A haplotype-resolved view of human gene regulation

Mitchell R. Vollger1*, Elliott G. Swanson2*, Shane J. Neph1, Jane Ranchalis1, Katherine M. Munson2, Ching-Huang Ho3, Adriana E. Sedeño-Cortés1, William E. Fondrie4, Stephanie C. Bohaczuk1, Yizi Mao1, Nancy L. Parmalee5, Benjamin J. Mallory2, William T. Harvey2, Younjun Kwon2, Gage H. Garcia2, Kendra Hoekzema2, Jeffrey G. Meyer6, Mine Cicek6, Evan E. Eichler2,7, William S. Noble2,8, Daniela M. Witten9, James T. Bennett10, John P. Ray2,3,11, Andrew B. Stergachis1,2,12,†

1. Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA, USA
2. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
3. Center for Systems Immunology, Benaroya Research Institute at Virginia Mason Franciscan Health, Seattle, WA, USA
4. Talus Bioscience, Seattle, WA, USA
5. Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, Seattle, WA, USA
6. Department of Laboratory Medicine and Pathology, Mayo Clinic Hospital, Rochester, Minnesota, USA
7. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA
8. Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA
9. Departments of Statistics & Biostatistics, University of Washington, Seattle WA, USA
10. Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, and Department of Pediatrics, University of Washington, Seattle, WA, USA
11. Department of Immunology, University of Washington School of Medicine, Seattle, WA, USA
12. Brotman Baty Institute for Precision Medicine, Seattle, WA USA

† Corresponding author(s).
* These authors contributed equally to this work.

Abstract

Most human cells contain two non-identical genomes, and differences in their regulation underlie human development and disease. We demonstrate that Fiber-seq Inferred Regulatory Elements (FIREs) enable the accurate quantification of chromatin accessibility across the 6 Gbp diploid human genome with single-molecule and single-nucleotide precision. We find that cells can harbor >1,000 regulatory elements with haplotype-selective chromatin accessibility (HSCA) and show that these elements preferentially localize to genomic loci containing the most human genetic diversity, with the human leukocyte antigen (HLA) locus showing the largest amount of HSCA genome-wide in immune cells. Furthermore, we uncover HSCA elements with sequence non-deterministic chromatin accessibility, representing likely somatic epimutations, and show that productive transcription from the inactive X chromosome is buttressed by clustered promoter-proximal elements that escape X chromosome inactivation.

Introduction

Advances in long-read sequencing (Jain et al., 2018; Wenger et al., 2019) have enabled the routine *de novo* assembly of 6 Gbp diploid human genomes at reference quality (Cheng et al., 2021; Rautiainen et al., 2023) resulting in the completion of the first human genome (Logsdon et al., 2021; Miga et al., 2020; Nurk et al., 2022) and pangenome (Guarracino et al., 2023; Liao et al., 2023; Vollger, Dishuck, et al., 2023). These advances have improved our understanding of the structure of the human genome and human genetic variation, enabling researchers to investigate genetic variants within their native haplotype context along the 6 Gbp diploid human genome. Now, our next challenge is studying the function of this 6 Gbp genome.

Despite the potential for studying the function of genetic variation along the diploid genome, most chromatin assays are reliant on collapsed 3 Gbp representations of a human genome. Specifically, given the inconsistent density of genetic variation within the human genome, short-read-based chromatin methods are poorly suited for uniquely mapping chromatin across the diploid genome, and even when an element happens to overlap a heterozygous single-nucleotide variant (SNV), read-mapping artifacts are known to confound measures of allele-specific chromatin accessibility (ASCA).

Long-read single-molecule chromatin assays have overcome these challenges by leveraging the ability of long reads to correctly map to the diploid genome (Grasberger et al., 2024; Lee et al., 2020; Vollger, Korlach, et al., 2023). For example, Fiber-seq uses a non-specific N^6-adenine methyltransferase (m6A-MTase) to stencil protein occupancy footprints along DNA molecules in the form of m6A-modified bases (Stergachis et al., 2020). These m6A-modified DNA molecules are sequenced using a single-molecule platform (Eid et al., 2009), enabling the synchronous readout of the genetic sequence, CpG methylation status, and chromatin architecture of each ~20kb sequencing read. The promise of single-molecule chromatin assays is to: (1) enable the accurate discrimination of single-molecule chromatin architectures without any *a priori* knowledge of the reference genome from which that read derives and (2) enable the accurate identification of accessible putative regulatory elements along the diploid human genome without any *a priori* knowledge of the regulatory landscape of that sample. Fiber-seq and a similar approach using GpC methyltransferases (nanoNOMe) have enabled the investigation of nucleosomal patterns within complex genomic loci (Battaglia et al., 2022; Dubocanin et al., 2022, 2023; Gershman et al., 2021), but the computational tooling to accurately call regulatory elements across the diploid human genome has yet to be developed.

To realize the full potential of long-read single-molecule chromatin assays, we developed a semi-supervised machine learning classifier for precisely mapping chromatin patterns along individual reads, as well as a statistical approach for identifying accessible putative regulatory elements along the diploid human genome with single-molecule and single-haplotype precision at near nucleotide resolution. Using this approach, we uncover basic principles guiding gene regulation, X chromosome inactivation, and human genome evolution.

Results

Single-molecule Fiber-seq inferred regulatory elements

Fiber-seq enables the stenciling of protein occupancy events along individual chromatin fibers at near base-pair resolution using m6A-modified A/T base pairs (**Fig. 1a**). The predominant signal produced by these m6A stencils is nucleosome occupancy footprints, which correspond to ~146 bp stretches of largely unmodified A/T base pairs separated by stretches of m6A-modified bases (i.e., MTase Sensitive Patches [MSPs]). The majority of MSPs are <100 bp in size, consistent with standard internucleosomal linker regions, with MSPs >150 bp enriched at known DNaseI hypersensitive sites (DHSs) (**Fig. S1**). However, the majority of MSPs >150 bp in size are located outside of known DHSs, likely reflecting unstable nucleosome occupancy events, RNA/DNA polymerase activity, and other chromatin features in addition to novel regulatory elements previously hidden from short-read assays. This indicates that MSP size alone is not an accurate classifier for actuated regulatory elements versus standard internucleosomal linker regions on a single-molecule level. However, given the near base-pair resolution of Fiber-seq, we hypothesized that MSPs within known DHSs would have other features in addition to length that distinguish them from MSPs outside of DHSs, including their m6A density, the distribution of m6As within them, etc.

To accurately classify chromatin along individual Fiber-seq reads, we sought to develop a machine learning classifier that could utilize Fiber-seq features at MSPs that differentiate actuated single-molecule regulatory elements from internucleosomal linkers (**Fig. S1**). Importantly, in order to make our classifier broadly applicable to diverse chromatin contexts, reference genomes, cell types, and organisms, we aimed to develop a classifier that enabled the accurate classification of individual chromatin fiber architectures without any knowledge of: (1) the underlying reference genome; (2) the exact primary sequence of the underlying fiber; (3) the configuration of chromatin at other positions along the same fiber; or (4) the CpG methylation status of the underlying fiber. By not including the primary read sequence or reference genome in our training, our model has the potential to generalize across different cell types and organisms that may employ unique sets of binding elements to regulate their accessible chromatin. In addition, in contrast to traditional short-read epigenetic classifiers that operate at the level of positions along a reference genome, we sought to enable single-molecule chromatin classification on each read prior to reference mapping, as this would improve chromatin classification within complex regions of the human genome, and enable our epigenetic annotations to readily extend to regions along each read that represent gaps within the GRCh38 reference assembly.

To create training data for this machine learning task, we generated multiple Fiber-seq datasets from the reference cell line GM12878 (totaling 136-fold coverage), as this cell line has extensive paired short-read epigenetic data. Importantly, we purposefully included sub-optimal Fiber-seq data in our training dataset from samples that were either undermethylated or overmethylated with the m6A-MTase, with the goal that our final model would extend to analytical fluctuations in the Fiber-seq method (**Fig. S1**). Although we aim to apply this method at the level of individual Fiber-seq reads irrespective of their reference genome, labels for the training data relied on GRCh38, as this enabled us to use the extensive short-read epigenetic data from GM12878 to

assign positive and negative labels to individual MSPs. Specifically, MSPs that overlapped GM12878 DNase-seq DHSs or GM12878 CTCF ChIP-seq peaks were assigned a positive label. In contrast, MSPs from short-read mappable regions of the genome without these short-read epigenetic peak annotations were assigned a negative label. Notably, as previously described (Lee et al., 2020; Shipony et al., 2020; Stergachis et al., 2020), regulatory elements identified using bulk methods like DNase-seq and ChIP-seq exhibit marked heterogeneity in their single-molecule architecture, including numerous Fiber-seq reads that merely contain internucleosomal linker regions. As such, in the initial training data, MSPs ascribed a positive label actually represent a mixture of MSPs arising from actuated regulatory elements and others arising from internucleosomal linker regions. Consequently, our training data is best described as mixed-positive labels and clean-negative labels, a training paradigm that is best approached using a semi-supervised machine learning framework (Fondrie & Noble, 2021; Jha et al., 2023; Käll et al., 2007).

To carry out a semi-supervised approach, we trained an XGBoost model with five-fold cross-validation (Chen & Guestrin, 2016), iteratively retraining the model using the learned prediction from the previous iteration's model to create positive labels at 95% estimated precision until the number of positive identifications at 95% estimated precision (*Methods*) in the validation set ceased to increase (15 iterations). We applied this final model to our held-out test data to create an estimated precision based on the model's prediction score (*Methods*). We then used the estimated precision to classify 1,959,888,668 MSPs across 24,771,424 chromatin Fiber-seq reads as either **F**iber-seq **I**nferred **R**egulatory **E**lements (FIREs; 90% estimated precision or greater; n=32,006,894) or internucleosomal linker regions (less than 90% estimated precision). Overall, MSPs classified as FIREs were enriched in DHSs compared to using MSP length alone as a classifier (**Fig. S1**). As no other methods exist for classifying per-molecule chromatin architectures, in order to appropriately evaluate and benchmark our FIRE classifications, we next needed to develop a method for statistically aggregating these per-molecule classifications at individual genomic regions.

Aggregating across Fiber-seq inferred regulatory elements.

Next, we sought to leverage these single-molecule FIRE classifications to identify genomic loci containing accessible chromatin elements (**Fig. 1b**). To accomplish this, we first created a coverage-normalized aggregate FIRE score (*Methods*) that combines the FIRE classification across multiple molecules for every base in the genome. We then calibrated these aggregate FIRE scores by computing a null aggregate FIRE score using reads randomly shuffled along the genome. This enabled us to establish a null distribution and calculate a false discovery rate (FDR, *Methods*, **Fig. S2**). Next, for genomic positions that exceeded the 5% FDR threshold, we identified local maximums and called narrow peaks, setting the bounds of the peak using the median start and end positions of the underlying single-molecule FIRE elements (**Fig. 1a,b**). Finally, at each peak, we calculated the percentage of all Fiber-seq reads with a FIRE element that overlapped the local maximum of the peak, hereafter referred to as the "percent actuation." Importantly, this enables an intuitive, biologically meaningful metric of chromatin accessibility for the first time. In total, we identified 204,965 FIRE peaks in GM12878, which overlapped 89% of the previously annotated GM12878 DNaseI hypersensitive sites (**Supplemental Table 1**).

However, the boundaries of each FIRE peak correspond to the nucleotide precise median start/stop positions of the underlying single-molecule data, and the intensity reflects the exact fraction of chromatin fibers with accessibility.

To test the generalizability of our approach, we generated two biological replicates of Fiber-seq from a different human cell line (COLO829BL) (totaling 319-fold coverage), which demonstrated that we can readily extend our model to other samples and that FIRE scores are highly concordant between replicates (R=0.97) (**Fig. 1c**). Furthermore, we showed that FIRE scores were stable after downsampling our GM12878 data from 135- to 30-fold coverage (**Fig. S2**). Overall, these findings demonstrate that FIRE enables the robust *de novo* identification of accessible chromatin elements solely using long-read epigenomic data, and unlike short-read approaches, FIRE enables the intuitive representation of chromatin accessibility measures as a percent actuation for every base in the genome.

Benchmarking FIRE against existing chromatin accessibility measures

We sought to benchmark our data using a single-cell ATAC-seq (scATAC-seq) dataset from GM12878, as this data set and data type were not used in our training. Specifically, we pseudo-bulked an entire 26,910 cell ENCODE GM12878 scATAC-seq dataset and then calculated peaks using MACS2 as well as the proportion of cells that have a read mapping to that peak. Overall, we found that peaks of accessibility identified using Fiber-seq, scATAC-seq, or DNase-seq significantly overlapped, with 78% of all FIRE peaks also called by one of these other technologies (**Fig. 1d**). Furthermore, we found both DNase-seq and scATAC-seq signal support for even the least actuated Fiber-seq peaks (10-15%) (**Fig. 1e**) and that DNase and scATAC signal increase with the fraction of Fiber-seq molecules in FIRE peaks (**Fig. 1e,f**), indicating that overall FIRE percent actuation scores are in agreement with short-read bulk measures of chromatin accessibility.

Next, we compared FIRE percent actuation scores against ENCODE DHS annotations, demonstrating that the FIRE peaks with the highest percent actuation are largely tissue invariant. In contrast, the more abundant lower accessibility peaks show additional signals of tissue-specific chromatin expected in this lymphoblastoid cell line (lymphoid, stromal B, and myeloid/erythroid), indicating the power of Fiber-seq to quantify tissue-specific and tissue-invariant chromatin signatures (**Fig. 2a**). Overall, these findings validate the accuracy of FIRE peak calls and chromatin actuation measures against existing short-read chromatin assays.

FIRE provides more accurate measures of chromatin accessibility

We next sought to investigate differences between FIRE measures of chromatin accessibility and those of existing short-read chromatin assays. First, we sought to test how quantitative scATAC-seq is in measuring the fraction of cells with chromatin accessibility, which in our GM12878 sample should be directly proportional to the percentage of fibers containing an accessible element (i.e., FIRE percent actuation). We find that overall scATAC-seq markedly underestimates the actual chromatin accessibility at a given element (**Fig. 1f**). For example, elements with 10-15% Fiber-seq actuation are, on average, detected in less than 1% of

scATAC-seq cells, likely reflecting known issues in scATAC-seq with tagmentation heterogeneity and read dropout. However, we also found that the relationship between scATAC-seq per-cell measures of chromatin accessibility and FIRE percent actuation is often not linear. Specifically, elements detected in 10-15% of cells as measured by scATAC-seq can have Fiber-seq actuation ranging from 10% to 90% of fibers (**Fig. 1f**), indicating more pervasive heterogeneity in how scATAC-seq quantifies chromatin accessibility between different peaks.

In exploring these differences in Fiber-seq measures of chromatin accessibility, we found that controlling for GC content improved the correlation between DNase-seq and Fiber-seq, likely reflecting GC biases that are induced during the PCR amplification steps of DNase-seq (**Fig. S2**).

Furthermore, of the 45,420 GM12878 peaks identified only in Fiber-seq, 78.4% were <200 bp in length, raising the possibility that short-read chromatin accessibility measurements may be underperforming at detecting shorter regulatory elements (**Fig. 2b**). To test this, we calculated the correlation of scATAC-seq pseudo-bulk signal to Fiber-seq percent actuation at peaks with a long (>250 bp) and short (<200 bp) average length, demonstrating that the correlation between scATAC-seq and Fiber-seq drops from 0.81 to 0.38 when comparing peaks ≥250 bp or peaks ≤200 bp, respectively (**Fig. 2b,c**). Notably, 26.5% (30,476/114,865) of peaks ≤200 bp corresponded to CTCF ChIP-seq peaks. Furthermore, scATAC-seq signal within CTCF ChIP-seq peaks is reduced up to 3.47-fold compared to other scATAC-peaks with a similar Fiber-seq percent actuation (**Fig. 2d**). Overall, these results demonstrate that short-read measures of chromatin accessibility are confounded by both GC biases as well as the size of the underlying element, resulting in inaccurate measures of bulk, and especially single-cell, chromatin accessibility (**Fig. 1f**). Although the exact cause of this phenomenon is unknown, it is likely driven by a combination of artifacts inherent to short-read chromatin methods, including: (1) available template region for Tn5 transposition/DNaseI nicking; (2) fragment size selection; (3) PCR amplification preferences; and (4) short-read sequencing preferences.

