A haplotype-resolved view of human gene regulation

Mitchell R. Vollger[1*], Elliott G. Swanson[2*], Shane J. Neph[1], Jane Ranchalis[1], Katherine M. Munson[2], Ching-Huang Ho[3], Y. H. Hank Cheng[2], Adriana E. Sedeño-Cortés[1], William E. Fondrie[4], Stephanie C. Bohaczuk[1], Maxwell A. Dippel[3], Yizi Mao[1], Nancy L. Parmalee[5], Benjamin J. Mallory[2], William T. Harvey[2], Younjun Kwon[2], Gage H. Garcia[2], Kendra Hoekzema[2], Jeffrey G. Meyer[6], Mine Cicek[6], Evan E. Eichler[2,7], William S. Noble[2,8], Daniela M. Witten[9], James T. Bennett[10], John P. Ray[2,3,11], Andrew B. Stergachis[1,2,12,†]

[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.
[*] These authors contributed equally to this work.

Abstract

Diploid human cells contain two non-identical genomes, and differences in their regulation underlie human development and disease. We present Fiber-seq Inferred Regulatory Elements (FIRE) and show that FIRE provides a more comprehensive and quantitative snapshot of the accessible chromatin landscape across the 6 Gbp diploid human genome, overcoming previously known and unknown biases afflicting our existing regulatory element catalog. FIRE provides a comprehensive genome-wide map of haplotype-selective chromatin accessibility (HSCA), exposing novel imprinted elements that lack underlying parent-of-origin CpG methylation differences, common and rare genetic variants that disrupt gene regulatory patterns, gene regulatory modules that enable genes to escape X chromosome inactivation, and autosomal mitotically stable somatic epimutations. We find that the human leukocyte antigen (HLA) locus harbors the most HSCA in immune cells, and we resolve the specific transcription factor (TF) binding events disrupted by disease-associated variants within the HLA locus. Finally, we demonstrate that the regulatory landscape of a cell is littered with autosomal somatic

epimutations that are propagated by clonal expansions to create mitotically stable and non-genetically deterministic chromatin alterations.

Introduction

Advances in long-read sequencing (*1*, *2*) have enabled the routine *de novo* assembly of 6 Gbp diploid human genomes at reference quality (*3*, *4*), resulting in the completion of the first human genome (*5–7*) and pangenome (*8–10*). 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, as this can shed light on how non-coding genetic variants contribute to human disease risk, uncover somatic epimutations that are specific to a given haplotype, and illuminate gene regulatory patterns in genetically complex regions of the genome that show the greatest sequence diversity between humans.

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 phasing 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 (*11–14*) have overcome these challenges by leveraging the ability of long reads to correctly map to the diploid genome (*12*, *15*, *16*). 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 (**Fig. 1a**) (*11*). These m6A-modified DNA molecules are sequenced using a single-molecule platform (*17*), enabling the synchronous readout of the genetic sequence, CpG methylation status, and chromatin architecture of each multi-kilobase sequencing read. Prior work has used this data to accurately identify the presence of transcription factors (*18*); however, despite the promise of single-molecule chromatin assays, current approaches for analyzing this new data type are sensitive to potential experimental batch effects, and are unable to identify putative regulatory elements *de novo*, resulting in a dearth of new knowledge about how gene regulation is structured across the *entire* 6 Gbp genome, and how this structure changes across different tissues and disease states. To realize the full potential of long-read single-molecule chromatin assays, we developed a robust machine learning classifier that enables the precise delineation of chromatin *architectures* across the entire 6 Gbp diploid genome with single-molecule and single-haplotype precision at near nucleotide resolution. Furthermore, we leverage this approach across multiple cohorts to uncover basic principles guiding gene regulation, imprinting, somatic epimutations, and X chromosome inactivation.

Results

Single-molecule Fiber-seq inferred regulatory elements

We sought to develop a machine learning classifier that could utilize Fiber-seq features at stretches of m6A-modified bases (*i.e.,* MTase Sensitive Patches [MSPs]) that differentiate actuated single-molecule regulatory elements from nucleosome footprints and internucleosomal linkers (**Fig. 1a**). Although MSP size is frequently used as a simple classifier for this process (*11*, *12*), this feature is highly sensitive to technical sample-to-sample variation in the total rate of adenine methylation (**Fig. S1a,b**), which can result in systematic biases in identifying Fiber-seq accessible chromatin (*19*) (**Fig. S1c**). Furthermore, the majority of MSPs >150 bp in size are located outside of known DNaseI hypersensitive sites (DHSs) (**Fig. S1d,e**), likely reflecting unstable nucleosome occupancy events, resulting in this simple classifier having relatively poor performance for the *de novo* identification of putative regulatory elements genome-wide (**Fig. S1d**). Finally, by solving this problem *de novo* for each individual chromatin fiber (sequencing read), we seek to measure chromatin accessibility as the percent of molecules with chromatin actuation at any given regulatory element in the genome. To resolve this fundamental challenge limiting the broader adoption of single-molecule chromatin stenciling methods, we develop a tool for the accurate *de novo* classification of putative regulatory elements using only Fiber-seq data that does not need any prior knowledge of the underlying reference genome, or the exact primary sequence of the underlying fiber. Importantly, this structure enables this tool to generalize across different cell types that employ unique sets of binding elements to regulate their accessible chromatin (*15*, *16*, *20*, *21*), complex genomic regions not present within the reference GRCh38 (*22*, *23*), as well as organisms that have completely distinct genomic architectures (*24*).

To create our training data, we generated multiple Fiber-seq datasets from the reference cell line GM12878 (totaling 136-fold coverage) that differed by over 2.3-fold in their global m6A methylation rate, enabling the final model to extend to analytical fluctuations in the Fiber-seq method (**Fig. S1a-c**). Although our final tool works irrespective of a 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 (**Fig. S2a**). Notably, regulatory elements identified using bulk methods can exhibit marked heterogeneity in their single-molecule actuation pattern (*i.e.,* can be an actuated element, or merely an internucleosomal linker region at the single-molecule level) (*11*, *12*, *14*) (**Fig. S3**). As such, MSPs ascribed a positive label based on their genomic location relative to bulk chromatin measurements likely 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 best approached using a semi-supervised machine learning framework (*25–27*). To implement a semi-supervised approach, we trained an XGBoost model with five-fold cross-validation (*28*), 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,* **Fig. S2a-c**) 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 markedly enriched in DHSs compared to using MSP length alone as a classifier (**Fig. S1d**).

Aggregating across Fiber-seq inferred regulatory elements.

As no other methods exist for classifying per-molecule chromatin architectures, to appropriately benchmark our FIRE classifications, we next needed to develop a method for performing the *de novo* identification of putative regulatory elements genome-wide using these FIREs (**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 using a null aggregate FIRE score distribution, enabling us to calculate a false discovery rate (FDR, *Methods*, **Fig. S4a,b**) and identify genomic positions that met a 5% FDR threshold. *De novo* Fiber-seq accessible chromatin peaks were directly identified based on local maximums that met the 5% FDR threshold, and the bounds of each peak were defined based on 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." In total, we identified 204,965 FIRE peaks in GM12878, which overlapped 89% of the previously annotated GM12878 DNaseI hypersensitive sites (**Supplemental Table 1**). Our approach provides an intuitive, biologically meaningful metric of chromatin accessibility and peak width for the first time, with the boundaries of each FIRE peak corresponding to the nucleotide-precise median start/stop positions of the underlying single-molecule data, and the intensity reflecting the exact fraction of chromatin fibers with accessibility.

Benchmarking FIRE single-molecule measures of chromatin actuation

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**) even in the setting of experimental differences in the overall methylation rate between replicates (3.81% vs. 4.49% in replicate 1 versus replicate 2, respectively). Furthermore, we showed that FIRE scores were stable after down-sampling our GM12878 data from 135- to 30-fold coverage (**Fig. S4c,d**). In addition, we generated Fiber-seq data from a different cell type (erythroleukemia cell line K562) and demonstrated that overall FIRE measures of chromatin actuation are strongly correlated with those obtained using DNase-seq (**Fig. 1d, S5a-b**). We also benchmarked our FIRE peaks and chromatin actuation measures against single-cell ATAC-seq (scATAC-seq) data, as this data

type was 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. Overall, we found that peaks of accessibility identified using Fiber-seq and scATAC-seq significantly overlapped, with 77% of all FIRE peaks also called by scATAC-seq (**Fig. 1e**). Furthermore, we found scATAC-seq signal support for even the least actuated Fiber-seq peaks (10-15%) (**Fig. 1f**) and that scATAC signal increases at peaks with higher Fiber-seq FIRE percent actuation (**Fig. 1f,g**), indicating that overall FIRE percent actuation scores are in agreement with short-read bulk measures of chromatin accessibility.

In total, 22.2% of the FIRE peaks were unique to FIRE and not called using either the scATAC-seq or DNase-seq data (*i.e.,* FIRE-specific peaks). These FIRE-specific peaks were on average narrower in terms of their chromatin accessibility (**Fig. 2a,b**), did not show major changes in GC content (**Fig. S5c**), and were significantly enriched in REST ChIP-seq peaks that demonstrated single-molecule TF occupancy patterns at these REST sites (**Fig. S5d-e**). To validate the biological relevance of these FIRE-specific peaks, we overlapped all FIRE peaks with genetic variants in GM12878 cells that show genome-wide significant associations (GWAS) with different human traits and diseases (*29*). We observed that FIRE peaks were significantly enriched for overlapping disease-associated GWAS variants, as expected (p-value = 8.40e-71, two-sided Fisher's exact test) (*30*). However, unlike scATAC-seq-specific or DNase-specific peaks, FIRE-specific peaks also showed a comparable enrichment for disease-associated GWAS variants (**Fig. 2c,** p-value = 1.32e-17, two-sided Fisher's exact test), showing that these Fiber-seq unique elements are of comparably high quality as Fiber-seq peaks that have ATAC-seq and DNase-seq support. Overall, these findings demonstrate that FIRE enables the accurate and robust *de novo* identification of accessible chromatin elements solely using long-read epigenomic data.

FIRE provides more accurate measures of chromatin accessibility

We next investigated differences between FIRE measures of chromatin accessibility and those of existing short-read chromatin assays. First, we observed that controlling for GC content improved the correlation between short-read chromatin assays and Fiber-seq, likely reflecting GC biases that are induced during the PCR amplification steps of these short-read assays (**Fig. S4e**). Second, we observed that 78.4% of the GM12878 peaks identified only in Fiber-seq were <200 bp in length, raising the possibility that short-read chromatin assays may be selectively biased against detecting accessibility at shorter regulatory elements (**Fig. 2d, S5c**). Consistent with this, we observed that although short-read measures of chromatin accessibility are strongly associated with Fiber-seq percent actuation for elements >250 bp in length, this association markedly deteriorates for elements <200 bp in length (**Fig. 2d,e**), which are enriched in CTCF ChIP-seq peaks (26.5% of all peaks ≤200 bp). This potential size-dependent bias also appeared to extend to scATAC-seq measures of chromatin accessibility. 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**), with scATAC-seq signal being 3.47-fold lower at CTCF ChIP-seq peaks compared to other scATAC-peaks with a similar Fiber-seq percent actuation (**Fig. 2f**). These findings suggest that existing short-read maps of chromatin accessibility are reporting inaccurate chromatin accessibility measurements at CTCF binding elements, and consistent

with this, we observed that FIRE actuation at a CTCF ChIP-seq peak was markedly better at predicting CTCF ChIP-seq signal at that site than ATAC-seq signal (R^2 of 0.41 vs. 0.17, respectively) (**Fig. S6a**). Overall, this size-dependent bias resulted in the width of a peak being more predictive of ATAC-seq Tn5 insertions at a site than the actual number of chromatin fibers on which that peak is actuated (**Fig. 2f, S5b**). 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 biases.

