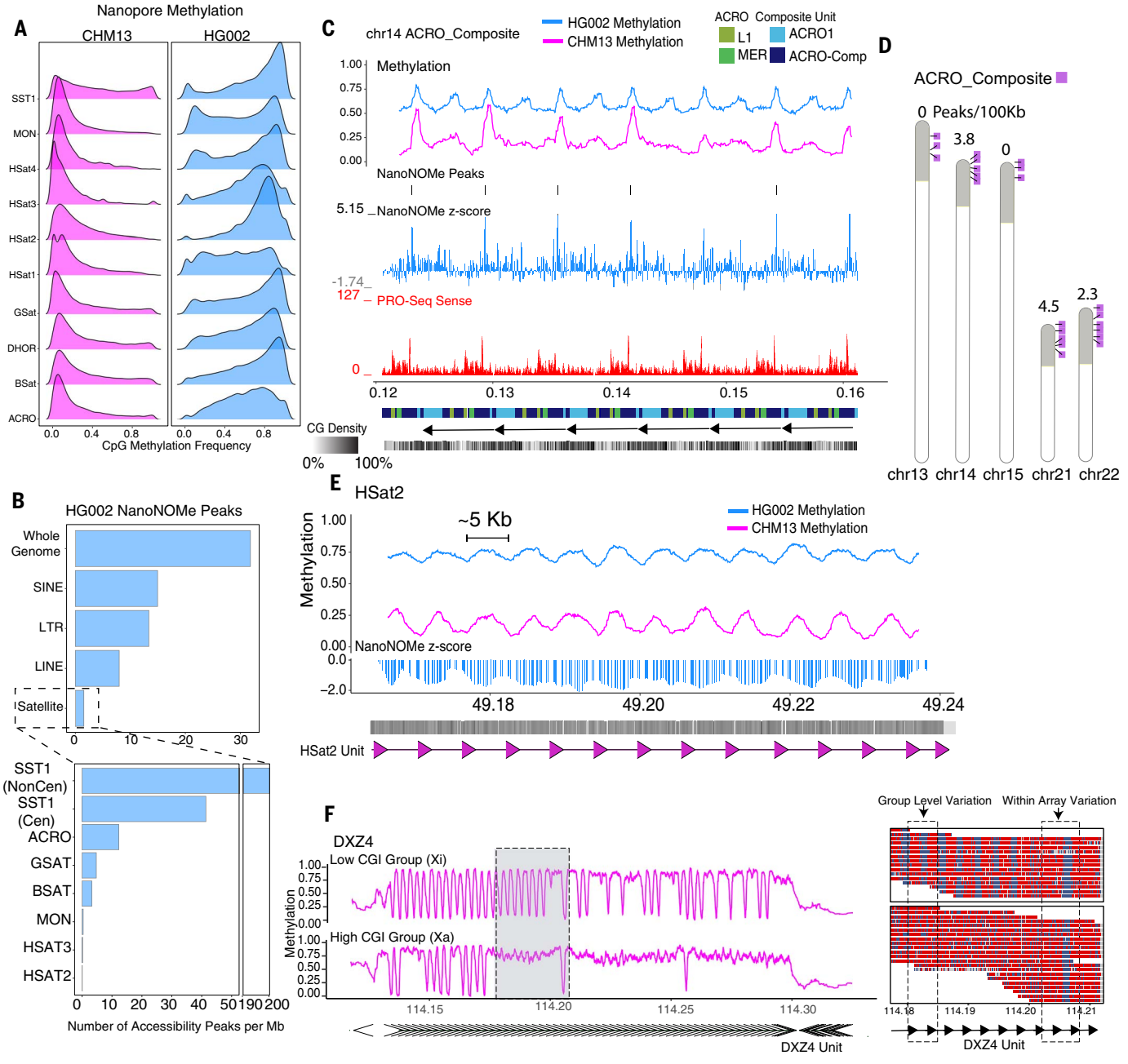


Fig. 3. Context-specific epigenetics in high-identity tandem repeats. (**A**) Nanopore methylation frequency of satellite repeat classes in CHM13 and HG002. (**B**) HG002 NanoNOMe statistically significant peak calls per 1 Mb of sequence in all major repeat classes compared with the whole genome (top) and within different satellite repeats (bottom). (**C**) Nanopore CpG methylation profiles, HG002 NanoNOMe accessibility peaks and *Z*-score (negative is inaccessible, positive is accessible), and non–kmer-filtered (multimapping) PRO-Seq coverage at the ACRO_Composite repeat (chromosome 14: 121,193 to 162,142). Annotation tracks at the bottom are the RepeatMasker V2 annotation from (*44*), monomeric annotations of the ACRO_Composites, and a GC density track. (**D**) Ideogram showing the arrayed locations of the ACRO_Composite across the acrocentric chromosomes (purple) within the acrocentric short arms (gray shaded). Listed above each chromosome is the nanoNOMe ACRO_composite peak density in peaks/100 kb. (**E**) Nanopore CpG methylation profiles and HG002 NanoNOMe accessibility *Z*-score of the HSat2 repeat (chromosome 16: 49,163,529 to 49,239,753). Annotation bars below represent CpG density and HSat2 repeat units on the bottom. (**F**) The DXZ4 locus on CHM13 clustered into two haplotypes (low CGI methylation and high CGI methylation) based solely on promoter methylation state. Left: methylation frequency plot of each haplotype. Right: single reads from the gray highlighted region on the left, with boxes highlighting CGI cluster group-level epigenetic variability and intra-array-level epigenetic variability between neighboring monomeric units.

