Supplementary Materials for

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†]

[†] Correspondence to: khmiga@ucsc.edu (K.H.M), wtimp@jhu.edu (W.T.)

This PDF file includes:

Materials and Methods [References (81-98)]

Figs. S1 to S28

Tables S1 to S9 Legends

Other Supplementary Materials for this manuscript includes the following:

Tables S1 to S9 as a separate Excel file

MDAR Reproducibility Checklist

MATERIALS AND METHODS

1. Methylation Processing

1.1 CHM13 and HG002 Nanopore

Nanopore reads were obtained from (*13*, *14*, *72*). Ultra-long nanopore reads were aligned to the CHM13 reference (*4*) with Winnowmap-v2.0 (*73*) with a k-mer size of 15. BAM files were filtered for primary alignments with SAMtools (v1.9), analysis of centromeric regions was done on reads >50kb. To measure CpG methylation in nanopore data we used Nanopolish (v0.13.2) (*34*). Nanopolish uses a Hidden Markov model on the nanopore current signal to distinguish 5mC from unmethylated cytosine. The methylation caller generates a log-likelihood value for the ratio of probability of methylated to unmethylated CGs at a specific k-mer. We filtered methylation calls using the nanopore_methylation_utilities tool (https://github.com/timplab/nanopore-methylation-utilities) (*79*), which uses a log-likelihood ratio of 1.5 as a threshold for calling methylation. CpG sites with log-likelihood ratios greater than 1.5 (methylated) or less than −1.5 (unmethylated) were considered high quality and included in the analysis. Reads that did not have any high-quality CpG sites were excluded from the subsequent methylation analysis. Nanopore_methylation_utilities integrates methylation information into the alignment BAM file for viewing in the bisulfite mode of Integrative Genomics Viewer (IGV) and creates Bismark-style files (*74*). Methylation data was plotted by binning the genome with the BSgenome R package (BSgenome_1.56.0) and taking the average of CG sites within each bin. Single-read plots were generated with the ggplot2 R package (ggplot2_3.3.3) using the single-read data in the tabix indexed single-read methylation bed files generated from nanopore_methylation_utilities, code for generating all figures is available at available on github (https://github.com/timplab/T2T-Epigenetics) and zenodo (*79*).

1.2 HG002 Bisulfite

Bisulfite FASTQs were collected from the an AWS open data set generated by ONT s3://ont-open-data/gm24385_mod_2021.09/ described here (https://labs.epi2me.io/gm24385-5mc). Paired-end FASTQs were aligned with Bismark (v0.22.2) (https://github.com/FelixKrueger/Bismark) (*74*) with default parameters to a reference comprised of CHM13 autosomes (chromosomes 1-22), HG002 T2T chromosome X (*4*) and GRCh38

chromosome Y using the “bismark” command with the key parameters “-p --bam --bowtie”. The reference genome was prepared by the Bismark command “bismark_genome_preparation” with default parameters. Methylation data was extracted using the Bismark command “bismark_methylation_extractor” with the following parameters: “-p --comprehensive --merge_non_CpG --bedGraph --gzip --remove_spaces --cytosine_report”. CpG methylation frequency for methylation map plot was generated by calculating the fraction of methylated reads to total coverage from the bismark CpG coverage bed file within bins in HG002 chromosome X with the BSGenome Bioconductor package (https://bioconductor.org/packages/BSgenome). Multiples of three bins were further smoothed with the “rollmean” function from the R package Zoo (https://cran.r-project.org/web/packages/zoo/index.html) (*81*).

1.3 HG002 NanoNOMe

HG002 nanoNOMe reads were aligned with with Winnowmap-v2.0 (*73*) with a k-mer size of 15 to both the T2T-CHM13+GRCh38 chromosome Y reference for whole genome analyses and CHM13 chromosomes 1-22+HG002 chromosome X+GRCh38 chromosome Y for HG002 chromosome X analyses. BAM files were filtered for primary alignments with SAM flag -F 256 and filtered for read lengths greater than 20kb. To measure CpG and GpC methylation in nanopore data we used Nanopolish (v0.13.2) 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. Reads that did not have any high-quality sites were excluded from the subsequent methylation analysis. Nanopore_methylation_utilities integrates methylation information into the alignment BAM file for viewing in the bisulfite mode in IGV and also creates Bismark-style files.

In order to choose the optimal bin size for accessibility analysis, we used the intrinsic smoothness test previously explained by (*82*). 15kb bins chosen by the intrinsic smoothness test were investigated for possible bias in CG and GC coverage **(fig. S12 A to C)**. To visualize the methylation and accessibility patterns, the bins were z-normalized across each chromosome. Nucleosome footprints were determined by counting the number of consecutive unlabeled GpC sites, or “Inaccessible Runs” (*16*).

For analysis of GpC accessibility we followed all methods outlined in (*16*). Briefly, we estimated profiles of measurements by fitting locally weighted generalized linear models across the genome for each sample as implemented by Bioconductor package bsseq v.1.20.0 (*83*). For GpC methylation, the minimum window was reduced to 100 bp and the number of sites to ten to account for rapid fluctuations in the accessibility profile due to nucleosome positioning. For visualization, we plotted the z-score of smoothed GpC methylation.

To find regions of high accessibility, continuous regions having smoothed accessibility greater than 99th percentile of the data were selected first. The significance of each accessible region was determined by performing a binomial test of the raw GpC methylation frequency, with overall accessibility frequency as the null probability. The probabilities were corrected for multiple testing using the Benjamini–Hochberg correction, and accessible regions with adjusted *p* values less than 0.01 and widths greater than 50 bps were determined to be accessibility peaks. Peaks called within 500bp of a GpC site that had anomalous coverage, i.e. coverage outside of the 5th-95th percentile, were removed from the repeat analysis. Biological replicates were processed individually and peak calls were merged across all three replicates.

1.4 Reduced representation bisulfite sequencing (RRBS):

RRBS raw sequencing data from early human embryos from (*35*), was obtained from SRA (GEO accession: GSE49828) with fastq-dump (v2.8.0, http://ncbi.github.io/sra-tools/). RRBS reads were trimmed with TrimGalore (v0.6.6, https://github.com/FelixKrueger/TrimGalore) using the “--rrbs” and “--paired” parameters. The reads were aligned with Bismark (v0.22.2) (*74*) using the “bismark” command with the key parameters “-p --bam --bowtie” to the CHM13 reference genome prepared by the Bismark command “bismark_genome_preparation” with default parameters. Methylation data was extracted using the Bismark command “bismark_methylation_extractor” with the following parameters: “-p --comprehensive --merge_non_CpG --bedGraph --gzip --remove_spaces --cytosine_report”. Methylation calls from technical replicates of the same biological replicate were combined by using the Bismark command “bismark2bedGraph” with the following parameters: “--buffer_size 20G --remove_spaces”. RRBS and CHM13 methylation data were imported into R using the “read.bismark” command from the “bsseq” package (v1.24.4) (*83*) using only CHM13 reference CpGs with the following parameters “strandCollapse = TRUE, rmZeroCov = FALSE”**.** CpG loci were retained if they were covered by at least one read in 90% of the samples analyzed.

Percent methylation was used to compare the samples by Euclidean distance with the R function “dist” with default parameters. Samples were clustered using the R function “hclust” using the “ward.D” method. The dendrogram (**fig. S11**) was plotted using “ggdendrogram” with default parameters. Full pipeline is available at https://github.com/timplab/T2T-Epigenetics/tree/main/rrbs_t2t and zenodo (*79*).

1.5 Methylation clustering:

Methylation clustering across the CHM13 X chromosome was performed on all CpG islands (CGI) 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 CGI and were longer than 5kb. Clustered reads were then intersected with known escape and XCI genes from (*51*). The same clustering procedure was performed at the DXZ4 locus.

1.6 Megalodon Methylation Calling

Megalodon (version 2.3.4) was run with the r9.4.1_450bps 5mC model with thresholding set as default.

2. NanoNOMe Library Preparation

2.1 HG002 Cell Culture:

NA24385 cells (HG002) were obtained from the Coriell Institute (https://www.coriell.org/). Cells were grown in T-25 flasks in RPMI 1640 media with L-glutamine (Gibco; 11875093) supplemented with 15% fetal bovine serum (Gibco; 26140079) and 1% penicillin-streptomycin (Gibco; 15140122). Cells were cultured at 37C with 5% CO_2 and were maintained by passaging ~⅓ into fresh media every three days. Cells tested negative for mycoplasma contamination with the LookOut Mycoplasma PCR Detection Kit (Sigma; MP0035) and Jumpstart Taq DNA Polymerase (Sigma; D9307). Cells at passage 11 were used in nanoNOMe sequencing.

2.2 NanoNOMe HG002 Sequencing:

NanoNOMe library preparation was performed according to the methods outlined in (*16*). Cells were collected by resuspension, then nuclei were extracted by incubating in resuspension buffer

(100 mM Tris-Cl, pH 7.4, 100 mM NaCl, 30 mM $MgCl_2$) with 0.25% NP-40 for 5 min on ice. Intact nuclei were collected by centrifugation for 5 min at 500*g* at 4 °C. Nuclei were subjected to a methylation labeling reaction using a solution of 1x M.CviPI Reaction Buffer (NEB), 300 mM sucrose, 96 µM *S*-adenosylmethionine (NEB) and 200 U of M.CviPI (NEB) in 500 µl volume per 500,000 nuclei. The reaction mixture was incubated at 37C with shaking on a thermomixer at 1,000 RPM for 15 min. The reaction was stopped by the addition of an equal volume of stop solution (20 mM Tris-Cl, pH 7.9, 600 mM NaCl, 1% SDS, 10 mM disodium EDTA). Samples were treated with Proteinase K (NEB) at 55C for >2 h and DNA was extracted via phenol:chloroform extraction and ethanol precipitation. After ethanol precipitation we enriched for HMW DNA with Circulomics Short Read Eliminator Extra Long (SRE-XL; SS-100-111-01).

Purified gDNA was prepared for nanopore sequencing following the protocol in the genomic sequencing by ligation kit LSK-SQK109 (ONT). Each nanopore library was prepared with 2 µg of input DNA. Fifteen libraries were generated and run on five PromethION flow cells with reloading (three libraries used per flow cell). Flow cells were flushed at 24 hours and 48 hours with the Oxford Nanopore's Flow Cell Wash kit (EXP-WSH003) and reloaded with fresh library. Sequencing runs ran for a total of 72 hours and were simultaneously basecalled with Guppy 4.0.11. All nanoNOMe data can be accessed at on Sequence Read Archive with BioProject Accession number PRJNA725525. Data was collected as three distinct biological replicates.

3. CUT&RUN

3.1 Library Generation

CUT&RUN was carried out as in (*75*), with some variations. Frozen pellets of 1.6 million HG002 cells or 1.2 million CHM13 cells were thawed on ice and centrifuged at 500x*g* for 5 minutes at 4C. Cells were washed with cold PBS twice. For nuclear extraction, each cell pellet was resuspended in 500 uL of Nuclear Extraction Buffer (NEB, 20 mM HEPES pH 7.9, 10 mM KCl, 0.5mM Spermidine, 0.1% NP40, 20% glycerol, Roche Proteinase Inhibitor tablets) by pipetting gently, and incubated on ice for 5 minutes. Cells were centrifuged and washed with Washing Buffer (WB, 20mM HEPES pH 7.5, 150mM NaCl, 0.5mM Spermidine, 0.1% BSA, 0.05% NP40, Roche Proteinase Inhibitor tablets), blocked in WB containing BSA, and incubated in primary antibody for 2h at 4C under rotation. Primary antibodies used were: mouse CENP-A (Abcam,

ab13939), rabbit CENP-B (Abcam, ab25734), rabbit H3K4me2 (Abcam, ab7766), and rabbit H3K27me3 (ThermoFisher, 39155).

Cells were washed twice with WB and incubated with pAG-MNase (Cell Signaling) for 1h at 4C under rotation. For pAG-MNase digestion, samples were incubated on ice for 10 minutes, then $CaCl_2$ was added to activate MNase to a final concentration of 2mM. Samples were then incubated for 30 minutes at 0C. To stop digestion, an equal volume of 2X STOP solution (200 mM NaCl, 20 mM EDTA, 4 mM EGTA, 0.1% NP40) was added. To recover low-salt fragments, samples were incubated for 1h at 4C under rotation, centrifuged at 500x*g* for 5 minutes and supernatant collected and labeled as low-salt fraction. RNase A was added following incubation for 20 minutes at 37C. Samples were treated with Proteinase K for 1h at 65C (or overnight), and DNA extraction was carried out with MasterPure Complete DNA Isolation kit (Lucigen) as indicated by the manufacturer. Samples were analyzed by a Fragment Analyzer.

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 was added as a control (obtained from the Henikoff lab). Libraries were purified using AMPure XP beads and the PCR reaction was carried out using NEBNext Ultra II Q5 master mix and NEBNext multiplex oligos for Illumina (12 cycles with annealing/extension for 15 seconds at 65C). Final libraries were purified using AMPure XP beads. Libraries were sequenced using NovaSeq 50PE sequencing.

3.2 Marker-assisted mapping:

Marker-assisted mapping of CUT&RUN data (CHM13 CENP-A, CHM13 H3K4me2, CHM13 H3K27me3, HG002 CENP-A, HG002 CENP-B) to the same genome (CHM13 to T2T-CHM13 or HG002 to CHM13 autosomes (chromosomes 1-22), HG002 T2T chromosome X and GRCh38 chromosome Y) was performed according to the methods outlined in (*56*). In brief, 150 bp paired-end CUT&RUN libraries were mapped with bwa-mem (*84*) and filtered with SAMtools (*77*) for unique 51-mers as follows:

```
bwa mem -k 50 -c 1000000
overlapSelect -overlapBases=51
```

Unique 51-mers were generated with Meryl software (*85*). This method differs from dynamic k-mer assisted mapping in that there is a set k-mer size of k=51 and indels are not accounted

for. More stringency can be placed on reads generated from a sample with its own genome (e.g. CHM13) as is the case with the CUT&RUN data and not with the ENCODE datasets.

4. ENCODE

4.1 ENCODE Dynamic k-mer assisted mapping:

We selected several ChIP-seq datasets generated as part of the ENCODE project (*1*) choosing ChIP-seq samples with at least 100 bp paired-end sequencing data and at least one matching input control (Methods). These criteria yielded 96 total sequencing libraries (**table S9**). All ENCODE-generated raw FASTQ files were downloaded from the ENCODE data portal (*86*). Prior to mapping, reads originating from a single library were combined. Reads were mapped with Bowtie2 (v2.4.1) (*76*) as paired-end with the arguments "--no-discordant --no-mixed --very-sensitive --no-unal --omit-sec-seq --xeq --reorder" **(fig. S1A)**. Alignments were filtered using SAMtools (v1.10) (*77*) using the arguments "-F 1804 -f 2 -q 2" to remove unmapped or single end mapped reads and those with a mapping quality score less than 2. PCR duplicates were identified and removed with the Picard tools "mark duplicates" command (v2.22.1, http://broadinstitute.github.io/picard) and the arguments "VALIDATION_STRINGENCY=LENIENT ASSUME_SORT_ORDER=queryname REMOVE_DUPLICATES = true".