FIRE provides a more complete genome-wide map of chromatin accessibility

We 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. 2g,h**). SDs comprise ~200 Mbp of genomic sequence (*31*, *32*) that are known to contribute to a variety of human diseases (*33*, *34*), but these regions have been challenging to study owing to mapping issues of short-read chromatin assays to these highly similar duplicated sequences (*35*). 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. 2h**). 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 haplotype-phasing our Fiber-seq reads could enable us to map chromatin across the 6 Gbp diploid human genome. For GM12878, each ~20kb read on average spans at least one heterozygous variant, and parental short-read sequencing data can bin these reads based on their maternal or paternal origin, enabling the accurate haplotype phasing of 87.9% of reads across GRCh38 (*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 genome-wide (**Fig. 3a**). Notably, we observed that elements showed more variability in actuation between haplotypes than between Fiber-seq replicates (**Fig. S6b**), highlighting the reproducibility of FIRE results between replicates and the central role of individual haplotypes in guiding chromatin accessibility patterns. We identified elements with haplotype-selective chromatin accessibility (HSCA) by comparing the percent actuation between the two haplotypes using a Fisher's exact test with a genome-wide FDR correction (Benjamini-Hochberg). This resulted in the identification of 9,773 elements with nominally significant (p-value <0.05) HSCAs in GM12878 cells, with 1,231 of these elements 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.

To further validate this, we performed Fiber-seq on 13 fibroblast cell lines for which parental genomic sequencing data was available to phase chromatin patterns to the maternal and paternal haplotype (**Fig. 3e**). Using this, we identified 30 FIRE peaks that showed consistent parent-of-origin haplotype-selective chromatin across the 13 fibroblast cell lines (**Fig. 3f**), 5 of which were previously unannotated as being within imprinted elements as defined using CpG methylation patterns. Notably, all 5 of these elements are within genomic regions known to be imprinted but do not directly overlap a differentially methylated CpG region (**Fig. 3g,h**), indicating that parent-of-origin chromatin accessibility at these sites is independent of the local CpG methylation. Overall, this demonstrates that only a fraction of the haplotype-selective peaks in a cell are caused by previously uncharacterized imprinted elements, and that elements with consistent parent-of-origin chromatin accessibility can be maintained even in the absence of underlying parent-of-origin CpG methylation.

Haplotype-selective chromatin marks disease-associated loci and elements

GM12878 cells contain 3.2 million autosomal heterozygous variants that distinguish each haplotype, and we next sought to evaluate which of these genetic variants contribute to the formation of haplotype-selective chromatin in GM12878 cells. Overall, HSCA elements are significantly enriched in directly overlapping heterozygous genetic variants, with 56% of all autosomal elements with HSCA containing at least one heterozygous genetic variant (**Fig. 3c**), and some containing over 10 heterozygous genetic variants. To determine whether the encompassed heterozygous variants may in fact be causal of the haplotype-selective chromatin signal, we quantified the overlap of these elements with variants known to have genome-wide significant associations with different human traits and diseases (*i.e.,* lead GWAS variants). Overall, elements with HSCA showed a 6-fold enrichment (36 versus 6) in overlapping lead GWAS variants compared to random non-HSCA FIRE peaks (**Fig. 4a**), indicating that haplotype-selective chromatin at these elements may, in fact, be directly resulting from the effect of these underlying genetic variants.

However, most lead GWAS variants are not thought to be disease-causal but rather are thought to be tagging a haplotype that contains neighboring variants that are mediating the disease or trait association, including neighboring rare variants with a large effect size (*36*). To evaluate whether these lead GWAS variants are tagging a putatively functional variant along the same haplotype, we quantified the incidence of haplotype-selective chromatin peaks within 500 kbp surrounding lead GWAS variants in GM12878. This demonstrated that lead GWAS variants were significantly enriched for being located within 40 kbp of a peak with haplotype-selective chromatin (**Fig. 4b**). Notably, unlike lead GWAS variants that directly overlap elements with HSCA, these adjacent haplotype-selective peaks preferentially contained rare genetic variants with a minor allele frequency (MAF) of less than 5%, indicating that Fiber-seq is identifying rare variants along these haplotypes that are potentially mediating the disease/trait associations (**Fig. 4c**). The disease associations of these GWAS lead variants preferentially localized

adjacent to GM12878 haplotype-selective chromatin elements were significantly enriched for diseases consistent with the biological function of lymphoblast cells (**Fig. 4d**).

The per-molecule and near-nucleotide resolution of Fiber-seq permitted the dissection of the specific TF binding elements and variants that are likely mediating these disease associations. 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 (rs4713235) directly disrupting single-molecule Fiber-seq TF occupancy at a CTCF binding element (**Fig. S7**). Furthermore, the 75 kb region surrounding the *HLA-DQA1* gene contains 3,199 heterozygous genetic variants that distinguish the two haplotypes in GM12878 (**Fig. S8a**). However, TF footprinting using Fiber-seq in GM12878 cells and primary T-cells enabled the identification of specific variants that are likely mediating these haplotype-selective chromatin features (**Fig. 4e**). For example, rs9271894, which is associated with Celiac disease risk in biobank PheWAS studies (*37*) (OR 2.84, p-value 1e-250), appears to disrupt protein occupancy at an E-box and an adjacent CCATT box selectively along T-cell haplotypes that harbor this variant (**Fig. 4f,g**). Notably, this element encodes a predicted low-affinity CCAAT box, and the single-molecule TF occupancy patterns indicate that protein occupancy at this CCAAT box is cooperatively dependent upon occupancy at the adjacent E-box that harbors rs9271894 (**Fig. 4g**). Consequently, although the 1 kbp region surrounding rs9271894 contains 54 heterozygous variants that distinguish the two haplotypes in one of the donors, Fiber-seq was able to identify that the putatively functional variant is likely rs9271894, illustrating how Fiber-seq can disentangle the relative contribution of non-coding variants to human disease/trait associations.

The HLA locus contains the highest rate of chromatin epigenomic diversity

Next, we sought to evaluate whether certain genomic loci are preferentially marked by haplotype-selective chromatin, irrespective of their disease association. To accomplish this, we quantified the enrichment of haplotype-selective peaks within rolling windows of 100 peaks (*Methods*) using Fiber-seq data from lymphoblastoid cells, activated and sorted $CD8^+$ T-cells, fibroblasts, and thyroid tissue from multiple unrelated donors. After removing imprinting regions, this demonstrated that the MHC region on chromosome 6 contained the most haplotype-selective chromatin in the entire human genome in $CD8^+$ T-cells (corrected p-value 0.0013, Benjamini-Hochberg corrected FDR < 5%) (**Fig. 4i**). Notably, whereas this enrichment for haplotype-selective chromatin was selectively localized to the MHC region HLA class II and HLA class I loci in lymphoblastoid cells, $CD8^+$ T-cells only showed this localization within the HLA Class II locus (**Fig. 4j**), reflecting differences in the biological roles of these loci in these distinct immune cell types. In contrast, although fibroblasts and thyroid tissue show numerous actuated elements within the MHC region (**Fig. S8b**), these samples did not show any enrichment for haplotype-selective chromatin within the HLA locus (**Fig. 4j, S8c**), indicating that although many MHC proteins are ubiquitously expressed across various cell types, the marked genetic diversity within the MHC region is only associated with haplotype-selective chromatin patterns in select cell types.

Given the marked divergence in chromatin accessibility between both haplotypes within the MHC region, we next sought to evaluate whether the haploid chromatin accessibility within this region showed more variability between both haplotypes within the same individual or between individuals. To accomplish this, we compared haploid Fiber-seq data from the COLO829BL and GM12878 lymphoblastoid cell lines, which were derived from two separate donors of different ages and sex (**Fig. 4k**). While the far majority of the haploid genome showed greater variability in chromatin accessibility between both donors, as would be expected, we did identify 40 extended genomic loci, including the MHC region, that showed greater chromatin accessible variability between both haplotypes within the same donor. Notably, these 40 genomic loci were significantly enriched for containing 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). This demonstrates that genomic regions containing some of the most genetically diverse human haplotypes (*38–40*) also show the most haplotype-selective chromatin patterns, consistent with the hypothesis that selective pressures and/or consequences of these diverse haplotypes are also at the level of altered gene regulatory patterns (*41*).

Somatic autosomal epimutations mark the chromatin landscape of cell lines and tissue

We next sought to understand the mechanistic basis underlying haplotype-selective chromatin at the 44% of GM12878 elements with HSCA that did not contain a genetic variant directly within the peak (*i.e.,* 'HSCA-noVar' elements) (**Fig. 3c**). Notably, unlike GM12878 elements with HSCA that overlap a heterozygous genetic variant (*i.e.,* 'HSCA-withVar' elements), HSCA-noVar elements did not show any enrichment for being adjacent to lead GWAS variants (**Fig. 4b, S6c**), suggesting that haplotype-selective chromatin at HSCA-noVar elements may not be genetically determined. Furthermore, HSCA-noVar elements overwhelmingly are not previously undiscovered imprinting sites (**Fig. 3f**), raising the possibility that haplotype-selectivity at these sites may arise from somatic epimutations causing random mono-allelic chromatin accessibility. To test this hypothesis, we determined the stability of haplotype-selectivity at these elements between multiple tissues derived from the same individual, under the assumption that somatic epimutations would preferentially show divergent chromatin patterns between different tissues that arose from the same zygote. Specifically, we performed Fiber-seq on a lymphoblastoid cell line (COLO829BL) and melanoma cell line (COLO829T) derived from the same donor (**Fig. 5a**). To ensure consistent haplotype phasing between these two samples, we generated a near telomere-to-telomere assembly of COLO829BL using Fiber-seq, ultra-long Oxford Nanopore, and Hi-C (contig N50 140.01 Mbp with 18 chromosomes having telomere-to-telomere scaffolds), enabling us to unambiguously phase 84.2% of all Fiber-seq reads from COLO829BL and 83.9% of Fiber-seq reads from COLO829T. 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, including at HSCA-noVar elements (Pearson's correlation > 0.9, p-value < 2.2e-16, two-sided *t*-test) (**Fig. 5b**). 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. 5c**). We find that for imprinted sites (R=0.87) or COLO829BL HSCA-withVar elements (R=0.76), there is a strong correlation in the HSCA between both cell types (*e.g.,* elements selective to haplotype 1 in COLO829BL are also selective to haplotype 1 in COLO829T) (**Fig.**

5d,e), consistent with imprinted and HSCA-withVar elements largely having deterministic chromatin actuation dependent upon the underlying haplotype. In contrast, COLO829BL HSCA-noVar elements had markedly divergent haplotype-selective actuation between the two cell types (*R*=0.15) (**Fig. 5d,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, findings consistent with haplotype selective-chromatin arising from somatic epimutations.

To validate these findings in a different individual, we performed Fiber-seq on primary lung and liver tissue that were obtained from a single individual. Furthermore, we generated a diploid genome assembly from this individual to consistently phase the chromatin data between the two tissues. Overall, we found that the lung tissue had 88 peaks with genome-wide significant HSCA (**Fig. 5f**), with 98.9% (87/88) of these HSCA elements in lung also demonstrating some degree of chromatin actuation in the liver tissue. Consistent with the COLO829 data, imprinted elements (*R*=0.88) and HSCA-withVar elements (*R*=0.87) showed a strong correlation in the haplotype-selective chromatin actuation between both tissues. In contrast, lung HSCA-noVar elements showed markedly divergent haplotype-selective chromatin actuation between the two tissues (*R*=-0.88) (**Fig. 5g**), consistent with haplotype-selective chromatin at these HSCA-noVar elements arising from somatic epimutations (**Fig. 5h**).