Centromere protein A (CENP-A) is an H3 variant that is enriched in centromeric nucleosomes and marks sites of kinetochore assembly (*58*). In HOR arrays, notable hypomethylation colocalizing with CENP-A enrichment at chromosomes X and 8 has been described (*13*, *14*). We extended this finding to all CHM13 centromeres, referring to this hypomethylation as the centromeric dip region (CDR) (Fig. 4A and table S9). We found that CDRs were present only in active HORs (fig. S25), and that active HORs were larger in size and had higher mean methylation frequency than inactive HORs, as exemplified by the chromosome 5 centromere (Fig. 4B).

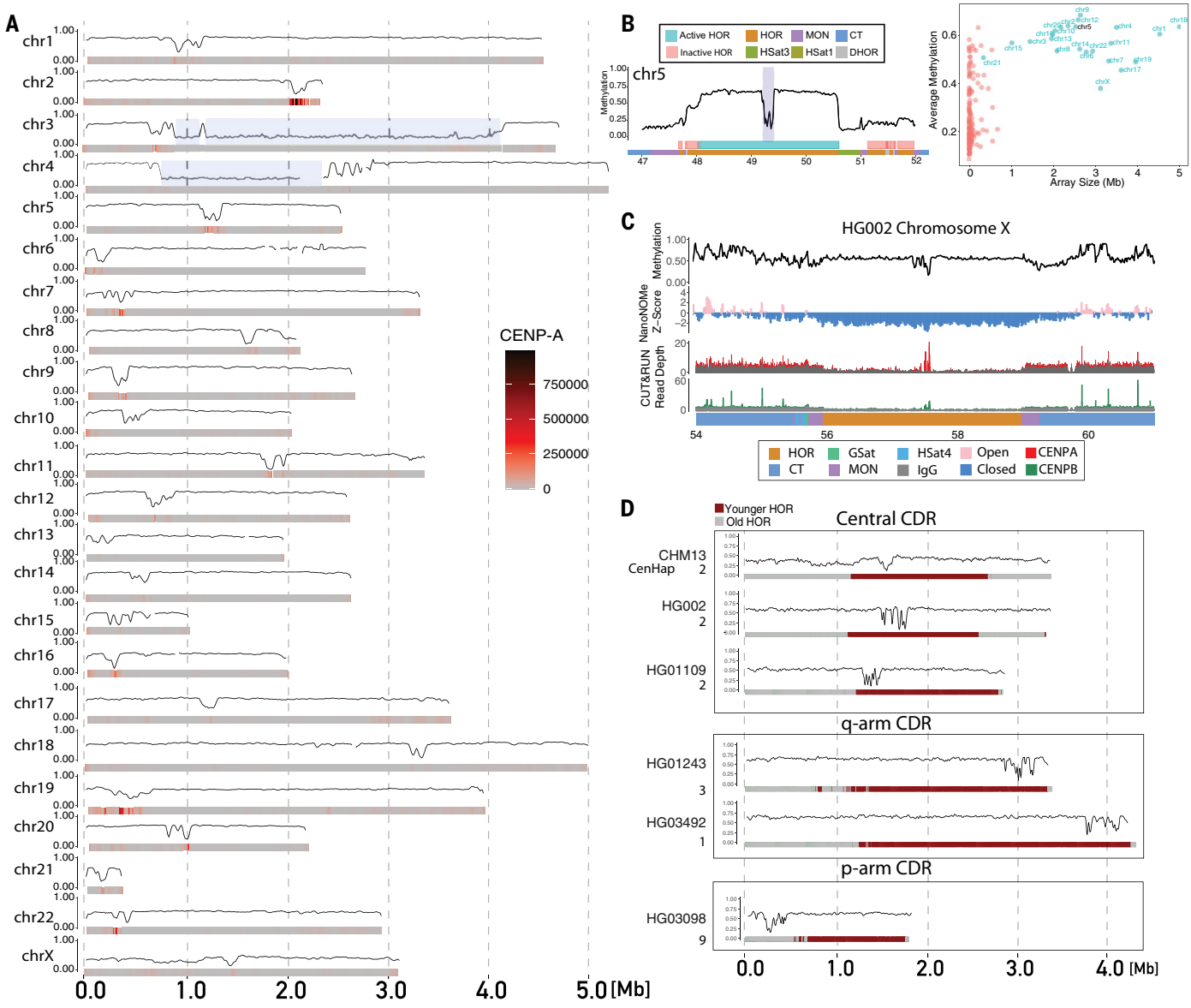


Fig. 4. Epigenetic maps within human centromeres. (**A**) Smoothed methylation frequency in 10-kb bins of the active HOR array for all CHM13 chromosomes. CENP-A enrichment from CUT&RUN data is shown as a heatmap under each plot. Chromosomes 3 and 4 have an HSat1 repeat (blue highlight) that breaks up the live HOR array. (**B**) Left: CHM13 methylation in the centromeric region of chromosome 5. Smoothed methylation frequency is plotted in 10-kb bins. HOR arrays are annotated as blue ("active") and pink ("inactive"). Right: scatter plot of average methylation within each HOR array versus size in megabase pairs. (**C**) Methylation, nanoNOMe accessibility, and CENP-A and CENP-B CUT&RUN data across the chromosome X centromeric array on HG002. Smoothed methylation and accessibility are plotted in 15-kb bins, and CUT&RUN is plotted as raw read counts with input shaded gray. Bottom bar annotates satellite regions indicating the location of the HOR, MON, GSat, HSat4, and CT regions. (**D**) Methylation in the active HOR array across diverse individuals. Coriell cell line sample ID and cenhap group are annotated to left. HORs are annotated as red (younger) and gray (older) computed on the basis of sequence divergence.