Alignments were then filtered for the presence of unique k-mers. Specifically, for each alignment, reference sequences aligned with template ends were compared to a database of k-mers unique in the whole genome. The size of the k-mers in the k-mer filtering step are dependent on the length of the mapped reference sequence. We generated k-mer databases for 50-100mers by multiples of five. If all 100bp of a read map then the 100mer database is used. However, if the reference sequence is longer or shorter than 100bp then we use the database that is shorter than the reference length to the nearest five. For example, when there is a 1bp insertion in the read compared to reference, the corresponding reference sequence is 99bp, therefore the 95-mer database is used **(fig. S1B)**. When there is a 1bp deletion in the read compared to the reference, the corresponding reference sequence is 101 bp long so we use the 100-mer database. Mismatches only impact the k-mer size if they occur in the first or last positions, otherwise reference sequence length is unchanged. Alignments were discarded if no unique k-mers occurred in either end of the read. k-mer databases were generated using KMC3

(v3.1.1)(*87*). Alignments from replicates were then pooled. Bigwig genome tracks were created using deepTools bamCoverage (v3.4.3) (*88*) with a bin size of 1bp and default for all other parameters. Across all cell lines and marks, initial Bowtie2 alignments only yielded a 0.6% average increase in aligned reads to CHM13 versus GRCh38p13, however after intersecting the alignments with unique variable-length k-mers*,* the percentage of new alignments increased to 2.33% on average **(table S1).**

Mapping differences were broken down into five categories: satellites, LINEs/SINEs, (SDs), all other repetitive element types, and non-repetitive sequence. The segmental duplication regions were obtained from (*39*). The remaining repeat categories were obtained from (*44*). LINE and SINE regions were extracted from the repeat masker annotation, merged so there were no overlapping elements, and intersected with the segmental duplication intervals to remove overlaps. The satellite regions were obtained the same way with the addition of intersecting them with both the segmental duplication and LINE/SINE intervals. The remaining repeat intervals were defined as all repeat masker annotations, merged and intersected with the previous three interval sets. The non-repetitive regions were defined as all intervals not covered by one of the above four tracks.

Peak calls were made using MACS2 (v2.2.7.1) (*78*) with default parameters and estimated genome sizes 3.03×10^9 and 2.79×10^9 for chm13v1 and GRCh38p13, respectively. GRCh38p13 peak calls were lifted over to chm13v1 using the UCSC liftOver utility, the chain file created by the T2T consortium, and the parameter “-minMatch=0.2”. Peak intersections were determined using bedtools (v2.26.0) (*89*) counting each liftOver peak only once if any intersection occurred. Peaks unique to CHM13 were generated by taking all non-intersecting peak calls from the chm13v1 peak calls and the GRCh38p13 liftover peak calls using bedtools intersect.

5. T2T-CHM13 Genomic Annotations

5.1 Repeat-Masking:

RepeatMasker annotations were generated in (*44*), and are available for both T2T-CHM13 and GRCh38 on the T2T UCSC genome assembly hub (http://t2t.gi.ucsc.edu/chm13/hub/hub.txt).

5.2 CENP-B motif annotation:

CENP-B sites in HG002 chromosome X were identified with fuzznuc software tool from EMBOSS (*90*) searching for the CENP-B consensus sequence as follows:

```
fuzznuc --sequence HG002_chromsomeX.fasta \
     --pattern NTTCGNNNNANNCGGGN -complement
```

5.3 Centromere annotations:

Centromere region repeat annotations are described in (*56*), and are available on the T2T UCSC genome assembly hub (http://t2t.gi.ucsc.edu/chm13/hub/hub.txt).

5.4 CDR Annotations

CDRs in T2T-CHM13 were manually annotated by labeling the entire span where CpG methylation is irregular, or lower than the flanking active array (which exhibits high, regular CpG methylation).

5.5 Genome Mappability

Minimum unique k-mer (MUK) lengths were calculated for T2T-CHM13. MUKs represent the minimum distance from a position in the genome needed to identify a unique sequence, either upstream (right-anchored) or downstream (left-anchored). To compute these values, all chromosomes of T2T-CHM13 were concatenated, followed by their reverse complements. A suffix array (SA) and longest common prefix (LCP) table were calculated from this single sequence using the algorithm and implementation adapted from Sapling (*91*). For each position in the SA, the LCP value plus one represents the MUK at that position. If the SA value indicated that the sequence was a reverse complement, the unique sequence was right-anchored; otherwise, the sequence was left-anchored. Any minimum unique sequences containing an N or overlapping a chromosome end were removed and those positions were marked as having no minimum unique sequence.

Mappability for 200 bp fragments was calculated by counting the fraction of 200-mers overlapping a given genomic position that have an MUK less than or equal to 200 for the first or last base in the fragment. All scripts for recapitulating this analysis can be found at https://github.com/msauria/T2T_MUK_Analysis and zenodo (*79*).

6. Gene expression and annotation

6.1 *NBPF* Analysis

CHM13 *NBPF* sequences were first mapped to the non-human primate assemblies and sequences less than 12.5Kb were removed to retain exclusively the VNTR sequence. The initial paralog used for mapping was *NBPF25P* (chr1:144,688,831-144,708,527). Next, a sequence alignment with mafftv7.453 (*92*) was performed with the following command:

```
mafft --reorder --maxiterate 1000 --thread 15 \
     ${T2T_NHP_NBPF.fa} > MSA.fa
```

For evolutionary timing estimates, extra copies that did not contribute to the topology of the tree or had obvious issues with alignment were removed. Lastly, the MSA was trimmed for all bases which had gaps in >=60% of sequences.

The phylogeny was generated by maximum likelihood with the following command:

```
iqtree2 -nt AUTO -m MFP -s ${trimmed_MSA-fasta} \
-o "{macaque_seq_outgroup}" \
--prefix "{NBPF_timing_estimate}" --redo-tree -alrt 1000 -b 100
```

Bayesian estimates with were performed with BEAST2 (*93*) using the trimmed MSA to estimate mutation rate at *NBPF* using two priors: macaque-human divergence of 25MYA +/- 2MYA, and chimp-bonobo divergence of 1.15 MYA +/- 0.3 MYA (*94*, *95*). Estimates of *NBPF* expansions were given with the 95% confidence interval of mutation rate.

6.2 PacBio Iso-seq

PacBio Iso-Seq on CHM13 (*13*) (accessions: SRX9009500, SRX9009501) was processed according to methods outlined in (*39*). In short, data was aligned with minimap2 v2.17 with the following command:

```
minimap2 -H -ax splice -uf -C5 --secondary=no --eqx
```

And filtered for primary alignments with:

```
samtools -F 2308
```

Iso-Seq coverage at *NBPF* loci was quantified with bedtools intersect.

6.3 HG002 RNA-seq

HG002 RNA-seq (accession: SRR13086640) was aligned with HISAT2 (*96*) and gene expression was quantified using StringTie2 (*97*) with the following command:

```
stringtie {input.bam} -G {gtf} --conservative -o {output.gtf} \
     -p 60 -B -e -A
```

6.4 Previously Unresolved Gene Annotations

Gene annotations were obtained from (*4*). Previously unannotated genes were extracted using the full T2T-CHM13-v1.1 gene annotation and extracting all gene IDs labeled with 'extra_paralog=True', then filtering the v1.0 gff file for these genes.

6.5 PRO-seq Data Analysis

PRO-seq data was generated and analyzed in (*44*).

Supplementary Figures:

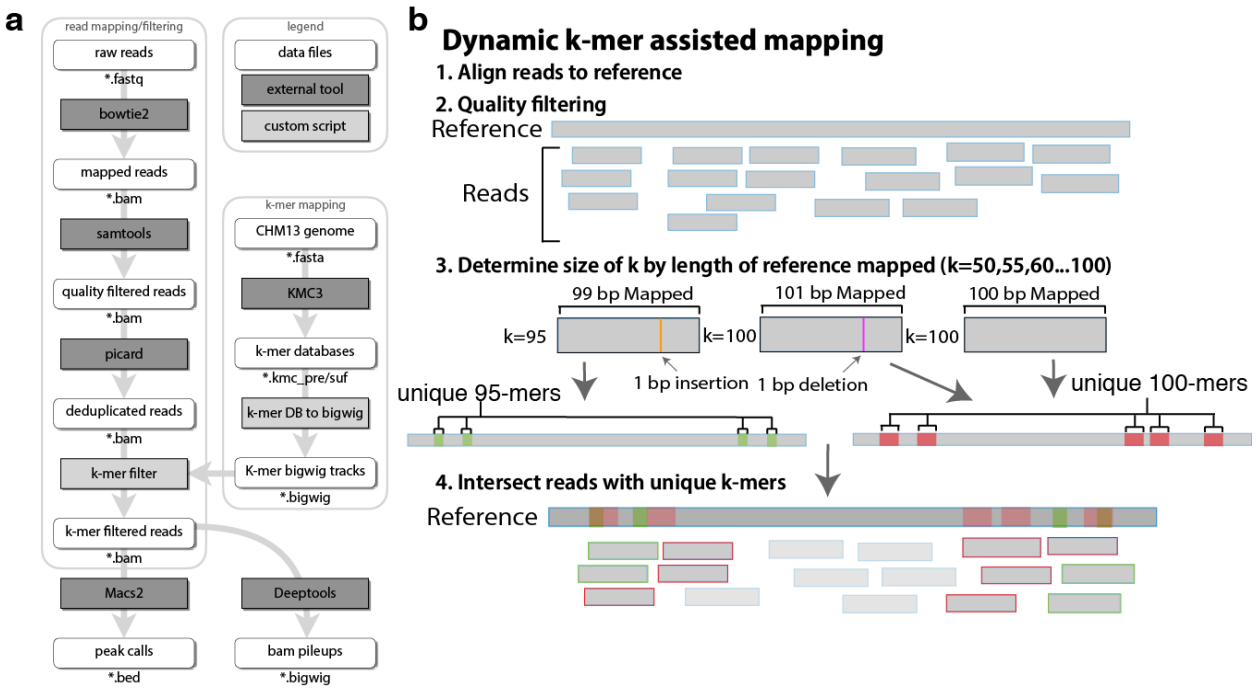

Figure S1. Overview of ENCODE dynamic k-mer mapping pipeline. a) *Bioinformatic pipeline for mapping ENCODE data to repetitive regions of the T2T-CHM13 genome.* ***b)*** *Schematic description of the dynamic k-mer assisted mapping pipeline.*

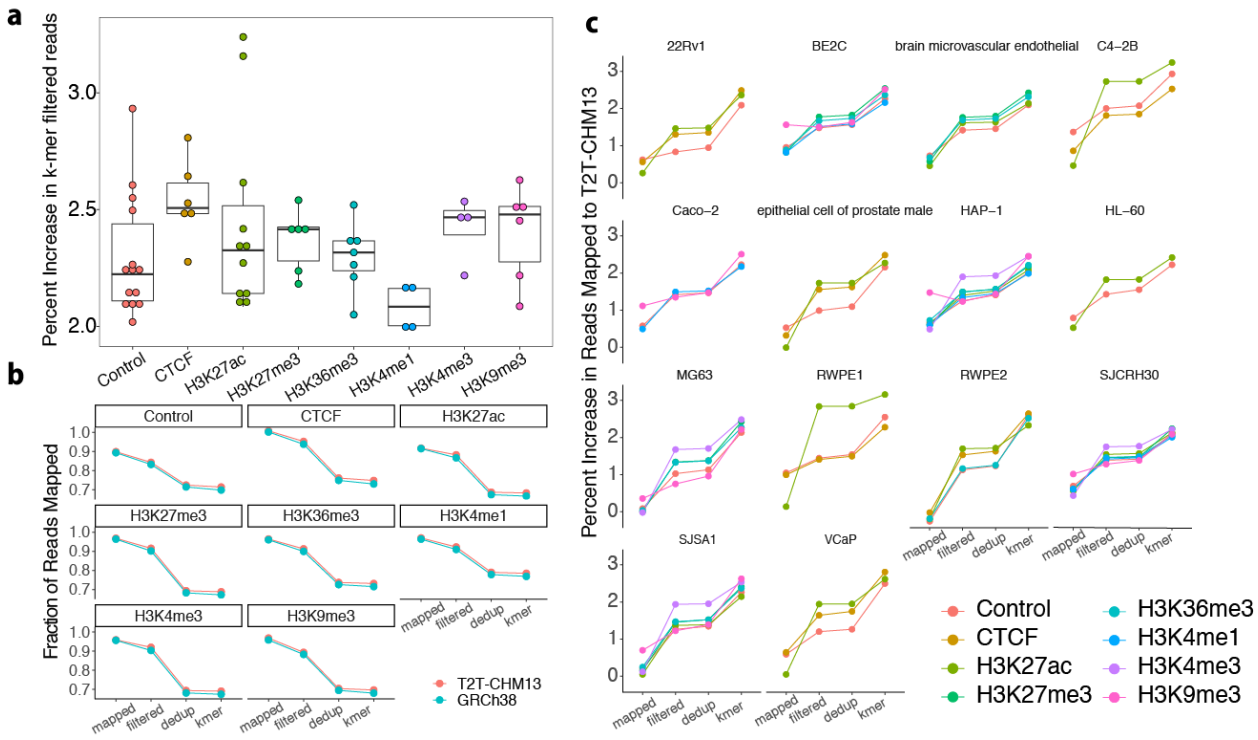

Figure S2. ENCODE mapping summary. a) *Relative percent increase in k-mer filtered aligned reads in T2T-CHM13 compared to GRCh38 for all datasets surveyed. Each point represents an ENCODE sample. Control samples are ChIP-seq matched input control for each library.* ***b)*** *Fraction of total aligned reads retained during each mapping step in the pipeline. Control samples are ChIP-seq matched input control for each library.* ***c)*** *Percent increase in alignments to T2T-CHM13 compared to GRCh38 at each mapping step separated by cell line. Control samples are ChIP-seq matched input control for each library.*