In addition to these <200 bp peaks, we also found that 8.66% (n=3,933) of the 45,420 GM12878 peaks identified only in Fiber-seq mapped to segmentally duplicated (SD) regions of the human genome (**Fig. 2e**). SDs comprise ~200 Mbp of genomic sequence (Bailey et al., 2002; Vollger et al., 2022) that is known to contribute to a variety of human diseases (Sharp et al., 2005, 2006), but these regions have been challenging to study owing to mapping issues of short-read chromatin assays to these highly similar duplicated sequences (Amemiya et al., 2019). Overall, 45% (3,933/8,700) of all Fiber-seq FIRE peaks in SDs are unique to FIRE, and the SDs with the highest sequence identity contained the highest fraction of FIRE unique peaks (**Fig. 2f**). Together, this demonstrates that not only can FIRE better resolve chromatin architectures within the traditionally mappable portion of the genome, but it is also able to uniquely resolve chromatin patterns across complex genomic regions that are largely impenetrable to short-read chromatin assays.

A haplotype-resolved view of human gene regulation

We next assessed whether Fiber-seq could further expand the mappable genome beyond a collapsed assembly by phasing our long reads into maternal and paternal haplotypes to add an

additional 3 Gbp of unexamined sequence. Specifically, for GM12878, each ~20kb read on average will span at least one heterozygous variant, and parental short-read sequencing data can be further used to bin these reads based on their maternal or paternal origin, enabling us to accurately haplotype phase 87.9% of GM12878 reads across entire GRCh38 chromosomes (*Methods*). Using these haplotype-phased reads, we pseudo-bulked our per-read chromatin architectures by haplotype, generating haplotype-specific chromatin actuation, CpG methylation, nucleosome positioning, and transcription factor (TF) occupancy patterns across the 6 Gbp diploid human genome (**Fig. 3a**). Notably, we observed that elements showed more variability in actuation between haplotypes than between Fiber-seq replicates (**Fig. S3**), highlighting the reproducibility of FIRE results between replicates, and the central role of individual haplotypes in guiding chromatin accessibility patterns.

To identify specific regulatory elements with haplotype-selective chromatin accessibility (HSCA), we compared the percent actuation between the two haplotypes for each element using a Fisher's exact test with a genome-wide FDR correction (Benjamini-Hochberg). This resulted in the identification of 9,773 peaks with nominally significant (p-value <0.05) HSCAs in GM12878 cells, with 1,231 of these peaks meeting genome-wide significance along the autosomes (FDR 5%) (**Fig. 3a,b**). As expected, haplotype-selective peaks were enriched at known imprinted sites, such as the *GNAS* locus, enabling the precise demarcation of actuated elements and TF binding events that are impacted by imprinting at these sites (**Fig. 3a**). However, known imprinting sites comprised only 5% of all haplotype-selective peaks (**Fig. 3c**), indicating that other features are the major drivers of haplotype-selective chromatin genome-wide.

Deterministic and non-deterministic haplotype-selective chromatin patterns

Next, we leveraged the paired long-read genomic and epigenetic information at each element to determine the contribution of genetic variation to HSCA signal. We found that only 36% of all HSCA elements contained a single genetic variant within the element (**Fig. 3c**), indicating that the majority of genome-wide significant HSCA elements would be poorly resolved using short-read approaches due to mapping issues associated with reads containing 0 or 2+ variants. Specifically, 44% of haplotype-selective elements contained no genetic variants within the peak, making these haplotype-selective architectures invisible to short-read allelic imbalance approaches. Notably, haplotype-selective elements lacking an underlying genetic variant had a CpG methylation pattern distinct from that of known imprinted variants (**Fig. 3d**), suggesting that these sites are not solely previously undiscovered imprinted elements. Moreover, 9% of all haplotype-selective elements contained three or more variants within the peak, with some containing over 10 variants – sequence differences that are ripe for mapping artifacts using short-read sequencing approaches.

To determine the stability of these haplotype-selective chromatin patterns, we performed Fiber-seq on two cell lines derived from the same donor (i.e., COLO829BL and COLO829T; **Fig. 4a**). As COLO829BL cells lack parental information, we haplotype-phased reads across entire chromosomes using a near telomere-to-telomere assembly that we constructed from our COLO829BL Fiber-seq data paired with COLO829BL ultra-long Oxford Nanopore and Hi-C data that we generated from the same cell line (contig N50 140.01 Mbp). This approach enabled us

to unambiguously phase 84.2% of all Fiber-seq reads from COLO829BL and use a consistent haplotype phasing for both COLO829BL and COLO829T, with 18 chromosomes having telomere-to-telomere scaffolds. Overall, we found that COLO829BL had 710 peaks with genome-wide significant HSCA and that the haplotype-selectivity of these peaks was in strong agreement across two COLO829BL Fiber-seq replicates, irrespective of the number of genetic variants within the HSCA element (Pearson's correlation > 0.9, p-value < 2.2e-16, two-sided *t*-test) (**Fig. 4b**). 29% (202/710) of these HSCA elements in COLO829BL also demonstrated some degree of chromatin actuation in the COLO829T cells, enabling us to evaluate the stability of HSCA at these 202 elements (**Fig. 4c**)

We find that for imprinted sites (R=0.87) or COLO829BL HSCA elements harboring genetic variants (R=0.76), there is a strong correlation in the HSCA between both cell types. Specifically, if one of these elements is selective to haplotype 1 in COLO829BL, it is also similarly selective to the same haplotype in COLO829T (**Fig. 4d,e**), a finding consistent with deterministic chromatin actuation that is dependent upon a site's underlying haplotype. In contrast, COLO829BL HSCA elements that lack an underlying genetic variant have markedly divergent haplotype-selective actuation between the two cell types (R=0.15). Specifically, a majority of these elements have similar chromatin actuation between the two haplotypes in COLO829T or completely flip their HSCA in the COLO829T cells (**Fig. 4d,e**), indicating that this class of elements overall displays a non-deterministic pattern of chromatin actuation that appears to be occurring irrespective of their underlying haplotype (**Fig. 4e,f**). Overall, this finding suggests that conservatively, on the order of ~100 regulatory elements in a cell may have non-deterministic behavior, creating somatically variable epialleles.

Resolving the chromatin impact of disease-associated variation within the HLA locus

Haplotype-selective sites with a large number of genetic variants often map to complex genomic regions like the HLA locus (**Fig. 5a**), which we could accurately haplotype phase using our long-read sequencing data despite the marked amount of sequence and structural variability between the two haplotypes (**Fig. S3**). To further explore haplotype-selective chromatin patterns within complex genomic regions like the HLA locus, we performed Fiber-seq on activated sorted $CD4^+$ T-cells and $CD8^+$ T-cells from 9 healthy donors. Shallow sequencing of each of these samples demonstrated that Fiber-seq enabled the reproducible mapping of genome-wide chromatin patterns along ~20kb long chromatin fibers from these primary cell types (**Fig. 5b**). Four of these samples were then sequenced to ~30x coverage, enabling the identification of haplotype-selective chromatin patterns. We observed that the HLA locus contained a marked amount of haplotype-selective chromatin patterns (**Fig. 5c**) and that the nucleotide-precise footprints obtained using Fiber-seq enabled us to elucidate specific SNVs that appear to be modulating TF occupancy within the HLA locus (**Fig. 5d**). For example, haplotype-selective Fiber-seq patterns within the *MOG1/HLA-F* locus demonstrated that among the lead SNPs associated with platelet counts, which is frequently an immune-mediated phenotype, only two appear to be present within putative regulatory elements, with one (rs29269) falling into a well-positioned Fiber-seq nucleosome footprint, and the other (rs4713235) directly disrupting single-molecule Fiber-seq TF occupancy at a CTCF binding element (**Fig. S4**). Together, these

findings highlight the utility of Fiber-seq for disentangling the relative contribution of non-coding variants within the HLA locus to human disease.

The HLA locus contains the highest rate of haplotype-selective chromatin

Next, we sought to extend this analysis genome-wide to evaluate whether certain genomic loci are preferentially marked by haplotype-selective chromatin. To accomplish this, we quantified the percentage of haplotype-selective peaks within rolling windows of 100 peaks and then calculated the enrichment of each window for haplotype-selective peaks (*Methods*). Aggregating all haplotype-selective peaks across the CD8+ T-cell experiments demonstrated that the MHC region contained the most haplotype-selective chromatin in the entire human genome (corrected p-value 0.0013, Benjamini-Hochberg corrected FDR < 5%) (**Fig. 5e**). Notably, this enrichment was largely localized to the HLA Class II locus of the MHC region (**Fig. 5f**). We next sought to see if more traditional antigen-presenting cells or non-immune cells display a similar enrichment for haplotype-selective chromatin within the HLA locus. Notably, lymphoblastoid cells showed significant enrichment of haplotype-selective chromatin within the HLA class II and HLA class I loci. In contrast, Fiber-seq data from fibroblasts or thyroid tissue did not display any enrichment in haplotype-selective chromatin within the MHC region, despite these regions having numerous actuated elements within these non-immune cell types (**Fig. S3**). Together, these findings indicate that although many MHC proteins are ubiquitously expressed across various cell types, the marked genetic diversity being maintained in the human population within the MHC region is largely driving haplotype-selective gene regulatory patterns at select loci in immune cells.

Given the marked divergence in chromatin accessibility between both HLA haplotypes within an individual, we next wanted to understand if there were genomic regions that showed higher variability in chromatin accessibility between both haplotypes within an individual (intra-individual) versus haplotypes between two individuals (inter-individual). To accomplish this, we compared haploid Fiber-seq data from the COLO829BL and GM12878 lymphoblastoid cell lines, which were derived from two separate donors. While the majority of the genome showed greater inter-individual variability, as would be expected, we did identify 40 extended genomic loci that did indeed show greater inter-individual similarity. Notably, these 40 loci were significantly enriched for alternative haplotypes, like the HLA locus (permutation test n=10,000, p<0.0001) and segmental duplications (permutation test n=10,000, p<0.0357). Together, these findings demonstrate that genomic regions containing haplotypes marked by the most genetic diversity within the human population (Dilthey et al., 2015; Sudmant et al., 2010, 2015) also show marked haplotype-selective chromatin patterns, suggesting that one of the selective pressures and/or consequences of these diverse haplotypes is also at the level of altered gene regulatory patterns (King & Wilson, 1975).

Comprehensive mapping of X chromosome inactivation using haplotype-selective chromatin

Finally, we sought to leverage haplotype-resolved chromatin patterns to generate comprehensive maps of gene regulation along the inactive X (Xi) and active X (Xa) chromosomes. Following random X chromosome inactivation (XCI) during early development, the inactivation states of each parental chromosome are consistently maintained throughout

each cell's lineage (Heard, 2004; Loda et al., 2022). Therefore, we reasoned that bottlenecking of the 46,XX cell line GM12878 would yield a clonal cell population with invariant Xi and Xa parental haplotypes (**Fig. 6a**). To validate this approach, we performed bulk full-length long-read transcript sequencing on bottlenecked GM12878 cells, enabling the haplotype phasing of 27% of all genes along the X chromosome, including the long non-coding RNA (lncRNA) *XIST*. *XIST* mediates XCI and is known to be specifically expressed from the inactive X (Borsani et al., 1991; Brockdorff et al., 1991; Brown et al., 1991; Panning et al., 1997), and we observed that 99% of *XIST* transcripts originate from the maternal haplotype (**Fig. S5**), confirming invariant inactivation of the maternal chromosome X haplotype within our bottlenecked GM12878 cells. This enabled us to use haplotype-specific genetic variants to unambiguously assign Fiber-seq reads to the Xa (paternal) or Xi (maternal).

We then quantified the proportion of Xa and Xi Fiber-seq reads that were accessible at each FIRE peak, which revealed Xa-biased chromatin accessibility chromosome-wide, with the exception of pseudoautosomal region 1 (PAR1; **Fig. 6b**). Notably, aside from the *XIST* promoter, the majority (57%) of elements selectively accessible on the Xi are within ICCE (Inactive-X CTCF-binding Contact Element) (Darrow et al., 2016), DXZ4 (Bonora et al., 2018; Giorgetti et al., 2016), and FIRRE (Functional Intergenic Repeating RNA Element) (Yang et al., 2015) – and 84.6% of these are CTCF structural elements known to organize the Xi into distinct three-dimensional mega-domains (Bansal et al., 2019) (**Fig. S6**). Next, we classified each FIRE peak as either 'Xa-specific,' 'Xi-specific,' or 'shared' (i.e., similar accessibility between both haplotypes) (**Fig. 6c**), and we stratified peaks by their location and by their overlap with GM12878 CAGE-positive transcriptional start sites (TSS), or GM12878 CTCF ChIP-seq peaks. Notably, only 64% of elements outside of PAR1 were Xa-specific, with 34% being shared and only 2.4% being Xi-specific elements, indicating that the chromatin compartment of the Xi is not inactivated but rather is punctuated by numerous accessible chromatin elements. However, 63% of these shared elements were CTCF elements, with the majority (58%) of X chromosome CTCF sites having similar accessibility along both the Xa and Xi. These shared CTCF sites were spread extensively along the entire Xi and to a greater degree than previously reported in mice (**Fig. 6d**) (Giorgetti et al., 2016).

Isoform- and nucleotide-precise mapping of XCI escaping transcripts and boundaries

In contrast to CTCF elements, we found that only 16% of TSSs have similar accessibility between Xa and Xi (i.e. escaping TSSs) (**Fig. 6c**), with the lone Xi-specific TSS being *XIST*. Notably, TSSs fully escaping XCI (i.e., equal promoter accessibility on Xa and Xi) clustered together and were confined to the p-arm. As transcript data from only 27% of the genes in GM12878 can be haplotype phased, to comprehensively validate the ability of Fiber-seq to identify genes escaping XCI, we compared our Xa and Xi chromatin accessibility data at each TSS against escape annotations from previous transcriptomic studies (Balaton et al., 2015; Oliva et al., 2020) (**Fig. 6e**). We found high concordance between Fiber-seq accessibility and escape status for all categories of escape annotations. Specifically, constitutively inactivated genes showed highly Xa-specific accessibility, genes known to escape variably between tissues and individuals showed a broad distribution, and PAR1 and consistently escaping genes primarily showed accessibility on both Xa and Xi.

Furthermore, for genes with multiple TSSs, Fiber-seq enabled the identification of individual TSSs that escape XCI. For example, the gene *UBA1* has four FIRE peaks that correspond to distinct *UBA1* TSSs based on the paired long-read transcript data. Whereas the three upstream *UBA1* TSSs are Xa-specific, the canonical *UBA1* TSS is fully accessible on both the Xa and Xi (**Fig. 6f**). Notably, the canonical *UBA1* TSS is bookended by an upstream CTCF ChIP-seq peak (**Fig. 6g**), and using the single-molecule m6A footprinting at the encompassed CTCF motif we find that CTCF is highly bound on both the Xa and Xi (90% and 63% of Fiber-seq reads respectively), suggesting that CTCF may be serving as a boundary element that insulates the canonical *UBA1* TSS from XCI, as proposed previously in mice (Giorgetti et al., 2016) (**Fig. 6h**). Of note, *UBA1* lacks heterozygous variants within the transcript sequence, making Fiber-seq the only possible haplotype-specific quantification of *UBA1* isoform usage.

Escaping promoter-proximal elements enable productive transcription from the Xi

Next, we sought to directly test the relationship between promoter accessibility and transcript production on the Xi using the paired haplotype-phased long-read transcript data. As expected, we observed highly skewed expression in genes with Xa-specific or Xi-specific TSS accessibility (**Fig. 6i**). However, genes with similar TSS accessibility between the Xa and Xi showed a distinct pattern, with some genes having similar transcript levels between the Xa and Xi (n=5), and others showing Xa-biased expression to a similar degree to that of genes with Xa-specific TSSs (n=4).