These results underscore the importance of methylation in proper centromere regulation and kinetochore assembly.

To investigate whether CDRs were confined only to early developmental samples, we examined HG002 nanopore-sequencing data to probe centromere methylation in an adult differentiated cell line. However, the high level of HOR array variability and the resulting inability to confidently phase and map reads from diploid chromosomes prevented us from using the T2T-CHM13 HOR reference for HG002 reads, as evidenced by the anomalous coverage that we observed for HG002 alignments in the HOR arrays (fig. S26) (*59*). Instead, we took advantage of the haploid nature of the HG002 X chromosome and used an HG002-specific X centromere reference (*4*, *56*), and were able to clearly observe a CDR (Fig. 4C). Furthermore, using nanoNOMe, the CDR was coordinated in this sample with a highly inaccessible region. When we examined the size of the inaccessible regions in the HOR versus the surrounding pericentromeric and centromeric transition (CT) regions, we found that the HORs were enriched in dinucleosomes compared with these other regions (fig. S27). Finally, looking at CUT&RUN CENP-A and centromere protein B (CENP-B) data, we observed a peak of CENP-A and CENP-B binding at the CDR. This is coordinated with a marked hypomethylation of the CENP-B motif within CDRs as opposed to outside the CDRs (fig. S28), and methylation is known to reduce CENP-B binding (*60*). This finding highlights the potential functional

importance of the CDR for kinetochore formation.

Taking this a step further, using Human Pangenome Reference Consortium (HPRC) data, we leveraged the assembled X chromosomes of four additional diverse male samples representing individuals included in the 1000 Genomes Project (Fig. 4D) (*56*, *61*). All arrays showed a distinct CDR in the X chromosome, with positional variability in the CDR location across individuals. Furthermore, CDR position was shared between individuals with more closely related centromere-spanning haplotypes (cenhap) assignments.

Cenhaps are long haplotypes that include centromere arrays due to reduced recombination in CenSat regions (*56*, *62*). Three of the samples, CHM13 (European), HG002 (European), and HG01109 (Puerto Rican), are within cenhap group 2, and all contain a centrally positioned CDR within an evolutionarily "younger" region of the HOR, as defined in (*56*). Two of the samples, HG01243 (Puerto Rican) and HG03492 (Pakistani), are within cenhap groups 3 and 1, which have been shown to be phylogenetically related (i.e., they share a clade with cenhaps 1 to 4) (*56*), and have a CDR positioned more toward the q-arm side of the centromere within the evolutionarily younger region of the HOR array. Finally, one of the samples, HG03098 (African), from the more distantly related cenhap group 9, has a CDR positioned toward the p-arm of the centromere in an older (more diverged) region of the HOR array (supporting the previous observation of an epi-allele in the region using available short-read datasets) (*56*). Therefore, we demonstrate the use of CDRs to identify epigenetic variability within human centromeres, variations that may influence centromere function during cell division. These variations show the critical importance of epigenetic profiling in the centromere, finding variation between individuals in a discrete, epigenetically defined region of the centromere.

DISCUSSION

This work provides a comprehensive view of the epigenetic organization of a complete human genome, uncovering complex epigenetic patterns in the previously unresolved 8% of the human genome. Functional annotation of these intractable regions has not been overlooked because of their lack of importance but rather because of technological limitations. Our study opens these regions to explore their epigenome, leaving no region of the genome unreachable. Here, with the combination of a complete genome assembly and technological advances in epigenetic profiling, we make substantial strides in functional genome assessment, expanding ENCODE (*1*) to include 3 to 19% more peak calls and increasing the number of CpG methylation calls by 10%. Long-read epigenetic methods, here focusing on nanopore methylation and chromatin accessibility, can resolve single-molecule epigenetic patterns within these regions, providing a foundational assessment of these areas. Long-read methylomes of distinctive developmental time points surveyed >99% of CpGs, establishing the CHM13 and HG002 methylomes as the most complete human methylomes to date (*3*). With these datasets, we profiled the additional 225 Mbp of sequence and 2680 gene annotations.

Of the previously unresolved genes, we found 57 with evidence of active promoters, including H3K4me3 or H3K27ac marks, in more than one cell type. We found 82 genes with a single cell type supporting active promoters, providing evidence that these previously unresolved gene annotations are functionally active across tissues. With more data from different tissue types, we may identify even more functional genes. More generally, we found that evolutionarily older gene paralogs were epigenetically repressed (similar to the epigenetic silencing of transposons), conferring genome stability and thus influencing genome evolution (*63*, *64*).

Examining satellite DNA, we integrated short- and long-read datasets to investigate complete satellite arrays, revealing that these regions vary in epigenetic and transcriptional activity despite high sequence identity and highlighting the importance of the local chromosome environment as a modulator of epigenetics. Repetitive DNA on the acrocentric short arms is known to play a role in nucleolar formation; however, the previous absence of these regions from the human reference has hampered research (*65*). Our findings suggest that, rather than acting in unison, the repeat families on these individual acrocentric chromosomes all have their own epigenetic identity, likely contributing to their unique functional roles in genome integrity and organization.