These findings raise the prospect that some autosomal accessible chromatin elements somatically acquire non-deterministic chromatin actuation patterns that are mitotically stable (**Fig. 5i**), and as humans are diploid, this results in each haplotype having a divergent chromatin actuation pattern. One prediction of this model would be that bottlenecked samples would have higher rates of elements with HSCA, as a greater proportion of the cells within the Fiber-seq reaction would have arisen from the same founder cell with somatic epimutation. Consistent with this prediction, we observed that fibroblast lines with greater evidence of clonal expansion demonstrated more HSCA-noVar elements (*R*=0.95) (**Fig. 5j,k**). Together, these findings indicate that the autosomal accessible chromatin landscape of a cell is molded by somatic epimutations that create mitotically stable and non-genetically deterministic chromatin alterations.

Somatically inactivated elements and domains along the X chromosome

The X chromosome in 46,XX cells is subjected to mitotically stable random X chromosome inactivation (XCI), creating an inactive X (Xi) and active X (Xa) chromosome within the same cell that is largely non-genetically deterministic (*42*, *43*). XCI represents an extreme case of mitotically stable somatic epimutations, and we sought to utilize Fiber-seq to explore the impact of XCI across every regulatory element along the X chromosome. To accomplish this, we applied Fiber-seq to two 46,XX cell lines [GM12878 (previously introduced) and MAN_1877] that have allelic XCI, enabling us to accurately map the chromatin architecture along the Xi and Xa based on parental haplotypes (**Fig. 6a**). Full-length long-read transcript sequencing of GM12878 cells showed that 99% of the long non-coding RNA *XIST* transcripts (*44–47*) originate from the maternal haplotype (**Fig. S9**), confirming that the Xi is invariantly the maternal haplotype in GM12878 cells. Using the Fiber-seq data, we categorized FIRE elements along the

X chromosome as to whether they were preferentially somatically silenced on the Xi (*i.e.,* 'Xa-selective' accessibility) or Xa (*i.e.,* 'Xi-selective' accessibility), or showed similar chromatin accessibility along both the Xa and Xi (*i.e.,* 'shared').

We observed that chromatin accessibility outside of the pseudoautosomal region 1 (PAR1) was predominantly impacted by XCI, with 63% of elements being Xa-specific, 34% being shared, and only 2.6% being Xi-specific elements (**Fig. 6b-d**). Aside from the *XIST* promoter, the majority (73.3%) of 'Xi-selective' elements were within *ICCE* (Inactive-X CTCF-binding Contact Element) (*48*), DXZ4 (*49*, *50*), and *FIRRE* (Functional Intergenic Repeating RNA Element) loci (*51*) – and 78.9% of these were CTCF elements (**Fig. S10**), which play a role in organizing three-dimensional mega-domains on the Xi (*52*). 66% of 'shared' elements were also CTCF elements, with the majority (59%) of X chromosome CTCF sites having similar accessibility along both the Xa and Xi. Notably, boundaries separating chromatin domains subjected to XCI and escaping XCI often contained precisely positioned CTCF elements with protein occupancy on both the Xi and Xa. For example, *UBA1* has four distinct accessible TSSs in GM12878 cells, with the three upstream TSSs being Xa-specific, and the canonical downstream TSS escaping XCI (**Fig. 6e**). The canonical XCI escaping *UBA1* TSS is bookended by an upstream CTCF element that shows single-molecule TF occupancy on both the Xa and Xi (90% and 59% occupancy, respectively) (**Fig. 6e**). Overall, these findings demonstrate that CTCF occupancy along the human X chromosome preferentially escapes XCI mediated somatic silencing and that CTCF occupancy along chromosome X may serve as boundary elements insulating chromatin domains from XCI, features that have been previously proposed in mice (*50*).

Escaping promoter-proximal elements enable productive transcription from the Xi

16% of TSSs for transcribed genes in GM12878 cells were also marked by 'shared' elements (**Fig. 6c**). Notably, these TSSs largely corresponded to genes known to escape XCI based on prior transcript sequencing data (*53*, *54*) (**Fig. 6d**), as well as paired GM12878 long-read transcript data. However, the GM12878 transcript sequencing data could only phase 27% of these genes owing to the paucity of heterozygous coding genetic variation along the human chromosome X in GM12878 cells. We sought to directly test the relationship between promoter accessibility and transcript production on the Xi using the paired haplotype-phased long-read transcript and Fiber-seq data. As expected, we observed highly skewed expression in genes with Xa-specific or Xi-specific TSS accessibility (**Fig. 6h**). However, genes with similar TSS accessibility between the Xa and Xi showed heterogeneous expression, 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=3).

Given the widespread accessibility that we observed along the Xi, we hypothesized that the escape of promoter-proximal regulatory elements may augment productive transcription from the Xi. 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. 6g**). This revealed enrichment of escaping regulatory elements within 100 Kbp of escaping TSSs, with the largest difference (51- to 71-fold increase) within 5 Kbp of the escaping TSS. Notably, 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 (12 vs 0, p-value = 0.04, one-sided Wilcoxon rank sum test; **Fig. 6i**). 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 how Fiber-seq can enable 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) and show that FIRE provides a more comprehensive and quantitative snapshot of a cell's accessible chromatin landscape, overcoming previously known and unknown biases inherent to our existing catalog of regulatory elements. Importantly, FIRE enables the accurate *de novo* construction of a cell's haplotype-resolved chromatin-accessible landscape directly using long-read sequencing data, enabling the advent of synchronous multi-omic long-read sequencing that can fine-map the chromatin mechanisms by which disease-associated haplotypes cause human disease (*15*, *16*), as well as resolve chromatin architectures within complex genomic loci like segmental duplications (*55*), centromeres (*56*, *57*), and transposons (*24*).

We show that Fiber-seq can localize the specific regulatory elements and TF binding events that underlie GWAS disease associations, including those within the HLA locus. The HLA locus has been definitively established as the primary genetic locus underlying autoimmune disease risk. However, the HLA locus is one of the most genetically complex loci in the human genome, with genetic variants occurring once every 20 bases along many disease-associated haplotypes – a rate of genetic diversity that stifles short-read-based approaches. Although it is largely assumed that only protein-coding variants mediate HLA locus autoimmune disease risk, emerging data has indicated that non-coding variants play a critical role in autoimmunity through altering HLA gene expression (*58–60*), which in turn can dramatically impact how immune cells respond to antigens. We show that Fiber-seq can pinpoint specific genetic variants within the HLA locus that drive disease associations, exposing that the Celiac disease-associated SNP rs9271894 disrupts cooperative TF occupancy at an E-box and adjacent CCAAT box immediately upstream of *HLA-DQA1* (**Fig. 4h**), which is one of the primary immune genes associated with Celiac disease (*61*). Although the exact impact of this variant on immune cell function is unknown, Fiber-seq nonetheless enabled us to localize which of the thousands of variants within the HLA region actually show a molecular phenotype at the chromatin level. Given these findings, we anticipate that the broader application of Fiber-seq to immune cells will dramatically alter how we study and understand the contribution of non-coding variation within the HLA locus to human disease risk.

It is well known that the establishment of X inactivation differs dramatically between mice and humans, both in terms of the order of events and the epigenetic silencing mechanisms (*43*). Although short-read epigenetic maps of the Xi and Xa can be readily established using hybrid

mice, this is not readily accomplishable in humans owing to the lack of genetic diversity along chromosome X. We demonstrate that Fiber-seq enables the precise delineation of chromatin elements along the Xa and Xi in humans, revealing widespread chromatin accessibility across the Xi and exposing that promoter accessibility along the Xi is necessary but often not sufficient for robust gene expression from Xi. Epigenetic editing approaches are actively being investigated for the treatment of X-linked diseases (*62*), and we anticipate that Fiber-seq will provide a roadmap for these translational efforts.

Outside of the X chromosome, we find that ~99% of autosomal accessible elements show nearly identical levels of chromatin accessibility between the two haplotypes in these bulk measurements. This demonstrates that, on average, 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 steady-state 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 (*63*). As described above, Fiber-seq exposed numerous elements with HSCA that appear to be primarily driven by underlying heterozygous genetic variants. However, ~40% of elements with HSCA lacked an underlying genetic variant. We show that these elements largely appear to have non-deterministic behavior (*i.e.,* their chromatin accessibility deviates between both haplotypes in a manner that cannot be explained by the genetic architecture of their underlying haplotype). Importantly, we prove that these elements are not imprinted sites and show that elements with non-deterministic behavior can be observed in primary human tissues. Together, these findings demonstrate that autosomal accessible chromatin elements can somatically acquire non-deterministic chromatin actuation patterns that are mitotically stable and divergent between haplotypes, a pattern consistent with these being sites of somatic epimutations (*64*). Notably, we find that clonal expansions increase the rate of somatic epimutations, consistent with these somatic epimutations arising from founder events. Together, these findings indicate that the autosomal accessible chromatin landscape of a cell is molded by somatic epimutations that create mitotically stable and non-genetically deterministic chromatin alterations.

Finally, we find that genomic loci marked by the most genetic diversity within the human population (*38–40*, *65*) 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 (*53*, *54*), this 6 Gbp view of human gene regulation will likely markedly improve our understanding of the regulatory architectures that make us humans.

Figure legends

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)** Correlation of FIRE score with bulk DNase-seq in K562 accessible peaks (two-sided *t*-test). **e)** Venn diagram showing the overlap of FIRE and scATAC peaks in GM12878. **f)** (Left) Average count of DNase I reads over FIRE peaks binned by their percent actuation (red-blue color scale). (Right) Percentile normalized scATAC and DNase I signal for 100 random FIRE peaks across each percent actuation bin. **g)** 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, 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)** Per-base enrichment of GWAS variants within shared peaks between FIRE and DNase/scATAC, FIRE only peaks, and peaks unique to DNase or scATAC as compared to shuffled random windows of the same size (p-value = 1.32e-17, two-sided Fisher's exact test). **d)** 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). **e)** Prediction of percent FIRE actuation from DNase peak as signal using a linear model for different bins of FIRE peak size. **f)** Comparison of scATAC-seq signal to FIRE in peaks with (green) and without (red-blue) CTCF ChIP-seq peaks. **g)** Genomic locus of *NOTCH2NLB* comparing the ability to map into repetitive regions between scATAC-seq, DNase–seq, and Fiber-seq. **h)** 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. a) The *GNAS* imprinted locus comparing the relationship between *GNAS* isoforms, scATAC-seq, DNase–seq, CTCF ChIP-seq, NFYA 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)** Difference between maternal and paternal accessibility for FIRE haplotype-selective peaks stratified by p-value (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. **e)** Schematic of sequencing 13 trios with parental short-reads for phasing and Fiber-seq on the probands to identify parent of origin effects (POE) in chromatin. **f)** Distribution of haplotype-selective chromatin accessibility (HSCA) peaks showing the fraction of fibroblast samples with consistent maternal or paternal bias. Dark red in the histogram indicates previously identified imprinted sites; the pie chart shows the proportion of sites with consistent POE. **g)** Browser views of two genomic regions (*MAGEL2* and *ZDBF2*) demonstrating consistent POE in fibroblasts. **h)** Relationship between parental bias in CpG methylation (y-axis) and FIRE actuation (x-axis). Purple points represent new imprinted sites, with three showing consistent POE in FIRE without evidence of differential CpG methylation.

Figure 4. Haplotype-selective chromatin in the major histocompatibility complex. a) The number of haplotype-selective peaks (red) or random windows (blue) in GM12878 that overlap GWAS variants that are heterozygous in GM12878. **b)** Top, the fraction of lead GWAS variants that can be found within a specific distance (kbp) of: a haplotype selective peak with a genetic variant (red)**,** a haplotype selective peak without genetic variants, and a random set of FIRE peaks of the same size. Bottom, the difference in the fraction of GWAS within a specific distance to a haplotype selective variant with a genetic variant versus a random set of FIRE peaks of the same size. **c)** Enrichment of disease-associated variants within 40 kbp of haplotype-selective FIRE peaks for different disease associations. The x-axis shows the log2 fold enrichment, and the y-axis represents the p-value of a two-sided Fisher's exact test. **e)** GM12878 haplotype-selective sites in the *HLA-DQA1*/*HLA-DQB1* locus. **f)** Haplotype-selective sites in the *HLA-DQA1*/*HLA-DQB1* locus for CD8+ T-cells sequenced to ~30-fold coverage across three individuals. **g)** Haplotype-selective Fiber-seq patterns showing disruption of single-molecule Fiber-seq TF occupancy and chromatin actuation in Donor 5 due to the underlying haplotype. **h)** Histogram of the percent of fibers with TF occupancy at the CCAAT box and E-box across both haplotypes of Donor 5. **i)** 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*). **j)** Cell-selective enrichment of HLA class I, II, and III for haplotype-selective elements (two-sided Fisher's exact test). **k)** 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 ($p < 1e\text{-}04$; permutation test n=10,000; *Methods*) and segmental duplications ($p = 0.0357$; permutation test n=10,000; *Methods*).