Given the widespread accessibility that we observed along the Xi, we hypothesized that productive transcription from the Xi may be augmented by the escape of promoter-proximal regulatory elements. To address this, we computed the average number of escaping non-TSS elements within 5-10 Kbp bins away from escaping or inactivated TSSs (**Fig. 6j**). This revealed enrichment of escaping regulatory elements within 100 Kbp of escaping TSSs, with the largest difference (126-fold increase) within 5 Kbp of the escaping TSS. Next, we tested whether the heterogeneity in transcript production from the Xi at genes with similar TSS accessibility between the Xa and Xi might be explained by heterogeneous escape of promoter-proximal enhancer elements. Consistent with this hypothesis, we found that genes that have similar TSS accessibility and transcript levels between the Xa and Xi had significantly more promoter-proximal escape elements compared with genes that have similar TSS promoter accessibility, but higher transcription from the Xa (11.2-fold increase, p-value = 0.037, one-sided Wilcoxon rank sum test; **Fig. 6k**). Overall, these results demonstrate that promoter accessibility is necessary but often not sufficient for robust gene expression from Xi, and that the simultaneous escape of promoter-proximal regulatory elements plays a pivotal role in maximizing the transcriptional potential of escape genes.

Discussion

We demonstrate that Fiber-seq enables a comprehensive view of chromatin accessibility across the 6 Gbp diploid human genome with single-molecule and single-haplotype precision at near nucleotide resolution. Specifically, we present a novel semi-supervised machine learning tool (FIRE), which we use to classify 225,183,940 actuated regulatory elements across 156,425,541 chromatin Fiber-seq reads. We demonstrate that FIRE enables the accurate *de novo*

construction of a cell's chromatin-accessible landscape directly using long-read sequencing data, enabling the advent of synchronous multi-omic long-read sequencing (Grasberger et al., 2024; Vollger, Korlach, et al., 2023) that can be used to fine-map the chromatin mechanisms by which disease-associated haplotypes and genetic variants contribute to human disease. Furthermore, we demonstrate that Fiber-seq FIRE provides a more comprehensive and quantitative snapshot of a cell's accessible chromatin landscape, overcoming previously known and unknown biases inherent to short-read bulk and single-cell approaches, such as the pervasive underreporting of chromatin accessibility at CTCF binding elements as well as the exploration of chromatin architectures within complex genomic loci. In addition, it accomplishes this with haplotype resolution.

This approach also disentangles gene regulation across the human inactive and active X chromosomes with nucleotide precision. We reveal that from an accessible chromatin perspective, the Xi is not inactive but rather is occupied by hundreds of accessible regulatory elements (n=777 in GM12878 cells), with 61% of these corresponding to CTCF elements. However, we find that promoter accessibility along the Xi is necessary but often not sufficient for robust gene expression from the Xi and that the simultaneous escape of promoter-proximal regulatory elements plays a pivotal role in maximizing the transcriptional potential of escape genes.

Overall, we find that ~99% of autosomal accessible elements show nearly identical levels of chromatin accessibility between the two haplotypes. Our findings demonstrate that chromatin accessibility is largely deterministic (i.e., a reflection of a cell's genomic sequence, cellular trans environment, and developmental history), indicating the potential of computational tools for accurately predicting the vast majority of a cell's chromatin accessibility landscape. As transvection is not known to impact human gene regulation, the chromatin patterns along each haplotype are independently formed during a cell's and organism's life span, a system ripe for stochastic deviations in the accessibility pattern between both haplotypes as a result of epigenetic noise and subsequent memory (Hathaway et al., 2012). We do conservatively uncover ~100 elements that appear to have non-deterministic behavior (i.e., their chromatin accessibility deviates between both haplotypes in a manner that cannot be readily explained by local genetic differences). Of note, our experimental design does not enable us to state definitively whether a specific element with HSCA has non-deterministic behavior. Rather, we find that, in general, elements with HSCA that lack an underlying genetic variant exhibit significantly higher variability in their haplotype-specific chromatin accessibility patterns between cells. This demonstrates that, unlike the other >99% of elements genome-wide, this class of elements lacks sequence deterministic behavior and potentially represents primary somatic epimutations (Horsthemke, 2006).

Finally, we find that genomic loci marked by the most genetic diversity within the human population (Dilthey et al., 2015; Sudmant et al., 2010, 2015) also contain the most haplotype-selective chromatin in the human genome. Combined with our findings that most regulatory elements show deterministic behavior, this finding indicates that genetic diversity at these loci is directly impacting an individual's gene regulatory network in a cell-selective manner. As changes in gene regulation are known to play a dominant role in human speciation (King &

Wilson, 1975), this 6 Gbp view of human gene regulation will likely markedly improve our understanding of the regulatory architectures that make us humans.

Figures

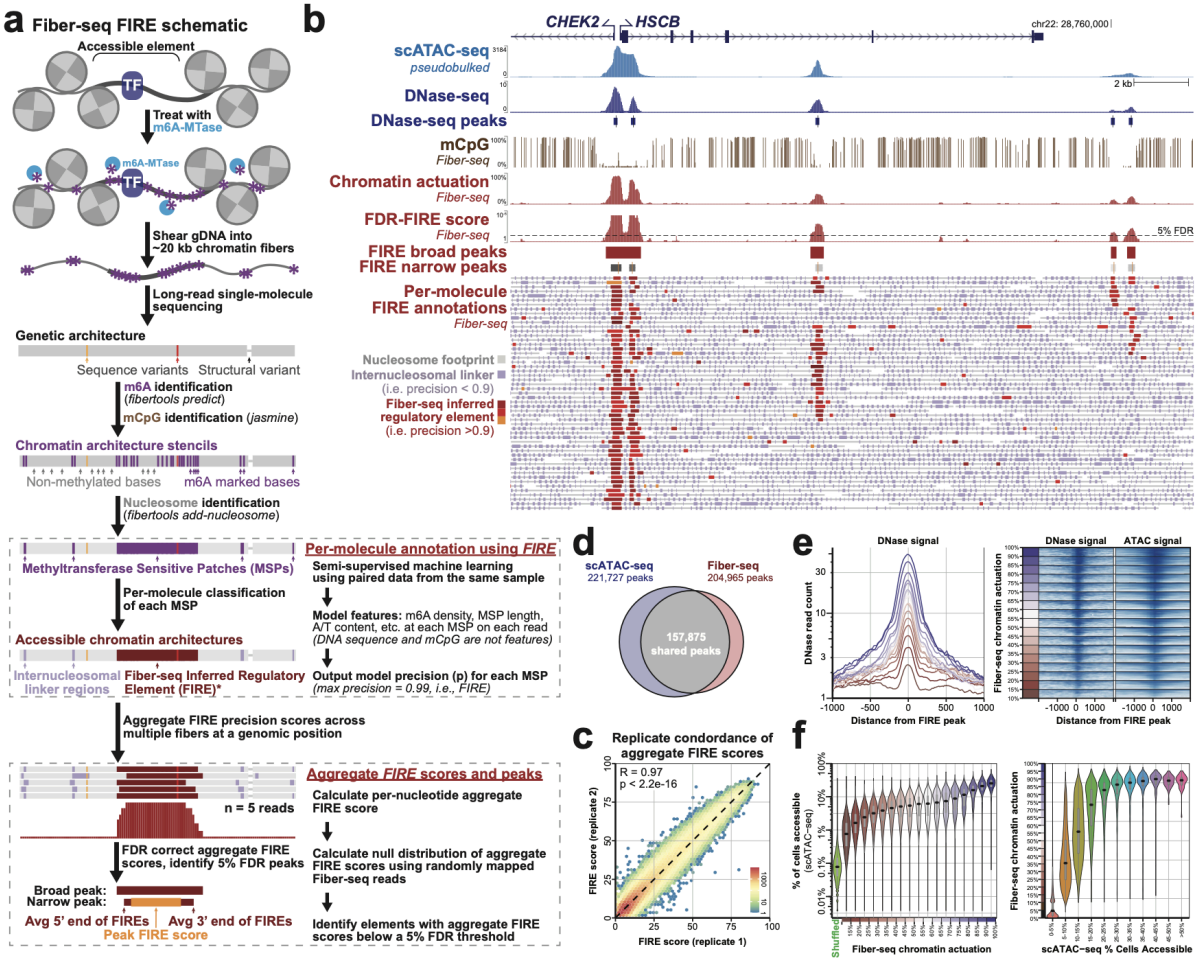


Figure 1. Fiber-seq Inferred Regulatory Elements and benchmarking against existing chromatin accessibility measures. a) A schematic of Fiber-seq experimental and computational processing, including the identification of Fiber-seq Inferred Regulatory Elements (FIREs). **b)** Genomic locus comparing the relationship between scATAC–seq, DNase–seq, mCpG, FIRE percent chromatin actuation, and FIRE peaks in GM12878. Below are individual Fiber-seq reads with MTase Sensitive Patches (MSPs) marked in purple, nucleosomes marked in gray, and FIRE elements marked in red. White regions separate individual reads. **c)** Correlation of FIRE score within the peaks of two technical replicates of COLO829BL (two-sided *t*-test). **d)** Venn diagram showing the overlap of FIRE and scATAC peaks in GM12878. **e)** (Left) Average count of DNase I reads over FIRE peaks binned by their percent actuation (red-blue color scale). (Right) Normalized scATAC and DNase I signal for 100 random FIRE peaks across each percent actuation bin. **f)** Comparison of percent actuation quantified by Fiber-seq and scATAC-seq. scATAC-seq accessibility values represent the fraction of single cells with at least one sequenced fragment overlapping the respective peak. FIRE peaks are binned by Fiber-seq percent actuation (left) and scATAC-seq percent actuation (right).

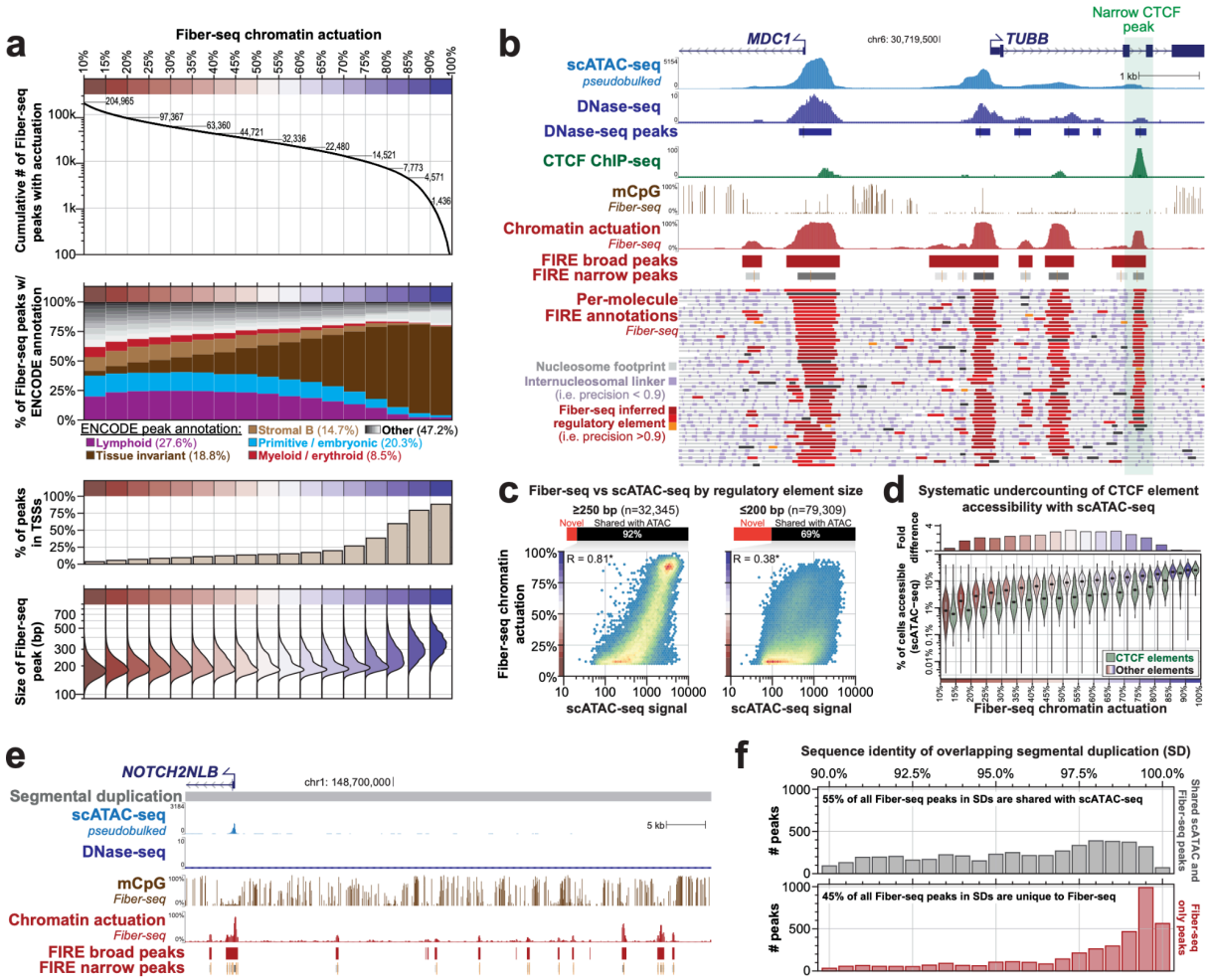


Figure 2. Chromatin features within FIRE-specific peaks. a) Features of FIRE peaks binned by percent actuation. **b)** Genomic locus comparing the relationship between scATAC-seq, DNase–seq, mCpG, CTCF ChIP-seq, and FIRE percent chromatin actuation and FIRE peaks in GM12878. A representative CTCF site with greater accessibility in Fiber-seq data than scATAC-seq and DNase–seq is highlighted in green (right). **c)** Correlation of FIRE percent actuation and scATAC-seq signal within FIRE peaks faceted by FIRE peak size (Pearson's correlation; $p < 2.2e-16$ two-sided *t*-test). **d)** Comparison of scATAC-seq signal to FIRE in peaks with (green) and without (red-blue) CTCF ChIP-seq peaks. **e)** Genomic locus of *NOTCH2NLB* comparing the ability to map into repetitive regions between scATAC-seq, DNase–seq, and Fiber-seq. **f)** FIRE peaks within segmental duplications stratified by the sequence identity of the underlying segmental duplication. FIRE peaks with a shared scATAC-seq peak are colored in gray, and peaks unique to FIRE are colored in red.