One of the features of our single-molecule epigenetic data is the ability to investigate single-molecule patterns of epigenetics. We used methylation alone to cluster reads in repetitive areas devoid of heterozygous polymorphisms. This includes the DXZ4 array, in which the methylation signature is critical to X chromosome inactivation (*66*, *67*). With the increase in resolution, our results show methylation variability between the clustered populations and intra-array epigenetic variation within adjacent monomers in the same array. Because satellite arrays are known to be hypervariable in the human population and linked to several human diseases, these results highlight the importance of long-read single-molecule epigenetic studies for understanding disease pathology.

Finally, the T2T-CHM13 genome assembly has opened exploration of the human centromere, enabling us to probe the epigenetic elements that define centromeric chromatin. We extended our original discovery of the CDR in chromosome 8 and chromosome X to all chromosomes, and found that CDRs denote the position of centromere-associated proteins (CENP-A and CENP-B in the HG002 genome) in differentiated cells (HG002, a lymphoblast). This provides evidence of CDRs outside of early developmental CHMs and emphasizes their importance in kinetochore positioning and epigenetic regulation of chromosome segregation. Expanding our CDR analysis to male X chromosomes representing diverse haplotypes, we uncovered variability in the localization of the CDR within the X HOR array. Such variability in active centromeric arrays has been explored through the presence of epi-alleles (*68*); however, we have been able to demonstrate the use of CDRs to precisely predict kinetochore site localization within an active array and report across individuals representing diverse ancestry. When combined with findings in other organisms, e.g., maize (*69*) and medaka (*70*), this suggests that the CDR is a conserved, functionally important feature of complex centromeres across vertebrate and plant lineages. Proper kinetochore formation is an essential process for eukaryotic cell division, a process that occurs in humans 330 billion times per day to sustain life. Our results lead to two major conclusions about the CDR: (i) CDR location on a given array is fixed in early development and maintained upon differentiation and (ii) there is a single stable CDR in each centromere. Our initial profile provides a multitude of avenues for future research, including how CDR position influences meiotic and mitotic stability, disease, and aneuploidy.

Our results act as a foundational study, expanding studies of the human genome through the use of the complete reference. There remain significant challenges to further exploring the epigenome in a larger and more diverse sample set to achieve optimal sequence alignment, especially among structurally variable repetitive regions, e.g., HORs. Efforts by the HPRC (*71*) to generate fully phased diploid genome assemblies will enable population-scale exploration of these areas. Limitations of short-read sequencing in unique regions can be supplemented by long-read epigenetic methods currently under rapid development (*16*, *17*). We are on the precipice of exploration into duplicated and repetitive portions of the genome. Further development of long-read epigenetic profiling across different populations and disease states will reveal more about regulation within the genome's most elusive regions.

METHODS SUMMARY

Methylation processing

Nanopore reads were obtained from (*13*, *14*, *72*). Ultra-long nanopore reads were aligned to the CHM13 reference (*4*) with Winnowmap version 2.0 (*73*) with a k-mer size of 15. BAM

files were filtered for primary alignments with SAMtools (version 1.9); analysis of centromeric regions was done on reads >50 kb. To measure CpG methylation in nanopore data, we used Nanopolish (version 0.13.2) with a log-likelihood ratio (LLR) cutoff of −1.5/1.5 (*34*). HG002 bisulfite FASTQs were collected from the Oxford Nanopore Technologies (ONT) open data repository https://labs.epi2me.io/gm24385-5mc. Paired-end FASTQs were aligned with Bismark (version 0.22.2) (*74*). For Nanopolish to Megalodon comparisons, Megalodon was run with the r9.4.1_450bps 5mC model with thresholding set as default.

NanoNOMe

HG002 cells were grown in culture and treated according to methods outlined in (*16*). Purified genomic DNA was prepared for nanopore sequencing following the protocol in the genomic sequencing by ligation kit LSK-SQK109 (ONT). To measure CpG and GpC methylation in nanopore data, we used Nanopolish (version 0.13.2) on the nanonome branch https://github.com/jts/nanopolish/tree/nanonome (*34*). We set an LLR threshold of −1/1 for GpC methylation calls and −1.5/1.5 for CpG methylation calls.

Methylation clustering

Methylation clustering was performed across the CHM13 X chromosome on all CGIs that overlap an annotated promoter of a protein-coding gene. Within the CGI, reads with an average methylation >0.2 were considered methylated, and reads with an average methylation <0.2 were considered unmethylated. Reads were only considered if they spanned the entirety of the CG islands and were longer than 5 kb. Clustered reads were then intersected with known escape and XCI genes from (*51*). The same clustering procedure was performed at the DXZ4 locus.

CUT&RUN

CUT&RUN was performed as detailed in (*75*) with some variations. For library preparation, NEBNext Ultra II End repair/A-tailing and ligation kits were used as indicated by the manufacturer, with 1.5 pg of Spike-in Yeast DNA added (obtained from the Henikoff laboratory at the Fred Hutchinson Cancer Research Center). Marker-assisted mapping of CUT&RUN data (CHM13 CENP-A, CHM13 H3K4me2, CHM13 H3K27me3, HG002 CENP-A, and HG002 CENP-B) to a sample-specific reference [CHM13 to T2T-CHM13 or HG002 to CHM13 autosomes (chromosomes 1 to 22), HG002 T2T (chromosome X), and GRCh38 (chromosome Y)] was performed according to the methods outlined in (*56*).