Figure 5. 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)** Experimental design showing sequencing of liver and lung primary tissues from donor 3. **g)** Comparison of haplotype-selective peak concordance between liver and lung primary tissues, analyzed as in panel **e**. **h)** Conceptual model distinguishing haplotype-invariant elements from haplotype-selective elements with deterministic and non-deterministic patterns. **i)** Model illustrating how population bottlenecking in cell populations may generate additional haplotype-selective chromatin accessibility. **j)** Experimental approach for sequencing nine female fibroblast cell lines, using X-inactivation allelic skewing as a proxy for cell clonality. **k)** Correlation analysis between the number of haplotype-selective FIRE peaks without genetic variants and the average allelic skew across all X chromosome promoters, excluding the pseudoautosomal regions (Pearson's correlation; two-sided t-test p-value = 7.9e−05).

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 in LCL cells (GM12878). Pseudoautosomal regions (PAR1 & PAR2) are highlighted in orange and gray, respectively. **c)** Counts of FIRE peaks categorized as Xa-specific, Xi-specific, or Shared between both haplotypes for LCL (top) and fibroblast cells (bottom). FIRE elements are stratified by their location within or outside of PAR1 (left), and non-PAR1 elements are further subsetted to those that overlap a CTCF site (middle) or TSS (right). **d)** Scatterplot of LCL 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 (*65*). **e)** *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 in LCLs (GM12878). **f)** Scatterplot of Fiber-seq percent actuation on the Xi in LCL (x-axis) and fibroblast (y-axis) cells. Points are colored as in panel **d**. **g)** The average number of escaping non-TSS FIRE peaks in LCL (left) and fibroblast (right) cells by absolute distance from TSSs. Counts are displayed separately for escaping TSSs (top, purple) and inactivated TSSs (bottom, blue). **h)** Full-length LCL transcript expression differences between the Xa and Xi for genes phased by Fiber-seq and displayed in **d**. 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**. **i)** The number of escaping LCL non-TSS FIRE peaks within 5 Kb of each TSS in the shared category in **h**. Shared TSSs were grouped into high or low log2 fold-change in expression, highlighted with blue and purple in **h** (*p = 0.04031; one-sided Wilcoxon rank sum test).

Supplemental figure legends

Figure S1. Motivation for the FIRE model: heterogeneity in single-molecule data. a) Range of m6A concentrations in Fiber-seq experiments used in training of the FIRE model. **b)** Heterogeneity of m6A concentrations between reads in the Fiber-seq experiments used in the training of the FIRE model. **c)**. Distribution of the length of methyltransferase-sensitive patches (MSPs) across the Fiber-seq experiments used in training the FIRE model. **d)** Number of actuated regulatory elements within DNase I hypersensitive sites (DHSs) on chromosome 20 (y-axis) vs the number of actuated regulatory elements outside of DHSs for different FIRE (red) and MSP length (gray) thresholds (x-axis). **e)** Distribution of MSP length within DNase peaks and outside of DNase peaks (grey), stratified by inferred FIRE elements (red) and nucleosomal linker regions (purple).

Figure S2. Training of the FIRE model. a) Schematic of training Fiber-seq inferred regulatory elements (FIREs) using XGBoost within the Mokapot framework. **b)** Schematic of windows along single reads used for calculating features in the FIRE model. Descriptions of each feature are listed in Table S3. **C)** Ranking the importance of features in the XGBoost model using the feature score (F score), which sums up how many times each feature is split on within the model.

Figure S3. FIRE elements and their underlying m6A calls. UCSC genome browser screenshots of a low accessibility regulatory element **(a)**, a high accessibility element **(b)**, and two sites without regulatory elements **(c,d)**. Shown in each panel in order are: the gene models, ATAC signal, percent of Fiber-seq reads with FIRE elements, wide and narrow FIRE peak calls, raw m6A calls for individual Fiber-seq reads, and FIRE calls for individual Fiber-seq reads in the same order.

Figure S4. The aggregate FIRE score and peak calling with the FIRE method. a) Schematic of calculating a false discovery rate (FDR) for the aggregated FIRE score. top) the aggregated FIRE score across multiple Fiber-seq molecules. middle) aggregate FIRE score for Fiber-seq reads with a shuffled start position within the chromosomes. bottom) FDR track calculated from the null distribution of FIRE scores and peak calls using a 5% FDR threshold. **b)** Number of base pairs genome-wide with an aggregate FIRE score above a threshold (x-axis) for both the observed FIRE elements (red) and the shuffled Fiber-seq reads (gray). Dashed lines are drawn at 1% and 5% FDR thresholds. **c)** The FDR score (y-axis) vs FIRE score threshold (x-axis) for 130-fold Fiber-seq coverage of G12878 and 30-fold coverage. **d)** The correlation of FIRE scores within FIRE peaks between 130-fold Fiber-seq coverage of G12878 (x-axis) and 30-fold coverage (y-axis). **e)** Correlation of FIRE percent actuation and DNase-peak signal within GM12878 for FIRE peaks faceted by FIRE peak size (left to right) and whether GC correction was (bottom) or was not (top) applied (Pearson's correlation; $p < 2.2e{-}16$ two-sided t-test).

Figure S5. Validation and unique features of FIRE elements. a) Correlation of FIRE percent actuation and DNase-peak signal within K562 for FIRE peaks faceted by FIRE peak size (left to right; Pearson's correlation; $p < 2.2e{-}16$ two-sided t-test). **b)** Association between the % of cells with single-cell ATAC-seq signal, versus the size of the regulatory element for only those elements that show a FIRE actuation of >70%. As expected, there is a strong association

between element width and scATAC-seq signal, but this does not approach 100%, demonstrating that even at very large elements, there is a limit of detection in scATAC-seq signal. **c)** Distribution of the GC content for unique FIRE peaks (top) and FIRE peaks shared with ATAC or DNase (bottom) stratified by peak size. Labels indicate the median GC content and the number of elements plotted in the distribution. **d)** Enrichment of transcription (TF) factor motifs within FIRE only peaks versus all FIRE peaks. Enrichments and p-values calculated using the HOMOR software. **e)** The proportion of methylated adenines over total adenines (either strand). All fibers with a FIRE element overlapping FIMO-predicted REST motif (green) are split by those with an overlapping footprint called by ft footprint (orange) and those that are not footprinted (blue). The x-axis represents the position relative to the start of the REST motif.

Figure S6. Genetic and chromatin diversity in GM12878 Fiber-seq across the major histocompatibility complex (MHC). a) Correlation analysis between CTCF ChIP-seq signal and three chromatin accessibility measures: ATAC-seq (left), DNase-seq (middle), and FIRE (right). Values represent Pearson's correlation coefficients with p-values determined by two-sided t-tests. **b)** Histogram of centered absolute difference in percent accessibility between replicates (blue) and haplotypes (red) for COLO829BL (Wilcoxon rank-sum test; W = 6.461e+09; p-value < 2.2e-16). **c)** Left, proportion of lead GWAS variants located within specific distances (kbp) of: haplotype-selective peaks containing genetic variants, haplotype-selective peaks without genetic variants, randomly selected FIRE peaks of equivalent size, and ten independent permutations of random genomic intervals of matching size. Right, identical analysis with all observations from within the MHC locus excluded.

Figure S7. Haplotype-selective Fiber-seq patterns within the MOG1/HLA-F locus. a) Haplotype-selective Fiber-seq patterns within the *MOG1/HLA-F* locus for 4 CD8+ cell lines. Annotated with red lines are lead GWAS SNPs associated with platelet counts. **b)** Fiber-seq results for an individual (donor 3) homozygous for the rs29269 SNP show that the SNP falls between regulatory elements and does not disrupt them. **c)** Fiber-seq results for an individual heterozygous (donor 1) for the rs4713235 SNP show that the SNP directly disrupts single-molecule Fiber-seq transcription factor occupancy at a CTCF binding element. **d)** FIRE haplotype-selective peaks for donors 1 and 3 (two-sided Fisher’s exact test). **e)** Chromatin accessibility for each haplotype across donors 1 (het), 3 (homo ref.), and 4 (homo alt.) at position rs4713235 (two-sided Fisher’s exact test).

Figure S8. Genetic and chromatin diversity in GM12878 Fiber-seq across the major histocompatibility complex (MHC). a) IGV screenshot of the genetic diversity at the *HLA-DRB1* locus for phased GM12878 Fiber-seq reads. **b)** Density of FIRE peaks for different samples across chromosome 6, where the color represents the cell type of the sample. Fibroblast samples are orange, Thyroid samples are blue, CD8+ samples are red, Lymphoblastoid cell lines are green, and the grey box highlights the MHC region. **c)** Ideogram depicting enrichment of haplotype-selective sites across chromosomes in sliding windows of 100 FIRE peaks for multiple cell types. Statistical significance was determined using a two-sided Fisher's exact test with Benjamini-Hochberg correction for multiple testing (*Methods*).

Figure S9. Haplotype-specific chromatin and gene expression of *XIST*. a) *XIST* promoter region comparing GM12878 scATAC-seq, CTCF ChIP-seq, mCpG, FIRE percent accessibility, FIRE peaks, and representative Fiber-seq reads from the paternal (Xa) and maternal (Xi) haplotypes. **b)** Barplot showing *XIST* transcript counts for the GM12878 paternal (Xa) and maternal (Xi) haplotypes.

Figure S10. Haplotype-specific chromatin at Inactive X structural elements. GM12878 browser tracks for the repeat-rich regions, which are known to organize the Xi into distinct three-dimensional mega-domains via Xi-specific CTCF binding sites: **a)** DXZ4, **b)** FIRRE (Functional Intergenic Repeating RNA Element), and **c)** ICCE (Inactive-X CTCF-binding Contact Element) loci. Shown are tracks for scATAC-seq, CTCF ChIP-seq, mCpG, FIRE percent accessibility, and FIRE peaks. Short-read CTCF data is not available for DXZ4 due to mapping limitations within segmental duplications.

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 and 100 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 (*66*). 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).

Nucleosome and MSP (methyltransferase sensitive patch) and calling.

Nucleosome calling is performed by identifying stretches of DNA that are protected from Hia5 (i.e. do not have m6A signal). The rate of false positive m6A calls in nucleosomes is very low when using fibertools (*65*) allowing for a heuristic to perform as well as or better than our

previous HMM caller (*66*). There are three parameters in our heuristic nucleosome calling that can be adjusted: the minimum nucleosome length (n, default 75), the minimum combined nucleosome length (c, default 100), and the minimum extension to nucleosome length (e, default 25). These three parameters impact the four phases in nucleosome calling.

1. Call all regions that have no m6A events for at least n bases a candidate nucleosome.
2. Call all regions of size c or more that have only one internal m6A (putative false positive) a candidate nucleosome.
3. Extend the length of nucleosomes identified in phases one and two if by spanning one additional m6A e bases of unmodified sequence would be added to the nucleosome length.
4. Recursively apply step three till no nucleosomes change.

After nucleosomes are called MSPs are operationally defined as all the regions between nucleosome footprints in the Fiber-seq data. This process is automated through the fibertools subcommand “add-nucleosomes”.

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 (*28*), 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 (*25*). 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**).

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 (*66*).