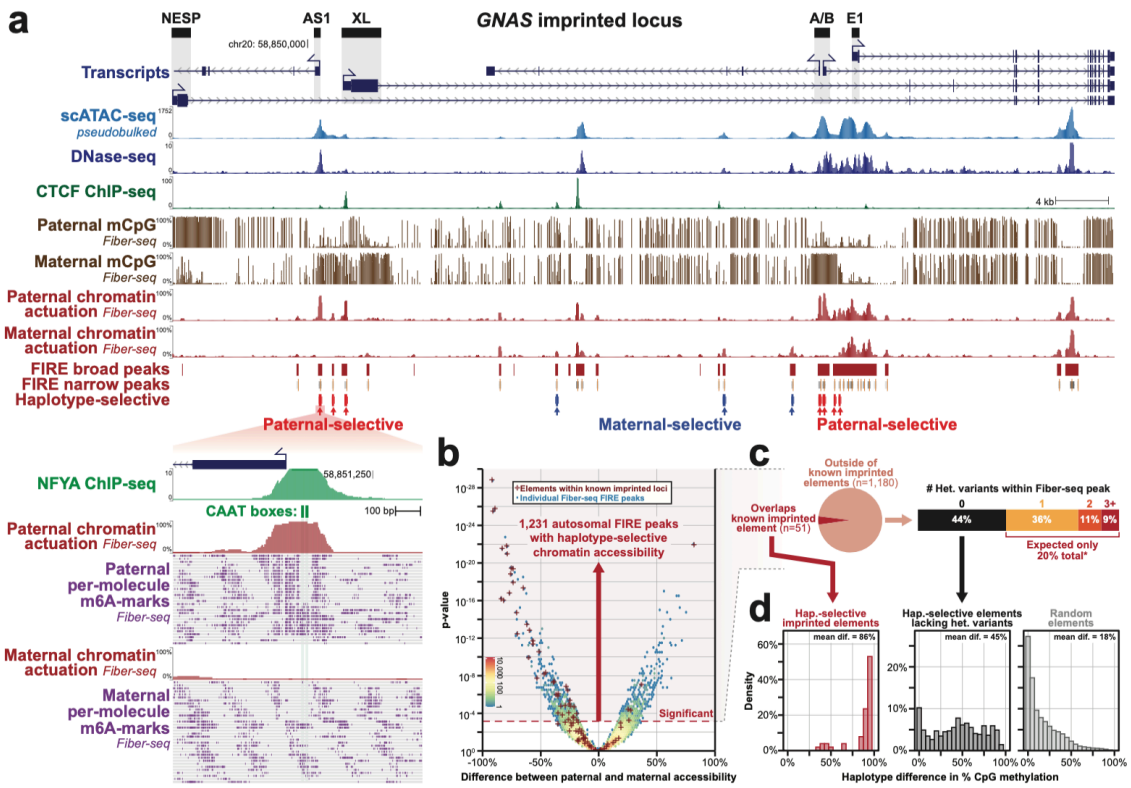


Figure 3. Haplotype-selective chromatin accessibility in GM12878. a) The *GNAS* imprinted locus comparing the relationship between *GNAS* isoforms, scATAC-seq, DNase–seq, CTCF ChIP-seq, mCpG, FIRE percent actuation, FIRE peaks, and maternally (blue) or paternally (red) haplotype-selective FIRE peaks in GM12878. Fiber-seq captures haplotype-selective chromatin architectures in both mCpG and chromatin actuation. **b)** FIRE haplotype-selective peaks (two-sided Fisher's exact test). The dashed line indicates genome-wide significance after applying a 5% FDR correction (Benjamini-Hochberg), and the red plus marks indicate known imprinted sites. **c)** Stratification of haplotype-selective peaks by imprinting status and the number of genetic variants within each peak. **d)** Histogram of the haplotype differences in percent CpG methylation for haplotype-selective peaks within imprinted sites, sites without genetic variants, and non-haplotype-selective peaks.

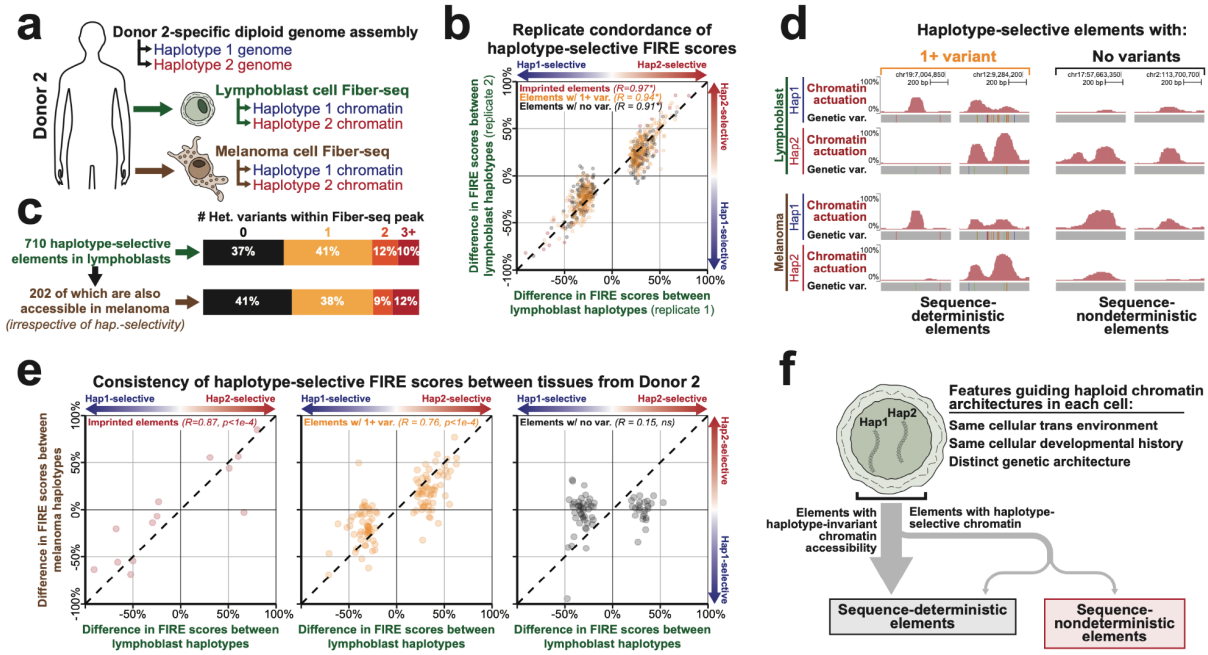


Figure 4. Deviation of haplotype-selective chromatin across cell types. a) Schematic of the sequencing of donor 2 of lymphoblast and melanoma cell lines. **b)** Replicate concordance of haplotype-selective percent actuation difference across two lymphoblast replicates for imprinted elements (red), elements with genetic variants between haplotypes (orange), and elements without genetic variants (black). Shown is the Pearson's correlation, all correlations are significant with $p < 2.2e-16$ (two-sided *t*-test). **c)** The overlap of 710 lymphoblast haplotype-selective peaks with genetic variants and the subset of those peaks (202) that also overlap peaks within the melanoma cell line. **d)** Example of shared haplotype-selective elements and unique haplotype-selective elements between lymphoblast and melanoma cells. **e)** Concordance of haplotype-selective peaks between lymphoblast and melanoma cells within imprinted sites, sites with genetic variants, and sites without genetic variants (Pearson's correlation; two-sided *t*-test: left $p = 1e-4$, middle $p < 2.2e-16$, right $p = 0.21$). **f)** Proposed model of haplotype-invariant elements versus haplotype-selective elements with deterministic and nondeterministic patterns.

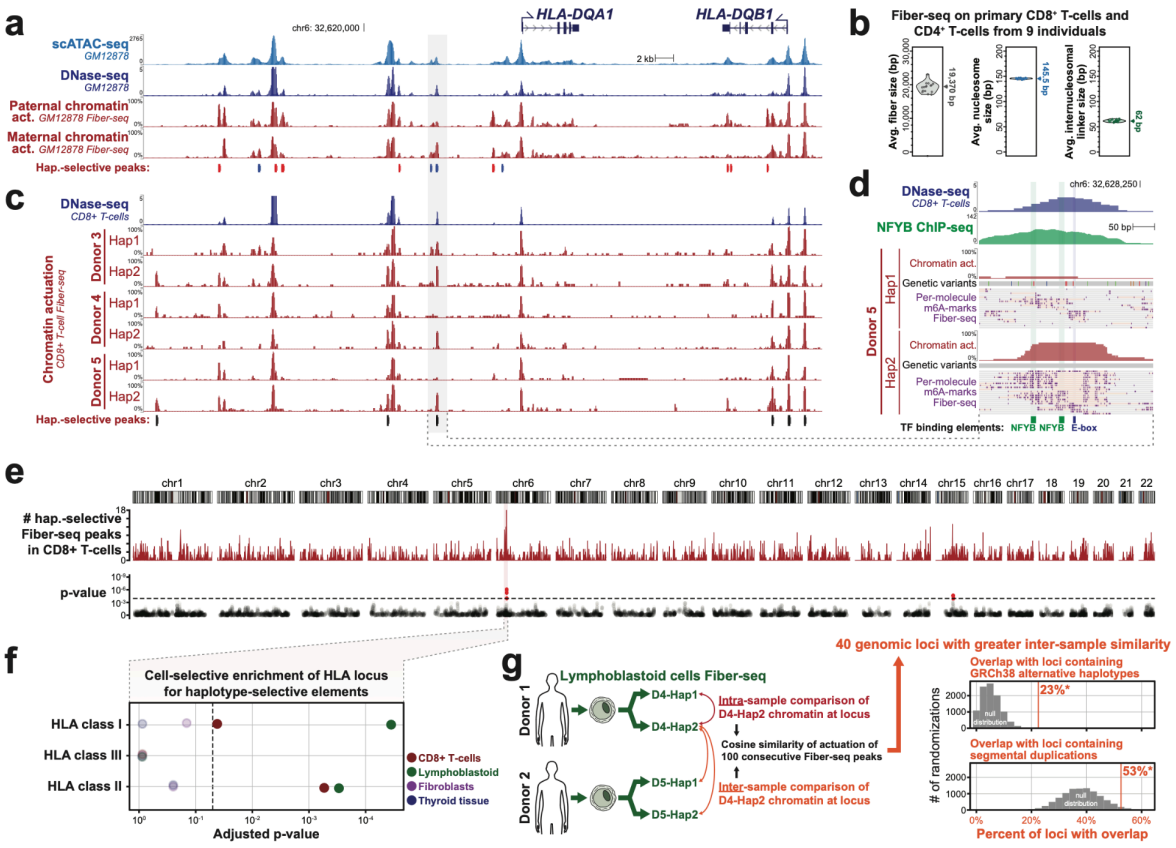


Figure 5. Haplotype-selective chromatin in the major histocompatibility complex. a) GM12878 haplotype-selective sites in the *HLA-DQA1*/*HLA-DQB1* locus. **b)** Average read length, nucleosome size, and internucleosomal linker size from Fiber-seq experiments on CD8+/CD4+ T-cells for 10 individuals. **c)** Haplotype-selective sites in the *HLA-DQA1*/*HLA-DQB1* locus for CD8+ T-cells sequenced to ~30-fold coverage across three individuals. **d)** Haplotype-selective Fiber-seq patterns showing disruption of single-molecule Fiber-seq TF occupancy and chromatin actuation due to the underlying haplotype. **e)** Ideogram of the number of haplotype-selective sites in high-coverage CD8+ T-cells and the two-sided Fisher's exact significance of the enrichment of haplotype-selective chromatin (*Methods*). **f)** Cell-selective enrichment of HLA class I, II, and III for haplotype-selective elements (two-sided Fisher's exact test). **g)** Schematic of testing intra- versus inter-sample cosine similarity between four haplotypes from two donors (GM12878 and COLO829BL) and enrichment of inter-sample similarity within GRCh38 alternative haplotypes ($p < 1e-04$; permutation test n=10,000; *Methods*) and segmental duplications ($p = 0.0357$; permutation test n=10,000; *Methods*).

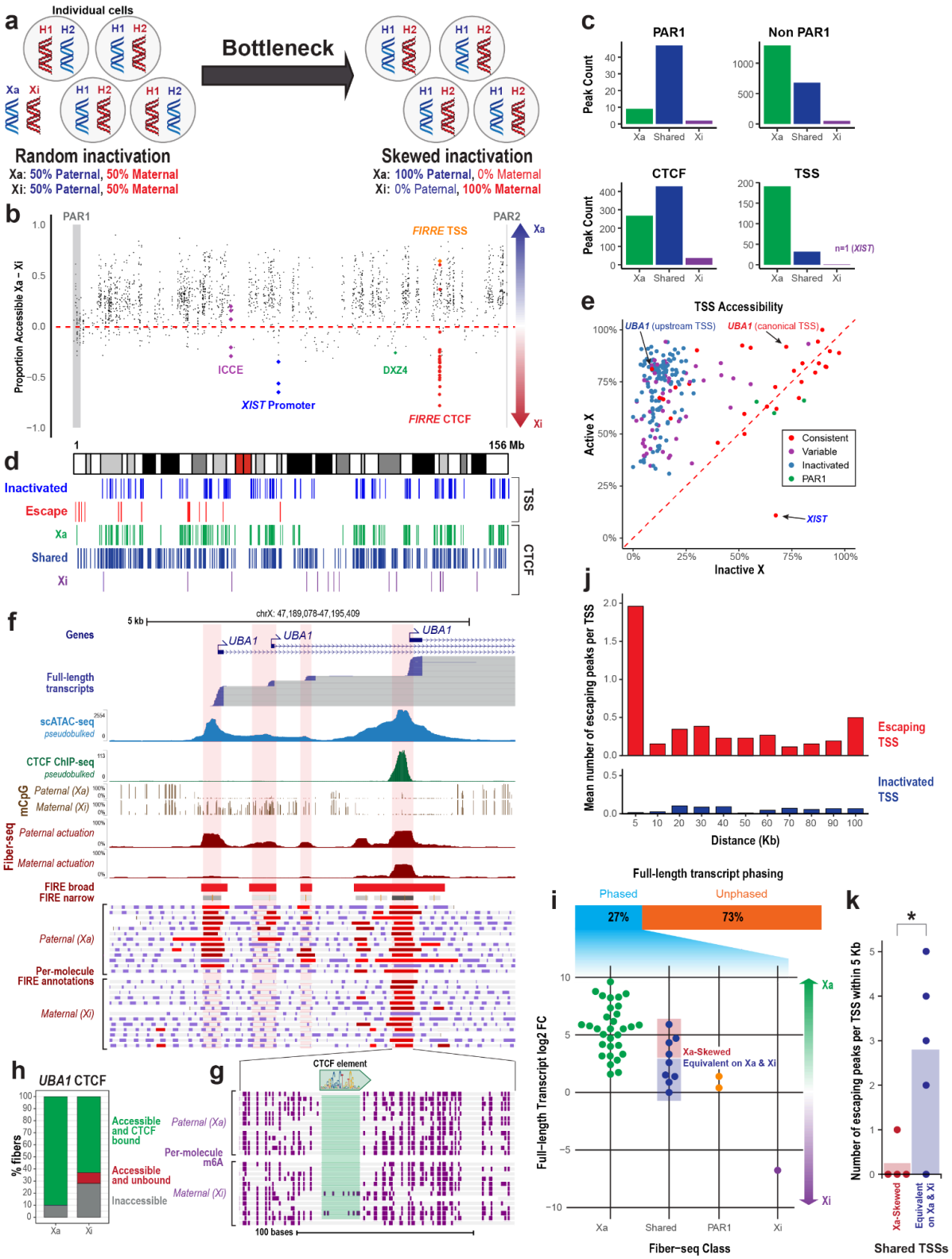

a
Individual cells
H1 H2
Xa Xi
Bottleneck
Random inactivation
Xa: 50% Paternal, 50% Maternal
Xi: 50% Paternal, 50% Maternal
Skewed inactivation
Xa: 100% Paternal, 0% Maternal
Xi: 0% Paternal, 100% Maternal
b
PAR1
PAR2
FIRRE TSS
Xa
Xi
Proportion Accessible Xa − Xi
ICCE
DXZ4
XIST Promoter
FIRRE CTCF
c
PAR1
Non PAR1
CTCF
TSS
Peak Count
Xa Shared Xi
n=1 (XIST)
d
1
156 Mb
Inactivated
Escape
Xa
Shared
Xi
TSS
CTCF
e
TSS Accessibility
UBA1 (upstream TSS)
UBA1 (canonical TSS)
Active X
Inactive X
Consistent
Variable
Inactivated
PAR1
XIST
f
chrX: 47,189,078-47,195,409
5 kb
Genes
UBA1
Full-length transcripts
scATAC-seq
pseudobulked
CTCF ChIP-seq
mCpG
Paternal (Xa)
Maternal (Xi)
Fiber-seq
Paternal actuation
Maternal actuation
FIRE broad
FIRE narrow
Per-molecule FIRE annotations
j
Mean number of escaping peaks per TSS
Escaping TSS
Inactivated TSS
Distance (Kb)
h
UBA1 CTCF
% fibers
Accessible and CTCF bound
Accessible and unbound
Inaccessible
g
CTCF element
Per-molecule m6A
100 bases
Full-length transcript phasing
i
Phased
Unphased
27%
73%
Full-length Transcript log2 FC
Xa-Skewed
Equivalent on Xa & Xi
Fiber-seq Class
k
Number of escaping peaks per TSS within 5 Kb
Xa-Skewed
Equivalent on Xa & Xi
Shared TSSs

Figure 6. Haplotype-specific features of X chromosome inactivation (XCI). a) Schematic of culture-derived XCI skewing. **b)** Chromosome-wide comparison between percent actuation of the paternal (Xa) and maternal (Xi) haplotypes at each FIRE peak. Pseudoautosomal regions (PAR1 & PAR2) are highlighted in gray. **c)** Counts of FIRE peaks categorized as Xa-specific, Xi-specific, or Shared between both haplotypes. FIRE elements are stratified by their location within or outside of PAR1 (top), and non-PAR1 elements are further subsetted to those that overlap a CTCF site or TSS (bottom). **d)** Ideogram showing the locations of TSSs classified as inactivated (blue) or full escapees (red) and CTCF sites classified as Xa-specific (green), shared (dark blue), and Xi-specific (purple). **e)** Scatterplot of Fiber-seq percent actuation on the Xi (x-axis) and Xa (y-axis) for each TSS. Points are colored by XCI escape annotations from previous studies (Balaton et al., 2015; Oliva et al., 2020). **f)** *UBA1* promoter region comparing full-length transcript reads, scATAC-seq, CTCF ChIP-seq, mCpG, FIRE percent actuation, FIRE peaks, and representative Fiber-seq reads from the paternal (Xa) and maternal (Xi) haplotypes. **g)** Zoom-in of an escaping *UBA1* TSS, which overlaps a highly-bound CTCF site. Representative Fiber-seq reads for each haplotype display individual m6A-modified bases as purple tick marks. **h)** Percentage of reads for each haplotype that contain a CTCF footprint (green), a FIRE element only (no CTCF footprint, red), or are inaccessible (no FIRE element, gray) at the CTCF site highlighted in h. **i)** Full-length transcript expression differences between the Xa and Xi for genes phased by Fiber-seq and displayed in e. Count differences are displayed as log2 fold-change between the haplotypes. Genes are stratified by the FIber-seq classifications of their TSS FIRE peaks as in c. **j)** The average number of escaping non-TSS FIRE peaks by absolute distance from TSSs. Counts are displayed separately for escaping TSSs (top, red) and inactivated TSSs (bottom, blue). **k)** The number of escaping non-TSS FIRE peaks within 5 Kb of each TSS in the shared category in **i**. Shared TSSs were grouped into high or low log2 fold-change in expression, highlighted with red and blue in **i**. Mean values are represented by colored bars. *$p = 0.03666$; one-sided Wilcoxon rank sum test.