ENCODE dynamic k-mer–assisted mapping

We selected several ChIP-seq datasets generated as part of the ENCODE project (*1*), choosing those with at least 100-bp paired-end sequencing data and at least one matching input control. These criteria yielded 96 total sequencing libraries (table S9). Reads were mapped with Bowtie2 [version 2.4.1 (*76*)], alignments were filtered using SAMtools [version 1.10 (*77*)], and polymerase chain reaction duplicates were identified and removed with Picard tools [version 2.22.1 (http://broadinstitute.github.io/picard)]. Alignments were then filtered for unique k-mers. Specifically, for each alignment, reference sequences aligned with template ends were compared with a database of k-mers unique in the whole genome. For each end of the paired-end sequencing reads, the k-mer length was determined by finding the largest multiple of 5 less than or equal to the aligned reference sequence length. Peak calls were made using MACS2 (version 2.2.7.1) (*78*) with default parameters and estimated genome sizes of 3.03×10^9 and 2.79×10^9 for chm13v1 and GRCh38p13, respectively.

REFERENCES AND NOTES

1. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
2. J. Dekker *et al*., The 4D nucleome project. *Nature* **549**, 219–226 (2017). doi: 10.1038/nature23884; pmid: 28905911
3. A. Kundaje *et al*., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi: 10.1038/nature14248; pmid: 25693563
4. S. Nurk *et al*., The complete sequence of a human genome. *Science* **376**, 44–53 (2022). doi: 10.1126/science.abj6987
5. D. Jost, C. Vaillant, Epigenomics in 3D: Importance of long-range spreading and specific interactions in epigenomic maintenance. *Nucleic Acids Res.* **46**, 2252–2264 (2018). doi: 10.1093/nar/gky009; pmid: 29365171
6. N. V. Fedoroff, Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758–767 (2012). doi: 10.1126/science.338.6108.758; pmid: 23145453
7. D. Gabellini, M. R. Green, R. Tupler, Inappropriate gene activation in FSHD: A repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell* **110**, 339–348 (2002). doi: 10.1016/S0092-8674(02)00826-7; pmid: 12176321
8. H. A. Bruce *et al*., Long tandem repeats as a form of genomic copy number variation: Structure and length polymorphism of a chromosome 5p repeat in control and schizophrenia populations. *Psychiatr. Genet.* **19**, 64–71 (2009). doi: 10.1097/YPG.0b013e3283207ff6; pmid: 19672138
9. D. Thoraval *et al*., Demethylation of repetitive DNA sequences in neuroblastoma. *Genes Chromosomes Cancer* **17**, 234–244 (1996). doi: 10.1002/(SICI)1098-2264(199612)17:4<234::AID-GCC5>3.0.CO;2-4; pmid: 8946205
10. Y.-T. Chen *et al*., Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. *Genes Chromosomes Cancer* **45**, 392–400 (2006). doi: 10.1002/gcc.20298; pmid: 16382448
11. D. T. Ting *et al*., Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **331**, 593–596 (2011). doi: 10.1126/science.1200801; pmid: 21233348
12. K. M. Hassan, T. Norwood, G. Gimelli, S. M. Gartler, R. S. Hansen, Satellite 2 methylation patterns in normal and ICF syndrome cells and association of hypomethylation with advanced replication. *Hum. Genet.* **109**, 452–462 (2001). doi: 10.1007/s004390100590; pmid: 11702227
13. G. A. Logsdon *et al*., The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021). doi: 10.1038/s41586-021-03420-7; pmid: 33828295
14. K. H. Miga *et al*., Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi: 10.1038/s41586-020-2547-7; pmid: 32663838
15. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020). doi: 10.1038/s41576-020-0236-x; pmid: 32504078
16. I. Lee *et al*., Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020). doi: 10.1038/s41592-020-01000-7; pmid: 33230324
17. A. B. Stergachis, B. M. Debo, E. Haugen, L. S. Churchman, J. A. Stamatoyannopoulos, Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020). doi: 10.1126/science.aaz1646; pmid: 32587015
18. Materials and methods are available as supplementary materials.
19. A. Janssen, S. U. Colmenares, G. H. Karpen, Heterochromatin: Guardian of the Genome. *Annu. Rev. Cell Dev. Biol.* **34**, 265–288 (2018). doi: 10.1146/annurev-cellbio-100617-062653; pmid: 30044650
20. J. Chen, Z.-H. Yuan, X.-H. Hou, M.-H. Shi, R. Jiang, LINC01116 promotes the proliferation and inhibits the apoptosis of gastric cancer cells. *Eur. Rev. Med. Pharmacol. Sci.* **24**, 1807–1814 (2020). pmid: 32141549
21. D. Gabellini *et al*., Facioscapulohumeral muscular dystrophy in mice overexpressing FRG1. *Nature* **439**, 973–977 (2006). doi: 10.1038/nature04422; pmid: 16341202
22. G. Giannuzzi *et al*., The human-specific BOLA2 duplication modifies iron homeostasis and anemia predisposition in chromosome 16p11.2 autism individuals. *Am. J. Hum. Genet.* **105**, 947–958 (2019). doi: 10.1016/j.ajhg.2019.09.023; pmid: 31668704
23. Z. Jiang *et al*., Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007). doi: 10.1038/ng.2007.9; pmid: 17922013
24. F. A. M. Maggiolini *et al*., Genomic inversions and GOLGA core duplicons underlie disease instability at the 15q25 locus. *PLOS Genet.* **15**, e1008075 (2019). doi: 10.1371/journal.pgen.1008075; pmid: 30917130
25. J. Schmutz *et al*., The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, 268–274 (2004). doi: 10.1038/nature02919; pmid: 15372022
26. T. W. Prior, Perspectives and diagnostic considerations in spinal muscular atrophy. *Genet. Med.* **12**, 145–152 (2010). doi: 10.1097/GIM.0b013e3181c5e713; pmid: 20057317
27. J. Hauke *et al*., Survival motor neuron gene 2 silencing by DNA methylation correlates with spinal muscular atrophy disease severity and can be bypassed by histone deacetylase inhibition. *Hum. Mol. Genet.* **18**, 304–317 (2009). doi: 10.1093/hmg/ddn357; pmid: 18971205
28. P. Cruz-Tapias, J. Castiblanco, J.-M. Anaya, "Major histocompatibility complex: Antigen processing and presentation," in *Autoimmunity: From Bench to Bedside* (El Rosario Univ. Press, 2013), pp. 271–284.
29. A. Sekar *et al*., Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016). doi: 10.1038/nature16549; pmid: 26814963
30. T. Tsukahara *et al*., Prognostic significance of HLA class I expression in osteosarcoma defined by anti-pan HLA class I monoclonal antibody, EMR8-5. *Cancer Sci.* **97**, 1374–1380 (2006). doi: 10.1111/j.1349-7006.2006.00317.x; pmid: 16995877
31. D. Bello, M. M. Webber, H. K. Kleinman, D. D. Wartinger, J. S. Rhim, Androgen responsive adult human prostatic epithelial cell lines immortalized by human papillomavirus 18. *Carcinogenesis* **18**, 1215–1223 (1997). doi: 10.1093/carcin/18.6.1215; pmid: 9214605
32. Z. Souri *et al*., HDAC inhibition increases HLA class I expression in uveal melanoma. *Cancers* **12**, 3690 (2020). doi: 10.3390/cancers12123690; pmid: 33316946
33. M. Karimzadeh, C. Ernst, A. Kundaje, M. M. Hoffman, Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018). doi: 10.1093/nar/gky677; pmid: 30169659
34. J. T. Simpson *et al*., Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017). doi: 10.1038/nmeth.4184; pmid: 28218898
35. H. Guo *et al*., The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014). doi: 10.1038/nature13544; pmid: 25079557
36. I. K. Suzuki *et al*., Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation. *Cell* **173**, 1370–1384.e16 (2018). doi: 10.1016/j.cell.2018.03.067; pmid: 29856955
37. X. Wu *et al*., CpG island hypermethylation in human astrocytomas. *Cancer Res.* **70**, 2718–2727 (2010). doi: 10.1158/0008-5472.CAN-09-3631; pmid: 20233874