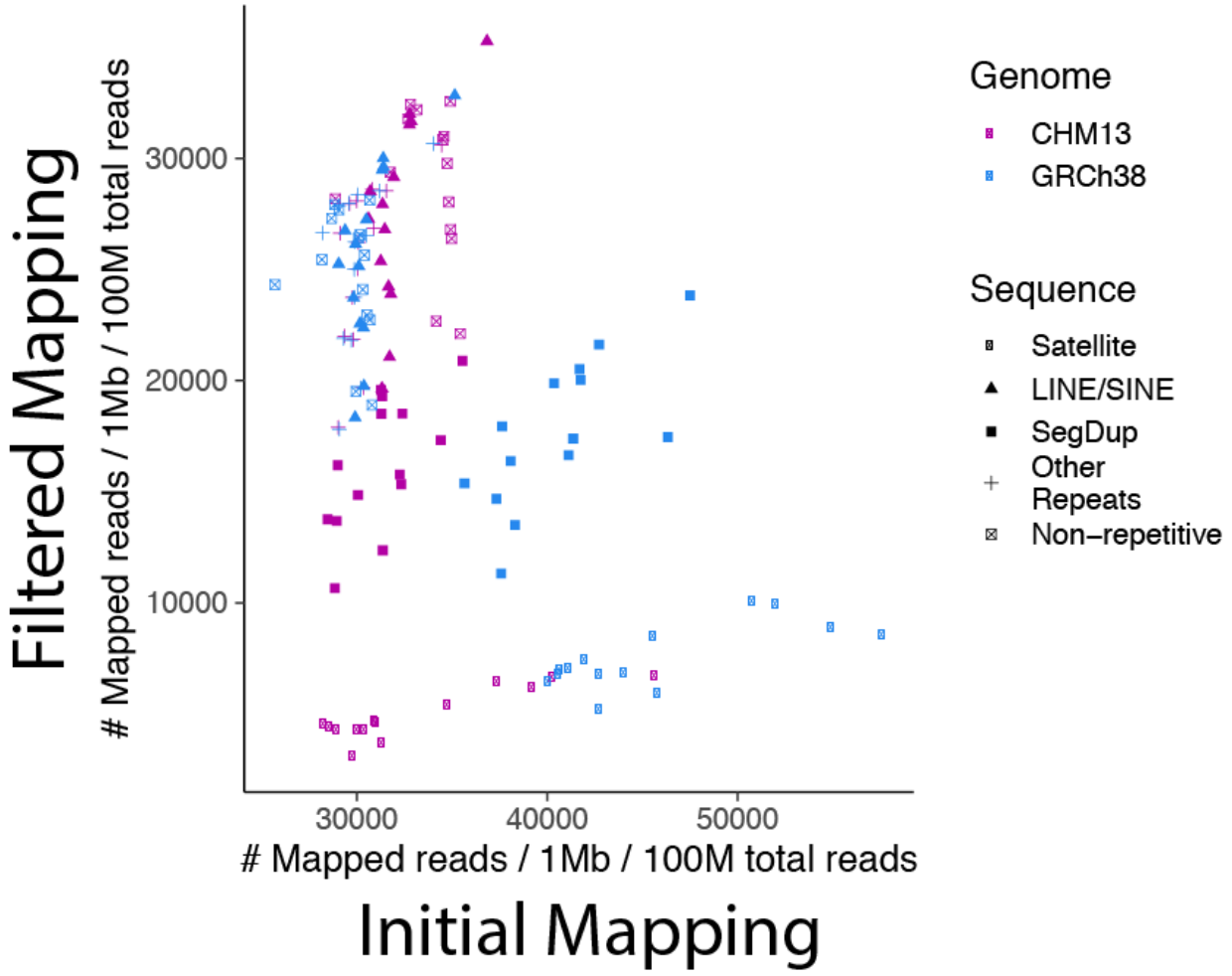

Figure S3. Dynamic k-mer mapping read filtering. *Normalized number of mapped reads filtered out of input control ChIP-seq library (x-axis) versus normalized number of mapped reads filtered out of input control ChIP-seq library after quality filtering, deduplication and k-mer filtering (y-axis) for each ENCODE cell line profiled. Different genomic regions are denoted by different shape markers, reference genome by color of marker. Number of reads is normalized to a post alignment read depth of 100M reads per 1Mbp of sequence.*

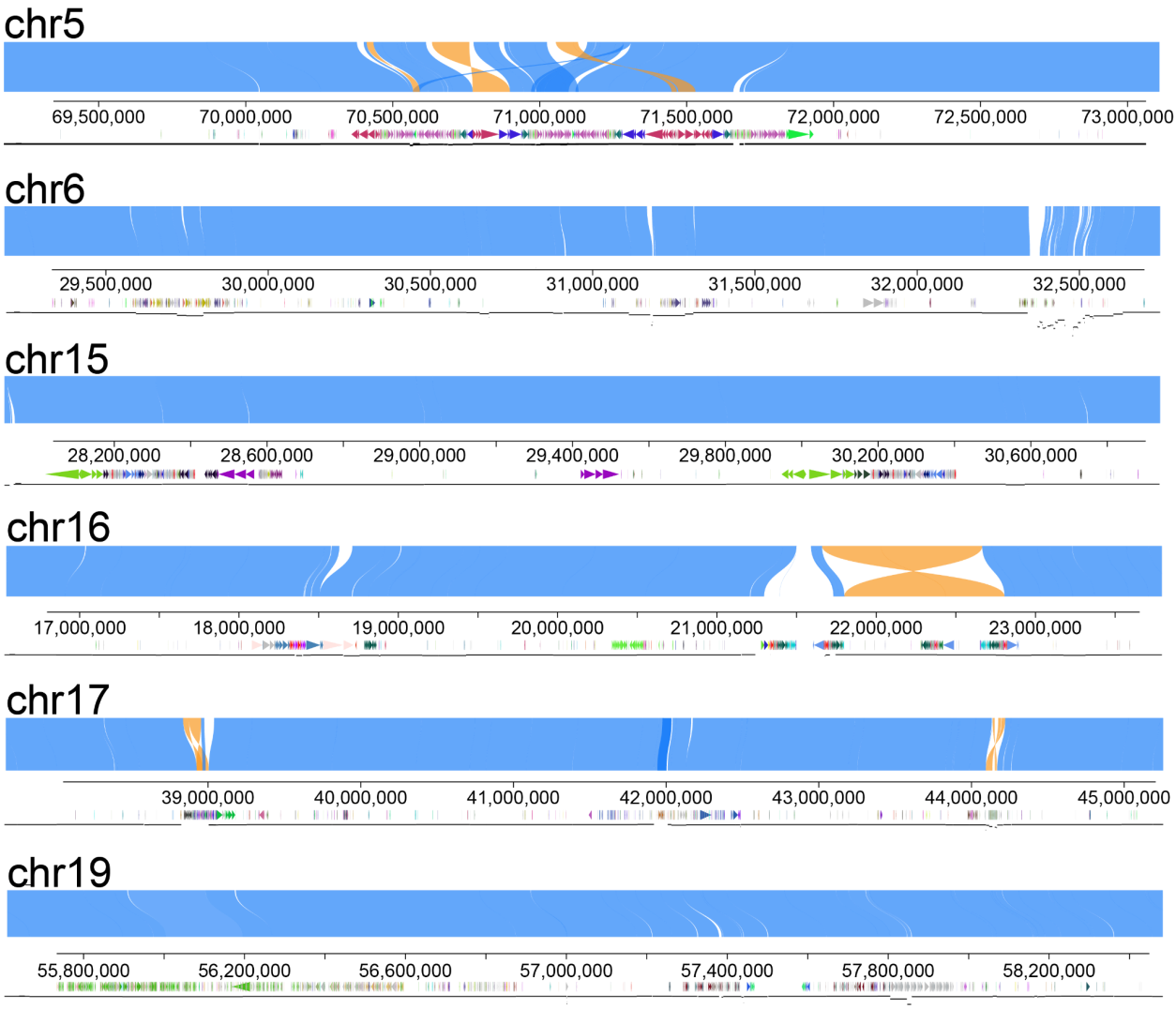

Figure S4. GRCh38 to T2T-CHM13 alignments at ENCODE enriched loci. *Alignments of GRCh38 (top) to T2T-CHM13 (bottom) illustrate structural variation between the two assemblies at regions designated with an orange triangle in Figure 1B. Sequence gaps are blank space (white), inversions are highlighted in yellow. Colored arrows are from the dupMasker track and identity plotted below as a line plot. Figures were produced with https://mrvollger.github.io/SafFire/ also available on zenodo (80).*

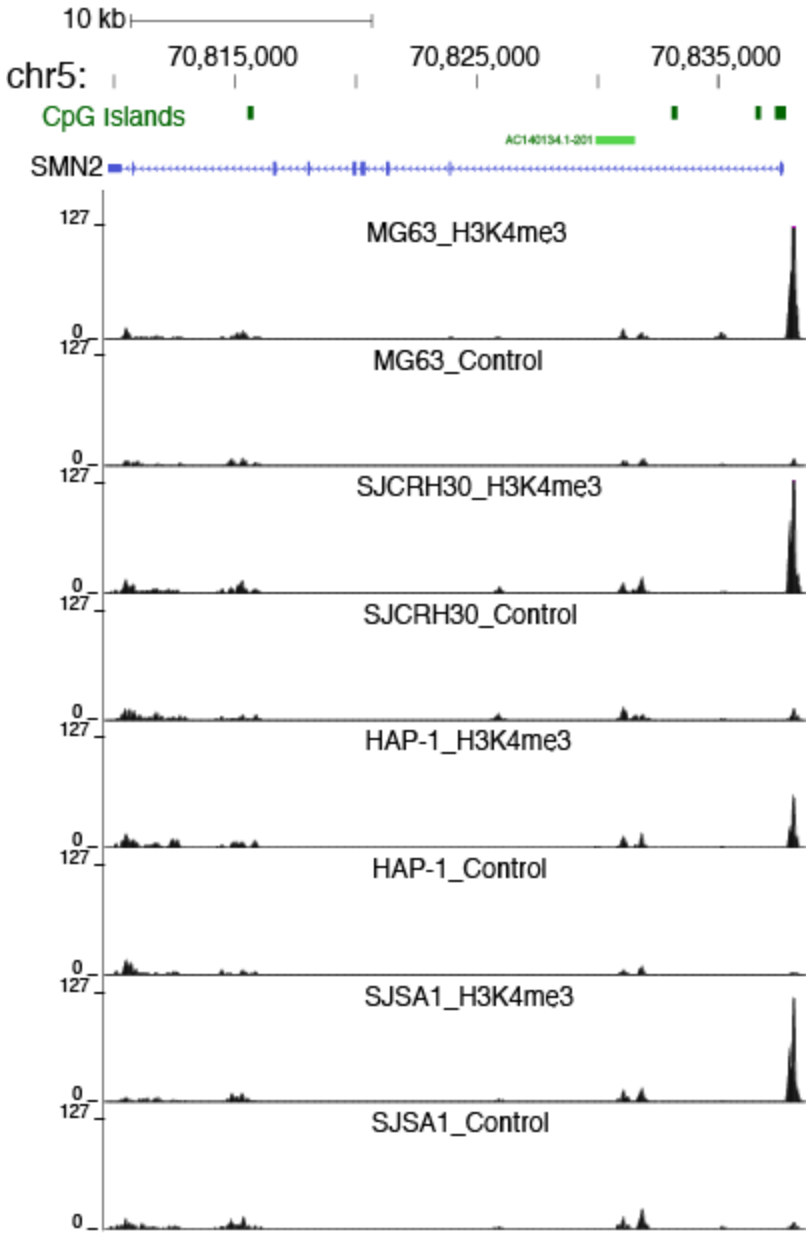

Figure S5. H3K4me3 peaks at SMN2. *ENCODE dynamic k-mer assisted alignments of ChIP-seq H3K4me3 marks from different cell lines to the SMN2 locus.*

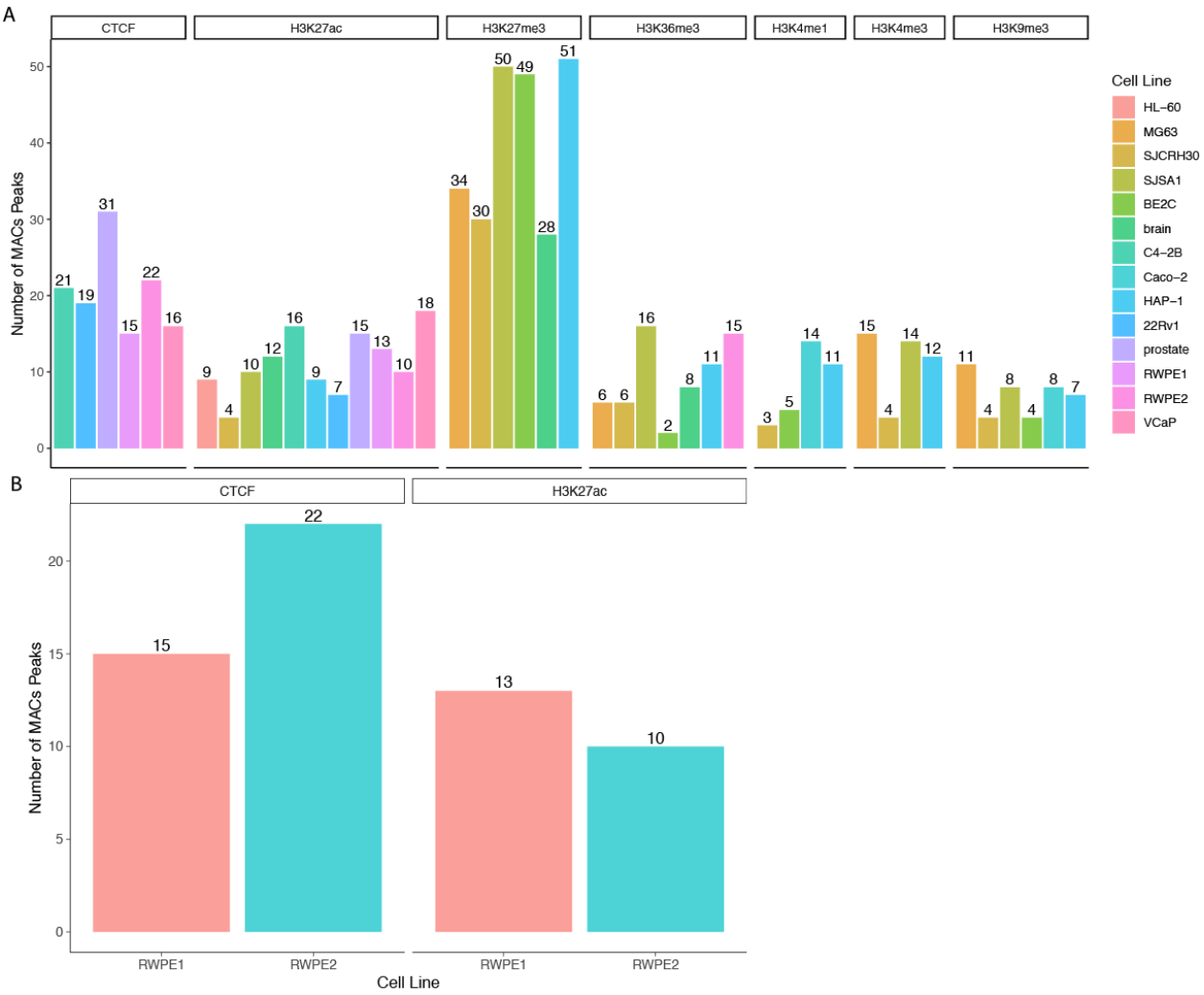

Figure S6. ENCODE peaks at HLA genes. *A) Total peak calls for all ENCODE samples at the HLA locus. B) Peak calls for CTCF and H3K27ac in the RWPE1 and RWPE2 cell lines at the HLA locus.*