Methods

Cell Culture

GM12878 were purchased from the Coriell Institute for Medical Research and cultured in RPMI 1640 Medium supplemented with 2mM L-glutamine,10% fetal bovine serum and100 I.U./mL penicillin/100 µg/mL streptomycin. Cell were maintained in a 37 °C humidified incubator under 5% carbon dioxide. Cell cultures were split every 3-5 days.

Fiber-seq Library Preparation and Sequencing

Cells were permeabilized and treated with Hia5 enzyme as previously described (Dubocanin et al., 2022). Specifically, 1 million cells were washed with PBS and then resuspended in 60 µl Buffer A (15 mM Tris, pH 8.0; 15 mM NaCl; 60 mM KCl; 1mM EDTA, pH 8.0; 0.5 mM EGTA, pH 8.0; 0.5 mM Spermidine) and 60 µl of cold 2X Lysis buffer (0.1% IGEPAL CA-630 in Buffer A for GM12878 and GM24385; 0.2% IGEPAL CA-630 in Buffer A for UDN318336 fibroblasts) was added and mixed by gentle flicking then kept on ice for 10 minutes. Samples were then pelleted, the supernatant removed, and then resuspended in 57.5 µl Buffer A and moved to a 25°C thermocycler. 0.5 µl of Hia5 MTase (100 U) and 1.5 µl 32 mM S-adenosylmethionine (NEB B9003S) (0.8 mM final concentration) were added, then carefully mixed by pipetting the volume up and down 10 times with wide bore tips. The reactions were incubated for 10 minutes at 25°C, then stopped with 3 µl of 20% SDS (1% final concentration) and transferred to new 1.5 mL microfuge tubes. High molecular weight DNA was then extracted using the Promega Wizard HMW DNA Extraction Kit A2920. PacBio SMRTbell libraries were then constructed using the Fiber-seq treated gDNA following the manufacturer's SMRTbell prep kit 3.0 procedure (https://www.pacb.com/wp content/uploads/Procedure-checklist-Preparing-whole-genome-and-metagenome-libraries-using SMRTbell-prep-kit-3.0.pdf).

Training of Fire-seq inferred regulatory elements.

For training data, we generated 21 different GM12878 Fiber-seq experiments with a range of under- and over-methylated experimental conditions to ensure we captured a broad range of percent m6A (**Fig. S1**; 5.8-13.3%) to ensure our model could generalize to new samples with varying levels of m6A. We merged these sequencing results and randomly selected 10% of the Fiber-seq reads from 100,000 randomly selected DNase I and CTCF ChIP-seq peaks for mixed-positive labels and 100,000 equally sized regions that did not overlap DNase or CTCF Chip-seq for negative labels. We then generated features for each of these MSPs, including length, log fold enrichment of m6A, the A/T content, and windowed measures of m6A around the MSP (**Supplemental Table 3**) and held out 20% of the Fiber-seq reads to be used as test data.

To carry out semi-supervised training, we used an established method, Mokapot (Fondrie & Noble, 2021; Käll et al., 2007), which we summarize below. In the first round of semi-supervised training, Mokapot identifies the feature that best discriminates between our mixed-positive and negative labels and then selects a threshold for that feature such that the mixed-positive labels can be discriminated from the negative labels with 95% estimated precision (defined below).

The subset of mixed-positive labels above this threshold is then used as an initial set of positive labels in training an XGBoost model with five-fold cross-validation (Chen & Guestrin, 2016). Then this process is iteratively repeated, using the learned prediction from the previous iteration's model to create positive labels at 95% estimated precision, until the number of positive identifications at 95% precision in the validation set ceases to increase (15 iterations, **Fig. S1**) (Fondrie & Noble, 2021; Käll et al., 2007).

Estimated precision of individual FIRE elements

We cannot compute the precision associated with a particular XGBoost score because we do not have access to a set of clean-positive labels. Instead, we define a notion of "estimated precision" using a balanced held-out test set of mixed-positive and negative labels (20% of the data). We defined the estimated precision (EP) of a FIRE element to be

$$EP = 1 - \frac{FP+1}{TMP}$$

where TMP is the number of "true" identifications from the mixed positive labels with at least that element's XGBoost score, and FP is the number of false positive identifications from negative labels with at least that element's XGBoost score. We add a pseudo count of one to the numerator of false positive identifications so as to prevent liberal estimates for smaller collections of identifications (Fondrie & Noble, 2021).

Aggregate FIRE score calculation

The FIRE score (S_g) for a position in the genome (g) is calculated using the following formula:

$$S_g = \frac{-50}{R_g} \sum_{i=1}^{C_g} \log_{10}(1 - min(EP_i, 0.99))$$

where C_g is the number of FIRE elements at the gth position, R_g is the number of Fiber-seq reads at the gth position, and EP_i is the estimated precision of the ith FIRE element at the gth position. The estimated precision of each FIRE element is thresholded at 0.99, such that the FIRE score takes on values between 0 and 100. Regions covered by less than four FIRE elements (i.e., if $C_g < 4$) are not scored and are given a value of negative one (**Fig. S2**).

Regions of unreliable coverage

Regions with unreliable coverage were defined as regions with Fiber-seq coverage that deviate from the median coverage by five standard deviations, where a standard deviation is defined by the Poisson distribution (i.e., the square root of the mean coverage).

FIRE score FDR calculation

We shuffle the location of an entire Fiber-seq read by selecting a random start position within the chromosome and relocating the entire read to that start position. Fiber-seq reads originating

from regions with unreliable coverage (defined above) are not shuffled, and reads from regions with reliable coverage are not shuffled into regions with unreliable coverage (bedtools shuffle -chrom -excl, v2.31.0) (Quinlan & Hall, 2010). We then compute FIRE scores associated with the shuffled genome. Recalling that the FIRE score for the gth position in the genome is denoted S_g, we divide the number of bases that have shuffled FIRE scores above S_g by the number of bases that have un-shuffled FIRE scores above S_g. This provides an estimate of the FDR associated with the FIRE score S_g.

Peak calling

Peaks are called by identifying FIRE score local-maxima that have FDR values below a 5% threshold and at least 10% actuation (C_g / R_g). Adjacent local maxima that share 50% of the underlying FIRE elements or have 90% reciprocal overlap are merged into a single peak, using the higher of the two local maxima. Then, the start and end positions of the peak are determined by the median start and end positions of the underlying FIRE elements. We also calculated and reported wide peaks by taking the union of the FIRE peaks and all regions below the FDR threshold and then merged the resulting regions that were within one nucleosome (147 bp) of one another.

Short-read accessibility data and comparisons

DNase I hypersensitivity sites and bigWig tracks for GM12878 were downloaded from the ENCODE data portal. We used accession ENCFF762CRQ for DNase peaks and ENCFF960FMM for the bigWig track (Meuleman et al., 2020). For CTCF CHiP-seq peaks, we used the union of ENCFF356LIU and ENCFF960ZGP (Neph, Vierstra, et al., 2012). When intersecting short-read and FIRE peaks, we required a 1 bp overlap, and when measuring the short-read signal within a FIRE peak, we used the maximum values of the short-read signal across the FIRE peak. scATAC-seq data was downloaded in fastq format from the ENCODE portal (**Supplemental Table 2**). Fastq data from each experiment were processed using cellranger-atac count, and the outputs were aggregated using cellranger-atac aggr (10x Genomics v.2.1.0). Aligned fragments from passing cell barcodes were intersected with FIRE peaks using bedmap (v2.4.41) (Neph, Kuehn, et al., 2012). Shuffled regions were generated using Bedtools shuffle (v2.31.0). scATAC-seq peaks were called using MACS2 callpeak (v2.2.7.1, parameters: -g hs -q 0.01 --nomodel --shift -75 --extsize 150 --keep-dup all -B --SPMR). scATAC-seq percent accessibility values were computed for each element in the FIRE and shuffled peak sets as the percentage of cell barcodes with at least 1 fragment overlapping that respective peak region.

Haplotype-selective peaks

For peaks with at least 10 Fiber-seq reads on both haplotypes, we tested the difference in percent actuation in each haplotype (fraction of reads with FIRE elements) using a two-sided Fisher's exact test. Specifically, the inputs for the test are the number of FIRE elements in haplotype one, the number of Fiber-seq reads without FIRE elements in haplotype one, the number of FIRE elements in haplotype two, and the number of Fiber-seq reads without FIRE

elements in haplotype two. We then apply an FDR correction (Benjamini-Hochberg) to correct for multiple testing.

The FIRE pipeline

The FIRE pipeline (https://github.com/fiberseq/FIRE v0.0.4) was applied to aligned and phased Fiber-seq BAM files with m6A predictions called using Fibertools-rs (v0.4) (Jha et al., 2023). The FIRE pipeline is a Snakemake (Mölder et al., 2021) workflow that applies the FIRE model to individual reads, calculates the aggregate FIRE scores, computes the FDR, calls peaks, and identifies haplotype-selective peaks. The only required inputs for the FIRE pipeline are an aligned Fiber-seq BAM file and the reference genome used for alignment.

Phasing and identification of genomic variants

Variant calling for single nucleotide variants (SNVs) and insertions/deletions (Indels) was conducted using DeepVariant version 1.5.0 (Poplin et al., 2018), while structural variants (SVs) were identified using pbsv (https://github.com/PacificBiosciences/pbsv). The Fiber-seq reads were haplotype-phased through a specialized pipeline available on GitHub (https://github.com/mrvollger/k-mer-variant-phasing) (Vollger, Korlach, et al., 2023). This pipeline employs SNVs detected by DeepVariant and utilizes the HiPhase (Holt et al., 2024) variant-based phaser to organize reads into phase blocks. These blocks are then attributed to either the maternal or paternal haplotype using parental short-read genome sequencing in conjunction with the trio k-mer-based phaser meryl (k=31) when available (Rhie et al., 2020).

Identification of CpG methylation

Base-level CpG methylation was called using jasmine (PacBio), and the percent CpG methylation at each genomic position was identified from a pileup of reads using pb-CpG-tools (https://github.com/PacificBiosciences/pb-CpG-tools).

Imprinted loci

Imprinted loci were defined using the 143 differentially methylated regions identified in 12 B-lymphocyte cell lines by Akbari et al. (Akbari et al., 2022).

Identifying regions enriched in haplotype selective elements

To test for regions enriched for haplotype-selective elements, we took consecutive windows containing 100 FIRE peaks (sliding by 10 peaks) and compared the number of nominally significant haplotype-selective elements in the 100 peak window to the number of haplotype-selective peaks genome-wide using a two-sided Fisher's exact test.

Intra- versus Inter-sample sample similarity in FIRE actuation

To measure the similarity between two haplotypes, we took genomic regions containing 100 peaks and measured the cosine similarity of the percent actuation between the two haplotypes. We then repeated this process across the genome in 100 peak windows, sliding 10 peaks at a time and comparing every unique pair of haplotypes across our two samples.

Testing for enrichment of inter-sample similarity within SDs and alternative haplotypes

To test for the enrichment of regions with greater inter-sample similarity, we compared the percent of sites intersecting with SDs or alternative haplotypes to 10,000 random shufflings of windows of the same size to calculate an empirical p-value.

Genome assembly of COLO829BL and estimation of the number of T2T scaffolds

To create a *de novo* assembly for COLO829BL, we used Verkko (v1.4.1) with downsampled PacBio HiFi reads (60-fold) and ONT ultra-long reads (60-fold) where ultra-long reads were downsampled in descending order based on read length to retain the longest reads. Additionally, we used 30-fold Illumina Hi-C data within Verkko to allow for long-range phasing of the assembly. To estimate the number of telomere-to-telomere (T2T) scaffolds, we required that a single scaffold covered more than 95% of a specific chromosome based on T2T-CHM13v2.0 alignment and that there were more than 20 occurrences of telomere sequences with 500bp of both ends of the contig.

X chromosome peaks

FIRE peaks used in XCI analyses were filtered to remove peaks overlapping ENCODE 2020 hg38 blacklist regions (accession: ENCFF356LFX). Transcriptional start site (TSS) FIRE peaks were identified as peaks intersecting both GENCODE v45 TSSs (padded to 20 bp) and ENCODE CAGE-seq peaks (**Supplemental Table 2**). CTCF FIRE peaks were identified as peaks intersecting ENCODE CTCF ChIP-seq peaks (**Supplemental Table 2**). Intersections were performed using Bedtools intersect (v2.31.0). Peaks included in XCI analyses were filtered to meet the following phasing and coverage criteria: <= 25% variance from the mean coverage, < 35% of reads assigned as unphased, and at least 10X coverage from each haplotype. In addition, peaks surrounding the MTMR1 TSS (chrX:150,631,401-150,738,995) were manually filtered due to substantial disagreement in kmer phasing. Peaks were required to be actuated on at least 30% of Fiber-seq reads on the Xa and/or Xi.

XCI Classifications

X chromosome FIRE peaks with a percent actuation difference of >= 50% between haplotypes were classified as “Xa-specific” or “Xi-specific” if the peaks were actuated on >= 30% of paternal or maternal Fiber-seq reads, respectively. FIRE Peaks were classified as “shared” if the peaks were actuated on >= 30% of Fiber-seq reads on both haplotypes or >= 30% of reads on one haplotype with a percent difference of < 50% between haplotypes. Transcriptional start site (TSS) FIRE peaks actuated on less than 30% of maternal Fiber-seq reads were classified as inactivated. TSS FIRE peaks were classified as fully escaping XCI if >= 30% of maternal Fiber-seq reads were actuated and the percent actuation difference between haplotypes was less than 25%.

Iso-seq analysis

Full-length 3’ transcriptomic iso-seq data was generated in a previous study (Vollger, Korlach, et al., 2023). Reads were clustered by isoform using Isoseq cluster 2 (PacBio v3.99.99) and

aligned to hg38 using pbmm2 (PacBio v1.11.99, parameters: --sort --min-gap-comp-id-perc 95.0 --min-length 50 --sample --report-json mapping_stats.report.json --preset ISOSEQ). Phased variant calls were used to haplotag clustered isoforms using WhatsHap v1.6 (Martin et al., 2016)). Clustered isoforms were then collapsed using Isoseq collapse v4.0.0 (PacBio). Isoforms were annotated using Pigeon v1.2.0 (PacBio). Isoforms and genes with >= 35% of transcripts not assigned to a haplotype were classified as unphased.