38. P. H. Sudmant *et al*., Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010). doi: 10.1126/science.1197005; pmid: 21030649
39. M. R. Vollger*et al*., Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022). doi: 10.1126/science.abj6965
40. I. T. Fiddes *et al*., Human-specific NOTCH2NL genes affect Notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369.e22 (2018). doi: 10.1016/j.cell.2018.03.051; pmid: 29856954
41. Z. Deng *et al*., A role for CTCF and cohesin in subtelomere chromatin organization, TERRA transcription, and telomere end protection. *EMBO J.* **31**, 4165–4178 (2012). doi: 10.1038/emboj.2012.266; pmid: 23010778
42. A. Ambrosini, S. Paul, S. Hu, H. Riethman, Human subtelomeric duplicon structure and organization. *Genome Biol.* **8**, R151 (2007). doi: 10.1186/gb-2007-8-7-r151; pmid: 17663781
43. C. Li *et al*., DNA methylation reprogramming of functional elements during mammalian embryonic development. *Cell Discov.* **4**, 41 (2018). doi: 10.1038/s41421-018-0039-9; pmid: 30109120
44. S. J. Hoyt *et al*., From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022). doi: 10.1126/science.abk3112
45. J. J. Yunis, W. G. Yasmineh, Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation. *Science* **174**, 1200–1209 (1971). doi: 10.1126/science.174.4015.1200; pmid: 4943851
46. C. Jiang, D. Liao, Striking bimodal methylation of the repeat unit of the tandem array encoding human U2 snRNA (the RNU2 locus). *Genomics* **62**, 508–518 (1999). doi: 10.1006/geno.1999.6052; pmid: 10644450
47. G. Nishibuchi, J. Déjardin, The molecular basis of the organization of repetitive DNA-containing constitutive heterochromatin in mammals. *Chromosome Res.* **25**, 77–87 (2017). doi: 10.1007/s10577-016-9547-3; pmid: 28078514
48. A. M. Cotton *et al*., Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.* **24**, 1528–1539 (2015). doi: 10.1093/hmg/ddu564; pmid: 25381334
49. A. Hellman, A. Chess, Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141–1143 (2007). doi: 10.1126/science.1136352; pmid: 17322062
50. X. Chen *et al*., Loss of X chromosome inactivation in androgenetic complete hydatidiform moles with 46, XX karyotype. *Int. J. Gynecol. Pathol.* **40**, 333–341 (2021). doi: 10.1097/PGP.0000000000000697; pmid: 33021557
51. P. Bansal, Y. Kondaveeti, S. F. Pinter, Forged by DXZ4, FIRRE, and ICCE: How tandem repeats shape the active and inactive X chromosome. *Front. Cell Dev. Biol.* **7**, 328 (2020). doi: 10.3389/fcell.2019.00328; pmid: 32076600
52. B. P. Chadwick, DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. *Genome Res.* **18**, 1259–1269 (2008). doi: 10.1101/gr.075713.107; pmid: 18456864
53. J. Giacalone, J. Friedes, U. Francke, A novel GC-rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes. *Nat. Genet.* **1**, 137–143 (1992). doi: 10.1038/ng0592-137; pmid: 1302007
54. H. F. Willard, J. S. Waye, Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987). doi: 10.1016/0168-9525(87)90232-0
55. 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). doi: 10.1016/j.gdata.2015.05.035; pmid: 26167452
56. N. Altemose *et al*., Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022). doi: 10.1126/science.abl4178
57. R. C. Allshire, G. H. Karpen, Epigenetic regulation of centromeric chromatin: Old dogs, new tricks? *Nat. Rev. Genet.* **9**, 923–937 (2008). doi: 10.1038/nrg2466; pmid: 19002142
58. 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). doi: 10.1242/jcs.114.19.3529; pmid: 11682612
59. K. H. Miga, Centromeric satellite DNAs: Hidden sequence variation in the human population. *Genes* **10**, 352 (2019). doi: 10.3390/genes10050352; pmid: 31072070
60. Y. Tanaka, H. Kurumizaka, S. Yokoyama, CpG methylation of the CENP-B box reduces human CENP-B binding. *FEBS J.* **272**, 282–289 (2005). doi: 10.1111/j.1432-1033.2004.04406.x; pmid: 15634350
61. The 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245
62. 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**, e42989 (2019). doi: 10.7554/eLife.42989; pmid: 31237235
63. Z. Lippman *et al*., Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004). doi: 10.1038/nature02651; pmid: 15269773
64. A. V. Badyaev, Epigenetic resolution of the 'curse of complexity' in adaptive evolution of complex traits. *J. Physiol.* **592**, 2251–2260 (2014). doi: 10.1113/jphysiol.2014.272625; pmid: 24882810
65. M. van Sluis *et al*., Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. *Genes Dev.* **33**, 1688–1701 (2019). doi: 10.1101/gad.331892.119; pmid: 31727772
66. E. M. Darrow *et al*., Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E4504–E4512 (2016). doi: 10.1073/pnas.1609643113; pmid: 27432957