Aggregate FIRE score calculation

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

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

where C_g is the number of FIRE elements at the g^{th} position, R_g is the number of Fiber-seq reads at the g^{th} position, and EP_i is the estimated precision of the i^{th} FIRE element at the g^{th} 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) (*67*). We then compute FIRE scores associated with the shuffled genome. Recalling that the FIRE score for the g^{th} 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 (*68*). 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) (*69*). 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.

FIRE Peaks Enrichment Transcription Factor Motifs

To determine the transcription factor (TF) binding motifs enriched in FIRE peaks we used Homer v3.12 (http://homer.ucsd.edu/homer/motif/) to discover TF binding sites represented in target sequences and compared them to a chosen background. Using findMotifsGenome.pl, we determined the enrichment of all FIRE peaks and all ATAC peaks to the whole genome as a background, then we compared ATAC-overlapping FIRE peaks and non-ATAC-overlapping FIRE peaks with all FIRE peaks as the background. We determined the REST motif genome coordinates in the non-ATAC-overlapping FIRE peaks for aggregate footprinting analysis using annotatePeaks.pl.

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 and select elements with a correct p-value less than or equal to 0.05 to be haplotype selective peaks.

884 **The FIRE pipeline**

885 The FIRE pipeline (https://github.com/fiberseq/FIRE v0.0.4) was applied to aligned and phased
886 Fiber-seq BAM files with m6A predictions called using Fibertools-rs (v0.4) (*69*). The FIRE
887 pipeline is a Snakemake (*70*) workflow that applies the FIRE model to individual reads,
888 calculates the aggregate FIRE scores, computes the FDR, calls peaks, and identifies haplotype-
889 selective peaks. The only required inputs for the FIRE pipeline are an aligned Fiber-seq BAM
890 file and the reference genome used for alignment.

891 **Phasing and identification of genomic variants**

892 Variant calling for single nucleotide variants (SNVs) and insertions/deletions (Indels) was
893 conducted using DeepVariant version 1.5.0 (*71*), while structural variants (SVs) were identified
894 using pbsv (https://github.com/PacificBiosciences/pbsv). The Fiber-seq reads were haplotype-
895 phased through a specialized pipeline available on GitHub (https://github.com/mrvollger/k-mer-
896 variant-phasing) (*71*). This pipeline employs SNVs detected by DeepVariant and utilizes the
897 HiPhase (*72*) variant-based phaser to organize reads into phase blocks. These blocks are then
898 attributed to either the maternal or paternal haplotype using parental short-read genome
899 sequencing in conjunction with the trio k-mer-based phaser meryl (k=31) when available (*73*).

900 **Identification of CpG methylation**

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

904 **Imprinted loci**

905 Imprinted loci were defined using the 143 differentially methylated regions identified in 12 B-
906 lymphocyte cell lines by Akbari et al. (*74*). Novel imprinted elements in the 13 fibroblast samples
907 were identified by taking the union of all haplotype-selective peaks across all 13 samples and
908 identifying sites where at least 10 of the samples showed significant chromatin actuation.
909 Subsequently, we analyzed these elements for sites where all samples showed the same
910 direction of maternal or paternal skew and classified these as putative novel imprinted elements.

911 **Identifying regions enriched in haplotype-selective elements**

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

916 **Intra- versus Inter-sample sample similarity in FIRE actuation**

917 To measure the similarity between two haplotypes, we took genomic regions containing 100
918 peaks and measured the cosine similarity of the percent actuation between the two haplotypes.
919 We then repeated this process across the genome in 100 peak windows, sliding 10 peaks at a
920 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 down-sampled PacBio HiFi reads (60-fold) and ONT ultra-long reads (60-fold) where ultra-long reads were down-sampled in descending order based on read length to retain the longest reads (*16*). 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.

Genome assembly of ST001.

To create a de novo assembly for ST001, we used hifiasm (v0.19.9) with down-sampled PacBio HiFi reads selected to retain the longest 70-fold of reads. The initial PacBio HiFi dataset combined the Fiber-seq liver and lung reads from ST001. We then created consistent phasing by aligning the Fiber-seq reads (pbmm2 v1.13.0) back to the diploid assembly of ST001 and assigned haplotypes based on alignment to either the H1 or H2 assembly.

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, GM12878 peaks surrounding the MTMR1 TSS (chrX:150,631,401-150,738,995) and a fibroblast sample peak at the PDHA1 TSS (chrX:19,343,661- 19,343,996) were manually filtered due to substantial disagreement in between the variant and k-mer phasing methods. 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 (*74*). 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 (*75*). 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 (*76*) (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 (*77*). FIRE elements that completely overlapped modules two and three of a motif were centered using *fibertools* ft center (*26*), and the number of m6A observed within the motif was quantified. FIRE-contained motifs with <= 1 m6A were classified as bound.

Data availability

Processed epigenetic (FIRE) results for all samples are publicly available through https://doi.org/10.5281/zenodo.14511246, as well as the Fiber-seq data portal https://stergachislab.github.io/Fiber-seq-publication-data/. Per-sample information is provided in **Table S4**. All raw and processed sequencing data generated for GM12878 (PRJNA1233341) and K562 (SRX20077598, SRX20077599, and SRX20077600) have been submitted to the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/). Restrictions apply to the availability of some genetic data generated or analyzed during this study to preserve subject confidentiality. Genomic data for these samples are accessible to the scientific community through the UDN, GREGoR, All of US, and SMaHT consortia, with per-sample information and accessions provided in **Table S4**. Individuals interested in accessing this should submit a data access request to the relevant consortia. Cell lines obtained from the National Institute of General Medical Sciences Human Genetic Cell Repository at the Coriell Institute for Medical Research include GM12878.

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

A.B.S. is a co-inventor on a patent relating to the Fiber-seq method (US17/995,058).

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. We thank Alan Beggs, Monica Wojcik, Anne O'Donnell Luria, Gail Jarvik, and Katrina Dipple for providing the fibroblast cell lines. The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM12878.

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, 1U01HG013744, and OT2OD002748 to A.B.S., a Brotman Baty Institute Catalytic Collaboration Grant to A.B.S., NIH grants 1DP2AI183504 and 1U01AI176320 to J.P.R., as well as a Crohn's and Colitis Foundation Senior Research Award #1158945 to J.P.R. Additionally, M.R.V. and S.C.B. were supported by a training grant (T32) from the NIH (2T32GM007454-46). M.R.V. was also supported by a Pathway to Independence award from the National Institute of General Medical Sciences (1K99GM155552-01).

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

1. M. Jain, H. E. Olsen, D. J. Turner, D. Stoddart, K. V. Bulazel, B. Paten, D. Haussler, H. F. Willard, M. Akeson, K. H. Miga, Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).

2. A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, M. W. Hunkapiller, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, doi: 10.1038/s41587-019-0217-9 (2019).

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

4. M. Rautiainen, S. Nurk, B. P. Walenz, G. A. Logsdon, D. Porubsky, A. Rhie, E. E. Eichler, A. M. Phillippy, S. Koren, Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).

5. K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. A. Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, A. M. Phillippy, Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).

6. G. A. Logsdon, M. R. Vollger, P. Hsieh, Y. Mao, M. A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P. C. Dishuck, A. Rhie, L. G. de Lima, T. Dvorkina, D. Porubsky, W. T. Harvey, A. Mikheenko, A. V. Bzikadze, M. Kremitzki, T. A. Graves-Lindsay, C. Jain, K. Hoekzema, S. C. Murali, K. M. Munson, C. Baker, M. Sorensen, A. M. Lewis, U. Surti, J. L. Gerton, V. Larionov, M. Ventura, K. H. Miga, A. M. Phillippy, E. E. Eichler, The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).

7. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N.-C. Chen, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W.

Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

8. W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, S. Buonaiuto, X. H. Chang, H. Cheng, J. Chu, V. Colonna, J. M. Eizenga, X. Feng, C. Fischer, R. S. Fulton, S. Garg, C. Groza, A. Guarracino, W. T. Harvey, S. Heumos, K. Howe, M. Jain, T.-Y. Lu, C. Markello, F. J. Martin, M. W. Mitchell, K. M. Munson, M. N. Mwaniki, A. M. Novak, H. E. Olsen, T. Pesout, D. Porubsky, P. Prins, J. A. Sibbesen, J. Sirén, C. Tomlinson, F. Villani, M. R. Vollger, L. L. Antonacci-Fulton, G. Baid, C. A. Baker, A. Belyaeva, K. Billis, A. Carroll, P.-C. Chang, S. Cody, D. E. Cook, R. M. Cook-Deegan, O. E. Cornejo, M. Diekhans, P. Ebert, S. Fairley, O. Fedrigo, A. L. Felsenfeld, G. Formenti, A. Frankish, Y. Gao, N. A. Garrison, C. G. Giron, R. E. Green, L. Haggerty, K. Hoekzema, T. Hourlier, H. P. Ji, E. E. Kenny, B. A. Koenig, A. Kolesnikov, J. O. Korbel, J. Kordosky, S. Koren, H. Lee, A. P. Lewis, H. Magalhães, S. Marco-Sola, P. Marijon, A. McCartney, J. McDaniel, J. Mountcastle, M. Nattestad, S. Nurk, N. D. Olson, A. B. Popejoy, D. Puiu, M. Rautiainen, A. A. Regier, A. Rhie, S. Sacco, A. D. Sanders, V. A. Schneider, B. I. Schultz, K. Shafin, M. W. Smith, H. J. Sofia, A. N. Abou Tayoun, F. Thibaud-Nissen, F. F. Tricomi, J. Wagner, B. Walenz, J. M. D. Wood, A. V. Zimin, G. Bourque, M. J. P. Chaisson, P. Flicek, A. M. Phillippy, J. M. Zook, E. E. Eichler, D. Haussler, T. Wang, E. D. Jarvis, K. H. Miga, E. Garrison, T. Marschall, I. M. Hall, H. Li, B. Paten, A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

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

10. A. Guarracino, S. Buonaiuto, L. G. de Lima, T. Potapova, A. Rhie, S. Koren, B. Rubinstein, C. Fischer, Human Pangenome Reference Consortium, J. L. Gerton, A. M. Phillippy, V. Colonna, E. Garrison, Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023).

11. A. B. Stergachis, B. M. Debo, E. Haugen, L. S. Churchman, J. A. Stamatoyannopoulos, Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020).

12. I. Lee, R. Razaghi, T. Gilpatrick, M. Molnar, A. Gershman, N. Sadowski, F. J. Sedlazeck, K. D. Hansen, J. T. Simpson, W. Timp, Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).

13. N. J. Abdulhay, C. P. McNally, L. J. Hsieh, S. Kasinathan, A. Keith, L. S. Estes, M. Karimzadeh, J. G. Underwood, H. Goodarzi, G. J. Narlikar, V. Ramani, Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9**, 1–23 (2020).

14. Z. Shipony, G. K. Marinov, M. P. Swaffer, N. A. Sinnott-Armstrong, J. M. Skotheim, A. Kundaje, W. J. Greenleaf, Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat. Methods*, 1–9 (2020).

15. H. Grasberger, A. M. Dumitrescu, X.-H. Liao, E. G. Swanson, R. E. Weiss, P. Srichomkwun, T. Pappa, J. Chen, T. Yoshimura, P. Hoffmann, M. M. França, R. Tagett, K.

Onigata, S. Costagliola, J. Ranchalis, M. R. Vollger, A. B. Stergachis, J. X. Chong, M. J. Bamshad, G. Smits, G. Vassart, S. Refetoff, STR mutations on chromosome 15q cause thyrotropin resistance by activating a primate-specific enhancer of MIR7-2/MIR1179. *Nat. Genet.*, doi: 10.1038/s41588-024-01717-7 (2024).