Promoter-proximal escape quantification

TSS FIRE peaks were grouped by escape status (fully escaping or inactivated). For each group, non-TSS FIRE peaks within 100 Kb in either direction (upstream and downstream) of each TSS were counted in 5 Kb distance bins. Counts within each bin were then normalized by the number of TSSs in the respective group.

CTCF footprinting

X chromosome CTCF FIRE peaks (overlapping ENCODE CTCF ChIP-seq peaks) that also fully overlap a FIMO-predicted CTCF motif (Grant et al., 2011) (v5.5) were used for footprinting analyses. We decided to limit CTCF footprinting to the most highly bound portions of CTCF, modules two and three (Yin et al., 2017). FIRE elements that completely overlapped modules two and three of a motif were centered using *fibertools* ft center (Jha et al., 2023), and the number of m6A observed within the motif was quantified. FIRE-contained motifs with <= 1 m6A were classified as bound.

Data availability

FIRE results and their associated trackHubs are available on Amazon Web Services at: https://s3-us-west-1.amazonaws.com/stergachis-manuscript-data/index.html?prefix=2024/Vollger_et_al/

Code availability

FIRE is available as a Snakemake pipeline on GitHub (https://github.com/fiberseq/FIRE) and Zenodo (https://zenodo.org/records/11075226). Fibertools is available on GitHub (https://github.com/fiberseq/fibertools-rs) and Zenodo (https://zenodo.org/records/10850620). The code used for phasing long-read data is available as a Snakemake pipeline on GitHub (https://github.com/mrvollger/k-mer-variant-phasing) and Zenodo (https://zenodo.org/records/10655527). The code for making figures and tables is available on GitHub (https://github.com/mrvollger/fire-figures) and Zenodo (https://zenodo.org/doi/10.5281/zenodo.10681988).

Competing interests

All authors declare no competing interests.

Acknowledgments

The authors thank Dr. John Stamatoyannopoulos for his assistance in reviewing this manuscript and Christine M. Disteche for feedback regarding the XCI escape portion of this study. The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM12878 and GM24385.

Funding

A.B.S. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and is a Pew Biomedical Scholar. This study was supported by National Institutes of Health (NIH) grants 1DP5OD029630, UM1DA058220, and OT2OD002748 to A.B.S. Additionally, M.R.V. and S.C.B. were supported by a training grant (T32) from the NIH (2T32GM007454-46).

Author contributions

Conceptualization and design: M.R.V., E.G.S., and A.B.S. Experimental design and execution: E.G.S., J.R., and A.B.S. Computational experiments: M.R.V., E.G.S., A.E.S., and S.J.N. Data generation: E.G.S., J.R., K.M.M., C.H., S.C.B., Y.M., N.L.P., B.J.M., G.H.G, K.H., J.G.M., M.C., E.E.E, J.T.B., and A.B.S. Genome Assembly: W.T.H., Y.K., and E.E.E. Conceptualization and design of the semi-supervised training regime and peak calling: M.R.V., D.M.W., W.E.F., W.S.N., and A.B.S. FIRE implementation: M.R.V. Supplemental material organization: M.R.V., E.G.S., and A.B.S. Display items: M.R.V., E.G.S., and A.B.S. Manuscript writing: M.R.V., E.G.S., and A.B.S. with input from all authors.

References

Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M. A., Hirst, M., & Jones, S. J. M. (2022). Genome-wide detection of imprinted differentially methylated regions using nanopore sequencing. *eLife*, *11*. https://doi.org/10.7554/eLife.77898

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*(1), 9354. https://doi.org/10.1038/s41598-019-45839-z

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., & Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science*, *297*(5583), 1003–1007. https://doi.org/10.1126/science.1072047

Balaton, B. P., Cotton, A. M., & Brown, C. J. (2015). Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biology of Sex Differences*, *6*, 35. https://doi.org/10.1186/s13293-015-0053-7

Bansal, P., Kondaveeti, Y., & Pinter, S. F. (2019). Forged by DXZ4, FIRRE, and ICCE: How Tandem Repeats Shape the Active and Inactive X Chromosome. *Frontiers in Cell and Developmental Biology*, *7*, 328. https://doi.org/10.3389/fcell.2019.00328

Battaglia, S., Dong, K., Wu, J., Chen, Z., Najm, F. J., Zhang, Y., Moore, M. M., Hecht, V., Shoresh, N., & Bernstein, B. E. (2022). Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. *Nature Genetics*, *54*(10), 1504–1513. https://doi.org/10.1038/s41588-022-01188-8

Bonora, G., Deng, X., Fang, H., Ramani, V., Qiu, R., Berletch, J. B., Filippova, G. N., Duan, Z., Shendure, J., Noble, W. S., & Disteche, C. M. (2018). Orientation-dependent Dxz4 contacts shape the 3D structure of the inactive X chromosome. *Nature Communications*, *9*(1), 1445. https://doi.org/10.1038/s41467-018-03694-y

Borsani, G., Tonlorenzi, R., Simmler, M. C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H. F., Avner, P., & Ballabio, A. (1991).

Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, *351*(6324), 325–329. https://doi.org/10.1038/351325a0

Brockdorff, N., Ashworth, A., Kay, G. F., Cooper, P., Smith, S., McCabe, V. M., Norris, D. P., Penny, G. D., Patel, D., & Rastan, S. (1991). Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature*, *351*(6324), 329–331. https://doi.org/10.1038/351329a0

Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R., & Willard, H. F. (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, *349*(6304), 38–44. https://doi.org/10.1038/349038a0

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Darrow, E. M., Huntley, M. H., Dudchenko, O., Stamenova, E. K., Durand, N. C., Sun, Z., Huang, S.-C., Sanborn, A. L., Machol, I., Shamim, M., Seberg, A. P., Lander, E. S., Chadwick, B. P., & Aiden, E. L. (2016). Deletion of *DXZ4* on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(31), E4504–E4512. https://doi.org/10.1073/pnas.1609643113

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., & McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, *47*(6), 682–688. https://doi.org/10.1038/ng.3257

Dubocanin, D., Sedeno Cortes, A. E., Hartley, G. A., Ranchalis, J., Agarwal, A., Logsdon, G. A.,

Munson, K. M., Real, T., Mallory, B. J., Eichler, E. E., O'Neill, R. J., & Stergachis, A. B. (2023). Conservation of chromatin organization within human and primate centromeres. In *bioRxiv* (p. 2023.04.20.537689). https://doi.org/10.1101/2023.04.20.537689

Dubocanin, D., Sedeno Cortes, A. E., Ranchalis, J., Real, T., Mallory, B., & Stergachis, A. B. (2022). Single-molecule architecture and heterogeneity of human telomeric DNA and chromatin. In *bioRxiv* (p. 2022.05.09.491186). https://doi.org/10.1101/2022.05.09.491186

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., … Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Fondrie, W. E., & Noble, W. S. (2021). mokapot: Fast and Flexible Semisupervised Learning for Peptide Detection. *Journal of Proteome Research*, *20*(4), 1966–1971. https://doi.org/10.1021/acs.jproteome.0c01010

Gershman, A., Sauria, M. E. G., Hook, P. W., Hoyt, S. J., Razaghi, R., Koren, S., Altemose, N., Caldas, G. V., Vollger, M. R., Logsdon, G. A., Rhie, A., Eichler, E. E., Schatz, M. C., O'Neill, R. J., Phillippy, A. M., Miga, K. H., & Timp, W. (2021). Epigenetic Patterns in a Complete Human Genome. In *bioRxiv* (p. 2021.05.26.443420). https://doi.org/10.1101/2021.05.26.443420

Giorgetti, L., Lajoie, B. R., Carter, A. C., Attia, M., Zhan, Y., Xu, J., Chen, C. J., Kaplan, N., Chang, H. Y., Heard, E., & Dekker, J. (2016). Structural organization of the inactive X chromosome in the mouse. *Nature*, *535*(7613), 575–579. https://doi.org/10.1038/nature18589

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* , *27*(7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Grasberger, H., Dumitrescu, A. M., Liao, X.-H., Swanson, E. G., Weiss, R. E., Srichomkwun, P.,

Pappa, T., Chen, J., Yoshimura, T., Hoffmann, P., França, M. M., Tagett, R., Onigata, K., Costagliola, S., Ranchalis, J., Vollger, M. R., Stergachis, A. B., Chong, J. X., Bamshad, M. J., … Refetoff, S. (2024). STR mutations on chromosome 15q cause thyrotropin resistance by activating a primate-specific enhancer of MIR7-2/MIR1179. *Nature Genetics*, *56*(5), 877–888. https://doi.org/10.1038/s41588-024-01717-7

Guarracino, A., Buonaiuto, S., de Lima, L. G., Potapova, T., Rhie, A., Koren, S., Rubinstein, B., Fischer, C., Human Pangenome Reference Consortium, Gerton, J. L., Phillippy, A. M., Colonna, V., & Garrison, E. (2023). Recombination between heterologous human acrocentric chromosomes. *Nature*, *617*(7960), 335–343. https://doi.org/10.1038/s41586-023-05976-y

Hathaway, N. A., Bell, O., Hodges, C., Miller, E. L., Neel, D. S., & Crabtree, G. R. (2012). Dynamics and memory of heterochromatin in living cells. *Cell*, *149*(7), 1447–1460. https://doi.org/10.1016/j.cell.2012.03.052

Heard, E. (2004). Recent advances in X-chromosome inactivation. *Current Opinion in Cell Biology*, *16*(3), 247–255. https://doi.org/10.1016/j.ceb.2004.03.005

Holt, J. M., Saunders, C. T., Rowell, W. J., Kronenberg, Z., Wenger, A. M., & Eberle, M. (2024). HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* , *40*(2). https://doi.org/10.1093/bioinformatics/btae042

Horsthemke, B. (2006). Epimutations in human disease. *Current Topics in Microbiology and Immunology*, *310*, 45–59. https://doi.org/10.1007/3-540-31181-5_4

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4). https://doi.org/10.1038/nbt.4060

Jha, A., Bohaczuk, S. C., Mao, Y., Ranchalis, J., Mallory, B. J., Min, A. T., Hamm, M. O.,

Swanson, E., Finkbeiner, C., Li, T., Whittington, D., Noble, W. S., Stergachis, A. B., & Vollger, M. R. (2023). Fibertools: fast and accurate DNA-m6A calling using single-molecule long-read sequencing. *bioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.04.20.537673

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*(11), 923–925. https://doi.org/10.1038/nmeth1113

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, *188*(4184), 107–116. https://doi.org/10.1126/science.1090005

Lee, I., Razaghi, R., Gilpatrick, T., Molnar, M., Gershman, A., Sadowski, N., Sedlazeck, F. J., Hansen, K. D., Simpson, J. T., & Timp, W. (2020). Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods*, *17*(12), 1191–1199. https://doi.org/10.1038/s41592-020-01000-7

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

Loda, A., Collombet, S., & Heard, E. (2022). Gene regulation in time and space during X-chromosome inactivation. *Nature Reviews. Molecular Cell Biology*, *23*(4), 231–249. https://doi.org/10.1038/s41580-021-00438-7

Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P. C., Rhie, A., de Lima, L. G., Dvorkina, T., Porubsky, D., Harvey, W. T., Mikheenko, A., Bzikadze, A. V., Kremitzki, M., Graves-Lindsay, T. A., Jain, C., … Eichler, E. E. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, *593*(7857), 101–107. https://doi.org/10.1038/s41586-021-03420-7

Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., Schöenhuth, A., & Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. In *bioRxiv* (p. 085050). https://doi.org/10.1101/085050

Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., Reynolds, A., Haugen, E., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Sandstrom, R., Vierstra, J., Kaul, R., & Stamatoyannopoulos, J. (2020). Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, *584*(7820), 244–251. https://doi.org/10.1038/s41586-020-2559-3

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*(33), 33. https://doi.org/10.12688/f1000research.29032.2

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics* , *28*(14), 1919–1920. https://doi.org/10.1093/bioinformatics/bts277

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., Maurano, M. T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., … Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints.

Nature, *489*(7414), 83–90. https://doi.org/10.1038/nature11212

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gewirtz, A. D. H., Cotter, D. J., Parsana, P., Kasela, S., Balliu, B., Viñuela, A., Castel, S. E., Mohammadi, P., Aguet, F., Zou, Y., Khramtsova, E. A., Skol, A. D., Garrido-Martín, D., Reverter, F., Brown, A., … Stranger, B. E. (2020). The impact of sex on gene expression across human tissues. *Science*, *369*(6509). https://doi.org/10.1126/science.aba3066

Panning, B., Dausman, J., & Jaenisch, R. (1997). X chromosome inactivation is mediated by Xist RNA stabilization. *Cell*, *90*(5), 907–916. https://doi.org/10.1016/s0092-8674(00)80355-4

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* , *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., Eichler, E. E., Phillippy, A. M., & Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01662-6

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1),

245. https://doi.org/10.1186/s13059-020-02134-9

Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., Stewart, H., Price, S. M., Blair, E., Hennekam, R. C., Fitzpatrick, C. A., Segraves, R., Richmond, T. A., Guiver, C., Albertson, D. G., Pinkel, D., Eis, P. S., Schwartz, S., Knight, S. J. L., & Eichler, E. E. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics*, *38*(9), 1038–1042. https://doi.org/10.1038/ng1862

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., & Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*, *77*(1), 78–88. https://doi.org/10.1086/431652

Shipony, Z., Marinov, G. K., Swaffer, M. P., Sinnott-Armstrong, N. A., Skotheim, J. M., Kundaje, A., & Greenleaf, W. J. (2020). Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nature Methods*, *17*(3), 319–327. https://doi.org/10.1038/s41592-019-0730-2

Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S., & Stamatoyannopoulos, J. A. (2020). Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science (New York, N.Y.)*, *368*(6498), 1449–1454. https://doi.org/10.1126/science.aaz1646

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., & Shendure, J. (2010). Diversity of human copy number. *Science* , *11184*(2009), 2–7. https://doi.org/papers2://publication/uuid/C37D7A2A-43D1-4164-A5DE-4447C58EBC0D

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81.

https://doi.org/10.1038/nature15394

Vollger, M. R., Dishuck, P. C., Harvey, W. T., DeWitt, W. S., Guitart, X., Goldberg, M. E., Rozanski, A. N., Lucas, J., Asri, M., Human Pangenome Reference Consortium, Munson, K. M., Lewis, A. P., Hoekzema, K., Logsdon, G. A., Porubsky, D., Paten, B., Harris, K., Hsieh, P., & Eichler, E. E. (2023). Increased mutation and gene conversion within human segmental duplications. *Nature*, *617*(7960), 325–334. https://doi.org/10.1038/s41586-023-05895-y

Vollger, M. R., Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K. M., Lewis, A. P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K. H., Phillippy, A. M., Timp, W., Ventura, M., & Eichler, E. E. (2022). Segmental duplications and their variation in a complete human genome. *Science*, *376*(6588), eabj6965. https://doi.org/10.1126/science.abj6965

Vollger, M. R., Korlach, J., Eldred, K. C., Swanson, E., Underwood, J. G., Cheng, Y.-H. H., Ranchalis, J., Mao, Y., Blue, E. E., Schwarze, U., Munson, K. M., Saunders, C. T., Wenger, A. M., Allworth, A., Chanprasert, S., Duerden, B. L., Glass, I., Horike-Pyne, M., Kim, M., … Stergachis, A. B. (2023). Synchronized long-read genome, methylome, epigenome, and transcriptome for resolving a Mendelian condition. In *bioRxiv* (p. 2023.09.26.559521). https://doi.org/10.1101/2023.09.26.559521

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Yang, F., Deng, X., Ma, W., Berletch, J. B., Rabaia, N., Wei, G., Moore, J. M., Filippova, G. N., Xu, J., Liu, Y., Noble, W. S., Shendure, J., & Disteche, C. M. (2015). The lncRNA Firre

anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biology*, *16*(1), 52. https://doi.org/10.1186/s13059-015-0618-0