67. R. J. L. F. Lemmers *et al*., Cis D4Z4 repeat duplications associated with facioscapulohumeral muscular dystrophy type 2. *Hum. Mol. Genet.* **27**, 3488–3497 (2018). doi: 10.1093/hmg/ddy236; pmid: 30281091
68. 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). doi: 10.1101/gr.206706.116; pmid: 27510565
69. D.-H. Koo, F. Han, J. A. Birchler, J. Jiang, Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome. *Genome Res.* **21**, 908–914 (2011). doi: 10.1101/gr.116202.110; pmid: 21518739
70. K. Ichikawa *et al*., Centromere evolution and CpG methylation during vertebrate speciation. *Nat. Commun.* **8**, 1833 (2017). doi: 10.1038/s41467-017-01982-7; pmid: 29184138
71. K. H. Miga, T. Wang, The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021). doi: 10.1146/annurev-genom-120120-081921; pmid: 33929893
72. K. Shafin *et al*., Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020). doi: 10.1038/s41587-020-0503-6; pmid: 32686750
73. C. Jain, A. Rhie, N. Hansen, S. Koren, A. M. Phillippy, A long read mapping method for highly repetitive reference sequences. *bioRxiv* (2020), p. 2020.11.01.363887. doi: 10.1101/2020.11.01.363887
74. F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011). doi: 10.1093/bioinformatics/btr167; pmid: 21493656
75. J. Thakur, S. Henikoff, Unexpected conformational variations of the human centromeric chromatin complex. *Genes Dev.* **32**, 20–25 (2018). doi: 10.1101/gad.307736.117; pmid: 29386331
76. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: 10.1038/nmeth.1923; pmid: 22388286
77. P. Danecek *et al*., Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021). doi: 10.1093/gigascience/giab008; pmid: 33590861
78. Y. Zhang *et al*., Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008). doi: 10.1186/gb-2008-9-9-r137; pmid: 18798982
79. A. Gershman *et al*., Code repositories used for T2T epigenetics for: Epigenetic patterns in a complete human genome, *Zenodo* (2022); https://doi.org/10.5281/zenodo.6046354.
80. M. R. Vollger, A. Lo, Version of SafFire used in T2T figures for: Epigenetic patterns in a complete human genome, Zenodo (2022); https://doi.org/10.5281/zenodo.5911863.
81. A. Zeileis, G. Grothendieck, zoo: S3 infrastructure for regular and irregular time series. *J. Stat. Softw.* **14**, 1–27 (2005). doi: 10.18637/jss.v014.i06
82. M. Sobecki *et al*., MadID, a versatile approach to map protein-DNA interactions, highlights telomere-nuclear envelope contact sites in human cells. *Cell Rep.* **25**, 2891–2903.e5 (2018). doi: 10.1016/j.celrep.2018.11.027; pmid: 30517874
83. K. D. Hansen, B. Langmead, R. A. Irizarry, BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012). doi: 10.1186/gb-2012-13-10-r83; pmid: 23034175
84. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi: 10.1093/bioinformatics/btp324; pmid: 19451168
85. J. R. Miller *et al*., Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008). doi: 10.1093/bioinformatics/btn548; pmid: 18952627
86. C. A. Davis *et al*., The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46** (D1), D794–D801 (2018). doi: 10.1093/nar/gkx1081; pmid: 29126249
87. M. Kokot, M. Dlugosz, S. Deorowicz, KMC 3: Counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017). doi: 10.1093/bioinformatics/btx304; pmid: 28472236
88. F. Ramírez *et al*., deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44** (W1), W160–W165 (2016). doi: 10.1093/nar/gkw257; pmid: 27079975
89. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi: 10.1093/bioinformatics/btq033; pmid: 20110278
90. P. Rice, I. Longden, A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000). doi: 10.1016/S0168-9525(00)02024-2; pmid: 10827456
91. M. Kirsche, A. Das, M. C. Schatz, Sapling: Accelerating suffix array queries with learned data models. *Bioinformatics* **37**, 744–749 (2021). doi: 10.1093/bioinformatics/btaa911; pmid: 33107913
92. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi: 10.1093/molbev/mst010; pmid: 23329690
93. R. Bouckaert *et al*., BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019). doi: 10.1371/journal.pcbi.1006650; pmid: 30958812
94. T. Marques-Bonet, O. A. Ryder, E. E. Eichler, Sequencing primate genomes: What have we learned? *Annu. Rev. Genomics Hum. Genet.* **10**, 355–386 (2009). doi: 10.1146/annurev.genom.9.081307.164420; pmid: 19630567
95. M. de Manuel *et al*., Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016). doi: 10.1126/science.aag2602; pmid: 27789843
96. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019). doi: 10.1038/s41587-019-0201-4; pmid: 31375807
97. S. Kovaka *et al*., Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019). doi: 10.1186/s13059-019-1910-1; pmid: 31842956