16. M. R. Vollger, J. Korlach, K. C. Eldred, E. Swanson, J. G. Underwood, S. C. Bohaczuk, Y. Mao, Y.-H. H. Cheng, J. Ranchalis, E. E. Blue, U. Schwarze, K. M. Munson, C. T. Saunders, A. M. Wenger, A. Allworth, S. Chanprasert, B. L. Duerden, I. Glass, M. Horike-Pyne, M. Kim, K. A. Leppig, I. J. McLaughlin, J. Ogawa, E. A. Rosenthal, S. Sheppeard, S. M. Sherman, S. Strohbehn, A. L. Yuen, A. W. Stacey, University of Washington Center for Rare Disease Research, Undiagnosed Diseases Network, T. A. Reh, P. H. Byers, M. J. Bamshad, F. M. Hisama, G. P. Jarvik, Y. Sancak, K. M. Dipple, A. B. Stergachis, Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition. *Nat. Genet.*, doi: 10.1038/s41588-024-02067-0 (2025).

17. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

18. T. W. Tullius, R. S. Isaac, J. Ranchalis, D. Dubocanin, L. S. Churchman, A. B. Stergachis, RNA polymerases reshape chromatin and coordinate transcription on individual fibers, *bioRxivorg* (2023)p. 2023.12.22.573133.

19. D. Pellerin, G. F. Del Gobbo, M. Couse, E. Dolzhenko, S. K. Nageshwaran, W. A. Cheung, I. R. L. Xu, M.-J. Dicaire, G. Spurdens, G. Matos-Rodrigues, I. Stevanovski, C. K. Scriba, A. Rebelo, V. Roth, M. Wandzel, C. Bonnet, C. Ashton, A. Agarwal, C. Peter, D. Hasson, N. M. Tsankova, K. Dewar, P. J. Lamont, N. G. Laing, M. Renaud, H. Houlden, M. Synofzik, K. Usdin, A. Nussenzweig, M. Napierala, Z. Chen, H. Jiang, I. W. Deveson, G. Ravenscroft, S. Akbarian, M. A. Eberle, K. M. Boycott, T. Pastinen, All of Us Research Program Long Read Working Group, B. Brais, S. Zuchner, M. C. Danzi, A common flanking variant is associated with enhanced stability of the FGF14-SCA27B repeat locus. *Nat. Genet.* **56**, 1366–1370 (2024).

20. S. C. Bohaczuk, Z. J. Amador, C. Li, B. J. Mallory, E. G. Swanson, J. Ranchalis, M. R. Vollger, K. M. Munson, T. Walsh, M. O. Hamm, Y. Mao, A. Lieber, A. B. Stergachis, Resolving the chromatin impact of mosaic variants with targeted Fiber-seq. *Genome Res.*, doi: 10.1101/gr.279747.124 (2024).

21. C. J. Peter, A. Agarwal, R. Watanabe, B. S. Kassim, X. Wang, T. Y. Lambert, B. Javidfar, V. Evans, T. Dawson, M. Fridrikh, K. Girdhar, P. Roussos, S. K. Nageshwaran, N. M. Tsankova, R. P. Sebra, M. R. Vollger, A. B. Stergachis, D. Hasson, S. Akbarian, Single chromatin fiber profiling and nucleosome position mapping in the human brain. *Cell Rep. Methods* **4**, 100911 (2024).

22. G. A. Hartley, M. Okhovat, S. J. Hoyt, E. Fuller, N. Pauloski, N. Alexandre, I. Alexandrov, R. Drennan, D. Dubocanin, D. M. Gilbert, Y. Mao, C. McCann, S. Neph, F. Ryabov, T. Sasaki, J. M. Storer, D. Svendsen, W. Troy, J. Wells, L. Core, A. Stergachis, L. Carbone, R. J.

O'Neill, Centromeric transposable elements and epigenetic status drive karyotypic variation in the eastern hoolock gibbon, *bioRxivorg* (2024)p. 2024.08.29.610280.

23. D. Dubocanin, G. A. Hartley, A. E. Sedeno Cortes, Y. Mao, S. Hedouin, J. Ranchalis, A. Agarwal, G. A. Logsdon, K. M. Munson, T. Real, B. J. Mallory, E. E. Eichler, S. Biggins, R. J. O'Neill, A. B. Stergachis, Conservation of dichromatin organization along regional centromeres. *BioRxiv*, 2023.04. 20.537689 (2023).

24. K. L. Bubb, M. O. Hamm, J. K. Min, B. Ramirez-Corona, N. A. Mueth, J. Ranchalis, M. R. Vollger, C. Trapnell, J. T. Cuperus, C. Queitsch, A. B. Stergachis, The regulatory potential of transposable elements in maize, *bioRxivorg* (2024). https://doi.org/10.1101/2024.07.10.602892.

25. W. E. Fondrie, W. S. Noble, mokapot: Fast and Flexible Semisupervised Learning for Peptide Detection. *J. Proteome Res.* **20**, 1966–1971 (2021).

26. A. Jha, S. C. Bohaczuk, Y. Mao, J. Ranchalis, B. J. Mallory, A. T. Min, M. O. Hamm, E. Swanson, D. Dubocanin, C. Finkbeiner, T. Li, D. Whittington, W. S. Noble, A. B. Stergachis, M. R. Vollger, DNA-m6A calling and integrated long-read epigenetic and genetic analysis with fibertools. *Genome Res.*, doi: 10.1101/gr.279095.124 (2024).

27. L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, M. J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).

28. T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2016)*KDD '16*, pp. 785–794.

29. K. E. Taylor, K. M. Ansel, A. Marson, L. A. Criswell, K. K.-H. Farh, PICS2: next-generation fine mapping via probabilistic identification of causal SNPs. *Bioinformatics* **37**, 3004–3007 (2021).

30. M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, J. A. Stamatoyannopoulos, Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

31. J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, E. E. Eichler, Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).

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

33. A. J. Sharp, D. P. Locke, S. D. McGrath, Z. Cheng, J. A. Bailey, R. U. Vallente, L. M. Pertz, R. A. Clark, S. Schwartz, R. Segraves, V. V. Oseroff, D. G. Albertson, D. Pinkel, E. E.

Eichler, Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).

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

35. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

36. K. Wang, S. P. Dickson, C. A. Stolle, I. D. Krantz, D. B. Goldstein, H. Hakonarson, Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.* **86**, 730–742 (2010).

37. G. McInnes, Y. Tanigawa, C. DeBoever, A. Lavertu, J. E. Olivieri, M. Aguirre, M. A. Rivas, Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* **35**, 2495–2497 (2019).

38. A. Dilthey, C. Cox, Z. Iqbal, M. R. Nelson, G. McVean, Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).

39. P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, Diversity of human copy number. *Science* **11184**, 2–7 (2010).

40. P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, J. O. Korbel, An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

41. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

42. E. Heard, Recent advances in X-chromosome inactivation. *Curr. Opin. Cell Biol.* **16**, 247–255 (2004).

43. A. Loda, S. Collombet, E. Heard, Gene regulation in time and space during X-chromosome inactivation. *Nat. Rev. Mol. Cell Biol.* **23**, 231–249 (2022).

44. G. Borsani, R. Tonlorenzi, M. C. Simmler, L. Dandolo, D. Arnaud, V. Capra, M. Grompe, A. Pizzuti, D. Muzny, C. Lawrence, H. F. Willard, P. Avner, A. Ballabio, Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**, 325–329 (1991).

45. N. Brockdorff, A. Ashworth, G. F. Kay, P. Cooper, S. Smith, V. M. McCabe, D. P. Norris, G. D. Penny, D. Patel, S. Rastan, Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351**, 329–331 (1991).

46. C. J. Brown, A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, H. F. Willard, A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**, 38–44 (1991).

47. B. Panning, J. Dausman, R. Jaenisch, X chromosome inactivation is mediated by Xist RNA stabilization. *Cell* **90**, 907–916 (1997).

48. E. M. Darrow, M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S.-C. Huang, A. L. Sanborn, I. Machol, M. Shamim, A. P. Seberg, E. S. Lander, B. P. Chadwick, E. L. Aiden, Deletion of *DXZ4* on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4504-12 (2016).

49. G. Bonora, X. Deng, H. Fang, V. Ramani, R. Qiu, J. B. Berletch, G. N. Filippova, Z. Duan, J. Shendure, W. S. Noble, C. M. Disteche, Orientation-dependent Dxz4 contacts shape the 3D structure of the inactive X chromosome. *Nat. Commun.* **9**, 1445 (2018).

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

51. F. Yang, X. Deng, W. Ma, J. B. Berletch, N. Rabaia, G. Wei, J. M. Moore, G. N. Filippova, J. Xu, Y. Liu, W. S. Noble, J. Shendure, C. M. Disteche, The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol.* **16**, 52 (2015).

52. P. Bansal, Y. Kondaveeti, S. F. Pinter, Forged by DXZ4, FIRRE, and ICCE: How Tandem Repeats Shape the Active and Inactive X Chromosome. *Front Cell Dev Biol* **7**, 328 (2019).

53. B. P. Balaton, A. M. Cotton, C. J. Brown, Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol. Sex Differ.* **6**, 35 (2015).

54. M. Oliva, M. Muñoz-Aguirre, S. Kim-Hellmuth, V. Wucher, A. D. H. Gewirtz, D. J. Cotter, P. Parsana, S. Kasela, B. Balliu, A. Viñuela, S. E. Castel, P. Mohammadi, F. Aguet, Y. Zou, E. A. Khramtsova, A. D. Skol, D. Garrido-Martín, F. Reverter, A. Brown, P. Evans, E. R. Gamazon, A. Payne, R. Bonazzola, A. N. Barbeira, A. R. Hamel, A. Martinez-Perez, J. M. Soria, GTEx Consortium, B. L. Pierce, M. Stephens, E. Eskin, E. T. Dermitzakis, A. V. Segrè, H. K. Im, B. E. Engelhardt, K. G. Ardlie, S. B. Montgomery, A. J. Battle, T. Lappalainen, R. Guigó, B. E. Stranger, The impact of sex on gene expression across human tissues. *Science* **369** (2020).

55. T. D. Real, P. Hebbar, D. Yoo, F. Antonacci, I. Pačar, M. Diekhans, G. J. Mikol, O. G. Popoola, B. J. Mallory, M. R. Vollger, P. C. Dishuck, X. Guitart, A. N. Rozanski, K. M. Munson, K. Hoekzema, J. E. Ranchalis, S. J. Neph, A. E. Sedeño-Cortes, B. Paten, S. R. Salama, A. B. Stergachis, E. E. Eichler, Genetic diversity and regulatory features of human-specific NOTCH2NL duplications, *bioRxivorg* (2025)p. 2025.03.14.643395.

56. D. Dubocanin, G. A. Hartley, A. E. Sedeño Cortés, Y. Mao, S. Hedouin, J. Ranchalis, A. Agarwal, G. A. Logsdon, K. M. Munson, T. Real, B. J. Mallory, E. E. Eichler, S. Biggins, R.

J. O'Neill, A. B. Stergachis, Conservation of dichromatin organization along regional centromeres. *Cell Genom.* **5**, 100819 (2025).

57. G. A. Hartley, M. Okhovat, S. J. Hoyt, E. Fuller, N. Pauloski, N. Alexandre, I. Alexandrov, R. Drennan, D. Dubocanin, D. M. Gilbert, Y. Mao, C. McCann, S. Neph, F. Ryabov, T. Sasaki, J. M. Storer, D. Svendsen, W. Troy, J. Wells, L. Core, A. Stergachis, L. Carbone, R. J. O'Neill, Centromeric transposable elements and epigenetic status drive karyotypic variation in the eastern hoolock gibbon. *Cell Genom.* **5**, 100808 (2025).

58. P. Raj, E. Rai, R. Song, S. Khan, B. E. Wakeland, K. Viswanathan, C. Arana, C. Liang, B. Zhang, I. Dozmorov, F. Carr-Johnson, M. Mitrovic, G. B. Wiley, J. A. Kelly, B. R. Lauwerys, N. J. Olsen, C. Cotsapas, C. K. Garcia, C. A. Wise, J. B. Harley, S. K. Nath, J. A. James, C. O. Jacob, B. P. Tsao, C. Pasare, D. R. Karp, Q. Z. Li, P. M. Gaffney, E. K. Wakeland, Regulatory polymorphisms modulate the expression of HLA class II molecules and promote autoimmunity. *Elife* **5** (2016).