Yin, M., Wang, J., Wang, M., Li, X., Zhang, M., Wu, Q., & Wang, Y. (2017). Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Research*, *27*(11), 1365–1377. https://doi.org/10.1038/cr.2017.131

Figure 1

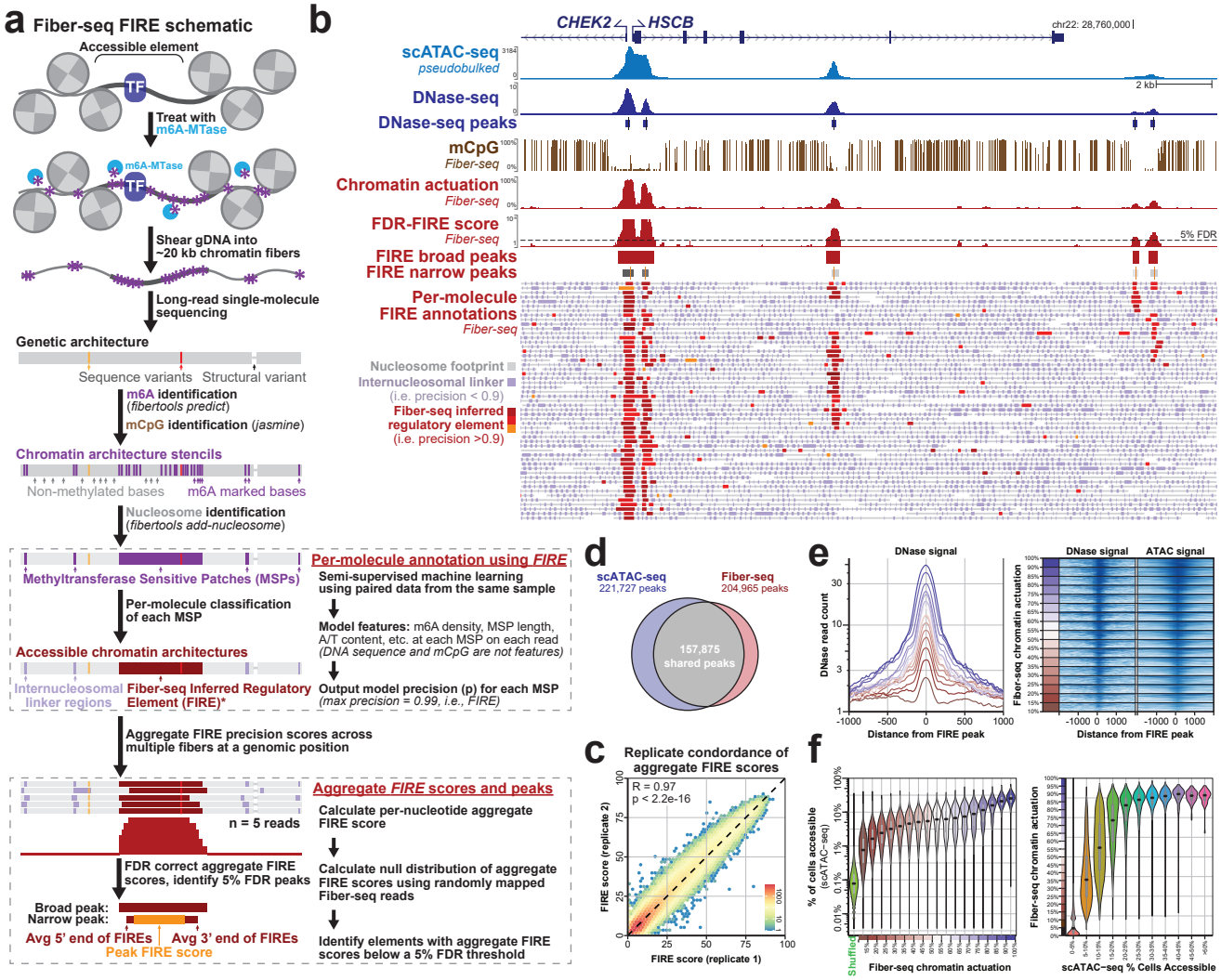

a Fiber-seq FIRE schematic
Accessible element
TF
Treat with m6A-MTase
m6A-MTase
Shear gDNA into ~20 kb chromatin fibers
Long-read single-molecule sequencing
Genetic architecture
Sequence variants
Structural variant
m6A identification (fibertools predict)
mCpG identification (jasmine)
Chromatin architecture stencils
Non-methylated bases
m6A marked bases
Nucleosome identification (fibertools add-nucleosome)
Methyltransferase Sensitive Patches (MSPs)
Per-molecule classification of each MSP
Accessible chromatin architectures
Internucleosomal linker regions
Fiber-seq Inferred Regulatory Element (FIRE)*
Per-molecule annotation using FIRE
Semi-supervised machine learning using paired data from the same sample
Model features: m6A density, MSP length, A/T content, etc. at each MSP on each read (DNA sequence and mCpG are not features)
Output model precision (p) for each MSP (max precision = 0.99, i.e., FIRE)
Aggregate FIRE precision scores across multiple fibers at a genomic position
n = 5 reads
FDR correct aggregate FIRE scores, identify 5% FDR peaks
Broad peak:
Narrow peak:
Avg 5' end of FIREs
Avg 3' end of FIREs
Peak FIRE score
Aggregate FIRE scores and peaks
Calculate per-nucleotide aggregate FIRE score
Calculate null distribution of aggregate FIRE scores using randomly mapped Fiber-seq reads
Identify elements with aggregate FIRE scores below a 5% FDR threshold
b
CHEK2
HSCB
chr22: 28,760,000
scATAC-seq pseudobulked
DNase-seq
DNase-seq peaks
mCpG Fiber-seq
Chromatin actuation Fiber-seq
FDR-FIRE score Fiber-seq
5% FDR
FIRE broad peaks
FIRE narrow peaks
Per-molecule FIRE annotations Fiber-seq
Nucleosome footprint
Internucleosomal linker (i.e. precision < 0.9)
Fiber-seq inferred regulatory element (i.e. precision >0.9)
2 kb
d
scATAC-seq 221,727 peaks
Fiber-seq 204,965 peaks
157,875 shared peaks
e
DNase signal
DNase read count
Distance from FIRE peak
Fiber-seq chromatin actuation
ATAC signal
c
Replicate condordance of aggregate FIRE scores
R = 0.97
p < 2.2e-16
FIRE score (replicate 1)
FIRE score (replicate 2)
f
% of cells accessible (scATAC-seq)
Shuffled
Fiber-seq chromatin actuation
scATAC-seq % Cells Accessible

Figure 2

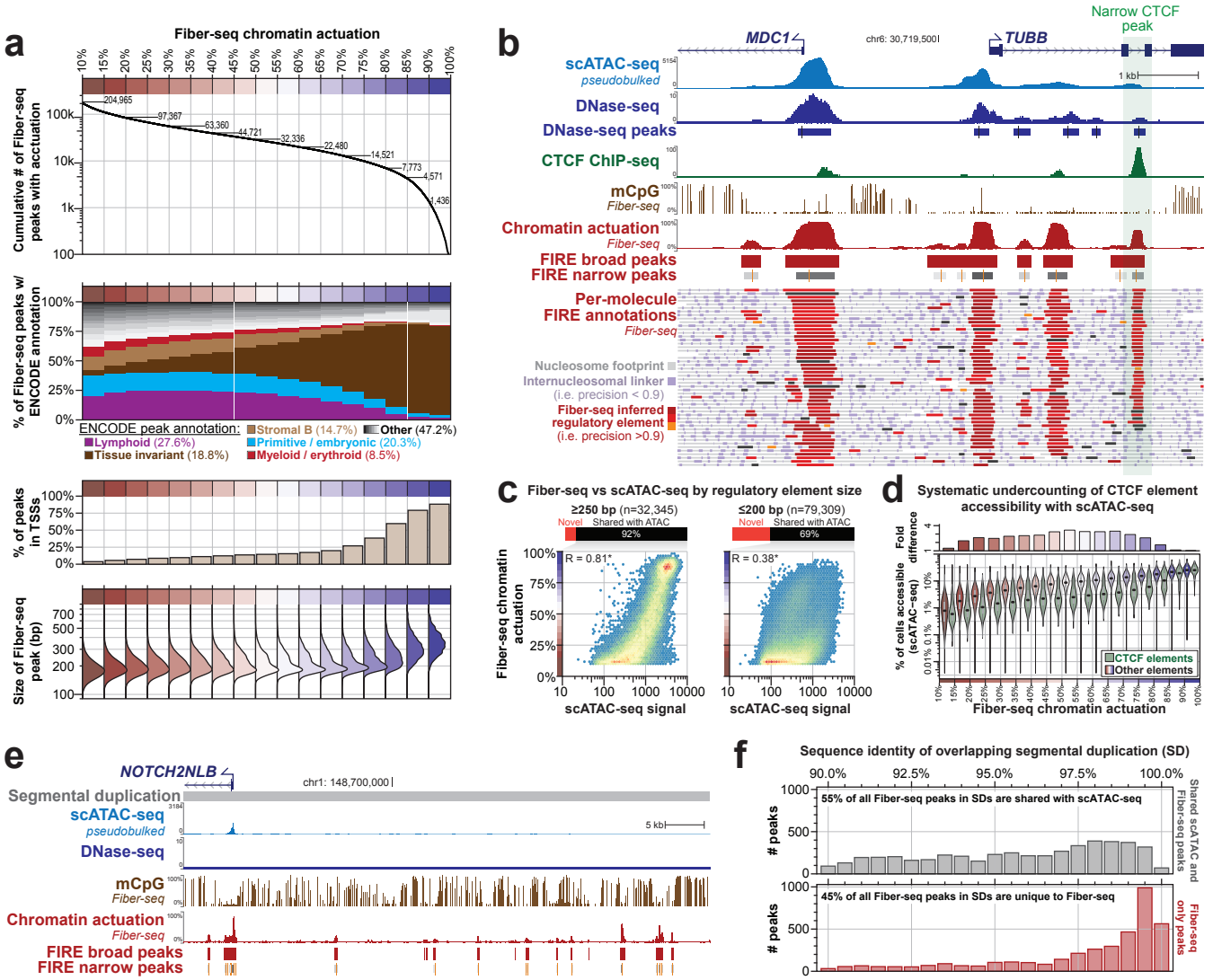

a
Fiber-seq chromatin actuation
Cumulative # of Fiber-seq peaks with acctuation
% of Fiber-seq peaks w/ ENCODE annotation
ENCODE peak annotation:
Stromal B (14.7%)
Other (47.2%)
Lymphoid (27.6%)
Primitive / embryonic (20.3%)
Tissue invariant (18.8%)
Myeloid / erythroid (8.5%)
% of peaks in TSSs
Size of Fiber-seq peak (bp)
b
MDC1
TUBB
Narrow CTCF peak
scATAC-seq
pseudobulked
DNase-seq
DNase-seq peaks
CTCF ChIP-seq
mCpG
Fiber-seq
Chromatin actuation
FIRE broad peaks
FIRE narrow peaks
Per-molecule FIRE annotations
Nucleosome footprint
Internucleosomal linker
(i.e. precision < 0.9)
Fiber-seq inferred regulatory element
(i.e. precision >0.9)
c
Fiber-seq vs scATAC-seq by regulatory element size
≥250 bp (n=32,345)
≤200 bp (n=79,309)
Novel
Shared with ATAC
92%
69%
R = 0.81*
R = 0.38*
Fiber-seq chromatin actuation
scATAC-seq signal
d
Systematic undercounting of CTCF element accessibility with scATAC-seq
Fold difference
% of cells accessible (scATAC-seq)
CTCF elements
Other elements
Fiber-seq chromatin actuation
e
NOTCH2NLB
chr1: 148,700,000
Segmental duplication
scATAC-seq
pseudobulked
DNase-seq
mCpG
Fiber-seq
Chromatin actuation
FIRE broad peaks
FIRE narrow peaks
f
Sequence identity of overlapping segmental duplication (SD)
55% of all Fiber-seq peaks in SDs are shared with scATAC-seq
45% of all Fiber-seq peaks in SDs are unique to Fiber-seq
peaks
Shared scATAC and Fiber-seq peaks
Fiber-seq only peaks

Figure 3

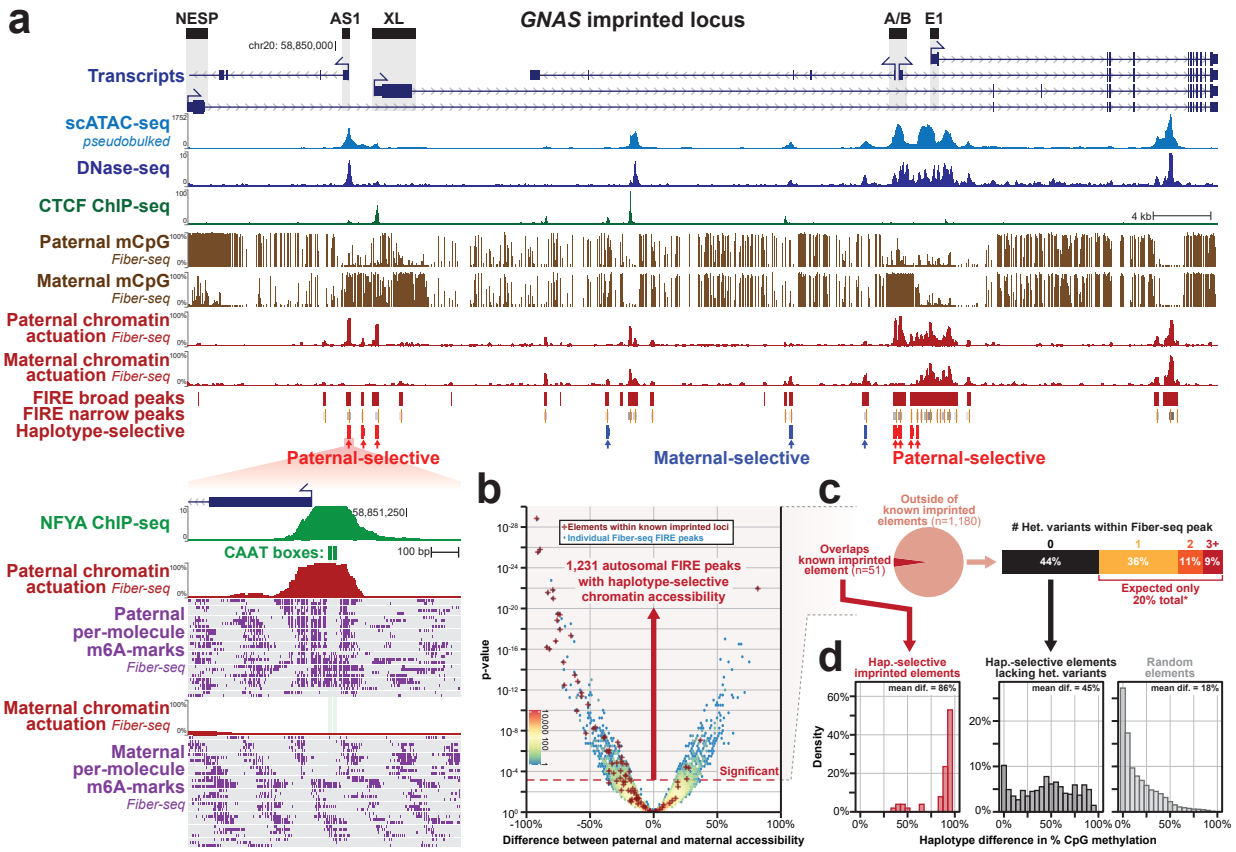

a
NESP
AS1
XL
GNAS imprinted locus
A/B
E1
chr20: 58,850,000
Transcripts
scATAC-seq
pseudobulked
DNase-seq
CTCF ChIP-seq
4 kb
Paternal mCpG
Fiber-seq
Maternal mCpG
Fiber-seq
Paternal chromatin actuation Fiber-seq
Maternal chromatin actuation Fiber-seq
FIRE broad peaks
FIRE narrow peaks
Haplotype-selective
Paternal-selective
Maternal-selective
Paternal-selective
NFYA ChIP-seq
58,851,250
CAAT boxes:
100 bp
Paternal chromatin actuation Fiber-seq
Paternal per-molecule m6A-marks Fiber-seq
Maternal chromatin actuation Fiber-seq
Maternal per-molecule m6A-marks Fiber-seq
b
Elements within known imprinted loci
Individual Fiber-seq FIRE peaks
1,231 autosomal FIRE peaks with haplotype-selective chromatin accessibility
Significant
p-value
Difference between paternal and maternal accessibility
c
Outside of known imprinted elements (n=1,180)
Overlaps known imprinted element (n=51)
Het. variants within Fiber-seq peak
0
1
2
3+
44%
36%
11%
9%
Expected only 20% total*
d
Hap.-selective imprinted elements
Hap.-selective elements lacking het. variants
Random elements
mean dif. = 86%
mean dif. = 45%
mean dif. = 18%
Density
Haplotype difference in % CpG methylation