ACKNOWLEDGMENTS

We thank R. Workman, S. Sholes, and A. Royles for reading and editing the manuscript; I. Lee for engaging in discussion about epigenetics and chromatin state; and Circulomics for help in ultra-high-molecular-weight DNA extraction. **Funding:** This study was supported by the National Institutes of Health (grant R01HG009190 to W.T., grant F32 GM134558 to G.A.L., grants R24 DK106766-01A1 and U24HG010263 to M.C.S., grants 1R01HG011274-01 and 1U01HG010971 to K.H.M., and an Intramural Research Program grant from the National Human Genome Research Institute to S.K., A.R., and A.M.P.); a Damon Runyon Postdoctoral Fellowship and PEW Latin American Fellowship to G.V.C.; and a Howard Hughes Medical Institute Hanna Gray Fellowship to N.A. **Author contributions:** K.H.M. and W.T. conceived the study. W.T., A.M.P., K.H.M., and A.G. coordinated the collaboration. K.H.M. and N.A. performed repeat characterization and satellite DNA assembly. A.G., M.E.G.S., and M.C.S. performed ENCODE and mappability analyses. G.V.C. and N.A. performed CUT&RUN. N.A., A.R., and S.K. performed ONT mapping and methylation calling. S.J.H. and R.J.O. performed TE and noncentromeric repeat annotations. A.G. and A.S. performed gene annotation liftover. M.R.V., G.A.L., and E.E.E. performed SegDup annotations. A.R. performed marker-assisted mapping of CUT&RUN data. A.G., P.W.H, and R.R. performed HG002 cell culture and nanoNOMe sequencing and analysis. A.G., M.R.V., X.G., and E.E.E. performed phylogenetic and aging analysis of *NBPF* genes. M.J. performed megalodon methylation calling. A.G., W.T., and K.H.M. developed figures. A.G. and W.T. drafted the manuscript. All authors provided critical feedback and read and approved the final manuscript. **Competing interests:** W.T. has two patents (8,748,091 and 8,394,584) licensed to ONT. K.H.M. and W.T. have received travel funds to speak at symposia

organized by ONT. K.H.M. is a scientific advisory board member of Centaura, Inc. **Data and materials availability:** Nanopolish methylation calls are available on Zenodo (*79*). HG002 nanoNOMe data can be accessed on the Sequence Read Archive BioProject with accession number PRJNA725525. CUT&RUN data on CHM13 and HG002 can be accessed on the Sequence Read Archive with BioProject accession numbers PRJNA559484 and PRJNA752795. All other datasets used in this study are properly cited with accessions referenced in the methods and materials. CHM13hTERT cells were obtained for research use through a material transfer agreement with U. Surti and the University of Pittsburgh. Code for all CHM13 and HG002 CpG and GpC methylation, the ENCODE analysis pipeline, mappability analysis (MUK), and NBPF timing analysis code and alignments are available on Zenodo (*79*). SatFire figures are also available at Zenodo (*80*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj5089
Materials and Methods
Figs. S1 to S28
References (81–97)
Tables S1 to S9
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

26 May 2021; resubmitted 29 November 2021
Accepted 14 February 2022
10.1126/science.abj5089

Epigenetic patterns in a complete human genome

Ariel GershmanMichael E. G. SauriaXavi GuitartMitchell R. VollgerPaul W. HookSavannah J. HoytMiten JainAlaina ShumateRoham RazaghiSergey KorenNicolas AltemoseGina V. CaldasGlennis A. LogsdonArang RhieEvan E. EichlerMichael C. SchatzRachel J. O'NeillAdam M. PhillippyKaren H. MigaWinston Timp

Science, 376 (6588), eabj5089. • DOI: 10.1126/science.abj5089