59. J. B. Kang, A. Z. Shen, S. Gurajala, A. Nathan, L. Rumker, V. R. C. Aguiar, C. Valencia, K. A. Lagattuta, F. Zhang, A. H. Jonsson, S. Yazar, J. Alquicira-Hernandez, H. Khalili, A. N. Ananthakrishnan, K. Jagadeesh, K. Dey, Accelerating Medicines Partnership Program: Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Network, M. J. Daly, R. J. Xavier, L. T. Donlin, J. H. Anolik, J. E. Powell, D. A. Rao, M. B. Brenner, M. Gutierrez-Arcelus, Y. Luo, S. Sakaue, S. Raychaudhuri, Mapping the dynamic genetic regulatory architecture of HLA genes at single-cell resolution. *Nat. Genet.* **55**, 2255–2268 (2023).

60. A. Alcina, M. D. M. Abad-Grau, M. Fedetz, G. Izquierdo, M. Lucas, O. Fernández, D. Ndagire, A. Catalá-Rabasa, A. Ruiz, J. Gayán, C. Delgado, C. Arnal, F. Matesanz, Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations. *PLoS One* **7**, e29819 (2012).

61. F. Megiorni, A. Pizzuti, HLA-DQA1 and HLA-DQB1 in Celiac disease predisposition: practical implications of the HLA molecular typing. *J. Biomed. Sci.* **19**, 88 (2012).

62. K. V. Good, J. B. Vincent, J. Ausió, MeCP2: The genetic driver of Rett syndrome epigenetics. *Front. Genet.* **12**, 620859 (2021).

63. N. A. Hathaway, O. Bell, C. Hodges, E. L. Miller, D. S. Neel, G. R. Crabtree, Dynamics and memory of heterochromatin in living cells. *Cell* **149**, 1447–1460 (2012).

64. B. Horsthemke, Epimutations in human disease. *Curr. Top. Microbiol. Immunol.* **310**, 45–59 (2006).

65. P. H. Sudmant, S. Mallick, B. J. Nelson, F. Hormozdiari, N. Krumm, J. Huddleston, B. P. Coe, C. Baker, S. Nordenfelt, M. Bamshad, L. B. Jorde, O. L. Posukh, H. Sahakyan, W. S. Watkins, L. Yepiskoposyan, M. S. Abdullah, C. M. Bravi, C. Capelli, T. Hervig, J. T. S. Wee, C. Tyler-Smith, G. Van Driem, I. G. Romero, A. R. Jha, S. Karachanak-Yankova, D. Toncheva, D. Comas, B. Henn, T. Kivisild, A. Ruiz-Linares, A. Sajantila, E. Metspalu, J. Parik, R. Villems, E. B. Starikovskaya, G. Ayodo, C. M. Beall, A. Di Rienzo, M. F. Hammer, R. Khusainova, E. Khusnutdinova, W. Klitz, C. Winkler, D. Labuda, M. Metspalu, S. A. Tishkoff, S. Dryomov, R. Sukernik, N. Patterson, D. Reich, E. E. Eichler, Global diversity, population stratification, and selection of human copy-number variation. *Science* **349** (2015).

66. D. Dubocanin, A. E. Sedeno Cortes, J. Ranchalis, T. Real, B. Mallory, A. B. Stergachis, Single-molecule architecture and heterogeneity of human telomeric DNA and chromatin, *bioRxiv* (2022)p. 2022.05.09.491186.

67. A. R. Quinlan, BEDTools: The Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1-34 (2014).

68. S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, J. A. Stamatoyannopoulos, An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).

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

70. F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, J. Köster, Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).

71. R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, M. A. DePristo, A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

72. J. M. Holt, C. T. Saunders, W. J. Rowell, Z. Kronenberg, A. M. Wenger, M. Eberle, HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* **40** (2024).

73. A. Rhie, B. P. Walenz, S. Koren, A. M. Phillippy, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

74. V. Akbari, J.-M. Garant, K. O'Neill, P. Pandoh, R. Moore, M. A. Marra, M. Hirst, S. J. M. Jones, Genome-wide detection of imprinted differentially methylated regions using nanopore sequencing. *Elife* **11** (2022).

75. M. Patterson, T. Marschall, N. Pisanti, L. van Iersel, L. Stougie, G. W. Klau, A. Schönhuth, WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).

76. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

77. M. Yin, J. Wang, M. Wang, X. Li, M. Zhang, Q. Wu, Y. Wang, Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res.* **27**, 1365–1377 (2017).