Figure 4

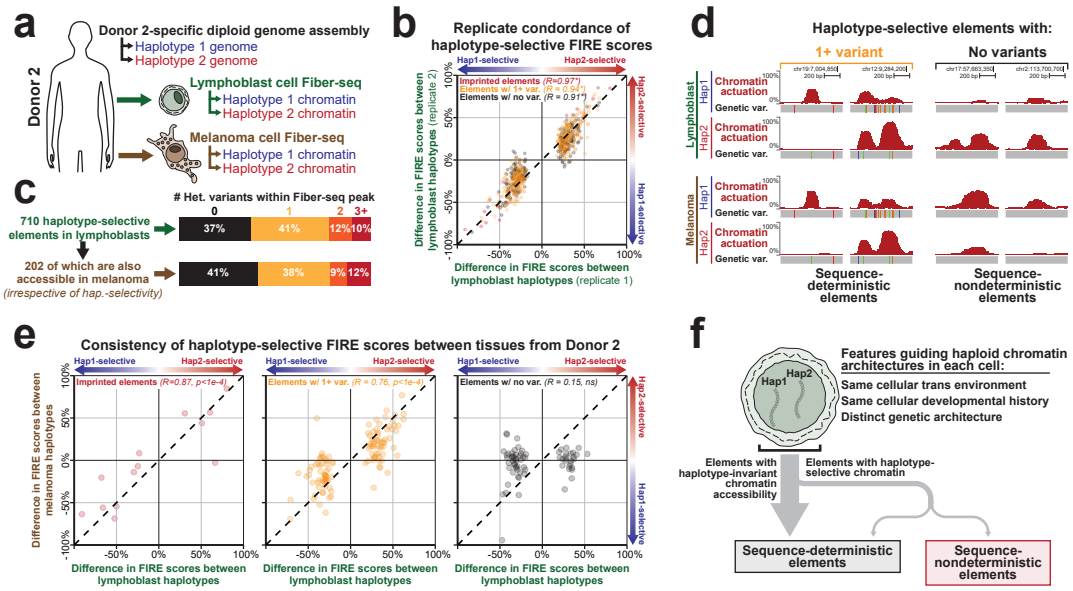
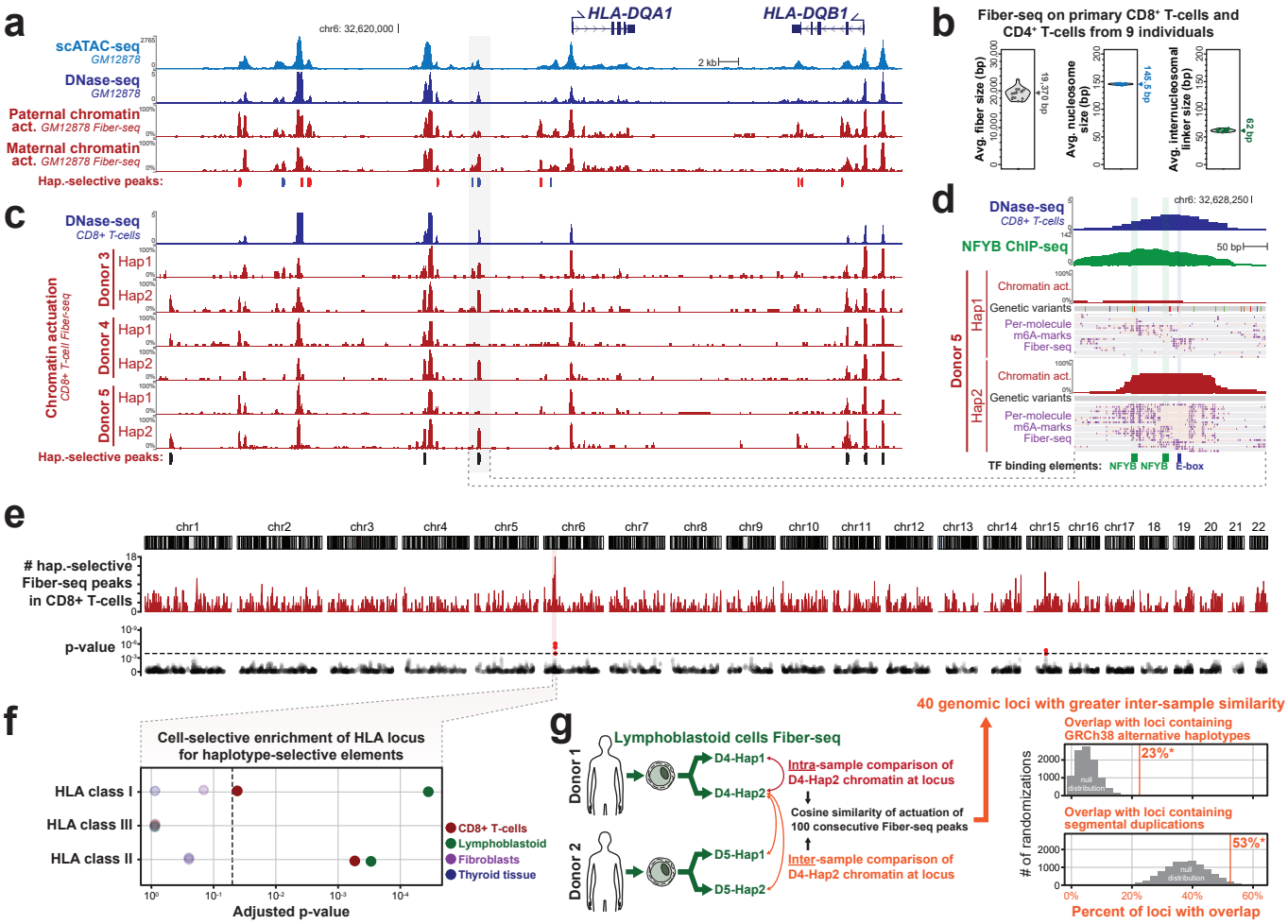
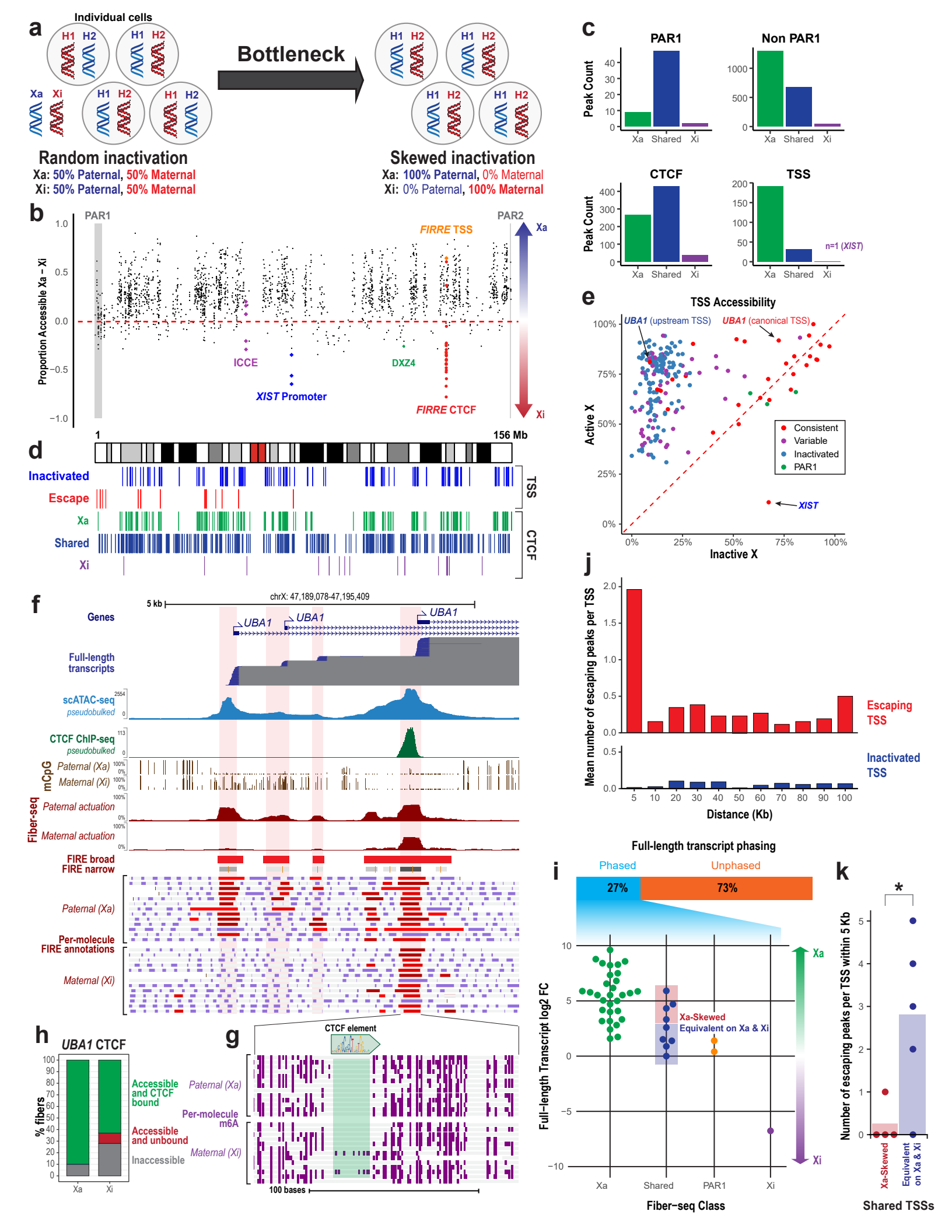

a
Donor 2
Donor 2-specific diploid genome assembly
Haplotype 1 genome
Haplotype 2 genome
Lymphoblast cell Fiber-seq
Haplotype 1 chromatin
Haplotype 2 chromatin
Melanoma cell Fiber-seq
Haplotype 1 chromatin
Haplotype 2 chromatin
b
Replicate condordance of haplotype-selective FIRE scores
Hap1-selective
Hap2-selective
Imprinted elements (R=0.97*)
Elements w/ 1+ var. (R = 0.91*)
Elements w/ no var. (R = 0.91*)
Difference in FIRE scores between lymphoblast haplotypes (replicate 2)
Difference in FIRE scores between lymphoblast haplotypes (replicate 1)
c
Het. variants within Fiber-seq peak
0
1
2
3+
710 haplotype-selective elements in lymphoblasts
37%
41%
12%
10%
202 of which are also accessible in melanoma
(irrespective of hap.-selectivity)
41%
38%
9%
12%
d
Haplotype-selective elements with:
1+ variant
No variants
Lymphoblast
Melanoma
Hap1
Hap2
Chromatin actuation
Genetic var.
Sequence-deterministic elements
Sequence-nondeterministic elements
e
Consistency of haplotype-selective FIRE scores between tissues from Donor 2
Hap1-selective
Hap2-selective
Imprinted elements (R=0.87, p<1e-4)
Elements w/ 1+ var. (R = 0.76, p<1e-4)
Elements w/ no var. (R = 0.15, ns)
Difference in FIRE scores between melanoma haplotypes
Difference in FIRE scores between lymphoblast haplotypes
f
Features guiding haploid chromatin architectures in each cell:
Hap1
Hap2
Same cellular trans environment
Same cellular developmental history
Distinct genetic architecture
Elements with haplotype-invariant chromatin accessibility
Elements with haplotype-selective chromatin
Sequence-deterministic elements
Sequence-nondeterministic elements

Figure 5


a
HLA-DQA1
HLA-DQB1
chr6: 32,620,000
2 kb
scATAC-seq GM12878
DNase-seq GM12878
Paternal chromatin act. GM12878 Fiber-seq
Maternal chromatin act. GM12878 Fiber-seq
Hap.-selective peaks:
b
Fiber-seq on primary CD8+ T-cells and CD4+ T-cells from 9 individuals
Avg. fiber size (bp)
Avg. nucleosome size (bp)
Avg. internucleosomal linker size (bp)
c
DNase-seq CD8+ T-cells
Chromatin actuation CD8+ T-cell Fiber-seq
Donor 3
Donor 4
Donor 5
Hap1
Hap2
Hap.-selective peaks:
d
DNase-seq CD8+ T-cells
chr6: 32,628,250
NFYB ChIP-seq
50 bp
Chromatin act.
Genetic variants
Per-molecule m6A-marks Fiber-seq
TF binding elements: NFYB NFYB E-box
e
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 18 19 20 21 22
hap.-selective Fiber-seq peaks in CD8+ T-cells
p-value
f
Cell-selective enrichment of HLA locus for haplotype-selective elements
HLA class I
HLA class III
HLA class II
CD8+ T-cells
Lymphoblastoid
Fibroblasts
Thyroid tissue
Adjusted p-value
g
Lymphoblastoid cells Fiber-seq
Donor 1
Donor 2
D4-Hap1
D4-Hap2
D5-Hap1
D5-Hap2
Intra-sample comparison of D4-Hap2 chromatin at locus
Cosine similarity of actuation of 100 consecutive Fiber-seq peaks
Inter-sample comparison of D4-Hap2 chromatin at locus
40 genomic loci with greater inter-sample similarity
Overlap with loci containing GRCh38 alternative haplotypes
23%*
Overlap with loci containing segmental duplications
53%*
null distribution
# of randomizations
Percent of loci with overlap


a
Individual cells
H1 H2
Xa Xi
Bottleneck
Random inactivation
Xa: 50% Paternal, 50% Maternal
Xi: 50% Paternal, 50% Maternal
Skewed inactivation
Xa: 100% Paternal, 0% Maternal
Xi: 0% Paternal, 100% Maternal
b
PAR1
PAR2
FIRRE TSS
Xa
Xi
Proportion Accessible Xa − Xi
ICCE
DXZ4
XIST Promoter
FIRRE CTCF
c
PAR1
Non PAR1
CTCF
TSS
Peak Count
Xa Shared Xi
n=1 (XIST)
d
1
156 Mb
Inactivated
Escape
Xa
Shared
Xi
TSS
CTCF
e
TSS Accessibility
UBA1 (upstream TSS)
UBA1 (canonical TSS)
Active X
Inactive X
Consistent
Variable
Inactivated
PAR1
XIST
f
chrX: 47,189,078-47,195,409
5 kb
Genes
UBA1
Full-length transcripts
scATAC-seq
pseudobulked
CTCF ChIP-seq
mCpG
Paternal (Xa)
Maternal (Xi)
Fiber-seq
Paternal actuation
Maternal actuation
FIRE broad
FIRE narrow
Per-molecule FIRE annotations
j
Mean number of escaping peaks per TSS
Escaping TSS
Inactivated TSS
Distance (Kb)
Full-length transcript phasing
i
Phased
Unphased
27%
73%
Full−length Transcript log2 FC
Xa-Skewed
Equivalent on Xa & Xi
Fiber−seq Class
Shared
PAR1
k
Number of escaping peaks per TSS within 5 Kb
Xa-Skewed
Equivalent on Xa & Xi
Shared TSSs
h
UBA1 CTCF
% fibers
Accessible and CTCF bound
Accessible and unbound
Inaccessible
g
CTCF element
Paternal (Xa)
Per-molecule m6A
Maternal (Xi)
100 bases