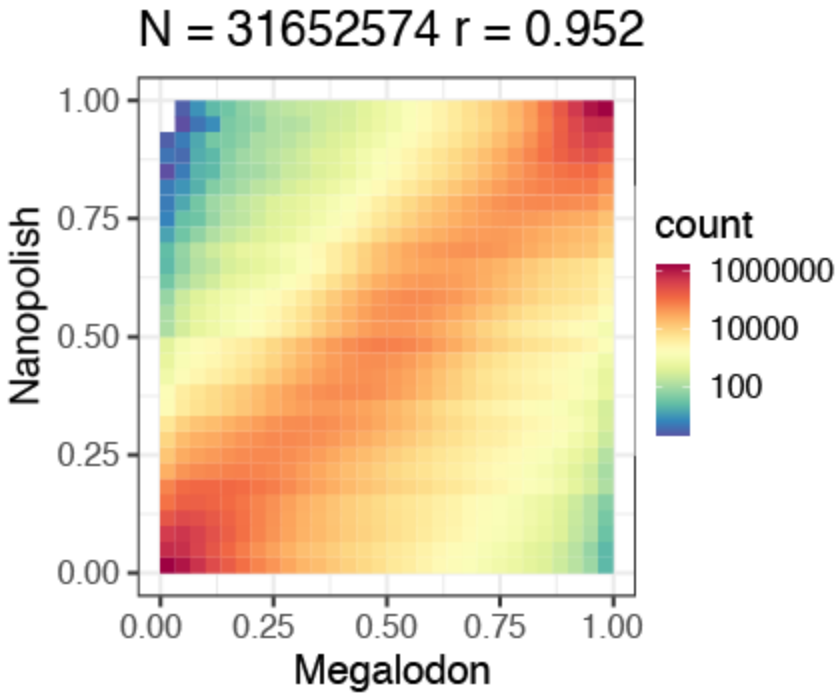

Figure S7. Nanopolish to Megalodon Comparison. *CHM13 CpG methylation frequency correlation plot between Nanopolish and Megalodon.*

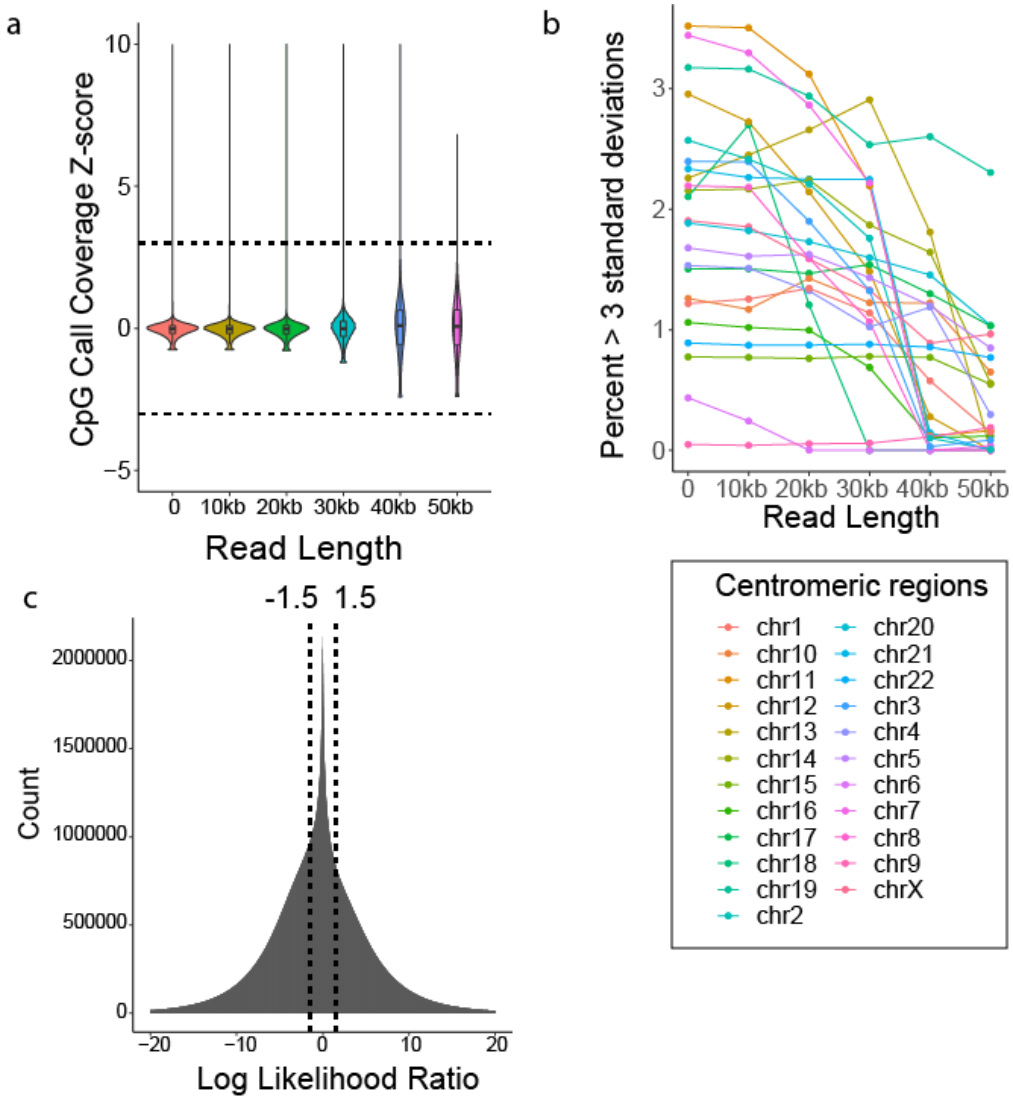

Figure S8. Thresholds for long-read nanopore methylation. a-b) *Z-score of CpG coverage as compared against the whole genome. With increasing read length the percentage of CpGs with coverage Z-scores greater than 3 and less than -3 decreased. >50kb reads were chosen to decrease coverage bias in centromeric regions, allowing for robust analysis of methylation calls through centromeric repeats.* ***c)*** *The distribution of all log-likelihood scores for methylation calls in CHM13. Calls with log likelihood ratios greater than 1.5 and less than -1.5 were considered high quality and used for all subsequent analysis.*

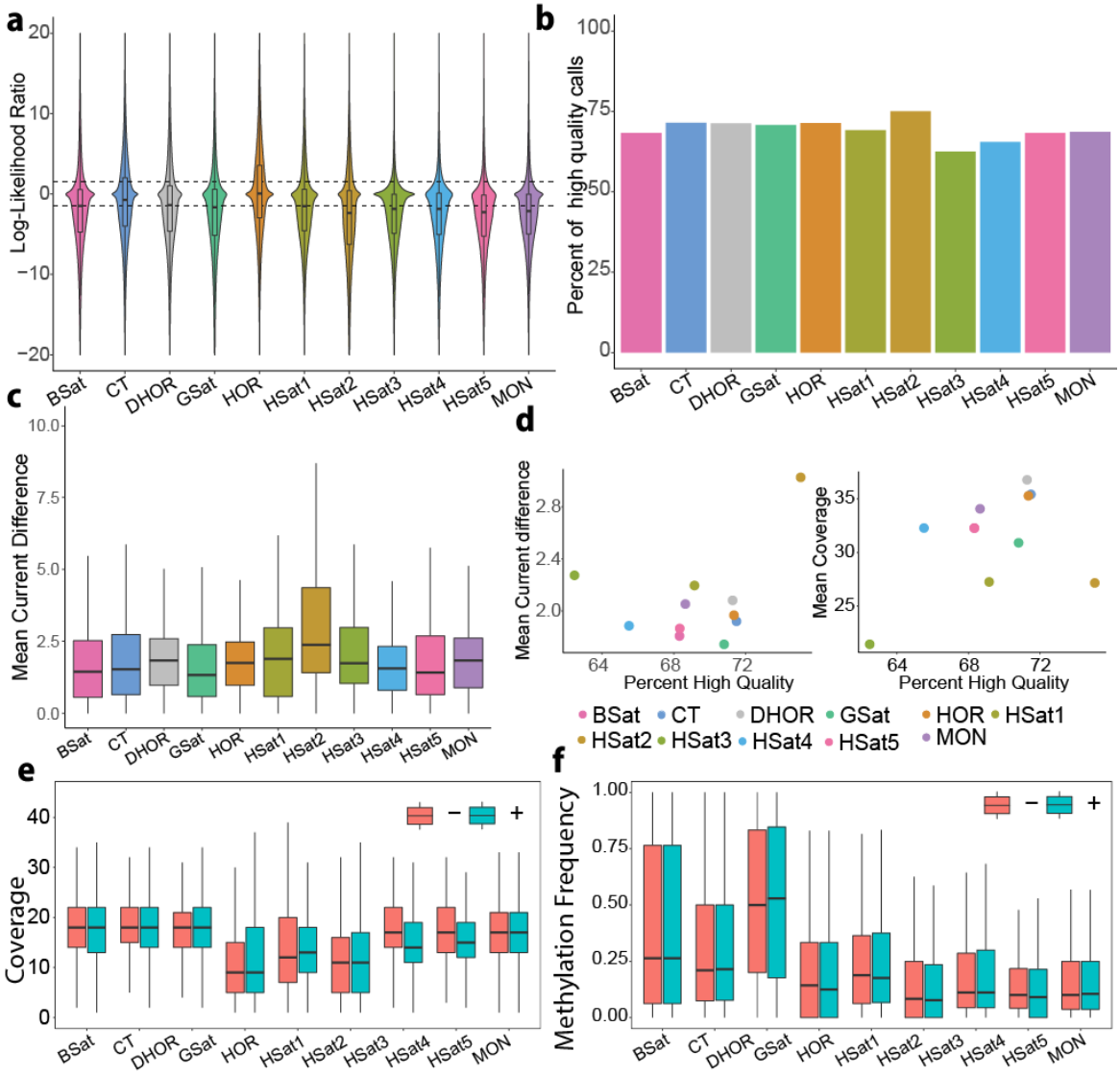

Figure S9. CHM13 CpG methylation quality control. ***a)*** *Distribution of log-likelihood ratios for each repeat type from nanopore reads filtered to only primary alignments >50kb.* ***b)*** *Percentage of high quality (|log likelihood| > 1.5) CpG calls within each satellite repeat.* ***c)*** *Distribution of the mean current difference between methylated and unmethylated k-mers possible in each repeat. For all CpG containing 6-mers per repeat the absolute value of the mean difference in methylated vs unmethylated current distribution was calculated. Boxplots are weighted by kmer frequency.* ***d)*** *(Left) Scatter plot of the percentage of high quality calls versus the mean current difference of methylated vs unmethylated 6-mers per repeat type weighted by kmer frequency, (Pearson Correlation, r = 0.36, p = 0.28). (Right) Scatter plot of the percentage of high quality methylation calls versus CpG coverage per repeat, (Pearson Correlation, r = 0.42, p = 0.19).* ***e)*** *Called CpG site coverage per read strand within each repeat type.* ***f)*** *Average CpG methylation frequency per read strand within each repeat type.*

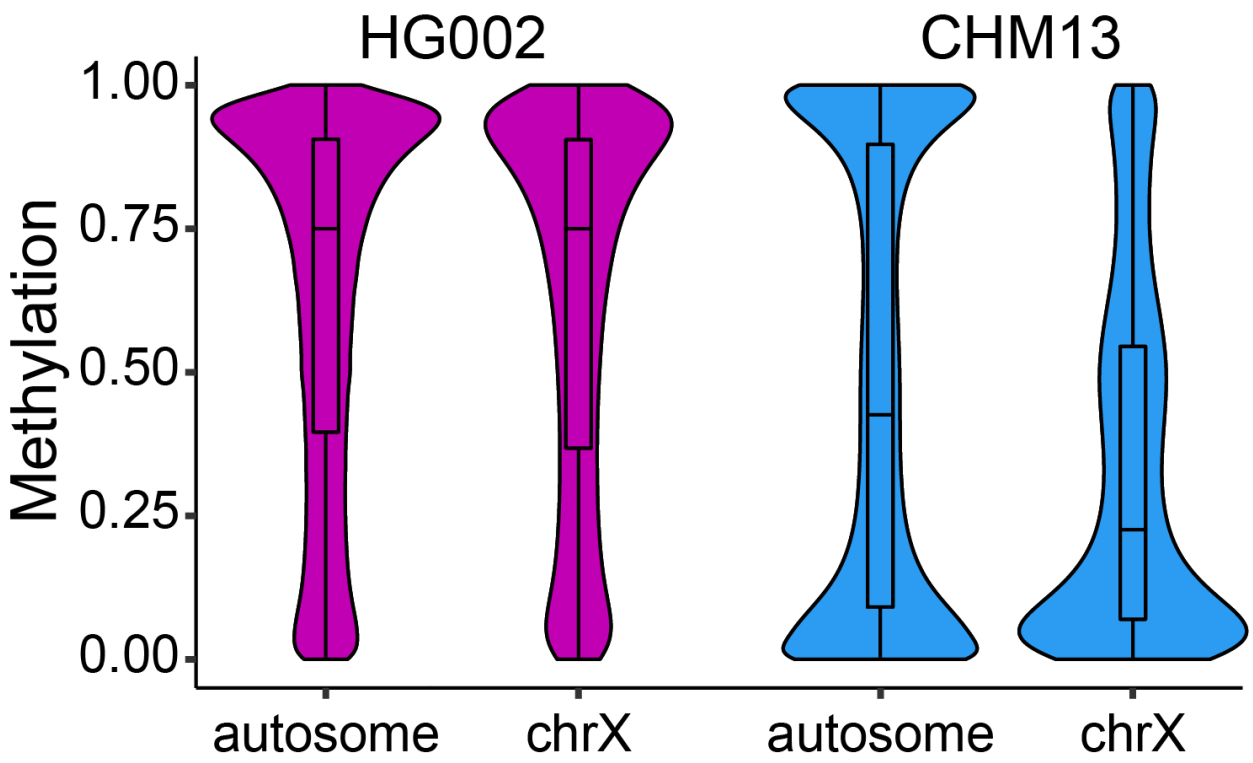

Figure S10. Whole genome methylation of CHM13 and HG002*. Distribution of nanopolish methylation frequency for HG002 and CHM13 on autosomes and chromosome X.*