Figure 1

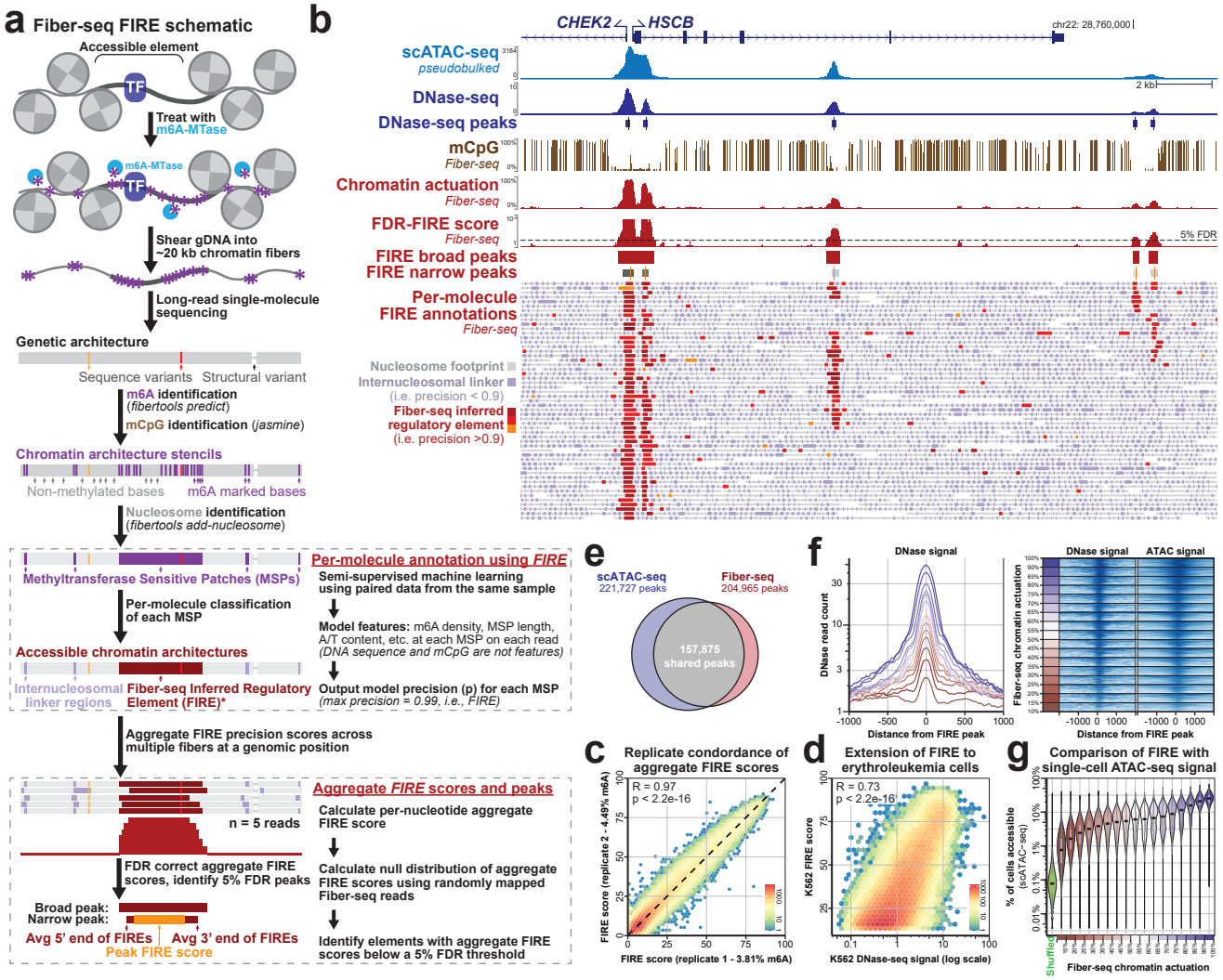

Figure 2

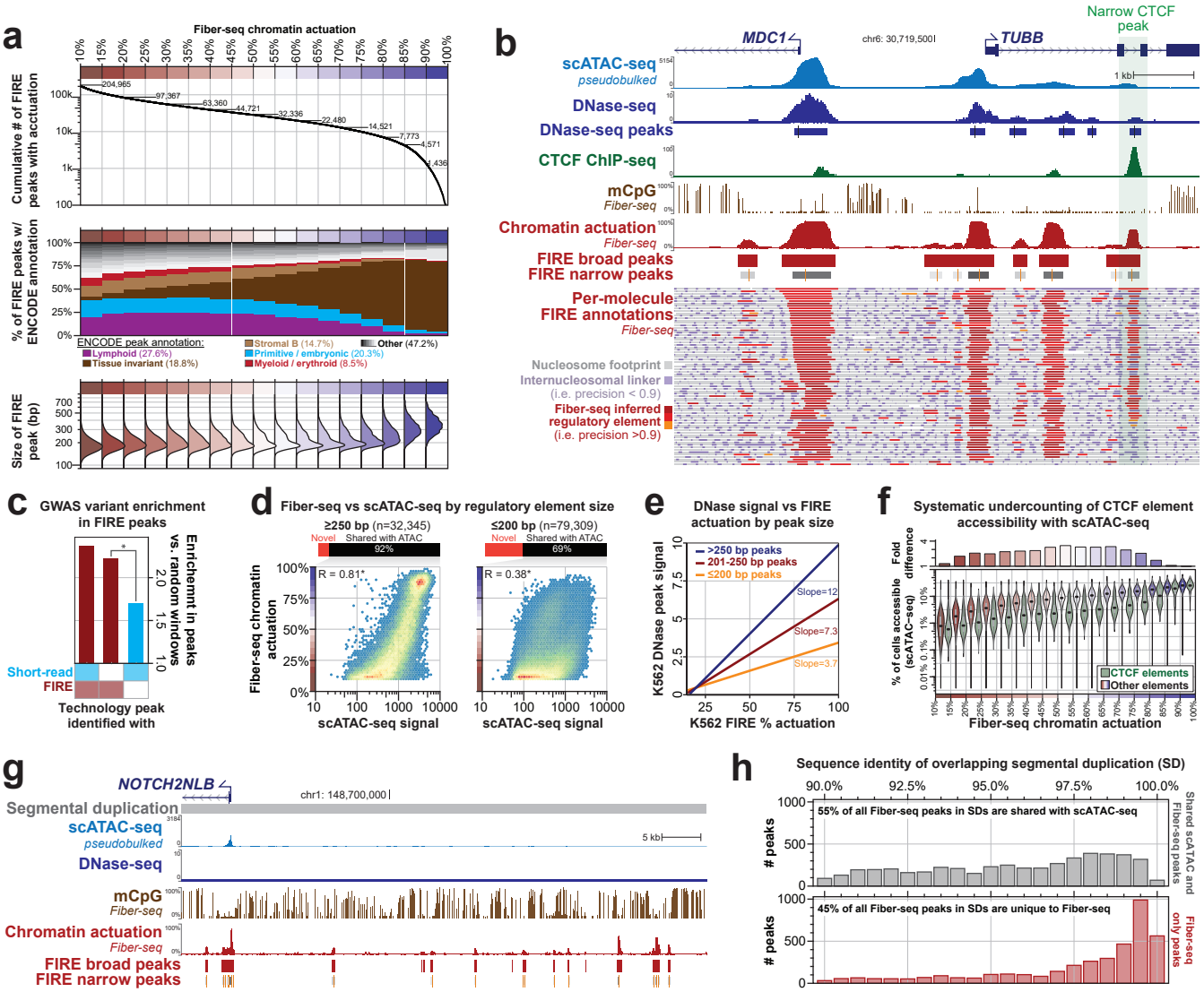

Figure 3

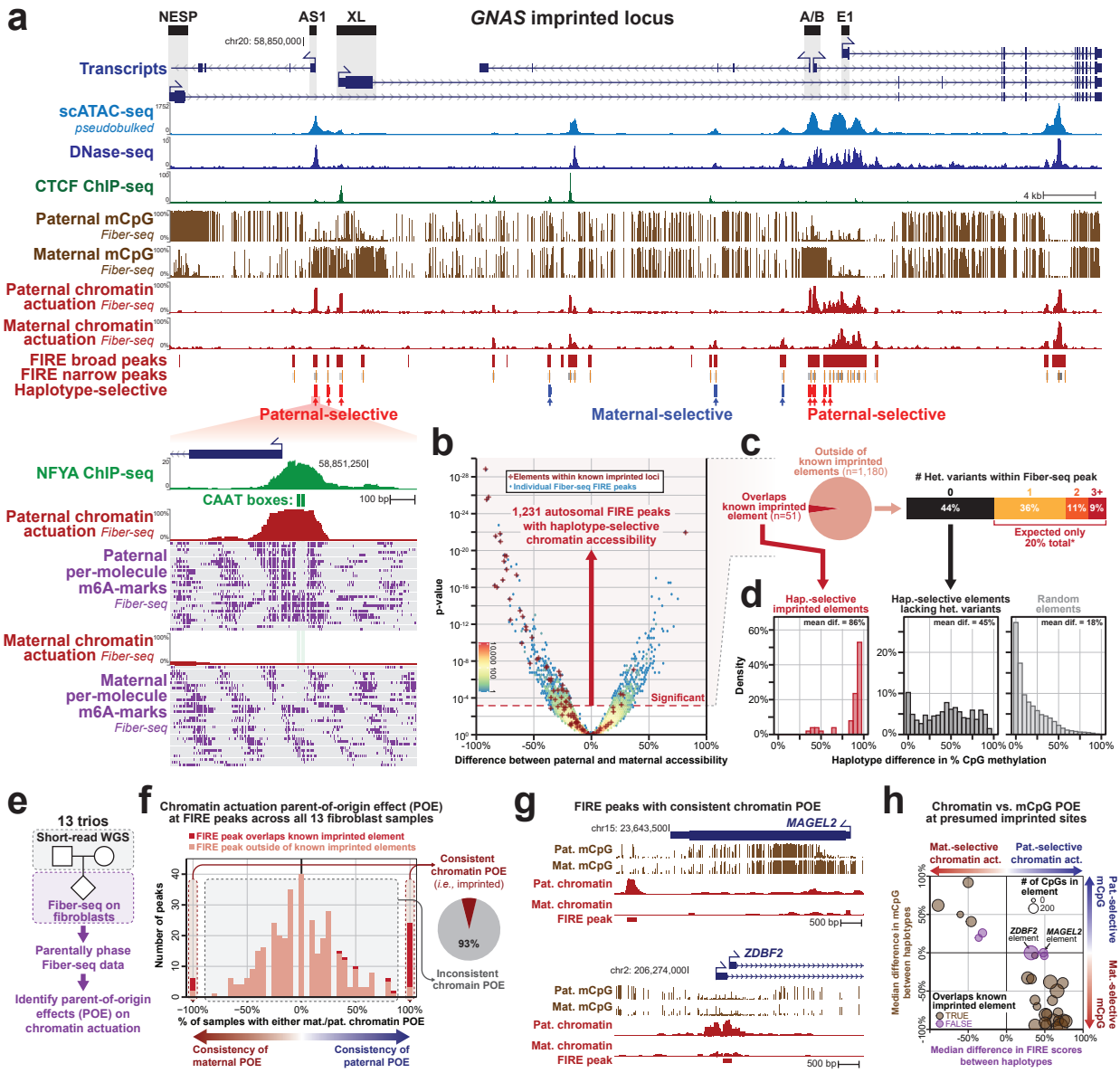

Figure 4

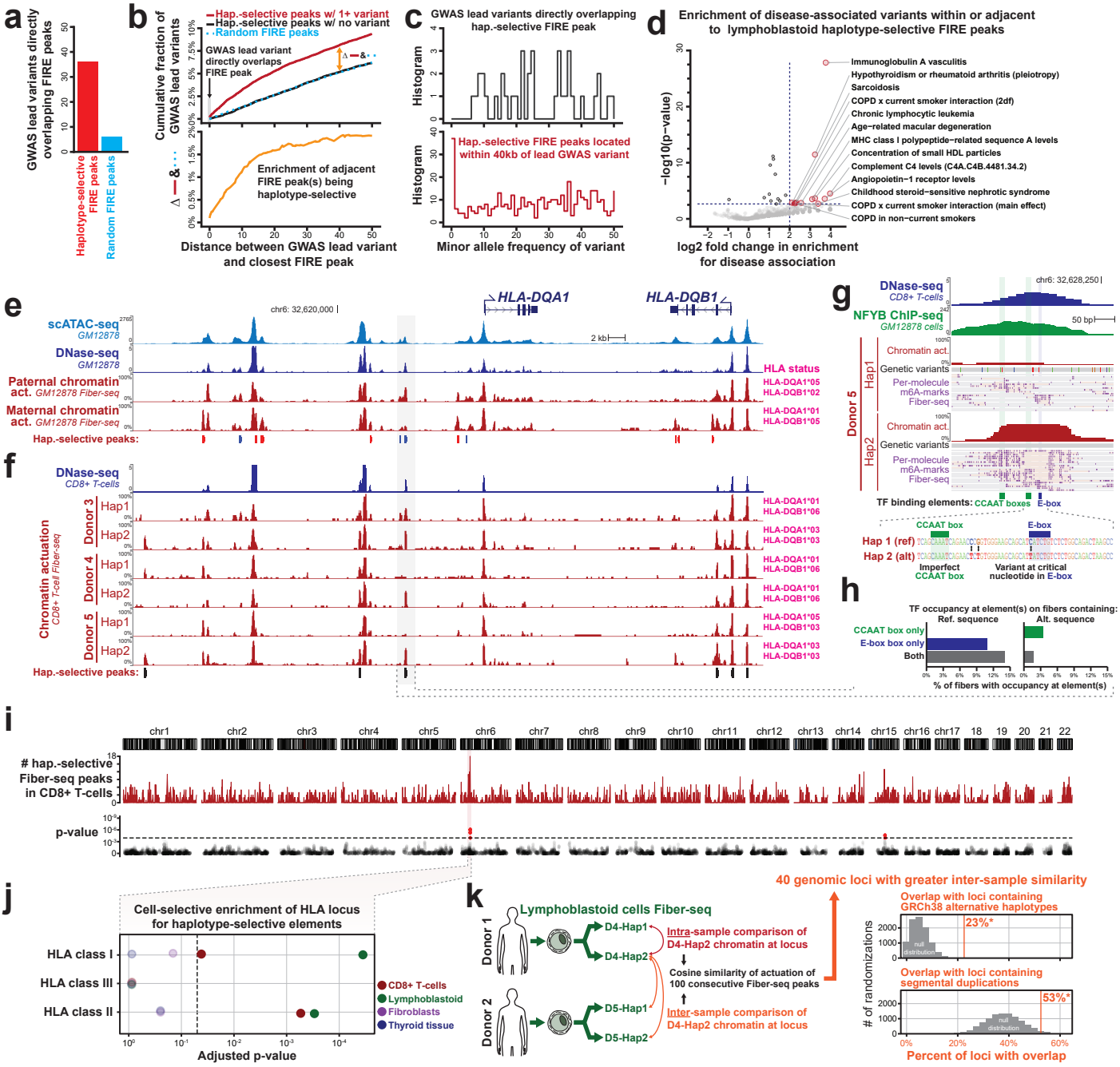

Figure 5

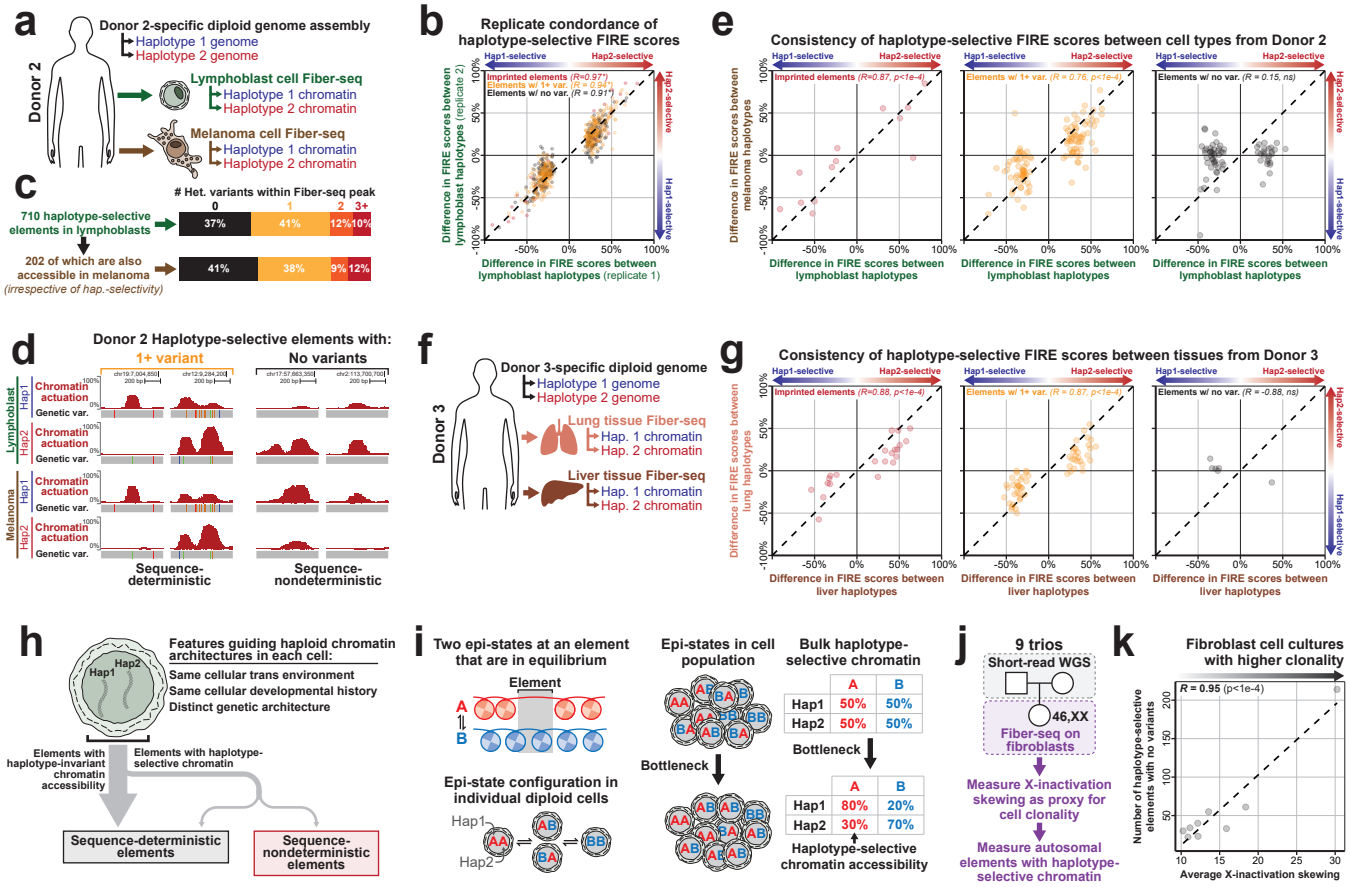

Figure 6