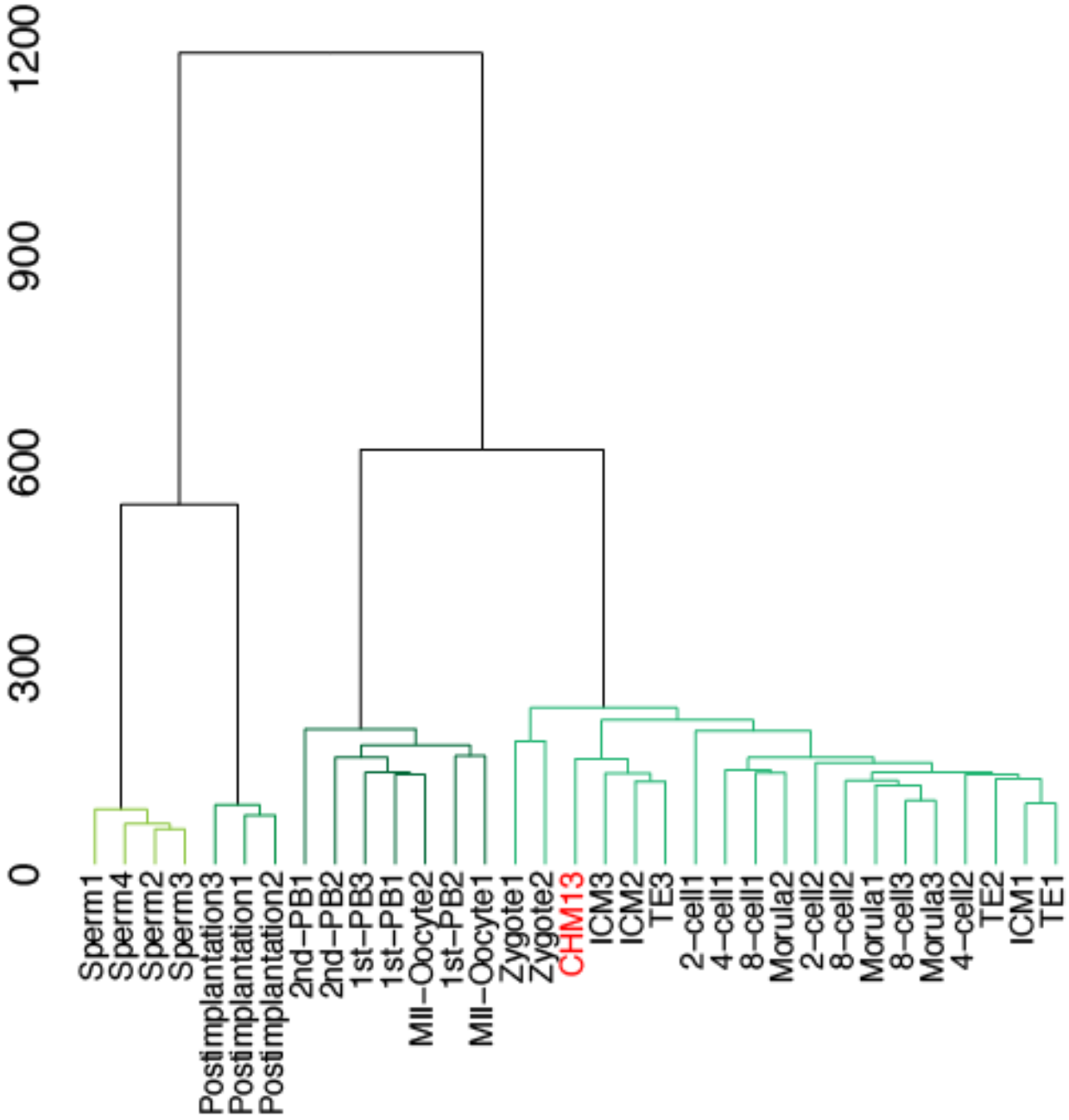

Figure S11. Comparison of methylation in early human embryo samples and CHM13. *Reduced representation bisulfite sequencing (RRBS) methylation data from 12 stages of human embryo development(35) compared to CHM13 methylation generated from nanopore data. T2T-CHM13 reference CpGs covered by at least 1 read in 90% of samples were used. Raw methylation percentages were clustered using Euclidean distance and ward.D clustering then plotted as a dendrogram.*

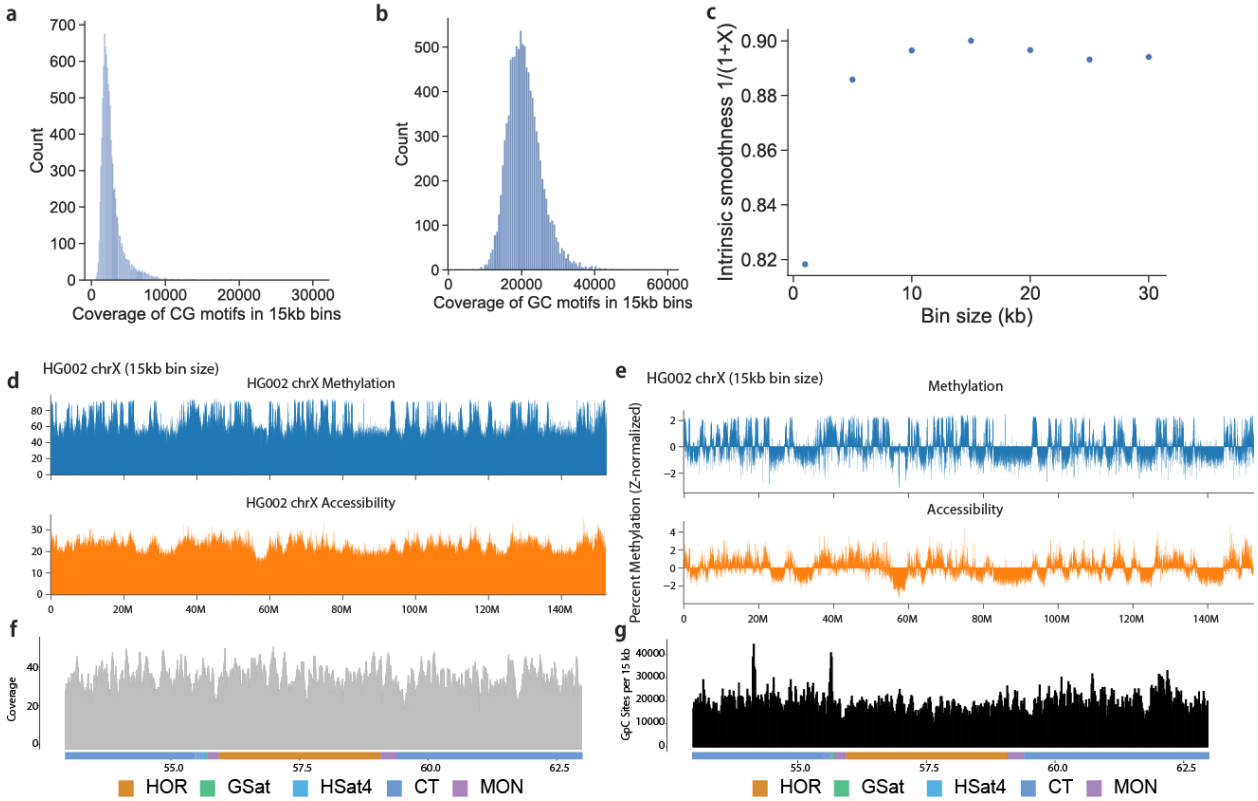

Figure S12. NanoNOMe alignment and GpC methylation calling. ***a)*** *Histogram of coverage for CpG sites in 15kb bins.* ***b)*** *Histogram of coverage for GpC sites in 15kb bins.* ***c****) Intrinsic smoothness as a function of bin size.* ***d)*** *Top panel shows percent CpG methylation for 15kb bins across chromosome X in HG002. Bottom panel shows percent GpC methylation (accessibility) for 15kb bins across chromosome X in HG002.* ***e)*** *Top panel shows Z-normalized methylation (CG) for 15kb bins across chromosome X in HG002. Bottom panel shows Z-normalized accessibility (GC) for 15kb bins across the chromosome X in HG002.* ***f)*** *NanoNOMe CpG call coverage in the centromeric region of the HG002 X chromosome.* ***g)*** *Number of GC sites per 15kb bin across the centromere of HG002 chromosome X.*

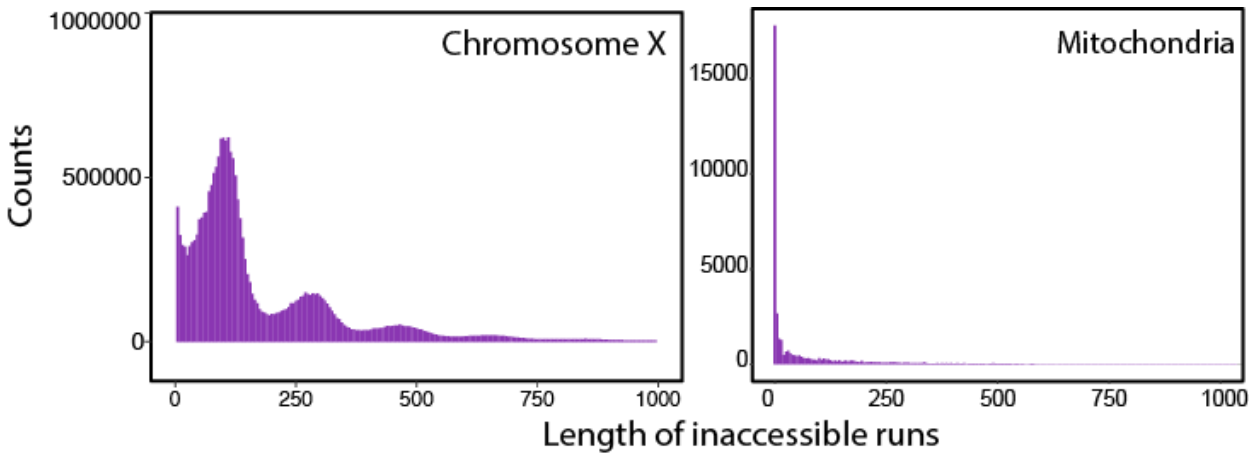

Figure S13. NanoNOMe Inaccessible Run Lengths. *Histograms of the length of inaccessible runs in the nanoNOMe data. Left histogram is all of chromosome X. Peaks correspond to mono-, di-, tri- and poly-nucleosomes. Inaccessible runs occur when nucleosomes or other proteins impede the ability of the GpC methyltransferase to label the DNA. Right panel shows histogram of runs on mitochondrial DNA that does not contain nucleosomes.*

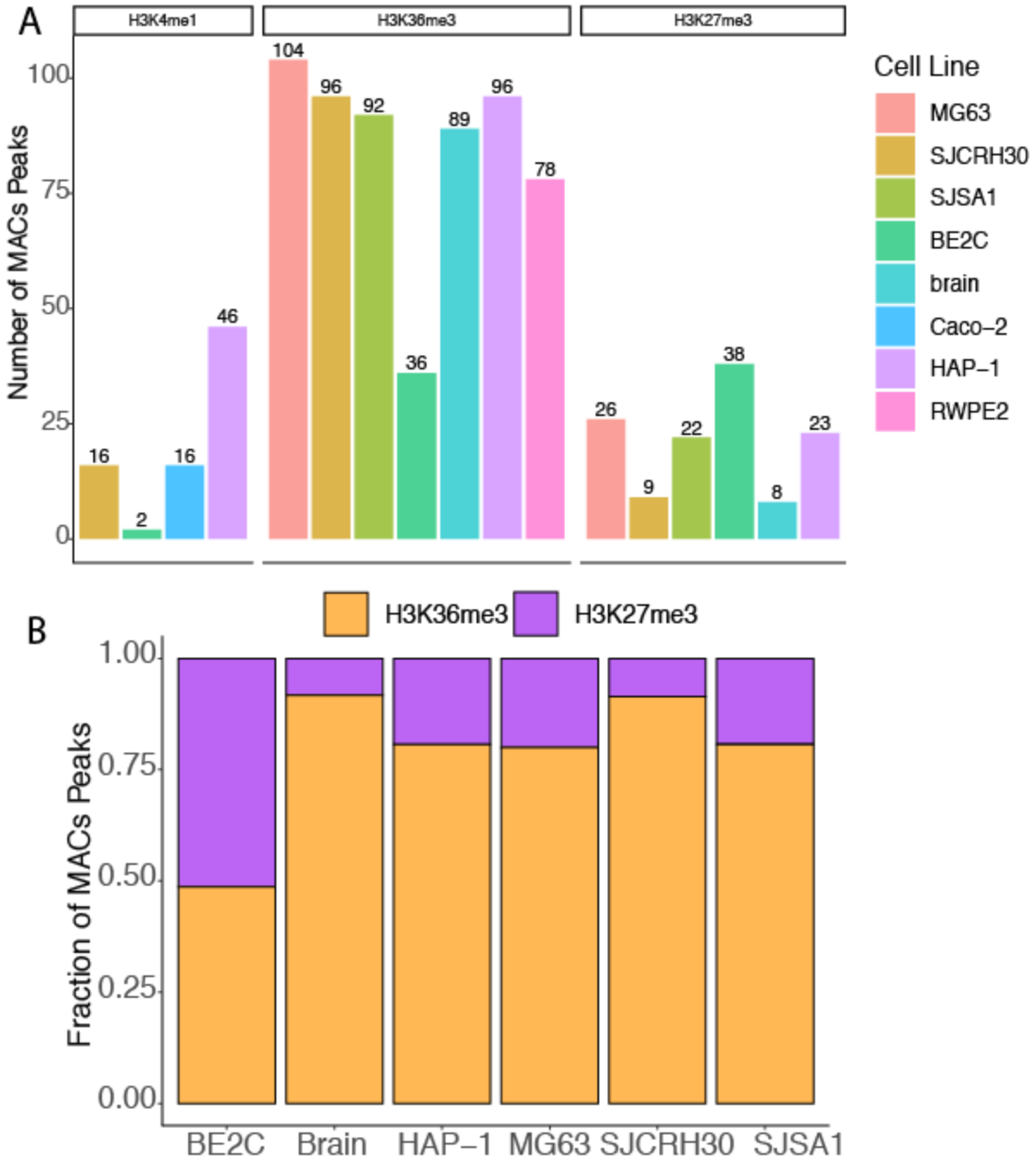

Figure S14. ENCODE peaks at NBPF genes. A) *All peak calls for all ENCODE samples in NBPF genes.* ***B)*** *Peak calls for H3K36me3 and H3K27me3 in NBPF genes.*

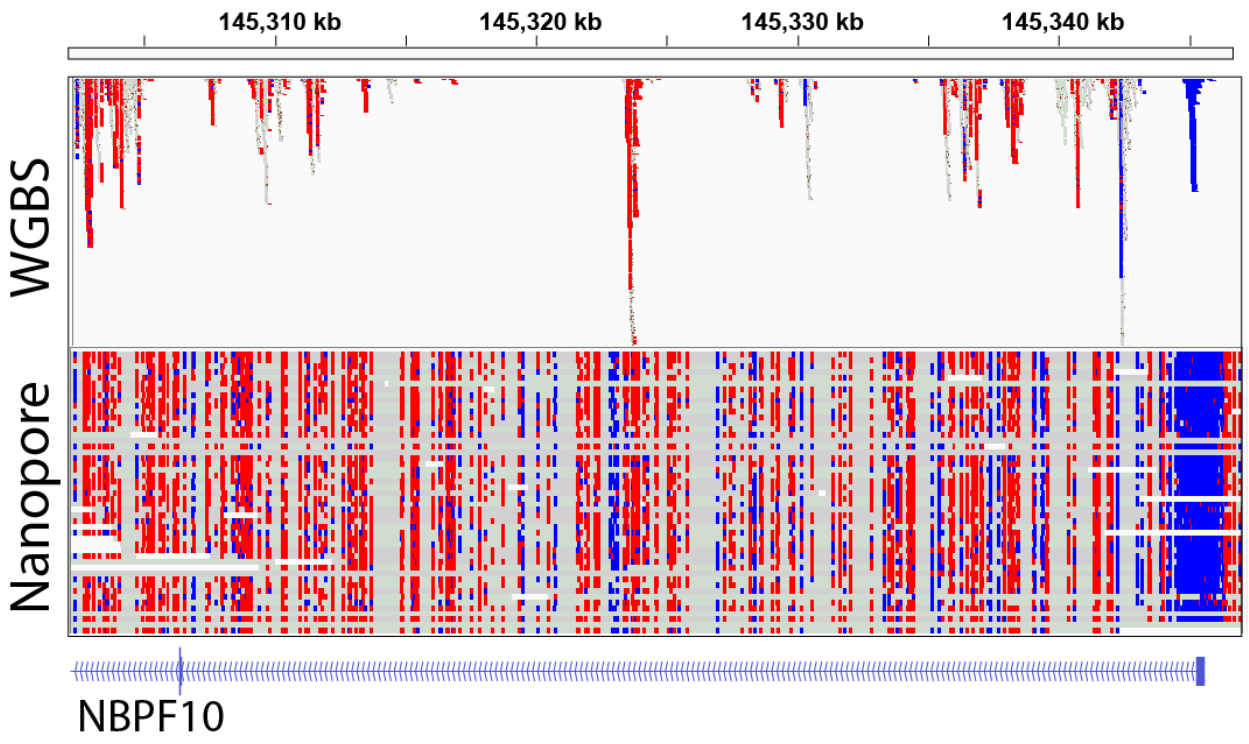

Figure S15. WGBS and nanopore alignments at NBPF10. *IGV plot of NBPF10 from HG002 nanopore and WGBS data. Red are methylated CpGs, blue unmethylated.*

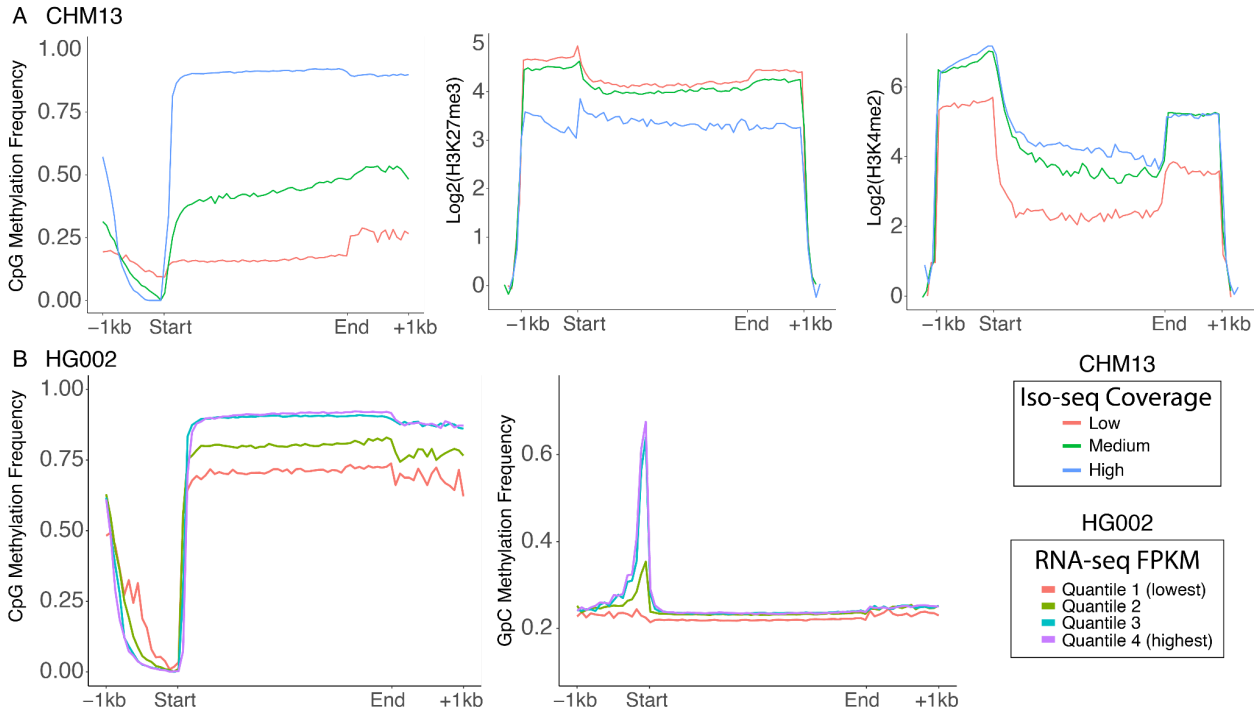

Figure S16. Genome wide expression profiling. ***A***) *Metaplots of all genes separated into high, medium, and low expression categories using CHM13 PacBio Iso-Seq coverage. (Left) CHM13 nanopore aggregated methylation frequency, (Middle) log2 of CHM13 marker-assisted mapping H3K27me3 CUT&RUN coverage, (Right) log2 of CHM13 marker-assisted mapping H3K4me2 CUT&RUN coverage.* ***B***) *Metaplots of HG002 methylation (Left) and GpC accessibility (Right) in all genes separated by FPKM quartile derived from Illumina RNA-seq.*

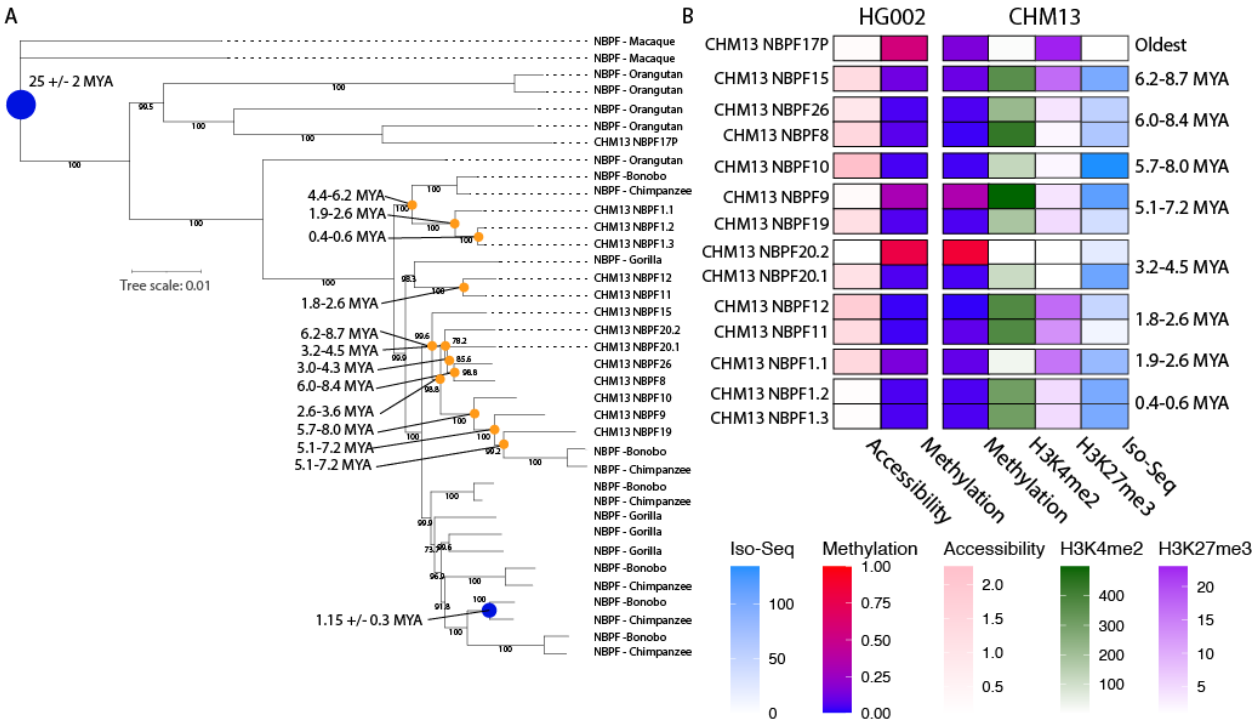

Figure S17. Phylogenetic aging of NBPF. A) *Phylogenetic tree of the NBPF gene family and 7 non-human primates. The scale is in substitutions per site and the numbers on the branches are the bootstrap support values.* ***B)*** *Heatmap characterizes epigenetics of human NBPF genes in CHM13 and HG002, including nanopore methylation, nanoNOMe accessibility, for both, and H3K4me2 CUT&RUN, H3K27me3 CUT&RUN and PacBio Iso-Seq for CHM13 only.*

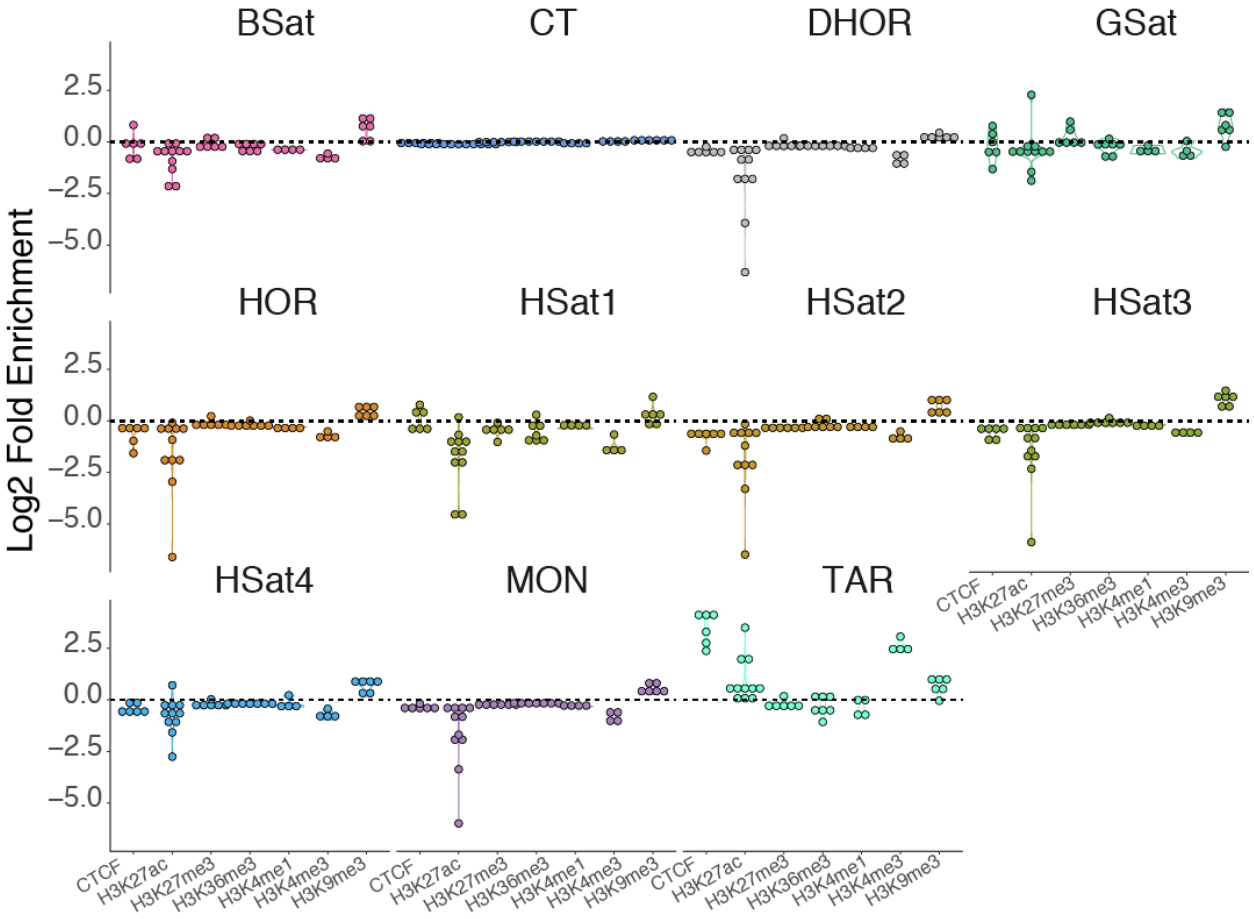

Figure S18. Enrichment of histone marks and CTCF across ENCODE samples. *Log2 fold enrichment of epigenetic mark versus input control for each satellite repeat class.*

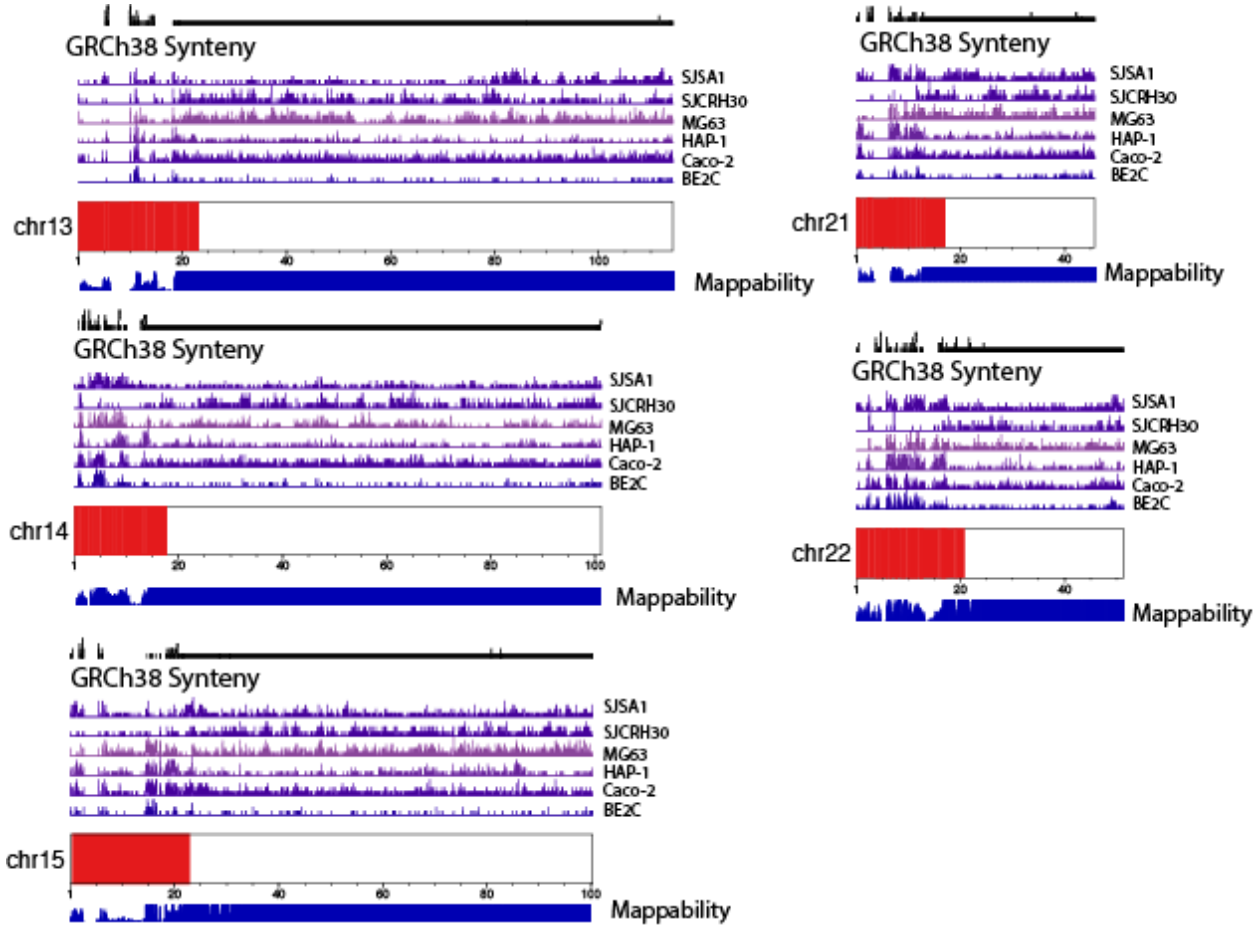

Figure S19. Previously unresolved H3K9me3 peaks in acrocentric chromosomes. *H3K9me3 previously unresolved peaks in T2T-CHM13 across acrocentric chromosomes in ENCODE cell lines. Y-axis peak count ranges from 0-5 for all cell lines. Red boxes denote centromeric regions. GRCh38 synteny shows syntenic regions between GRCh38 and T2T-CHM13, gaps are non-syntenic. Mappability tracks show the percentage of possible 200bp reads overlapping each base position that are uniquely mappable.*

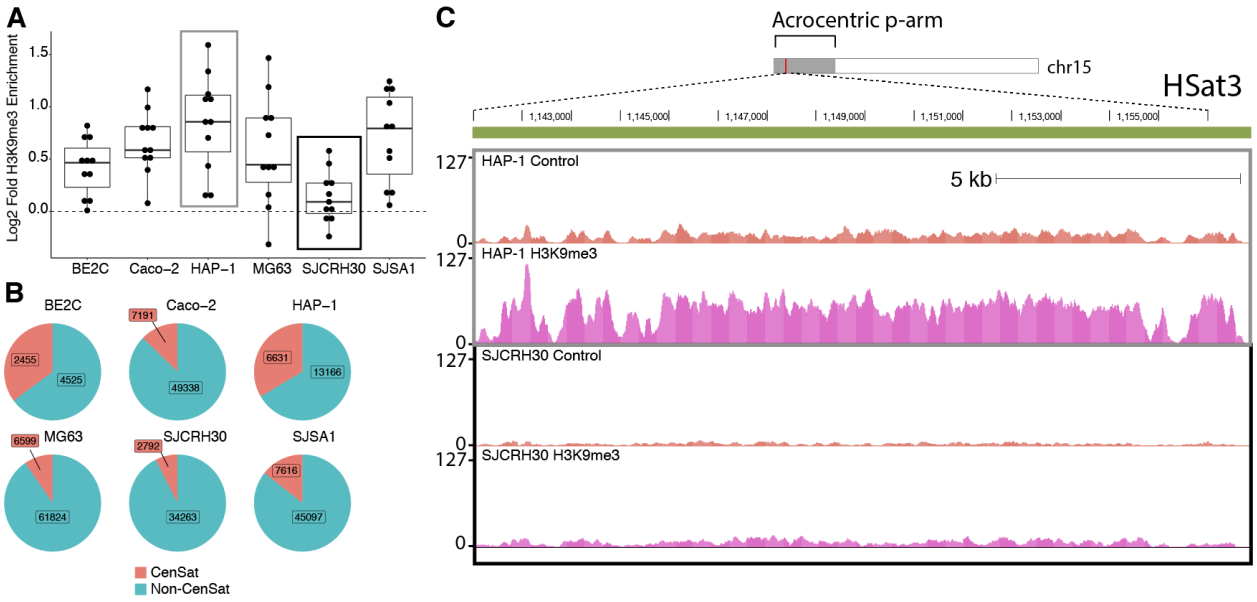

Figure S20. H3K9me3 enrichment in satellite repeats. A) *Log2 enrichment of H3K9me3 ChIP-seq normalized to input control for each satellite repeat. Each dot represents a satellite repeat class: HSat1, HSat2, HSat3, HSat4, GSat, HOR, DHOR, BSat, MON, TAR and CT.* ***B)*** *Pie chart illustrating the position of peak calls for H3K9me3 across the six cell lines, either within (pink) or without (blue) the CenSat regions.* ***C)*** *UCSC genome browser tracks of H3K9me3 coverage versus input control for HAP-1 (highest enrichment) and SJCRH30 (lowest enrichment).*

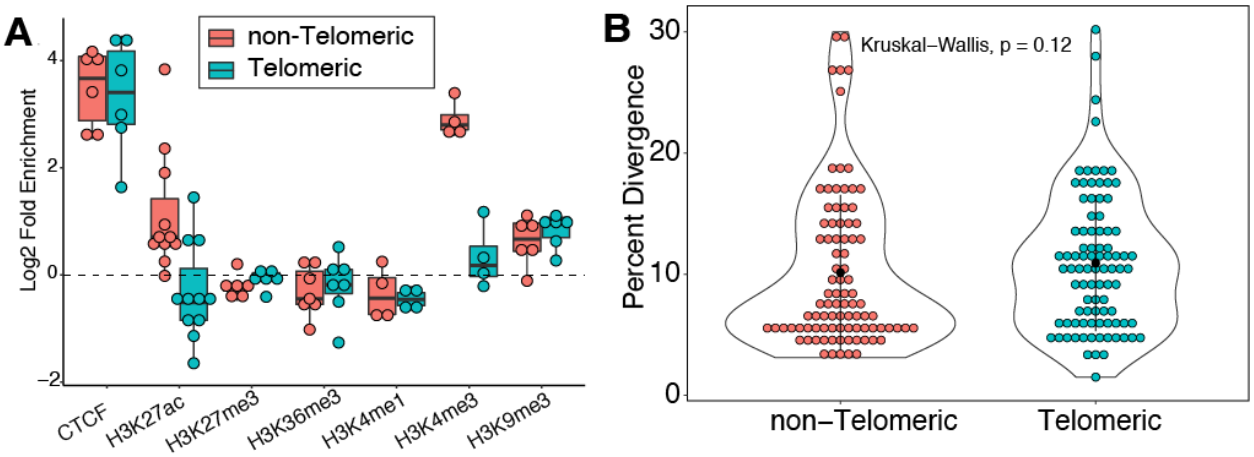

Figure S21. Percent divergence of the Telomere Associated Repeat (TAR). A) *Enrichment of chromatin marks at Telomeric TAR repeats (within 2kb of the chromosome end) and non-telomeric TAR repeats (outside 2kb of the chromosome end). A CTCF site in the TAR loci drives transcription of the TERRA lncRNA (41)); a negative regulator of telomerase-mediated telomere elongation. In T2T-CHM13 we observed TAR at chromosome ends, as expected, however we also observed TAR outside of the sub-telomeric regions (42). Both were enriched for CTCF binding, however, the non-telomeric TAR sequences are more enriched for activating chromatin marks H3K27ac and H3K4me3, suggesting differences in activity of TERRA between telomeric and non-telomeric TAR repeats.* ***B)*** *Percent divergence calculated by RepeatMasker from the TAR consensus sequence. There was no significant sequence divergence between telomeric and non-telomeric TAR (p = 0.12, Kruskal-Wallis).*

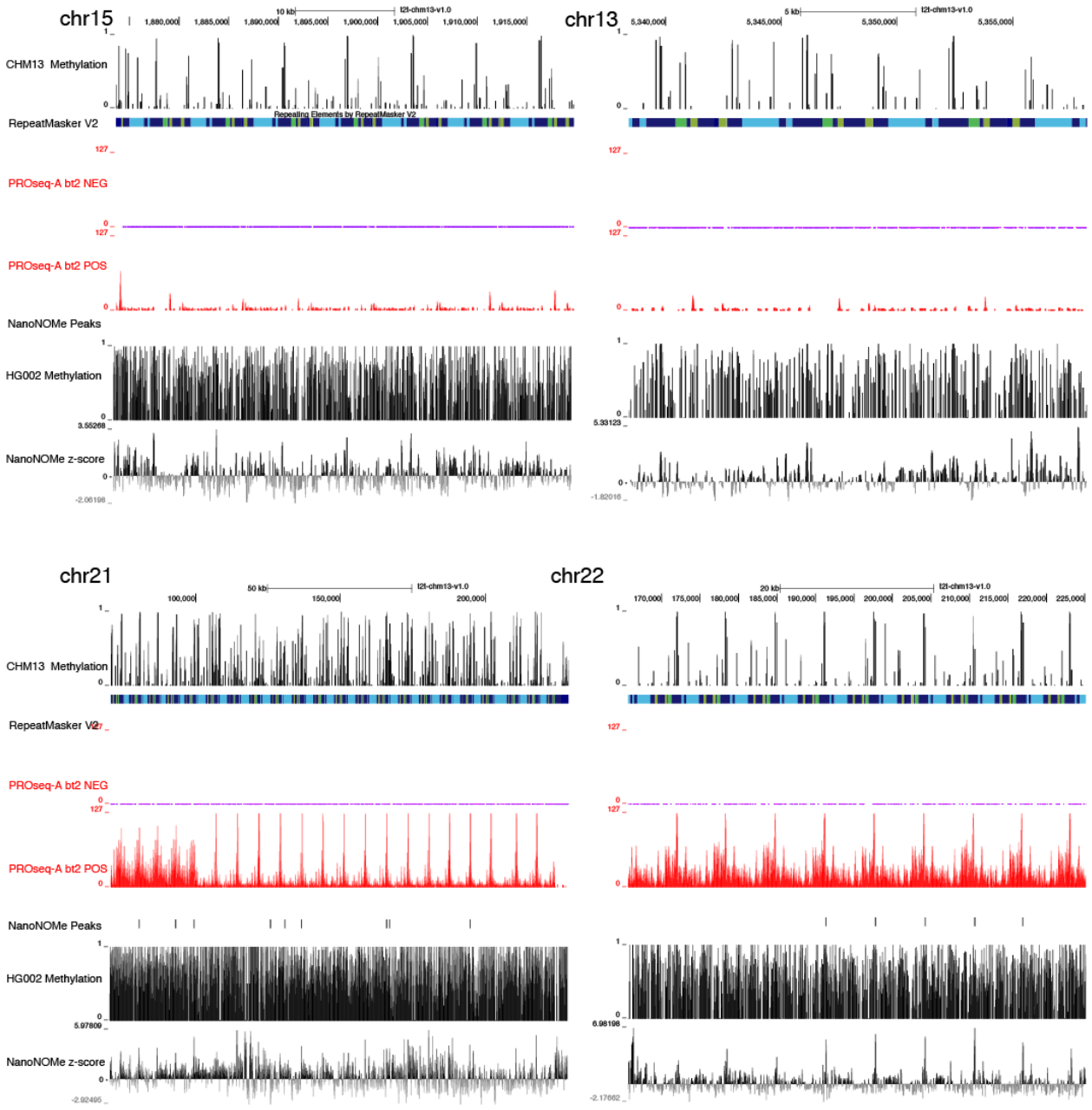

Figure S22. ACRO_Composite comparison across chromosomes. *Genome browser tracks of the ACRO_Composite across the acrocentric chromosomes (chr15, chr13, chr21, chr22). CHM13 and HG002 Methylation, PRO-seq (+ and - strand) and nanoNOMe z-score and peaks are shown, along with the RepeatMasker Annotation track. Across the different chromosomes more nanoNOMe peak calls are associated with higher transcriptional activity (PRO-seq).*

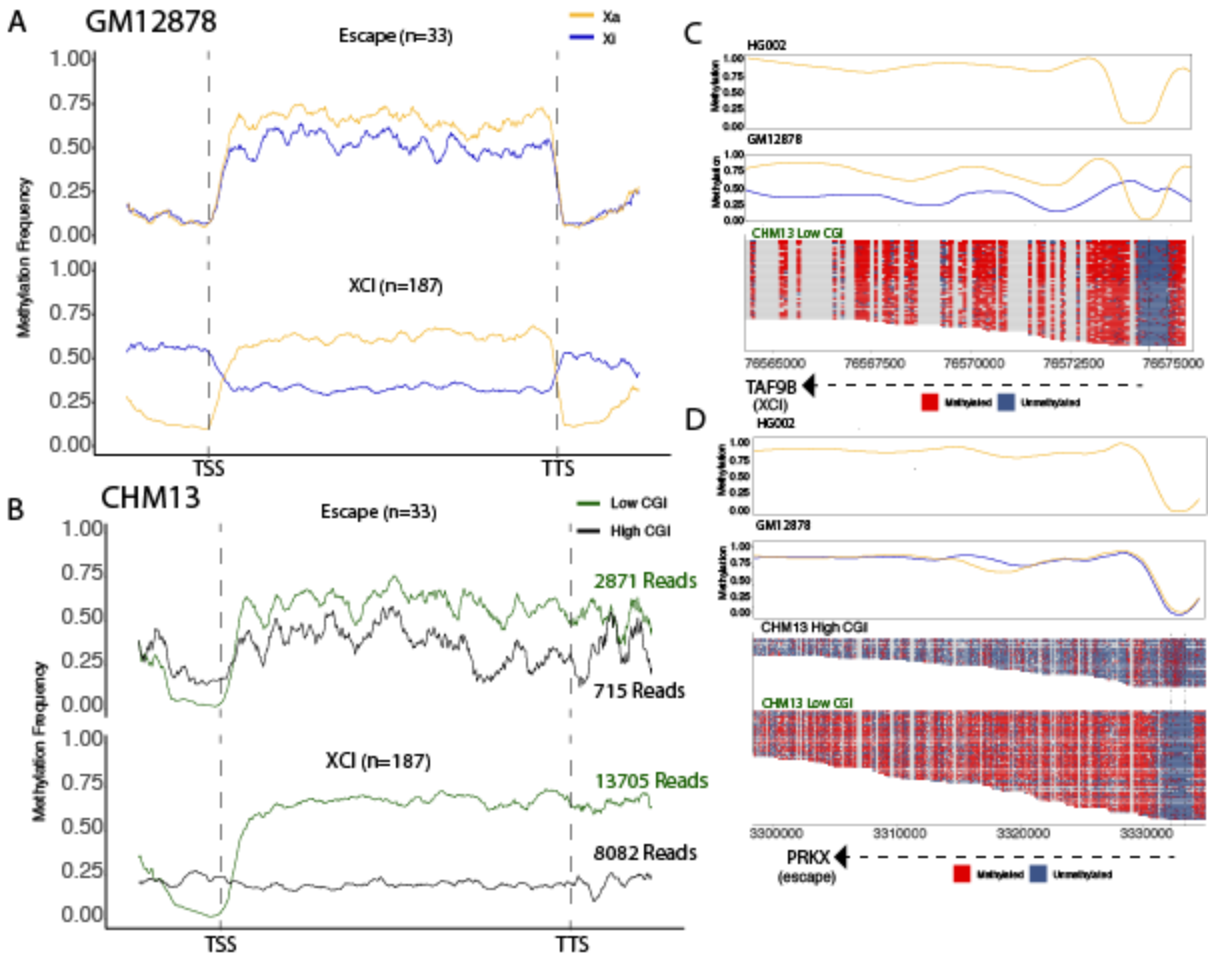

Figure S23. Characteristization of XCI in CHM13. A) *NanoNOMe CpG methylation data from (16) of GM12878, a clonal female lymphoblast cell line representing normal XCI lifted over to T2T-CHM13. Data is shown as ametaplot of methylation at genes known to be prone to X chromosome inactivation (XCI) and genes known to escape XCI (escape genes) (51).* ***B)*** *A metaplot of CHM13 single read clusters at XCI and escape genes. Reads are clustered using methylation at CpG island promoters for either low (green) or high (black) methylation.* ***C)*** *TAF9B is subject to XCI across diverse tissues and in GM12878, but in CHM13 all the reads cluster into the "Low CGI" cluster suggesting it is active on both alleles.* ***D)*** *PRKX is an escape gene and only has one methylation state (hypomethylated promoter) in GM12878, but in CHM13 30% of reads cluster into the "High CGI" group, indicating inactivity.*

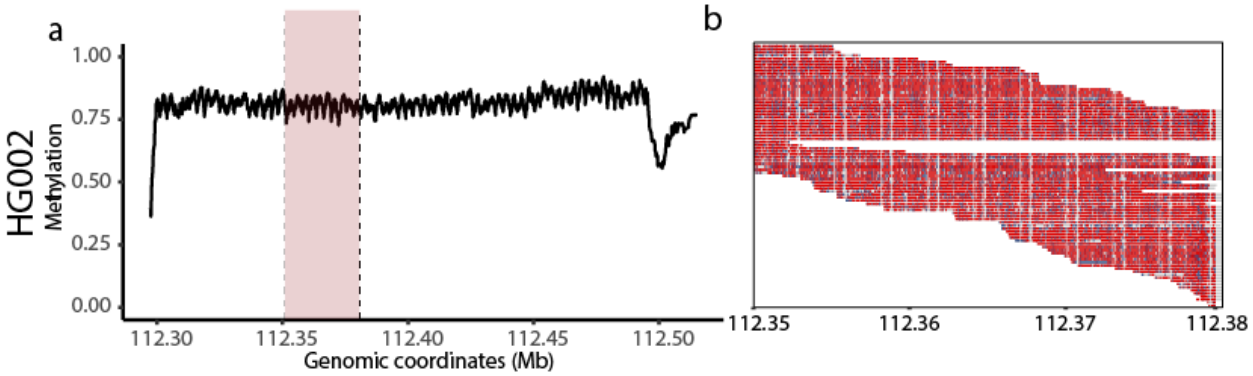

Figure S24. DXZ4 methylation frequency in HG002. a) *DXZ4 methylation frequency in HG002* ***b)*** *Single read methylation plot of HG002 DXZ4 at the pink highlighted region from* ***a****.*

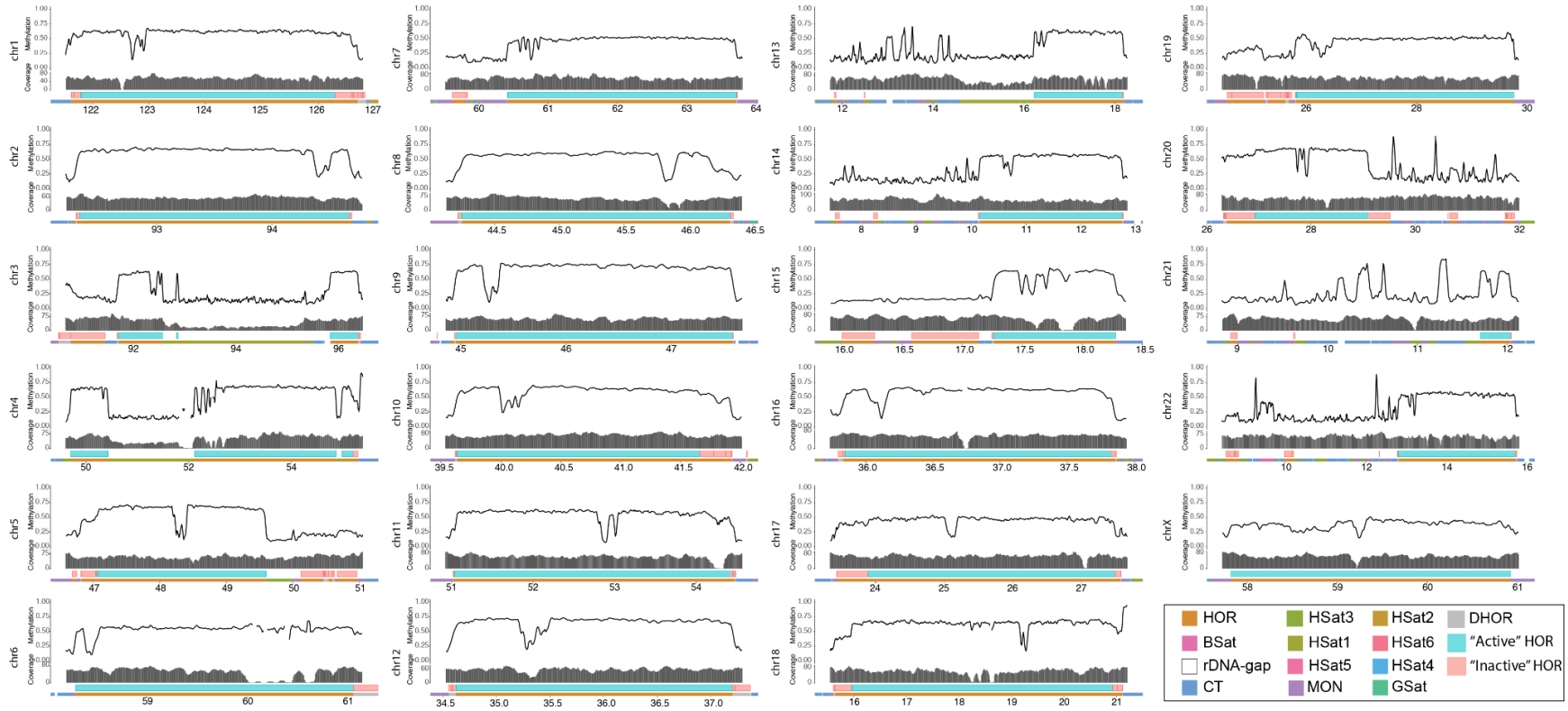

Figure S25. Methylation frequency of CHM13 centromeric regions. *Panels describing methylation in the centromeric regions of each chromosome of CHM13. CHM13 methylation frequency is plotted in 10kb bins smoothed with a rolling average over three bins. Coverage plot represents the number of aligned nanopore reads containing at least one high quality CpG methylation call. "Active" and "Inactive" HOR arrays are annotated by salmon and teal track below coverage plot. Bottom thin line annotates satellite repeats.*

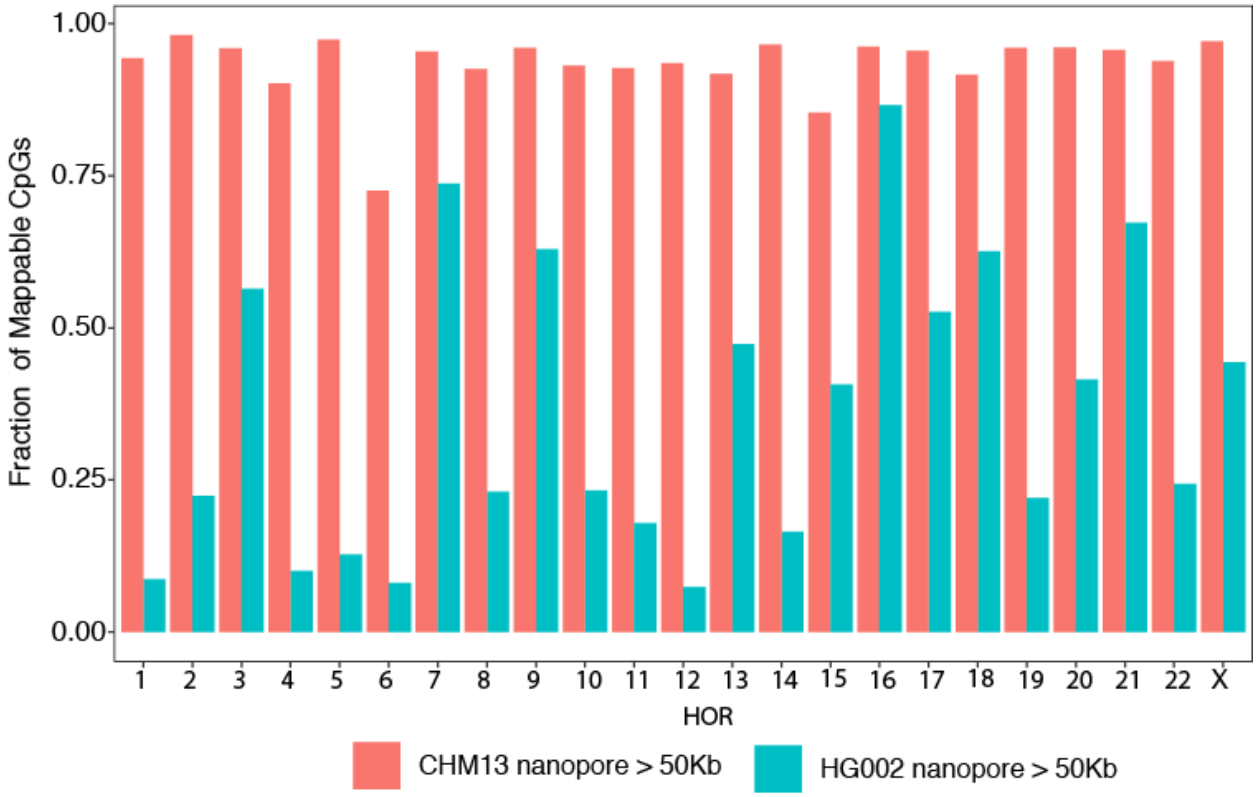

Figure S26. Mappability of HOR. *The fraction of mappable CpG sites for HG002 nanopore and CHM13 nanopore reads >50kb aligned to T2T-CHM13 at the HOR in each chromosome. Mappability was calculated by calculating the fraction of CpG sites that had coverage greater than the 5th percentile and less than the 95th percentile of total coverage.*

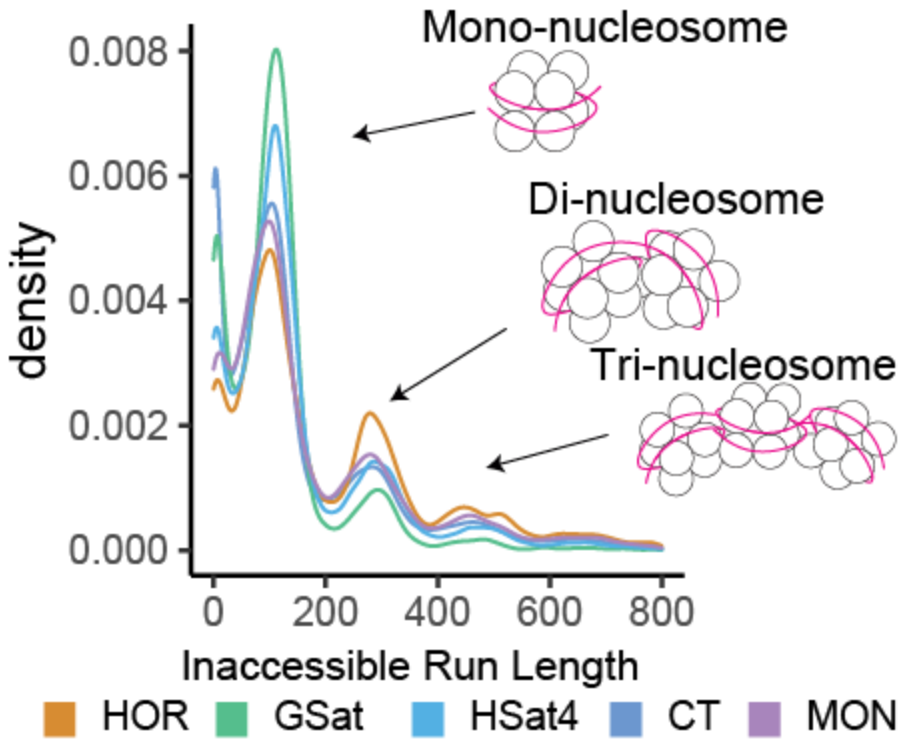

Figure S27. Nucleosome Positioning in HG002 cenX. *Density plots of length of inaccessible run lengths within different genomic regions on HG002 chromosome X.*

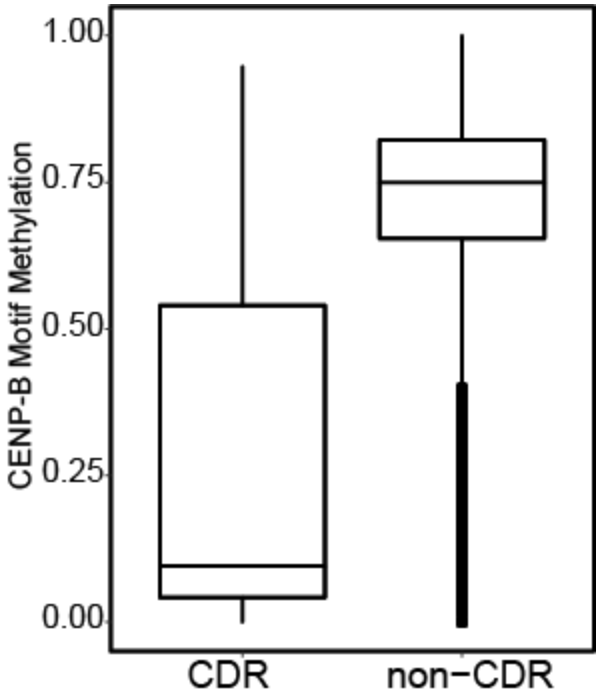

Figure S28. Methylation at CENP-B Boxes. *Methylation frequency of the CpG sites within the CENP-B motif inside and outside the CDR (p < 1e-15, Kruskal-Wallis) on HG002 chromosome X.*

SUPPLEMENTAL TABLES:

S1. ENCODE Accessions. *Accession numbers for ENCODE datasets used.*

S2. ENCODE alignments. *A summary of dynamic k-mer assisted mapping of ENCODE data against T2T-CHM13 and GRCh38.*

S3. Functional annotation of previously unannotated genes. *H3K27ac and H3K4me3 peaks at the TSS of previously unannotated genes in T2T-CHM13.*

S4. Early development RRBS data. *A summary of the early development RRBS data including the tissue and accession numbers for the datasets.*

S5. *NBPF* gene and duplicon coordinates for phylogeny. *Summary of the NBPF gene annotations used for all NBPF analyses. Coordinates for the duplicons were generated by mapping an initial paralog (NBPF25P chr1:144,688,831-144,708,527) back to the T2T-CHM13 genome to generate a high identity gene set for aging analysis.*

S6. nanoNOMe peaks in all repeats. *Number of nanoNOMe statistically significant peak calls in repeat classes normalized by genomic size.*

S7. nanoNOMe peaks in SST1 repeats. *Number of nanoNOMe statistically significant peak calls in SST1 repeat classes normalized by genomic size. SST1 regions are grouped as arrayed or monomeric and within the centromere or outside the centromere.*

S8. CHM13 XCI clustering. *Fraction of reads in low vs high CGI groups in escape genes and genes subject to XCI.*

S9. CHM13 CDRs. *Coordinates of the CHM13 CDRs.*